Non Stationarity and Market Structure Dynamics in Financial Time Series

Pier Francesco Procacci

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

February 18, 2023

Declaration

I, Pier Francesco Procacci, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Pier Francesco Procacci

Copyright

Abstract

This thesis is an investigation of the time changing nature of financial markets. Financial markets are complex systems having an intrinsic structure defined by the interplay of several variables. The technological advancements of the 'digital age' have exponentially increased the amount of data available to financial researchers and industry professionals over the last decade and, as a consequence, it has highlighted the key role of iterations amongst variables.

A critical characteristic of the financial system, however, is its time changing nature: the multivariate structure of the systems changes and evolves through time. This feature is critically relevant for classical statistical assumptions and has proven challenging to be investigated and researched. This thesis is devoted to the investigation of this property, providing evidences on the time changing nature of the system, analysing the implications for traditional asset allocation practices and proposing a novel methodology to identify and predict 'market states'.

First, I analyse how classical model estimations are affected by time and what are the consequential effects on classical portfolio construction techniques. Focusing on elliptical models of daily returns, I present experiments on both in-sample and out-of-sample likelihood of individual observations and show that the system changes significantly through time. Larger estimation windows lead to stable likelihood in the long run, but at the cost of lower likelihood in the short-term. A key implication of these findings is that the optimality of fit in finance needs to be defined in terms of the holding period. In this context, I also show that sparse models and information filtering significantly cope with the effects of non stationarity avoiding the typical pitfalls of conventional portfolio optimization approaches.

Having assessed and documented the time changing nature of the financial system, I propose a novel methodology to segment financial time series into market states that we call ICC - Inverse Covariance Clustering. The ICC methodology allows to study the evo-

lution of the multivariate structure of the system by segmenting the time series based on their correlation structure. In the ICC framework, market states are identified by a reference sparse precision matrix and a vector of expectation values. In the estimation procedure, each multivariate observation is associated to a market state accordingly to a minimisation of a penalized distance measure (e.g. likelihood, mahalanobis distance). The procedure is made computationally very efficient and can be used with a large number of assets. Furthermore, the ICC methodology allows to control for temporal consistency,S making it of high practical relevance for trading systems. I present a set of experiments investigating the features of the discovered clusters and comparing it to standard clustering techniques. I show that the ICC methodology is successful at clustering different states of the markets in an unsupervised manner, outperforming baseline standard models. Further, I show that the procedure can be efficiently used to forecast off-sample future market states with significant prediction accuracy.

Lastly, I test the significance of increasing number of states used to model equity returns and how this parameter relates to the number of observations and the time consistency of the states. I present experiments to investigate a) the likelihood of the overall model as more states are spanned, b) the relevance of additional regimes measured by the number of observations clustered. I found that the number of "market states" that optimally define the system is increasing with the time spanned and the number of observations considered.

Pier Francesco Procacci

Supervisor: Prof. Tomaso Aste

Impact Statement

This research contributes to the fields of financial time series analysis and complex systems and offers potential benefits both inside and outside academia.

A novel framework is proposed to deal with non-stationarity and to define, analyse and forecast market states. This offers significant advantages over standard approaches, as it can efficiently scale large multivariate datasets and enforces temporal consistency, opening up the field for further research on the role of correlation structure in modelling non-stationarity while managing the practical need for stable state definition.

The evidence presented in support of information filtering role in improving estimates and long-term model stability offers potential benefits in both academia and commercial activity as it can easily be integrated to improve classical and novel financial applications that rely on multivariate modelling of the financial system.

Outside academia, this research has potential impacts across the financial and investment management industry, professional practice, and public policy design. The ICC framework aims at helping portfolio managers and traders in tackling some of the main pitfalls of classical, widely used models. The information filtering approach can be integrated in most common financial applications, improving model accuracy and stability and significantly improving processing time, serving well the increasing low latency and high frequency needs. Lastly, the methods, experiments, and novel frameworks discussed throughout the thesis aim at improving the scalability of classical models and allow handling large multivariate datasets, which is a critical objective and current focus of many financial institutions.

Other than investment management, this research has the potential to impact public policy design and financial regulation. The proposed methods could help policy makers better understand and analyse the correlated nature of financial markets and help in designing policies that promote market stability and fairness.

Acknowledgements

Pursuing a PhD is a challenging and difficult task. I would like to overstate my gratitude to my PhD supervisor, Prof. Tomaso Aste. He made this research possible with his friendly support, patience and technical expertise. During the challenging lock-down times in London he has been much more than a supervisor to me, helping me when I needed the most and always encouraging me.

I would like to express my gratitude to UCL and the EPSRC for providing support and resources to conduct this research throughout the very unique conditions experienced in 2020/2021.

I would like to acknowledge Bloomberg and Carolyn Phelan for their technical help during the data acquisition phase at UCL.

I would like to thank Claudio ”the Boss” Marchetti for supporting me in taking the decision of leaving my country and a good job to take risks pursuing self fulfilment. He's been a second father to me and he shaped my personality in ways he cannot imagine.

All that happened during the last 5 years of my life would not have been possible without the help of Mariano Gambaro. My first boss and the first person trusting me and my skills in the financial industry when I was just a kid. The first mentor, selfless advisor and friend. Thank you.

This research has been for roughly 2/3rd conceived and written when London was restricted under Covid lock-downs. These have no doubt been challenging times for my physical and mental health and motivation, but it would have been much worse without the support of my friends and flatmates Pietro “Pit” Marone and Riccardo “Dino” Cesari.

I would like to thank Prof. Gianni Pola for his example, friendly suggestions and passion for the industry.

I would like to thank Citigroup, in particular the Equity Quant team headed by Chris Montagu and including Nik Joint, James Murray, Josie Gerken, Cosimo Recchia and David Chew for their support, patience and technical guidance.

Lastly, I would like to dedicate this thesis to my family, who showed love and support throughout my life. They have persuaded me to follow a PhD degree and have guided me continuously towards a solid education.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The objective of this chapter is to present an overview of this thesis by discussing the motivation behind the research problem, the objectives, experiments and contributions of this study and the structure of this thesis. The chapter starts by briefly introducing background information on stationarity and its implications for modelling purposes and suggesting that identifying homogeneous "states" or clusters can be a solution to this problem. The chapter then outlines the objectives, experiments, and contributions of this work and concludes with the thesis structure.

1.1 Motivation

Quantitative approaches to trading have widely developed over the last decades, with an increasing share of the institutional assets managed systematically. This practice is particularly acute for HFT Hedge Funds, with algorithmic high frequency trading systems alone accounting for more than 50% of US Equity trades [2]. This tendency has more recently paired with the vast availability of data, which is increasingly seen as the new information hedge and, therefore, leading financial institutions to build entire data departments to collect and store alternative data that may provide an hedge versus the wider market participants. In other words, systematic approaches that leverage the interaction among different variables is becoming the new gold race of financial modelling.

In this picture, systematic approaches to trading and portfolio allocation must be based on a robust and scalable multivariate modelling and forecasting of financial markets and the economy. The multivariate structure of the systems, however, is an entropic object that changes and evolves through time. This feature is critically relevant for classical statistical assumptions and has proven challenging to be investigated and researched.

So far, financial researchers have proposed different ways to tackle this problem, which

is becoming particularly acute with the emergence of machine learning based trading strategies. Some of the time series models proposed in literature try to account for these effects, but in most of cases these models become quickly unfeasible as the number of variables increases and, often, requires complex estimation procedures making inference inefficient in a multivariate context. Other models, like latent variable models, assume that returns are generate from a mixture of distributions and aim at modelling this effect by considering a latent state. These models are flexible and efficient to be estimated in high dimensions, but, in common formulations, do not describe the temporal dynamics characterizing financial time series. In the following chapter (Background and Literature Review) we review many of these attempts, highlighting their pros and cons in different scenarios.

The ambition of this thesis is to establish a modelling approach being able to capture different states of financial markets in an efficient way, while considering the temporal evolution of the data. We require the model to cope with high dimensionality, allowing to exploit the information content of the correlation structure.

1.2 Research Objectives

The main objectives of this research are:

1. Define and quantitatively measure the effects of non-stationarity on portfolio allocation, outlining drivers and impact on financial performances.

   While non-stationarity is a well known feature of financial markets, it is often difficult to measure its impact on trading systems and portfolio performances. The starting point of this research is a likelihood-based analysis playground to asses the evolution of the financial system and study of the impact of non-stationarity on the goodness of estimates and how it affects performances in common portfolio construction frameworks.

2. Investigate information filtering and sparsity in dealing with non-stationarity and study the role of these methods as a remedy against the entropic nature of the financial system.

3. Introduce a novel methodology to deal with the impact of non-stationarity on conven-

tional asset management practices, recognizing the importance of multivariate interactions and temporal consistency.

Only few research proposals directly tackle the change of distribution through time for the variables being modelled and most of the methods proposed in literature are highly exposed to the curse of dimensionality. Understanding the dependency structure of the many variables characterizing financial markets and its evolution with time is essential to capture the collective behaviour of the system. I propose an efficient methodology allowing to model the multivariate dynamics of markets, taking into account the correlation structure while delivering stable results which translates into minimizing transaction costs.

1.3 Research Experiments and Scientific Contribution

This research contributes to the existing literature in a number of ways:

1. Detailed exploration of the effects of non-stationarity on portfolio performances.

   I study the relationship between models likelihood and portfolio performances and present several experiments. First, I study the evolution of parameters' likelihood through time, in- and out- of sample, and how this depends on the estimation window length. Secondly, I analyse the impacts on portfolio performances and how the effects of different parameters likelihoods affect performances at different investment horizons.

2. Information Filtering in Financial Modelling.

   I explore the effect of information filtering and propose several experiments investigating the impact of filtering on a) the estimated parameters; b) portfolio construction and corresponding portfolio performances and c) model stability. Throughout the thesis, I outline how information filtering can improve estimates, avoid classical portfolio construction pitfalls leading to improved financial performance and improve long term model stability.

3. Proposal of novel methodology.

I propose a novel methodology called ICC where classification into states is constructed from a likelihood measure associate with a referential sparse precision matrix (inverse covariance matrix). We also enforce temporal coherence by penalizing frequent switches between market states and favouring model stability. Our approach simplifies and clarifies the definition of 'market state' by identifying each state with a sparse precision matrix and a vector of expectation values which are associated to a set of multivariate observations with largest adjusted likelihood. A sparse precision matrix provides an easily interpretable and intuitive structure of the market state with all the most relevant dependencies directly interconnected in a sparse network.

4. Comparison with industry standard estimation procedures and clustering methods.

   I compare the ICC methodology to commonly used clustering techniques used as baseline model and test the significance of the discovered clusters and the corresponding financial features, showing that the ICC methodology delivers financially meaningful clusters while traditional models do not.

5. Forecasting and Testing of established Machine Learning methods applied to a new application domain.

   A section of this thesis is devoted to forecasting market states, as identified and clustered using the ICC methodology. To this extent, I apply and test different machine learning and statistical methods in order to develop a framework for state based investing and daily trading. While the methods considered are known in literature, this is the first times they have been applied to this domain.

6. Number of States.

   Having defined a robust procedure to cluster observations into market states, I analyse how the likelihood of the overall model is impacted by the number of states spanned by the model. I present several experiments, both in- and out-of-sample and over different estimation and test windows testing how the likelihood is impacted overall and observation-wise. I found that the evolution of the financial system through time implies that the number of clusters that optimally describe the system increases through time as well: the more observations are considered and the larger the time window spanned, the higher the number of market states to be considered. These findings

support the statement that optimal estimation in finance is dependent on the holding period and time spanned.

7. Comprehensive US stock universe.

   We constructed a dataset of daily closing prices of 2490 US stocks entering among the constituents of the Russel 1000 index (*RIY index*) traded between 02/01/1995 and 31/12/2020. For each asset, we considered the corresponding daily log-returns. Looking at all the historical constituents has a number of advantages, including not exposing to survivorship bias and avoiding selection biases from index inclusion. Also, having taken as reference index the Russel 1000 allows us to consider about 90% of the total US market cap, avoiding size and liquidity biases.

8. Accurate generalized resampling procedures.

   In all of the experiments presented throughout this thesis, I used a randomised resampling procedure to avoid possible biases affecting the experiments presented and the conclusions drawn. The resampling procedure consists in sampling a fixed number of stocks (typically 100) at random and a random trading day spanned in our dataset. Train and test sets are then defined, respectively, using the observations prior and following to the randomly selected trading day. This procedure is then reiterated a fixed number of times (minimum 100), so that a different subset of stocks and different train and test sets are considered, spanning different market cycle phases.

1.4 Thesis Structure

The structure of this thesis is organised as follows:

- Chapter 2 - Background and Literature Review. The relevant literature and the key concepts in the areas of this research are reviewed in order to introduce the reader to the problems and methodological frameworks of this thesis. Our main goal is to highlight the current gap in the literature that this thesis aims to fill. Hence, I start with the most widely used approaches to financial market modelling and then focus on structural breaks and states models. After reviewing these traditional venues, I outline the role and information content of correlation structure and current research

on information filtering to efficiently leverage it.

- Chapter 3 - Portfolio Construction and Filtered Likelihood. I investigate the effects of non-stationarity on common maximum likelihood estimates used to describe financial assets features and further I analyse the effects of using such estimates as inputs for common asset allocation and trading practices.

- Chapter 4 - Market States and Stationarity Regimes. I introduce a novel methodology to define, analyse and forecast market states. After briefly reviewing the relevant literature, I introduce the novel ICC methodology, which efficiently allows to identify market states by means of a reference sparse precision matrix and a vector of expectation values while ensuring temporal consistency. The chapter also presents two experiments: in a first experiment I use the methodology to classify in-sample observations, using 100 assets and spanning 15 years of daily returns. The ICC model is compared to the common Gaussian Mixture model, delivering clusters significantly more homogeneous other than ensuring temporal consistency. Lastly, in a second example, I use the ICC methodology to study the short term market dynamics during Covid outbreak.

- Chapter 5 - Market States Forecasting. In this chapter, I present a set of experiments where I use the ICC methodology to forecast future states of the market form previous observations and assess the robustness of the forecast. In a second set of experiment, I present a simple trading strategy that times the market based on the forecasted state.

- Chapter 6 - Number of States. Having shown that we can efficiently identify states in which the financial system behaves differently, a natural question is how many states should we consider. In this Chapter, I investigate in-sample and out-of-sample likelihood of the parameters associated with each state as more states are considered and how the number of observations and time spanned impact the solution.

- Chapter 7 - Conclusions. The final chapter provides an overall conclusion of this research with a summary of the key findings of this work, and what can be learned from the results of its models and experiments. The thesis ends with our recommendations for future work to be done in this area.

1.5 Published Papers

In this section I list all the published as well as working papers related to this thesis:

[153] Procacci, P. F. and Aste, T. (2019). Forecasting market states. *Quantitative Finance*, 19(9):1491–1498

[154] Procacci, P. F. and Aste, T. (2022b). Portfolio construction and sparse multivariate modelling. *Journal of Asset Management*, 23(6):445–465

[155] Procacci, P. F., Phelan, C. E., and Aste, T. (2020). Market structure dynamics during COVID-19 outbreak. *ArXiv preprint 2003.10922*

[152] Procacci, P. and Aste, T. (2022a). States characterisation and evolution of the financial system. *Working Paper*

The research addressed in this thesis has been presented and discussed in the following seminar and conferences:

- EPSRC CDT Computer Science seminar series (18 April 2019). Non stationarity in financial time series. London.
- Citigroup Global Quantitative Research Conference (14 June 2019). Machine Learning in Financial Forecasting. Valencia.
- Financial computing and analytics seminar series, Department of Computer Science, UCL (8 April 2020). Regime detection in financial time-series and further results in portfolio construction. London.
- AIDA Trading international series (18 September 2020). Big Data and Systematic Investing. With C. A. Lehalle and P. Puggioni. London.
- Sole 24 ore BS Data Series (28 January 2022). On the impact of data availability on market practices. Milan.

Chapter 2

Background and Literature Review

In this chapter, I describe the body of scientific literature scrutinized by this research. The main goal is to provide a background for some key approaches that will be referenced along the thesis and to highlight the current gap in the literature that this research aims to fill. Hence, I start with the main modelling approaches for financial time series and introduce mixture models. I then review 'market states' models, and in particular structural breaks and Markov switching models. Lastly, I provide an overview on the role of the correlation structure, its persistency and information filtering networks.

2.1 Modelling financial time series

Modelling financial time series is most of the times centred on the study of returns, instead of prices. This is for two main reasons: *(a)* returns are a scale-free 'summary' of the investment decisions and *(b)* returns have far more attractive statistical properties than prices. Let S_t be the price of an asset at time t. Different definitions of returns may be considered. The log returns or *continuously compounded* returns r_t are defined as

$$r_t = \log(1+Z_t) = \log\frac{S_t}{S_{t-1}} = \log\frac{S_t}{C} - \log\frac{S_{t-1}}{C} \tag{2.1}$$

where log denotes the natural logarithm, Z_t the simple return at time t and C is a scale constant. In most of quantitative finance studies and applications, log returns are preferred to other definitions because they are more tractable (for example, multi-periods returns are given by the sum of single period returns) and because of their statistical properties [40, 170].

Considering, then, a panel $\mathbf{R}$ of T log returns for N assets $\{r_{i,t}; i = 1,...,N; t = 1,...,T\}$, some financial theories focus on the dynamic evolution over time for a single (univariate) asset, other theories emphasize the joint distribution of the N returns at a single time t. In

both cases, the most general approach to modelling is to consider the probability distribution describing the returns, under the common assumption of returns being independent and identical distributed (IID). In the following, I shall refer and consider example with univariate time series, but the material covered and the conclusions drawn are generally valid.

The distribution most commonly considered for log returns is the normal distribution with mean μ and variance σ^2. In this framework, simple returns are then IID lognormal random variables with mean and variance given by

$$\mathbb{E}(Z_t) = \exp(\mu + \frac{\sigma^2}{2}) - 1 \tag{2.2}$$

and

$$\mathrm{Var}(Z_t) = \exp(2\mu + \sigma^2)[exp(\sigma^2) - 1] \,. \tag{2.3}$$

Notice that the lower bound for simple return is -1.

However, in practical applications, the lognormal assumption is most of times inconsistent with historical stock returns. In particular, many stocks exhibit an excess kurtosis [170]. Alternatives are constituted by more complex distributions (not easily tractable) or the use of mixture models. For instance, the Gaussian mixture distribution can be written as linear superposition of K Gaussian distributions

$$p(\mathbf{R}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{R}; \mu_k, \Sigma_k) \,. \tag{2.4}$$

Each Gaussian density $\mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$ is called a *component* of the mixture and the parameters π_k are called *mixing coefficients*. We show in Appendix A.1 that these models can be formulated in terms of a discrete, unobservable latent variable. Mixture models maintain the tractability of simple distributions (*e.g.*, normal), have finite higher moments and capture excess kurtosis. Indeed, they provide an interpretation for the observed excess kurtosis: returns do not come from *one* distribution, but from a collection of (at least two) distributions. In latent formulations we assume that these different distributions are due to different, unobservable 'states' of the system, describing different regimes of interactions among variables.

2.1.1 Linear Time Series Models

Time series models have dominated the quantitative finance literature over the past decades. Indeed, focusing on the time evolution of the series, they provide a natural framework to

analyse the dynamic structure of returns. Let $\{r_t\}$ be a series of log returns treated as a discrete collection of random variables (stochastic process). A time series model for the observed data $\{r_t\}$ is a specification of the joint distribution of a sequence of random variables $\{R_t\}$ of which $\{r_t\}$ is one realisation.

The simplest time series models consider only the series $\{r_t\}$ and its information content and attempt to capture the linear relationship between r_t and the previous realizations $\{r_{t-1}\}$. A series r_t is said to be linear if it can be written as

$$r_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} \, . \tag{2.5}$$

where μ is the mean of r_t; ψ_i are the coefficients governing the dynamics of r_t and $\{\varepsilon_r\}$ is a series of IID random variables with zero mean, finite variance and known distribution (*i.e.* white noise series). ε_t is often referred as shock or innovation.

A well-known example of linear time series are the so called ARMA models [35, 179] that take the form

$$r_t = \sum_{p=1}^{P} \phi_p r_{t-p} + \varepsilon_t + \sum_{q=1}^{Q} \theta_{t-q} \varepsilon_{t-i} \tag{2.6}$$

where ϕ_p are the *autoregressive* (AR) coefficients; θ_q are the *moving average* (MA) coefficients and ε_t is a white noise series. ARMA processes describe a linear relationship between each observation r_t and previous observations together with previous shocks. Given this linear setting, it is straightforward to generalize the model in Eq. (2.5) to include other variables (*e.g.*, economic indicators) in the process while maintaining the same assumptions, in particular stationarity of the process and IID innovations.

Real financial time series are characterized by serial correlation and a dependency structure. Fitting an AR model accounts for the serial correlation of returns and has proven useful in many empirical application [38]. The dependency structure, instead, refers to the autocorrelation path that can be inferred from the autocovariance function of the squared returns r_t^2 or absolute returns $|r_t|$. This feature violates the stationarity conditions of linear models and requires further treatment. Moreover, as many authors show [46, 60, 77, 118], the behaviour of squared returns r_t^2 suggests the presence of clusters of volatility that need to be accounted for.

2.1.2 Stationarity

In order to make meaningful inference and forecasting on time series, a crucial role is played by stationarity. In loose terms, a process is stationary if its properties, or some of them, do not vary with time.

In particular, the time series $\{r_t\}$ is strictly stationary iff

$$(r_1, ..., r_n) \stackrel{d}{=} (r_{1+h}, ..., r_{n+h}) \tag{2.7}$$

for all integers h and $n \geq 1$ and where $\stackrel{d}{=}$ indicates that the two random vectors have the same joint distribution function.

The time series $\{r_t\}$ is weakly stationary iff

a) $\mathbb{E}(r_t) = \mu$ is independent of t ,

and (2.8)

b) $\text{Cov}(r_t, r_{t-l}) = \gamma_l$ is independent of t for each l ,

where l is an arbitrary integer. par

Intuitively, a strictly stationary time series maintains its multivariate structure constant through time, while a weakly stationary time series maintain only some key properties. Time series analysis consists on finding a mathematical structure describing the evolution of the phenomenon being studied, estimating the parameters governing such dynamics and using it to draw conclusions and make predictions on the phenomenon itself. Not surprisingly, if we wish to make predictions, then clearly we must assume that something does not vary with time. Thus, most of the classical time series models assume at least a weakly stationary process.

2.1.3 Non-Linear Time Series Models

Non-linear models are motivated by the need to reflect properties or stylized features of time series that violate the assumptions in linear models. In financial time series, these properties include tail heaviness, asymmetry and serial dependence. Also financial assets' volatility presents some commonly observed characteristics, including volatility clustering (i.e. volatility tends to be high for certain time periods and low for other periods) and asymmetric reaction to large price increases and decreases (referred to as the leverage effect). These properties play an important role in modelling financial time series and require

flexible structures that can extend the linear assumption.

A general non-linear model can be formulated in a convenient way in terms of its conditional moments [170] and I shall consider this formulation in the remaining of this section.

Let F_{t-1} be σ-field set of information available at time $t-1$. Typically, F_{t-1} is a series of linear combination of $\{r_{t-1}\}$ and $\{\varepsilon_{t-1}\}$. Then the conditional mean and variance of r_t given F_{t-1} are

$$\mu_t = \mathbb{E}(r_t|F_{t-1}) = g(F_{t-1}), \qquad \sigma_t^2 = \mathrm{Var}(r_t|F_{t-1}) = h(F_{t-1}) \tag{2.9}$$

where $g(\cdot)$ and $h(\cdot)$ are non-linear functions with $h(\cdot) > 0$. Thus, we can formulate a general model with nonlinearities restricted to mean and variance functions as

$$r_t = g(F_{t-1}) + \sqrt{h(F_{t-1})}a_t \tag{2.10}$$

where $a_t = \varepsilon_t/\sigma_t$. Notice that that the linear model in Eq. (2.5) is obtained if $g(\cdot)$ is a linear function of the elements in F_{t-1} and $h(\cdot) = \sigma_\varepsilon^2$. For non-linear $g(\cdot)$, the model is said to be non-linear in mean, whereas if $h(\cdot)$ is non-linear (*e.g.*, time variant), the model is said to be non-linear in variance.

To account for the dependency structure and volatility clustering, Bollerslev (1990) proposed the generalized autoregressive conditional heteroskedasticity (GARCH) model. Considering a log returns series r_t of the form in Eq. (2.10) and letting $\varepsilon_t = r_t - g(F_t)$, then ε_t follows a GARCH(*m*,*s*) model if

$$\varepsilon_t = \sigma_t a_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i \varepsilon_{t-t}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}\,. \tag{2.11}$$

The model is stable provided that $\sum_{i=1}^{\max(m,s)}(\alpha_i + \beta_i) < 1$ implying that the unconditional variance of ε_t is finite, while the conditional variance evolves through time. The model is coherent with the volatility clusters previously described and provides a simple parametric formulation to describe the volatility evolution. It is worth empathizing, however, that model in Eq. (2.11) is a univariate formulation. When considering a multivariate setting, not only the variances, but also the correlations among variables must be modelled. In particular, for N assets being modelled, there are $N(N-1)/2$ correlations to be considered other than N variances. This clearly shows how the GARCH model suffers the curse of

dimensionality in a multivariate context, as discussed in Section 2.2.

Many non-linear time series models have been proposed in literature with the underlying idea of modelling the conditional mean based on a parametric function. It is the case of, for example, the state-dependent model [151], the bilinear models [76] or deterministic dinamycal linear systems [91].

2.1.4 Market States

In financial markets, changes of regimes are often caused by factors different from the variables being modelled (*e.g.*, economics variables, political factors, ecc...). For this reason, regimes or states are often considered unobservable or referred as latent and this is reflected in the time-varying nature of parameters. Modelling regimes has, therefore, a great appeal from an economic perspective.

Many time series models presented in literature tried to describe this phenomenon. Among the most well-known, it is worth mentioning the TAR model of Tong (1978) and the the Markov switching model by Hamilton (1989). In TAR models, the goal is to estimate two sets of parameters corresponding to different regimes segmented by means of a threshold 'k' referred as *structural break*. Formally, this is modelled by means of a piecewise-linear autoregression of the form

$$r_t = \begin{cases} \alpha_1 + \beta_1 r_{t-1} + \varepsilon_t & \text{if } r_{t-1} < k \\ \alpha_2 + \beta_2 r_{t-1} + \varepsilon_t & \text{if } r_{t-1} \geq k \end{cases} \tag{2.12}$$

However, unlike typical engineering application, we cannot say with certainty when a structural break has occurred in economic time series and the prior knowledge of major economic events could lead to bias in inference [40].

The Markov switching model, instead, models the change in regime by means of an unobserved state variable which is typically modelled as a Markov chain:

$$r_t = \begin{cases} \alpha_1 + \beta_1 r_{t-1} + \varepsilon_{1,t} & \text{if } s_t = 1 \\ \alpha_2 + \beta_2 r_{t-1} + \varepsilon_{2,t} & \text{if } s_t = 2 \end{cases} \tag{2.13}$$

where s_t is an unobservable Markov chain with some transition probability $\mathbf{P}$. Hamilton (1989) proposes an estimation method based on the EM algorithm. However, for slightly more complex dynamics considered, we need to rely on variational inference

techniques or MCMC methods [170]. This implies that, in a multivariate context and particularly if we aim to extract information on the switching from the correlation structure, estimation becomes difficult to perform.

More recently, advances in computational techniques and computing power have allowed researchers to investigate market regimes in a multivariate context, using the information and facing the computational problem of the correlation structure. Next Section is devoted to a review of these techniques.

2.2 Correlation Structure

Understanding the correlation structure of financial returns has proven crucial for a wide range of applications such as risk management [7, 45, 51, 62, 79, 96, 168], option pricing [32, 61, 114, 185] and asset allocation [25, 28, 121, 183].

Most popular approaches in the industry assume - for convenience - a stationary correlation structure [28, 62]. However, it is well established that correlations among stocks are not constant over time [4, 113, 138] and increase substantially in periods of high market volatility, with, asymmetrically, larger increases for downward moves (see, for example, [6, 45, 162]). Being able to predict future correlation structure would provide very powerful tools for risk management, option pricing and asset allocation. Indeed, various approaches have been proposed in the literature to model and predict time-varying correlations. Examples are, for instance, the generalized autoregressive conditional heteroskedasticity (GARCH) models [30] described in previous section or the Dynamic Conditional Correlation (DCC) model by Engle (2002). However, most of these models are not able to cope with more than a few assets due to the curse of dimensionality having number of parameters that increases super-linearly with the number of variables [54]. Other approaches have been focusing on the study of changes in a time-varying correlation matrix computed from a rolling window. This is, for instance, the case of estimators like the RiskMetrics [116] or [110]. However, since these approaches use only a small part of the data, these estimators have large variances and, in case of high dimensionality, may lead to inconclusive estimates [102]. [134] proposed a comparison between correlation matrices from different windows by computing a relative distance between these time-varying correlation matrices.

This approach demonstrated that market have patterns in time that are persistent and sometime recurrent. Other approaches [70, 75, 81, 87, 184] considered instead a segmentation of the observation window by assigning each multivariate observation at each time instance t to a cluster accordingly to a distance metric.

2.2.1 Information Filtering Network

Financial markets are complex systems and, as such, are characterized by the interaction of many elements. As previously noted, understanding the dependency structure of these variables and its evolution with time is essential to capture the collective behaviour of the systems. One possible approach to represent the set interactions in a complex system is a network structure where the vertices are the system's elements and edges between vertices indicate the interactions between the corresponding elements. In the extraction of information from observed correlation, two major challenges are faced: *(a)* observed correlations are often spurious and subject to random fluctuations making uncorrelated events to appear correlated and vice-versa. This phenomenon is referred as 'noise dressing' [73, 102]; *(b)* In correlation-based graphs and in the absence of any filtering procedure, all links among elements are present. This is likely to contain redundant and less-relevant information that do not provide valuable insight, other than making the computations less efficient and exposing the estimates to overfitting [15, 27, 106].

Information filtering networks aims at retrieving the *relevant* sub network of interactions among the elements of the system. In the pioneering work of Mantegna (1999), the author proposed to investigate financial systems by the extraction of a minimal set of relevant interactions associated with the strongest correlations belonging to the Minimum Spanning Tree (MST). The MST structure is, however, a drastic filtering tool and is likely to discard valuable information. Tumminello et al. (2005) and Aste and Di Matteo (2006) considered the geometrical and topological structure associated to the network to be constructed. In particular, they show that graphs of different complexities can be constructed by iteratively linking the most strongly connected nodes under the constraint of generating *planar* graphs, obtaining a structure defined Planar Maximally Filtered Graph (PMFG).

There is now a large body of literature proving network filtering to be a powerful tool to associate a sparse network to a high-dimensional dependency measure with applications ranging from financial markets [15] to biological systems [166] and econophysics [119].

2.2.2 Triangulated Maximally Filtered Graph - TMFG-LoGo

The Triangulated Maximal Filtered Graph (TMFG) is a family of information filtering networks introduced in Massara et al. (2015a), and Massara et al. (2017). These are planar graphs, but with the advantage of being *decomposable* graphs, other than being generated in a computationally efficient way. Decomposable graphs are clique forests, made of n cliques connected by separators (cliques of smaller size). Decomposable graphs have the property that, when the vertices of the separators are disconnected, the graph becomes divided into into disconnected components.

Barfuss et al. (2016) show how, given the decomposable graphical structure, we can produce global sparse inverse covariance matrices from a sum of local inversions. In particular, let's consider a graph $\mathcal{G}$ made of N_c cliques C_n, with $M = 1,...,M_c$ and N_s separators S_m, with $m = 1,...,N_s$. In $\mathcal{G}$, the p vertices represent the variables $\mathbf{X} = (X_1, X_2, ..., X_p)$ and the edges represent the couple of conditionally dependent variables. Given this network, one can write the joint probability density function $f(\mathbf{X})$ of the set of p variables $\mathbf{X}$ by means of the factorization [106]

$$f(\mathbf{X}) = \frac{\prod_{n=1}^{N_c} f_{C_n}(\mathbf{X_{C_n}})}{\prod_{m=1}^{N_s} f_{S_m}(\mathbf{X_{S_m}})^{k(S_m)-1}} \tag{2.14}$$

where f_{C_n} and f_{S_m} are the marginal density functions of the variables in C_m and S_m; the term $k(S_m)$ is the numerosity of the disconnected components obtained by removing the separators S_m. Equation (2.14) is a consequence of the Bayes theorem and, as such, is generally valid and applicable. Looking for the functional form of $f(\mathbf{X})$ using the maximum entropy method [93], we obtain

$$f(\mathbf{X}) = \frac{1}{Z} \exp\left(-(\mathbf{X}-\boldsymbol{\mu})\mathbf{J}\,(\mathbf{X}-\boldsymbol{\mu})^{\top}\right) \tag{2.15}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the vector of expectation values; the elements of $\mathbf{J} \in \mathbb{R}^{pxp}$ are the Lagrange multipliers associated with the second moments of the distribution that are the coefficient of the covariance matrix $\boldsymbol{\Sigma}$ and $'$ is the transpose operator. Notice that if we want to reproduce all the second moments $\Sigma_{i,j}$, then the solution for the distribution parameters is $\mathbf{J} = \boldsymbol{\Sigma}^{-1}$. In the multivariate normal case, it follows from Eq. (2.14) that the network $\mathcal{G}$ coincides with the structure of non zero coefficients of $\mathbf{J}$ in Eq. (2.15) and the elements $J_{i,j}$ can be computed by considering the local inversion of the covariance matrices associated

with separators and cliques

$$\mathbf{J}_{i,j} = \sum_{C \ s.t. \ \{i,j\} \in C} \left(\mathbf{\Sigma}_{\boldsymbol{C}}^{-1}\right)_{i,j} - \sum_{S \ s.t. \ \{i,j\} \in S} (k(S)-1) \left(\mathbf{\Sigma}_{\boldsymbol{S}}^{-1}\right)_{i,j} \tag{2.16}$$

with $\mathbf{J}_{i,j} = 0$ if i, j are not both part of a common clique. Equation (2.16) reduces the global problem of a p x p matrix inversion into a sum of local inversions of matrices of size of the separators and cliques (max three or four for TMFG graphs [124]). This implies that to obtain a nonsingular global estimate of the inverse covariance four observations would be enough. The moments $J_{i,j}$ to be retained (*i.e.*, non-zero) are chosen in order to maximize the likelihood associated with the multivariate distribution 2.15. For all decomposable graphs, to maximize the log likelihood associated with the distribution 2.15, only $\log|\mathbf{J}|$ needs to be maximized [106]

$$\log|\mathbf{J}| = \sum_{m=1}^{N_s} [k(s)-1] \ \log|\hat{\Sigma}_{S_n}| - \sum_{m=1}^{N_c} [k(s)-1] \ \log|\hat{\Sigma}_{C_n}| \ . \tag{2.17}$$

The TMFG-LoGo construction starts with a tetrahedron $C_1 = \{v_1, v_2, v_3, v_4\}$ with smallest correlation determinant $|\hat{\mathbf{R}}_C|$. Then are iteratively introduced, inside the existing triangular faces, the vertex the maximizes $|\hat{\mathbf{R}}_S| - \log|\hat{\mathbf{R}}_C|$, where S and C are the *new* separators and cliques produced by the vertex insertion. Algorithm 1 reports the TMFG-LoGo construction procedure outputting a decomposable graph which is constituted by four-clique connected with three-clique separators.

Algorithm 1 TMFG-LoGo algorithm

Input
$\hat{\Sigma} \in \mathbb{R}^{p \ x \ p}$, covariance matrix estimated from a set of observations $\mathbf{X}$
$\hat{\mathbf{R}} \in \mathbf{R}^{p \ x \ p}$, correlation matrix associated with $\hat{\Sigma}$

Initialize
$\mathbf{J}$ = array of p x p zeros
C_1 = Tetrahedron, v_a, v_b, v_c, v_d with smallest $|\hat{\mathbf{R}}_{\mathbf{C_1}}|$
$\mathcal{T}$ = assign the four triangular faces in C_1
$\mathcal{V}$ = assign the remaining $p-4$ vertices not in C_1
while $\mathcal{V}$ is not empty **do**
find the combination $\{v_a, v_b, v_c, v_d \in \mathcal{T}$ and $v_d \in \mathcal{V}$ with the largest $|\hat{\mathbf{R}}_{\mathbf{v_a},\mathbf{v_b},\mathbf{v_c}} / \hat{\mathbf{R}}_{\mathbf{v_a},\mathbf{v_b},\mathbf{v_c},\mathbf{v_d}}|$
Remove v_d from $\mathcal{V}$
Remove $\{v_a, v_b, v_c\}$ from $\mathcal{T}$
Add $\{v_a\}$, $\{v_b, v_c\}$, $\{v_a, v_c, v_d\}$, $\{v_b, v_c, v_d\}$ to $\mathcal{T}$
Compute $J_{i,j} = J_{i,j} + \left(\mathbf{\Sigma}^{-1}_{\{v_a,v_b,v_c,v_d\}}\right)_{i,j} - \left(\mathbf{\Sigma}^{-1}_{\{v_a,v_b,v_c\}}\right)_{i,j}$
return $\mathbf{J}$, sparse estimation of $\hat{\Sigma}^{-1}$

2.2.3 Persistence of Correlation Structure

Financial time series are characterized by volatility clustering. This phenomenon has been assessed and documented by many authors (see Section 2.1.1) with many corresponding modelling proposals. Recently, some authors analysed a similar clustering or persistence effect in the correlation structure [13, 21, 138, 139] offering valuable insights to multivariate analysis and forecasting of financial time series.

Information filtering networks provide a valuable tool to the study of this effect. In this section we used a persistence measure proposed by Musmeci et al. (2016b) to investigate correlation persistence in the data panel of interest. We make use of the TMFG networks structure described previous Section and constructed using the TMFG-LoGo procedure (Algorithm 1).

We considered the whole dataset length, between 01/02/1995 and 12/31/2015, for a subset of 100 stocks chosen at random among those that have been continuously traded throughout the observed period. To calculate the correlation between different variables and to analyse its evolution through time, we considered n rolling time windows T_a, with $a = 1, ..., n$. Each time window contains θ log returns for each asset. Within each rolling window, we calculated the the correlation matrix $\rho(T_a)$ using an exponential smoothing method with $\alpha = 0.99$ smoothing factor [116, 149]. From each correlation matrix $\rho(T_a)$ we computed the corresponding TMFG network, obtaining n TMFGs, $G(T_a)$ with $a = 1, ..., n$. To analyse the evolution of the correlation structure, we considered a persistence measure $\langle ES \rangle(T_a)$ [138] built from the graph's edges *survived* at each rolling window T_a

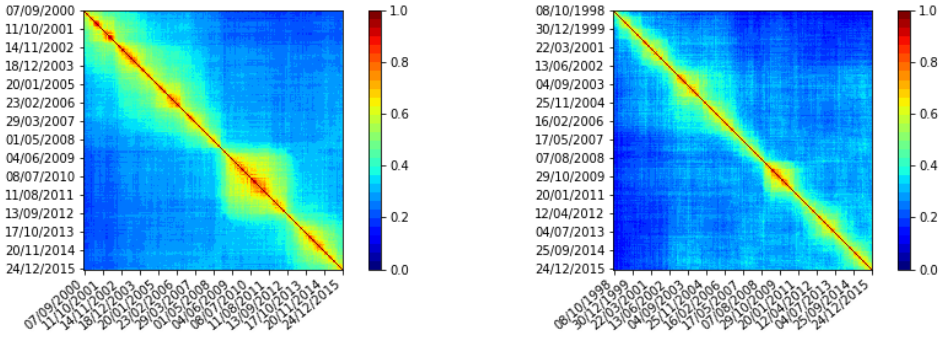
$$\langle ES \rangle(T_a) = \sum_{b=a-L}^{a-1} w(T_b) ES(T_a, T_b) \tag{2.18}$$

where $w(T_b) = \exp\left(\frac{b-a-1}{L/3}\right)$; L is a parameter and $ES(T_a, T_b)$ is the *edges survival ratio* measured as the fraction of edges in common between $G(T_a)$ and $G(T_b)$

$$ES(T_a, T_b) = \frac{1}{N_{\text{edges}}} |E^{T_a} \cup E^{T_b}| \tag{2.19}$$

where N_{edges} is the number of edges in the two graphs (fixed and equal to $3(N-2)$ for TMFG) and E^{T_a}, E^{T_b} are the edge-set of the graphs at T_a, T_b.

$\langle ES \rangle(T_a)$ relies on past data and indicates how slowly the correlation structure at time window T_a is different from previous time windows. This is a weighted average of similarity



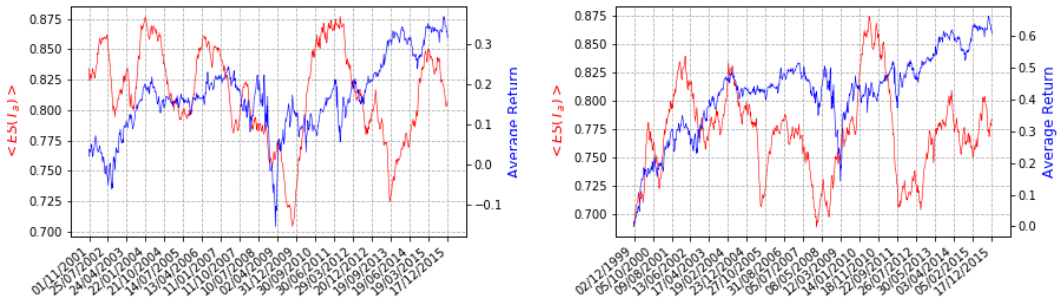
(a) $ES(T_a, T_b)$, $\alpha = 0.99$ and $\theta = 1000$ **(b)** $ES(T_a, T_b)$, $\alpha = 0.99$ and $\theta = 500$

Figure 2.1: Edge survival ratio, $ES(T_a, T_b)$. EWMA smoothing parameter $\alpha = 0.99$ and $\theta = 1000$ (left) and $\theta = 500$ (right). This matrix-form representation provides a visual of the persistent correlation clusters observed from 1999 to 2015.

(edge survival ratio) between $G(T_a)$ and previous L TMFG networks, weighted following an exponential smoothing scheme to give more weight to networks closer to T_a. We computed the measure using time windows T_a of different length θ and we report the results for $\theta = 1000$ and $\theta = 500$. In general, for higher values of θ, the network estimation is more robust, but $\langle ES \rangle (T_a)$ is less reactive.

Figure 2.1 presents the edge survival ratio $ES(T_a, T_b)$ computed between each pair of window T_a,T_b with $a, b = 1, ..., n$; $a \neq b$. The figure shows that there are clear similarity clusters, with higher similarity values in correspondence of crisis periods (2002-2003, 2005-2006, 2009-2012). This is particularly clear from panel(a), where $\theta = 1000$ estimation window provide a clearer representation.

Figure 2.2 shows the evolution of $\langle ES \rangle (T_a)$ in time as compared to the cumulative average return of the stocks considered. We found high levels of persistence in correspondence of crisis events with peaks of $\langle ES \rangle (T_a) = 0.88$. As expected, $\langle ES \rangle (T_a)$ computed with $\theta = 500$ (panel(b)) is more volatile and reactive, but similar evidences are obtained with $\theta = 1000$ (panel(a)). We make use of this feature in the definition of the trading strategy described in Chapter 5.2.



(a) $< ES > (T_a), \theta = 1000, \alpha = 0.99$ **(b)** $< ES > (T_a), \theta = 500, \alpha = 0.99$

Figure 2.2: Correlation structure persistence. $< ES > (T_a)$ with EWMA smoothing parameter $alpha = 0.99$ and $\theta = 1000$ (left) and $\theta = 500$ (right)

Chapter 3

Portfolio Construction and Filtered Likelihood

Portfolio optimization approaches rely on multivariate-modelling of markets and the economy. In this section, we address three sources of error related to the modelling of these complex systems: 1. oversimplifying hypothesis; 2. parameters' sampling-error; 3. intrinsic non-stationarity.
For what concerns point 1. we propose a L_0-norm sparse elliptical modelling and show that sparsification is effective. We quantify the effects of points 2. and 3. by studying the models' likelihood in- and out-of-sample for parameters estimated over different train windows. We show that models with larger off-sample likelihoods lead to better performing portfolios only for shorter train sets. For larger train sets, we found that portfolio performances deteriorate and detach from the models' likelihood, highlighting the role of non-stationarity. Investigating the out-of-sample likelihood of individual observations we show that the system changes significantly through time. Larger estimation windows lead to stable likelihood in the long run, but at the cost of lower likelihood in the short-term: the 'optimal' fit in finance needs to be defined in terms of the holding period. Lastly, we show that sparse models outperform full-models and conventional GARCH extensions by delivering higher out of sample likelihood, lower realized volatility and improved stability, avoiding typical pitfalls of conventional portfolio optimization approaches.

3.1 Introduction

Quantitative approaches to asset management have accumulated unprecedented popularity over the last few decades. Of all the algorithms and strategies developed, portfolio selection models are among those that have received wider attention. The essence of portfolio investing is to find the best way of assigning weights to a given set of assets to maximize future portfolio returns while minimizing the investment risk. The exploration of this field starts with Markowitz's mean-variance optimization process [121].

The theory of mean–variance-based portfolio selection is still today a cornerstone of mod-

ern asset management. It rests on the presumption that rational investors choose among risky assets purely on the basis of expected return and risk, with risk measured as portfolio variance. The theoretical foundation of this framework is sound if either: investors exhibit quadratic utility, in which case they ignore non-normality in the data [74], or all the higher moments of the portfolio distribution can be expressed as a function of mean and variance and hence all optimal solutions satisfy the mean-variance criterion. Also, the "optimality" of the mean-variance portfolios is based on the assumption that investors live in a one-period world, while in reality they have an investment horizon that lasts longer than one period. Markets, indeed, constantly change over time and investors are subject to inflows/outflows forcing them to adjust their allocation and take corrective actions.

Form a general, high level, perspective all portfolio optimization approaches are based on a multivariate model of the variables in the market and the economy. The optimization strategies are devised to maximize profits and minimize risks based on such models. In modelling these complex systems there are, however, several sources of inaccuracies and errors with the three main ones being: 1. oversimplifying hypothesis (such as the use of normal distributions); 2. uncertainties resulting from the estimation of the parameters from datasets of limited sizes; 3. intrinsic non-stationarity of these systems, which makes in-sample estimations, based on past observations, inadequate for the estimation of off-sample, future properties. Most likely all three of these factors – and others – contribute to undermining the predictive power of any attempt of modelling markets.

While a large deal of literature has been devoted to relaxing some of the most unrealistic model assumptions (point 1.) the current main pitfall of portfolio optimization is attributed to error maximization (point 2.). This effect has long been established in literature [132, 140]. Essentially, inputs into the mean-variance optimization are measured with uncertainty, and the optimization procedure tends to pick those assets which appear to have the most attractive features – but these are outlying cases where estimation error is likely to be the highest, hence maximizing the impact of estimation error on portfolios' weights. The estimation error is also amplified by market evolution which makes the training on the past not fully representative of future market behaviour (point 3.).

In this Chapter, I address all three sources of inaccuracies. For what concerns point 1. we propose a L_0-norm topologically regularized sparse elliptical modelling [9] and

show that sparsification is effective. We quantify the effects of estimation error and non-stationarity on portfolio performances (point 2. and 3.) by assessing the goodness of models' statistical likelihood for estimates over train sets of different lengths. Specifically, we study how the realized portfolio variance reacts to different out-of-sample likelihoods of the input parameters and, particularly, to sparse models. Further, we analyse how sparse precision matrices impact the magnitude and the stability of portfolio weights.

The remainder of this Chapter is organized as follows: in Section 3.2 we briefly review the theory around portfolio construction, highlighting the pitfalls on assumptions and estimation error and the main solutions proposed in literature; in Section 3.3 we outline our methodology and experiments design and in Section 3.4 we present the results. Appendix D is devoted to recalling some useful aspects of Elliptical distributions.

3.2 Literature Review

3.2.1 Modern Portfolio Theory

Considering a portfolio of n assets with weights $\mathbf{w} = (w_1, ..., w_n)$, returns $\mathbf{R} = (\mathbf{r}_1, ..., \mathbf{r}_n)$ and portfolio returns

$$\mathbf{r}_p = \mathbf{w}\mathbf{R}^\top, \tag{3.1}$$

the standard mean-variance optimization problem consists in minimizing the portfolios' variance σ_p for fixed levels of expected returns $\mathbb{E}[\mathbf{r}_p] = \bar{r}_p$

$$\begin{aligned} \min_{\mathbf{w}} \quad & \sigma_p^2 = \mathbf{w}\boldsymbol{\Sigma}\mathbf{w}^\top \\ \text{s.t.} \quad & \mathbb{E}[\mathbf{r}_p] = \bar{r}_p, \\ \text{and} \quad & \mathbf{w}\mathbb{1} = 1, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is the assets' covariance matrix and $\mathbb{1} \in \mathbb{R}^{n \times 1}$ is a basis column vector with all elements equal 1. Solving for $\mathbf{w}$ for different values of $\bar{r}_p$, one can obtain the optimal weights (i.e. the weights that minimize the portfolio variance) corresponding to different portfolio expected returns $\bar{r}_p$ yielding the so-called efficient frontier – i.e. the set of optimal weights which provide the lowest variance for each level of expected return.

As discussed in the introduction, this optimization is only concerned with the first two moments of the distribution of portfolios' returns and it does not deal with multiperiod investment decisions. These pitfalls have largely been discussed in literature. [101] provides a clear review of what are the assumptions under which repeatedly investing in one-period-efficient portfolios will also result in multiperiod efficiency. Also, many models have been proposed to deal explicitly with multiperiod optimality (see, for example, [128], [111] or [8]). With respect to non-normality of returns, a rich literature is available on both alternative parametrization of data [12] and optimisation frameworks that consider other distribution moments [20, 90, 105] or other measures of risk/return [85, 156, 181].

From the optimization problem in Eq.(3.2) it is also clear that the optimization does not treat the error and uncertainty around the parameters $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$. The difference between the estimated and true distribution parameters is called estimation error. It arises from both the sampling procedure or availability of data and non-stationarity. The error coming from sampling, also referred to as sampling error, is due to parameters used in the portfolio optimization process being typically point estimates – we can only expect these estimates to equal the true distribution parameters if our sample is infinitely large. Assuming stationary data, sampling error could be fixed by increasing the number of observations in the estimation sample. Indeed, the convergence rate is in the inverse of the square-root of the sample size, as dictated by the law of large numbers. This would come handy in our times of increasing data availability. However, a second source of estimation error comes from non-stationarity. A time series is said to be non-stationary if its distribution parameters (or the distribution itself) changes over time – in this case, extending the length of observations might reduce the contribution of sampling error to estimation error, but at the same time, it could increase that of non-stationarity [37].

Many techniques have been proposed in literature to deal with this phenomenon, both relying on heuristic methods and decision-theoretic foundations [160]. Heuristic approaches mainly propose to constrain the optimization problem in order to impose feasible optimal weights.

Michaud and Michaud (1998) addresses explicitly the sampling error proposing a Monte Carlo based procedure called resampling. In order to model the randomness of the input mean vector and covariance matrix, portfolio resampling consists in repeatedly drawing from the return distribution given by the point estimates and creating n artificial

new samples. For each sample, an efficient frontier is estimated and the final, resampled-efficient frontier is given by the average weight across all of the resampled portfolios.

From a decision-theoretic perspective, Bayesian techniques have recently played a primary role in literature. The rationale behind Bayesian statistics for portfolio construction is to include non-sample information to tackle the effect of parameter uncertainty on optimal portfolio choice. Instead of a point estimate, Bayesian approaches produce a density function for the parameters involved, by combining sample information (likelihood) with prior belief, potentially coming from non-sample information. A special case of this general approach is the seminal work of Black and Litterman (1992). In their pioneering work, the authors assume assets' returns to be normally distributed with mean equal to the 'equilibrium returns' (that is, the mean returns that would output the market portfolio if used in a mean-variance optimization) and combine this "sample" information with investors' views on the assets. In this way, in absence of an informative prior from investors, the model would return the market or 'equilibrium' portfolio. In presence of investors' priors, instead, the allocation would diverge from the equilibrium portfolio accounting for investors' views and, proportionally, to their confidence level. Other than being highly appealing from a practitioner's perspective, the model proposed in [28] highlights the flexibility of the Bayesian framework, with many sources of information that could potentially be used in combination or to update the in-sample information. This is a very active area of research with recant notable examples including Scherer et al. (2012) and De Franco et al. (2019).

More recently, entropy is receiving increasing attention as alternative measure of uncertainty in information theory, econometrics, and finance [16]. Starting from the pioneering work of Philippatos and Wilson (1972), entropy based portfolio allocation models are increasingly popular in the financial literature. Entropy in place of variance as measure of uncertainty and diversification for the portfolio selection problem has proven to provide greater diversification and stability, avoiding classical corner solutions of the mean-variance approach [16, 148]. Further entropy is a non-parametric function designed to accommodate non-normality and asymmetry and no covariance estimation is required as the joint entropy dependence structure can be captured in the objective function [129]. Lastly, entropy provides a flexible framework also in mixing multiple sources of information into the joint probability definition [131].

3.2.2 Conditional Correlation Models

Modelling volatility in financial time series has been the object of much attention ever since the introduction of the autoregressive conditional heteroskedasticity (ARCH) model in the seminal paper of Engle (1982). Numerous variants and extensions of ARCH models have been proposed investigating and leveraging different effects observed in financial time series - see Bollerslev et al. (1992) and Bera and Higgins (1993) for a survey of ARCH-type models. Stochastic volatility (SV) models in continuous time are at the foundation of modern derivatives pricing [49, 80, 89], aiming at resolving the shortcomings from constant volatility assumption of the Black–Scholes [29] based approaches.

While modelling volatility of the returns has been the main centre of attention, understanding the co-movements of financial returns is of great practical importance. A large body of literature has therefore developed, studying the evolution and temporal dependence of correlations, with the main approaches being multivariate extensions of the GARCH model. Similarly to the univariate case, many different model specifications have been proposed trying to balance flexibility and number of parameters. For a survey, please refer to Boudt et al. (2019) and Bauwens et al. (2006).

More recently, copula [165] based models are increasingly emerging as useful tools to deal with non standard multivariate distributions remedying to various shortcomings of the GARCH structures, with the copula approach being effective in describing the non-linear, asymmetric, and possible tail dependence between markets. Copula-GARCH models combine the use of GARCH models and a copula function to allow flexibility on the choice of marginal distributions and dependence structures and particularly the vine-copula method has been gaining attention recently in that a multi-dimensional density can be decomposed into a product of conditional bivariate copulas and marginal densities [95]. Vine structure is an approach to effectively solve the problem of the dynamic correlation structure between multiple variables, and it provides an effective solution to the matter of variable correlation with complex dependency patterns. The vine-copula method has been gaining attention recently in that a multi-dimensional density can be decomposed into a product of conditional bivariate copulas and marginal densities [57, 86]. Several authors show that compared to the traditional methods the vine structures are better in capturing the dependence between variables and in risk management applications [36, 182].

Given the wide adoption among financial academics and practitioners, we will consider the Orthogonal GARCH (O-GARCH) as baseline method to compare our results (Section 3.5). Considering a dataset of T *returns* $\times$ n *assets* the observations are assumed to be generated by an orthogonal transformation of n (or a smaller number of) univariate GARCH processes. The matrix of transformation is the orthogonal matrix (or a subsection) of eigenvectors of the covariance matrix of the returns. In the generalized version, this matrix must only be invertible.

In the Orthogonal GARCH model of [3], the $n \times n$ time-varying variance matrix $\mathbf{H}_t$ is generated by n univariate GARCH models

$$\mathbf{r}_t = \mathbf{G}\mathbf{z_t} \tag{3.3}$$

where $\mathbf{G}$ is a non-singular $n \times n$ matrix. In the generalized specification of the O-GARCH model of [173], the uncorrelated factors $\mathbf{z}_t$ are standardized to have unit unconditional variances ($\mathbb{E}[\mathbf{z_t}\mathbf{z_t}^\top] = \mathbb{1}$). The principal components (i.e. unobservable factors) are estimated from the data through $\mathbf{G}$ and the factors $\mathbf{z}_t$ are assumed to follow a GARCH process with the $n \times n$ diagonal matrix of conditional variances of $\mathbf{z}_t$ defined as

$$\mathbf{H}_t^{\mathbf{z}} = (\mathbf{I} - \mathbf{A} - \mathbf{B}) + \mathbf{A} \odot (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\top) + \mathbf{B}\mathbf{H}_{T-1}^{\mathbf{z}} \tag{3.4}$$

where $\mathbf{A}$ and $\mathbf{B}$ are diagonal $n \times n$ parameter matrices and $\odot$ denotes the Hadamard (i.e. element-wise) product. Therefore the conditional covariance matrix of $\mathbf{r}_t$ can be expressed as

$$\mathbf{H}_t = \mathbf{G}\mathbf{H}_t^{\mathbf{z}}\mathbf{G}^\top . \tag{3.5}$$

The linear mapping $\mathbf{G}$ is constructed via singular value decomposition of the returns covariance matrix $\mathbb{E}[\mathbf{r}_t\mathbf{r}_t^\top] = \mathbf{\Sigma}$

$$\mathbf{G} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V} \tag{3.6}$$

where the $\mathbf{U}$ is the matrix of the eigenvectors of $\mathbf{\Sigma}$ and the diagonal matrix $\mathbf{\Lambda}$ holds its eigenvalues.

It is worth emphasizing that while these models have the potential to offer great flexibility, they are inevitably exposed to the curse of dimensionality in that as the number of

assets and/parameters increases, the estimation of these models becomes quickly unfeasible [41].

3.3 Methodology

The estimation error is quantified by measuring how well the functional form of the multivariate probability distribution, $f_{\boldsymbol{\theta}}(\mathbf{X})$ defined via the parameters $\boldsymbol{\theta}$ estimated in-sample, describes the actual data out-of-sample. The statistical measure that describes how likely are observations to belong to the estimated probability function is the likelihood. The likelihood principle is a cornerstone of statistical analysis in that maximum likelihood estimators are guaranteed to be asymptotically efficient under mild conditions ([177], [50], [53]). In this section, we introduce our methodology and discuss our results for the multivariate normal case. In Section 3.4.3 I further discuss the generality of this approach and show that it extends to other distributions of the elliptical family including, in particular, the multivariate Student-t.

The logarithm of the likelihood for the normal case is proportional to

$$\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}_t) = \log |\mathbf{J}| - (\mathbf{x}_t - \boldsymbol{\mu})\mathbf{J}(\mathbf{x}_t - \boldsymbol{\mu})^\top + k \ , \tag{3.7}$$

where $\mathbf{x}_t = (x_{t,1}, x_{t,2}, ..., x_{t,n})$ is the n-dimensional multivariate returns observation vector at time t; $\boldsymbol{\theta}$ is the model parameter set, which includes $\boldsymbol{\mu}$ the vector of means and $\mathbf{J}$ the generalized precision matrix and; k is a constant which is independent from $\boldsymbol{\mu}, \mathbf{J}$ or $\mathbf{x}_t$ (see Section 3.4.3). In the multivariate normal case $\mathbf{J} = \boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance. Our results generalize to other elliptical distributions with defined covariance where $\mathbf{J}$ is proportional to the inverse covariance, which we assume is defined and invertible. Results for the Student-t are explicitly reported in appendix 3.4.3.

Our goal is to study the log-likelihood in Eq. (3.7) using different estimation windows and comparing how the precision matrices, estimated through maximum-likelihood and TMFG-LoGo, perform. We considered a **dataset** of daily closing prices of US stocks entering among the constituents of the S&P 500 index between 02/01/1997 and 31/12/2015. After screening for those continuously traded and those not displaying abnormal returns, we

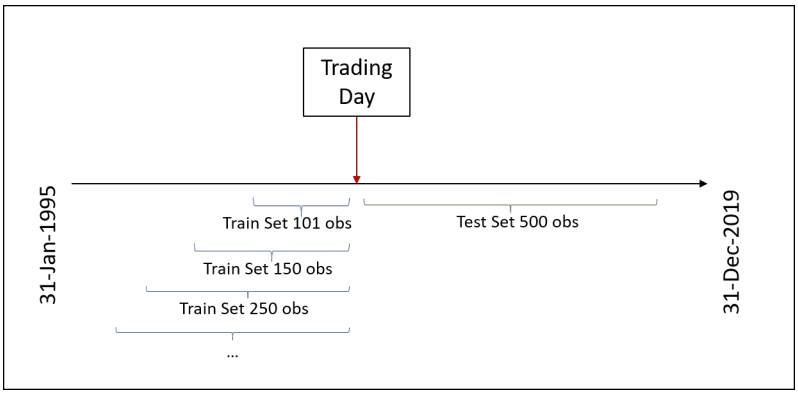


Figure 3.1: Training and Testing scheme.We randomly sample the ending date of the training period, the 'trading day'. We then estimate the model parameters considering different training windows using observations up to the randomly selected trading day. The subsequent 500 observations are used for testing.

reached a final dataset of 342 stocks. For each asset $i = 1, ..., n$, we calculated the corresponding daily returns $x_{t,i} = S_{t,i}/S_{(t-1),i} - 1$, where $S_{t,i}$ is the closing price of stock i at time t, for a total of 4026 daily multivariate observations.

We designed a resampling experiment in which we select 100 stocks at random among the 342 and a random trading day spanned in our dataset. Starting from the randomly selected trading day and going back in time, we define five train sets of different sizes by including an increasing number of observations. We start at 101 observations, then 150, 250, 500, 1000 and finally 1500 observations. We then use a fixed-length test set of 500 observations following the randomly selected trading day. We keep the test set length fixed to avoid biases and selected 500 observations so that, for all estimation windows, the main crisis event (i.e. Global Financial Crisis in 2008) can be randomly included in or out of sample. Figure 3.1 shows a sketched example of our train/test split with different estimation windows.

We use the train set to estimate the mean vector $\boldsymbol{\mu}$, the maximum-likelihood covariance matrix $\boldsymbol{\Sigma}$ and the sparse TMFG LoGo covariance matrix $\boldsymbol{\Sigma}_{TMFG}$. These parameters are then used to compute the log-likelihood in Eq.(3.7) for both in-sample and out-of-sample observations. We then investigate how the different estimates used in a portfolio optimization procedure affect the optimal weights and portfolio characteristics. To this extent, we considered the standard, unconstrained Markowitz optimization problem described in Section 3.2.1. This is done to avoid any bias coming from constraints in our analysis and to keep the framework as plain as possible. We focus therefore our analysis on the minimum

variance portfolio, that is the efficient portfolio that minimizes the expected variance. To obtain the solution for the minimum variance portfolio, the portfolio optimization problem in Eq. (3.2) rewrites as

$$\begin{aligned} \min_{\mathbf{w}} \quad & \sigma_p^2 = \mathbf{w}\boldsymbol{\Sigma}\mathbf{w}^\top \\ \text{s.t.} \quad & \mathbf{w}\mathbb{1} = 1, \end{aligned} \tag{3.8}$$

which gives the optimal, minimum variance weights

$$\mathbf{w}^*_{\text{min}} = c\ \mathbb{1}\boldsymbol{\Sigma}^{-1}, \tag{3.9}$$

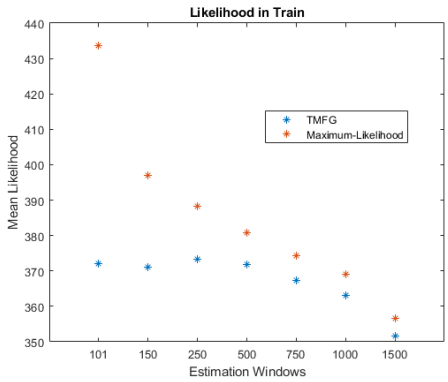
where $c = \frac{1}{\mathbb{1}^\top \boldsymbol{\Sigma}^{-1}\mathbb{1}}$ is a normalization constant. Considering the estimation scheme described above and outlined in fig. 3.1, the estimated covariance matrices are used as inputs in the minimum variance optimal portfolio, Eq. (3.9), to compare the different log-likelihood levels obtained out-of-sample and the corresponding effects on portfolio performances.

3.4 Results

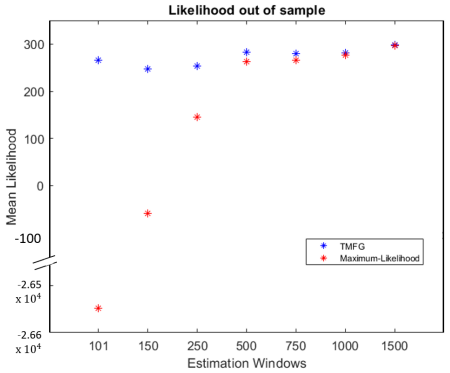
3.4.1 Likelihood Comparison

Figure 3.2 reports the average log-likelihood for the train data (fig. 3.2a) and for the test data (fig. 3.2b) computed across 100 resamplings. The larger the log-likelihood is, the better the parameters $\boldsymbol{\theta}$ are at describing the data, for the assumed model. Fig. 3.2a shows that, as expected and by definition, the maximum-likelihood estimate of the covariance matrix provides a higher in-sample likelihood as compared to the TMFG covariance, although the latter tracks quite closely the maximum-likelihood. Also, one might observe that the likelihood is strictly decreasing with the number of observations included in the estimation window. Indeed, as the number of observations decreases relative to the parameters, the model overfits the sample yielding larger in-sample likelihoods. Filtering the covariance matrix and reducing the number of parameters clearly limits the overfitting potential of the model as shown by the lower levels of likelihood attained by TMFG when fewer observations are used which, therefore, results in a larger gap in likelihood relative to the maximum-likelihood covariance.

Perhaps more interestingly, fig. 3.2b reports the likelihoods obtained out-of-sample using the two different in-sample estimates of the covariance matrix. The first observation is that TMFG-LoGo provides a substantially larger log-likelihood, especially for short estimation windows. This result is exacerbated by the fact that when 101 observations are considered, the number of stocks is very close to the number of observations in our samples. While the resulting covariance is still full-rank (*number of observations* $>$ *number of variables*), it leads to unstable estimation in the maximum-likelihood covariance (i.e. the so-called 'the curse of dimensionality') whereas TMFG-LoGo is still well defined. Note that there is a break y-axis of the figure to allow a better inspection of the results. The figure shows that for longer estimation windows, the out-of-sample log-likelihood computed with the maximum-likelihood covariance tends to converge to the TMFG likelihood which, however, a) always provides the best out-of-sample likelihood in our experiment and b) provides quite stable likelihood values also for shorter estimation windows. We conclude that the TMFG-LoGo algorithm does a good job at filtering the correlation structure providing higher out-of-sample likelihood and stable results with shorter estimation windows, confirming the results with stationary time series previously reported in [15].



(a) Likelihood comparison in-sample



(b) Likelihood comparison out-of-sample. Note the y-axis break to fit the scale for 101 days estimation window.

Figure 3.2: Log Likelihood computed in- and out-of-sample. Both the likelihood computed using the maximum likelihood and the TMFG covariances decrease in sample as the sample size increases. The maximum likelihood covariance delivers by construction the highest likelihood, but the TMFG likelihood tracks it closely. In test, instead, the TMFG covariance always attain the highest likelihood and delivered good results also when the number of observations becomes close to the number of variables.

3.4.2 Impact of precision matrix estimate on optimal portfolios

We now address empirically the question of what is the impact of different parameter estimates on portfolios weights and performances, when these parameters are used as inputs in the portfolio optimization problem in Eq. (3.2). Having focused our attention on the minimum variance portfolio on the efficient frontier, we report in Figure 3.3 the realized standard deviation of portfolios obtained using the same parameters which provided the log-likelihood displayed in Figure 3.2. The chart shows that, overall, the out-of-sample portfolio variance decreases as the likelihood increases up until when 750 observations are used. This is coherent with respect to the likelihood results that reported, indeed, increasing likelihoods for the same estimation windows. In particular, for shorter estimation windows, the TMFG-LoGo covariance matrix provides portfolios with significantly lower realized variance. Also, little changes are observed in the realized variance when observations from 101 to 750 are included, signalling that the TMFG-LoGo extract the relevant dependency links also when few observations are available. The gap in performance tends to reduce as the number of observations in the estimation window increases, with the TMFG-LoGo portfolios always displaying lower volatility. However, when more than 750 daily observations are included, while the out of sample likelihood remains flat or slightly increases, the portfolios' variance tends to increase.

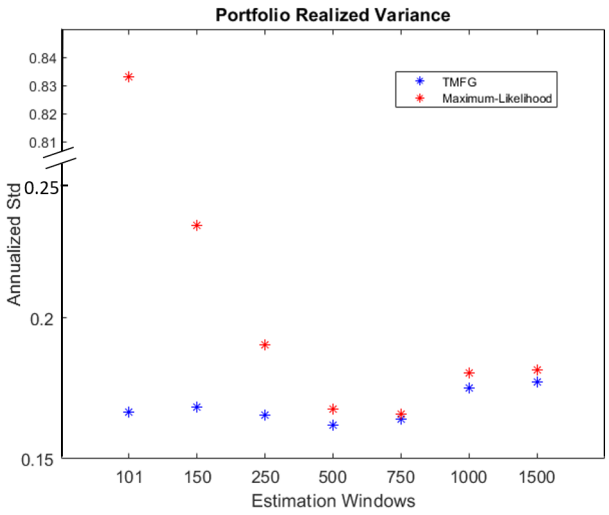
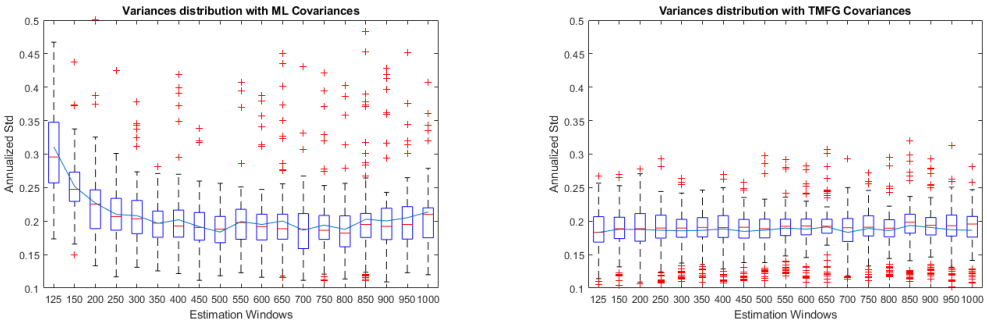


Figure 3.3: Realized Standard Deviation. Increasing the estimation window and for higher values of likelihood (figure 3.2), the realized standard deviation of portfolios decreases. Y-axis break to fit the scale for 101 days estimation window.

To further investigate this pattern, we report in Figure 3.4 the volatilities for all 100 resamplings and considering steps of 25 observations in the estimation windows. The figure confirms that the TMFG-LoGo covariances delivered overall less volatile portfolios across resamplings and estimation windows. Secondly, the figure shows that the portfolios obtain the lowest out-of-sample variance when approximately 2 to 3 years of daily observations (450 to 700 observations) are included in the train set. This pattern is clear for the Maximum likelihood portfolios, with means, quintiles and outliers drifting upwards when more than 750 observations are included. The TMFG filtered covariance regularizes and smooths this effect as well, but still when more than 750 observations are included, the resulting portfolios exhibit a slightly higher variance. This is consistent with the literature showing that longer estimation windows provide worse forecasts in financial time series due to the regime-changing nature of financial markets [153]. It is also worth emphasizing that the different features of TMFG-LoGo and the Maximum-Likelihood portfolios are due solely to the different estimates of the covariance matrix as these are the only inputs used in the Minimum Variance optimization 3.9



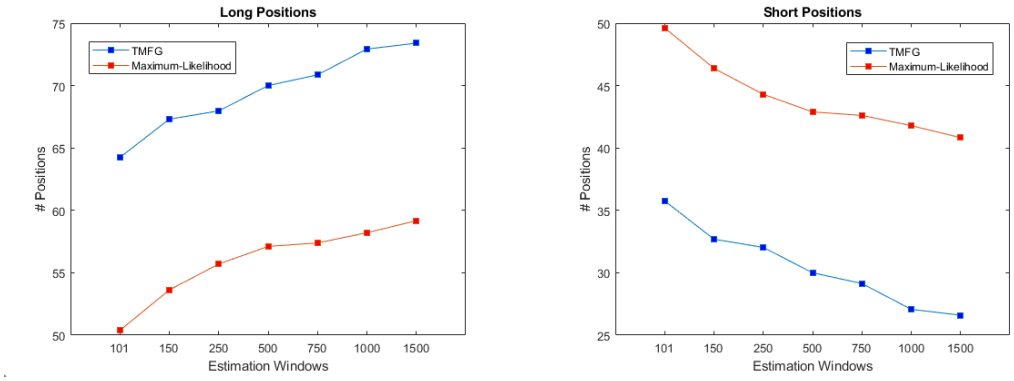
(a) Realized volatilities obtained with Maximum Likelihood covariances

(b) Realized volatilities obtained with TMFG filtered covariances

Figure 3.4: Portfolio realized volatility across resamplings for different estimation windows. The box-plot shows the distribution of the variances obtained for 100 resampled portfolios and the the blue line overlaid shows the average variance (i.e. the mean of the variances distribution).

Finally, we address the impact of sparsity on optimal portfolio weights. Figure 3.5 reports the number of Long (fig. 3.5) and Short (fig. 3.5a) positions (i.e. positive and negative weights assigned to the stocks in portfolio) on average across the 100 resamplings. The first observation is that the number of long positions tends to increase as the estimation window increases and coherently the short positions diminish accordingly. Using TMFG-LoGo precision matrices anticipates this behaviour, in that TMFG portfolios always display a greater number of long positions also for short estimation windows. Recalling that the Minimum-Variance optimisation (Eq. 3.2) is constrained to sum to 1, the intuition behind this phenomenon is that the fewer the observations used in the estimation of the covariance, the higher is the tendency of the Minimum-variance portfolios to exhibit extreme negative and positive weights. In other words, the weights still sum up to 1, but with a combination of large long and short bets. Over the long term, estimates are more stable and possible outliers in assets' variances and correlations are polished, leading to more stable portfolios. This intuition is confirmed by looking at the distribution of weights across resamplings in Figure 3.6. This chart (note the different scales) shows that using the TMFG-LoGo covariance matrix significantly improves the stability of the optimal solutions, reducing outliers and avoiding "corner", i.e. extreme solutions which are a typical pitfall of the unconstrained Markowitz optimization. This results shows that the correlation coefficient among assets plays an important role in that for high correlation levels, the optimization procedure would prefer one stock in place of another for slightly more appealing variance features. Having filtered the correlation structure in the TMFG-LoGo procedure, we obtained a portfolio

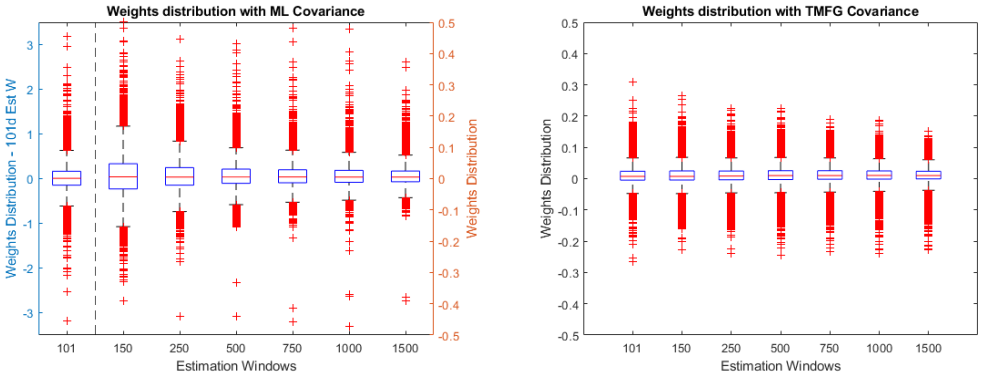
that is much more general (hence the anticipated larger number of Long positions) and less sensitive to single assets features given the filtered correlation among stocks. Lastly, considering the standard unconstrained optimization problem in Eq. (3.2), both the maximum-likelihood and the TMFG matrices produce portfolios that are in the vast majority of cases investing in all assets. In other words, even considering a sparse precision matrix like in the TMFG-LoGo case, we very rarely found weights equal to zero assigned to some assets.



(a) Number of "Buy" positions

(b) Number of "Sell" positions

Figure 3.5: Comparison of Buy/Sell Active Positions. As the number of training observations increases, the optimizations delivers an increasing number "Long" positions. This tendency is anticipated when using TMFG fiiltered covarinace which always delivers an higher number of Long positions.



(a) Distribution of optimal weights using Maximum-Likelihood covariance

(b) Distribution of optimal weights using TMFG covariance

Figure 3.6: Optimal Weights Distribution. Using the TMFG filtered covariance in the optimization provides stable weights as compared to the maximum-likelihood covariance, avoiding "corner solutions" and enhancing diversification.

3.4.3 Elliptical Distributions

The methodology outlined so far and the experiments performed assume that returns are normally distributed in that we considered the classical mean-variance optimization setting

and we used the normal log likelihood functional form 3.7 to assess the goodness of the estimates. In this section we discuss the general validity of our findings for the class of elliptical distributions by addressing both of these choices. We begin our discussion by briefly introducing a general definition for the probability density function of elliptical distributions. Consider an n-dimensional vector of multivariate returns $\mathbf{x} = (x_1, x_2, ..., x_n)$. If $\mathbf{x}$ is elliptical distributed, then its probability density function is defined as

$$f_x(\mathbf{x}) = c_n |\mathbf{J}|^{1/2} g_n \left[(\mathbf{x} - \boldsymbol{\mu}) \mathbf{J} (\mathbf{x} - \boldsymbol{\mu})^\top \right], \tag{3.10}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{1 \times n}$ is the vector of location (mean) parameters and c_n is a normalization constant. The matrix, $\mathbf{J} = \boldsymbol{\Omega}^{-1} \in \mathbb{R}^{n \times n}$ is the generalized precision matrix, a positively defined matrix which is the inverse of the dispersion matrix $\boldsymbol{\Omega}$. When the covariance is defined then $\boldsymbol{\Omega} = (-\psi'(0))^{-1} \boldsymbol{\Sigma}$, that is, $\boldsymbol{\Omega}$ is proportional to the covariance matrix and the proportionality factor is the inverse of the first derivative of the characteristic generator evaluated at 0. The function, $g_n(\cdot)$ is called density generator.

Also, let us stress that $(\mathbf{x} - \boldsymbol{\mu}) \mathbf{J} (\mathbf{x} - \boldsymbol{\mu})^\top$ - i.e. the generalized, square Mahalanobis distance - is a quadratic term and hence a non-negative quantity provided that the matrix $\boldsymbol{\Omega}$ is positive definite. To ease the notation, for the remaining of the Chapter we shall refer to the generalized Mahalanobis distance as d^2

$$d^2 = (\mathbf{x} - \boldsymbol{\mu}) \mathbf{J} (\mathbf{x} - \boldsymbol{\mu})^\top . \tag{3.11}$$

For different density generators $g_n(\cdot)$ we obtain different distributions of the elliptical family. It is easy to see, for example, that the normal distribution is obtained by using:

$$g(u) = e^{-u/2}, \tag{3.12}$$

and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}$.

Similarly the Student-t distribution is obtained by using:

$$g_n(u) = \left(1 + \frac{u}{v}\right)^{-\frac{n+v}{2}}, \tag{3.13}$$

where v is the degrees of freedom, and $\boldsymbol{\Omega} = \frac{v-2}{v} \boldsymbol{\Sigma}$.

The validity of the **mean-variance** framework for elliptical distributions has long been

established in literature [145]. This proposition is derived easily from two properties of the elliptical distributions. First, for every elliptical distribution with defined mean and variance, the distribution is completely specified by them ([145] or [42]), with all the higher moments being either zero or proportional to the first or second moment. Second, any linear combination of multivariate elliptically distributed variables is also an elliptically distributed variable. In the case of normal distribution and Stiudent-t distribution they also have the same density generator function. Further details on these properties are provided in Appendix D.

It follows that, if asset returns have a multivariate elliptical distribution $\mathbf{x} \sim \mathcal{E}_n(\boldsymbol{\mu}, \boldsymbol{\Omega}, g_n)$, then the portfolio expected return and dispersion are given by, respectively, $\mathbb{E}[r_p] = \mathbf{w}\boldsymbol{\mu}^\top$ and $\sigma_p = \mathbf{w}\boldsymbol{\Omega}\mathbf{w}^\top$, matching the optimization framework outlined in Section 3.2.1.

With respect to our **likelihood analysis**, considering distributions with probability density function of the form specified in Eq. (3.10), the corresponding likelihood function is of the form

$$\mathcal{L}_{ED}(\boldsymbol{\theta}; \boldsymbol{x}) = |\mathbf{J}|^{1/2} \; g_n\left(d^2\right) \; . \tag{3.14}$$

where ED denotes the general Elliptical Distributions and we omitted the constant of integration. To stress the general validity of our analysis for other elliptical distributions, we repeated the experiments discussed in Section 3.3 considering the t-student generator.

Assuming a **Student - t** distribution of the log returns, the log likelihood (Eq. (3.14)) is

$$\log \mathcal{L}_{\text{Student}} = \frac{\log|\mathbf{J}|}{2} - \frac{n+\nu}{2} \log\left(1 + \frac{d^2}{\nu - 2}\right) \tag{3.15}$$

where n is the sample size and ν is the degree of freedom. Figure 3.7 reports the likelihood comparison for the same resamplings as in Figure 3.2 but using a student-t log likelihood as in Eq. (3.15). Here we used $n = 500$ observations (i.e. the out-of-sample size) and $\nu = 3$. We verified that this findings are robust across different degrees of freedom in the range $\nu = [2.1, 4]$.

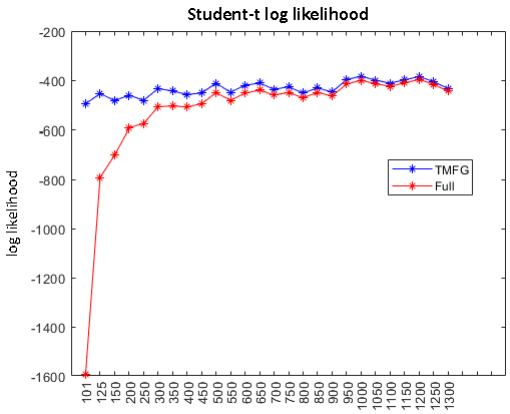


Figure 3.7: Log likelihood assuming a t-student distribution of log returns. Out-of-sample log likelihood likelihood computed using the maximum likelihood and the TMFG covariances. These results are coherent with the findings related to the normal distribution presented in Section 3.4.1.

3.5 O-GARCH Comparison

Having assessed the impact of sparsity on portfolio weights' stability, in this section I compare the features of portfolios obtained with TMFG-LoGo covariances with a baseline standard approach. As discussed in Section 3.2, time varying conditional models represent the most widely used models in dealing with covariance structure. As such, I focus my attention on the general O-GARCH model recalled in Section 3.2.2.

Following the approach of [173] outlined in in Section 3.2.2, for each resampling I estimated the maximum likelihood covariance eigenvectors and the corresponding principal components (PC). The PCs are assumed to follow a GARCH process and the corresponding parameters are estimated for each principal component time series. We tested for different model specifications, allowing for both the ARCH and GARCH parameters to range from 1 to 3 lags and selected the best model based on the AIC and BIC criteria. Table 3.1 and table 3.2 reports the average AIC and BIC statistics for every model specifications across different estimation windows. For all estimation windows, both the AIC and BIC criteria support the GARCH(1,1) specification. Furthermore, in Appendix E I report the median, 5^{th} and 95^{th} percentiles of all the AIC and BIC statistics. The table shows that the GARCH(1,1) specification delivered the lowest AIC and BIC statistics for each of the main percentiles considered. In other words, the GARCH(1,1) specification selected in our experiment is the

preferred specification according to the AIC and BIC criteria in all cases, and not only in mean across resamplings.

	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$	$(2,3)$	$(3,2)$	$(3,3)$
$\mathbf{q = 101}$	-664	-661	-660	-660	-659	-659	-657
$\mathbf{q = 125}$	-806	-803	-803	-803	-800	-801	-799
$\mathbf{q = 250}$	-1495	-1492	-1491	-1491	-1489	-1490	-1488
$\mathbf{q = 500}$	-2914	-2908	-2911	-2911	-2908	-2909	-2907
$\mathbf{q = 750}$	-4316	-4305	-4313	-4313	-4308	-4311	-4310
$\mathbf{q = 1000}$	-5643	-5628	-5641	-5641	-5635	-5639	-5639
$\mathbf{q = 1500}$	-8062	-8040	-8061	-8061	-8056	-8059	-8060

Table 3.1: Average AIC information criterion for different GARCH specifications across 100 resamplings for different estimation windows.

Train Obs	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$	$(2,3)$	$(3,2)$	$(3,3)$
101	-657	-651	-648	-648	-643	-644	-639
125	-798	-792	-789	-789	-784	-784	-780
250	-1485	-1478	-1474	-1474	-1469	-1469	-1464
500	-2902	-2892	-2890	-2890	-2883	-2884	-2878
750	-4302	-4287	-4290	-4290	-4281	-4284	-4278
1000	-5629	-5609	-5617	-5617	-5606	-5610	-5605
1500	-8046	-8019	-8035	-8035	-8024	-8028	-8023

Table 3.2: Average BIC information criterion for different GARCH specifications across 100 resamplings for different estimation windows.

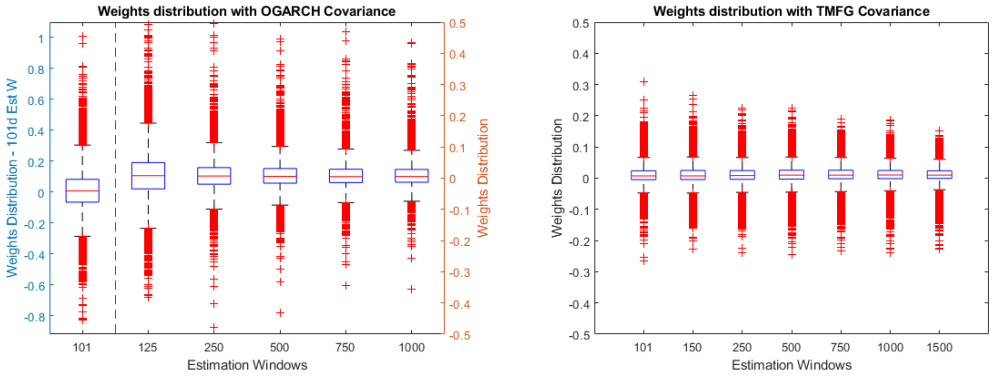
Having selected the GARCH(1,1) specification, for each resampling we estimated the model parameters (Table 3.5 reports the estimated parameters and corresponding p-Values), forecasted one steps ahead principal components and then reconstruct the covariance matrix as in Eq. (3.5).

	$ARCH(1)$	$pValue$	$GARCH(1)$	$pValue$
$\mathbf{q = 101}$	0.0688	0.430	0.505	0.386
$\mathbf{q = 125}$	0.0633	0.381	0.552	0.318
$\mathbf{q = 250}$	0.0630	0.166	0.625	0.182
$\mathbf{q = 500}$	0.0632	0.102	0.705	0.115
$\mathbf{q = 750}$	0.0604	0.049	0.766	0.065
$\mathbf{q = 1000}$	0.0564	0.031	0.806	0.041
$\mathbf{q = 1500}$	0.0445	0.0007	0.917	0.0084

Table 3.3: GARCH(1,1) - Average Parameter and p-value. Mean parameters and corresponding p-Values for the GARCH(1,1) model estimated for the Principal Components across 100 resamplings.

We carried out the same stability experiments described in Section 3.4.2 using the O-GARCH(1,1) covariance in the portfolio optimization. Figure 3.8 presents the distribution of weights across resamplings for the optimal minimum-variance weights obtained using the O-GARCH covariance and compared to the TMFG and maximum-likelihood covariances. The figure shows that the O-GARCH covariance presents instability problems similar to the full covariance (Figure 3.6). Similarly, Figure 3.9 in extends Figure 3.5 by comparing the number of Long (fig. 3.9a) and Short (fig. 3.9b) positions when also the optimal minimum variance weights obtained using the O-GARCH forecasted covariance matrix are considered. The O-GARCH positions closely track the full covariance behaviour, thus the same conclusions drawn in Section 3.4.2 apply.

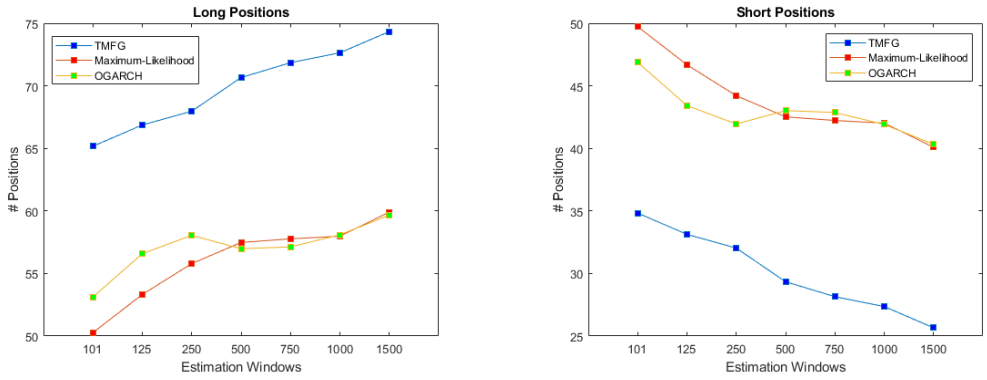
In essence, all the stability problems that apply to the Maximum-Likelihood covariance still apply to the O-GARCH forecasted covariance matrix when used for portfolio construction.



(a) Distribution of optimal weights using O-GARCH covariance

(b) Distribution of optimal weights using TMFG covariance

Figure 3.8: Optimal Weights Distribution. O-GARCH comparison.



(a) Number of "Buy" positions

(b) Number of "Sell" positions

Figure 3.9: Buy/Sell Active Positions. O-GARCH comparison. As the number of training observations increases, the optimizations delivers an increasing number "Long" positions. This tendency is anticipated when using TMFG fiiltered covarinace which always delivers an higher number of Long positions.

3.5.1 Backtest

To further compare the features of sparse TMFG portfolios to the conventional O-GARCH, we backtest a simple trading strategy and compared performance results. The strategy follows a rolling estimation scheme with daily rebalance: we pick a random trading day, we use the previous q days to estimate the parameters and we compute the optimal minimum variance portfolio weights (Eq. (3.9)). We then roll the estimation window on a daily basis and rebalance the portfolio accordingly, assuming therefore to buy at the close price and hold until close the following day. This process is reiterated through 500 observations following the randomly picked starting trading day. In accordance with the testing framework considered throughout the thesis, for each estimation window we run 100 resampling where we pick a different set of 100 stocks and a different start trading day. We do no include transactions costs in the performance presented, as we separately investigate the turnover of both the strategies.

In terms of parameters re-estimation, on a daily basis we consider the previous q observations and re-estimate the means vector and covariance matrix. For the O-GARCH, we forecast the one day ahead conditional volatility of the principal components and reconstruct the corresponding covariance matrix. The model is fit only once, with the first estimation window, and we then keep the ARCH and GARCH parameters fixed. As the new observations come in, we estimate the new covariance matrix, the corresponding principal components and use it forecast the one day ahead covariance matrix.

Table 3.4 presents the out-of-sample annualized standard deviation of the strategy with minimum variance portfolios constructed based on TMFG precision matrices compared to the OGARCH based portfolios in Table 3.5. For each estimation window, we report the median, 5^{th} and 95^{th} percentiles across 100 resamplings.

As shown in the tables, TMFG based portfolios delivered more stable performances with lower realized variance in the vast majority of cases.

	σ		
Train Obs	5^{th}	Median	95^{th}
101	0.078	0.097	0.197
125	0.079	0.106	0.198
250	0.080	0.100	0.196
500	0.084	0.112	0.209
750	0.085	0.097	0.214
1000	0.088	0.114	0.212
1500	0.092	0.119	0.223

Table 3.4: TMFG portfolios performance metrics. Annualized standard deviation of a daily rebalancing minimum volatility strategy. Optimal weights computed using using TMFG precision matrices. Median, 5^{th} and 95^{th} percentiles across 100 resamplings.

	σ		
Train Obs	5^{th}	Median	95^{th}
101	0.104	0.168	0.415
125	0.101	0.178	0.384
250	0.097	0.137	0.293
500	0.098	0.151	0.316
750	0.099	0.117	0.325
1000	0.103	0.164	0.314
1500	0.105	0.161	0.334

Table 3.5: OGARCH portfolios performance metrics.

More important than performance in the context of this thesis is the stability of portfolios weights through rebalances. As previously mentioned, the performance metrics presented in Table 3.4 and Table 3.5 do not take into account transaction costs to allow a separate investigation on the impact of the precision matrix on portfolios stability. We turned our attention to daily turnover, computed for each day t as

$$\tau_t = \sum_{i=1}^{100} |w_{i,t} - w_{i,t-1}| \tag{3.16}$$

where $w_{i,t}$ is the weight allocated at time t to the $i-th$ stock among the 100 randomly sampled.

Figure 3.10 reports the average daily turnover of the two strategy across our 100 resamplings. Across all the estimation windows, TMFG portfolios reported significantly lower turnover, in most of cases two to three times less than O-GARCH portfolio. Plotting all daily turnover data across all resamplings in Figure 3.11 further shows that turnover on TMFG portfolios is actually less than 5% for most of the days across our resamplings, with a distribution that is heavily positively skewed. O-GARCH portfolios, on the other hand, displays a much less favorable turnover distribution, with extremes that reach 300% turnover. These results come with no surprise given the much higher portfolio stability already observed in Figure 3.8.

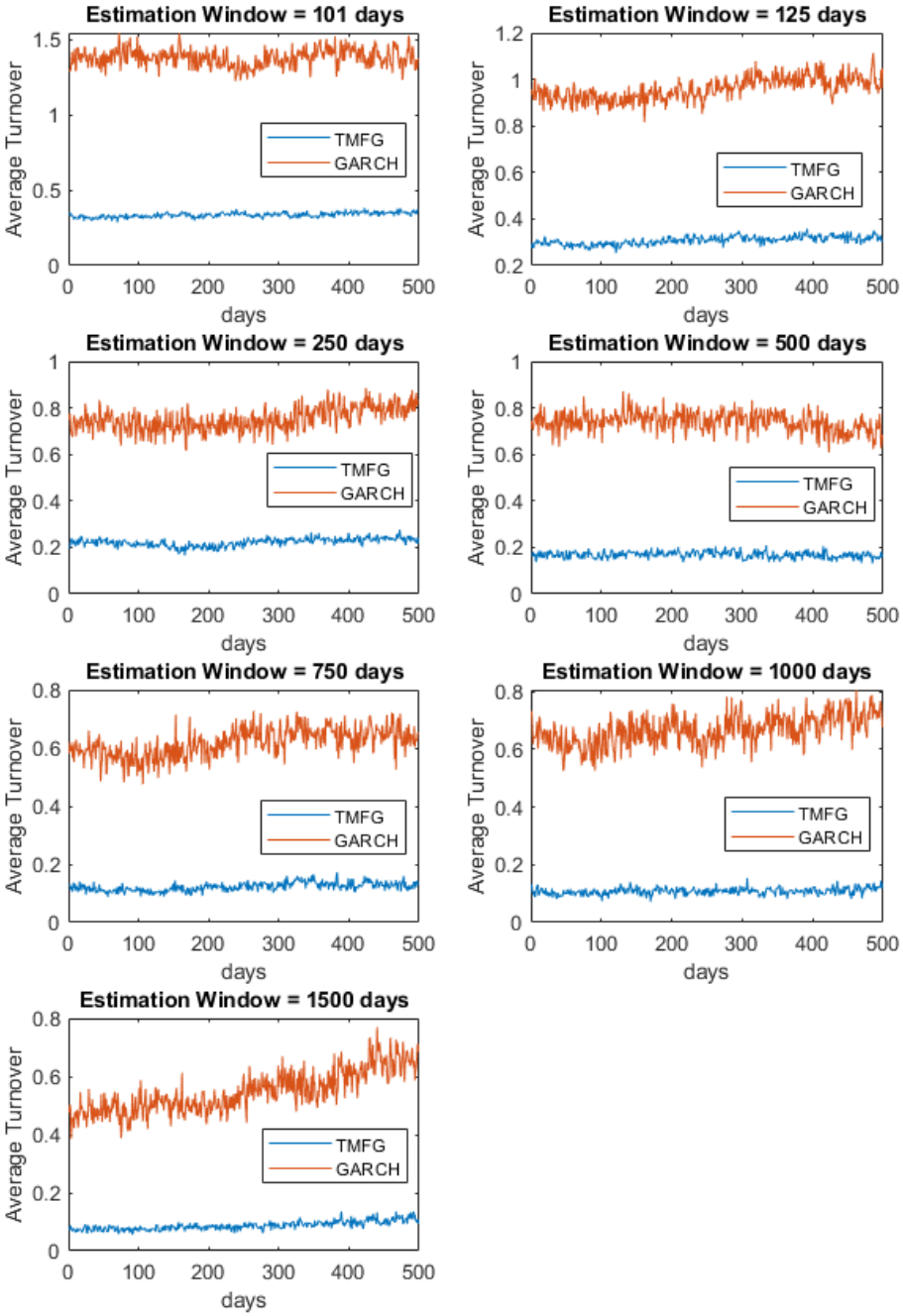


Figure 3.10: Average daily turnover across 100 resamplings for different estimation window lengths q. Comparison of minimum variance portfolios constructed based on TMFG (blue line) and O-GARCH (orange line) precision matrices.

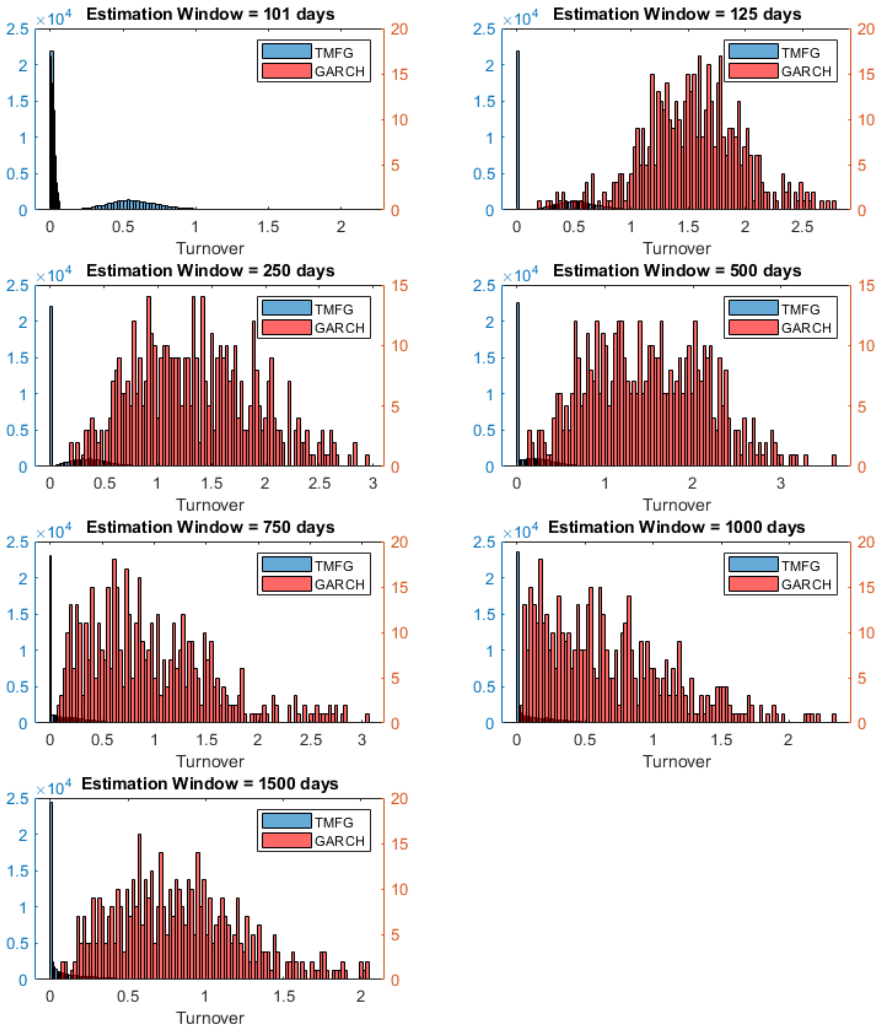
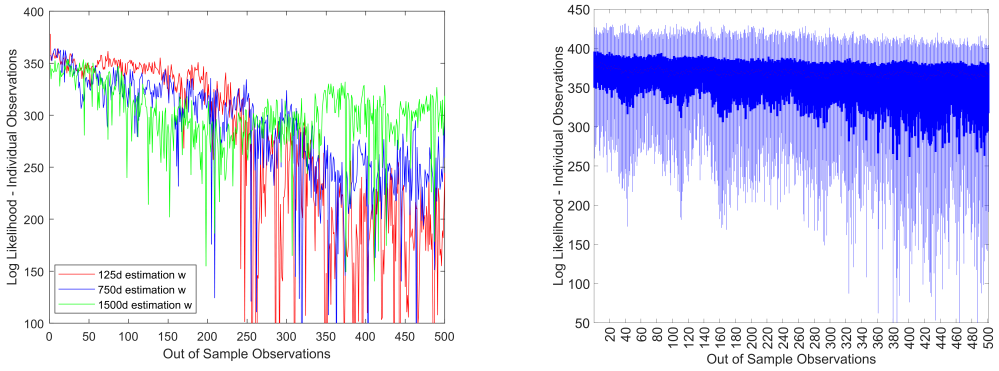


Figure 3.11: Daily turnover histogram across 100 resamplings for different estimation window lengths q. Comparison of minimum variance portfolios constructed based on TMFG (blue bars) and O-GARCH (red bars) precision matrices.

3.6 Non Stationarity

From the results discussed in the previous Section and shown in Figure 3.3 and Figure 3.2, we found that the portfolio performances improve coherently with the likelihood up until approximately 3 years of observations are used in the train set to estimate the models parameters. However, when more observations are included in the train set, the likelihood of the parameters detaches from the portfolio performances and we speculate that this is due to the role of non-stationarity. To further investigate this phenomenon, Figure 3.12 reports the likelihood corresponding to each out-of-sample observation in our experiment.



(a) Mean Likelihood for each observation across resamplings. Comparison of likelihoods obtained when 125, 750 and 1500 days are used in the train set.

(b) Boxplot of likelihoods representing the quartiles and min-max levels for each observation across resamplings, having removed outliers. The plot is for 750 days train set (blue plot on the left).

Figure 3.12: Out-of-sample likelihood measured observation-by-observation

Figure 3.12a shows the average likelihood across 100 resamplings for each out-of-sample observation. We note that when shorter estimation windows are used to estimate the models' parameters (i.e. 125 days) the likelihood is higher in the days immediately following the estimation window, but tends to rapidly decrease as the observations depart from the training window. Larger estimation windows (i.e. 750 or 1,500 days) instead, lead to a more stable likelihood in the long run, but at the cost of a lower likelihood for the observation closer to the estimation set. Figure 3.12b shows the observation-wise box-plot of the likelihood computed across the resamplings when 750 observation are used in the train set. The box plot reports the 25%-75% quantile interval (dark blue) and the max-min interval ('whiskers' light blue) having excluded the 'outliers' that are below the whiskers' [104]. Other than decreasing means, the figure shows that as the observations depart from the train set, the amount of observations posting a significantly lower likelihood increases,

together with the downside volatility. In other words, it is more likely to have observations that are far from the model estimated in sample, supporting the conclusions drawn from Figure 3.12a.

As we discussed in Procacci and Aste (2019), market states tend to be persistent in daily observations. Shorter estimation windows, therefore, tend to better describe the system belonging to the same 'state' which is likely to be persistent for adjacent observations. Notice that by considering the aggregate behaviors across 100 resamplings, we want to avoid specific market conditions and state shifts, but rather focus on the general behaviour. The evolution of the financial system is obviously very dynamic and the goodness of parameters is certainly dependent on both systemic and idiosyncratic events.

These results also provide further insights on the findings discussed in Figure 3.2 and Figure 3.3 in that our conclusions are dependent on the number of out-of-sample observations that in our case coincides with the portfolio holding period - i.e. 500 days in our experiments. Shorter estimation windows provide better fit in the short term, while larger estimation windows provide robustness in the long run. The optimal balance between these two effects depends on the holding period and in our experiments it is achieved with approximately 3 years observations in the estimation window. Short holding periods do not require robustness in the long run (i.e. shorter estimation windows would deliver better results). As the holding period increases, the long term robustness becomes more relevant than the short term fit and larger estimations windows have to be preferred.

3.6.1 A Closer Look at Likelihood

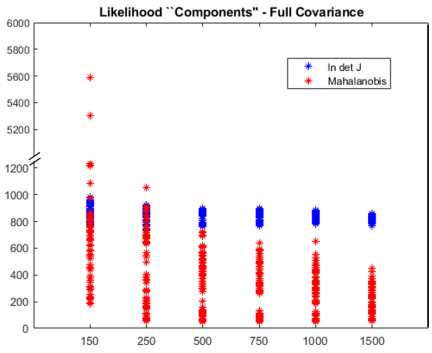
In our discussion on non-stationarity, we focused on likelihood as a standard measure of "goodness" of the estimated parameters in describing observations. As discussed in Section 3.4.3, without loss of generality for variables following a multivariate elliptical distribution of the form in Eq (3.10), the likelihood function is essentially given by the determinant of the precision matrix minus some function of the Mahalanobis distance (plus constant terms, see Eq (3.14)). In other words, we could think of the likelihood as a measure of 'distance' between the determinant and a function of the mahalanobis distance.

Figure 3.13 report the Determinant and Mahalanobis distance values computed for all resamplings and across different estimation windows (x-axis) for both the Full covariance

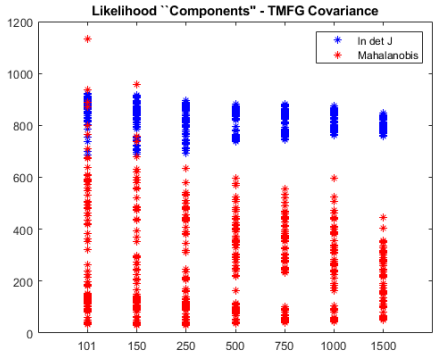
(Figure 3.13a) and the TMFG covariance (Figure 3.13b). Looking at this two objects separately, therefore, I aim at gaining intuition on a) how the two components interplay in the likelihood results presented in this Chapter and b) the effects of the TMFG-LoGo filtering.

From Figure 3.13, the first observation is that, across resamplings, the Mahalanobis distance displays a much higher variability then the determinant. The determinant (when defined) is, indeed, quite stable across resamplings both for TMFG and Full covariance matrices. The variance in both determinat and Mahalanobis decrease as the number of observations in the training set increases (leading to a more stable parameters' likelihood as described in 3.4.1). Figure 3.14, provides more interpretable insights on the effects of information filtering by reporting the difference between the Full and TMFG based Mahalanobis (Figure 3.14a) and determinants (Figure 3.14c) computed for each resampling and across all estimation windows. Here we can observe that the Full covariance actually always delivers larger Mahalanobis and larger determinant in absolute terms. Clearly the largest differences in both Mahalanobis distance and log determinant are displayed when shorter estimation windows are used. Interestingly, the difference in Mahalanobis varies significantly across resamplings, being sometimes close to zero even for shorter estimation windows, while the difference in log determinant remains more neat.

The key observation from these charts, however, is the y-axis of Figure 3.14a and Figure 3.14c in that it clearly shows that the difference in Mahalanobis distance is, in most of the cases, larger than the difference in log determinant, explaining the higher likelihood observed with the TMFG-LoGo precision matrix is used. In other words, the sparse TMFG precision matrix leads to a smaller determinant and Mahalanobis distance. The reduction in the log determinant, however, is more then compensated by a larger reduction in the Mahalanobis distance, leading to a larger likelihood.

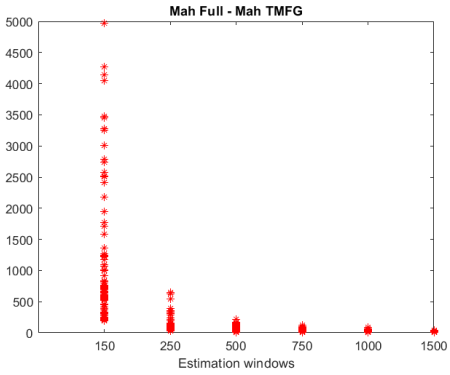


(a) Determinant and Mahalanobis distance with Full covariance. Note the y axis break.

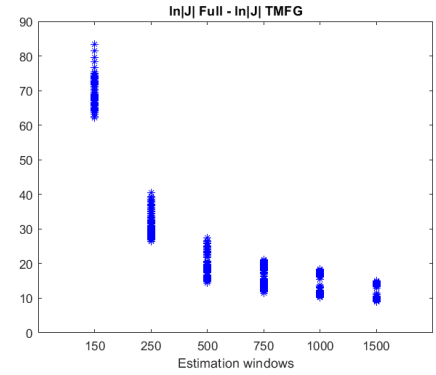


(b) Determinant and Mahalanobis distance with TMFG covariance.

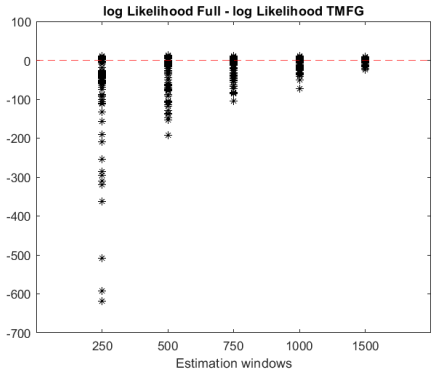
Figure 3.13: Visualizing Likelihood: Determinant and Mahalanobis distance



(a) Difference in Mahalanobis distance computed with Full and TMFG covariances



(b) Difference in log determinant computed with Full and TMFG covariances.



(c) Difference in overall log Likelihood computed with Full and TMFG covariances.

Figure 3.14: Comparison Full vs TMFG on individual likelihood components. Differences in likelihood components computed with Full and TMFG covariance for all resamplings and across estimation windows.

3.7 Discussion

Portfolio construction is a cornerstone of financial theory and practice. However, it is still today a controversial topic for both academics and practitioners. Any portfolio optimization strategy relies on assumptions and modelling of the future market structure. However, inferring such structure from past observations is a very challenging task, plagued by uncertainty around parameters estimation and relying on some non fully satisfied assumptions.

We identify three main sources of inaccuracies and errors: 1. model oversimplification; 2. limited size of the estimation set; 3. non-stationarity. We address oversimplification by introducing a modelling that uses a L_0-norm regularized elliptical multivariate distribution, demonstrating that it over-performs traditional models both in likelihood and in portfolio variance performances. We test the effect of sample size by training the models on windows of different sizes and find that performances initially increase with sample size but then eventually decrease for windows above 750 days. We attribute the initial improvement in performance to sampling error, which is reduced when more observations are included, and we interpret the decay in performance when more the 750 observations are included as an instance of non-stationarity. We further investigate this phenomenon by studying the likelihood corresponding to individual observations out-of-sample and show that shorter estimation windows deliver higher out-of-sample likelihood in the days immediately following the train window, but it tends to rapidly decrease afterwards. As more observations are included in the training set, the out-of-sample likelihood gains stability, with larger values in the long term, but at the cost of lower likelihood in the short term. We conclude that the financial system changes significantly through time and the 'optimal' fit in finance needs to be defined in terms of the holding period.

Our main contribution to the literature on portfolio construction is the demonstration of the relationship between the goodness of the model, measured as out-of-sample likelihood, and the realized portfolio volatility. We show that higher likelihood obtained with filtered TMFG-LoGo precision matrices correspond to lower portfolio volatility out-of-sample. The relationship between larger likelihood and lower realized volatility is also verified in the maximum-likelihood estimate of the covariance matrix when computed over train sets of different lengths. Further, we show that sparse, filtered covariance matrices can signifi-

cantly reduce estimation errors coming from both sampling error and non-stationarity. It also reduces many of the instability problems related to mean-variance optimal weights.

Finally, all of the analysis and conclusions drawn in this Chapter are based on different estimates of the covariance matrix. While forecasting future returns remains of primary importance in trading and wealth management, we showed that the correlation structure, sometimes overlooked in the asset allocation literature, plays a key role in portfolio construction and a good deal of performances depend upon it.

Chapter 4

Market States and Stationary Regimes

We propose a novel methodology to define, analyse and forecast market states. In our approach market states are identified by a reference sparse precision matrix and a vector of expectation values. In our procedure each multivariate observation is associated to a given market state accordingly to a minimisation of a penalized Mahalanobis distance. The procedure is made computationally very efficient and can be used with a large number of assets. We demonstrate that this procedure is successful at clustering different states of the markets in an unsupervised manner. In particular, we describe an experiment with one hundred log-returns and two states in which the methodology automatically associates states prevalently to pre- and post- crisis periods with one state gathering periods with average positive returns and the other state periods with average negative returns, therefore discovering spontaneously the common classification of 'bull' and 'bear' markets. In another experiment, with again one hundred log-returns and two states, we demonstrate that this procedure can be efficiently used to forecast off-sample future market states with significant prediction accuracy. This methodology opens the way to a range of applications in risk management and trading strategies in the context where the correlation structure plays a central role.

4.1 Introduction

Markets do not always behave in the same way. In common terminology, there are periods of 'bull' market in which prices are more likely to rise and periods of 'bear' market in which prices are more likely to fall. These different 'states' of markets are commonly attributed in literature to unobservable, or latent, regimes representing a set of macroeconomic, market and sentiment variables.

In our paper [153], Prof. Aste and I build on Hallac et al. (2017) and propose a similar Covariance based Clustering. However, we consider single observations and do not enforce Toeplitz structure on the precision matrix. We, therefore, call this methodology

ICC - Inverse Covariance Clustering. We also enforce temporal coherence by penalizing frequent switches between market states and favouring temporal consistency. Further, our framework is fully flexible to optimize any gain measure: in the experiments presented in this Chapter we do not directly maximise likelihood, but rather assign states to clusters according to their Mahalanobis distance [56]. We experiment with this methodology in the context of financial time series and provide a detailed analysis of the role played by sparsity and temporal consistency, while assessing the significance of the clusters. Finally, we show that the cluster classification can be used for one step ahead off-sample prediction.

Our approach simplifies and clarifies the definition of 'market state' by identifying each state with a sparse precision matrix and a vector of expectation values which are associated to a set of multivariate observations clustered together accordingly with a given procedure. In the following, the precision matrix of market state 'k' is denoted with $\mathbf{J}_k$ and it represents the structure of partial correlations between the system's variables. In the multivariate normal case, two nodes are conditionally independent if and only if the corresponding element of $\mathbf{J}_k$ is equal to zero. A sparse precision matrix provides an easily interpretable and intuitive structure of the market state, with all the most relevant dependencies directly interconnected in a sparse network. Furthermore, sparsity reduces the number of parameters from order n^2 (with n the number of variables) to order n preventing overfitting [106] and filtering out noisy correlations [15, 139].

The remainder of this Chapter is organized as follow: in Section 4.2 I briefly review previous works on states classification and sequence clastering and the general Baum–Welch algorithm. In Section 4.3 I introduce the ICC methodology, following the E-step and M-step typical of the EM algorithm (see Section A) and highlighting the role of sparsity in the M-step. In Section 4.4 I present a clustering experiment, comparing the ICC and GMM models in segmenting the daily returns of Russel 1000 members. In Section 4.5 I present a second experiment, in which the ICC model is used to find market states during the COVID-19 outbreak. Lastly, in Section 4.6 I highlight the main findings and conclusions derived throughout the chapter.

4.2 Literature Review

4.2.1 State Models in Time Series

Many time series models presented in literature tried to capture this phenomenon. Among the most popular methods, it is worth mentioning the TAR models [169], trying to estimate 'structural breaks' in the time series process, and the Markov switching models [83], where the change in regimes are parametrized by means of an unobserved state variable typically modelled as Markov chain. However, the application of TAR models in finance is frequently criticized since it cannot be established with certainty when a structural break has occurred in economic time series and the prior knowledge of major economic events could lead to bias in inference [40]. Markov switching models, on the other hand, are highly affected by the curse of dimensionality. In particular, for slightly more complex dynamics than the original proposal [83], we need to rely on variational inference techniques or MCMC methods [98, 170]. This implies that, in a multivariate context and particularly if we aim to extract information on the switching from the correlation structure, estimation becomes difficult to perform.

Other approaches focus on clustering of observations into groups: 'similar' data objects are discovered on the basis of some criteria for comparisons. Most works related to clustering of time series are classified into two categories: subsequence time series clustering and point clustering. Subsequence clustering involves the clustering of sliding windows of data points and usually aim at discover repeated patterns. Example are Dynamic Time Warping [112], Hierarchical methods [141] or pattern discovery [158]. In point clustering methods, instead, each multivariate observation at each time instance t is assigned to a cluster. In most popular approaches, however, this is done based on a distance metric [70, 75, 81, 87, 184].

In a multivariate context, different 'states' of markets are not only reflected in the gains and losses, but also in the relative dynamics of prices. Indeed, the correlation structure changes between bull and bear periods indicating that there are structural differences in these market states. Most common approaches in the industry assume -for convenience- a stationary correlation structure [28, 62]. However, it is well established that correlations among stocks are not constant over time [4, 113, 138] and increase substantially in periods

of high market volatility, with, asymmetrically, larger increases for downward moves (see, for example, [6, 45, 162]). Indeed, various approaches have been proposed in literature to model and predict time-varying correlations. Examples are, for instance, the generalized autoregressive conditional heteroskedasticity (GARCH) models by Bollerslev (1990) or the dynamic conditional correlation (DCC) model by Engle (2002). However, most of these models are not able to cope with more than a few assets due to the curse of dimensionality having numbers of parameters that increases super-linearly with the number of variables [54]. Other approaches have been focusing on the study of changes in a time-varying correlation matrix computed from a rolling window. This is, for instance, the case of estimators like the RiskMetrics Longerstaey and Spencer (1996) or Lee and Stevenson (2003). However, since these approaches use only a small part of the data, these estimators have large variances and, in case of high dimensionality, may lead to inconclusive estimates [102].

Hallac et al. (2017) introduced a clustering algorithm called TICC (Toeplitz Inverse Covariance Clustering), originally proposed for electric vehicles, where classification into states is constructed from a likelihood measure associated with a referential sparse precision matrix (inverse covariance matrix). Instead of considering each observation in isolation, however, in their approach they cluster short subsequences of observations so that the covariance matrix constructed on the subsequences provides a representation of the cross-time partial correlations. In this setting, then, by imposing a Toeplitz constraint to the precision matrix of each regime, the cross-time partial correlations are constrained to be constant and, hence, covariance-stationarity is enforced. This method has a number of appealing features from a financial perspective, although the structure of data considered by the authors is significantly different from noisy data in finance.

4.2.2 HMM and the Baum-Welch Algorithm

A hidden Markov model [157] describes the joint probability of a collection of 'hidden' and observed discrete random variables and relies on the assumption that the $i-th$ hidden variable given the $(i-1)-th$ hidden variable is independent of previous hidden variables, and the current observation variables depend only on the current hidden state.

Consider a discrete hidden random variable $\mathbf{X_t}$ with a finate number K of possible hidden states. Assuming $P(\mathbf{X_t}|\mathbf{X_{t-1}})$ is independent on time t, the time-independent transition

matrix is given by

$$A = \{a_{ij}\} = P(\mathbf{X_t} = j \mid \mathbf{X_{t-1}} = i) \ . \tag{4.1}$$

Consider now the discrete observation variables $\mathbf{Y}_t$ and assume it can take one of K possible values. Assuming also that any observation $\mathbf{Y_t} = y_i$ is independent on the hidden state, then the probability of any observation y_i at time t for the state $\mathbf{X}_t = j$ is given by

$$b_j(y_i) = P(\mathbf{Y_t} = y_i \mid \mathbf{X_{t-1}} = i) \ . \tag{4.2}$$

Taking into account all the values spanned by $\mathbf{Y}_t$ and $\mathbf{X}_t$, the NxK state-conditional probability matrix is given by $\mathbf{B} = \{b_j(y_i)\}$. Thus, a hidden Markov chain can be described by $\theta = (\mathbf{A}, \mathbf{B}, \pi)$.

Considering an observation sequence $\mathbf{Y}_t = (\mathbf{Y}_1 = y_1, ..., \mathbf{Y}_N = y_N)$ The Baum–Welch algorithm [17, 18, 157] leverages EM approach (discussed in Appendix A) to find the maximum likelihood estimate of the parameters of a hidden Markov model given a set of observed feature vectors - i.e. $\theta^* = \max_\theta P(\mathbf{Y} \mid \theta)$.

The algorithm starts with an initial selection for the parameters θ. Then, we estimate $\alpha(t) = P(\mathbf{Y}_1 = y_1, ..., \mathbf{Y}_t = y_t, \mathbf{X_t} = i \mid \theta)$, i.e. the joint probability of observing all of the data up to time t and state i at time t, and $\beta(t) = P(\mathbf{Y}_1 = y_1, ..., \mathbf{Y}_t = y_t \mid \mathbf{X_t} = i, \theta)$, i.e. the conditional probability of all future data from time $t+1$ to N . This is found via recursive procedure, sometimes referred to as the forward-backward algorithm and concludes the 'E' step of the procedure. Now, similarly to the 'M' step discussed in Appendix A, the results $\alpha(t)$ and $\beta(t)$ are used to find the new set of parameters $\theta^{\mathbf{new}}$. The algorithm then continue to alternate between E and M steps until convergence is satisfied.

It is worth noticing that the forward-backward algorithm is computationally expensive. For each iteration, the α recursion is a $O(K^2)$ operation and the β recursion is $O(K^2N)$. In our ICC approach discussed in the remainder of this Chapter, we propose a penalized Viterbi procedure to approximate the maximum likelihood solution for our latent variable problem. Despite not providing the full conditional likelihood of the hidden parameters, the Viterbi algorithm discussed in Section 4.3.2 allows to significantly improve the computational efficiency of the ICC methodology, making it a better fit for trading application and for handling high dimensional datasets.

4.3 Methodology

Consider a time series $\mathbf{X}$ of T multivariate observations,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 = [x_{1,1}, x_{1,2}, ..., x_{1,n}] \\ \mathbf{x}_2 = [x_{2,1}, x_{2,2}, ..., x_{2,n}] \\ \vdots \\ \mathbf{x}_T = [x_{T,1}, x_{T,2}, ..., x_{T,n}] \end{bmatrix} \tag{4.3}$$

where $\mathbf{x}_t \in \mathbb{R}^n$ is the multivariate observation at time t constituted by the log-returns of the n stocks in the panel. Our goal is to assign each of the T multivariate observation $\mathbf{x}_t$ to K clusters using the information content of the correlation structure and avoiding the curse of dimensionality [27]. We identify each state by a sparse precision matrix and a vector of expectation values which are associated to a set of multivariate observations with largest adjusted likelihood. In the following, the precision matrix of market state 'k' is denoted with $\mathbf{J}_k$ and it represents the structure of partial correlations between the system's variables. In the multivariate normal case, two nodes are conditionally independent if and only if the corresponding element of $\mathbf{J}_k$ is different form zero. A sparse precision matrix provides an easily interpretable and intuitive structure of the market state with all the most relevant dependencies directly interconnected in a sparse network. Furthermore, sparsity reduces the number of parameters from order n^2 (with n the number of variables) to order n preventing overfitting [106] and filtering out noisy correlations [15, 139].

Our approach is inspired by latent variable models, but we account for the temporal dimension by encouraging adjacent observations to belong to the same cluster. The clustering procedure uses a redesigned version of the Expectation Maximization (EM) algorithm [58, 127] reviewed and discussed in Appendix A. It starts by setting the number of clusters K (in this Chapter we limit to $K = 2$, see Chapter 6 for a discussion on the optimal number of states) and assigns multivariate observations to clusters randomly. From these K sets of data we compute the sample means $\boldsymbol{\mu}_k$ and the precision matrices $\mathbf{J}_k$ and we then iteratively re-assign points to the cluster with smallest

$$\mathcal{M}_{t,k} = d^2_{t,k} + \gamma \mathbb{1}\{\mathcal{K}_{t-1} \neq k\} \ . \tag{4.4}$$

where $d^2_{t,k} = (\mathbf{x}_t - \boldsymbol{\mu}_k)^T \mathbf{J}_k \, (\mathbf{x}_t - \boldsymbol{\mu}_k)$ is the the square Mahalanobis distance of observation

$\mathbf{x}_t$ in cluster k with respect to the cluster centroid $\boldsymbol{\mu}_k$; $\mathbf{x}_t = [x_{t,1}, x_{t,2}, ..., x_{t,n}]$ is the n-stocks multivariate observation at time t $(= 1, ..., T)$; $\boldsymbol{\mu}_k$ is the vector of the means for cluster k; $\mathbf{J}_k$ is the (sparse) precision matrix for cluster k; γ is a parameter penalizing state switching; $\mathcal{K}_{t-1}$ is the cluster assignment of the observation at time $t-1$. We considered as well clustering with respect to maximum likelihood and minimum Euclidean distance, however we report only about the procedure with Mahalanobis distance which is the one providing the best results. Specifically, Euclidean distance is very efficient in distinguishing positive and negative returns but does not distinguish well between pre- and post-crisis periods. The maximum likelihood, instead, identifies very well the crisis period but then it is much less clean in classifying the 'bull' and 'bear' market states. Let us note that the used Mahalanobis distance clustering is producing high likelihood although not maximal.

The clustering assignment procedure is made computationally efficient by using the Viterbi algorithm [27, 176] that transforms an otherwise $O(K^T)$ procedure into $O(KT)$ (Section 4.3.2). Further, the sparse precision matrix $\mathbf{J}_k$ is computed efficiently from the observations in each cluster by means of the TMFG-LoGo network filtering approach [15, 125].

In the following section, I discuss the details of both the "E-step", where we estimate parameters using TMFG-logo and the "M-step", where we assign observation to the "k-th" cluster while enforcing temporal consistency using the Viterbi algorithm.

4.3.1 M-step: TMFG-LoGo

In the *M-step* of the algorithm, the parameters associated with the distribution of each state are estimated. In our design, these distributions are defined solely in terms of the inverse covariance matrix $\mathbf{J}$ and the vector of expected values $\boldsymbol{\mu}$. As discussed in Section 2.2, the empirical estimate of the correlation structure suffers many pitfalls mostly related to overfitting and to the number of available observations. To overcome these difficulties and obtain a reliable representation of the correlation structure associated to each state, we considered a filtered, sparse inverse covariance computed by means of the TMFG-LoGo approach. This is an efficient algorithm [124] that produces a chordal graph with $3(p-2)$ edges, where p is the number of variables.

The TMFG-LoGo approach reviewed in Section 2.2.2 has proven to perform better than other filtering approaches including GLasso and Ridge providing the additional advantages of efficiency and fixed sparsity level with no need to calibrate hyperparameters

[125]. In this Section we motivate the choice of TMFG-LoGo filtering procedure in terms of statistical significance by comparing the performances of the TMFG-LoGo to the cross-validated Ridge l_2 penalized inverse covariance (Ridge) on our dataset. We considered the widely used Ridge penalization as robust estimate of the empirical inverse covariance matrix and compared it to TMFG-LoGo and show that, when applied to our dataset, TMFG-LoGo produces more stable likelihood results than Ridge. We used 40% of the data (from 31/12/2007 to 31/12/2015) as test set, and we considered as train sets the q observations preceding the test set (until 30/12/2007). The penalization parameter of Ridge was defined by cross validating within the train set. To compare TMFG-LoGo and the cross-validated Ridge we computed the log-likelihoods $\mathcal{L}_{s,k} = 1/2(\log|\mathbf{J}_k| - d^2_{s,k} - p\log(2\pi))$ using the two covariance estimates and compared them. Figure 4.1 shows the likelihood observation-wise computed in train and in test using the TMFG-LoGo and Ridge precision matrices estimated over $q = 500$ observations. The TMFG-LoGo likelihoods are much more stable over time suggesting that the procedure was successful in filtering out noise. Table 4.1 reports details on mean, 5^{th} and 95^{th} percentiles of the likelihoods computed in the train and test set. As previously mentioned, TMFG-LoGo likelihoods are much more stable with 5^{th} and 95^{th} varying a few percent only for TMFG-LoGo and instead varying of more than one order of magnitude in Ridge. We found similar results for TMFG-LoGo and Ridge when different values of q are considered. Note that Ridge log likelihoods have large differences between train and test. This is a typical indication of overfitting. Conversely, TMFG presents small differences indicating that the LoGo procedure acts as a topological penalize.

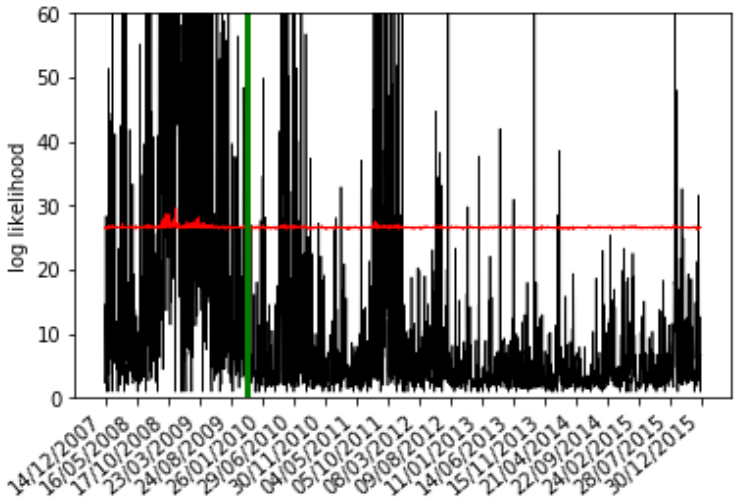


Figure 4.1: Train and test log likelihood observation-wise using TMFG (red line) and Ridge (black line) precision matrices. The green vertical line divides train and test set. Ridge peaks reach values outside the range up to 320.

Train set

	Average	$5^{th} percentile$	$95^{th} percentile$
$\mathcal{L}_{Ridge}$	41.70	2.19	188.85
$\mathcal{L}_{TMFG}$	26.71	26.53	27.22

Test Set

	Average	$5^{th} percentile$	$95^{th} percentile$
$\mathcal{L}_{Ridge}$	8.08	1.39	27.64
$\mathcal{L}_{TMFG}$	26.55	26.44	26.73

Table 4.1: TMFG and Ridge log likelihood metrics - means, 5^{th} and 95^{th} percentiles - computed in train (top panel) and test (bottom panel) set. TMFG and Ridge precision matrices are estimated using $q = 500$ observations.

4.3.2 E-step: The Viterbi Algorithm

Figure 4.3 provides a visualization of the problem of assigning points to clusters. Based on the parameters estimates ($\boldsymbol{\mu}_k$ and $\mathbf{J}_k$ via TMFG-LoGo) from the E-step of the Expectation Maximization procedure, we compute the likelihood of every multivariate observation obtaining, for each cluster k and for each observation t, a value $\mathcal{L}_{t,k}$. If we assume observations to be independent, maximizing the overall likelihood corresponds to maximize the individual likelihood at each time t. In Figure 4.3, this means choosing the cluster k that provides the highest individual likelihood $\mathcal{L}_{t,k}$ at each time-step.

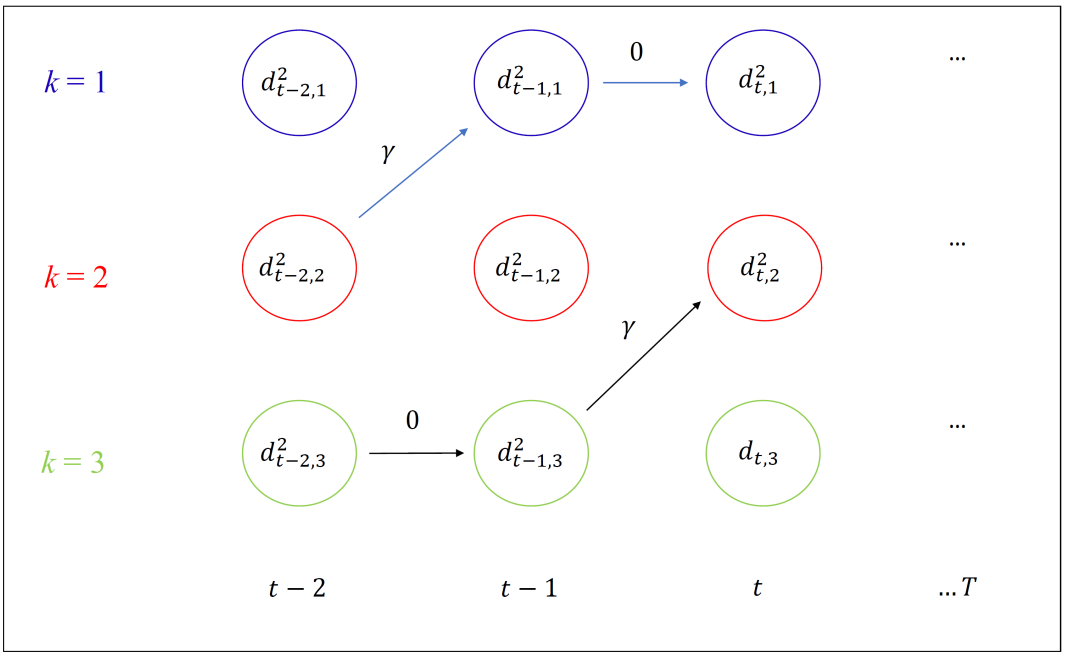


Figure 4.3: Cluster assignment paths - Sketched example. Example of two among the K^T possible paths considering $K = 3$ clusters and T observations. $\mathcal{L}_{t,j}$ represents the log likelihood of the multivariate observation at time t if assigned to cluster j. If an observation is assigned to same cluster as the previous one, no penalty is applied, otherwise a *cost* weighted by the parameter γ is added.

However, when we analyse latent states through time, we need to consider the most probable *sequence* of latent states which is not the set of most probable individual states. In particular, if we introduce a cost parameter γ that penalizes cluster switching, the problem complexity becomes combinatorial, since we need to account for the whole sequence or *path* of assignations. In particular, given K potential cluster assignment of T points (multivariate observations), the number of potential paths grows exponentially with the length of the chain to K^T possible assignments of points to clusters. Based on a dynamic programming

approach, the Viterbi algorithm [176] provides an efficient solution with complexity $O(KT)$ (*i.e., linear*) to this problem, searching the space of the paths and finding the most efficient path. The Viterbi algorithm in the convenient formulation by [82] is sketched in 2.

Algorithm 2 Viterbi algorithm

Input
$\mathcal{L}_{i,j}$ = negative log likelihood of observation i if assigned to state j
γ = time consistency parameter

Initialize
PreviousCost = array of K zeros
CurrentCost = array of K zeros
PreviousPath = array of K elements
CurrentPath = array of K elements

for each observation $i = 1, ..., T$ **do**
  for each state $j = 1, ..., K$ **do**
    MinVal = index of minimum value of PreviousCost
    if PreviousCost$[MinVal] + \gamma >$ PreviousCost$[j]$ **then**
      CurrentCost$[j]$ = PreviousCost$[j] - \mathcal{L}_{i,j}$
      CurrentPath$[j]$ = PreviousPath$[j]$.append$[j]$
    else
      CurrentCost$[j]$ = PreviousCost$[MinVal] + \gamma - \mathcal{L}_{i,j}$
      CurrentPath$[j]$ = PreviousPath$[MinVal]$.append$[j]$
    PreviousCost=CurrentCost
    PreviousPath=CurrentPath
FinalMinVal=index of minimum value of CurrCost
FinalPath=CurrPath[FinalMinVal]

For our purposes, we cannot calibrate the hyperparameter γ by cross validation. This due to the fact that the states are unobservable and model dependent. We selected, therefore, the parameter by grid searching the space of parameter γ in the range $[0,3]$ with steps 0.2 and selecting the value that maximizes the penalized joint likelihood of the sample

$$\max_{\gamma} \sum_{t=0}^{T} \mathcal{M}_{t,k} - \gamma \mathbb{1}\{\mathcal{K}_{t-1} \neq k_t\} \ , \tag{4.5}$$

where k_t is the cluster assignment of the t^{th} observation. For both experiments the maximum was found for $\gamma = 1$. Let us note that this is a meaningful result because, from an entropic perspective, a switch of state should 'cost' about one bit of information.

A more general formulation can be implemented by describing the paths as Markov chains and introducing a transition probability between the states. However, under the Markov chain formalism the expression in Eq. (5.1) for the likelihood ratio is no longer consistent because it implies implicitly IID multivariate observations.

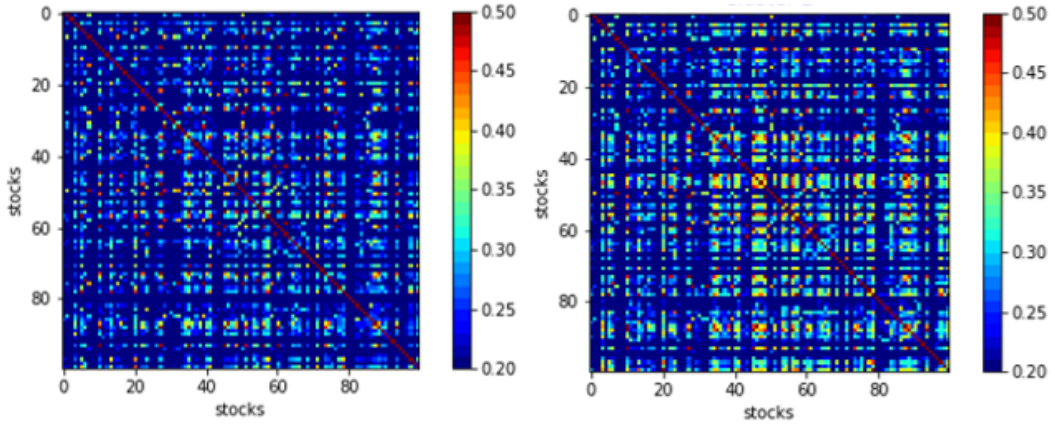
4.4 Experiment - Market States Identification

In this Section, I report results for one experiment performed over a **dataset** of daily closing prices of $n = 2490$ US stocks entering among the constituents of the Russel 1000 index (*RIY index*) traded between 02/01/1995 and 31/12/2015. For each asset $i = 1,...,n$, we calculated the corresponding daily log-returns $r_i(t) = \log(S_i(t)) - \log(S_i(t-1))$, where $S_i(t)$ is the closing price of stock i at time t.

As mentioned in the introduction of the Chapter, our primary goal is to efficiently cluster noisy, multivariate time series into meaningful regimes, while controlling for temporal consistency. In this experiment, I considered the entire dataset between 02/01/1995 and 31/12/2015 and estimated two referential market states. In order to explore the role of each building block of our algorithm and to compare it to a traditional baseline method, we investigate five models:

- a) ICC Model - Sparse precision matrix and temporal consistency
- b) ICC Model - Full precision matrix and temporal consistency
- c) ICC Model - Sparse precision matrix
- d) ICC Model - Full precision matrix
- e) Gaussian Mixture Model - Full covariance

Model (a) is the present proposed ICC methodology. Model (b) considers full precision matrices $\mathbf{J}_k$ instead of sparse ones. Model (c) relaxes temporal consistency allowing for $\gamma = 0$ in Eq. (4.6). Model (d) has $\gamma = 0$ full precision matrices. Finally, Model (e) is a conventional Gaussian Mixture Model [27] that has been chosen as a baseline method given the similarities with the ICC approach. We analysed and compared the resulting clusters both in terms of market properties to which the two clusters are associated and in terms of temporal consistency. First, we focused on a subset of 100 stocks chosen at random among those that have been continuously traded throughout the observed period. Random choice of the basket is to avoid selection bias. We then consider random resamplings to assess the robustness when different stocks are considered.



(a) *Bull* state correlation structure

(b) *Bear* state correlation structure

Figure 4.4: Estimated sparse correlation matrices for *bull* and *bear* clusters. Given the TMFG algorithm used in constructing these structures, the number of edges (non zero entries of the correlation matrix) is constant and equal to $3N-6$, where N is the number of variables or nodes in the graphical representation. The comparison among panel(a) and panel(b) shows that many partial correlations increase significantly in Cluster 2 (stress state) panel(b), consistently with the other clusters features analysed and validating our findings coherently with previous results [6, 45, 138]

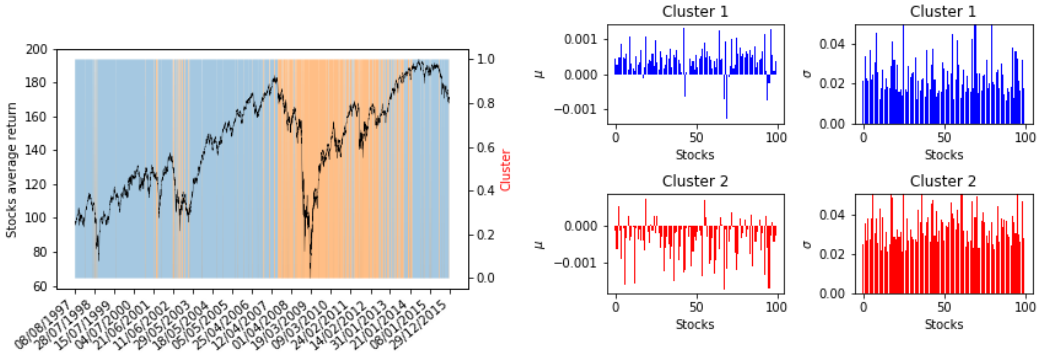
.

4.4.1 ICC Clusters Evaluation

We optimized the temporal consistency parameter by grid-searching as described in Section 4.3.2 and used $\gamma = 16$ for ICC Sparse (a) and $\gamma = 14.7$ for ICC Full (b) in both the experiments presented in this Chapter. The two referential precision matrices, $\mathbf{J}_1$ and $\mathbf{J}_2$, obtained with this experiment and had 344 non-zero entries (dependency network edges) of which 181 were common to both states showing a good level of differentiation but also significant overlaps between the two market states. Figure 4.4 presents the the two corresponding correlation matrix showing a significant higher correlation in cluster 2 than in cluster 1.

The number of points assigned to each cluster were respectively 3295 for cluster 1 and 1704 for cluster 2. Figure 4.7 reports with colored background the points' assignment for the two clusters. We can observe there is a good spatial consistency. For instance, the average number of consecutive days in cluster 1 is 27.6 days. We also note that cluster 1 (blue background) tends to be associated with periods of rising market prices whereas cluster 2 (orange background) appears more present during crisis and market downturns. We indeed discovered that -automatically- the methodology assigns *'bull'* market periods (positive mean returns) to cluster 1 and *'bear'* market periods (negative mean returns) to

cluster 2. We can for instance observe in Figure 4.5a that 52 consecutive observations during the 2001-2002 *.com* bubble crisis and 211 consecutive observations during the 2007-2008 global financial crisis have been assigned to the *bear* cluster 2. From Fig.4.5b we observe that the *bull* cluster 1 has, indeed, average positive returns for all stocks whereas the *bear* cluster 2 has average negative returns. Furthermore, also the standard deviations are different between the two cluster assignments.



(a) Time series segmentation results

(b) Mean (*left*) and standard deviation (*right*) of each stock for each temporal cluster

Figure 4.5: Clustering segmentation for experiment 1 over the whole dataset. Panel (a) reports the cumulative average return at each time t across the 100 stocks; in this picture, the blue background corresponds to time instances assigned to Cluster 1 and the orange background correspond instead to time instances assigned to Cluster 2. Panel (b) reports mean and standard deviation of each of the 100 stocks respectively computed using the returns assigned to each of the 2 clusters. We observe that Cluster 1 exhibits positive mean returns (*'bull'* state) and lower levels of volatility for all the considered stocks, while for cluster 2 all the stocks present negative mean returns (*'bear'* state) and higher levels of volatility.

To compare the two clusters on a risk-adjusted basis, we computed the Sharpe ratio [163, 164] for each stock in each cluster. We found for the *bull* cluster an average annualized Sharpe ratio equal to 1.2, with 5^{th} and 95^{th} percentiles respectively equal to 0.84 and 1.78, while the *bear* cluster had average -0.96, with -1.03 and -0.24 as 5^{th} and 95^{th} percentiles. It is, therefore, clear that the two clusters have very different risk-return profiles. Figure 4.6 reports the Sharpe ratios in the two clusters for the 100 stocks. In order to verify robustness and generality of the results we computed the same quantities for 100 other randomly chosen baskets of 100 stocks. For all resampled baskets of stocks we found a consistent clusterization in *bull* and *bear* regimes with Sharpe ratios for at least 75% of stocks larger than zero for the bull state and significantly smaller than zero for the bear state. Across the 100 resamplings, the two clusters had average number of elements respectively

equal to 3451 and 1293.

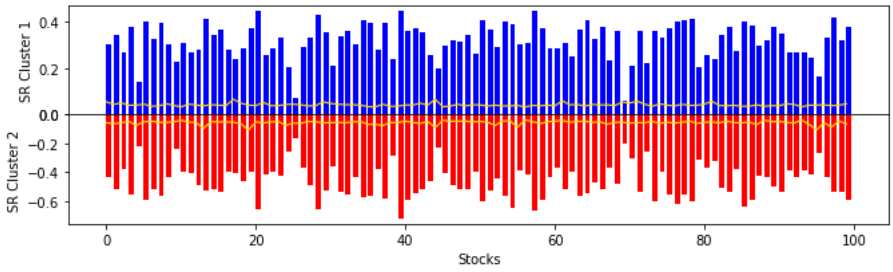
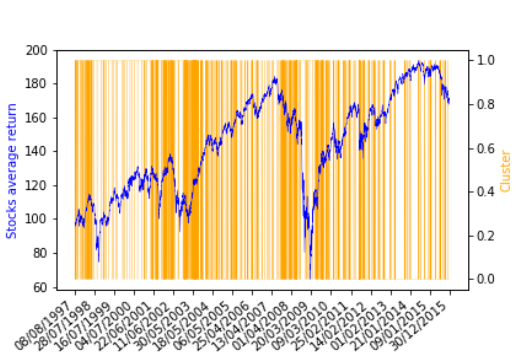


Figure 4.6: Estimated Sharpe Ratio (SR) for each of the 100 stocks in the sample. The blue bars report the SR computed from log-returns in Cluster 1, whereas the red bars report the SR computed from log-returns in Cluster 2. The gold lines represent the significance levels for which $|SR|$ is significantly different form zero in cluster 1 or cluster 2 at a significance level of 0.01.

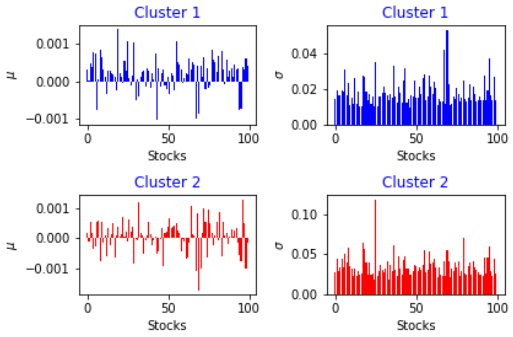
4.4.2 GMM Clusters Evaluation

We estimated clusters and parameters based on the Gaussian Mixture Model (denoted as GMM) as baseline method and we performed the same analyses and testing procedures as in previous section. The segmentation and estimation procedure follows the Expectation Maximization algorithm as described in Section A.1. In this classical model, full precision matrices, $\mathbf{J}_1$ and $\mathbf{J}_2$, are considered and temporal consistency is not enforced. It is worth empathizing that by neglecting the temporal dynamics we treat each observation as independent and, therefore, maximizing the likelihood of the sample is equivalent to maximizing the likelihood of each observation. Thus, when assigning clusters' points in the E-step, each observation is assigned to the cluster that maximizes his likelihood yielding a problem with complexity $O(T)$.

Figure 4.7 reports the point assigned to cluster 1 with white background and orange background the points' assignment for clusters 2 with, respectively, 2766 and 2033 number of points assigned to each cluster. The average number of consecutive days in cluster 1 is 9.3 and 14.4 in cluster 2, revealing a lower temporal consistency than observed in previous section. From Figure 4.7b we can observe mixed average returns in the two clusters and similar levels of volatility. Computing the Sharpe ratio for each stock in each cluster, we found for cluster 1 an average Sharpe ratio equal to 0.015, with 5^{th} and 95^{th} percentiles respectively equal to -0.023 and 0.055, while cluster 2 had average 0.003, with -0.02 and 0.03 as 5^{th} and 95^{th} percentiles. Figure 4.8 presents the computed Sharpe ratios for each



(a) Time series segmentation results



(b) Mean (*left*) and variance (*right*) of each stock for each temporal cluster

Figure 4.7: Clustering segmentation for experiment 1 over the whole dataset. Panel (a) reports the cumulative average return at each time t across the 100 stocks; the white background corresponds to time instances assigned to Cluster 1 and the orange background correspond instead to time instances assigned to Cluster 2. Panel (b) reports mean and standard deviation of each of the 100 stocks respectively computed using the returns assigned to each of the 2 clusters. Differently from the results is Figure 4.5b, stocks do not exhibit a structural different behaviour across the two clusters.

stocks in the two clusters and the corresponding 0.01 significance levels. It is possible to observe that only seven stocks out of 100 in the sample present a Sharpe ratio significantly bigger than 0 in cluster 1 while none is significantly bigger or smaller than 0 in cluster 2.

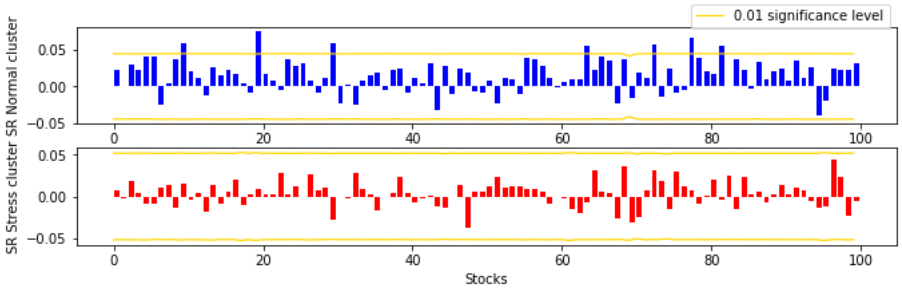


Figure 4.8: Estimated Sharpe Ratio (SR) using GMM clusters. SR for each of the 100 stocks in the sample considering the GMM clusters. The gold lines represent the significance levels at 0.01.

4.4.3 Sparsity and Temporal Consistency

In order to assess the role of sparsity and temporal consistency, we performed the same analysis on the 'alternative' ICC Models (b)-(d) and the GMM (e).

Table 4.2 summarizes the number of stocks having positive/negative Sharpe ratio in both clusters over 100 resamplings. In the table, each couple refers to the number of stocks having positive SR in *bull* (left) and negative SR in *bear* (right) states. We found that, in absence of temporal consistency constraints, both the ICC models (c, d) meaningfully clas-

sify clusters with and without sparsity. However, when temporal consistency is considered, ICC Full (b) is significantly affected by the constraint while ICC Sparse (a) provides robust results. GMM delivered the worst clusters in terms of risk/return significance.

	Median	5^{th} **percentile**	95^{th} **percentile**
GMM	(69,64)	(48,53)	(75,81)
ICC **Full,** $\gamma = 0$	(77,78)	(67,71)	(92,98)
ICC **Sparse,** $\gamma = 0$	(85,87)	(69,75)	(96,95)
ICC **Full,** $\gamma = 14.7$	(73,74)	(68,65)	(78,80)
ICC **Sparse,** $\gamma = 16$	(75,81)	(65,69)	(86,90)

Table 4.2: Positive/Negative Sharpe ratio for (‘*bull*’,‘*bear*’) states. Median, 5^{th} and 95^{th} percentiles obtained from 100 random resamples of the stocks composing the dataset.

Focusing on temporal consistency, Table 4.3 reports the number of switches and the segment length resulting from the cluster assignments of the five models. When no temporal consistency is enforced (c,d), ICC provides the less temporal consistent results with small differences related to sparsity. This also explains the good results obtained by the models in terms of risk/return significance. When constrained to be temporal consistent, ICC Full (b) shows large variability in temporal consistency across samples with some having only a few switches over the whole period and others having several hundreds. ICC Sparse (a) is instead more consistent with a few hundred switches over the whole period which are less than 1/3 of the switches in GMM (e).

	Number of Switches		
	Median	5^{th} **percentile**	95^{th} **percentile**
GMM	785	540	874
ICC **Full,** $\gamma = 0$	1203	992	2176
ICC **Sparse,** $\gamma = 0$	1157	727	1421
ICC **Full,** $\gamma = 14.7$	204	120	306
ICC **Sparse,** $\gamma = 16$	208	54	298

	Segment length		
	Median	5^{th} **percentile**	95^{th} **percentile**
GMM	5.07	2.4	11.8
ICC **Full,** $\gamma = 0$	3.3	1.68	4.38
ICC **Sparse,** $\gamma = 0$	3.5	2.8	6.65
ICC **Full,** $\gamma = 14.7$	22.64	14.6	38.26
ICC **Sparse,** $\gamma = 16$	23.6	18	55.27

Table 4.3: Temporal consistency metrics. Number of switchings and Segment lengths over 100 resampligs.

4.5 Market structure dynamics during COVID-19 outbreak

COVID-19 outbreak is an unprecedented event in modern human history with potentially catastrophic human consequences. The pandemic has had and is still having profound effects on society, the economy and the financial system. In [155], we investigated the how markets reacted to the covid-19 outbreak and particularly how correlation-driven market states evolved and reacted to the instability and volatility of the market.

4.5.1 Methods

We considered the ICC methodology described in Section 4.3 to automatically extract four inherent market-structures associated with a set of 623 equities continuously traded in the US market during the period from February 1999 to March 20, 2020. The clustering was performed by maximising the following adjusted log-likelihood:

$$\tilde{\mathcal{L}}_{t,k} = -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_k)^T \mathbf{J}_k (\mathbf{x}_t - \boldsymbol{\mu}_k) + \frac{1}{2}\log|\mathbf{J}_k| - \gamma \mathbb{1}\{\mathcal{K}_{t-1} \neq k\}. \tag{4.6}$$

where $\mathbf{x}_t \in \mathbb{R}^{n,1}$ is the vector of log-returns at time t; $\boldsymbol{\mu}_k \in \mathbb{R}^{n,1}$ is the vector of the expected values for cluster k; $\mathbf{J}_k \in \mathbb{R}^{n,n}$ is the sparse precision matrix for cluster k computed via the TMFG-LoGo method ; γ is a parameter penalizing state switching. In the present analysis we use $\gamma = 100$, but results are consistent across a large range of values of this parameter. Note that the present approach is slightly different from methodology outlines in Section 4.3 where the Mahalanobis distance was minimized instead. For the purpose of this experiment, I focused on four clusters but the outcomes are robust with respect to the number of clusters and analogous results can be obtained for two or six clusters as well.

4.5.2 Results

Fig 4.10 reports the clustering structure obtained. In the chart, the bars height illustrates the daily mean market price (y-axis, in 10^3 \$ units) while the bars' color intifies the state to which each daily observation is allocated. Note the central part of the 2008 crisis is associated with a state (blue bars) that has again become prevalent during the last few weeks spanned by the dataset (see inset). We compare the likelihood of this 'crisis' state with the likelihood associated with the state which is instead prevalent during the long *'bull'* period post 2008 (green bars). The result is shown in Fig 4.11 where the logarithm of the ratio between the likelihoods of the crisis and *bull* states is reported. We note that

the *'bull'* state prevails until February 2020 producing a negative log-ratio, afterwards the crisis-state becomes more representative and eventually becomes extremely dominant in March. The timing of the surge in the dominance of the crisis-state is consistent with that of the surge of US confirmed cases. It must be noted that this experiment focuses on the first 'outbreack' covid phase and thus the number of covid related observations is clearly very limited. However, the evidences from this experiment suggest the market features experiences during the initial covid outbreack may ultimately be classified as distinct from that of the 2008 crisis with some similarities with the late 90' states.

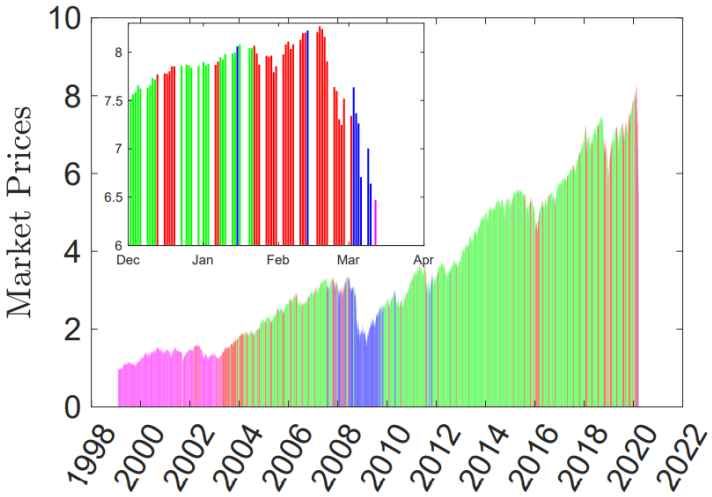


Figure 4.10: Market states during the period 02/1999 to 03/2020. The y-axis, in 10^3 $ units, reports and average daily market price and the color of the bars correspond to the market-state assigned by the ICC procedure.

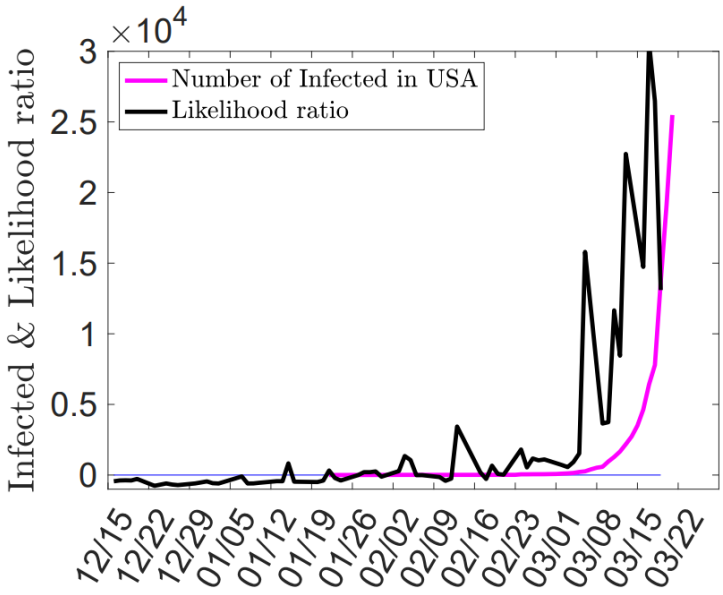


Figure 4.11: COVID-19 infected cases vs. Market States likelihood. Comparison between the number of COVID-19 infected cases in US and the logarithm of the ratio between the likelihood of the bull-state (green bars in Fig.4.10) and the stress-state (blue bars in Fig.4.10). The log-likelihood ration is scaled by a factor 5 to keep the same y-axis. The period is from December 15 2019 to March 20 2020

4.6 Discussion

In this Chapter I presented a novel methodology to define, identify and classify market states. The ICC methodology should be intended as an open framework with several methodological choices that can be modified and further investigated in future work. For instance, the segmentation with the Mahalanobis distance turned out to be a powerful tool in the reported experiments, however there is a broad range of possible metrics for clustering and experiments with Euclidean distance or Likelihood also produce interesting results. Further, the choice of TMFG network over other possible information filtering networks or other sparsification methodologies can be investigated. All these and other methodological choices have been motivated by simplicity and intuitiveness.

Lastly, I reported two experiments to illustrate that the method is efficient and reliable in identifying accurate and interpretable structures in multivariate, non-stationary financial datasets. In the first experiment discussed in Section 4.4 we imposed only two clusters motivated by simplicity. The fact that they turned out to be respectively populated mostly with average positive and negative returns associated with pre- and post-crisis periods was unexpected by us and opens potentials for completely novel ways to use multivariate analytics

for the forecasting of stock market returns. This also greatly simplified the interpretation of these states as *'bull'* and *'bear'* markets. In the COVID-19 case study in Section 4.5 we allow for higher flexibility. Of course, in reality, there are more than two market states and common definition of bull and bear markets are often blurry. Chapter 6 is devoted to the analysis of how the number of states spanned by the model impacts the states as the system evolves through time.

Chapter 5

Market States Forecasting and Trading

In this Chapter I apply the ICC methodology to forecast future states of the market form previous observations and present a simple application to equity systematic trading. The first section of the chapter focuses on forecasting. I experiment using the states likelihood ratio as predictor to forecast one-day-ahead market states. Experiments with both Logistic regression and SVM framewroks are presented delivering an accuracy higher than 50%. The second section presents a simple trading buy/sell strategy based on the forecasted market state and shows that the strategy outperformed the buy-and-hold benchmark.

5.1 States Forecasting

In this Chapter I present a set of experiments where the ICC methodology is used to forecast future states of the market from previous observations. To this end, we used the first 60% of the data (from 01/02/1995 to 12/31/2007) as train set from which we extracted the two referential precision matrices and means ($\mathbf{J}_1$, $\boldsymbol{\mu}_1$) and ($\mathbf{J}_2$, $\boldsymbol{\mu}_2$). We then forecasted the probability that, given an observation at time t, the observation at a following time $t+h$ would belong to state k.

We used the log likelihood ratio of the two clusters [142] from a rolling window of length Δ:

$$\mathcal{R}_t = \sum_{s=t-\Delta+1}^{t} \mathcal{L}_{s,1} - \mathcal{L}_{s,2} \ , \tag{5.1}$$

where $\mathcal{L}_{s,k}$ are the same as the adjusted log-likelihood $\tilde{\mathcal{L}}_{t,k}$ in Eq. (4.6) but with $\gamma = 0$. In our experiment, we considered $\Delta = 28$ days since this is the average length of segments obtained from our clustering procedure in the first experiment. Figure 5.1 provides a visual representation of the likelihood ratio computed for each cluster and of its evolution as compared to market movements. The vertical line divides the train set from the test set.

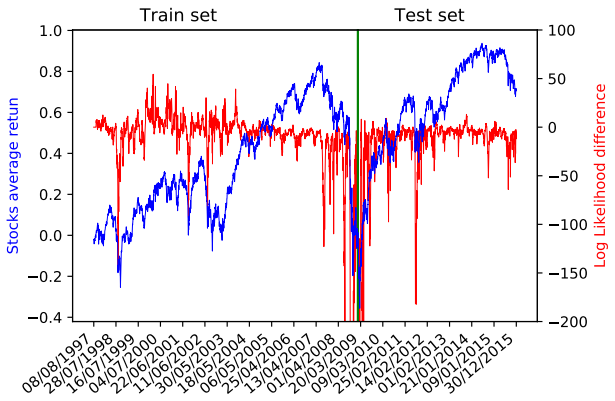


Figure 5.1: Log likelihood ratio and mean returns across train and test sets. The log likelihood ratio of the two states $\mathcal{R}_t$ was computed using using the $\Delta = 28$ *days*. The green vertical bar indicates the end of the train set and the beginning of the test set. We estimated $mathbf{J}_1$ and $\mathbf{J_2}$ in train and held it fixed for the computation of $\mathcal{R}_t$ also in the test set. The black horizontal line identifies $\mathcal{R}_t = 0$ level, *i.e.* the level above which the *bull* state is more likely. Coherently with previous findings, we can identify persistent market states with a more frequent bull market and regions of bear market.

We learned two models: a Logistic Regression and a Support Vector Machine with radial basis function (RBF) kernel. To assess the goodness of our approach we compared test set predictions with the classification performed over the whole period in the first experiment (see Figure 4.7). We used three metrics [92] to assess the performance of our classification method: the True Positive Rate TPR (number of elements correctly assigned to cluster 1 divided by total number of elements in cluster 1), the True Negative Rate TNR (number of elements correctly assigned to cluster 2 divided by total number of elements in cluster 2) and Accuracy ACC (number of correct predictions in cluster 1 or 2 divided by total number of elements).

5.1.1 Logistic Regression

In a first experiment, we fit a logistic regression of market states $\mathcal{K}_t$ against the log likelihood ratio $\mathcal{R}_t$. This model can be written as

$$P(\mathcal{K}_{t+h} = 1,2 \,|\, \mathcal{R}_t = x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad , \tag{5.2}$$

where the parameters β_0 and β_1 are estimated through maximum likelihood [27]. We estimated all parameters ($\mathbf{J}_1$, $\mathbf{J}_2$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, γ, β_0 and β_1) in the train set and then we used these

parameters to predict, in the test set, the next day state given the log-likelihood ratio $\mathcal{R}_t = x$. Specifically, we predict $\widehat{\mathcal{K}}_{t+1} = 1$ if $P(\mathcal{K}_{t+1} = 1 \,|\, \mathcal{R}_t = x) > 0.5$ and $\widehat{\mathcal{K}}_{t+1} = 2$ otherwise. For instance, for the day 30-Jan-2008 (test set) we predicted a *bear* state with probability $P(\mathcal{K}_{30-Jan} = 2 \,|\, \mathcal{R}_{29-Jan}) = 0.72$, where $\mathcal{R}_{29-Jan}$ was computed using the observations from 02-Jan to 29-Jan-2008 ($\Delta = 28$ days, all in the test set) and the parameters $\boldsymbol{\mu}_k$, $\mathbf{J}_k$, γ, β_0 and β_1 were the ones calibrated on the train set with data until 12/31/2007.

Computing the performance metrics, we obtained $TPR = 0.94$, $TNR = 0.28$ and $ACC = 0.66$ in the test set. In order to test for the robustness of our method, we randomly resampled the 100 stocks and performed the classification experiment considering the new dataset. We repeated this process 100 times and stored the three performance metrics TPR, TNR and ACC. Table 5.1.1 presents a summary of the results obtained. As we can see, ACC is higher than 50% and TPR is higher than 80% at the 5^{th} percentile, however TNR is low with median 38% and above 52% only at the 95^{th} percentile. This indicates that there is a tendency to over-assign time-instances to cluster 1 (*bull* state) and conversely missing predictions for the less frequent *bear* state. Nonetheless, we verified (by using the hypergeometric distribution as reported in [11]) that, despite their low values, these TNR are statistically significant at 0.01 level indicating that there is, indeed, significant prediction power also for the bear state. Let us stress that the present forecasting exercise is not optimized and there are several ways these performances can be improved. For instance, we verified that by introducing an adjustable threshold different from 0.5 in the logistic regression we obtain better results for TNR and ACC. However, this is beyond the purpose of this experiment where we privileged simplicity over performances.

	Median	5^{th} **percentile**	95^{th} **percentile**
TPR	0.93	0.85	0.99
TNR	0.38	0.13	0.52
ACC	0.69	0.53	0.85

Table 5.1: Out-of-sample performance metrics of LR classifier. Median, 5^{th} and 95^{th} percentiles obtained from 100 random resamples of the stocks composing the dataset.

5.1.2 Support Vector Machine

In a second experiment, we learned a Support Vector Machine considering market states $\mathcal{K}_t$ as class labels and the log likelihood ratio $\mathcal{R}_t$ as feature. In order to exploit and analyse non linear relations, we considered a RBF kernel. Considering our two classes classification problem, we shall define the $n-th$ target class (clusters) label as $t_n \in \{-1,1\}$ such that $t_n = -1$ if the $n-th$ observation belongs to cluster 0 and $t_n = 1$ otherwise. The Support Vector Machine model learns a decision bound of the form:

$$y(x) = w^T \phi(x) + b \tag{5.3}$$

where w is the parameter to be estimated; b is an explicit bias term; $\phi(x)$ is a feature space transformation such that the kernel function is a RBF kernel: $K(\mathbf{x},\mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}') = k(||\mathbf{x}-\mathbf{x}'||)$. The RBF kernel nonlinearly maps observations into a higher dimensional (virtually infinite) space and it can thus handle cases in which the relationship between features and class labels is non linear. Moreover, the RBF kernel is a convenient choice due to the lower number of parameters to be estimated (C,γ) with respect to a polynomial kernel and due to the fewer numerical difficulties [175].

The model is estimated solving a quadratic programming optimization problem described in Appendix C. First, we optimized the kernel parameters (C,γ) by grid searching the parameter space $\gamma \in [1,200]$ and $C \in [1,100]$ with unitary steps and by cross-validating within the train set. Given the optimal parameters $\gamma^* = 7$ and $C = 12$, we then estimated the parameters (w and b) of Eq. (5.3) in the training set and used these parameters to forecast the test set according to the sign of $y(x)$. Specifically, we predict $\widehat{\mathcal{K}}_{t+1} = 1$ (*that is*, $t_{t+1} = 1$) if $y(x_t) > 0$ and $\widehat{\mathcal{K}}_{t+1} = 0$ otherwise. In practice, the expression for $y(x)$ is computed considering the *dual representation* and by means of the support vectors (Eq. (C.16)) coming from the dual Lagrangian as described in Appendix C.

	Median	5^{th} **percentile**	95^{th} **percentile**
TPR	0.60	0.52	0.79
TNR	0.40	0.23	0.58
ACC	0.53	0.5	0.78

Table 5.2: Out-of-sample performance metrics of SVM classifier. Median, 5^{th} and 95^{th} percentiles obtained from 100 random resamples of the stocks composing the dataset.

We computed the same prediction metrics as from previous section obtaining $TPR = 0.63$, $TNR = 0.41$ and $ACC = 0.55$ in the test set. Table 5.1.2 presents, instead, the results, all statistically significant at 0.01 level, for the same metrics for the 100 resampligs. ACC and TPR have, respectively, 0.53 and 0.6 median, with TPR being above 23% at 5^{th} percentile. SVM produced clearly more balanced results and the over prediction of cluster 1 noticed in the LR experiment does not appear.

5.2 Trading Strategy and Backtesting

In this section, I present a simple trading strategy coming from a direct application of our methodology and inspired by the discovered features of the two clusters. As for the prediction experiments in Section 5, we used the first 60% of the data (from 01/02/1995 to 12/31/2007) as train set from which we extracted the two referential precision matrices and means ($\mathbf{J}_1$, $\boldsymbol{\mu}_1$) and ($\mathbf{J}_2$, $\boldsymbol{\mu}_2$) and we held them fixed throughout the whole test set. Note that this is a conservative and convenient assumption motivated by simplicity and expository clarity, since in real trading conditions new information would be included and a rolling or sliding window scheme would be considered to include each daily new observation in the estimation procedure. In order to focus on the results of our market states forecast ability and to obtain a performance that is not affected by other allocation decisions, we limit our analysis to buy/sell timing decisions without considering optimal allocation procedures. To this extent, we considered the equally weighted portfolio:

$$w_i^{eq} = \frac{1}{N} \tag{5.4}$$

where N is the number of stocks in portfolio, and, in our application, $N = 100$ and the the return at time t on the equally weighted portfolio is given by the average return on the market:

$$r_t^{eq} = \sum_{i=1}^{N} w_i^{eq} x_{i,t} = \frac{1}{N} \sum_{i=1}^{N} x_{i,t} \tag{5.5}$$

Other than being a convenient assumption for our experiment, the equally weighted portfolio provide good diversification and robust performances [78, 167] other than being often considered as benchmark among academics and practitioners [63].

In Section 5.2.1 we present the strategy and the backtest in both training and test set as compared to the equally weighted portfolio; in Section 5.2.2 we performed a series of tests to asses the general validity of our strategy and robustness of the results presented.

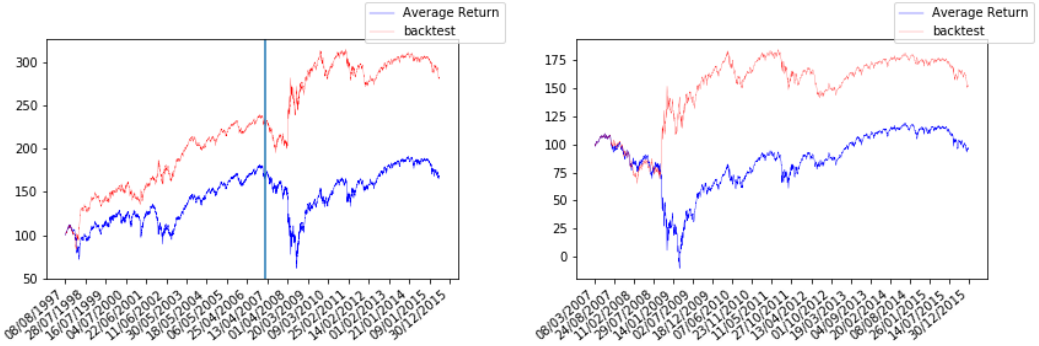
5.2.1 Strategy and Results

One of the main improvements of our methodology with respect to classical latent variables modelling approaches when applied to financial data is the spatial consistency of the discovered states. As discussed in Section 4.3, adjacent observations are encouraged to belong to same state and we obtained an average cluster length (subsequent observations assigned to the same cluster) of 28 days, as outlined in Section 4.4.1. This result is coherent with the documented persistence of returns' correlation structure (Section 2.2.3) and with the investment objective of minimizing rebalancing of portfolios. To exploit this facts, we designed a simple trading strategy considering again the log likelihood ratio $\mathcal{R}_t$ defined in Eq. (5.1) with rolling windows of length $\theta = 28$ days. The trading rule is a decision threshold on the value of $\mathcal{R}_t$:

$$TS(\mathcal{R}_t) = \begin{cases} buy & \text{for } \mathcal{R}_t \geq \frac{\sigma_{\mathcal{R}}}{4} \\ sell & \text{for } \mathcal{R}_t < \frac{\sigma_{\mathcal{R}}}{4} \end{cases} \tag{5.6}$$

where $\sigma_{\mathcal{R}}$ is the standard deviation of $\mathcal{R}_t$. Following Eq. (5.2.1), at time $t+1$ we are long the equally weighted portfolio if $\mathcal{R}_t \geq \frac{\sigma_{\mathcal{R}}}{4}$ and we are short otherwise. The financial intuition behind this trading rule is that *bull* states are more frequent and last longer than *bear* states. Moreover, short (*sell*) positions are riskier than long position. We require, therefore, a significant evidence to enter a short position.

We estimated the parameters ($\mathbf{J}_1$, $\mathbf{J}_2$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, γ and $\sigma_{\mathcal{R}}$) in the training set and used this to compute $\mathcal{R}_t$ in both training and test set and to execute the strategy Eq. (5.2.1).



(a) Backtesting, train + test sets. **(b)** Backtesting, test set only.

Figure 5.2: Strategy cumulative return compared to equally weighted portfolio. In-sample and out-of-sample performance of the proposed strategy compared to the benhmark equally weighted portfolio. Panel (a) shows the result over the whole time series and Panel (b) present the performance considering only the test set.

Figure 5.2 presents the cumulative return of the strategy as compared to the equally weighted portfolio in both training and test set while Table 5.2.1 presents a summary of the risk-return performances in test set only. We found that our strategy outperformed the equally weighted portfolio by 69.46% within the 8 years spanned by the test set, from 01/01/2008 to 30/12/2015, and by more than 120% considering also the training set. It is interesting to notice how a good portion of the overperfomance is due to the ability of the strategy to predict the main crisis events (1998, 2001-2003, 2008-2009, 2011). The two performances presented a similar level of volatility in test set with standard deviations of 1.57% for our strategy and 1.59% for the equally weighted portfolio leading to a Sharpe ratio of, respectively, 1.75% and -0.145% in test set. It is worth empathizing that these are daily metrics and that a daily SR of 1.75% corresponds to an annual of 0.3 considering 252 trading days a year.

	Average Return	**Volatility (σ)**	**Sharpe Ratio**
Strategy	0.029%	1.57%	1.75%
EQ Portfolio	-0.0025%	1.59%	-0.145%

Table 5.3: Performance metrics of the strategy and equally weighted portfolio. Out-of-sample risk-return performance metrics of the proposed strategy compared to the benchmark equally weighted portfolio. Daily average return, volatility and Sharpe Ratio obtained from the backtested series.

5.2.2 Robustness testing

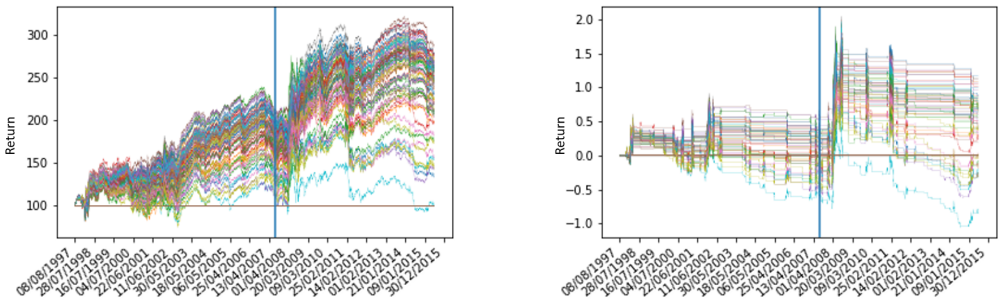
Backtesting procedures and results are often considered to be not reliable by practitioners and academics since the obtained performances might be the result of specific conditions or selected variables that undermine the general validity of a strategy and do not apply in other circumstances. However, when considering financial time series, overperforming the 'average market return' (*that is*, the 'average investor') implies the existence of information or structures not considered by the market [63, 66] and the existence of sources of inefficiencies that can be specific to only certain asset classes or individual securities. A complete discussion of market efficiency is beyond the scope of this thesis, but it is important to remind that the *general validity* of a strategy or investment approach should be assessed carefully and that profitable strategies can be, and often are, not *generally valid*.

In order to test for the robustness of our strategy, in this section we present two tests in which we tweak relevant variables of our procedure and a third test to compare the results obtained by our methodology to a trivial clustering procedure.

Boostrapping

The first test conducted aims at validating our procedure and the obtained results considering different stocks. To this extent, following an approach similar to the validating procedure in Chapter 5, we randomly resampled the 100 stocks, segmented and clustered the observations following the procedure in Section 4.3 and backtested the strategy described in previous section. We repeated this experiment 200 times and reported the corresponding backtest results as compared to the equally weighted portfolio. Figure presents the backtest results for all of the 200 resamplings. We found that in 172 out of 200 cases the strategy outperformed the equally weighted portfolio. Panel (a) shows that the backtests tend to perform in a similar fashion and in all cases to perform well during market crises (for example 2002-2003 or 2008-2009) confirming a good ability of the strategy to interpret the main crises events with different stocks considered, while Panel (b) suggests that underperformances are mainly accumulated during bull market periods (1998-2000, 2003-2007, 2010-2015) suggesting a possible high sensitivity to short term market volatility.

Table 5.2.2 presents a summary of risk and return metrics computed for the 200 resampled sets on the test set. We found a similar volatility levels among the resamplings (1.63% and 1.45% as 95^{th} and 5^{th} percentiles), coherently with the similar backtest paths discussed from Figure 5.3 panel (a). On the daily average return size, the mean average return was



(a) Backtest for 200 resampled different sets of 100 stocks.

(b) Performance spread with respect to an equally weighted portfolio

Figure 5.3: Boostrapping backtest results. 200 resampled equity sets with dimensionality fixed at 100 stocks. Estimation window 28 days. Panel (a) shows the backtests in both train and test sets for each resampling; Panel (b) the results relative to the equally weighted portfolio.

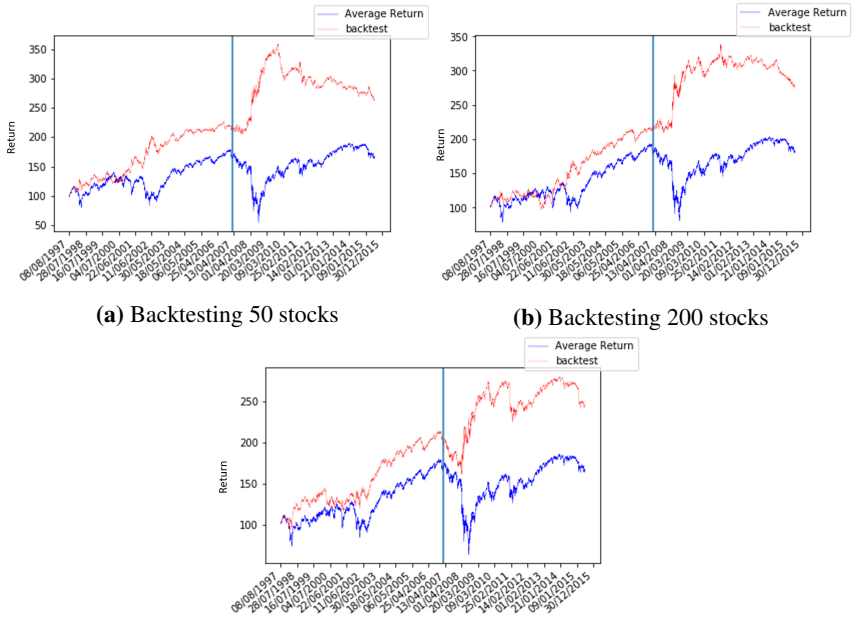
0.015% with a 95^{th} and 5^{th} percentiles of, respectively, 0.032% and 0.003%, showing that the 5^{th} percentile still provides a higher average returns than the equally weighted portfolio.

	Average Return	**Volatility (σ)**
Mean	0.015%	1.59%
5^{th}	0.003%	1.45%
95^{th}	0.032%	1.63%

Table 5.4: Out-of-sample risk-return performance metrics. 200 resamplings considered. Mean, 5^{th} and 95^{th} percentiles of daily average return and volatility from the backtested series of the 200 resamplings.

Dimensionality Effect

In this second test, we analysed the effect of changing the number of stocks included in sample. We considered data panels composed by 50, 200 and 500 stocks chosen at random from the entire dataset. Other than adding or subtracting individual stocks, changing the number of variables affects the estimated correlation structure given the fixed sparsity level imposed via TMFG-Logo (see Section 2.2.2). Figure 5.4 presents the cumulative return of the strategy obtained for each data panel (panel(a) refers to the 50 stocks sample, panel(b) to 200 and panel(c) to 500). The figure shows that, in all cases, our strategy significantly outperformed the equally weighted portfolio, proving it robust to dimensionality.



(a) Backtesting 50 stocks

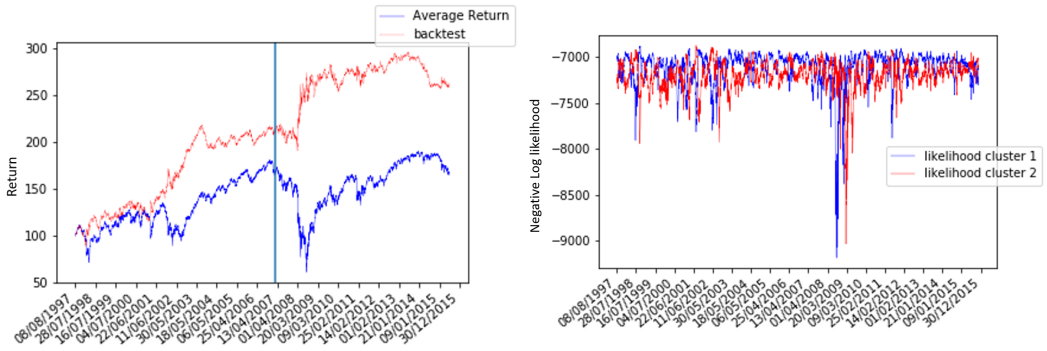
(b) Backtesting 200 stocks

(c) Backtesting 500 stocks

Figure 5.4: Dimensionality Effect. Strategy performance with varying number of stocks included in sample: 50 (panel(a)), 200 (panel(b)) and 500 (panel(c)) stocks, compared to the sample average return. Estimation window 28 days

Trivial comparison

In this section we compared the results of our methodology with the performance obtained using the parameters from a trivial clustering procedure. We artificially created the two clusters by assigning to cluster 1 observations corresponding to positive returns and negative returns to cluster 2. We computed the trivial clusters in the train set and used the same strategy as in Eq. (5.2.1). Figure 5.5 panel(a) shows that the backtest produced worse results, but comparable to those obtained with our clustering procedure. It is crucial to notice, however, that this procedure provided much more unstable results, with 1149 switching in the backtest as compared to the 338 of the our methodology. Indeed, Figure 5.5 panel(b) presents the negative log likelihood values of the two states computed using the trivial clustering parameters and showing the noisy behaviour with frequent overlapping. This is a significant difference with respect to our methodology confirming the goodness of our temporal consistency and filtering approaches.



(a) Backtesting using trivial clusters

(b) Trivial clusters likelihood observations

Figure 5.5: Trivial clusters comparison. Cumulative return vs weighted average portfolio (panel (a)) and clusters' likelihood for 100 stocks sample (panel (b)) obtained considering the trivial clusters.

5.3 Discussion

In this Chapter I applied the ICC methodology to forecast future states of the market form previous observations and present a simple application to equity systematic trading.

First, I experimented on the forecasting ability of our procedure using the log-likelihood ratio from the two clusters as sole feature to predict the next day market state represented by a binary class label. We fit two models, a logistic regression and a Support Vector Machine, to describe the relationship among these two variables. In both cases we obtained an accuracy level higher than 50%. Using the logistic regression model, we obtained an over-assignment of the bull state, but still significant prediction power for the bear state. SVM provided more balanced results, but still providing below 50% accuracy in predicting the bear state. This experiment shows that our procedure provides significant prediction power. Provided that this accuracy levels are achieved with the use solely of the information from past returns, as discussed in Chapter 7, one of the main contributions of our approach to the forecasting problem is represented by the efficiency of the procedure in exploiting information from the correlation structure.

In the second part of the chapter, I presented an application of the ICC methodology to equities trading. The application consists of a long/short strategy without optimal allocation. We backtested the strategy and compared the results to the performance of a buy-and-hold strategy with equally weighted portfolio allocation. We found that our methodology significantly outperformed the equally weighted portfolio and, in particular, the strategy has shown very reactive to main crises events. I presented a series of tests to assess the ro-

bustness of our methodology. The resulting performances are stable and robust to different variables tweaks, proving the methodology to be reliable when considering U.S. large cap equities datasets.

Chapter 6

Number of States

In this Section we analyse the main hyperparameter of the ICC methodology: the number of market states. The number of regimes underlying the financial system is certainly not a new problem. However, the literature developed on the topic is either leading to conflicting conclusions or based on the common belief in the financial industry that markets are only driven by a few types of states. In this chapter I present a series of experiments aimed at testing the significance of increasing number of regimes used to model equity returns and how this parameter relates to the number of observations and the time consistency of the states. The experiments investigate a) the likelihood of the overall model as more states are spanned and b) the relevance of additional regimes measured by the number of observations clustered. I conclude that the multivariate structure of the system changes through time, leading to new "states" being required to describe the system accurately as new observations are considered.

6.1 Introduction

In Section 4 I have introduced and defined the ICC methodology as an efficient and stable method to segment multivariate time series into homogeneous clusters. An additional advantage of the method is that it only requires the input of two hyperparameters: the temporal consistency γ and the number of states k. The number of clusters to be considered - i.e. the number of market states into which observations are clustered - is no doubt the most influential hyperparamenter in the ICC method, other than being, more generally, a natural research question of high theoretical and practical relevance.

Considering as reference point the literature on hidden Markov models and state space modelling in finance, however, it does not provide a clear answer to this question, showing that the selection of number of states even within the realm of classical time series approaches is still today an unanswered question. Many authors *assume* that two regimes are enough to correctly capture the evolution of financial markets (see, for example, [1, 5, 88]),

borrowing from common believes and industry practices of *'bull'* and *'bear'* markets. This clearly remains a rough simplification of a much more complex reality and the subject suffers a lack of attention despite its obvious interest for both academics and practitioners.

To the best of our knowledge, only a few authors have addressed directly this issue, leading, however, to conflicting results. [178] selects the number of states based on the marginal improvement of the model in-sample log likelihood, leading to the selection of three states. [72] analyses the number of states required to explain returns' characteristics of different asset classes independently, whilst still from a univariate standpoint. They conclude that two states suffices to explain the returns on only a few asset classes (exchange rates, commodities, US and EMU government bonds, etc..) and up to five states are required to capture the dynamics of other asset classes (high yield bonds). [134], which uses a definition of market states in the context of complex systems and thus leveraging the correlation structure of multiple variables, proposes a top down subsequent in-sample clustering (subject to a threshold hyperparameter), leading to 8 different market states being selected.

6.2 Methodology

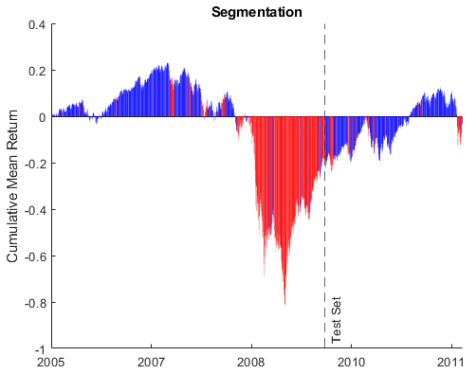
In this Chapter, I directly address the selection of Number of States in the context of the ICC methodology. I study the overall likelihood delivered by the model both in- and out-of-sample as the number of states changes and for different values of temporal consistency γ. Moreover, as I have discussed in Section 3, inference decisions on financial variables are intimately linked to the time component both in- and out-of-sample. As such, I also analyse what is the impact of different numbers of observations and time spanned across samples.

In our ICC methodology, the number of states serves in essence as a selection criteria for the parametrization of the model. Every 'market state' represents a probability distribution into which observations can be clustered and, therefore, considering an increasing number of market states translates into introducing more parameters into the model. To better understand the underlying mechanics, consider the example provided in Figure 6.1. This is a one resampling, two states example - i.e. I just considered two states, sampled randomly 100 stocks and a starting date. I then trained the ICC model over 1000 daily

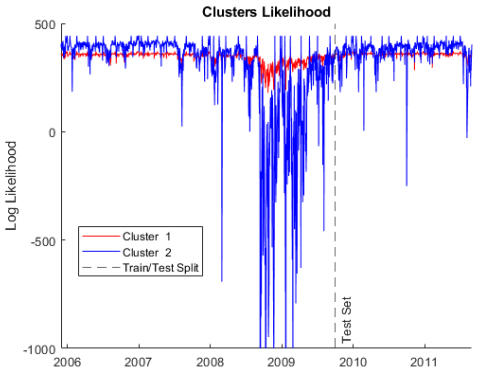
observations in-sample, resulting into two sets of means and covariances being estimated. I then tested the features of the states over 500 observations out-of-sample. This is done by maximizing the out-of-sample likelihood subject to the time consistency constraint, but using the parameters estimated in sample. In other words, we simply assign the observations to one of the two clusters by running the Viterbi algorithm (as described in Section 4.3.2).

Figure 6.1 aims at displaying how different states impact the likelihood of the model, the features of the corresponding clusters and highlights the relevance of time spanned and temporal consistency. Figure 6.1a presents the segmentation returned by the model. The chart reports the cumulative average return of the 100 sampled stocks with each daily observations coloured according to the market state ('Cluster 1' or 'Cluster 2') into which it has been clustered. By simply looking at this chart, it is intuitive that 'Cluster 2' contains what resemble bullish observations, with positive returns and up-warding trend, and 'Cluster 1' contains the GFC crisis period and more generally negative returns observations. Figure 6.1c presents the distribution of the two states , both in-sample and out-of-sample, and Table 6.1d reports the corresponding annualized Mean, Standard deviation and Sharpe Ratios, confirming the the *bull* and *bear* intuition. Perhaps more interestingly, Figure 6.1b highlights the underlying decision process of the model, showing the observation-wise likelihood of each cluster. In other words, this figure is the 'real world' representation of the sketched Figure 4.3. It is worth emphasizing that this is the final result of the training process discussed in Section 4. Nevertheless it provides clarifying insights on how the number of states impact the goodness of the model: in essence, clusters compete with each other to deliver the highest likelihood given the observation. Looking at Figure 6.1b, the blue 'Cluster 2' likelihood is significantly higher than the other in most of the bullish observations. A second observation is that the 'Cluster 2' likelihood is also much more volatile than that of Cluster 1, with spikes or outliers signalling that the Cluster 2 parameters well describe the 2008-2009 GFC period and only a few other observations in our sample. Our intuition is that Cluster 1 mostly represents the crisis period which is indeed different from other bearish moves of the market.

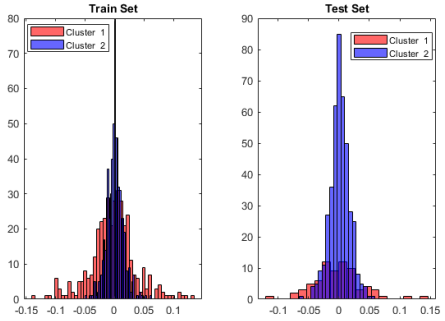
In the remainder of this Chapter, I will address how likelihood and clusters' performances are affected by the number of states at first keeping the temporal consistency parameter $\gamma = 0$, and then studying the impact of γ and the time spanned on the decision



(a) Time series segmentation results



(b) Log Likelihood of both clusters observation-wise



(c) Empirical distribution of each cluster

		μ	σ	SR
Train Set	Cluster 1	-27%	0.58	-0.47
	Cluster 2	14%	0.20	0.73
Test Set	Cluster 1	-12%	0.61	-0.20
	Cluster 2	24%	0.24	0.98

(d) Mean, Standard Deviation and Sharpe Ratio for 2 estimated market states, in- and out-of-sample.

Figure 6.1: Impact of Market States - Two States Example. Panel (a) reports the cumulative daily average return across the 100 randomly sampled stocks and the bars colour scheme reports the cluster assignation. Panel (b) reports the daily observation-wise likelihood computed for each of the two states. Panel (c) reports the histogram of the observations clustered into each market state in- and out-of-sample. Panel (d) reports the performance metrics for 2 estimated market states, in- and out-of-sample.

process. To obtain results as unbiased as possible, we will employ the resampling procedure discussed multiple times throughout this thesis: 500 resamplings, each time I randomly sample 100 stocks and a starting date. To assess the quality of the clusters and the model as whole, I will consider the average likelihood across 500 resamplings, average mean, standard deviation and Sharpe ratio, testing the significance of the Sharpe ratios across resamplings.

I considered the same dataset described and used in Section 3, consisting of daily closing prices of US stocks entering among the constituents of the S&P 500 index between 02/01/1997 and 31/12/2015. Same resampling methodology is also used in conducting the experiments described through the Chapter, with 500 resamplings in which I randomly select 100 stocks and a random trading day indicating the end of the training set.

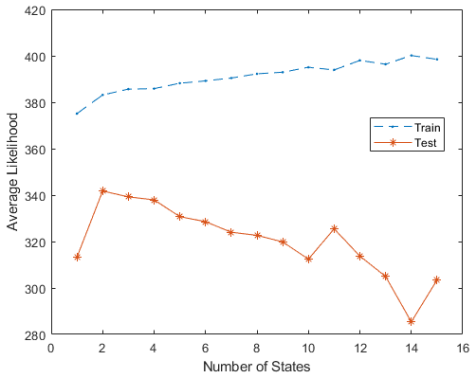
6.3 Results

6.3.1 Likelihood

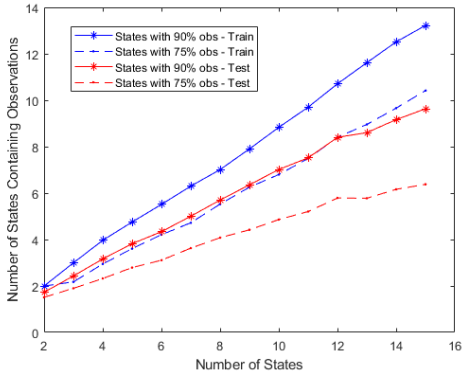
The first measure we studied to asses an "optimal" number of states is the average likelihood in-sample and out-of-sample as we allow the model to consider more states. In this first experiment, I keep the number of observations constant to 500 in-sample and 500 out-of-sample observations and the temporal consistency parameter $\gamma = 0$ in order to only analyse the effects of increasing number of states on the mechanics of the ICC estimation keeping other variables fixed. Figure 6.2 summarises our findings.

Figure 6.2a reports the average likelihood across the 500 resamplings and considering the whole sample of 500 observations in- and out-of-sample, as the number of states considered increase. In train set, as expected, the likelihood increases monotonically as the number of states increases. This is to be expected as adding states indeed introduces more parameters and thus tends to overfit the train set. More interestingly, out-of-sample the model exhibits a sharp increases in likelihood from 1 to to 2 states and then decreases almost monotonically. This implies that, on average when $\gamma = 0$ and 500 observations in- and out-of-sample are considered, two states delivered the overall best likelihood. It is also worth noticing that the likelihood of the model remains high, albeit not at his maximum, for 3, 4 and 5 states as well, declining sharply after that, suggesting a clear overfit of the train set.

An additional question to asses the relevance of the states is whether a good number of observations are contained in each cluster, signalling therefore that they capture a meaningful market feature, or weather only a handful of observations are clustered, signalling that the cluster is somehow 'residual' and made up of outliers. Figure 6.2b reports on the y-axis how many clusters contain 90% and 75% of the observations as the model is allowed to consider more states (x-axis). As previously noted, the training procedure clearly tends to overfit the training set efficiently using all clusters as the number of states increase. Even when 15 states are considered, 90% of the train observations are distributed across approximately 14 clusters. Very different picture for the test set. Only 2 to 4 states are used efficiently, with 75% and 90% of the observations being contained in maximum, respectively 4 and 6 states even when 15 states are trained.



(a) Time series segmentation results



(b) Mean (*left*) and variance (*right*) of each stock for each temporal cluster

Figure 6.2: Model likelihood and number of observations allocated to each state across 500 resamplings. Panel (a) reports the overall model likelihood (y axis), in- and out-of-sample as the number of states considered by the model increases (x axis). Panel (b) reports the number of states (y axis) containing 75% and 95% of observations as the number of states considered by the model increases (x axis).

6.3.2 Performance Metrics

While the models' likelihood is paramount for the purposes of our analysis and certainly of primary importance for my conclusions, it is still relevant to observe what features characterize market states. To this extent, Table 6.3.2 reports the average annualized mean (μ), standard deviation (σ) and Sharpe Ratio for each cluster corresponding to different number of states spanned by the model. The columns report the number of states considered by the model and each row identifies a different cluster. It is worth remarking that these metrics are uniquely concerned with the mean and variance of the discovered clusters which certainly do not offer a complete market overview. Nevertheless, being these the most commonly observed market features, this section aims at gaining intuition on the market states discovered by the model and on their effects as the number of states considered increases.

Number of States Trained

		2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cluster1	μ	-2.83%	-3.01%	1.61%	10.12%	-5.87%	-0.52%	7.93%	0.68%	-0.58%	1.06%	4.81%	1.16%	6.21%	9.99%
	σ	0.551	0.361	0.357	0.349	0.363	0.345	0.338	0.341	0.344	0.351	0.331	0.336	0.321	0.32
	SR	-0.9	-1.4	0.75	4.65	-2.55	-0.3	3.75	0.3	-0.3	0.45	2.25	0.6	3	4.95
	SR % significance	98.2	96.138	28.166	97.204	97.16	84.166	97.162	5.186	0.152	16.14	76.136	13.16	71.148	96.182
Cluster 2	μ	3.06%	6.16%	4.5%	0.09%	5.3%	5.81%	1.53%	2.56%	1.41%	9.71%	11.13%	3.8%	13.83%	2.75%
	σ	0.34	0.355	0.352	0.352	0.357	0.334	0.328	0.345	0.353	0.344	0.334	0.333	0.313	0.312
	SR	0.45	2.7	2.1	0	2.4	2.7	0.75	1.2	0.6	4.5	5.25	1.8	3.45	1.35
	SR % significance	99.1	97.186	98.17	25.182	76.156	68.168	21.166	6.172	3.15	98.152	99.178	23.15	99.126	31.146
Cluster 3	μ		1.75%	0.83%	-2.94%	0.71%	5.21%	2%	1.78%	7.65%	-0.9%	-2.79%	3.75%	1.77%	5.34%
	σ		0.369	0.357	0.363	0.365	0.341	0.337	0.34	0.353	0.34	0.334	0.328	0.321	0.328
	SR		0.75	0.3	-1.45	0.3	2.4	0.9	0.9	3.45	-0.45	-1.35	1.8	0.9	2.55
	SR % significance		1.16	2.194	98.162	0.186	66.174	23.156	14.17	99.142	13.172	66.162	43.128	23.158	57.162
Cluster 4	μ			-2.01%	1.03%	9.67%	-3.4%	4.52%	-0.6%	-1.12%	6.22%	-3.12%	16.74%	1.24%	0.32%
	σ			0.343	0.341	0.347	0.34	0.328	0.357	0.359	0.35	0.331	0.337	0.317	0.324
	SR			-0.45	0.45	4.35	1.05	2.25	-0.3	-0.45	2.85	-1.5	7.95	0.6	0.15
	SR % significance			98.134	13.176	99.174	98.168	69.158	12.176	12.168	88.148	78.176	99.134	17.162	4.134
Cluster 5	μ				0.98%	7.73%	5.19%	1.71%	9.65%	-1.6%	-8%	11.7%	-5.94%	5.97%	11.02%
	σ				0.351	0.358	0.346	0.335	0.345	0.355	0.347	0.335	0.336	0.321	0.315
	SR				0.45	3.45	3.3	0.75	4.5	-0.75	-3.6	5.55	-2.85	3	5.55
	SR % significance				2.156	99.184	87.162	27.154	98.132	2.144	99.48	92.61	88.146	63.152	89.41
Cluster 6	μ					3.07%	13.81%	-2.09%	-3.24%	-4.12%	2.45%	-1.13%	-0.82%	2.14%	4.79%
	σ					0.349	0.338	0.336	0.344	0.355	0.348	0.318	0.328	0.304	0.315
	SR					1.15	6.45	-1.05	-0.6	-1.8	1.05	-0.6	-0.45	1.05	2.4
	SR % significance					16.172	99.44	0.154	86.182	98.142	22.134	29.162	11.6	18.146	48.142
Cluster 7	μ						2.9%	3.62%	4.01%	3.84%	8.83%	-4.13%	8.7%	1.78%	-0.8%
	σ						0.338	0.333	0.344	0.346	0.342	0.325	0.33	0.319	0.32
	SR						1.35	1.8	4.2	1.8	4.05	-1.95	4.2	0.9	-0.45
	SR % significance						26.182	12.174	22.168	20.13	88.156	91.124	96.134	2.16	3.174
Cluster 8	μ							8.07%	0.09%	0.25%	4.57%	11.83%	-2.29%	3.86%	0.18%
	σ							0.34	0.342	0.353	0.354	0.324	0.338	0.317	0.324
	SR							3.75	0	0.15	2.1	5.85	-1.05	1.95	0.15
	SR % significance							84.174	2.4	3.6	49.158	98.156	85.136	72.4	1.152
Cluster 9	μ								-1.29%	8.91%	1.8%	7.84%	-7.5%	5.23%	-0.6%
	σ								0.34	0.349	0.351	0.325	0.328	0.311	0.315
	SR								-0.6	4.05	0.75	3.9	-3.6	2.7	-0.3
	SR % significance								2.176	83.178	2.12	79.11	98.128	92.14	8.124
Cluster 10	μ									0.52%	-2.88%	-2.59%	8.79%	-7.04%	0.71%
	σ									0.342	0.36	0.329	0.335	0.313	0.316
	SR									0.3	-1.2	-1.2	4.2	-3.6	0.3
	SR % significance									4.16	26.13	32.148	96.13	99.154	6.5
Cluster 11	μ										4.3%	-2.31%	-6.57%	-7.81%	3.54%
	σ										0.353	0.324	0.331	0.324	0.318
	SR										1.95	-1.2	-3.15	-3.9	1.8
	SR % significance										76.15	68.134	97.148	99.13	42.142
Cluster 12	μ											-2.92%	-0.35%	0.33%	-4.67%
	σ											0.327	0.323	0.315	0.322
	SR											-1.35	-0.15	0.15	-2.25
	SR % significance											64.0	2.14	5.162	74.148

		Number of States Trained													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cluster 13	μ												-1.41%	-5.43%	5.08%
	σ												0.335	0.309	0.317
	SR												-0.6	-2.85	2.55
	SR % significance												48.136	68.14	56.144
Cluster 14	μ													-1.05%	4.38%
	σ													0.313	0.321
	SR													-0.6	2.1
	SR % significance													22.142	38.154
Cluster 15	μ														1.7%
	σ														0.315
	SR														0.9
	SR % significance														16.2

Table 6.1: TMFG portfolios performance metrics. Average annualized returns, standard deviation and Sharpe ratio of a daily rebalancing minimum volatility strategy. Optimal weights computed using using TMFG precision matrices. Median, 5^{th} and 95^{th} percentiles across 100 resamplings.

The first observation from Table 6.3.2 is that, as observed in multiple experiments throughout this thesis (e.g. Section 4.4.1 or Section 6.3.1), when only two clusters are considered by the model, they tend to align to the common definitions of *bull* and *bear* markets. The table shows that with two states (first column), the two clusters exhibit respectively positive mean returns / low variance and negative mean return / high variance. As the number of states considered by the model increases, it is still possible to identify a *bull* and *bear* state, but their features are accentuated - i.e. higher (lower) mean return and lower (higher) variance. In other words, the observations get more concentrated, with the clusters becoming more specific. The additional states, at the contrary, tend to display non extreme mean-return features. In the 3 states case, 'Cluster 1' and 'Cluster 2' clearly resemble our description of *bear* and *bull* states, while 'Cluster 3' displays a moderately positive mean return, a standard deviation slightly above that of the *bull* state and a Sharpe Ratio often not significantly higher than zero - features that could be interpreted as those of a sideways, directionless

market. These less extreme features are displayed by most of the additional states when the number of clusters spanned by the model increases until around 8 states. Similarly, while the SR significance remains high (never below 95%) for the *bull* and *bear* states , I found a non significant SR for most of the additional market states. When more than 8 states are spanned by the model, other 'extreme' states are identified by the procedure, i.e. more than one very high (very low) mean return states are discovered. The SR significance of all the "extreme" states diminishes, with the additional one being low in significance. Our interpretation is that these additional states are used to capture outliers, or, more generally, only a few observations of the *bull* and *bear* clusters with particularly extreme features. Taking out outliers from the *bull* and *bear* states increases the likelihood of the model and homogeneity if the clusters, being however increasingly prone to overfitting.

6.3.3 Effects of number of observations and time spanned

Having assessed the in- and out-of-sample likelihood of the model when different numbers of states are considered in a static and basic setting (i.e. $\gamma = 0$ and fixed estimation and testing windows of 500 observations), we turned our analysis to the effect of time.

Figure 6.3 reports the average likelihood (6.3.a and 6.3.b) and the number of states containing 90% of the observations (6.3.c and 6.3.d) both in- and out-of-sample. The in-train results are coherent with what we reported in Section 6.3.1 and Section 3.4.1 and thus come perhaps with no surprise. The likeilhood in train is increasing with the number of states and decreasing with the number of observations: the higher the number of parameters and the lower the number of observations is, the easier is for the model to overfit the train set. Similarly the model is pretty efficient at exploiting the increasing number of states, with all states included in the model containing a significant amount of observations (fig 6.3.c).
The right-hand charts in the figure present the out-of-sample results. Figure 6.3.b reports the average out-of-sample likelihood. When only a few observations are considered, a smaller number of states suffices in describing the system dynamics, delivering the highest average likelihood. In this case, adding more states leads very quickly to overfitting the train-set, with an exponential deterioration of the out-of-sample likelihood (purple and yellow lines). When more observations are considered, however, more states do provide a better overall model performance: when 750 (orange line) and 1000 (blue line) observations are considered, the maximum average likelihood is obtained, respectively, with 3 and 4 states. Figure 6.3.d coherently complements these findings showing that new states are efficiently used only when more observations are included. With 1000 and 750 observations, when up to 4 states are included in the model, all of the states are used to allocate 90% of the observations. When 150 and 250 observations are considered, new states become less relevant and more than 5 and 6 states are, in essence, never efficiently allocated.

6.3.4 Effects of varying γ

A second parameter linked to the number of states in our modelling is the time consistency parameter γ. While of high practical relevance, enforcing exogenously temporal consistency as we do in our ICC model obviously affects the overall optimality of the model from a likelihood perspective and forces observations to aggregate, intuitively allowing for *less*

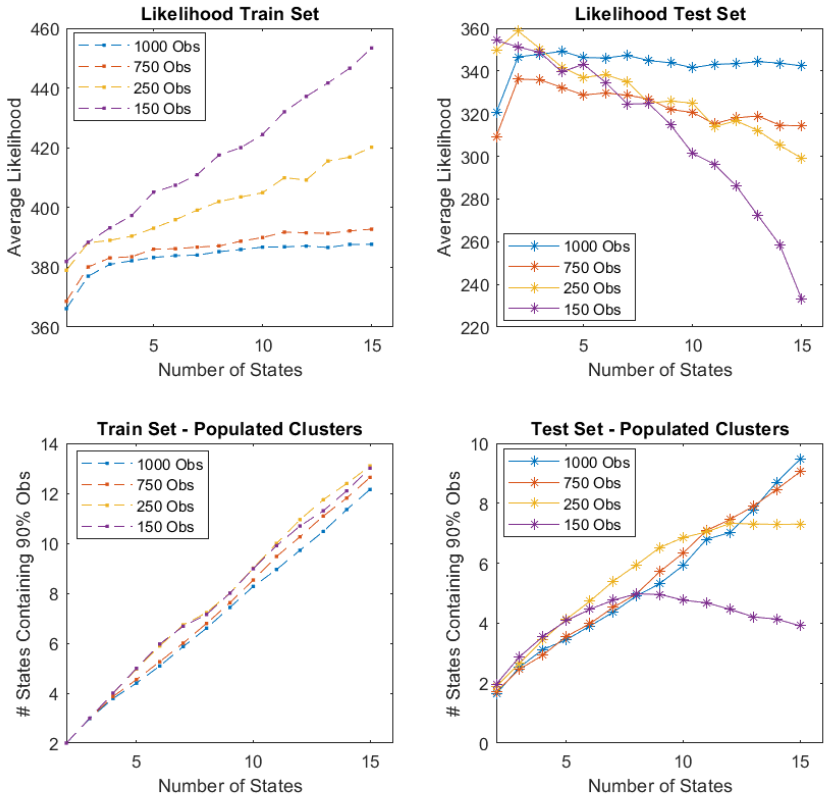


Figure 6.3: Average likelihood and observations allocation for different train and test lengths.

similar observations to be included in the same cluster. To investigate the effects of varying γ on the number of states, the run the same experiment described in Section 6.3.3 but letting γ vary from 1 to 50. Figure 6.4 reports the average likelihood (6.4.a and 6.4.b) and the number of states containing 90% of the observations (6.4.c and 6.4.d) both in- and out-of-sample for γ values of 1, 10, 30 and 50. Focusing first on likelihood, Figure 6.4.a and .b do not show any clear pattern related to varying time consistency values. Both the in- and out-of-sample likelihood behave as documented for the case in which $\gamma = 0$ (see Figure6.2a). The higher γ is, the lower the likelihood is, particularly out-of-sample, which is to be expected given that enforcing temporal consistency constraints the likelihood optimisation in our model. The findings presented in Figure 6.4.c and 6.4.d also support the intuitions on the effects of γ. The charts show that the higher the temporal consistency enforced, the less homogeneous is the population density of the clusters. In-sample (Figure 6.4.c) the effects are less significant, with only $\gamma = 50$ materially affecting the number of states used to allocate 90% of the

observations. The out-of-sample results (Figure 6.4.d), instead, provide a clear picture of the effect. In particular, for very high values of γ (e.g. 30 or 50 in our experiment), 90% of the observations are contained in maximum of 3 states on average with the blue and orange lines in Figure 6.4.c immediately platooning as more states are spanned by the model.

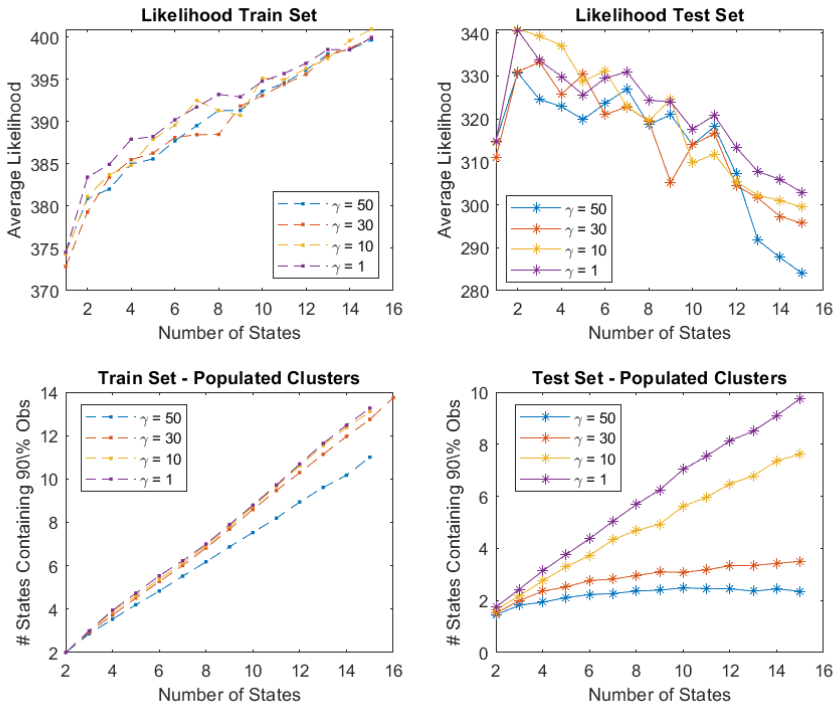


Figure 6.4: Average likelihood and observations allocation for different temporal consistency parameters.

6.4 Discussion

Asking how many states describe the dynamics of an asset class or, more generally, the financial system is an incomplete question. In this section, we show that the financial system evolves through time. Other than the univariate distribution of individual variables, the way the variables interact with each other and, more generally, the multivariate structure of the system changes through time, leading to new 'states' being required to describe the system accurately. We conclude that the 'optimal' number of states needs to be analysed

and defined in terms of the time span considered. This is indeed coherent and supports our findings described in Chapter 3 in that the optimal fit in finance needs to be defined (also) in terms of the investment horizon.

Chapter 7

Conclusions and Future Work

This chapter concludes the thesis and revisits the key findings, highlights the achievements and contributions and opens to future investigation by listing few potential extensions and directions for further research.

7.1 Conclusions

This thesis investigated the time changing nature of financial markets. Financial markets are complex systems having an intrinsic structure defined by the interplay of several variables which, however, changes and evolves through time. This feature is critically relevant for classical statistical assumptions and has proven challenging to be investigated and researched.

The main motivations of the topic and investigations presented in this thesis were: (i) non-stationarity is a key feature of the financial system, difficult to quantify but that impacts practitioners and researchers across financial markets; (ii) the increasing data availability and consequential adoption of data-driven methodologies across markets is demanding new modelling paradigms being able to exploit the correlation and interaction of a large number of variables in an efficient and scalable way; and (iii) information filtering networks is a modelling paradigm that can prevent many of the pitfalls of classical estimation techniques when dealing with non-stationarity, however with only few evidences are provided literature on the impacts in the finance domain.

Therefore, to explore these issues, this research is divided in three studies: (i) Study of non-stationarity from through the lens of parameters likelihood and its impact on portfolio performances through time; (ii) ICC methodology to define and forecast market states; (iii)

Integration of information filtering.
In our view, each study, respectively, generated the following major contributions: (i) When considering financial time series, optimisations on parameters and consequential investment frameworks must take into account the holding period and are otherwise incomplete; (ii) A novel methodology to define, analyse and forecast market states; and (iii) Evidences in support of information filtering role in avoiding classical pitfalls of standard estimation approaches applied to financial, non-stationary time series.

By tackling such important problems, this thesis aimed to unveil the dynamic nature of the financial system through time.

Finally, this thesis also show that Financial, Tech and Mathematical considerations are indeed complementary in the contemporary financial landscape and a new hybrid skill set will reshape the current financial markets practice. The forces driving these changes can be identified: a) easy access to vast amounts of data; b) availability of virtually unlimited computing resources and c) open-access state-of-art in AI/ML algorithm libraries. The developments have led to a scramble for talent across the Investment Banking world, with Data Scientists poached from technology and retail companies. Therefore, the new ”Quants” are and will increasingly be from hybrid backgrounds with a strong Computer Science component.

7.2 Main Contributions

Based on our results, background and literature review, the main contributions that this research offers are:

1. A new framework to measure non-stationarity and its impact on portfolio performances. I show that the observation-wise parameters likelihood drifts downward with time and significantly increases in variance. I derive the new principle that, when considering financial time series, optimisations on parameters and consequential investment frameworks (e.g. portfolio constructions) must take into account the holding period and are otherwise incomplete.

2. A novel methodology to define, analyse and forecast market states called ICC. The procedure assigns each observation to a cluster or *state* based on its correlation struc-

ture and defines each cluster as a sparse Markov random field, making the results highly interpretable. The ICC methodology offers significant advantages with respect to standard approaches as it can efficiently scale large multivariate datasets and enforces temporal consistency, opening the field for further research on the role of the correlation structure in modelling non-stationarity while managing the practical need for stable states definition.

3. Evidences in support of information filtering role in improving estimates and long term model stability. Information filtering can easily be integrated to improve classical and novel financial applications that rely on a multivariate modelling of the system.

7.3 Main Experiments

The conclusions and contributions highlighted above are derived and supported by a series of exploratory studies.

1. Non-stationarity and sparse optimal portfolio.

   I report on two sets of experiments to study: a) the in- and out-of-sample parameters' likelihood through time and the dependency on estimation window; b) the impacts on portfolio performances. First, I found that shorter estimation windows deliver higher out-of-sample likelihood in the observations immediately following the train window, but it tends to rapidly decrease afterwards. As more observations are included in the training set, the out-of-sample likelihood gains stability, with larger values in the long term, but at the cost of lower likelihood in the short term. Secondly, I demonstrate the relationship between the goodness of the model, measured as out-of-sample likelihood, and the realized portfolio volatility. Further, I compare portfolio performances and features obtained when sparse precision matrices are used as input for portfolio construction and show that sparsity can significantly reduce estimation errors coming from both sampling error and non-stationarity and avoid many of the classical portfolio construction pitfalls.

2. ICC states discovery.

I present one experiment where I apply the ICC methodology with two clusters to a random set of equity returns and compare it to standard GMM derived clusters used as baseline model. I found that the ICC clusters display neat, interpretable financial features with a good time consistency. Further, I integrate the use of the TMFG-LoGo information filtering to estimate the referential state precision matrix and show that it provides improved clusters significance and better model stability by reducing the cluster switches, other than improving the algorithm efficiency.

3. Market states forecasting.

   I report two sets of experiments where I apply the ICC methodology to forecast future market state and build a simple equity trading strategy. I study the forecasting ability of our procedure using the log-likelihood ratio from two market states to predict the next day cluster. I experimented with two models, a logistic regression and a Support Vector Machine and in both cases obtained an accuracy level higher than 50%. Following that, I present a simple equity trading strategy consisting of a binary long/short investment rule without optimal allocation and compared the strategy to the performance of an equally weighted portfolio. I show that our methodology significantly outperformed the equally weighted portfolio and report on a series of tests to assess the robustness of our methodology, proving that performances are stable and robust to different variables tweaks.

4. Number of states.

   Having defined a robust procedure to cluster observations into market states, I present again two sets experiments to study how the in- and out-of-sample likelihood is impacted by different number of states considered by the model. The experiments investigate the dependency with a) the estimation window length; and b) the time consistency enforced. I found that the evolution of the financial system through time also implies that the number of clusters that optimally describe the system increases through time: the more observations are considered and the larger the time window spanned, the higher the number of market states to be considered. These findings also reinforce the statement that optimal estimation in finance is dependent on the holding period and time spanned.

7.4 Implications for financial practices

A relatively large set of state space models and time series clustering techniques have been proposed in literature and reviewed in Chapter 2. Two main pitfalls, however, typically affect these models: a) Curse of dimensionality and b) Model instability. With the ICC methodology, we aimed at solving both these issues. The motivations for this research are deeply rooted in financial practices. Dealing with non-stationarity per se impacts virtually any data related process in the financial industry as classical statistical assumptions fail and ad hoc practical strategies are often employed. I report below three direct practical application of particular relevance within the investment management domain:

1. Trading strategies.

   Regime shifting and timing strategies, particularly at high frequency, are widely used and researched. Common criticism for these strategies is the typical instability of the models that leads to frequent rebalances and consequential high trading costs. The ICC method is of high practical relevance in this sense, since it allows explicit control of temporal consistency as part of the states definition problem.

2. Causality, correlation and variables interaction.

   Financial markets are complex systems having an intrinsic structure defined by the interplay of several variables. The large and exponentially increasing amount of data available today has played a significant role in accelerating the 'scientification' of the investment process, with virtually any investment decision or recommendation today being supported by data. The ICC methodology provides a flexible and scalable framework to study the interaction among variables and leverage the large amount of alternative data.

3. Risk premia interaction and timing.

   Starting from the pioneering work of [67], the risk premia literature has disrupted the financial industry over the last decade. In very loose terms, risk premia are very relevant in that provide an interpretable source of risk/return for a given asset or investment product. As such, risk premia are nowadays extensively used both to assess and optimise the risk/return exposures and to define a whole new set of financial products (e.g. smart beta). A key implication of this practice, however, is that different

risk premia behave differently during different stages of the economic cycle and their correlation is time-changing and challenging to analyse. The ICC methodology provides a convenient framework to analyse the multivariate structure of risk premia (and other variables) through time, making it highly relevant for the state-of-art portfolio optimisation practices and for the wide array of factor based investment strategies and products.

7.5 Further Work

Finally, our work also opens new avenues to future investigations. Below, I list a few potential extensions and directions for further research.

7.5.1 ICC extensions

In this thesis I introduced the ICC methodology and applied it to daily equity returns to discover and test market states features, predictability and likelihood behaviour. However, the methodology is fully flexible and easily scalable. As such, there are several directions for further testing and expansion of the methodology. In particular:

- Distance measures

  Throughout this thesis, I presented several experiments with application of the ICC methodology where clusters are optimised and discovered by optimising either the parameters' Likelihood or the Mahalanobis distance. This choice is theoretically inspired, since I looked at clusters by investigating how well parameters describe a segment of returns, hence leading to parameters' likelihood being the objective to be optimised. However, many other statistical or financial measures could be considered, such us the relative entropy or the clusters Sharpe ratio, just to mention two. Further investigation of this would also shed light on other ways to look at non-stationarity and the evolution of the financial system through time.

- Alternative data

  One of the motivating drivers for the ICC methodology is the efficiency of the algorithm. The procedure is highly scalable, and can be used with hundreds of variables, without suffering the curse of dimensionality as other, more popular models do. This

open new avenues of investigation on how alternative data can be used to better describe the financial system. A natural evolution would be to consider other asset classes and macroeconomic data. Similarly, a wide set of alternative data such as text data coming from financial reports and sentiment analysis or online users interactions can be a easily included in the model framework.

- Portfolio construction

  All the portfolio construction experiments presented in this thesis focus on the simplest possible optimisation techniques and solutions. The goal of these experiments is to test and compare the impact of different estimation approaches and forecasting techniques. Using unsophisticated portfolio construction approaches allows a fair and clearer comparison, avoiding unnecessary biases. However, the flexibility of the ICC methodology is easily applicable with more sophisticated optimisation techniques that could lead to better performances. As an example, one output of the ICC procedure is the likelihood of each cluster for each observation. A robust portfolio could be built by ensambling the clusters parameters, with a weighting scheme proportional to the clusters' likelihood. Another approach could be to treat the iterations of the ICC methodology as Bayesian updates for the portfolio construction parameters.

7.5.2 Market states and system entropy

In all the investigations and experiments presented through this thesis, I always considered a framework with a fixed number of states and analysed the states features, their likelihood and evolution through time. One feature that has been highlighted is that, even considering a high number of states, their likelihood is volatile and the interaction among variables continuously changing. One possible interpretation of this is that the financial system is characterised by a random disorder, with a set of (potentially infinite) reference 'states' that evolve through time. The system moves continuously towards the reference states, but never settles in one. This view of the financial system resembles frameworks described in physics and chemistry, particularly the spin glasses magnetic states, and modelling approaches could be inspired by them. A general theory of 'inherent states' would generalise the framework discussed in this thesis and, more generally, any state dependent theory of financial markets.

Bibliography

[1] Al-Anaswah, N. and Wilfling, B. (2011). Identification of speculative bubbles using state-space models with Markov-switching. *Journal of Banking & Finance*, 35(5):1073–1086.

[2] Aldridge, I. and Krawciw, S. (2017). *Real-time risk : what investors should know about fintech, high-frequency trading, and flash crashes*. Wiley.

[3] Alexander, C. and Chibumba, A. (1997). Multivariate orthogonal factor garch. *University of Sussex, Mimeo*.

[4] Ang, A. and Bekaert, G. (2002). International asset allocation with time-varying correlations. *Review of Financial Studies*, 15:1137–1187.

[5] Ang, A. and Bekaert, G. (2015). International asset allocation with regime shifts. *Review of Financial Studies*, 15(4):1137–1187.

[6] Ang, A. and Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63(3):443–494.

[7] Ardia, D. and Meucci, A. (2015). Parametric stress-testing in non-normal markets via copula-marginal entropy pooling. *Risk Magazine*, June:1–5.

[8] Arditti, F. and Levy, H. (1976). Portfolio efficiency analysis in three moments: the multiperiod case. *Journal of Finance*, (30):797–809.

[9] Aste, T. (2020). Topological regularization with information filtering networks. *arXiv preprint arXiv:2005.04692*.

[10] Aste, T. and Di Matteo, T. (2006). Dynamical networks from correlations. *Physica A Statistical Mechanics and its Applications*, 370:156–161.

[11] Aste, T. and Di Matteo, T. (2017). Causality network retrieval from short time series. *arXiv preprint arXiv:1706.01954*.

[12] Bamberg, G. and Dorfleitner, G. (2001). Fat tails and traditional capital market theory. *Working Paper, University of Augsburg*.

[13] Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. *Machine Learning*, 56(1):89–113.

[14] Barbieri, A., Dubikovsky, V., Gladkevich, A., Goldberg, L. R., and Hayes, M. Y. (2010). Central limits and financial risk. *Quantitative Finance*, 10(10):1091–1097.

[15] Barfuss, W., Massara, G. P., di Matteo, T., and Aste, T. (2016). Parsimonious modeling with information filtering networks. *Physical Review E*, 94:062306.

[16] Batra, L. and Taneja, H. C. (2020). Portfolio optimization based on generalized information theoretic measures. *Communications in Statistics - Theory and Methods*, 0(0):1–15.

[17] Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73:360–363.

[18] Baum, L. E., Petrie, T., Soules, G. W., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

[19] Bauwens, L., Laurent, S., and Rombouts, J. V. K. (2006). Multivariate garch models: a survey. *Journal of Applied Econometrics*, 21(1):79–109.

[20] Bawa, V. S. (1978). Safety-first, stochastic dominance, and optimal portfolio choice. *Journal of Financial and Quantitative Analysis*, 13(2):255–271.

[21] Becker, H. (2005). A survey of correlation clustering. *Advanced Topics in Computational Learning Theory*, pages 1–10.

[22] Beg, A. B. M. R. A. and Anwar, S. (2014). Detecting volatility persistence in garch models in the presence of the leverage effect. *Quantitative Finance*, 14(12):2205–2213.

[23] Bennett, K. P. (1992). Robust linear programming discrimination of two linearly separable sets. *Optimization Methods and Software*, 1:23–34.

[24] Bera, A. K. and Higgins, M. L. (1993). Arch models: properties, estimation and testing. *Journal of Economic Surveys*, 7(4):305–366.

[25] Bergen, V., Escobar, M., Rubtsov, A., and Zagst, R. (2018). Robust multivariate portfolio choice with stochastic covariance in the presence of ambiguity. *Quantitative Finance*, 18(8):1265–1294.

[26] Berkane, M. and Bentler, P. (1986). Moments of elliptically distributed random variates. *Statistics & Probability Letters*, 4(6):333 – 335.

[27] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

[28] Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43.

[29] Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637.

[30] Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The review of economics and statistics*, pages 498–505.

[31] Bollerslev, T., Chou, R., and Kroner, K. F. (1992). Arch modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*, 52(1-2):5–59.

[32] Borland, L. and Bouchaud, J.-P. (2004). A non-Gaussian option pricing model with skew. *Quantitative Finance*, 4(5):499–514.

[33] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM.

[34] Boudt, K., Galanos, A., Payseur, S., and Zivot, E. (2019). Chapter 7 - multivariate garch models for large-scale applications: A survey. In *Conceptual Econometrics Using R*, volume 41 of *Handbook of Statistics*, pages 193–242. Elsevier.

[35] Box, G., Jenkins, G., and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*. Forecasting and Control Series. Prentice Hall.

[36] Brechmann, E. C., Hendrich, K., and Czado, C. (2013). Conditional copula simulation for systemic risk stress testing. *Insurance: Mathematics and Economics*, 53(3):722–732.

[37] Broadie, M. (1993). Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, 45:215–229.

[38] Brockwell, P. and Davis (2002). *Introduction to Time Series and Forecasting*. Springer International Publishing.

[39] Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009). Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272.

[40] Campbell, J. Y., , Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

[41] Caporin, M. and McAleer, M. (2012). Do we really need both BEKK and DCC? A tale of two multivariate GARCH models. *Journal of Economic Surveys*, 26(4):736–751.

[42] Chamberlain, G. (1983). A characterization of the distributions that imply mean–variance utility functions. *Journal of Economic Theory*, 29(1):185–201.

[43] Chiang, T. C., Tan, L., and Li, H. (2007). Empirical analysis of dynamic correlations of stock returns: evidence from Chinese A-share and B-share markets. *Quantitative Finance*, 7(6):651–667.

[44] Christie, S. (2005). Is the Sharpe ratio useful in asset allocation? *MAFC Research Papers*, pages 1–48.

[45] Cizeau, P., Potters, M., and Bouchaud, J.-P. (2010). Correlation structure of extreme stock returns. *Quantitative Finance*, 1(2):217–222.

[46] Cont, R. (2007). *Volatility Clustering in Financial Markets: Empirical Facts and Agent-Based Models*, pages 289–309. Springer.

[47] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[48] Coulon, J. and Malevergne, Y. (2011). Heterogeneous expectations and long-range correlation of the volatility of asset returns. *Quantitative Finance*, 11(9):1329–1356.

[49] Cox, J. C. (1997). The constant elasticity of variance option pricing model. *Journal of Portfolio Management*, 23(5):15–17.

[50] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Landmarks in Mathematics and Physics. Princeton University Press.

[51] Daluiso, R. and Morini, M. (2017). Hedging efficiently under correlation. *Quantitative Finance*, 17(10):1535–1547.

[52] Danaher, P., Wang, P., and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(2):373–397.

[53] Daniels, H. E. (1965). The asymptotic efficiency of a maximum likelihood estimator. *Matematika*, 9(1):149–161.

[54] Danielsson, J. (2011). *Financial risk forecasting: the theory and practice of forecasting market risk with implementation in R and Matlab*. Wiley-Blackwell.

[55] De Franco, C., Nicolle, J., and Pham, H. (2019). Bayesian learning for the Markowitz portfolio selection problem. *International Journal of Theoretical and Applied Finance*, 22(07):1–40.

[56] De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18.

[57] DeLuca, G., Rivieccio, G., and Corsaro, S. (2020). Value-at-Risk dynamics: a copula-VaR approach. *European Journal of Finance*, 26(2-3):223–237.

[58] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[59] Diks, C. (2004). The correlation dimension of returns with stochastic volatility. *Quantitative Finance*, 4(1):45–54.

[60] Ding, Z. and Granger, C. W. (1996). Modeling volatility persistence of speculative returns: a new approach. *Journal of Econometrics*, 73(1):185 – 215.

[61] Duan, J.-C., Popova, I., and Ritchken, P. (2002). Option pricing under regime switching. *Quantitative Finance*, 2(2):116–132.

[62] Duffie, D. and Pan, J. (1997). An overview of value at risk. *Journal of Derivatives*, 4(3):7–49.

[63] Elton, E., Gruber, M., and Brown, S. (2013). *Modern Portfolio Theory and Investment Analysis, 9th Edition: Ninth Edition*. Wiley Global Education.

[64] Engle, R. (2002). Dynamic conditional correlation. *Journal of Business & Economic Statistics*, 20(3):339–350.

[65] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

[66] Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, 49(3):283 – 306.

[67] Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

[68] Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

[69] Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric multivariate and related distributions*. Number 36 in Monographs on statistics and applied probability. Chapman & Hall.

[70] Focardi, S. M. and Fabozzi, F. J. (2004). A methodology for index tracking based on time-series clustering. *Quantitative Finance*, 4(4):417–425.

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

[72] Gatumel, Mathieuand Ielpo, F. (2014). The number of regimes across asset returns: identification and economic value. *International Journal of Theoretical and Applied Finance*, 17(06):1450040.

[73] Giada, L. and Marsili, M. (2001). Data clustering and noise undressing of correlation matrices. *ArXiv e-prints*, 63(6):061101.

[74] Gollier, C. (2001). Wealth Inequality and Asset Pricing. *Review of Economic Studies*, 68(1):181–203.

[75] Grabarnik, P. and Särkkä, A. (2001). Interacting neighbour point processes: some models for clustering. *Journal of Statistical Computation and Simulation*, 68(2):103–125.

[76] Granger, C. and Andersen, A. (1978). *An introduction to bilinear time series models*. Angewandte Statistik und Ökonometrie. Vandenhoeck and Ruprecht.

[77] Granger, C. W. J. and Ding, Z. (1995). Some properties of absolute return: an alternative measure of risk. *Annales d'Économie et de Statistique*, (40):67–91.

[78] Grinblatt, M. and Titman, S. (1989). Mutual fund performance: an analysis of quarterly portfolio holdings. *Journal of Business*, 62(3):393–416.

[79] Grobys, K. (2018). Risk-managed 52-week high industry momentum, momentum crashes and hedging macroeconomic risk. *Quantitative Finance*, 0(0):1–15.

[80] Hagan, P. S., Kumar, D., Lesniewski, A., and Woodward, D. E. (2002). Managing smile risk. *Wilmott Magazine*, September:84–108.

[81] Hallac, D., Nystrup, P., and Boyd, S. (2016). Greedy Gaussian segmentation of multivariate time series. *ArXiv e-prints*.

[82] Hallac, D., Vare, S., Boyd, S. P., and Leskovec, J. (2017). Toeplitz inverse covariance-based clustering of multivariate time series data. *CoRR*, abs/1706.03161.

[83] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

[84] Han, C. (2020). A nonparametric approach to portfolio shrinkage. *Journal of Banking & Finance*, 120:105953.

[85] Harlow, W. V. (1991). Asset allocation in a downside-risk framework. *Financial Analysts Journal*, 47(5):28–40.

[86] He, C., Li, G., Fan, H., and Wei, W. (2021). Correlation between Shanghai crude oil futures, stock, foreign exchange, and gold markets: a garch-vine-copula method. *Applied Economics*, 53(11):1249–1263.

[87] Hendricks, D., Gebbie, T., and Wilcox, D. (2016). Detecting intraday financial market states using temporal clustering. *Quantitative Finance*, 16(11):1657–1678.

[88] Henry, O. T. (2009). Regime switching in the relationship between equity returns and short-term interest rates in the UK. *Journal of Banking & Finance*, 33(2):405–414.

[89] Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2):327–343.

[90] Hogan, W. W. and Warren, J. M. (1974). Toward the development of an equilibrium capital-market model based on semivariance. *Journal of Financial and Quantitative Analysis*, 9(1):1–11.

[91] Hsieh, D. A. (1991). Chaos and nonlinear dynamics: application to financial markets. *Journal of Finance*, 46(5):1839–1877.

[92] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer.

[93] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106:620–630.

[94] Jobson, J. D. and Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and treynor measures. *Journal of Finance*, 36(4):889–908.

[95] Jondeau, E. and Rockinger, M. (2006). The copula-garch model of conditional dependencies: an international stock market application. *Journal of International Money and Finance*, 25(5):827–853.

[96] Kaut, M., Vladimirou, H., Wallace, S. W., and Zenios, S. A. (2007). Stability analysis of portfolio management with conditional value-at-risk. *Quantitative Finance*, 7(4):397–409.

[97] Kenett, D. Y., Huang, X., Vodenska, I., Havlin, S., and Stanley, H. E. (2015). Partial correlation analysis: applications for financial markets. *Quantitative Finance*, 15(4):569–578.

[98] Kim, C.-J. and Nelson, C. (1999). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT Press, Cambridge, MA.

[99] King, M. A. and Wadhwani, S. (1990). Transmission of volatility between stock markets. *Review of Financial Studies*, 3(1):5–33.

[100] Kremer, P. J., Lee, S., Bogdan, M., and Paterlini, S. (2020). Sparse portfolio selection via the sorted L1-norm. *Journal of Banking & Finance*, 110:105687.

[101] Kritzman, M. P. (2000). *Puzzles of finance: six practical problems and their remarkable solutions*. Wiley investment series. Wiley.

[102] Laloux, L., Cizeau, P., Bouchaud, J.-P., and Potters, M. (1999). Noise dressing of financial correlation matrices. *Physical Review Letters*, 83:1467–1470.

[103] Landsman, Z. M. and Valdez, E. A. (2003). Tail conditional expectations for elliptical distributions. *North American Actuarial Journal*, 7(4):55–71.

[104] Langford, E. (2006). Quartiles in elementary statistics. *Journal of Statistics Education*, 14(3).

[105] Lassance, N. and Vrins, F. (2021). Portfolio selection with parsimonious higher comoments estimation. *Journal of Banking & Finance*, 126:106115.

[106] Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

[107] Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

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

[109] Ledoit, O. and Wolf, M. (2020). The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*.

[110] Lee, S. and Stevenson, S. (2003). Time weighted portfolio optimisation. *Journal of Property Investment & Finance*, 21:233–249.

[111] Li, D. and Ng, W.-L. (2000). Optimal dynamic portfolio selection: Multi-period mean-variance formulation. *Mathematical Finance*, 10:387–406.

[112] Liao, W. T. (2005). Clustering of time series data: a survey. *Pattern Recognition*, 38(11):1857–1874.

[113] Lin, W.-L., Engle, R., and Ito, T. (1994). Do bulls and bears move across borders? International transmission of stock returns and volatility. *Review of Financial Studies*, 7(3):507–38.

[114] Linders, D. and Stassen, B. (2016). The multivariate variance gamma model: basket option pricing and calibration. *Quantitative Finance*, 16(4):555–572.

[115] Lo, A. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58:36–52.

[116] Longerstaey, J. and Spencer, M. (1996). *RiskMetrics Technical Document*. J.P. Morgan/Reuters.

[117] Lwin, K. T., Qu, R., and MacCarthy, B. L. (2017). Mean-VaR portfolio optimization: a nonparametric approach. *European Journal of Operational Research*, 260(2):751 – 766.

[118] Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, 36(4):394–419.

[119] Mantegna, R. and Stanley, H. (2000). *An introduction to econophysics: correlations and complexity in finance*. Cambridge University Press.

[120] Mantegna, R. N. (1999). Hierarchical structure in financial markets. *European Physical Journal B - Condensed Matter and Complex Systems*, 11(1):193–197.

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

[122] Maruyama, Y. and Seo, T. (2003). Estimation of moment parameter in elliptical distributions. *Journal of the Japan Statistical Society. Japanese issue*, 33:215–229.

[123] Massara, G. P. and Aste, T. (2019). Learning clique forests. *arXiv preprint arXiv:1905.02266*.

[124] Massara, G. P., di Matteo, T., and Aste, T. (2015a). Network filtering for big data: Triangulated maximally filtered graph. *CoRR*, abs/1505.02445.

[125] Massara, G. P., di Matteo, T., and Aste, T. (2015b). Network filtering for big data: Triangulated maximally filtered graph. *CoRR*, abs/1505.02445.

[126] Massara, G. P., Di Matteo, T., and Aste, T. (2017). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, 5(2):161–178.

[127] McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley.

[128] Mei, X., DeMiguel, V., and Nogales, F. J. (2016). Multiperiod portfolio optimization with multiple risky assets and general transaction costs. *Journal of Banking & Finance*, 69:108–120.

[129] Mercurio, P. J., Wu, Y., and Xie, H. (2020). An entropy-based approach to portfolio optimization. *Entropy*, 22(3).

[130] Mertens, E. (2002). Variance of the IID estimator in Lo (2002). working paper. *University of Basel, Department of Finance*.

[131] Meucci, A. (2010). Fully flexible views: Theory and practice. *arXiv preprint*.

[132] Michaud, R. (1989). The Markowitz optimization enigma: is optimized optimal? *Working Paper, University of Augsburg*, (45):31–42.

[133] Michaud, R. and Michaud, R. (1998). *Efficient Asset Management: A practical Guide to Stock Portfolio Optimization and Asset Allocation*. Harvard Business School Press.

[134] Münnix, M., Shimada, T., Schäfer, R., Leyvraz, F., Seligman, T., Guhr, T., and Stanley, H. (2012). Identifying states of a financial market. *Scientific Reports*, 2:644.

[135] Musmeci, N., Aste, T., and Di Matteo, T. (2014). Risk diversification: a study of persistence with a filtered correlation-network approach. *ArXiv e-prints*.

[136] Musmeci, N., Aste, T., and Di Matteo, T. (2015). Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods. *PLoS ONE*.

[137] Musmeci, N., Aste, T., and Di Matteo, T. (2016a). Interplay between past market correlation structure changes and future volatility outbursts. *Scientific reports*, 6.

[138] Musmeci, N., Aste, T., and Di Matteo, T. (2016b). What does past correlation structure tell us about the future? An answer from network filtering. *ArXiv e-prints*.

[139] Musmeci, N., Nicosia, V., Aste, T., Di Matteo, T., and Latora, V. (2016c). The multiplex dependency structure of financial markets. *ArXiv e-prints*.

[140] Nawrocki, D. (1996). Portfolio analysis with a large universe of assets. *Applied Economics*, (28):1191–1198.

[141] Nevill-Manning, C. G. and Witten, I. H. (1997). Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, 7(1):67–82.

[142] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337.

[143] Opdyke, J. (2007). Comparing Sharpe ratios: So where are the p-values? *Journal of Asset Management*, 8:308–336.

[144] Osuna, E., Freund, R., and Girosi, F. (1997). *Support Vector Machines: Training and Applications*. MIT Press.

[145] Owen, J. and Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance*, 38(3):745–752.

[146] Philippatos, G. C. and Wilson, C. J. (1972). Entropy, market risk, and the selection of efficient portfolios. *Applied Economics*, 4(3):209–220.

[147] Platt, J. C. (1999). *Advances in Kernel Methods*, chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press.

[148] Pola, G. (2016). On entropy and portfolio diversification. *Journal of Asset Management*, 17(4):218–228.

[149] Pozzi, F., Di Matteo, T., and Aste, T. (2012). Exponential smoothing weighted correlations. *European Physical Journal B*, 85(6):175.

[150] Preis, T., Kenett, D. Y., Stanley, H. E., Helbing, D., and Ben-Jacob, E. (2012). Quantifying the behavior of stock correlations under market stress. *Scientific Reports*, 2:750–752.

[151] Priestley, M. B. (1980). State-dependent models: a general approach to non-linear time series analysis. *Cahiers du Centre d'Études de Recherche Opérationnelle*, 22:285–307.

[152] Procacci, P. and Aste, T. (2022a). States characterisation and evolution of the financial system. *Working Paper*.

[153] Procacci, P. F. and Aste, T. (2019). Forecasting market states. *Quantitative Finance*, 19(9):1491–1498.

[154] Procacci, P. F. and Aste, T. (2022b). Portfolio construction and sparse multivariate modelling. *Journal of Asset Management*, 23(6):445–465.

[155] Procacci, P. F., Phelan, C. E., and Aste, T. (2020). Market structure dynamics during COVID-19 outbreak. *ArXiv preprint 2003.10922*.

[156] Quaranta, A. G. and Zaffaroni, A. (2008). Robust optimization of conditional value at risk and portfolio selection. *Journal of Banking & Finance*, 32(10):2046–2056.

[157] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[158] Ren, L., Wei, Y., Cui, J., and Du, Y. (2017). A sliding window-based multi-stage clustering and probabilistic forecasting approach for large multivariate time series data. *Journal of Statistical Computation and Simulation*, 87(13):2494–2508.

[159] Satinover, J. B. and Sornette, D. (2012). Cycles, determinism and persistence in agent-based games and financial time-series: part i. *Quantitative Finance*, 12(7):1051–1064.

[160] Scherer, B. (2007). *Portfolio Construction and Risk Budgeting*. Risk Books.

[161] Scherer, B., Winston, K., and O'Cinneide, C. (2012). *Bayesian Methods In Investing*. Oxford University Press.

[162] Schmitt, T. A., Chetalova, D., Schäfer, R., and Guhr, T. (2013). Non-stationarity in financial time series: generic features and tail behavior. *Europhysics Letters*, 103:58003.

[163] Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business*, 39(1):119–138.

[164] Sharpe, W. F. (1994). The Sharpe ratio. *Journal of Portfolio Management*, 21(1):49–58.

[165] Sklar, M. (1959). *Fonctions de répartition à N Dimensions et leurs marges*. Université Paris 8.

[166] Song, W.-M., Di Matteo, T., and Aste, T. (2012). Building complex networks with platonic solids. *Physical Review*, 85:046115.

[167] Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, 22(3):353–363.

[168] Tchernitsernd, A. and Rubisov, D. H. (2009). Robust estimation of historical volatility and correlations in risk management. *Quantitative Finance*, 9(1):43–54.

[169] Tong, H. (1978). *On a Threshold Model*. NATO ASI Series, Applied Science. Sijthoff & Noordhoff.

[170] Tsay, R. (2005). *Analysis of financial time series*. Wiley series in probability and statistics. Wiley-Interscience, 2. ed. edition.

[171] Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. N. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Science*, 102:10421–10426.

[172] Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative review. *Journal of Machine Learning Research*, 10:66–71.

[173] Van der Weide, R. (2002). GO-GARCH: a multivariate generalized orthogonal GARCH model. *Journal of Applied Econometrics*, 17(5):549–564.

[174] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag.

[175] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

[176] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

[177] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20(4):595–601.

[178] Wang, M., Lin, Y.-H., and Mikhelson, I. (2020). Regime-switching factor investing with hidden Markov models. *Journal of Risk and Financial Management*, 13(12).

[179] Whittle, P. (1951). *Hypothesis testing in time series analysis*, volume 4 of *Uppsala universitet. Statistics*. Almquist and Wiksell.

[180] Xing, X., Hu, J., and Yang, Y. (2014). Robust minimum variance portfolio with L-infinity constraints. *Journal of Banking & Finance*, 46:107–117.

[181] Yao, H., Huang, J., Li, Y., and Humphrey, J. (2021). A general approach to smooth and convex portfolio optimization using lower partial moments. *Journal of Banking & Finance*, pages 106–167.

[182] Zhang, B., Wei, Y., Yu, J., Lai, X., and Peng, Z. (2014). Forecasting VaR and ES of stock index portfolio: A vine copula method. *Physica A: Statistical Mechanics and its Applications*, 416:112–124.

[183] Zhang, H. and Yan, C. (2018). Modelling fundamental analysis in portfolio selection. *Quantitative Finance*, 18(8):1315–1326.

[184] Zolhavarieh, S., Aghabozorgi, S., and Wah Teh, Y. (2014). A review of subsequent time series clustering. *Scientific World Journal*, 2014.

[185] Zumbach, G. and Fernández, L. (2014). Option pricing with realistic ARCH processes. *Quantitative Finance*, 14(1):143–170.

Appendix A

Expectation Maximization

The *Expectation Maximization* (EM) algorithm [58, 127] is a technique for finding maximum likelihood solutions for probabilistic models having latent variables.

Consider a model in which the observed variables are collectively denoted by $\mathbf{X}$ and all the latent (unobserved) variables are denoted with $\mathbf{Z}$. The joint probability distribution of the system $p(\mathbf{X},\mathbf{Z}|\theta)$ is described in terms of the set of parameters θ. The EM algorithm aims at maximizing the likelihood function for $\mathbf{X}$ given θ that is

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta) \ , \tag{A.1}$$

(assuming $\mathbf{Z}$ discrete). The maximization of A.1 is, in most cases, a complex problem and a closed form solution is often not attainable. Instead, the maximization of the complete-data likelihood $p(\mathbf{X},\mathbf{Z}|\theta)$ is often significantly easier.

Defining a generic distribution $q(\mathbf{Z})$ over the latent variables, the following decomposition holds

$$\log p(\mathbf{X}|\theta) = \mathcal{L}(q,\theta) + KL(q||p) \ , \tag{A.2}$$

where:

$$\mathcal{L}(q,\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\} \ , \tag{A.3}$$

$$KL(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \right\} \ . \tag{A.4}$$

Decomposition A.2 is general and it holds for any choice of $q(\mathbf{Z})$. $KL(q||p)$ is the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X},\theta)$ while $\mathcal{L}(q,\theta)$ is a functional of the distribution $q(\mathbf{Z})$.

The EM algorithm is a two-stages optimization method for finding maximum likelihood solutions which makes use of the identity in Eq. (A.2). In the **E-step**, $\mathcal{L}(q,\theta)$ (Eq. (A.3)) is

maximized with respect to $q(\mathbf{X})$ considering $\theta = \theta^{old}$ fixed. Since $\log p(\mathbf{X}|\theta^{old})$ does not depends on $q(\mathbf{Z})$, the solution of this maximization problem is obtained when the Kullback-Leibler distance vanishes, that is when $q(\mathbf{Z})$ is equal to $p(\mathbf{Z}|\mathbf{X},\theta^{old})$.

In the subsequent **M-step**, we fix the distribution $q(\mathbf{Z})$ and $\mathcal{L}(q,\theta)$ is maximized with respect to θ to obtain the new, 'updated values' θ^{new}. The operation is then iterated. As shown in [27], the quantity being maximized in the M-step is, indeed, the expectation of the complete-data log-likelihood. This will cause $\mathcal{L}$ to increase (unless it is already at its maximum) and, hence, the log-likelihood function to increase as well, converging eventually to its maximum.

A.1 EM for Gaussian Mixtures

When it comes to modelling real data, simple distributions are often unable to capture the probabilistic structure of the dataset, while linear superimposition of two or more distribution can give better results. *Mixture distributions* are probabilistic models in which we assume that data are generated by a linear combination of basic distributions.

In a Gaussian mixture model, we consider the superimposition of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x};\mu_k,\Sigma_k)\ . \tag{A.5}$$

Each Gaussian density $N(\mathbf{x};\mu_k,\Sigma_k)$ is called a *component* of the mixture and the parameters π_k are called *mixing coefficients*, where it is trivial to notice that

$$0 \leq \pi_k \leq 1 \quad \textit{and} \quad \sum_{i=1}^{K} \pi_k = 1 \tag{A.6}$$

In order to obtain a convenient representation involving an explicit latent variable, consider a K-dimensional binary random variable $\mathbf{z}$ such that $z_k \in \{0,1\}$ and $\sum_k z_k = 1$. The marginal distribution $p(\mathbf{z})$ is defined in terms of the mixing coefficients such that

$$p(z_k = 1) = \pi_k\ . \tag{A.7}$$

Given the binary representation of $\mathbf{z}$ we can write the marginal distribution over $\mathbf{z}$ and the

conditional distribution $p(\mathbf{x}|\mathbf{z})$ as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{A.8}$$

and

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)^{z_k} \, . \tag{A.9}$$

Using the results A.8 and A.9 we can obtain the marginal distribution over $\mathbf{x}$ as

$$p(\mathbf{x}) = \sum_z p(\mathbf{x}, \mathbf{z}) = \sum_z p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)^{z_k} \, , \tag{A.10}$$

hence the marginal distribution over $\mathbf{x}$ is a Gaussian mixture of the form A.5.

Consider an *N observations x D variables* dataset that we wish to model using a mixture of Gaussians. The corresponding latent variables will be denoted by a *N x K* matrix $\mathbf{Z}$. If we assume the data points are IID, then the log likelihood function is given by

$$\log p(\mathbf{X}; \pi, \mu, \Sigma) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_k; \mu_k, \Sigma_k) \right\} \, . \tag{A.11}$$

Given the presence of the summation over k inside the logarithm in A.11, we cannot derive a closed form solution for the maximization of the likelihood function. To see this, if we take the derivative of $\log p(\mathbf{X}; \pi, \mu, \Sigma)$ with respect to the means $\boldsymbol{\mu}_k$ and set it to zero, we obtain

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n; \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n; \mu_j, \Sigma_j)} \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) = 0 \tag{A.12}$$

where the term

$$\gamma(z_n k) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n; \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n; \mu_j, \Sigma_j)} \tag{A.13}$$

called *responsability*, depends on the parameters μ_k in a complex way. The responsabilities denote, indeed, the conditional probability of $\mathbf{z}$ given $\mathbf{x}$. Rearranging A.12 and using the definition of responsabilities

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_n k) \mathbf{x}_n \tag{A.14}$$

where $N_k = \sum_{n=1}^{N} \gamma(z_n k)$. Following a similar approach, we can maximize with respect to Σ_k and π_k (taking into account the constraints A.6)

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_n k) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \tag{A.15}$$

where μ_k is obtained from A.14, and

$$\pi_k = \frac{N_k}{N} \, . \tag{A.16}$$

One way to solve this problem is the EM algorithm discussed in previous section. First we initialize the means $\boldsymbol{\mu}_k$, the covariances $\boldsymbol{\Sigma}_k$ and the mixing coefficients π_k and compute the initial value of the log likelihood. In the **E step** we evaluate the responsabilities using the current parameters values using A.13. In the **M step** we re-estimate the parameters using the current responsablities and A.14, A.15 and A.16. Given the estimated parameters, we evaluate the log likelihood A.11 and check for convergence of the parameters or the log likelihood. If convergence is not satisfied, we iterate from the E step.

Appendix B

Testing Sharpe Ratio

Considering the basic definition of Sharpe ratio in common usage

$$SR = \frac{\mu}{\sigma} \tag{B.1}$$

where μ is the sample mean of returns and σ is the sample standard deviation, [94] and, more recently [115], derived the asymptotic distribution of SR under the assumption of normal and IID returns

$$SR \overset{a}{\sim} \mathcal{N}\left(\frac{\mu}{\sigma}, \frac{1}{n}\left(1+\frac{\mu^2}{2\sigma^2}\right)\right) . \tag{B.2}$$

Mertens (2002) relaxed the normality assumption and presented a derivation that is valid under IID *generally*

$$\sqrt{T}\left(SR-\widehat{SR}\right) \overset{a}{\sim} \mathcal{N}\left(0, 1+\frac{1}{2}SR^2 - SR\,\gamma_3 + SR^2\left[\frac{\gamma_4-3}{4}\right]\right) \tag{B.3}$$

where $\gamma_4 = \frac{\mu_3}{\sigma^3}$ and $\gamma_4 = \frac{\mu_4}{\sigma^4}$. In other worlds, to relax the normality assumption we need to adjust for kurtosis and skewness. Recently, Christie (2005), using a GMM approach, derived the asymptotic distribution of SR relaxing also the IID requirement, considering the variance of $\widehat{SR}$

$$Var(\sqrt{T}\widehat{SR}) = \mathbb{E}\left(\frac{SR^2\mu_4}{4\sigma^4} - \frac{SR\left[(R_t-R_{ft})(R_t-\mu)^2-(R_t-R_{ft})\sigma^2\right]}{\sigma^3} + \frac{(R_t-\mu)^2}{\sigma^2} - \frac{2(R_t-\mu)}{\sigma} + \frac{3SR^2}{4}\right) . \tag{B.4}$$

Opdyke (2007) provided a more convenient formulation for the same general case. Deriving the equivalence among B.4 and B.5

$$Var(\sqrt{T}\widehat{SR}) = 1+\frac{SR^2}{4}\left[\frac{\mu_4}{\sigma^4}-1\right] - SR\frac{\mu_3}{\sigma_3} \tag{B.5}$$

provided the asymptotic distribution of $\widehat{SR}$ as in B.6

$$\sqrt{T}\left(SR-\widehat{SR}\right) \overset{a}{\sim} \mathcal{N}\left(0, 1+\frac{SR^2}{4}\left[\frac{\mu_4}{\sigma^4}-1\right]-SR\frac{\mu_3}{\sigma^3}\right) \tag{B.6}$$

and from which the estimated standard error is

$$\widehat{SE(\widehat{SR})} = \sqrt{\left[1+\frac{\widehat{SR^2}}{4}\left(\frac{\hat{\mu}_4}{\hat{\sigma}^4}-1\right)-\widehat{SR}\frac{\hat{\mu}_3}{\hat{\sigma}^3}\right]/(T-1)}\,. \tag{B.7}$$

Based on B.7, we can easily derive the confidence bound B.8 for $\widehat{SR}$ significance testing

$$\widehat{SR} \pm z_{crit}\ \widehat{SE(\widehat{SR})} \tag{B.8}$$

where z_{crit} is the critical value of the standard normal distribution corresponding to the significance level α of choice.

It is also worth to empathize that we cannot directly test for comparison among Sharpe ratios of the two clusters as in Opdyke (2007) since the number of observations in each cluster is not ensured to be the same.

Appendix C

Optimization in Support Vector Machines

Support Vector Machines (SVM) finds a linear separating hyperplane with the maximal margin in the features' higher dimensional space defined by the a feature mapping $\phi(\boldsymbol{x})$. Given a training set of feature-label pairs $(x_n, t_n), i = 1, ..., N$ where $t_n \in \{-1, 1\}$ and $x_n \in \mathbb{R}^p$, the SVMs [33, 47, 175] require the solution to the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & t_n(\boldsymbol{w}^T\phi(\boldsymbol{x}+b)) \geq 1-\xi_n, \\ & \xi_n \geq 0 \end{aligned} \tag{C.1}$$

where $(\boldsymbol{w}^T\phi(\boldsymbol{x}+b))$ is the model of the form $y(\cdot)$ in Equation (5.3); $C > 0$ is a penalty parameter controlling the trade off between the margin and variables ξ_n; xi_n is the *slack variable*, with one slack variable training data point. The slack variables where introduced by Bennett (1992) and Cortes and Vapnik (1995) and these are defined as $xi_n = 0$ if the $n-th$ data point is inside or on the correct margin boundary and $xi_n = |t_n - y(\boldsymbol{x_n})|$ otherwise. Slack variables where introduced to remove the assumption that training data points are linearly separable in the feature space $\phi(\cdot)$, defining the *soft margin* SVM presented in C.1. In this way we can maximize the margin while penalizing the points that lie on the wrong size of the margin boundary.

The Lagrangian corresponding to the constrained maximization problem in C.1 is given by

$$L(\mathbf{w}, b, \xi, \mathbf{a}, \boldsymbol{\tau}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\tau_n\xi_n \tag{C.2}$$

where $a_n \geq 0$ and $\tau_n \geq 0$ are the Lagrange multipliers. The corresponding set of *Karush-*

Kuhn-Tucker (KKT) conditions are

$$a_n \geq 0 \tag{C.3}$$

$$t_n y(x_n) - 1 + \xi_n \geq 0 \tag{C.4}$$

$$a_n(t_n y(x_n) - 1 + \xi_n) = 0 \tag{C.5}$$

$$\tau_n \geq 0 \tag{C.6}$$

$$\xi_n \geq 0 \tag{C.7}$$

$$\xi_n \tau_n = 0 \tag{C.8}$$

$$\tag{C.9}$$

where $n = 1, ..., N$. Optimizing with respect to $\mathbf{w}$, b and ξ_n and using the definition 5.3 for $y(\cdot)$, we obtain [27]

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \tag{C.10}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} a_n t_n = 0 \tag{C.11}$$

$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow a_n = C - \tau_n \tag{C.12}$$

using these results we obtain the dual Lagrangian form

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{p=1}^{P} a_n a_p t_n t_p k(\mathbf{x}_n, \mathbf{x}_p) \tag{C.13}$$

subject to the constraints:

$$0 \leq a \leq C \tag{C.14}$$

$$\sum_{n=1}^{N} a_n t_n = 0 \tag{C.15}$$

where $k(\mathbf{x}_n, \mathbf{x}_p)$ is the kernel function, as discussed in Section 5.1.2. This is, again, a quadratic programming problem. The dual formulation allowed to turn the optimization C.1 over P variables (number of features) into the dual problem C.13, over N variables (length of features' examples). It is worth emphasizing that in case the set of basis functions $\phi(\mathbf{x})$ is fixed with $P < N$, this conversion could appear disadvantageous. However, it allows to reformulate the model using kernels and, therefore, to efficiently map features into

feature spaces whose dimensionality exceeds data points, including infinite feature spaces (as it is the case of RBF kernels in the experiment discussed in Chapter 5.1.2). Moreover, in order to classify new data points using the trained model, we can express $y(\boldsymbol{x})$ defined in 5.3 in terms of the parameters a_n and the kernel function obtaining:

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \ . \tag{C.16}$$

Given the KKT conditions, for every data point we have that either $a_n = 0$ or $a_n \geq 0$. If $a_n \geq 0$, then by C.5 these data points must satisfy

$$t_n y(\boldsymbol{x_n}) = 1 - \xi_n \tag{C.17}$$

If $a_n < X$ then the point lies on the margin since by C.12 $\tau_n > 0$ that from C.8 implies $\xi_n = 0$. Points with $a_n = C$ lie inside the margin and they can have either $\xi_n \leq 1$ (correctly classified) or $\xi_n > 1$ (misclassified). It is crucial to notice that the subset of data points for which $a_n = 0$ does not contribute to the predictive model C.16 and hence plays no role in making predictions for new data points. The remaining data points constitute the *support vectors*. This property is central to the efficient applicability of SVMs and provided an added advantage to the use of the dual representation since, once the model is trained, a significant amount of data can be discarded and only support vectors used to make predictions.

In general, the solution of a quadratic programming problem in M variables has computational complexity $O(M^3)$ [27] and, although predictions are made using only support vectors, the training phase uses the whole dataset. Different efficient algorithms to solve the quadratic programming problem C.13 have been proposed and most popular approaches break down the optimization into a series of smaller quadratic programming problems. Among these it is worth mentioning the *chunking* method [174], which exploits the fact that the value of the Lagrangian is unchanged by removing the columns and rows of the kernel corresponding to zero valued Lagrange multipliers a_n, and *Decomposition Methods* [144], which solves smaller quadratic programming problems of *fixed* size in a numerical way. One of the most widely used approach and the one that we implemented in the experiment discussed in Section 5.1.2 is the *sequential minimal optimization* (SMO) [147] which considers just two Lagrange multipliers at time in order to identify the non zero ones, providing an alternative and efficient approach to the same goal of the *chunking* method. In this case, in fact, the subproblem can be solving analytically avoiding numerical solutions.

SMO is found to have a complexity that is either linear or quadratic with the number of data points, depending on the application [27].

Having solved the quadratic programming problem and obtained a solution for **a**, to determine the parameter b 5.3 we note that support vectors a_n for which $0 < a_n < C$ have $\xi_n = 0$ and, therefore, $t_n y(x_n) = 1$. Using C.16

$$t_n \left(\sum m \in \mathcal{S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1 \tag{C.18}$$

where $\mathcal{S}$ denotes the set of indices of the support vectors. Instead of solve the equation for b by considering an arbitrary support vector, a numerical stable solution by multiplying the whole set (S) by t_n (notice that $t_n^2 = 1$) and then averaging over all the support vectors obtaining

$$b = \frac{1}{N_{\mathcal{M}}} \sum n \in \mathcal{M} \left(t_n \sum m \in \mathcal{S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right) \tag{C.19}$$

where $\mathcal{M}$ indicates the set of indices for which $0 < a_n < C$.

Appendix D

Properties of Elliptical Distributions

In this section we recall some useful properties of Elliptical Distribution which we referred to in our discussion and particularly in Section 3.4.3.

Property 1 (Distribution Definition). *Consider an n-dimensional random vector* $\mathbf{X} = (X_1, ..., X_n)$. $\mathbf{X}$ *has a multivariate elliptical distribution with location parameter* $\boldsymbol{\mu}$ *and dispersion parameter* $\boldsymbol{\Omega}$*,written as* $\mathbf{X} \sim \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ *if its characteristic function* ϕ *can be expressed as*

$$\phi_X(\mathbf{w}) = \mathbb{E}(e^{i\mathbf{w}\mathbf{X}}) = e^{i\mathbf{w}\boldsymbol{\mu}}\,\psi\left(\frac{1}{2}\mathbf{w}\boldsymbol{\Omega}\mathbf{w}^T\right), \tag{D.1}$$

for some location parameter $\boldsymbol{\mu} \in \mathbb{R}^{1\times n}$*, positive-definite dispersion matrix* $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$ *and for some function* $\psi(\cdot) : [0,\infty) \to \mathbb{R}$ *such that* $\psi\left(\sum_{i=1}^{n} w_i^2\right)$ *is a characteristic function, which is called characteristic generator. If* $\mathbf{X} \sim \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ *and if its density* $f_X(\mathbf{X})$ *exists, it is of the form defined in Eq. (3.10).*

Property 2 (Density Generator). *The function* $g(\cdot)$ *defined in Section 3.4.3 is guaranteed to be density generator if the following condition holds*

$$\int_0^\infty x^{n/2-1} g_n(x)dx < \infty. \tag{D.2}$$

Property 3 (Affine Equivariance). *If* $\mathbf{X} = (X_1, ..., X_n)$ *is an n-dimensional elliptical random variable with location parameter* $\boldsymbol{\mu}$ *and dispersion parameter* $\boldsymbol{\Omega}$ *so that* $\mathbf{X} \sim \mathcal{E}_X(\boldsymbol{\mu}, \boldsymbol{\Omega})$*, then for any vector* $\mathbf{a} \in \mathbb{R}^{1\times m}$ *and any matrix* $\mathbf{B} \in \mathbb{R}^{m\times n}$ *the following affine equivariace holds*

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim \mathcal{E}_Y(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Omega}\mathbf{B}). \tag{D.3}$$

In other words, any linear combination of multivariate elliptical distributions is another elliptical distribution.

In the special cases of normal, Student-t and Cauchy distributions, the induced density generators are m-dimensional version of the original generator of $\boldsymbol{X}$.

For the proof of Properties 1,2 and 3, we refer to [69].

This implies that any portfolio $Y = \beta_1 X_1 + ... + \beta_n X_n$ of elliptically distributed variables is distributed accordingly with a (univariate) elliptical distribution, which is a location-scale distribution. Furthermore, for any univariate elliptical distribution all moments can be obtained from the first and second moments (if defined). In particular, for centered variables with zero mean ($\mu_Y = 0$), the resulting distribution of Y is symmetrical around zero and it has all odd moments equal to zero and all even moments given by

$$\mu_{2m} = c_m \mu_2^m,$$

with

$$c_m = \frac{(2m)!}{(2^m m!)} \frac{\psi^{(m)}(0)}{(\psi^{(1)}(0))^m}.$$

Where $\psi^{(m)}(0)$ indicated the m^{th} derivative of $\psi(\omega)$ computed at $\omega = 0$.
As an example, in the normal (0,1) case, $\mu_2 = 1$, $c_m = 0$ for all $m = 1, 2, ...$, the kurtosis is $\mu_{(4)} = \frac{4!}{2^4 4!} = 3$, and $\mu_{(2m)} = \frac{(2m)!}{(2^m m!)}$. For the proof we refer to [26], which derived this property by succesive differentiations of $\phi(\cdot)$, and to [122], which attained the same result by expressing the elliptical distribution in terms of a random vector with uniform distribution on the unit sphere.

Therefore the mean-variance optimization is of general applicability and relevance for any portfolio generated from multivariate elliptically distributed variables.

Appendix E

Orthogonal GARCH Estimation

In this section, I report details on the O-GARCH estimation discussed in Section 3.5. In particular, Tables 3.1 and 3.2 report the average AIC and BIC statistics across the 100 resamplings I considered in the experiment. Table E.1 below reports Median, 5^{th} and 95^{th} percentiles all the statistics obtained. The table shows that the GARCH(1,1) specification delivered the lower AIC and BIC statistics for each of the main percentiles considered. In other words, the GARCH(1,1) specification selected in our experiment is the preferred specification according to the AIC and BIC criteria in ALL cases, and not only in mean across resamplings.

		AIC			BIC		
	Train Obs	5^{th}	Median	95^{th}	5^{th}	Median	95^{th}
GARCH(1,1)	**101**	-726	-679	-539	-718	-672	-531
	125	-866	-831	-643	-857	-823	-635
	250	-1610	-1561	-1250	-1599	-1551	-1240
	500	-3111	-2978	-2562	-3098	-2965	-2550
	750	-4624	-4309	-3869	-4610	-4295	-3855
	1000	-6092	-5684	-5017	-6077	-5669	-5002
	1500	-8735	-8006	-7388	-8719	-7990	-7372
GARCH(2,2)	**101**	-723	-676	-535	-709	-663	-522
	125	-862	-827	-640	-848	-813	-626
	250	-1606	-1558	-1247	-1588	-1540	-1230
	500	-3107	-2975	-2561	-3086	-2954	-2539
	750	-4620	-4307	-3867	-4597	-4284	-3843
	1000	-6088	-5681	-5015	-6063	-5656	-4991
	1500	-8733	-8007	-7387	-8707	-7980	-7361
GARCH(3,3)	**101**	-719	-672	-532	-701	-654	-514
	125	-858	-824	-637	-838	-804	-617
	250	-1602	-1554	-1245	-1578	-1530	-1220
	500	-3103	-2972	-2558	-3074	-2942	-2528
	750	-4616	-4304	-3864	-4584	-4272	-3832
	1000	-6085	-5678	-5013	-6050	-5644	-4979
	1500	-8732	-8007	-7386	-8695	-7969	-7349

Table E.1: AIC and BIC information criteria corresponding to different GARCH specifications. Median, 5^{th} and 95^{th} percentiles across 100 resamplings.