Politecnico di Torino

Collegio di Ingegneria Elettronica, delle Telecomunicazioni e Fisica (ETF)

Firm Dynamics and Lead-Lag Effects: Statistical Analysis of Growth Rates

Master's Thesis in Physics of Complex Systems

Candidate:
Pietro Bicocchi

Supervisors:
Prof. L. Dall'Asta
Prof. D. Farmer
PhD L. Mungo
PhD J. Moran

October 2023

Acknowledgements

I would like to express my deepest gratitude to all those who have supported and guided me throughout my academic journey. This accomplishment would not have been possible without the invaluable assistance and encouragement of many individuals and institutions.

I am profoundly grateful to Luca who arranged my visit and supported me with his guidance, support, and patient mentorship that have shaped not only my research but also my personal growth. Luca is the type of person who always finds a way to help others. Despite being one of the most brilliant individuals I've had the privilege to meet, Luca remains one of the humblest. This speaks volumes about him. My sincere gratitude goes to José. His remarkable blend of expertise and dedication have been essential in making this thesis a reality. His constructive feedback forms the basis of this research. The generous sharing of knowledge, the countless hours spent discussing research ideas, and the willingness to go above and beyond in support of my work are examples of Luca and José's exceptional dedication to mentoring young scholars.

I am deeply thankful to Professor Dall'Asta, whose supervision and support have been essential in making this thesis a reality.

I want to express my heartfelt thanks to Professor Farmer, who gave me this opportunity, for the incredible experience I had as a visitor at the Institute for the New Economic Thinking. The biggest lesson I absorbed is the importance of asking fundamental questions in research. Whether it was in seminars, talks, or even casual conversations, I couldn't help but notice the passion and commitment of everyone at INET to delve deep and explore fundamental issues. A special thanks to Dorothy, who is truly the soul of the institute with her dedication and attention to detail.

During my academic journey, I had the privilege of crossing paths with an exceptional group of individuals, friends and colleagues who have enriched my life in countless ways. My heartfelt gratitude to Bianca, Luca, Marta, Michele, Niccolò, and Pietro. A special thanks to Michele, for his friendship, support and inspiration throughout this journey.

Last but certainly not least, I'd like to express my gratitude to those who supported me from the very beginning. Thank you Mariah for being the incredible person you are. Your love, support, and presence in my life mean more to me than words can express. My heartfelt thanks to my family, without your help and support this journey would have not been possible. To my parents, who have always been my role models, thank you for believing in me. Your sacrifices, both seen and unseen, have paved the way for my education, and your belief in me has been a constant source of motivation. To my brother, his presence has been a source of strength and joy.

Abstract

This thesis explores the complex dynamics of companies within supply chains, focusing on economic growth rates and the interplay between price and sales returns. The primary objective is to identify sectorial correlations while effectively filtering out noisy correlations in multivariate datasets. The research employs methodologies rooted in Random Matrix Theory and non-linear dynamics tools to separate meaningful relationships ("signal") from random fluctuations ("noise").

An in-depth investigation of principal modes reveals clear sectoralization patterns and strong nonlinearity features, supporting the idea that these time series are products of non-linear dynamical processes. This claim is further validated through surrogate data techniques and sensitivity analyses with market indicators, highlighting visible correlations.

A critical aspect of this study revolves around the disparity between sales and price returns. Initially, differences in structures were attributed to variations in sampling frequencies. However, this research demonstrates that such disparities persist even when downsampling price data. A novel finding emerges, suggesting that these two types of returns are not merely different realizations of the same distribution process but rather result from separate generative mechanisms. This result encourages further exploration.

Additionally, this research delves into temporal correlation dynamics, particularly lead-lag effects. Various techniques, including the analysis of the lagged correlation matrix and Hilbert Principal Component Analysis, provide insights into significant lags and identify leading and lagging companies within the network of interconnected variables.

In conclusion, this thesis underscores the need for a deeper understanding of the connection between sales and price returns and the influence of external factors. It sets the stage for future investigations and the development of fine-grained macroeconomic models, such as Agent-Based Models (ABMs), that can replicate detailed economic time series data.

Keywords: Econophysics, Complex Systems, Random Matrix Theory, Non-linear Dynamics.

Abstract in lingua italiana

Questa tesi esplora le intricate dinamiche delle companies all'interno delle supply chain, concentrandosi sui tassi di crescita economica e sull'interazione tra prezzi degli azioni e vendite. L'obiettivo principale è identificare correlazioni settoriali, filtrando efficacemente le correlazioni rumorose in set di dati multivariati. La ricerca utilizza metodologie avanzate basate sulla Random Matrix Theory (RMT) e strumenti di dinamica non lineare per separare le relazioni significative ("segnale") dalle fluttuazioni casuali ("rumore").

Un'approfondita indagine delle principal modes rivela chiari modelli settoriali e forti proprietà di non linearità, che sostengono il fatto che queste serie temporali sono il prodotto di processi dinamici non lineari. Questa considerazione viene ulteriormente convalidata attraverso tecniche di surrogate data e analisi di sensitività con indicatori di mercato, evidenziando correlazioni evidenti.

Un aspetto critico di questo studio riguarda la disparità tra le vendite e i prezzi delle azioni. Inizialmente, le differenze nelle strutture dei dati venivano attribuite a variazioni nelle frequenze di campionamento. Tuttavia, questa ricerca dimostra che tali disparità persistono anche quando i dati dei prezzi vengono campionati a frequenze diverse. Emerge un nuovo risultato, suggerendo che questi due tipi di ritorni non sono semplicemente diverse realizzazioni dello stesso processo di distribuzione, ma derivano da meccanismi generativi differenti. Questo risultato incoraggia ulteriori approfondimenti.

Inoltre, questa ricerca approfondisce le dinamiche temporali delle correlazioni, in particolare gli effetti lead-lag. Diverse tecniche, tra cui l'analisi della lagged correlation matrix e l'Hilbert PCA, forniscono informazioni sui time lag significativi e identificano le aziende leading e lagging nella rete di variabili interconnesse.

In conclusione, questa tesi sottolinea la necessità di una comprensione più approfondita del legame tra vendite e prezzi di azioni e dell'influenza di fattori esterni. Questo lavoro prepara il terreno per future indagini e lo sviluppo di modelli macroeconomici dettagliati, come gli Agent Based Models (ABM), in grado di replicare dati dettagliati sulle serie temporali economiche.

Contents

Introduction

In recent decades, extensive research has been dedicated to exploring **complex systems** across various scientific domains, including the physical, biological, social, and economic sciences. These systems are characterized by numerous interconnected components, connected through complex relationships, giving rise to large-scale collective structures or behaviors.

A fundamental concept, introduced by physicist Philip W. Anderson in his seminal essay "More Is Different" [3], is that as we ascend the ladder of scientific complexity, transitioning from the microscopic to the macroscopic scale, new phenomena and principles emerge. These principles cannot be deduced from the lower-level constituents alone. This idea led to a profound shift in scientific thinking, highlighting the recognition that within the intricate network of interactions within complex systems, "more" indeed becomes "different".

Econophysics is a multidisciplinary field that combines the principles of physics and statistical mechanics with the complexities of economics and finance. This concept first appeared in the late 20th century in response to the realization that traditional economic models often struggled to capture the behaviors of financial markets and economic systems [42]. This approach is rooted in the belief that many aspects of economic and financial systems can be understood and modeled using the same mathematical and statistical principles that describe physical systems, such as the behavior of particles in a gas or the diffusion of molecules in a liquid. Economic systems are inherently complex and composed of a large number of interacting components, including individual traders, investors, institutions, banks, and assets. These components collectively give rise to emergent behaviors that are often difficult to predict using traditional economic models.

Within this landscape, our research aims to explore the entangled dynamics of companies in the supply chain within the context of economic growth rates, with a specific focus on return prices and sales. We employ methodologies rooted in Random Matrix Theory and non-linear dynamics tools to analyze these multivariate datasets. These highly volatile quantities provide an ideal experimental setting to test methods from Random Matrix

Theory, as the interactions between individual components are obscured by noisy observations. We seek to determine if any meaningful correlation exists among these firms. Merely computing the empirical correlation matrix between firms is insufficient to address these questions, as the empirical correlation matrix tends to contain a significant amount of noise. Consequently, we aim to familiarize ourselves with such phenomena and develop the ability to discern and separate signals from noise, which serves as a valuable exercise in this context. We do this by highlighting the distinctions between the sample covariance matrix and the population correlation matrix while exploring its eigenvalues and eigenvectors.

Our research begins with a comprehensive literature review, laying the groundwork for understanding the analytical tools and principles that underpin our investigation, such as the Marcenko-Pastur distribution. To make sense of these large multivariate datasets, we address the practical aspects of data, including collection, coverage, and preprocessing, ensuring the empirical foundation for our analysis is robust.

Subsequently, we move on to the statistical analysis of logarithmic returns and their correlations. We employ the Principal Component Analysis method on the correlation matrix to help us investigate its significance and investigate the distribution of eigenvalues and eigenvectors. Our analysis extends to the study of mode signals, both in the time and frequency domains. When analyzed, these mode signals appear to exhibit strong nonlinearity features. At this stage, we intend to evaluate whether the modes obtained through eigendecomposition can be attributed to "noise" or if they exhibit features of non-linearity, using surrogate data methods for validation.

A key question addressed in this thesis is whether sales returns and price returns share a common generative process, despite their structural differences and disparities. Initially, we hypothesized that the different sampling frequencies of the time series might account for this discrepancy. However, we have demonstrated that this is not the case. We will demonstrate this by delving into the comparison of the eigenstructure of the empirical correlation when downsampling our price data. Additionally, we recreate the synthetic time series by introducing an element of randomness into the realization of the mode signals. In doing so, we ensure that they display a comparable correlation structure and retain certain statistical properties of the original data. Comparing the distribution of the spectral distance of the synthetic time series to the original one, and the spectral distance between the sales and prices we aim to determine whether the sales and prices are just two "draws" from the same generative distribution or not.

Finally, we explore lead-lag effects, trying to understand temporal dependencies in economic data. The goal is to provide a coherent perspective on the complete temporal correlation dynamics within a network of interconnected variables, public companies in our case. How can we approach this general problem? In other words, how can we gain insights into the network dynamics at different points in time? How do changes in one node simultaneously and differentially influence other nodes? Our primary focus is to address these questions, along with several related ones. We introduce the concept of the lagged correlation matrix to identify time lags that maximize lead-lag relationships and leverage the singular value decomposition of this matrix to gain insights. We also employ an unconventional approach called Hilbert Principal Component Analysis, inspired by climate science, to explore potential lead-lag relationships between variables. Our objective is to provide a theoretical framework for analyzing large datasets that consider both the behavior of individual time series and the inter-correlations between units.

We believe that our research, which combines innovative methodologies with empirical investigation, contributes novel insights into understanding the interplay between sales and prices, ultimately enhancing our comprehension of economic and financial systems.

1 | Literature Review

1.1. Random Matrix Theory in a nutshell

In recent years, Random Matrix Theory (RMT) has emerged as a powerful and versatile tool for understanding complex systems in various fields like finance, econophysics, and network dynamics. It provides a mathematical framework that allows researchers and practitioners to make sense of large correlation matrices, which are often characterized by intricate structures and noisy elements. In this chapter, we aim to introduce you to the world and tools of RMT and to show how it plays a crucial role in studying the interactions between firms in the supply chain, well known to include a lot of noise.

When we analyze the supply chain of firms, we encounter numerous challenges. Traditional statistical methods struggle to capture the underlying patterns due to the presence of noise. That's where RMT comes into the picture. By applying its principles, we can uncover meaningful information hidden within the complexity of the correlation matrices. These matrices provide insights into how firms within the supply chain influence each other, helping us understand the dynamics at play.

In this chapter, we will explore the theoretical foundations of RMT. We'll explain the key concepts, techniques, and mathematical tools that underpin its framework. This theoretical foundation will enable us to explore practical applications in macroeconomics and finance. However, it is important to note that this treatment is not exhaustive. Furthermore, we have chosen to present these results in a narrative style rather than a strictly rigorous Theorem-Lemma manner. To address these limitations, we offer references where more precise and detailed statements can be found.

1.1.1. Setting the stage

In today's era of "Big Data" the exponential growth of data sets across diverse fields, including physics, image analysis, genomics, epidemiology, engineering, economics, and finance, constitutes a significant challenge. A question to be addressed is: how to identify common factors or causes that explain the joint dynamics of numerous quantities?

Imagine analyzing daily returns of stocks, temperature variations across different regions or various health indicators within a population. To quantify the similarities and relations between these observables, we rely on an essential mathematical object: the $N \times N$ correlation matrix C. By examining its eigendecomposition, i.e. its eigenvalues and eigenvectors, we can capture the most influential common dynamical "modes" or linear combinations of the original variables that exhibit the largest variance. This widely-known method is called Principal Component Analysis (PCA).

However, a major concern arises when applying these statistical techniques in practice. The expectation value in calculating the standard Pearson correlation matrix:

$$C_{ij} = E[y_i y_j], \quad i, j \in [1, N]$$

is seldom precisely computable because the underlying distribution of the variables is often unknown. Empirically, we attempt to estimate $\mathbf{C}$ by collecting a substantial number of realizations of the N variables, forming a data matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N) \in \mathbb{R}^{N \times T}$. Then, assuming to have a large number T of realizations, we compute the sample correlation matrix estimator $\mathbf{E}$ as

$$E_{ij} = \frac{1}{T} \sum_t Y_{it} Y_{jt}$$

where Y_{it} represents the realization of the i-th observable at "time" t.

While classical multivariate statistics demonstrate that $\mathbf{E}$ converges to $\mathbf{C}$ when N is significantly smaller than the number of observations T, problems arise when N becomes large. Estimating all elements of the correlation matrix or even just its eigenvalues becomes challenging when the number of observations is not very large compared to N itself. For example, in stock returns, T represents the total number of trading days. Unfortunately, data are not always available and sometimes to make sure that our data are stationary, i.e. that the underline statistics it is not changing during the time of observation, we have to restrict the time in which we analyze them. For these reasons, T is likely to be not large enough. Therefore, it becomes crucial to distinguish between the empirical correlation matrix $\mathbf{E}$ and the "true" correlation matrix $\mathbf{C}$ of the underlying statistical process.

A common measure to quantify the dimensionality of the problem, which plays a crucial role in understanding the behavior of random matrices in the large dimension limit, is given by the ratio between the number of variables (N) and the number of observations (T): $q = \frac{N}{T}$.

When the ratio q is small, i.e. $q \to 0$, meaning there are many more observations compared to the number of variables, the random matrix behavior is relatively well understood. In this regime, certain statistical properties, such as the eigenvalue distribution, can be described using classical results from multivariate statistics.

However, as q becomes larger, and the number of variables (N) approaches or exceeds the number of observations (T), the behavior of random matrices becomes more complex. In this high-dimensional regime, new phenomena emerge, and the classical statistical results may no longer hold. Analyzing the properties of random matrices in this regime is a challenging problem in RMT.

For instance, in the case of stocks, where N is typically ~ 500 and T is ~ 2500 (representing 10 years of daily data), q equals 0.2. This example highlights the importance of reporting for the effects induced by non-negligible q values in various applications.

In the next sections, we explore the differences between $\mathbf{E}$ and $\mathbf{C}$ and discuss the efficacy of reconstructing $\mathbf{C}$ from the knowledge of $\mathbf{E}$ in situations where both N and T become large. This scenario, known as the large dimension limit (LDL) or the "Kolmogorov regime" requires cautious analysis.

1.1.2. Overview and analytical tools

The subsequent section aims to provide an overview of key results that are crucial for comprehending the methods and tools employed in the upcoming sections. While the scope of these results extends beyond what we will be mentioning, our discussion will primarily focus on the applications of Random Matrix Theory (RMT) to financial markets. This topic has gained significant attention in the last decade, with numerous papers dedicated to it [10, 14, 17, 18, 39, 52].

One of the central problems in high-dimensional statistics is accurately estimating the population (true) covariance matrix. This problem was first addressed by J. Wishart in 1928, who derived the conditional distribution of the sample covariance matrix, denoted by $\mathbf{E}$, for the case of T i.i.d. Gaussian realizations of the set $\mathbf{y} = \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$ [63]. The distribution of $\mathbf{E}$ given $\mathbf{C}$ can be expressed as follows:

$$\mathcal{P}_W(\mathbf{E} \mid \mathbf{C}) = \frac{T^{NT/2}}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{E})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\operatorname{Tr}^{-1}\mathbf{E}} \tag{1.1}$$

Here, $\Gamma_N()$ represents the multivariate Gamma function with parameter N. T. W. Anderson later derived the marginal probability density distribution of the eigenvalues for finite N and T as:

$$\rho_N(\lambda) = \frac{1}{N}\sum_{k=0}^{N-1} \frac{k!}{T-N+k} \left[L_k^{T-N}(\lambda)\right]^2 \lambda^{T-N} e^{-\lambda} \tag{1.2}$$

This expression assumes that $T > N$, and L_k^l denotes the Laguerre polynomials. These results provide insights into characterizing the behavior of $\mathbf{E}$ under finite T and N.

However, an important question arises: how well does the estimator $\mathbf{E}$ perform as we increase N? This forms another central topic in RMT. Charles Stein demonstrated that the sample matrix $\mathbf{E}$ becomes increasingly inaccurate as the dimension of the system, N, grows. This phenomenon, known as "Stein's Paradox" challenges the intuitive belief that combining multiple sources of information will always enhance the accuracy of an estimate. Stein's Paradox highlights that in certain scenarios, the combined estimator can exhibit a larger mean squared error (MSE) than separate estimators that solely rely on the original data [58].

Stein's Paradox arises due to the interplay between estimators and the correlation structure of the data. In the context of RMT, it led to the recognition that the empirical covariance matrix obtained solely from the data may not be the optimal estimator when estimating multiple parameters. Instead, alternative estimators, such as shrinkage estimators, have gained prominence in RMT for estimating the population covariance matrix and its associated properties [40]. These estimators strike a balance between bias and variance, resulting in more accurate and robust estimates. The celebrated "linear shrinkage estimator" proposed in the 1980s takes the form:

$$\Xi = \alpha_s \mathbf{E} + (1-\alpha_s)\,\mathbf{I}_N \tag{1.3}$$

Here, Ξ represents an estimator of $\mathbf{C}$, and $\alpha_s \in (0,1)$ denotes the shrinkage intensity parameter. The shrinkage estimator presented in Equation 1.3 interpolates between extreme shrinkage towards the null hypothesis $\mathbb{I}_N$ (when $\alpha_s = 0$) and no shrinkage, using the empirical "raw" matrix $\mathbf{E}$ (when $\alpha_s = 1$).

In summary, Stein's Paradox illustrates that a combined estimator, which incorporates additional information beyond the data itself, exhibits improved performance as the dimension of the system grows. The efficacy of employing the simple estimator (1.3), rather than relying solely on the sample covariance matrix $\mathbf{E}$, was eventually quantified by Ledoit in 2004 for the $N \to \infty$ regime [40].

In the LDL (Largest Dimension Limit) the seminal work of Marčenko and Pastur [43] emerged. The **Marchenko-Pastur distribution** characterizes the behavior of eigenvalues in the limit of large-dimensional random matrices with independent and identically distributed elements. It provides a probability density function that describes the density of eigenvalues and offers valuable insights into their statistical properties, such as the concentration around a central region. This result comes from the fact that they obtained a self-consistent equation for the spectrum of E given C as N goes to infinity. Notably, this equation encapsulates the impact of the quality ratio $q = \frac{N}{T}$, which appears here precisely. Indeed was already known by the high dimensional statistics community, thanks to the work of Anderson in the 1963 [4], that in the classical limit $T \to \infty$ and N fixed, the sample eigenvalues converge to the population eigenvalues. A result that aligns with Marčenko-Pastur distribution for $q = 0$. Nevertheless, the interesting result comes when we consider $q = \mathcal{O}(1)$, i.e. when the ratio is fixed. The result shows that, no matter how large T, the sample eigenvalues serve as noisy estimators of the true population ones. Specifically, as q increases, the distortion between the spectrum of matrix E and its "true" counterpart becomes progressively more pronounced. This result is also called the *curse of dimensionality*. While each individual coefficient of the covariance matrix $\mathbf{C}$ can be estimated with minimal error when the sample size T is sufficiently large (under the assumption of temporal stationarity), the situation changes when both N and T assume large values, a common occurrence in various scenarios. In such cases, the sample estimator $\mathbf{E}$ becomes "inadmissible". More specifically, the substantial number of simultaneous noisy variables introduces substantial systematic errors during the computation of matrix eigenvalues.

The Marčenko-Pastur theorem has had a momentous impact on our understanding of the "curse of dimensionality." A significant development emerged in 1995 [57], when it was established that this theorem exhibits a remarkable degree of **universality** as $N \to \infty$ and $q = \mathcal{O}(1)$. This universality can be likened to the renowned Wigner semi-circle law, wherein the Marčenko-Pastur equation demonstrates robust validity across an extensive range of random measurement processes and general population covariance matrices $\mathbf{C}$ [57, 65].

Drawing from financial data sets known for their non-Gaussian nature [16], in the context of sample covariance matrices, lots of empirical evidence have emerged [38, 53]. For what concerns the eigenvalues of such covariance matrices, has been noticed that the majority of them conform, at an initial level, to the null hypothesis $\mathbf{C} = \mathbf{I}$, while a finite number of outliers, referred to as "spikes," reside outside the bulk. This observation serves as the foundation for the *spiked covariance matrix model*, which owes its name to the seminal work by Johnstone in 2001 [36]. This model sheds light on another universal property in RMT, the **Tracy-Widom distribution** [60], that governs the top bulk eigenvalues in the spiked covariance matrix. This finding underscores the remarkable rigidity of the eigenvalue edge within the bulk, exhibiting fluctuations on the order of $T^{-2/3}$ [36, 60]. Consequently, a simple yet effective methodology, known as "eigenvalue clipping" [53], emerges to distinguish meaningful eigenvalues (beyond the edge) from their noisy counterparts (within the bulk). The approach involves treating eigenvalues residing within the Marčenko-Pastur spectrum's bulk as noise, replacing them with a constant value, while leaving the principal components outside the bulk (the spikes) intact. Remarkably, this straightforward approach demonstrates robust performance in out-of-sample scenarios [17], underscoring the crucial role of regularization and cleaning techniques in such settings.

1.1.3. Marcenko-Pastur distribution

A first example of universality in RMT is that the asymptotic behavior of random matrices frequently becomes independent of the specific probability distribution of their entries. In addition, the resulting limiting distribution generally exhibits non-zero values exclusively within a confined interval, displaying distinct and well-defined boundaries.

For example, Wigner's semicircle law for the distribution of a symmetric or Hermitian matrix with i.i.d. entries is universal since the limiting distribution converges to the same density regardless of the underlying distribution of the matrix entries [7].

The presence of sharp edges in the eigenvalue distribution of random matrices holds significant importance for practical applications, particularly in the context of extracting signals from noise. However, when dealing with finite-sized matrices, adapting asymptotic results originally derived for infinite-sized matrices becomes challenging. Nonetheless, an eigenvalue that deviates significantly from the asymptotic range serves as a reliable indicator of non-random behavior. In contrast, when the asymptotic distribution lacks compact support, employing similar heuristics necessitates a more comprehensive understanding of the convergence rate.

While there has been a growing interest in studying the eigenvectors [39] of random matrices in recent times, a majority of the established results primarily focus on the spectra or eigenvalue distributions of such matrices, encompassing both the global regime, which involves statistical analysis of the entire set of eigenvalues, and the local regime, which is concerned with the spacing between individual eigenvalues.

When dealing with financial data sets, it is often useful to distinguish the eigenvalues that lie within the spectrum of $\rho_{\mathbf{M}}$, referred to as the **bulk** of the eigenvalues from those that are well separated from it that we will call **outliers** or **spikes**.

In this section, we are interested in exploring the limiting distribution of the spectra of random matrices $\mathbf{M}$ belonging to the Wishart ensemble, i.e. we want to provide an elegant proof of how to find the celebrated Marčenko-Pastur (MP) law for the distribution of the eigenvalues of a Wishart matrix, i.e. the sample covariance matrix, as defined in multivariate statistics.

This proof has been published by Joel Bun et al. in their paper about Cleaning large Correlation Matrices [21]. Even though the organization of the proof could be slightly different the idea is the same. We will first introduce the main (mathematical) tools needed to tackle this proof.

It's time to formalize the problem. Let us consider the $N \times T$ matrix $\mathbf{Y}$ consisting of T independent realizations of random centered Gaussian vectors of size N and covariance $\mathbf{C}$, then the Wishart matrix is defined as the $N \times N$ matrix $\mathbf{M}$ as $\mathbf{M} = T^{-1}\mathbf{Y}\mathbf{Y}^*$.

Note that the symmetry of the considered matrices ensures that the eigenvalues of $\mathbf{M}$ are defined on the real line.

Definition 1. *A matrix $\boldsymbol{M}$ is said to be rotationally invariant if the probability is invariant under the transformation $\boldsymbol{M} \to \boldsymbol{\Omega M \Omega}^\dagger$ for any matrix $\boldsymbol{\Omega}$ belonging to the Orthogonal group $\mathbf{O}(N)$, i.e. $\mathcal{P}_\beta(\boldsymbol{M}) = \mathcal{P}_\beta(\boldsymbol{\Omega M \Omega}^\dagger)$, $\forall \boldsymbol{\Omega} \in \mathbf{O}(N)$.*

Where $\mathcal{P}_\beta(\mathbf{M})$ is a certain probability measure and β is the Dyson's threefold way index that specifies the symmetry properties of the ensemble ($\beta = 1$ for Orthogonal, $\beta = 2$ for Unitary and $\beta = 4$ for Symplectic ensembles). Linking to the concept of universality, a property is said to be *universal* if it does not depend on the specific probability measure $\mathcal{P}_\beta(\mathbf{M})$.

A typical example of invariant measure in the physics literature is that $\mathcal{P}_\beta(\mathbf{M})$ is of the form of a Boltzmann distribution:

$$\mathcal{P}_\beta(\mathbf{M})\mathcal{D}\mathbf{M} \propto e^{-\frac{\beta N}{2}\mathrm{Tr}V(\mathbf{M})}\mathcal{D}\mathbf{M} \tag{1.4}$$

with V the so called *potential* function and $\mathcal{D}\mathbf{M} = \prod_{i=1}^N d\mathbf{M}_{ii} \prod_{i<j}^N d\mathbf{M}_{ij}$ denotes the (Lebesgue) flat measure. On the other hand, the distribution (1.4) can be rewritten in terms of the eigenvalues and eigenvectors of $\mathbf{M}$ as:

$$\mathcal{P}_\beta(\mathbf{M})\mathcal{D}\mathbf{M} \propto e^{-\frac{\beta N}{2}\sum_{i=1}^N V(\nu_i)} \prod_{i<j}^N |\nu_i - \nu_j|^\beta \Big(\prod_{i=1}^N \mathrm{d}\nu_i\Big)\Big(\mathrm{d}\Omega\Big), \tag{1.5}$$

where the Vandermonde determinant ($\prod|\nu_i - \nu_j|^{\beta}$)comes from the change of variables (from the $\mathbf{M}_{ij}$ to the ν_i and Ω_{ij}).

The distribution of the eigenvalues $\{\nu_i\} : i = \{1, \ldots, N\}\}$ can be characterized through the *Empirical Spectral Distribution* (ESD) (also known as the "Eigenvalue Distribution"):

$$\rho^N_{\mathbf{M}}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - \nu_i) \tag{1.6}$$

with δ the Dirac delta function.

On the other hand, we define as the *Limiting Spectral Density*, if there exists, the *deterministic* limit $\rho^N_{\mathbf{M}} \to \rho_{\mathbf{M}}$ as $N \to \infty$.

Two of the most fundamental features of large random matrices are that:

- One expects the ESD to converge (almost surely in many cases) to a unique and *deterministic* limit $\rho^N_{\mathbf{M}} \to \rho_{\mathbf{M}}$ as $N \to \infty$.
- The predicted *self-averaging* (sometimes call *ergodicity* or *concentration*) property of the LSD: when the dimension N becomes very large, a single sample of $\mathbf{M}$ spans the whole eigenvalue density function, independently of the specific realization of $\mathbf{M}$. The consequence of this self-averaging property is that we can replace the computation of the ESD (1.6) for a specific $\mathbf{M}$ by the average according to the probability measure of $\mathbf{M}$ (e.g. over the measure (1.4)):

$$\rho_{\mathbf{M}}(x) = \lim_{N\to\infty}\rho^N_{\mathbf{M}}(x), \quad \text{with} \quad \rho^N_{\mathbf{M}}(x) = \left\langle \frac{1}{N}\sum_{i=1}^{N}\delta(x - \nu_i)\right\rangle_{\mathbf{M}}. \tag{1.7}$$

Along this section and henceforth we will adopt the following convention. We shall denote by $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_N$ the eigenvalues of $\mathbf{M}$ with associated eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$; which can be reindexed, by convenience, as $\mathbf{w}_i \equiv \mathbf{w}_{\nu_i}$ for any integer $1 \leq i \leq N$.

In order to prove the Marcenko-Pastur theorem it's necessary to introduce several transforms that often appear in RMT literature.

Definition 2. *The* ***resolvent*** *of* $\boldsymbol{M}$ *is defined as*[1]

$$\mathbf{G}_{\boldsymbol{M}}(z) := (z\mathbf{I}_N - \boldsymbol{M})^{-1}, \tag{1.8}$$

[1] Note that in the mathematical literature, it differs by a minus sign.

with $z := x - \mathrm{i}\eta \in \mathbb{C}^-$, where $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}$. It can be shown that:

- It is a continuous function of z and is easy to differentiate (we will use it to not work directly on the ESD).
- It contains the complete information about the eigenvalues $\{\nu_i\}$ *and* the eigenvectors $\{\mathbf{w}_i\}$ since it can be rewritten as:

$$\mathbf{G}_{\mathbf{M}}(z) = \sum_{i=1}^{N} \frac{\mathbf{w}_i \mathbf{w}_i^*}{z - \nu_i}. \tag{1.9}$$

It is easy to see that the number of singularities of the resolvent is equal to the number of eigenvalues of $\mathbf{M}$.

We define now the normalized trace of Eq. (1.8) as

$$\mathfrak{g}_{\mathbf{M}}^{N}(z) := \frac{1}{N} \mathrm{Tr}\left[\mathbf{G}_{\mathbf{M}}(z)\right], \tag{1.10}$$

We shall skip the index $_{\mathbf{M}}$ as soon as there is no confusion about the matrix we are dealing with.

Definition 3. *In the limit of large dimension, one can define the* ***Stieltjes*** *(or* Cauchy*)* ***transform*** *of ρ as:*

$$\mathfrak{g}^{N}(z) \underset{N\to\infty}{\sim} \mathfrak{g}(z), \qquad \mathfrak{g}(z) := \int \frac{\rho(u)}{z - u} \mathrm{d}u. \tag{1.11}$$

Let's show now a few interesting properties of the Stieltjes transform. For instance, if the density function ρ does not contain Dirac masses, then this is the unique solution of the so-called *Riemann-Hilbert* problem, i.e :

1. $\mathfrak{g}(z)$ is analytic in $\mathbb{C}^+$ except on its branch cut on the real axis inside $\mathrm{supp}[\rho_{\mathbf{M}}]$;
2. $\lim_{|z|\to\infty} z\mathfrak{g}(z) = 1$;
3. $\mathfrak{g}(z)$ is real for $z \in \mathbb{R}\setminus \mathrm{supp}[\rho_{\mathbf{M}}]$;
4. When near the branch cut, two different values for $\mathfrak{g}(z)$ are possible, depending on whether the cut is approached from above or from below, i.e.:

$$\lim_{\eta\to 0^+} \mathfrak{g}(x \pm \mathrm{i}\eta) = \mathfrak{h}(x) \mp i\pi\rho(x), \qquad x \in \mathrm{supp}[\rho] \text{ and } \rho(x) \in \mathbb{R}^+, \tag{1.12}$$

where the function $\mathfrak{h}$ denotes the *Hilbert* transform of ρ defined by

$$\mathfrak{h}(x) := \fint_{\mathrm{supp}[\rho]} \frac{\rho(u)}{x-u} \mathrm{d}u \tag{1.13}$$

with $\fint$ denoting Cauchy's principal value.

It is now immediate to see that if one knows $\mathfrak{g}(z)$ in the complex plane, the density ρ can be retrieved by inverting the last property of the Riemann-Hilbert problem:

$$\rho(x) \equiv \frac{1}{\pi} \lim_{\eta \to 0^+} \mathrm{Im}(\mathfrak{g}(x - \mathrm{i}\eta)), \qquad x \in \mathrm{supp}[\rho]. \tag{1.14}$$

The continuous limit of $\mathfrak{g}(z)$ in the large N limit thus allows to investigate the distribution of the eigenvalues that lie in the bulk component.
The main question now is, how to compute the limiting value of the Stieltjes transform? There exist several techniques to compute it, for example: (i) Coulomb gas methods, (ii) Feynman diagrammatic expansion [20], (iii) method of moments [2], (iv) Replicas, (v) Dyson's Brownian motion.

In the next insert, we will briefly explain the main idea behind the **Coulomb Gas Method**. For the sake of rigor we provide a more rigorous and self-contained reference by Mehta [44].

Coulomb Gas Analogy Here we introduce the Coulomb gas analogy, a powerful conceptual framework. In this analogy, the eigenvalues of matrix $\mathbf{M}$ are metaphorically treated as the spatial positions of fictitious charged particles. These particles interact via a two-dimensional Coulomb potential, characterized by its logarithmic nature.

Within this section, our primary objective is to establish a link between the potential function and the Stieltjes transform $\mathfrak{g}(z)$. This linkage arises specifically when the probability measure governing the matrix ensemble exhibits rotational invariance, conforming to the form specified in Equation (1.4).

We commence by expressing the partition function of the model, denoted as $\mathcal{Z}$, in terms of the model's potential function, $V(\mathbf{M})$, as derived from Equation (1.4):

$$\mathcal{Z} \propto \int e^{-\frac{\beta N}{2} \mathrm{Tr} V(\mathbf{M})} \mathcal{D}\mathbf{M}.$$

To obtain the Limiting Spectral Distribution (LSD), or rather its Stieltjes transform, we employ a saddle point method [20].

We express the partition function in terms of the eigenvalues, denoted as ν_i, and eigenvectors of matrix $\mathbf{M}$, utilizing Equation (1.5):

$$\mathcal{Z} \propto \int \Big(\prod_{i=1}^{N} \mathrm{d}\nu_i\Big) \exp(-N \sum_{i=1}^{N} \left[V(\nu_i) - \frac{\beta}{2N} \sum_{i \neq j} \log|\nu_i - \nu_j| \right]).$$

Here, we introduce the *action*, denoted as $S(\nu_i)$, which allows us to rewrite the partition function as follows:

$$\mathcal{Z} \propto \int \prod_{i=1}^{N} \mathrm{d}\nu_i \exp(-N^2 S(\nu_i)), \tag{1.15}$$

with the action defined as:

$$S(\nu_i) = \frac{1}{N} \sum_{i=1}^{N} V(\nu_i) - \frac{\beta}{2N^2} \sum_{i \neq j} \log|\nu_i - \nu_j|.$$

It is worth noting that the action is normalized so that its large N limit is of order 1. This formulation treats the eigenvalues as a thermal gas of one-dimensional particles subject to an external potential $V(z)$ and a logarithmic repulsive interaction. This interpretation forms the basis of the Coulomb gas analogy. In thermal equilibrium, the eigenvalues tend to cluster around potential wells, but the repulsive force prevents them from accumulating near the minimum, maintaining a characteristic spacing of order $\mathcal{O}(N^{-1})$. As depicted in Figure 1.1, for instance, in the case of a quadratic potential $V(x) = x^2/2$, all the particles tend to concentrate around zero.

In the large N limit, we compute the integral over the eigenvalues using the saddle-point method, leading to the following condition of "force equilibrium":

$$V'(\nu_i) = \frac{\beta}{N} \sum_{j=1; j \neq i}^{N} \frac{1}{\nu_i - \nu_j}, \quad \forall\, i = 1, \ldots, N. \tag{1.16}$$

While it may seem daunting to find the eigenvalues λ_i that satisfy these N equations, we can expect to determine the Limiting Spectral Distribution (LSD), denoted as $\rho_{\mathbf{M}}$, in the limit as $N \to \infty$. The LSD corresponds to the eigenvalue configuration that satisfies these saddle-point equations. Under the assumption of a single-cut density, the result, as established in [20], can be expressed as:

$$\mathfrak{g}(z) = V'(z) - Q(z)\sqrt{(z-\nu_+)}\sqrt{(z-\nu_-)}, \tag{1.17}$$

where $\nu_- < \nu_+$ denote the edges of supp$[\rho]$ (the support of the density) and $Q(z)$ is a Laurent polynomial with degree $d-1$ and order ℓ. Here, d corresponds to the degree of $V'(z)$, which is the degree of the polynomial $P(z)$ obtained by rewriting $V'(z) = z^{-\ell}P(z)$. If $V'(z)$ is a polynomial, then $\ell = 0$.

In the determination of the LSD, we are faced with $d+1$ unknowns to be determined: the coefficients of $Q(z)$, ν_-, and ν_+. The detailed procedure on how to obtain these unknowns is shown in [21].

It is important to note that the characterization of the potential function $V(z)$ governing the entries of $\mathbf{M}$ allows us to determine the corresponding LSD $\rho_{\mathbf{M}}$. Throughout the rest of this section, we demonstrate how this Coulomb gas analogy facilitates the retrieval of the Marchenko-Pastur distribution.

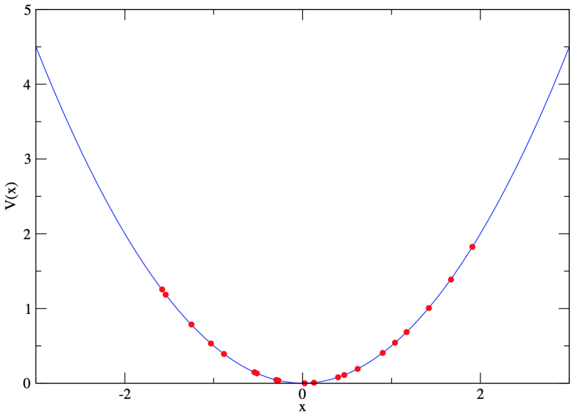


Figure 1.1: Typical configuration of a repulsive Coulomb gas with $N = 20$ particles (red dots) in the potential $V(x) = x^2/2$ as a function of x.

Let's finally dive into the proof of the Marchenko-Pastur law. Given the sample covariance matrix $\mathbf{M}$, for *any* N and $T > N$, Wishart derived the exact PDF of the entries $\mathbf{M}$ which reads:

$$\mathcal{P}_{\mathrm{w}}(\mathbf{M}|\mathbf{C}) = \frac{1}{2^{NT/2}\Gamma_N(T/2)} \frac{\det(\mathbf{M})^{\frac{T-N-1}{2}}}{\det(\mathbf{C})^{T/2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{C}^{-1}\mathbf{M}}. \tag{1.18}$$

We say that $\mathbf{M}$ (given $\mathbf{C}$) follows a Wishart$(N, T, \mathbf{C}/T)$ distribution. If we focus on the "isotropic" case, i.e., when $\mathbf{C} = \mathbf{I}_N$, we can derive from (1.18)

$$\mathcal{P}_{\mathrm{w}}(\mathbf{M}|\mathbf{I}_N) \propto \det(\mathbf{M})^{\frac{T-N-1}{2}} e^{-\frac{T}{2}\mathrm{Tr}\mathbf{M}} := e^{-\frac{T}{2}\mathrm{Tr}\mathbf{M} + \frac{T-N-1}{2}\mathrm{Tr}\log\mathbf{M}}, \tag{1.19}$$

which clearly belongs to the class of Boltzmann ensembles (1.4). Throughout the following, we shall denote by $\boldsymbol{\mathcal{W}}$ the $N \times N$ matrix whose distribution is given by (1.19). The corresponding potential function is given by (sub-leading terms have been ignored):

$$V(z) = \frac{1}{2q}\left[z - (1-q)\log z\right], \qquad \text{with} \qquad q := N/T. \tag{1.20}$$

Deriving we obtain a Laurent polynomial in z as we have

$$V'(z) = \frac{1}{2qz}\left[z - (1-q)\right].$$

Following our convention, $V'(z)$ is a Laurent polynomial of degree 1 and order $\ell = -1$. We provide the computation of the Stieltjes transform $\mathfrak{g}(z)$ in the appendix A. The final result reads:

$$\mathfrak{g}(z) = \frac{(z+q-1) - \sqrt{z-\nu_-}\sqrt{z-\nu_+}}{2qz}, \qquad \nu_\pm \ := \ (1 \pm \sqrt{q})^2, \tag{1.21}$$

and this is the solution found by Marčenko and Pastur in [43] in the special case $\mathbf{C} = \mathbf{I}_N$. Using the inversion formula (1.14) we find the celebrated Marčenko-Pastur (MP) law (for $q \in (0,1)$)

$$\rho_{\mathrm{MP}}(\nu) = \frac{\sqrt{4\nu q - (\nu+q-1)^2}}{2q\pi\nu}, \qquad \forall\, \nu \in \left[\nu_-, \nu_+\right]. \tag{1.22}$$

When $q \geqslant 1$, it is easy to see that $\mathbf{M}$ has $N-T$ zero eigenvalues that contribute $(1-q)\delta_0$ to the density Eq. (1.23). We can show that the convergence of the ESD towards the asymptotic MP law occurs, for $q < 1$, as $N^{-2/5}$ in the present case where the random elements of $\mathbf{Y}$ are Gaussian (for a full discussion of this issue, see [9]).

It's time to wrap up the result obtained.

Theorem 1.1. *Given a Wishart matrix* $\mathbf{M}$, *the Marčenko-Pastur law states that, as* $N \to \infty$, *for* $q \in (0,1)$, *the Empirical Spectral Distribution (ESD) converges to the following Limiting Spectral Density (LSD) supported in* $\left[\nu_-, \nu_+\right]$*:*

$$\rho_{MP}(\nu) = \frac{\sqrt{4\nu q - (\nu+q-1)^2}}{2q\pi\nu}, \qquad \forall\, \nu \in \left[\nu_-, \nu_+\right], \qquad \nu_\pm \ := \ (1 \pm \sqrt{q})^2. \tag{1.23}$$

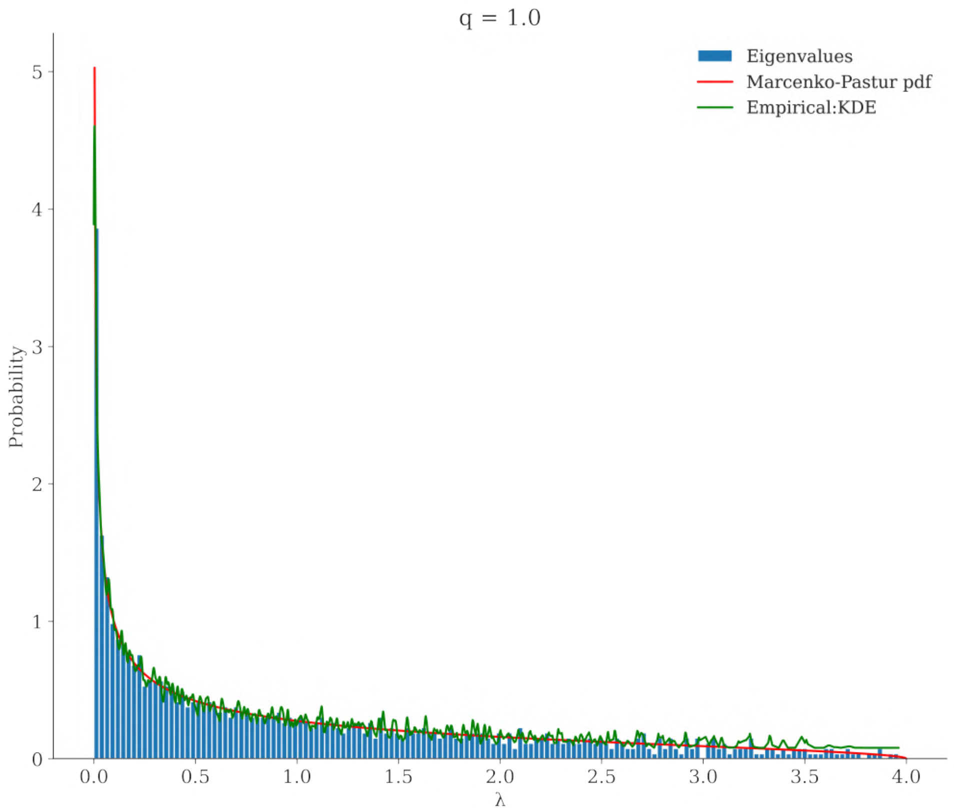


Figure 1.2: Simulation of the Marcenko-Pastur law with q = 1 for 1000 × 1000 matrices. The red curve corresponds to theoretical Marčenko-Pastur density (1.23), while the blue area corresponds to the real eigenvalues density

1.1.4. Real Sample Covariance Matrix and its Spectrum

Once we have formally introduced the Marčenko-Pastur law we can safely dive into its applications to real data sets. This law, as alluded to in its introduction, is a fundamental and well-established tool to study large sample covariances matrices. We have studied in detail the special case in which the null hypothesis $\mathbf{C} = \mathbf{I}_N$ (isotropic case). In literature, there are several studies in which we allow the population correlation matrix $\mathbf{C}$ to be *anisotropic*, i.e. not proportional to the identity matrix. Unfortunately, the final result is not as simple as Eq. (1.21) but many properties can be inferred from it [45].

This tool is essential to study the statistics of the spectrum of real sample covariance matrices. As mentioned earlier, the theoretical predictions for the eigenvalue spectra of random matrices have practical significance due to their universality with respect to the distribution of the underlying random variables. One notable feature is the presence of sharp edges in the spectrum, which, in the case of the existence of eigenvalues lying outside the expected region, is a departure against a simple “null hypothesis” benchmark.

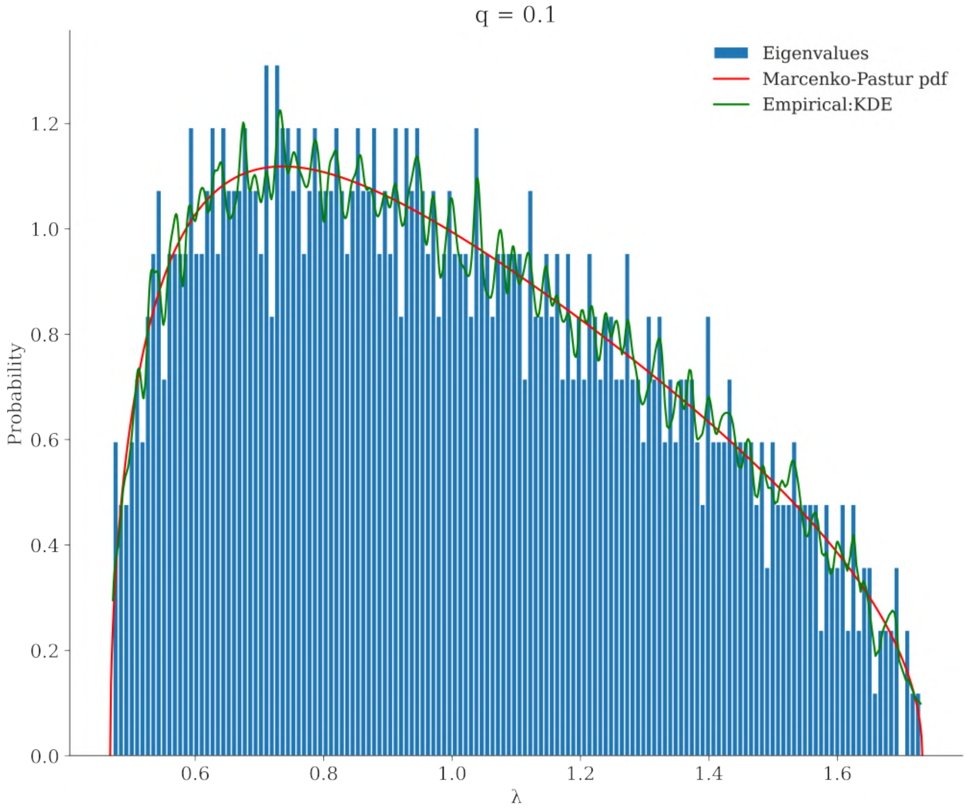


Figure 1.3: Simulation of the Marcenko-Pastur law with q = 0.1 for 1000 × 1000 matrices. The red curve corresponds to theoretical Marčenko-Pastur density (1.23), while the blue area corresponds to the real eigenvalues density

Figure 1.4 illustrates this point by comparing the empirical spectral density of a financial correlation matrix (corresponding to $N = 406$ and $T = 1300$, yielding $q \approx 0.31$) to the simplest Marčenko-Pastur spectrum in the null hypothesis case of $\mathbf{C} = \mathbf{I}_N$. While the bulk of the distribution is approximately accounted for, there appear to be a finite number of eigenvalues lying outside the Marčenko-Pastur sea, which can be referred to as outliers or spikes.

If there's one single take-home message from this thesis is that:

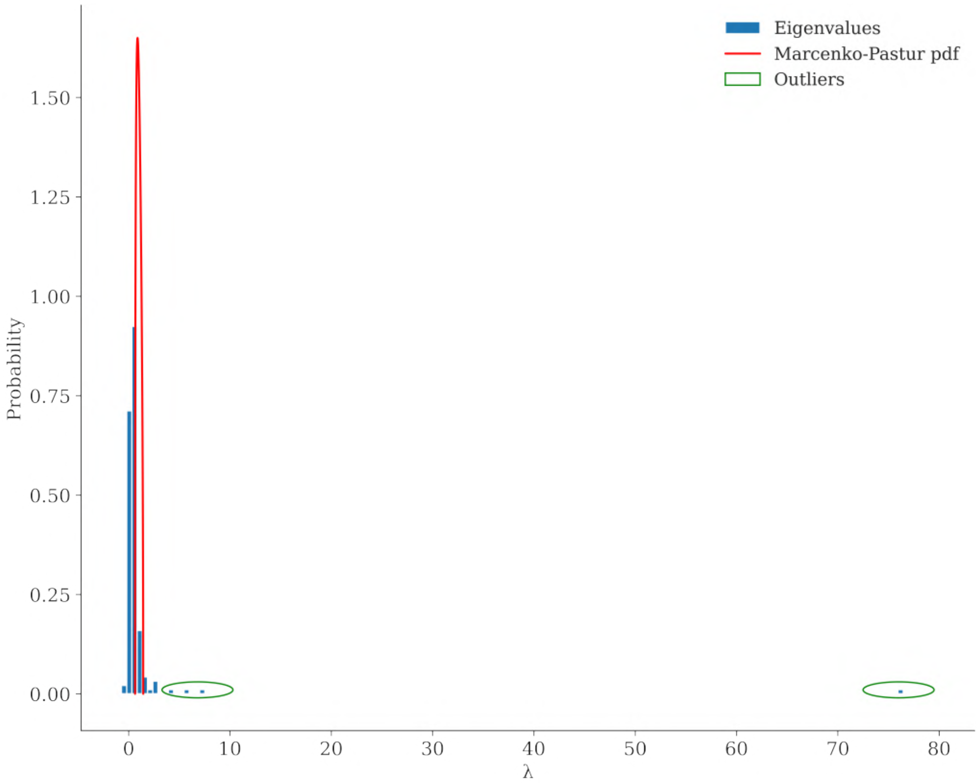


Figure 1.4: Test of the null hypothesis on the empirical correlation matrix **E** using global stocks' data with $N = 183$ and $T = 4525$ and $q \approx 0.04$.

In certain cases, it is possible for a finite number of eigenvalues to genuinely reside outside the Marčenko-Pastur sea, or more generally, outside the bulk region, even in the limit as N tends to infinity. This phenomenon challenges the null hypothesis that all eigenvalues should lie within the expected range dictated by the Marčenko-Pastur distribution.

Remarkably, the empirical data presented in Figure 1.4 provides compelling evidence for the existence of these **true outliers**. These outliers possess significant financial interpretation, as they can be interpreted in terms of specific economic sectors of activity. This suggests that the eigenvalues residing outside the expected range carry **meaningful information** about the underlying system under investigation. This makes the Marchenko-Pastur law an essential tool to distinguish signal from noise.

Understanding and characterizing these outliers are crucial for obtaining a comprehensive understanding of the statistical properties of eigenvalues in random matrices. Such deviations from the expected behavior for **financial data sets** have been extensively studied in this thesis, leading to valuable insights into the underlying dynamics.

Most of the statistical information of these systems lies in the outliers.

However, even if there are no such spikes in the spectrum of $\mathbf{C}$, for finite N, one can still expect to observe some eigenvalues beyond the upper edge of the Marčenko-Pastur distribution for finite size effects. But ... how likely is it to find outliers with finite N and T due to finite-size effects? Can we quantify this?

Finite Size Effects and the Tracy-Widom Region

The existence of sharp edges that separate a region where non-zero eigenvalue density and a region where no eigenvalues should be found, is a characteristic observed in the asymptotic limit of infinite matrix size $(N, T \to \infty)$ and in the absence of heavy-tailed distributions in the matrix elements (see Arous et al. [8]). However, for large but finite matrix sizes, the probability of finding eigenvalues beyond the Marčenko-Pastur sea remains very small but finite. Lots of investigations have been made on the width of the transition region and the tail behavior of the density of states, with notable contributions by Bowick et al. [19] and Tracy and Widom [60].

The Tracy-Widom distribution, another manifestation of universality, describes the fluctuations of macroscopic observables in numerous large-dimensional systems. Its derivation primarily relies on orthogonal polynomials, which fall beyond the scope of this thesis (see Tracy and Widom [60], Nadal et al. [47]). Precisely characterizing the distance between the largest eigenvalue λ_1 of $\mathbf{E}$ and the upper edge of the spectrum denoted by λ_+, the Tracy-Widom result can be formally stated as follows: the rescaled distribution of $\lambda_1 - \lambda_+$ converges to the Tracy-Widom distribution, typically denoted as F_1:

$$\mathcal{P}\left(\lambda_1 \leqslant \lambda_+ + \gamma N^{-2/3} u\right) = F_1(u), \tag{1.24}$$

where γ is a constant dependent on the specific problem. For the isotropic Marčenko-Pastur problem, $\lambda_+ = (1+\sqrt{q})^2$ and $\gamma = \sqrt{q}\lambda_+^{2/3}$.

The Tracy-Widom density $f_1(u) = F_1'(u)$ is well-characterized, particularly in terms of its left and right tails:

$$\ln f_1(u) \propto -u^{3/2} \quad \text{as } u \to +\infty, \quad \text{and} \quad \ln f_1(u) \propto -|u|^3 \quad \text{as } u \to -\infty. \tag{1.25}$$

It is noteworthy that the left tail is much thinner.

Regarding the distribution of the smallest eigenvalue $\lambda_{\min}$ around the lower edge λ_-, it also follows the Tracy-Widom distribution, except for the specific case of Marčenko-Pastur matrices with $q = 1$. In this case, $\lambda_- = 0$, constituting a "hard" edge since all eigenvalues of the empirical matrix must be non-negative. The treatment of this particular case can be found in Peche [51].

2 | Data

In this section, we provide a comprehensive description of the **Compustat dataset**, which serves as the primary source of financial data for our study. Compustat is a widely recognized and extensively used financial database that provides comprehensive coverage of public companies in the United States and other regions. It offers a rich set of financial and non-financial variables, enabling researchers to explore various aspects of corporate performance and behavior. In this thesis, we utilize the Compustat dataset to investigate relations between a firm's prices and sales.

2.1. Data Collection and Coverage

The Compustat dataset is maintained by Standard & Poor's (S&P) Global Market Intelligence, a leading provider of financial data and analytics. It aggregates information from various sources, including regulatory filings (e.g., 10-K and 10-Q reports), company press releases, and other public disclosures. The data undergoes rigorous quality control procedures to ensure accuracy and consistency.

The dataset includes a vast range of variables, capturing financial statements, market data, executive compensation, corporate governance information, and more. The financial statements cover key aspects such as income statements, balance sheets, and cash flow statements, providing a comprehensive view of a company's financial performance. Market data includes stock prices, trading volumes, and other market-related variables, enabling researchers to examine stock market dynamics and investor behavior.

The coverage of the Compustat dataset extends beyond U.S. companies, incorporating global firms operating in various industries. Additionally, the dataset covers companies of different sizes, from small-cap to large-cap.

Despite its extensive coverage, the Compustat dataset may have certain limitations. These limitations could arise from data reporting errors, missing values, or inconsistencies in data formats across companies and periods. It is common to perform data preprocessing steps to address these issues, including data cleaning, imputation of missing values, and standardization of variables.

2.2. Description of the Dataset

The raw Dataset consists of data from 26349 companies identified by the GVKEY in the timeframe that goes from 1961-06-30 to 2023-03-31 for 248 quarters.

In Table 2.1 we can see the meaning of each of the variables that appear in the Dataset.

In our study we focused our attention on two main variables of our financial data.

- **Sales**: The revenue generated by the company in the fiscal quarter. This imposes a "lower bound" in the choice of the time interval. As we will see in the following sections, even though the timeframe in which the data are available is quite large, we are still interested in capturing fine-grained information at high frequency. What if there is interesting information in sales at higher frequencies? As we will see in Chapter 3 where we will attempt to compare the eigenstructure at different frequencies, downsampling from daily to quarterly prices still results in the recovery of the same structure.

- **Prices**: Closing price of the company's stock in the fiscal quarter in the CompuStat Dataset. We recovered from the GVKEY the same companies in the well-known **Yahoo Dataset** in order to study the Adjusted Close Price for different frequencies (daily, weekly etc.)

To identify the sector of each firm we used the **SIC (Standard Industrial Classification)** code. It is a four-digit numerical code that categorizes the industries that companies belong to based on their business activities. SIC codes have a hierarchical, top-down structure. The first two digits represent the highest level business classification, while the subsequent two digits are used to further refine the identification.

The SIC system provides a hierarchical structure that allows for the classification of industries based on their level of downstream linkages. Downstream linkages refer to the flow of goods and services from one industry to another, where the output of one industry becomes an input for another. By organizing industries according to their primary activities, the SIC system helps identify the industries that are directly involved in the production and distribution of finished goods and services.

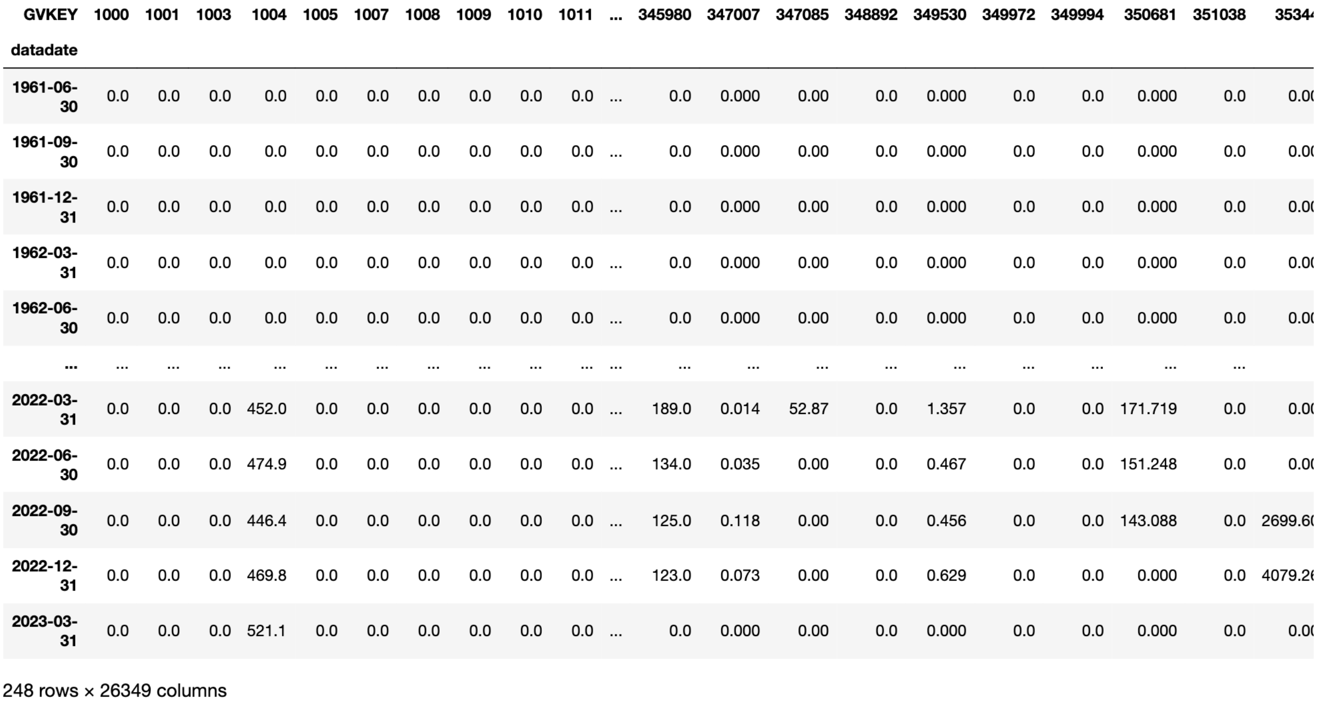
At the top of the downstream chain are industries classified under broad categories such as manufacturing, construction, and transportation. These sectors are responsible for producing final goods and delivering them to end consumers. As we move further downstream in the SIC system, industries become more specialized and focus on specific activities related to the production, assembly, or distribution of components or finished products.

2.3. Data preprocessing

During the data cleaning process, we encountered several issues that required specific treatments. In order to ensure the quality and consistency of the data, we adopted the following preprocessing conventions:

- Exclusion of companies with less than 160 quarters of activity: To focus on companies with a significant presence in the market during the studied period, we decided to exclude those that were active for less than 160 quarters.
- Data merging: To avoid duplication and ensure uniformity, we employed a merging strategy. Sales values within a quarter were aggregated by summing them, while duplicate prices were resolved by selecting the first value.

After applying these preprocessing steps, the dataset consists of 768 firms. The final DataFrame we are working with has the following structure:



GVKEY	1000	1001	1003	1004	1005	1007	1008	1009	1010	1011	...	345980	347007	347085	348892	349530	349972	349994	350681	351038	3534
datadate																					
1961-06-30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000	0.00	0.0	0.000	0.0	0.0	0.000	0.0	0.0
1961-09-30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000	0.00	0.0	0.000	0.0	0.0	0.000	0.0	0.0
1961-12-31	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000	0.00	0.0	0.000	0.0	0.0	0.000	0.0	0.0
1962-03-31	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000	0.00	0.0	0.000	0.0	0.0	0.000	0.0	0.0
1962-06-30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000	0.00	0.0	0.000	0.0	0.0	0.000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2022-03-31	0.0	0.0	0.0	452.0	0.0	0.0	0.0	0.0	0.0	0.0	...	189.0	0.014	52.87	0.0	1.357	0.0	0.0	171.719	0.0	0.0
2022-06-30	0.0	0.0	0.0	474.9	0.0	0.0	0.0	0.0	0.0	0.0	...	134.0	0.035	0.00	0.0	0.467	0.0	0.0	151.248	0.0	0.0
2022-09-30	0.0	0.0	0.0	446.4	0.0	0.0	0.0	0.0	0.0	0.0	...	125.0	0.118	0.00	0.0	0.456	0.0	0.0	143.088	0.0	2699.6
2022-12-31	0.0	0.0	0.0	469.8	0.0	0.0	0.0	0.0	0.0	0.0	...	123.0	0.073	0.00	0.0	0.629	0.0	0.0	0.000	0.0	4079.2
2023-03-31	0.0	0.0	0.0	521.1	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000	0.00	0.0	0.000	0.0	0.0	0.000	0.0	0.0

248 rows × 26349 columns

Figure 2.1: Pandas DataFrame for the sales

Table 2.1: Main Variables in the Compustat Dataset

Variable	Description
GVKEY	Unique identifier for each company in the Compustat dataset.
datadate	The date when the financial data was recorded.
fyearq	The fiscal year when the financial data was recorded.
fqtr	The fiscal quarter when the financial data was recorded.
indfmt	Industry format code indicating the type of industry classification system used.
consol	Consolidation code indicating the level of consolidation of financial data.
popsrc	Population source code indicating the source of the data.
datafmt	Data format code indicating the format of the financial data.
tic	Ticker symbol representing the company's stock on a stock exchange.
cusip	Unique identifier for each security issued by a company.
conm	The official name of the company.
curcdq	The currency code used for financial data reporting.
datacqtr	The fiscal quarter and year when the financial data was recorded.
datafqtr	The fiscal quarter when the financial data was recorded.
invfgq, invoq, invrmq, invtq, invwipq	Various inventory-related fields for different inventory types.
saleq	Net sales or revenue generated by the company in the fiscal quarter.
costat	Current status of the company, such as active or inactive.
mkvaltq	Market value of the company's equity in the fiscal quarter.
prccq	Closing price of the company's stock in the fiscal quarter.
gsector	Global Industry Classification Standard (GICS) sector code.
naics	North American Industry Classification System (NAICS) code.
sic	Standard Industrial Classification (SIC) code.

3 | Statistics of log-returns and its correlations

The investigation of correlations among different stocks holds significance not only from a scientific standpoint, aiming to comprehend the economy as a complex dynamical system, but also from practical perspectives like asset allocation and portfolio-risk estimation [16, 26, 42]. Distinguishing it from numerous physical systems where correlations are linked to fundamental interactions between constituent parts, the stock market presents a unique challenge as the underlying "interactions" remain elusive. In this study, we employ the principles and techniques of random matrix theory (see Chapter 1 for an introduction), originally developed in the realm of quantum systems where the precise nature of inter-subunit interactions is unknown, to examine cross-correlations among stocks.

Computing correlations, in this case, allows us to overcome the limitations arising from the lack of explicit knowledge about the driving forces behind these interactions.

By treating the correlation matrix of stock returns as a random matrix, we can extract meaningful statistical properties of the companies' returns. This approach enables us to explore patterns that might not be apparent through traditional methods. By analogy with the behavior of quantum particles, which can exhibit collective behavior despite having no direct interactions, we seek to uncover emergent phenomena within the stock market that are driven by the intricate interplay of numerous factors, including market sentiment, investor behavior, and macroeconomic indicators among the others.

Furthermore, RMT provides robust statistical tools for analyzing large datasets, which is particularly relevant in the context of the stock market, where a vast number of stocks are traded daily. The conventional methods of correlation analysis may struggle to handle such high-dimensional data, leading to biased estimates and unreliable results including a lot of noise in the estimations.

To analyze correlations, we begin by calculating the logarithmic price change, or "return", of stock $i = 1, \ldots, N$ over a time scale Δt, using the equation:

$$r_i(t) = \log \frac{s_i(t+\Delta t)}{s_i(t)}, \tag{3.1}$$

Here, $S_i(t)$ represents the price of stock i at time t. Since the dynamics of the prices of financial assets are known to be non-stationary, one cannot simply use them to examine the relationship between different assets but it becomes necessary to use logarithmic returns, which are generally assumed to be weakly stationary [56].

However, since different stocks exhibit varying levels of volatility (standard deviation), we further define the centered and rescaled returns (or growth rates) as

$$\widetilde{r}_i(t) := \frac{r_i(t) - \mathbb{E}_{t'}\left[r_i\left(t'\right)\right]}{\sqrt{\mathbb{V}_{t' \neq t}\left[r_i\left(t'\right)\right]}} \tag{3.2}$$

In this equation, $\mathbb{V}_{t' \neq t}$ denotes the variance of r_i, calculated as $\langle r_i^2 \rangle - \langle r_i \rangle^2$, and $\mathbb{E}$ signifies the time average over the studied period. By computing the equal-time cross-correlation matrix, denoted as C, we obtain its elements:

$$C_{ij} = \mathbb{E}\left[\widetilde{r}_i(t)\widetilde{r}_j(t)\right] \tag{3.3}$$

That can also be written using $\langle \cdots \rangle$ as:

$$C_{ij} = \langle r_i(t) r_j(t) \rangle \tag{3.4}$$

The elements C_{ij} are constrained within the range of $-1 \leqslant C_{ij} \leqslant 1$. A value of $C_{ij} = 1$ indicates perfect positive correlations between the stocks, $C_{ij} = -1$ signifies perfect negative correlations and $C_{ij} = 0$ corresponds to uncorrelated pairs of stocks.

3.1. Principal Component Analysis of the Correlation Matrix

Given the correlation matrix C, we can study it under the lens of RMT. The idea is to do the eigendecomposition of C, to study its eigenvalues and eigenvectors and to compare them to those of a randomly generated cross-correlation matrix.

So, let's consider our empirical correlation matrix (in matrix notation):

$$\mathsf{C} = \frac{1}{T}\,\mathsf{R}^{\mathsf{T}}\,\mathsf{R}\,, \tag{3.5}$$

where R is an $T \times N$ matrix with elements $\{R_{m\,i} \equiv R_i(m\Delta t)\,; i = 1, \ldots, N\,; m = 0, \ldots, T-1\}$, and R^T denotes the transpose of R. Moreover, we consider a "random" correlation matrix

$$\mathsf{R} = \frac{1}{T}\,\mathsf{W}^{\mathsf{T}}\,\mathsf{W}\,, \tag{3.6}$$

where W is an $T \times N$ matrix containing N time series of T random elements with zero mean and unit standard deviation, mutually uncorrelated. It follows that R belongs to the Wishart matrices in RMT.

Beginning with this definite positive matrix with dimensions $N \times N$, where n is the number of variables or companies in our case, we perform the eigendecomposition, also called Principal Component Analysis (PCA) [35] in the literature. This step involves finding the eigenvalues and eigenvectors of $\mathbf{C}$. The eigenvalues, denoted as $\lambda_1, \lambda_2, \ldots, \lambda_n$, represent the variances along the principal components of the data. The corresponding eigenvectors, denoted as $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, define the directions of these principal components. We sort the eigenvalues in descending order and arrange the corresponding eigenvectors accordingly. This step is important as it determines the significance and contribution of each principal component to the data's variance.

We can decompose the empirical correlation matrix into:

$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{3.7}$$

where $\mathbf{V}$ is an $n \times n$ matrix whose columns are the eigenvectors of $\mathbf{C}$, and $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{C}$. The diagonal elements of $\mathbf{D}$ represent the eigenvalues.

In the original notation, one may write the diagonalized matrix as:

$$\mathbf{C} = \sum_{\lambda} \lambda \left|v_\lambda\right\rangle \left\langle v_\lambda\right|$$

where the eigenvectors $|v_\lambda\rangle$ form an orthogonal basis of $\mathbb{R}^N$, i.e. $\langle v_\lambda \mid v_{\lambda'}\rangle = \delta_{\lambda,\lambda'}$.

And we can eventually write $|r(t)\rangle$ as a projection:

$$|r(t)\rangle = \sum_{\lambda} \langle v_\lambda \mid g(t)\rangle \, |v_\lambda\rangle = \sum_{\lambda} a_\lambda(t) |v_\lambda\rangle$$

That is, we can rewrite the returns as a linear combination of the eigenvectors, i.e. we rewrite the returns on the eigenvector basis. We refer to the coefficient $a_\lambda(t)$ as **mode signal** of the λ-th eigenmode. The mode signals represent the temporal behavior of the eigenmodes, simplifying it tells you each time what the importance of the λ-th eigenmode explaining the whole multivariate signal.

Wrapping up everything, from the correlation matrix C we have obtained the eigenvectors $\mathbf{V}$, the eigenvalues denoted as $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the mode signals that for each eigenvalue correspond to a temporal series showing the importance in time of the former.

3.2. Significance of Empirical Correlation matrix

The interpretation and significance of the empirical cross-correlation coefficients, denoted as C_{ij}, poses several challenges stemming from various factors. Here are a couple of them:

(i) The dynamic nature of market conditions implies that cross-correlations between pairs of stocks may not exhibit stationarity over time [14].

(ii) The finite length of the available time series used for estimating cross-correlations introduces inherent "measurement noise".

To overcome the limitations associated with the finite length of time series, one might consider employing a longer time span. However, this approach is susceptible to the non-stationarity of cross-correlations, thereby affecting the accuracy of the estimates. Consequently, the empirically measured cross-correlations include stochastic contributions, making it challenging to discern the non-random correlations from the overall structure of the matrix C.

Determining which stocks exhibit consistent correlations throughout the studied period necessitates a systematic approach. To address this issue, we subject the statistical properties of C to a null hypothesis test. Specifically, we compare the properties of C with those of a randomly generated correlation matrix. The latter is constructed using uncorrelated time series. If the characteristics of C align with those of a random correlation matrix, it implies that the observed content within C is random. Conversely, deviations in the properties of C from those of a random correlation matrix provide valuable information about genuine correlations. Thus, our objective lies in contrasting the properties of C with those of a random correlation matrix, enabling us to partition the contents of C into two distinct groups: (a) the portion of C that adheres to the statistical properties of random correlation matrices, often referred to as "noise," and (b) the portion of C that deviates from these properties, representing valuable "information".

Why unrelated phenomena can appear to be correlated?

This issue arises due to the inherent limitation of the number of observations available for analysis. This is a well-known phenomenon in the economic literature called the problem of sunspot variables [64].

As a result, it is possible for two entirely unrelated phenomena, such as stock prices and sunspots, to exhibit apparent correlations over a specific time interval T. In other words, the correlation coefficient, which would ideally be zero when studying very long time series, can actually be on the order of $\frac{1}{\sqrt{T}}$ and may accidentally appear substantial.

When attempting to systematically correlate N input variables with M output variables, the total number of pairs considered becomes $N \times M$. In the absence of any genuine correlation between these variables, the largest among these $N \times M$ empirical correlation coefficients, for Gaussian variables, can be approximated as $\approx \sqrt{2ln(NM)/T}$. Notably, this value increases with the product of N and M. For instance, for $N = M = 25$ and $T = 200$, it is approximately 0.25 [14].

However, when the input and output variables have non-Gaussian distributions with fat tails, the dimensionality curse becomes even more pronounced. In these cases, the aforementioned value can be even larger. This means that if two randomly fluctuating variables happen to take large values at the same time, it can greatly contribute to the observed correlation, even though the true correlation should be close to zero for larger time intervals (T).

To put it simply, when dealing with non-Gaussian and fat-tailed variables as in the financial returns case [33] (as shown in figure 3.1), it becomes more challenging to distinguish between genuine correlations and random fluctuations. The statistical properties of these variables, combined with limited data, make the analysis more complex.

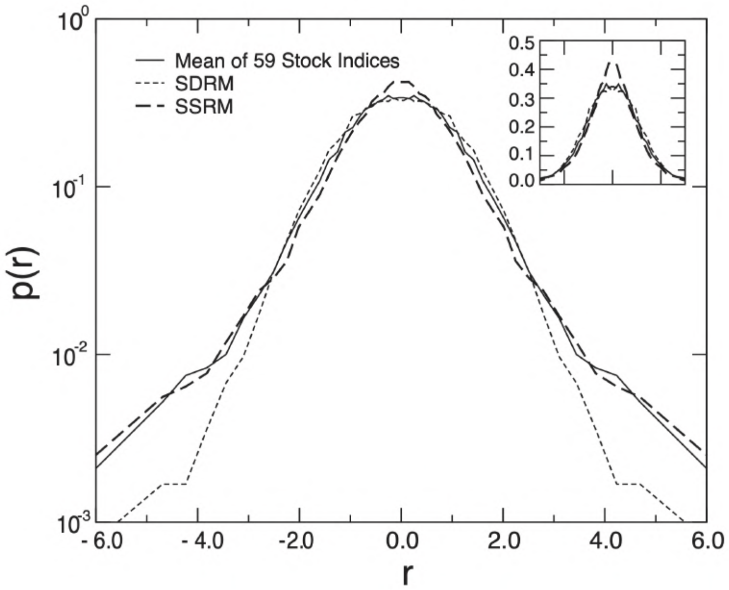


Figure 3.1: Symmetrized probability density distribution $p(r)$ of the returns r measured over periods $=$ 10 years for 59 stock indices. Also shown are the SSRM and SDRM for $\tau = 10$d. Image taken from "The origin of fat-tailed distributions in financial time series" [61].

3.3. Statistics of correlation coefficients

We delve into the analysis of the distribution, denoted as $P(C_{ij})$, which characterizes the elements $C_{ij};, i \neq j$ comprising the empirical correlation matrix, C. To gain insights, we examine $P(C_{ij})$ using quarter returns extracted from the CompuStat database.

Our initial observation reveals that $P(C_{ij})$ exhibits asymmetry and centers around a positive mean value ($\langle C_{ij} \rangle > 0$). This finding suggests that positively-correlated behavior tends to be more prevalent compared to negatively-correlated (anti-correlated) behavior.

3.4. Eigenvalues Distribution of the Correlation Matrix

In this section, we analyze the spectrum of the empirical correlation matrix and we focus on the dominant correlation modes in financial markets and their implications on market co-movement. Comparing the empirical eigenvalues to the correlation matrix to the theoretical upper edge of the Marchenko-Pastur (MP) spectrum allows one to extract statistically significant information [38]. As well known at this point "" deviations from the universal predictions of RMT identify system-specific, non-random properties of the

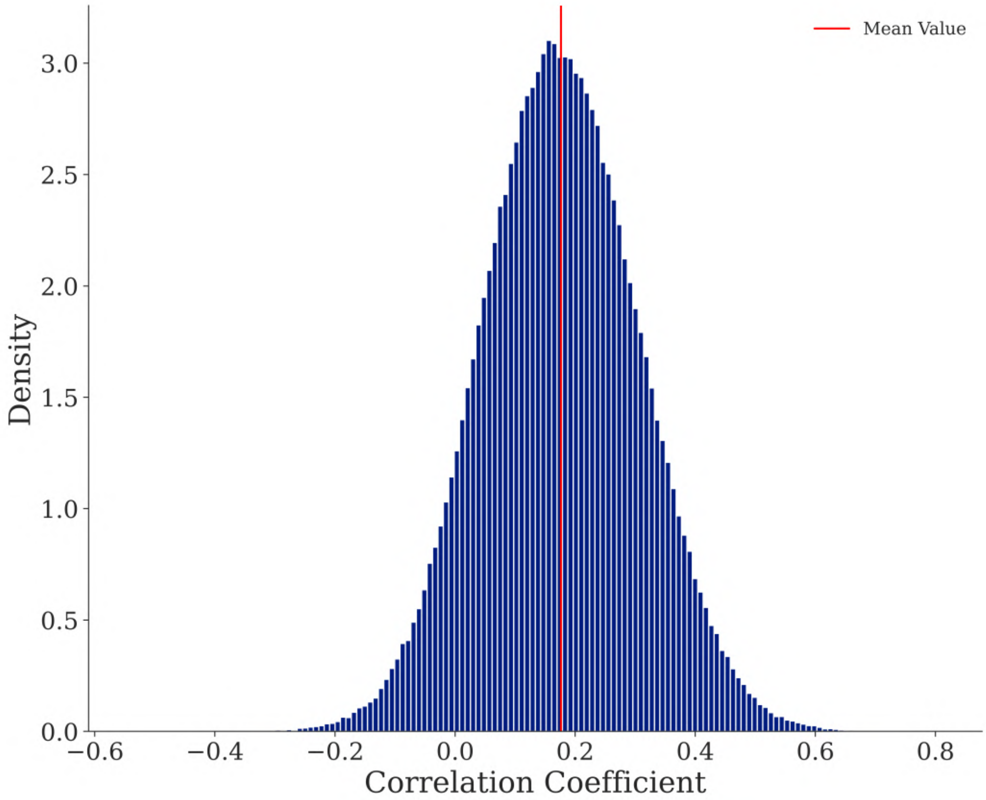


Figure 3.2: $P(C_{ij})$ for CompuStat prices dataset sampled quarterly.

system under consideration, providing clues about the underlying interactions. "" [53]

We have studied numerically the density of eigenvalues of the correlation matrix of returns prices of $N = 768$ assets of the CompuStat Dataset, based on quarter variations during the years $1961 - 2023$ (see the chapter **??**), for a total of $T = 248$ quarters (the corresponding value of q is 3.09). Finally, we compared it to the Marchenko-Pastur distribution.

In line with Laloux et al.'s research [39], we can construct a fit of the Marčenko-Pastur density with a modified parameter, $\sigma^2 = 1 - \lambda_1/N \approx 0.83$. This adjustment reflects the acknowledgment that the top eigenvalue and corresponding eigenvector are not random. Thus, it is reasonable to subtract the contribution of the market eigenvalue from the variance of the random component. The fit of this modified model is presented in Figure 3.3 for visual reference.

As a common fact in the literature [14, 16, 52] we found (see Figure 3.3) that the majority of eigenvalues from stock returns cross-correlation matrix overlap with the bulk of RMT eigenvalues, except:

- The first and largest eigenvector λ_1,

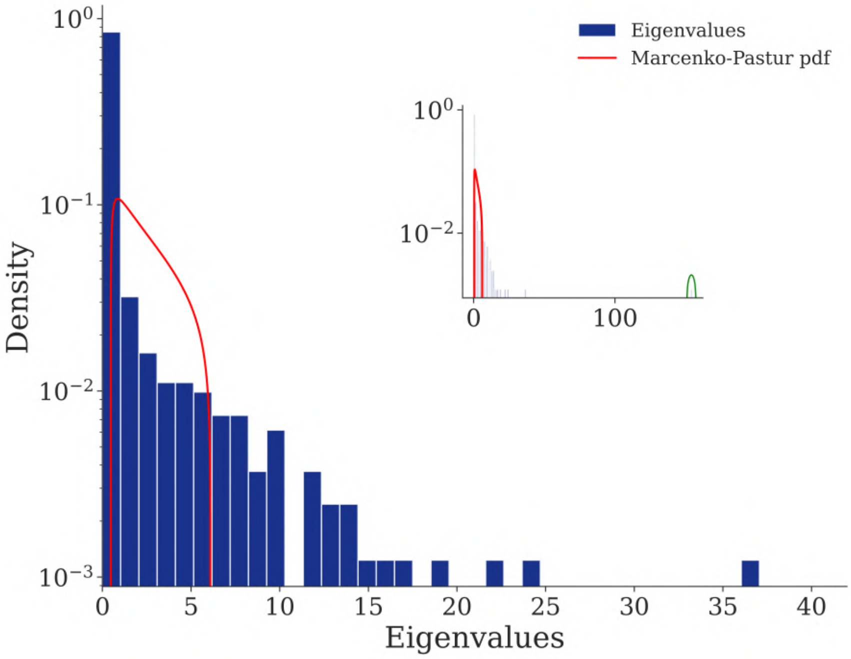


Figure 3.3: Distribution of empirical eigenvalues compared to Marchenko-Pastur distribution with $\sigma^2 = 0.83$ excluding λ_1. The y-axis of the histogram is in the log scale. Inset: same plot, but including the highest eigenvalue corresponding to the 'market', which is found to be 25 times greater than λ_{max}.

- Several large eigenvalues higher than the maximum theoretical eigenvalue λ_+,
- A density of eigenvalues around zero.

The last point is explained by degeneration since the value of q, in our case, is strictly larger than 1, so we expect to have a concentration of eigenvalues around zero, as explained in Chapter 1.

As we explore the characteristics of eigenvalue distributions we aim to determine the extent of randomness present in the spectrum. Notably, a significant number of eigenvalues in $\mathbf{C}$ align with the predicted distribution suggesting randomness in the matrix's contents, apart from those that exhibit deviations. In particular, we observe that the largest eigenvalue, denoted as λ_1, significantly exceeds the subsequent eigenvalues. In RMT terms The simplest 'pure noise' hypothesis is therefore inconsistent with the value of λ_1 since it's way larger than accepted by the Tracy-Widom theorem. This observation suggests that most stocks exhibit a strong positive mutual correlation, indicating the presence of a dominant correlation mode in the market. We explain this phenomenon from both a statistical and financial perspective, highlighting its coherence with the Capital Asset Pricing Model (CAPM) theory.

Eigenvalues play a crucial role in the analysis of correlation matrices, as they represent the variance explained by each corresponding eigenvector, as explained in the previous sections. In our context, the correlation matrix captures the interdependence between stock returns. The dominance of λ_1 over other eigenvalues signifies that a substantial portion of the total variance in the system can be accounted for by a single underlying mode (see figure 3.4).

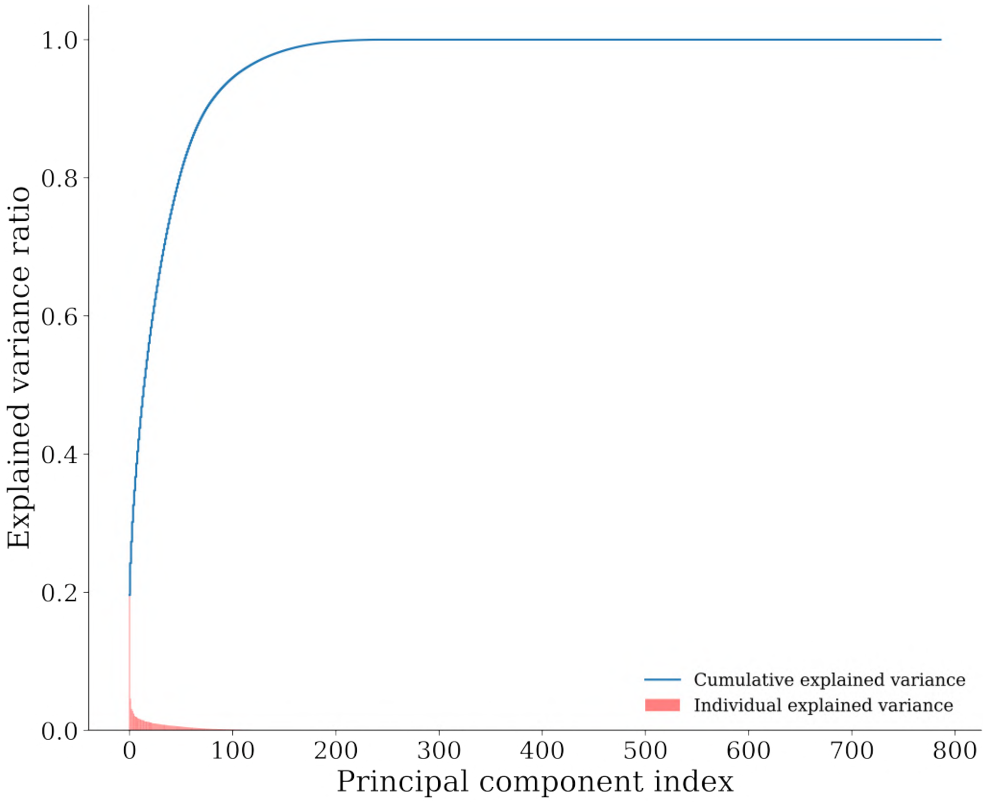


Figure 3.4: Explained Variance of the eigenvalues

From a statistical standpoint, the prevalence of a dominant correlation mode suggests that collective movements in the market are driven by a common factor that influences most, if not all, listed companies. This factor is likely associated with macroeconomic conditions, reflecting the shared exposure of companies to the overall economy. Consequently, the positive mutual correlation of stocks aligns with the fundamental understanding that economic changes impact various entities within the market similarly.

To elaborate further on the financial interpretation, consider the interconnectedness of companies operating within the same economic environment. As businesses navigate through economic cycles, they encounter similar market forces, regulatory changes, and consumer behavior. Consequently, their performances tend to exhibit similar trends, leading to a heightened positive correlation among their stock returns. This alignment with the broader market, known as the market mode, is indicative of the co-movement of stocks as a whole. [17, 38, 39]

The concept of a dominant correlation mode finds support in the Capital Asset Pricing Model (CAPM) [25], which says that the systematic risk of an asset is primarily determined by its correlation with the market portfolio. In this context, the market mode embodies the collective behavior of all stocks and serves as a proxy for the systematic risk that influences individual asset returns.

It has been shown, as well, that some information may also be buried below the band edge of the MP distribution [22], but we will not discuss it in our work.

3.5. Statistics of the Eigenvectors

In this section, we analyze the distribution of eigenvector components. The deviations of $P(\lambda)$ from the Marchenko-Pastur distribution suggest that these deviations should also be displayed in the statistics of the corresponding eigenvector components.

To fix the notation we will refer to the α component of the i-th eigenvector $|v_i\rangle$ as v^i_α.

3.5.1. Distribution of Eigenvector Components

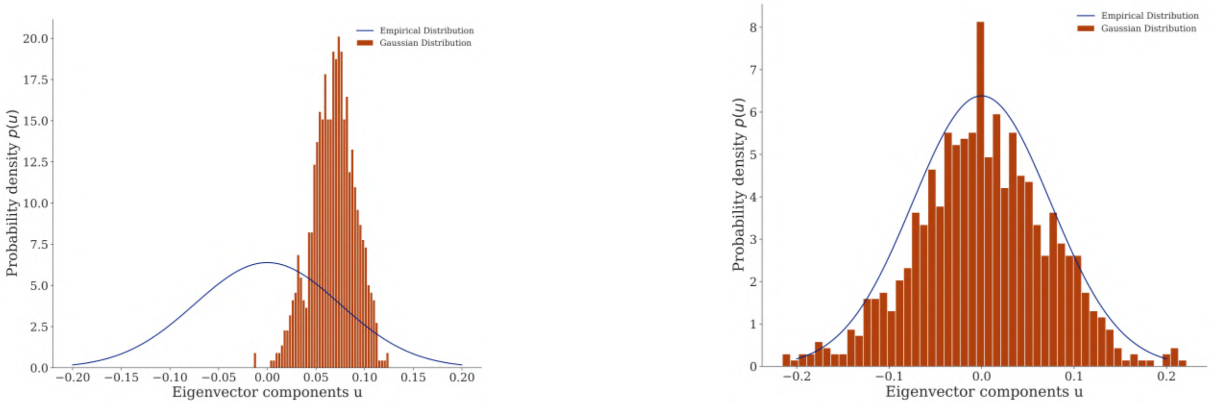
The idea that the low-lying eigenvalues exhibit randomness can be further examined by studying the statistical properties of the corresponding eigenvectors. Following the principles of Random Matrix Theory (RMT), if there is no specific information contained in the eigenvector $|v_i\rangle$, we expect that the distribution of its components v^i_α (as α varies) would follow a maximum entropy distribution (subject to the normalization constraint $\sum \alpha^N v^i_\alpha = 1$) [32]. This leads to the well-known Porter-Thomas distribution [13, 38], where the components $v^i_\alpha; \alpha = 1, \ldots, N$ of eigenvectors from a random correlation matrix conform to a Gaussian distribution with mean zero and unit variance:

$$P(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u}{2}\right) \tag{3.8}$$

where $P(u)$ represents the probability density function of the components of $|v_i\rangle$.

As depicted in Figure 3.5b, this theoretical distribution aligns remarkably well with the empirical histogram of the eigenvector components belonging to a typical eigenmode within the bulk of the Marchenko-Pastur (MP) distribution. Consistent results are obtained for other eigenvectors in the bulk region. However, deviations from the RMT result are observed for eigenvectors corresponding to the highest eigenvalues (Figure 3.5a), or more generally, for those lying beyond the theoretical edge λ_+, as illustrated for the case of λ_1. Specifically, the distribution of eigenvector components $p(u)$ systematically deviates from the Porter-Thomas distribution in these cases.

For instance, the top eigenvector (depicted in Figure 3.5a) exhibits significant deviations from a Gaussian distribution and appears to be approximately uniform, indicating that all stocks participate in it. Additionally, almost all components of this eigenvector share the same sign, leading to a shift in the distribution $p(u)$ towards one side. This suggests that the significant participants of the eigenvector share a common component that influences all of them with the same bias.



(a) First eigenvector

(b) 40-th (bulk) eigenvector

Figure 3.5: Distribution of the components u of eigenvectors compared to the Porter-Thomas distribution in the cases where (i) we consider the top eigenvector (deviating) and (ii) the 40-th eigenvector (bulk)

3.5.2. Interpretation of Eigenvectors

Structure of the Largest Eigenvector

As pointed out by several authors (see for instance, [38]), the dominant eigenvector is associated with the “market portfolio”, in the sense that all the coefficients of $|v_1\rangle$: v_i^1, $i = 1, \ldots, N$ are positive.

Structure of Deviating Eigenvectors

Having studied the interpretation of the largest eigenvalue which deviates significantly from RMT results, we next focus on the remaining eigenvalues.

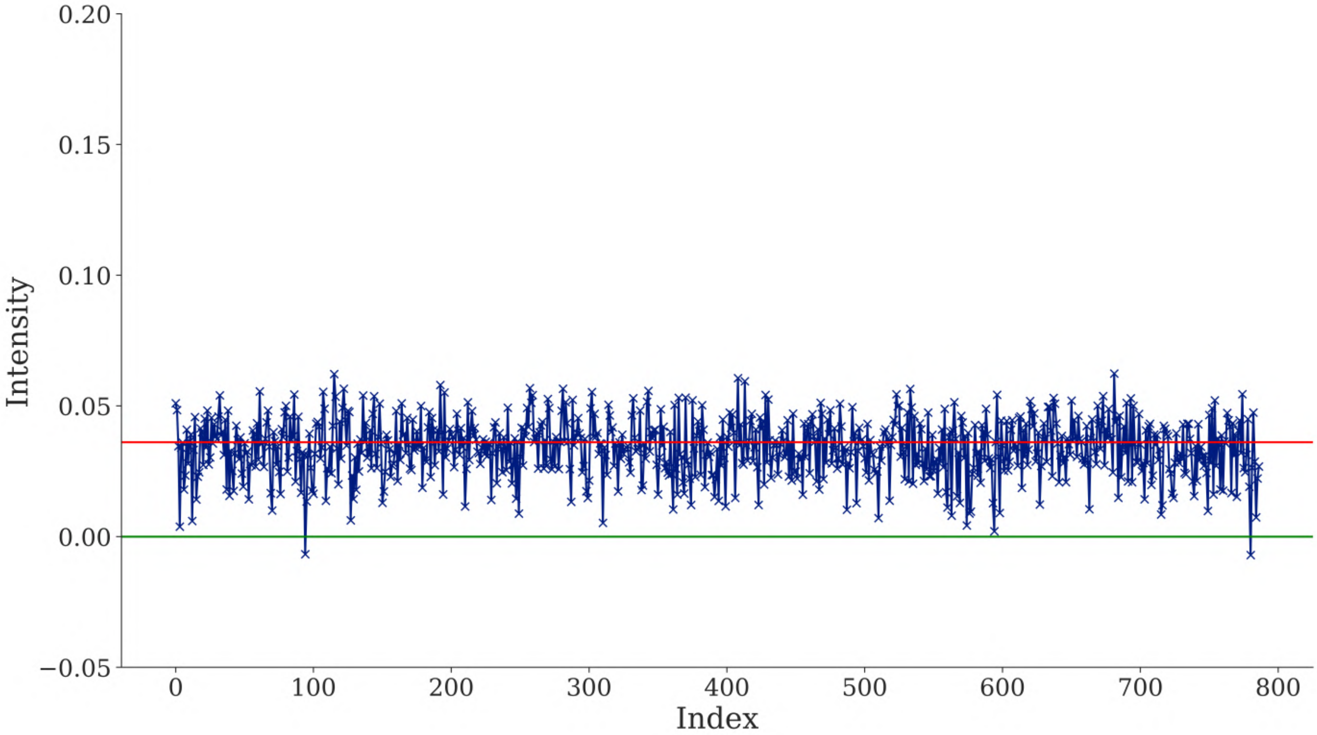


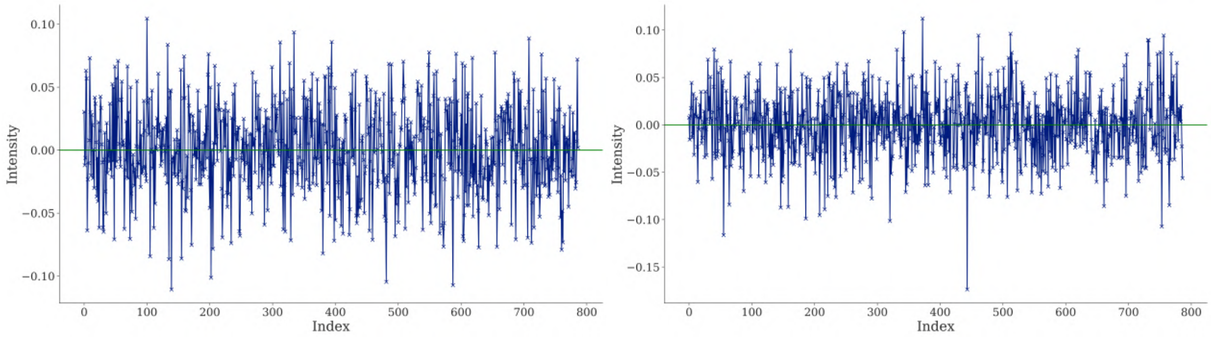
Figure 3.6: First mode. The red line represents $y = 1/\sqrt{N}$. Since the eigenvector is normalized and all the companies have similar participation in the market mode, we expect the components to have intensity close to $1/\sqrt{N}$

As a preliminary fact, to understand the eigenvectors that differ from the largest eigenvalue, we leverage the orthogonality property. We observe that these remaining eigenvectors must have negative components to ensure orthogonality with respect to the dominant eigenvector $|v_1\rangle$.

When examining successive eigenvectors, we anticipate that the "information" they carry is not uniformly distributed among all components. Instead, different components are expected to contribute differently to consecutive idiosyncratic oscillations of the market, also known as modes.

However, at first glance, the information encoded in these eigenvectors might not be readily apparent. As exemplified in Figure 3.7, where we present two modes—one residing outside the bulk, "far from the noise," and another within the bulk—their visual similarity is striking. This similarity arises due to the indices assigned to companies, which lack a sectorial distinction. In the subsequent section, we will demonstrate how sectorizing the index reveals valuable information and facilitates structural analysis.

Nonetheless, the deviations of the distribution of components within an eigenvector $|v_k\rangle$



(a) 2-th mode (deviating)

(b) 100-th mode (bulk)

Figure 3.7: We show the (i) 2-th mode that deviates from the bulk, (ii) 100-th that resides in the bulk.

from the RMT-predicted Gaussian distribution become more pronounced as the separation from the RMT upper bound $\lambda_k - \lambda_+$ increases. As proximity to λ_+ enhances the effects of randomness, we require a measure to quantify the number of components significantly participating in each eigenvector, thereby reflecting the degree of deviation from the RMT prediction concerning the distribution of eigenvector components. For this purpose, we employ the notion of the inverse participation ratio (IPR) [28], commonly employed in localization theory [28, 32, 46].

Inverse Participation Ratio (IPR) The IPR of the eigenvector $|v_k\rangle$ is defined as:

$$I^k \doteq \sum_{\alpha=1}^{N} [v_\alpha^k]^4, \tag{3.9}$$

where $v_\alpha^i, \alpha = 1, \dots, N$ are the components of eigenvector $|v_k\rangle$.

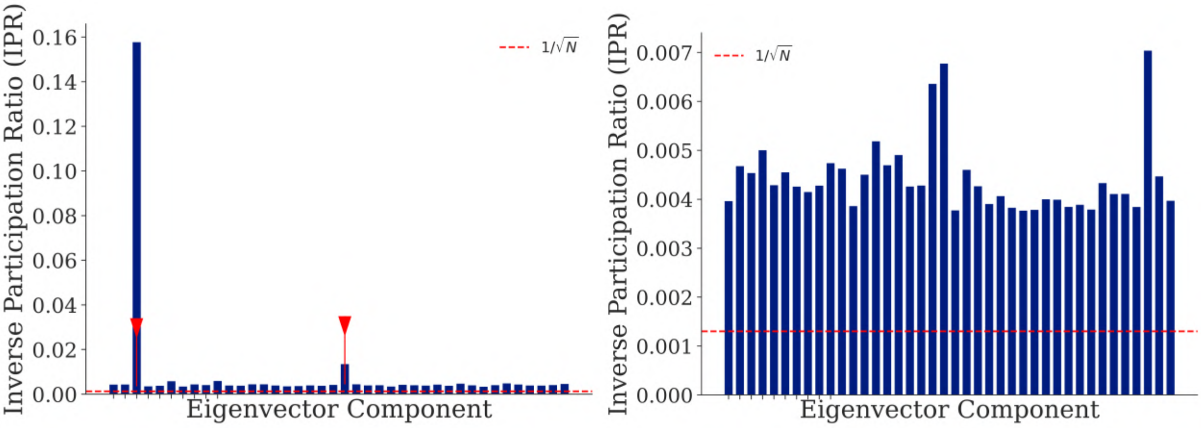
The meaning of I^k can be illustrated by two limiting cases:

- A vector with identical components $v_\alpha^k \doteqdot 1/\sqrt{N}$ for each α has $I^k = 1/N$.
- A vector with one component $v_l^k = 1$ and the remainder zero has $I^k = 1$.

The Inverse Participation Ratio (IPR) serves as a metric to assess the density of eigenvectors, reflecting the degree of significance attributed to their components. In essence, it quantifies the reciprocal of the number of eigenvector components that play a substantial role in characterizing the eigenvector, effectively gauging the localization of information within it.

When an eigenvector is densely concentrated in a few components, with the majority of its weight focused on specific elements, the corresponding IPR assumes higher values. On the other hand, if the eigenvector is more dispersed, involving a broader distribution of components, the IPR approaches the value of $1/\sqrt{N}$, indicative of an *extended* eigenvector.

To assess the density of eigenvectors for both sales and prices in our dataset, we compute the IPR for the first 40 eigenvectors, including both deviating and "random" modes. The results are visualized in Figure 3.8. For sales data, we observe two modes with notably higher density, suggesting the existence of a specific group of companies responsible for most of the variance in that mode of oscillation, capturing some idiosyncratic movements. These modes correspond to the third and 21-th modes. In the subsequent section, we will explore potential financial interpretations related to this observation.



(a) IPR for the first 40 modes of sales data

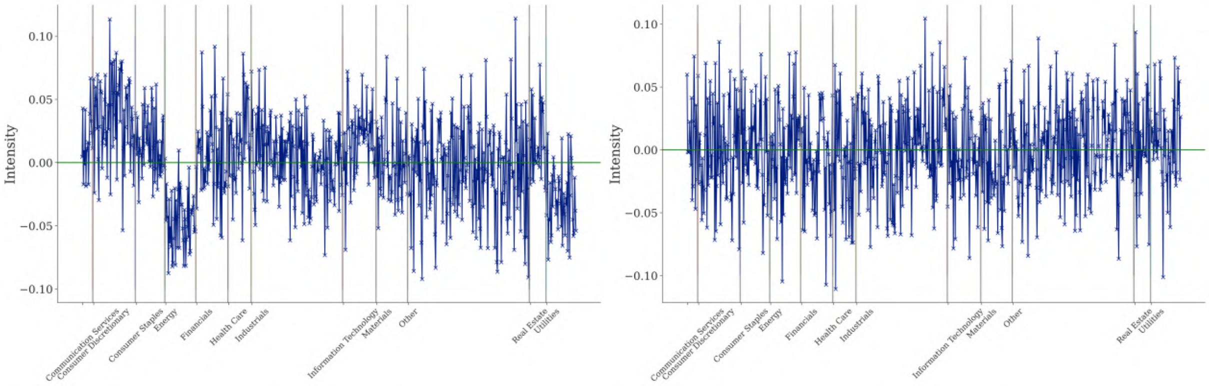
(b) IPR for the first 40 modes of prices data

Figure 3.8: We show the inverse participation ratio for (i) sales and (ii) prices. The IPR quantifies the reciprocal of the number of eigenvector components that contribute significantly. In (i) the two arrows point to the two modes that are more dense, respectively 3-rd and 21-th

Regarding the price data, we notice that there is no significant difference in terms of density between deviating and bulk modes, as indicated by the values of I, which are generally around $1/N$ with minor deviations. This finding suggests that the vectors are spread and extended, with nearly all components contributing to them.

It is worth mentioning that this observation has an interesting interpretation in the context of Anderson Localization theory, as discussed in [28, 49, 53]. In the context of localization theory, matrices containing "random band matrices" often contain extended states with small I^k in the bulk of the eigenvalue spectrum, whereas edge states are localized and have larger values of I.

Insights from Sectorization of Components To characterize the structure of deviating eigenvectors, we adopted a reframing approach by reindexing the components according to the sectors to which they belong. This reindexing allows us to identify the specific industrial sectors that significantly contribute to each of these deviating eigenvectors. In other words, this segmentation based on industrial sectors allows us to discern distinct patterns and correlations that might be otherwise obscured. To achieve this, we categorized the companies into industrial sectors using the Standard Industrial Classification (SIC) codes. Remarkably, our analysis reveals a distinct sector-clustered pattern within the eigenvectors. As anticipated, stocks belonging to similar industrial sectors demonstrate cohesive behavior, exhibiting comparable contributions to the explanation of business cycles. This sectorial representation of eigenmodes will be particularly valuable when studying lead-lag effects between sales and prices, offering a clearer interpretation of the underlying sector-specific dynamics.



(a) Sales

(b) Prices

Figure 3.9: Second eigenvector reindexed by sectors for the sales (i) and returns (ii).

Figure 3.9 displays the structure of the second eigenvector for both sales data and price data after sectorization. A noticeable distinction emerges between the two. In the case of sales data, a clear correlation structure becomes evident between the energy and utilities sectors, both of which are negatively correlated with other stocks in this mode. However, in the price data, such a coherent structure appears to be absent, making it challenging to identify meaningful correlations among the sectors. A similar lack of structure in prices is visible also analyzing the following deviating eigenmodes.

One intriguing question to address is whether there are significant lead-lag relationships within these sectorized eigenvectors. Despite their complex nature, these deviating principal components play a crucial role in constructing an efficient portfolio in the theory of Markowitz optimization [16].

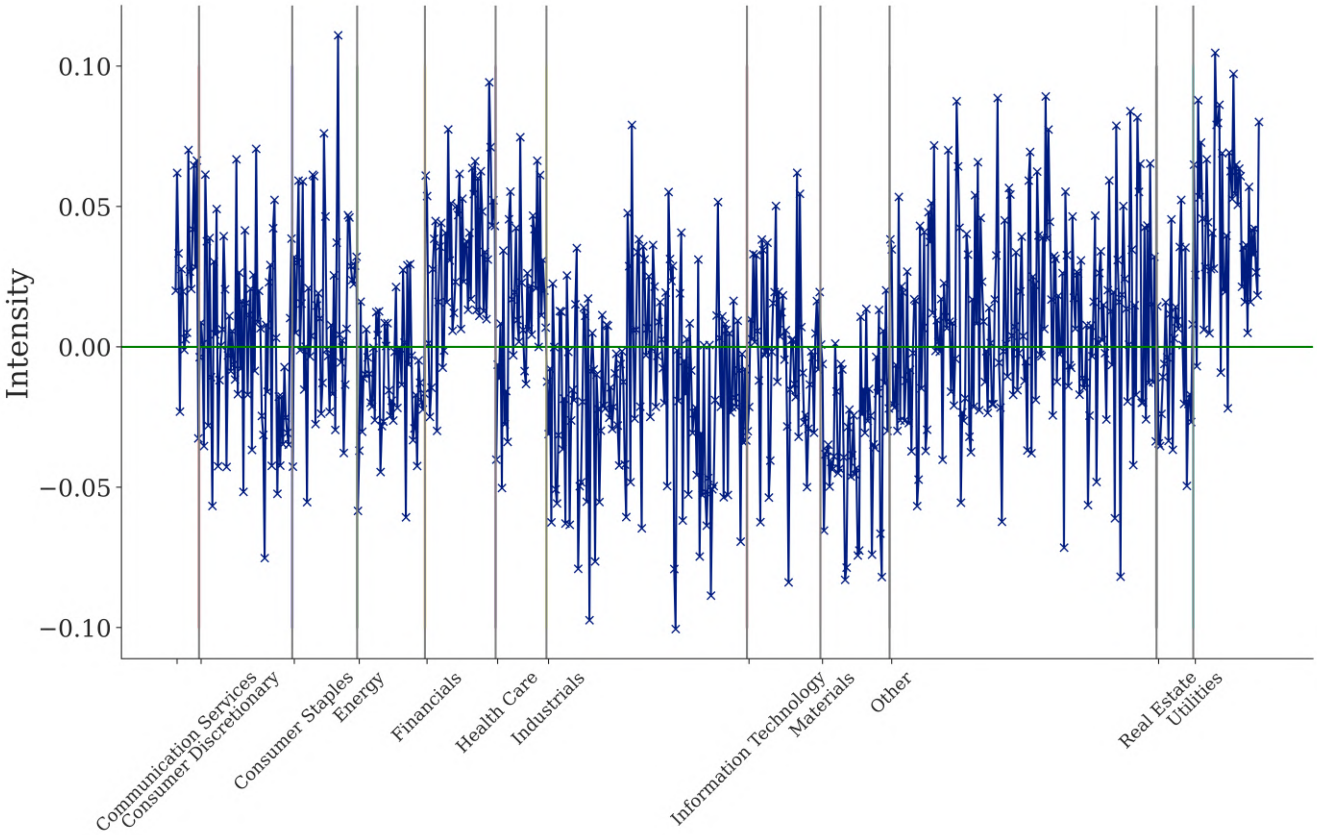


Figure 3.10: Second eigenvector reindexed by sectors for the sales

Let us delve deeper into the analysis of the third eigenvector for sales data (see Figure 3.10). As anticipated by the IPR analysis, a clear pattern emerges, and the eigenvector visibly contains valuable information. At this level of oscillation, we observe a distinct anti-correlation between the financials and industrial sectors, as well as between the energy and utilities sectors. However, it is essential to reiterate that as we move further along the eigenvectors, their explanatory power diminishes. Thus, this third eigenvector captures more intricate details of the correlation structure present in the data. Nevertheless, it is crucial to note that the primary mode, i.e., the largest eigenvalue, explains the majority of the variance, exhibiting notably higher magnitude compared to the rest.

For the sake of completeness, we also investigated the 21-th eigenvector (see Figure 3.11), which during the IPR analysis displayed intriguing characteristics. As this eigenvector resides at the edges of the bulk, the highest IPR value detected is likely due to randomness rather than a genuine pattern. Consequently, we do not observe any discernible patterns within this mode.

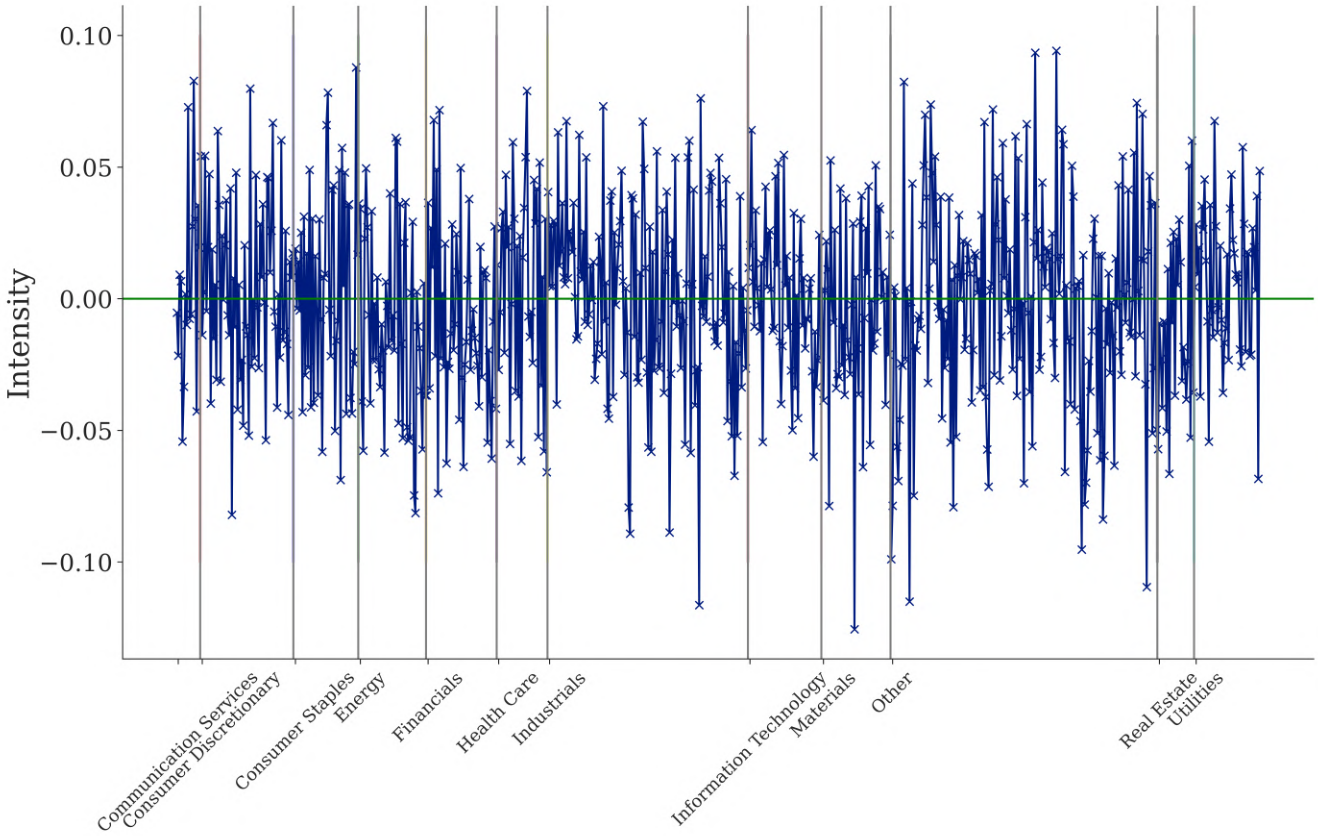


Figure 3.11: 21-th eigenvector reindexed by sectors for the sales

What about the last eigenvectors? In our aim of comprehending the underlying dynamics of the market, our focus now shifts towards the last eigenvectors, where the information content may not be immediately apparent. As we know from Random Matrix Theory (RMT), not all eigenvalues reside within the bulk of the Marchenko-Pastur distribution, with some deviating from the Gaussian orthogonal ensemble of random matrices. These smaller eigenvalues, clustering near the lower edge of the bulk, deserve to be explored.

First of all, we show in Figure 3.12 the Inverse Participation Ratio for the last 30 eigenvectors and we note, confirming our hypothesis of non-randomness for the first eigenvalues that the first components have a nontrivial value of IPR. The first component, which carries some meaningful information and is confined within a few components, it's visibly far from random.

Continuing our analysis, In figure 3.13 we show the structure of the last eigenvector as an example. Notably, this mode exhibits a fine-grained and idiosyncratic nature, capturing

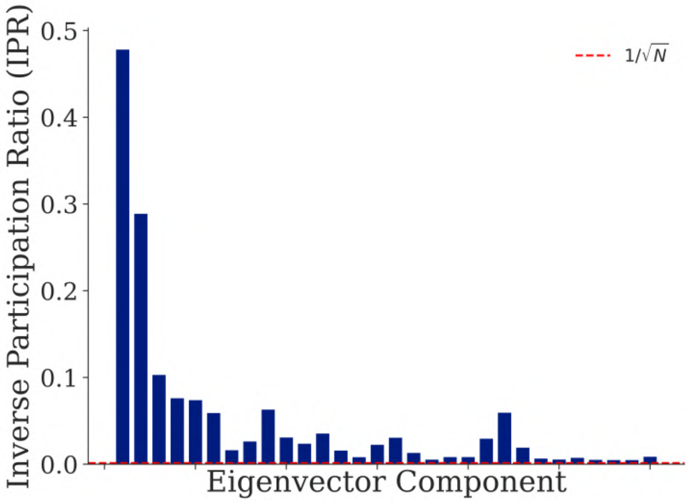


Figure 3.12: We show the inverse participation ratio for sales for the last 30 eigenvectors, in increasing order, in this specific case.

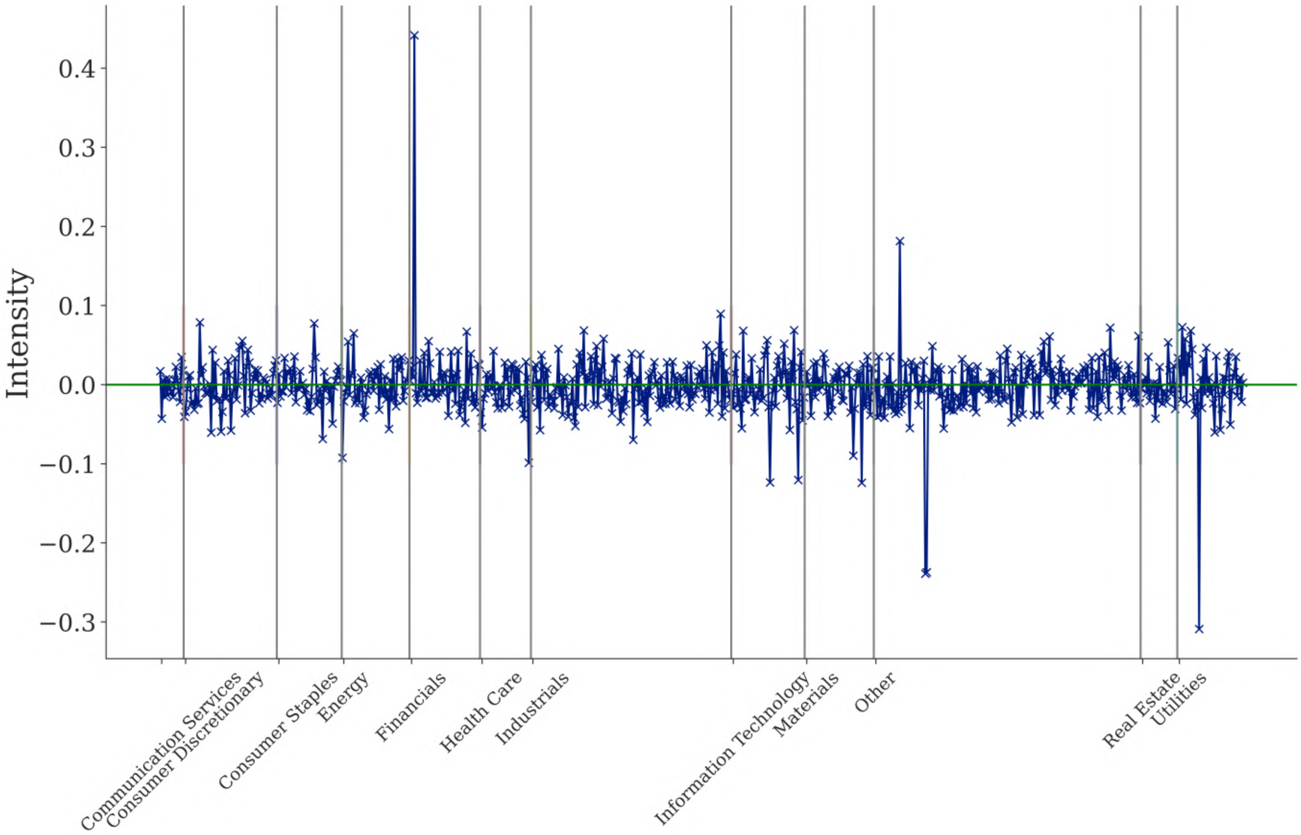


Figure 3.13: Last eigenvector reindexed by sectors for the sales

intricate patterns that may have evaded earlier scrutiny. Notably, the 5-th and 6-th components emerge as the most influential, displaying a strong anti-correlation effect.

Further examination reveals that these components correspond to AVX Corporation and Pinnacle West Capital, respectively. The former, a prominent American manufacturer of electronic components, and the latter, an electric utility holding company, both operate in interdependent sectors, potentially influencing each other's market dynamics.

Concluding, the sectorized representation of these eigenvectors shows the interplay between different industrial sectors within the market. Such information is instrumental for constructing well-balanced portfolios and understanding the underlying dynamics of market cycles.

3.6. Mode signals

In this section, we aim to explore the time-dependent coefficients that multiply each eigenvector when reconstructing the returns in the new basis obtained through Principal Component Analysis (PCA). To maintain self-consistency, let us recall that we can express the returns as a linear combination of the eigenvectors, effectively transforming the returns into the eigenvector basis:

$$|r(t)\rangle = \sum_{\lambda} a_{\lambda}(t)|v_{\lambda}\rangle, \tag{3.10}$$

Here, we refer to the coefficients $a_{\lambda}(t)$ as the **mode signals** of the λ-th eigenmode. These mode signals represent the temporal behavior of the eigenmodes, indicating the significance of the λ-th eigenmode in explaining the entire multivariate signal at each time instant.

Furthermore, we define the *relative mode intensity* $I_{\lambda}(t)$ as:

$$I_{\lambda}(t) = \frac{|a_{\lambda}(t)|^2}{\sum_{\lambda=1}^{N} |a_{\lambda}(t)|^2}, \tag{3.11}$$

which quantifies the fractional contribution of each eigenmode to the overall strength of price fluctuations at different time points.

In this section, our main focus is to analyze the mode signals corresponding to the most important eigenvectors, particularly the deviating ones. We seek to understand whether these mode signals exhibit specific structures, indicating potential non-linear mechanisms or randomness.

As shown in Figure 3.14, the intensity of the mode signals varies over time, with higher

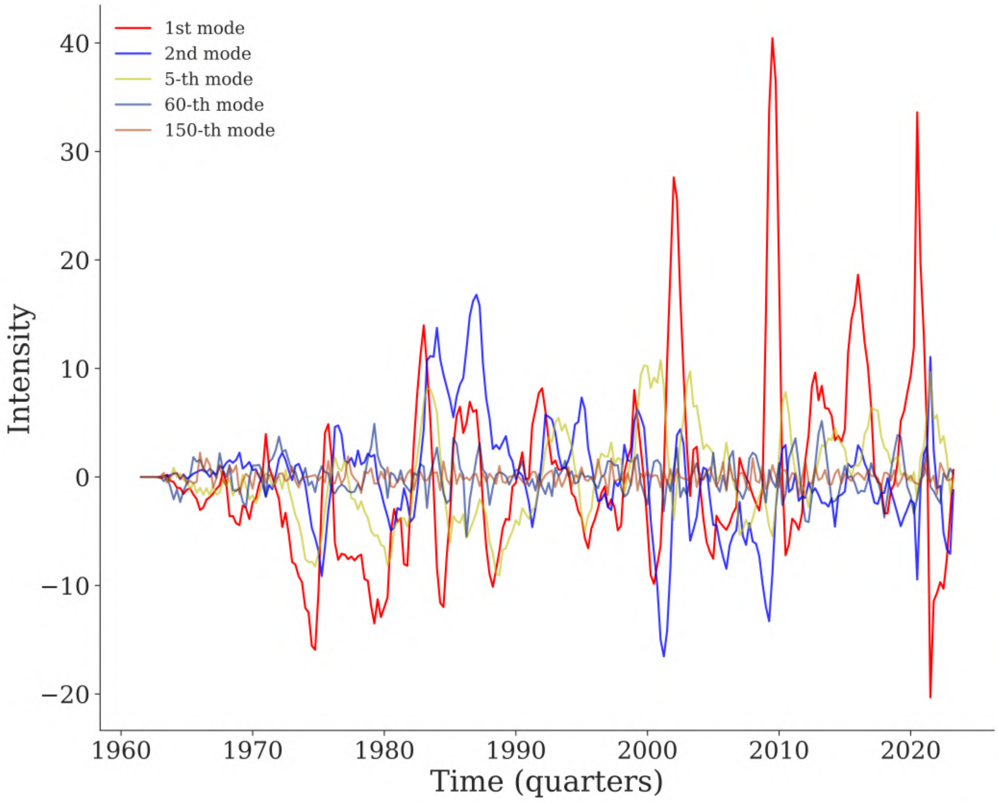


Figure 3.14: Multiple mode signals for sales data are depicted to illustrate the decreasing intensity with higher modes

modes generally exhibiting lower average intensities. Additionally, the mode signals are entirely uncorrelated as they belong to orthogonal modes.

Let us now focus on the first mode signal (Figure 3.15), denoted as $a_1(t)$. It is widely acknowledged in the literature that large values of the largest eigenvalue correspond to periods of high market volatility, suggesting the presence of strong collective behavior during volatile regimes [53]. This observation aligns with the idea that significant market events often lead to irrational market sentiment, resulting in more pronounced collective behavior [54]. To strengthen this correlation, we plot the first mode signals for sales data against the volatility of the United Kingdom stock price index in Figure 3.16. The results reveal a clear correlation between the mode signals and market volatility.

Since the first mode primarily represents the market, it is expected to be correlated, to some extent, with essential market indicators. We, therefore, examine its correlation with the Gross Domestic Product (GDP) and subsequently with the Interest Rate (3-Month Rates and Yields: Bank Bills for Australia).

In Figure 3.17, we aim to discern any correspondence between peaks in the GDP percent-

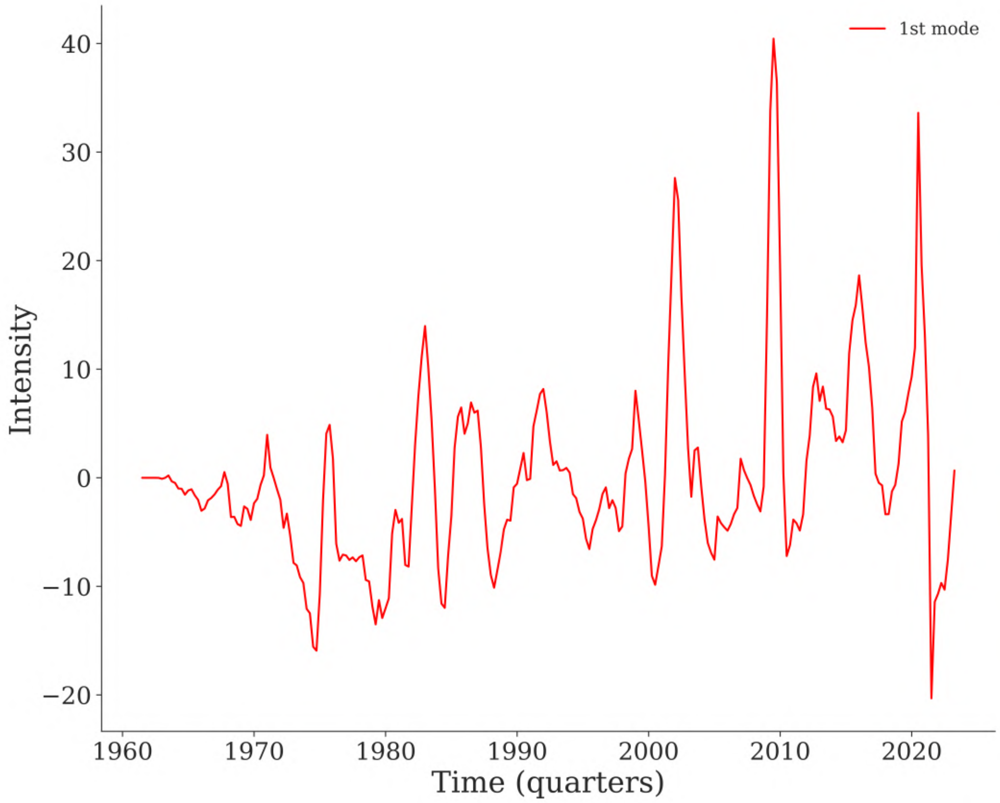


Figure 3.15: First mode signals for sales data.

age change and peaks in the market mode signal, while also comparing the GDP with the cumulative sum of the market mode signal. On the other hand, Figure 3.18 showcases the comparison between the Interest Rate and the market mode. In all figures, we properly rescaled the market mode signal to ensure comparability with the quantities under analysis.

3.6.1. Frequency-domain analysis of mode signals

In the context of economic analysis, the frequency-domain methods are a tempting approach, particularly when dealing with variables that exhibit cyclical behavior and are influenced by seasonal effects. As a well known fact, economic indicators such as GDP and unemployment often display patterns that can be linked to periodic fluctuations. [31]

By applying spectral analysis tools and Fourier transforms, we can gain insights into the temporal dynamics of the underlying processes. Specifically, these techniques allow us to discern the lengths of cycles, such as business cycles, and identify different phases of economic activity, such as periods of expansion or recession.

Examining the characteristics of an economic variable in the time-domain involves em-

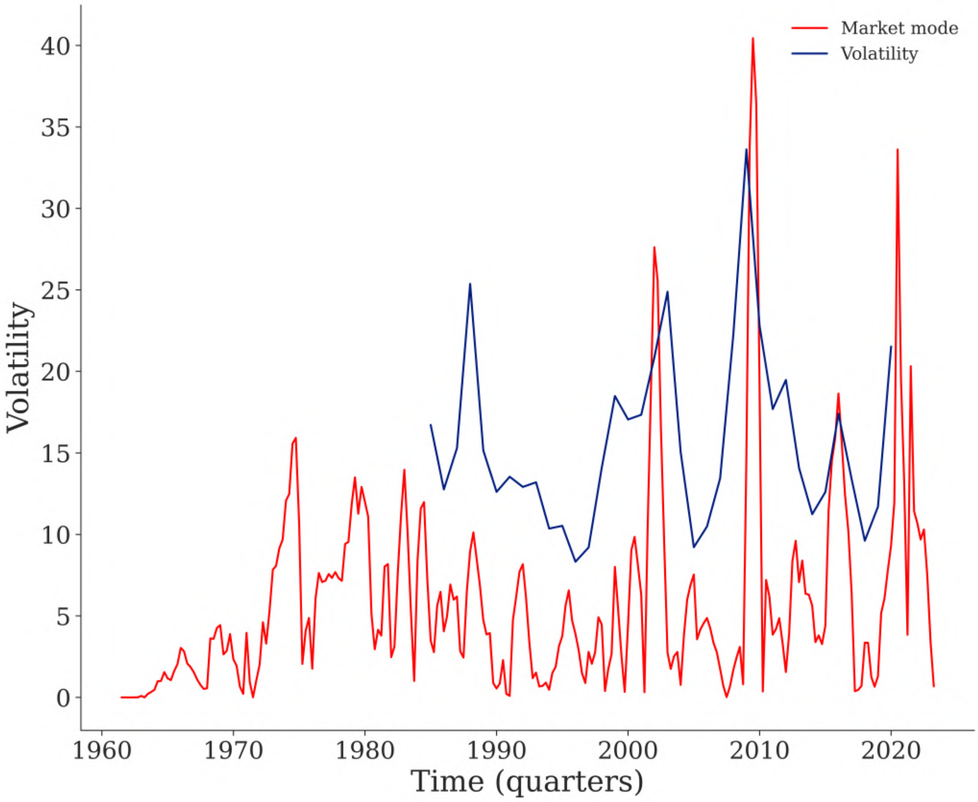
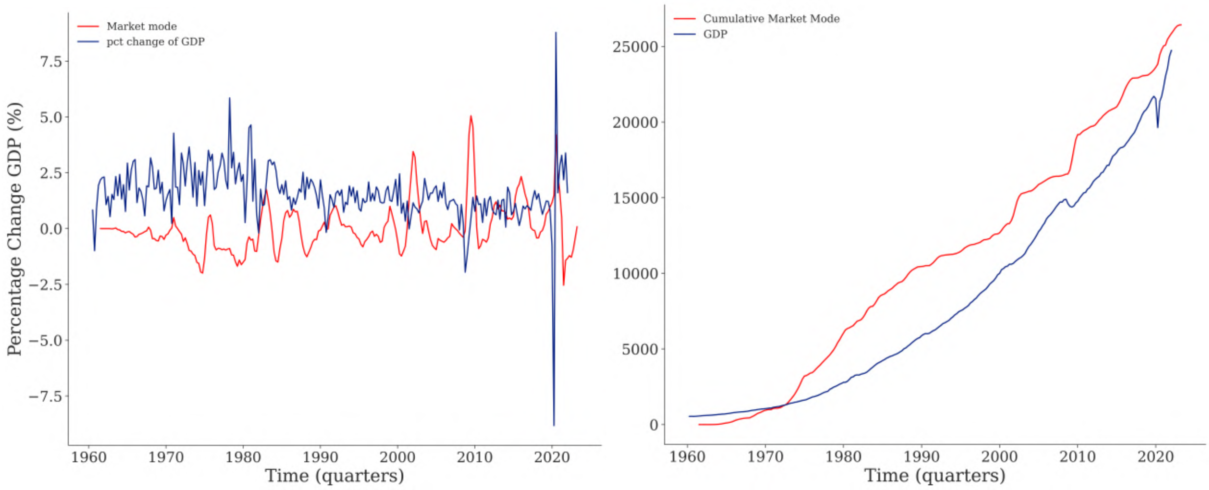


Figure 3.16: Absolute value of the first mode signals for sales data vs Volatility of Stock Price Index for United Kingdom, i.e. the 360-day standard deviation of the return on the national stock market index. (Bloomberg) - Source Code: GFDD.SM.01

ploying time-series analysis. On the other hand, spectral analysis focuses on investigating the behavior of the same economic variable across the frequency spectrum, thus operating in the frequency-domain. A fundamental aspect of spectral analysis lies in estimating the population spectrum or power spectrum, also known as the energy-density spectrum. The primary objective here is to understand how the variance of the studied variable can be partitioned into distinct frequency components.

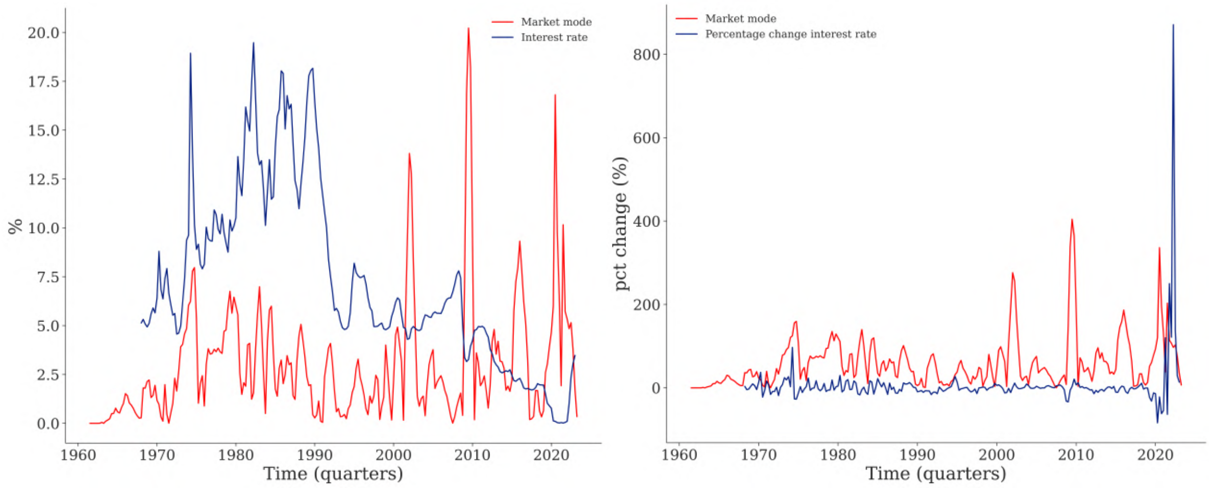
Fourier Transform The fundamental concept of spectral analysis revolves around transforming the original time series, denoted as $x(t)$, into a new sequence $X(f)$ that characterizes the significance of each frequency component f in the dynamics of the original series. This transformation is achieved through the discrete version of the Fourier transform (discreteness due to the fact that the time series is recorded at discrete time intervals), given by:



(a) Market mode vs pct change GDP

(b) Cumulative sum market mode vs GDP

Figure 3.17: (i) First mode signals for sales data vs percentage change of the Gross domestic product (GDP), the featured measure of U.S. output, is the market value of the goods and services produced by labor and property located in the United States. - National Income and Product Accounts of the United States (NIPA). (ii) Cumulative sum of the Absolute value of the first mode signals for sales data vs Gross domestic product (GDP).



(a) Market mode vs Interest rate

(b) Market mode vs pct change interest rate

Figure 3.18: (i) First mode signal for sales data vs Interest Rates: 3-Month Rates and Yields: Bank Bills: Total for Australia - OECD. (ii) first mode signal for sales data vs. percentage change Interest rate.

0.3

Figure 3.19: Typical examples of frequency spectra of some periodic time series composed of sinusoidal components.

$$X(f) = \sum_{t=-\infty}^{\infty} x(t)e^{-i2\pi ft} \tag{3.12}$$

where f represents the frequency at which $X(f)$ is evaluated. To gain a deeper understanding of this process, one can consider Euler's theorem, which allows us to express $e^{-i2\pi ft}$ as:

$$e^{-i2\pi ft} = \cos(2\pi ft) - i\sin(2\pi ft)$$

Hence, applying the formula 3.12 is equivalent to projecting the original signal $x(t)$ onto a set of sinusoidal functions, with each sinusoidal function corresponding to a particular frequency component.

Moreover, the inverse Fourier transform allows us to reconstruct the original signal $x(t)$ from $X(f)$:

$$x(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(f)e^{-i2\pi ft}df. \tag{3.13}$$

The formula 3.13 demonstrates that $X(f)$ reveals how much of each frequency component is required to synthesize the original signal $x(t)$. This decomposition process indicates the dominant frequency components.

Population Spectrum We define the population spectrum of a covariance-stationary process $y(t)$ as follows:

$$s_y(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}, \tag{3.14}$$

where γ_j represents the j-th autocovariance of $y(t)$ which is given by:

$$\gamma_j = E[(y(t) - \mu)(y(t-j) - \mu)]$$

, where μ denotes the expected value of $y(t)$.

$\omega = 2\pi f$ is a real scalar, and it is associated with the frequency $f = 1/\tau$ at which the spectrum is evaluated. Here, τ denotes the period length of one cycle at frequency f. Notably, the right part of equation 3.14 corresponds to the discrete-time Fourier transform of the autocovariance series. Additionally, there exists a close connection between this expression and the autocovariance generating function, defined as:

$$g_Y(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j, \tag{3.15}$$

where z denotes a complex scalar. As a consequence, the autocovariance generating function can be easily retrieved from the spectrum. Furthermore, employing the inverse discrete-time Fourier transform enables a direct estimation of the autocovariances based on the population spectrum.

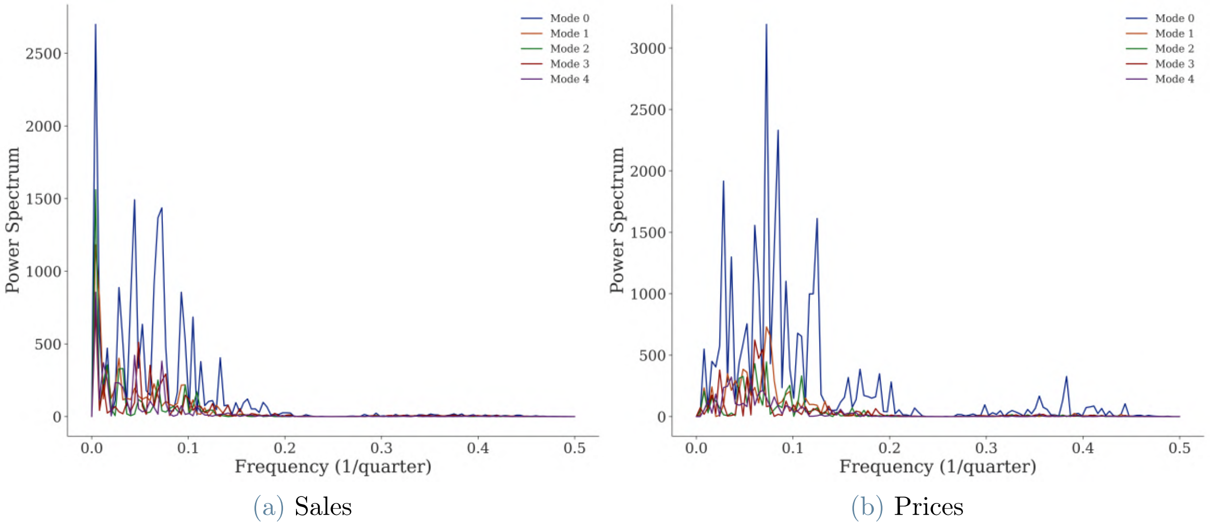
In practical applications, the typical and easiest approach employed to estimate the sample periodogram $\widehat{s}_y(\omega)$ is a non-parametric one, where the spectrum is inferred from a sample of realizations of the variable y, without assuming an explicit structure for the underlying data generating process. The estimation of the sample periodogram is straightforward, as it is directly related to the squared magnitude of the discrete-time Fourier transform $|Y(f)|$ of the time-series $y(t)$:

$$\widehat{s}_y(\omega) = \frac{1}{2\pi}\frac{1}{T}|Y(f)|^2, \tag{3.16}$$

where T represents the length of the time-series $y(t)$. The quantity $|Y(f)|^2$ is also referred to as the power spectrum of $y(t)$. This method is commonly known as the "periodogram" approach.

Let us now present the findings from analyzing the mode signals for the first eigenvectors. Our objective is to identify any discernible patterns and fundamental periods in these signals through frequency-domain analysis.

In Figure 3.20, we show the periodogram of the first 5 modes, revealing the contribution of each frequency component to the overall signal, as discussed earlier. The power values obtained from the periodogram highlight prominent frequencies, and we are keen on characterizing these periods to better understand any periodicities present in the data. This knowledge will be also valuable in studying lead-lag effects, as it allows us to identify



(a) Sales

(b) Prices

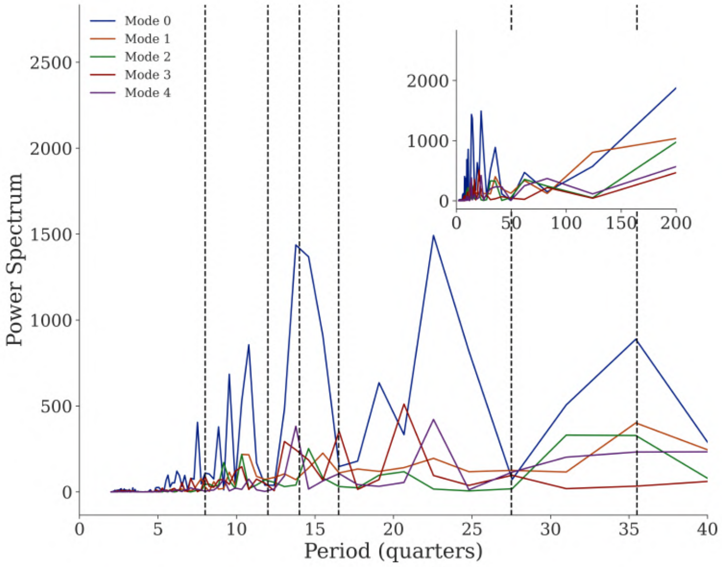
Figure 3.20: Periodogram for the first 5 modes, i.e. the contribution of each frequency component to the overall signal, in (i) sales data and (ii) prices data.

meaningful candidates' lag-time (τ) values.

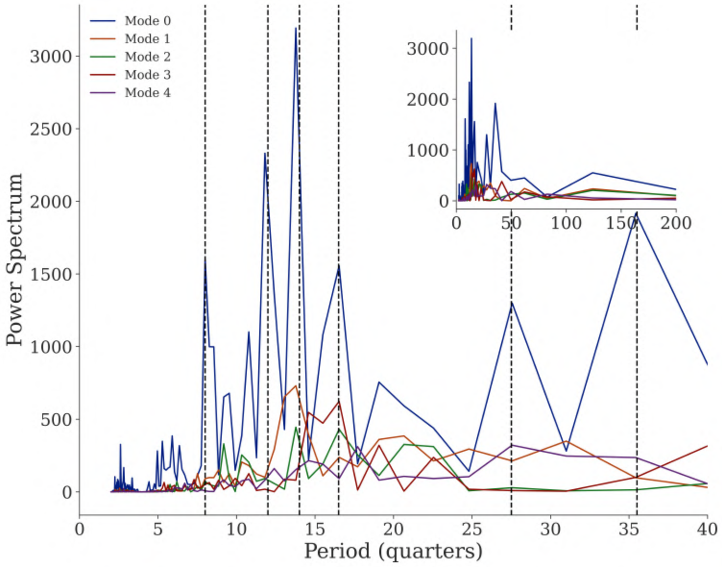
It is worth mentioning that we also include the periodogram plot for price data, which exhibits similarities with the sales data for the first modes. This suggests a complex periodicity in the market that is captured by frequency analysis. The presence of multiple peaks indicates several relevant periods in the mode signals. One particular observation is that the sales data shows a significant peak around the frequency corresponding to a period of $T = 248$ quarters. However, this peak is a finite-size effect, known as spectral leakage, that occurs due to the sample size.

Furthermore, we find it interesting that different modes share similar peaks, signifying that these frequencies effectively represent the underlying variability of returns over time. This is evidenced by the ability to reconstruct a large part of the variance by using just the first few modes.

In Figure 3.21, we directly plot the corresponding periods to the highest peaks in frequency, obtained simply by computing the inverse of the frequency $T = 1/f$. This visualization allows us to observe the most significant periods related to the data, providing valuable insights into the cyclic behavior and seasonality present in the time series. The black dotted lines in both (i) and (ii) represent periods $T = [8, 12, 14, 16.5, 27.5, 35.5]$, which are particularly noteworthy for the sales data. Comparing these significant periods with the prices periodogram reveals some common periods in both cases, indicating potential synchronized behaviors between sales and prices.



(a) Prices



(b) Sales

Figure 3.21: Periodogram for the first 5 modes but representing the period instead of the frequency. ($T = 1/f$). In (i) we show the prices data most relevant periods. In (ii) we show the same for sales data. In both figures, we plot the black dotted lines corresponding to the periods $T = [8, 12, 14, 16.5, 27.5, 35.5]$. These are the most significant periods for the sales data. Comparing them to the prices periodogram we see how a few periods are in common in the two cases.

3.6.2. Testing nonlinearities using the surrogate data methods

In this section, our primary focus lies in investigating the potential presence of nonlinearity in the mode signals extracted through Principal Component Analysis (PCA) from our returns data. The key question we aim to address is whether these mode signal time series can be attributed to a nonlinear dynamic or are more likely to be generated by a simple linear stochastic process. To achieve this, we will employ the method of surrogate data, a powerful statistical approach capable of detecting nonlinearity in time series.

To ensure the efficacy of the surrogate data method in capturing nonlinearity, we will initially apply it to well-known nonlinear processes, such as the Lorentz attractor. By subjecting these chaotic system's numerical data to the method, we can assess whether it successfully captures the underlying nonlinearity in these cases.

The method of surrogate data involves generating surrogate time series that maintain the linear properties of the original data while eradicating any nonlinear structures. Comparing statistical properties between the original time series and the surrogate data will enable us to gauge the presence of significant nonlinearity.

The formal implementation of the surrogate data method follows the principles of statistical hypothesis testing. It consists of two essential components: a null hypothesis, which serves as a candidate explanation for the data, and a discriminating statistic that quantifies specific aspects of the time series. The objective is to assess whether the null hypothesis adequately explains the data or if it falls short. The null hypothesis represents a potential explanation that we wish to challenge and demonstrate as insufficient for explaining the observed data. The discriminating statistic is a numerical measure that captures certain characteristics of the time series. When this statistic deviates significantly from what would be expected under the null hypothesis, the null hypothesis can be rejected. To estimate the distribution of the discriminating statistic, the surrogate data method employs direct Monte Carlo simulation. This approach involves generating multiple surrogate data sets that preserve certain properties of the original data while introducing randomness according to the null hypothesis. By comparing the observed discriminating statistic to the ensemble of statistics derived from surrogate data, the method enables us to evaluate the likelihood of the null hypothesis being valid.

Method of surrogate data The method starts by establishing a **null hypothesis** based on a chosen linear process. Subsequently, it generates multiple **surrogate data** sets that conform to this null hypothesis. Next, a **discriminating statistic** is computed for both the original data and each surrogate data set. If the calculated value for the original data significantly deviates from the ensemble of values obtained from the surrogate data, the null hypothesis is discarded, signaling the presence of nonlinearity.

To increase the method's effectiveness, we explore diverse null hypotheses and discriminating statistics. Even if this method avoids analytical derivations which can be very difficult if not impossible, the price to pay is that can be computationally intensive.

The method can be summarized with the following steps:

- Null hypothesis: The method starts by defining a **null hypothesis**, specifying the candidate process that may or may not adequately explain the data. The null hypothesis preserves certain properties of the original data, such as mean, variance, and possibly the Fourier power spectrum, while assuming no further structure in the time series.
- Surrogate data: An ensemble of surrogate data sets is generated from the original signal, adhering to the preserved properties dictated by the null hypothesis. The surrogate data sets are otherwise random, providing a basis for comparison with the actual data.
- Discriminating statistic: For each surrogate data set, a **discriminating statistic** is computed. This statistic serves as a measure of interest, and it can be chosen based on its relevance to the underlying dynamics or physical quantities of interest.
- Comparison and significance: The computed discriminating statistic for the actual sample time series is compared to the distribution derived from the surrogate data. By assessing the level of deviation, a measure of "significance" is determined.

Tailored Surrogate Data Method

Let us now specify the method we have built to test nonlinearities in our specific case. We started postulating the simplistic null hypothesis that our data come from a linear Gaussian process. The algorithm we have used to generate surrogate data is built on this assumption and it is called Unwindowed Fourier transform (FT) algorithm. To construct surrogate data sets, we ensure that their Fourier spectra match those of the raw data. To do so we just randomly shift the phases.

Algorithm 3.1 Unwindowed Fourier transform (FT) algorithm

```
Null Hypothesis: Assume data comes from a linear Gaussian process.
Fourier Transform: Compute the Fourier transform for positive and negative frequencies: $f = 0, 1/N, 2/N, ..., \frac{1}{2}$, where $N$ is the data size.
Randomize Phases: For each frequency, multiply the complex amplitude by $e^{i\phi}$, where $\phi$ is an independently chosen phase from the interval $[0, 2\pi]$.
Symmetrize Phases: To ensure the inverse Fourier transform results in real data, enforce $\phi(-f) = -\phi(f)$.
Inverse Fourier Transform: Perform the inverse Fourier transform using the symmetrized amplitudes to obtain the surrogate data.
```

The surrogate data method offers versatility in selecting a discriminating statistic. From a formal perspective, the key criterion for rejecting a null hypothesis is that the statistic demonstrates a distinct distribution for the data compared to the surrogate data. However, the method becomes more valuable when the chosen statistic provides an accurate estimation of a physically meaningful quantity. Given our interest in the possibility of chaotic underlying dynamics, we have opted to employ the correlation dimension [50], computed using the Grassberger-Procaccia algorithm [29], as our discriminating statistic. The correlation dimension has proven to be a valuable tool in distinguishing between random and chaotic time series in the existing literature [30].

$$\mid \mathbf{x}_i - \mathbf{x}_j \mid$$

The Grassberger-Procaccia Algorithm serves as a means of estimating the correlation dimension D for a given set of points randomly distributed according to a fractal measure μ.

Correlation Dimension Consider N points denoted by $\mathbf{x}_1, \ldots \mathbf{x}_N$ in a metric space with distances $\mid \mathbf{x}_i - \mathbf{x}_j \mid$ between any pair of points. For a positive number r, the correlation sum $C(r)$ is defined as the fraction of pairs whose distance is smaller than r:

$$\hat{C}(r) = \frac{2}{N(N-1)} \sum_{i<j} \theta\left(r - |\mathbf{x}_i - \mathbf{x}_j|\right),$$

where $\theta(x)$ is the Heaviside step function. This $\hat{C}(r)$ serves as an unbiased estimator of the correlation integral:

$$C(r) = \int d\mu(\mathbf{x}) \int d\mu(\mathbf{y})\theta(r - |\mathbf{x} - \mathbf{y}|).$$

Both $\hat{C}(r)$ and $C(r)$ monotonically decrease to zero as $r \to 0$. If $C(r)$ follows a power-law decrease, $C(r) \sim r^D$, then D is referred to as the correlation dimension of μ. Formally, the dimension is defined as $D = \lim_{r\to 0} \frac{\log C(r)}{\log r}$.

Consider a univariate (scalar) time series denoted as $x_1, \ldots, x_N$, where x_1 represents the measurement of the quantity x_i at time $t_i = t_0 + i\delta t$. Assuming stationarity, the statistical properties of the set x_i remain unchanged under time translation. However, unless the measurements are independent and identically distributed (i.i.d.), there will be correlations between consecutive measurements.

For chaotic systems, where the data is sampled from a trajectory on a strange attractor (or strange repeller), these correlations are weak and short-ranged. As a result, if the time series is of sufficient length N, we can consider the data to be effectively independent and randomly sampled from the invariant natural measure on the attractor. In such cases, we can directly apply the equations for analysis.

Moreover, employing Takens' time delay embedding theorem [48], we can transform a series of $N + m - 1$ univariate measurements into a time series of N delay vectors:

$$\mathbf{x}_i = (x_{i-m+1}, x_{i-m+2}, \ldots x_i) \in R^m$$

where m is the embedding dimension.

Lorenz Attractor

To gain familiarity with the surrogate data method and validate its effectiveness in detecting nonlinearity, we will apply this approach to the Lorenz attractor. The Lorenz attractor is a well-known chaotic system that exhibits intricate and sensitive dynamics. By focusing on the x-coordinate of the Lorenz attractor, we aim to assess whether our method can identify the underlying nonlinearity and chaotic behavior within this system.

The Lorenz system is described by a set of nonlinear ordinary differential equations, leading to a visually captivating three-dimensional butterfly-shaped trajectory, known as the Lorenz attractor.

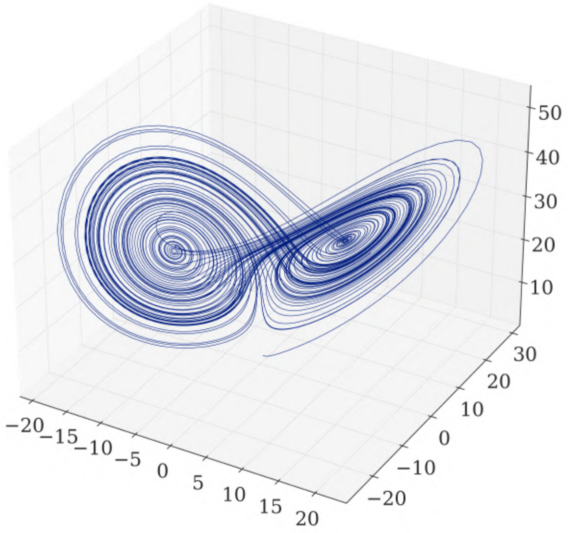


Figure 3.22: A visualization of a numerical simulation of the three-dimensional trajectory of the Lorenz equations.

$$\frac{dx}{dt} = \sigma(y - x), \tag{3.17}$$
$$\frac{dy}{dt} = x(\rho - z) - y, \tag{3.18}$$
$$\frac{dz}{dt} = xy - \beta z, \tag{3.19}$$

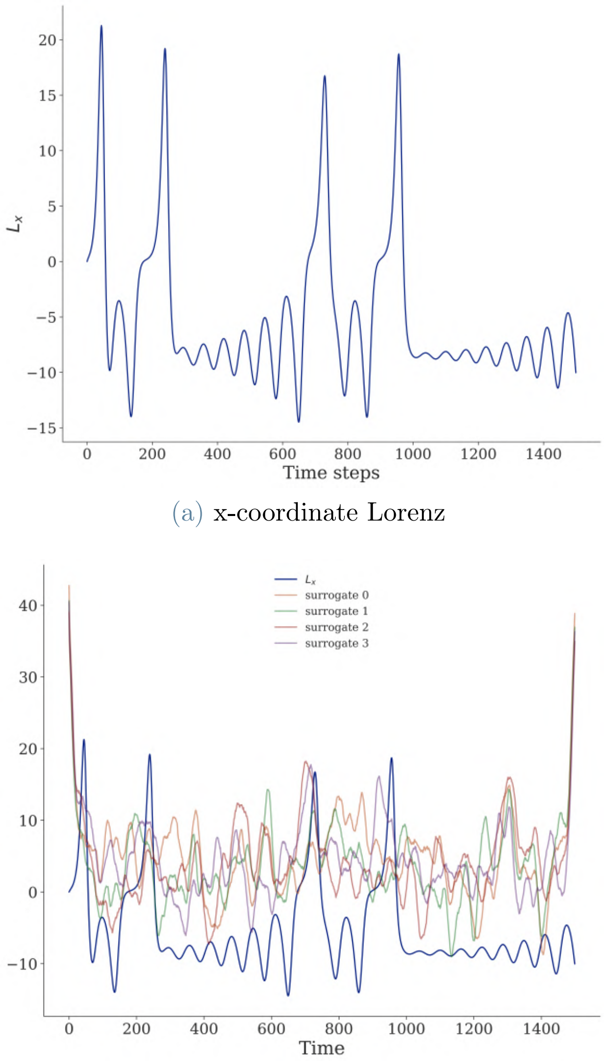
where x, y, and z represent the state variables, and σ, ρ, and β are parameters that dictate the system's behavior.

We will analyze the time series of the x-coordinate of the attractor 3.19 (see Figure 3.23 and investigate whether it originates from a nonlinear process, potentially indicating chaotic dynamics, using the tailored surrogate method data developed in the last section.

We hence compute generate 50 surrogates data and we compute (in a Monte Carlo fashion) the distribution of the discriminating statistic (correlation distance) for this ensemble. We then compare it to the value of the discriminating statistics computed on the x-coordinate.

The result is shown in Figure 3.24. It's clearly visible that the value of the correlation distance is far away from the distribution. This indicates that we can reject the null hypothesis that the inherent originating process is linear. Instead, we can confidently infer that the system exhibits non-linear characteristics, as indeed it does.

This confirmation underscores the method's efficacy in detecting and capturing non-



(a) x-coordinate Lorenz

(b) x-coordinate Lorenz and surrogates

Figure 3.23: (i) Time Series of the x-Coordinate of the Lorenz Attractor. (ii) x-coordinate with 5 generated surrogate data using the *Unwindowed Fourier transform (FT) algorithm*

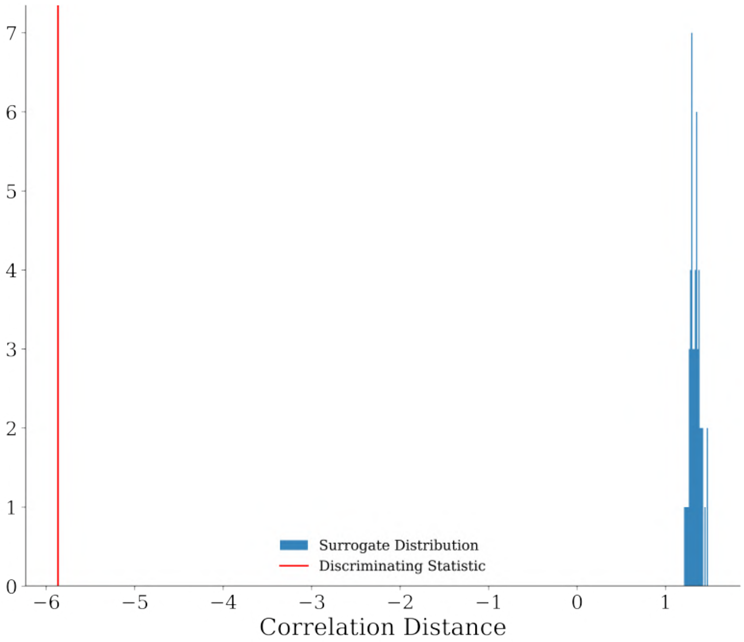


Figure 3.24: Distribution of the discriminating statistic compared for the original time-series and the surrogate data in the case of the Lorenz attractor.

linearities.

Non-Linearities in Mode Signals

Subsequently, we proceed to apply the method to the time series of the mode signals. As initially stated, our primary objective is to ascertain whether these mode signal time series exhibit non-linear dynamics. For the sake of illustration, we will initially focus on applying the method to the first mode. However, it is essential to note that the method can be readily extended to analyze subsequent modes as well.

Let us then apply the Method of Surrogate data machinery to the time series of the first mode signal (a sample of surrogates is shown in Figure 3.25). To do so, we applied the following steps:

1. Generate 100 surrogates data for the market mode.
2. Compute the discriminating statistic, i.e. the correlation distance for the surrogate data and the original time series.
3. Compare the distribution of the discriminating statistic for the surrogates with the value obtained for the original time series.
4. As a measure of significance we use the difference between the original and the mean

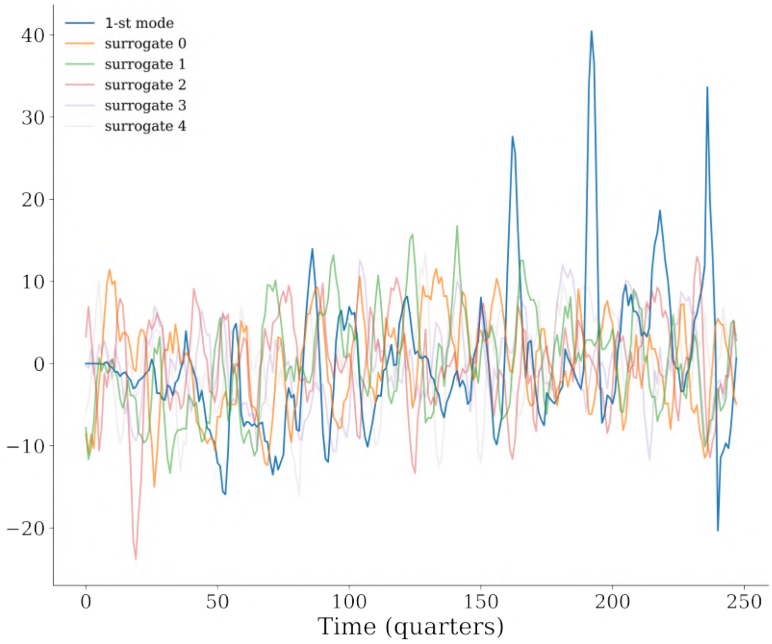


Figure 3.25: Time Series of the first mode signal and a few surrogates time series generated using the Unwindowed Fourier transform (FT) algorithm

surrogate value of the statistic, divided by the standard deviation of the surrogate values, following the approach defined in [59]. The larger the value the greater the difference between the two statistics.

$$\zeta = |\frac{\rho_D - \rho_S}{\sigma_S}|$$

The results obtained are shown in Figure 3.26. The picture makes it evident that the statistic significantly deviates from the surrogate data distribution. Let us better quantify this by computing the *measure of significance* in this specific case, which turns out to be:

$$\zeta = |\frac{\rho_D - \rho_S}{\sigma_S}| \approx 16.2$$

Similar results are observed for the successive modes. Figure 3.27 displays the discriminating statistics for the 2-nd and 3-rd modes, both exhibiting features of non-linearity, as expected. These outcomes are consistent across successive deviating modes, i.e. modes that reside outside the bulk of the Marchenko-Pastur distribution, which we will omit to show here for brevity.

Notably, as we analyze higher modes, we observe a reduction in the distance between the two statistics (using our measure of significance). This trend suggests that as we

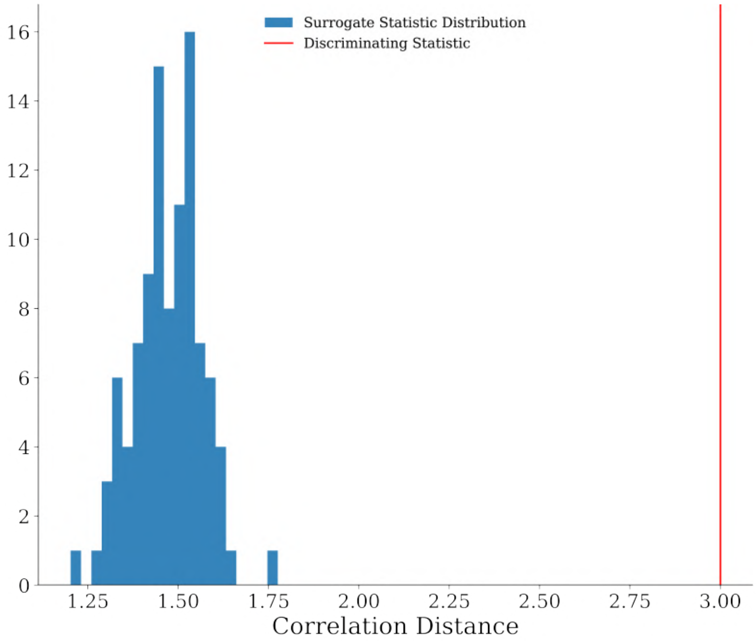
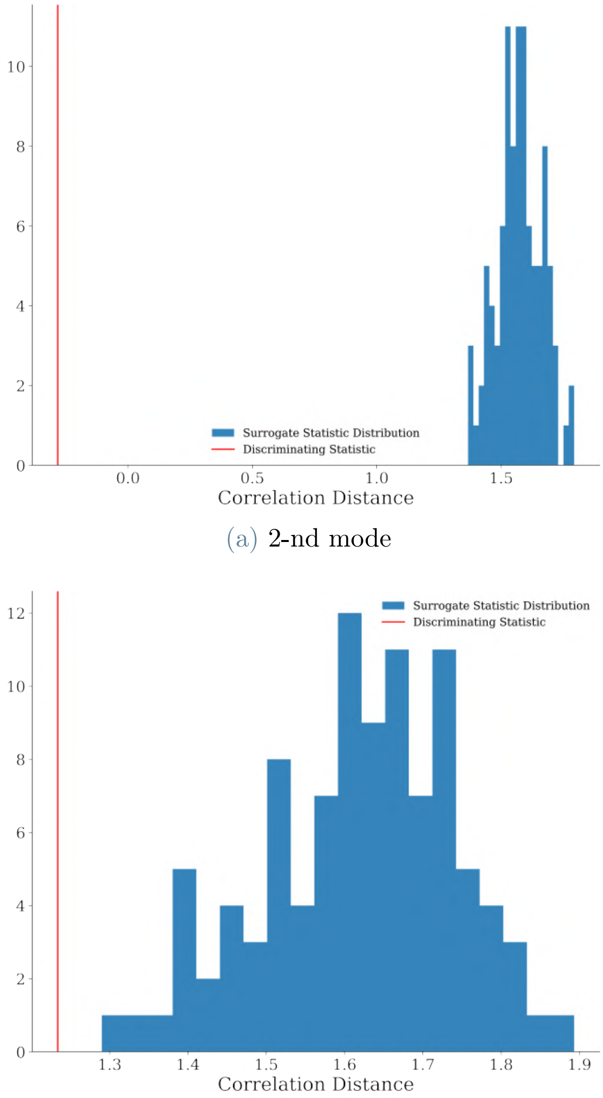


Figure 3.26: Distribution of the discriminating statistic compared for the original time-series and the surrogate data in the case of the first mode signal, i.e. the market mode.

move towards "random" modes, characterized by eigenvalues residing inside the bulk, the signal becomes less non-linear, approaching the null hypothesis of a linear Gaussian stochastic process. In particular, for modes inside the bulk of the MP distribution, there's no evidence of non-linearity, following the method of surrogate data's scheme. For the sake of conciseness, we will not show them here.

Concluding, it's out of our scope to find and investigate a possible non-linear model that explains the mode, but it's worth saying that this method, at least, suggests that the signal is nonlinear and hence it's possible to investigate this theoretical model. This is the essence of the inverse problem for a nonlinear system, i.e. to determine the underlying dynamical process in the practical situation where all that is available is a time series of data. In our case, however, we took a more modest goal the detection of nonlinear structure in a stationary time series.



(a) 2-nd mode

(b) 3-rd mode

Figure 3.27: Distribution of the discriminating statistic compared for the (i) 2nd mode and (ii) 3rd mode

4 | Do sales returns and prices returns originate from the same generative process?

A significant finding of our study relates to the disparity between sales and price returns time series. Notably, these two series exhibit distinct structures. A major limitation in studying sales arises from their fixed frequency of sampling, as public companies report sales every quarter. Consequently, when computing the correlation between sales series, we are constrained to use this coarse-grained data and rely on empirical correlation to infer the true value of the correlation between sales.

In the previous chapter, we observed that the sales and price modes differ in structure, particularly in the sectorial representation of the second mode, which is the most informative mode apart from the market mode. Despite representing the same companies, the price mode demonstrates significant differences compared to the sales mode. Surprisingly, we found no identifiable structure within the price eigenvectors. This starkly contrasts with the eigenvectors derived from sales returns time series, which display clear associations with various industrial sectors. Understanding the underlying reasons and implications of this discrepancy is crucial and represents a fundamental aspect of our subsequent analysis.

In this chapter, we want to address the following questions:

- Initially, we hypothesized that the disparity in structure between the time series could be attributed to the different frequencies of sampling. We wondered whether prices might not reveal their true correlation structure when sampled quarterly. To investigate this, we explored downsampling the price data at various frequencies, including daily prices. However, our findings indicate that the observed phenomena are not similar when using different sampling frequencies. Even when we attempted to replicate the low-frequency characteristics of sales growth time series by downsampling from daily to quarterly prices, the same structure persisted. Hence, we can't attribute the absence of structure in prices to the quarterly sampling alone; rather, it seems to be intrinsically linked to the nature of the price data itself.

- Despite the difficulty in inferring the influences between sales and prices, particularly in understanding how the two processes mutually affect each other, we sought to answer a simpler question: whether the two processes could hypothetically originate from the same generative process. Our significant finding reveals that the difference between two "draws" of sales and prices, assuming they stem from the same "true" generative process plus noise, is inconsistent with two draws from the same distribution. In other words, the two quantities are not merely different realizations of the same distribution process. Although we observed some correlations between the two, we discovered that they originate from distinct generative processes.

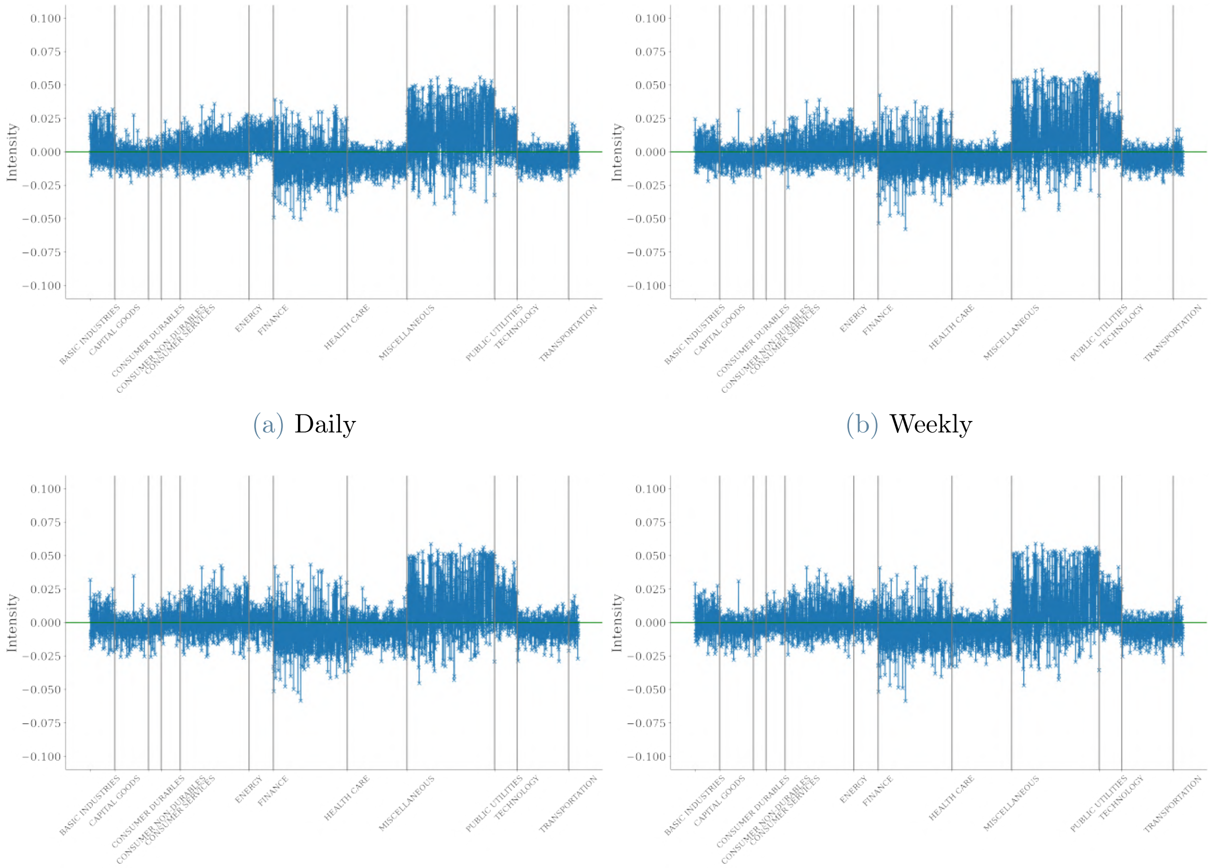
4.1. Consistent Structure of Prices' Mode Across Various Sampling Frequencies

In our analysis of (log) returns, we calculate the difference between (log) prices at various time intervals. While the frequency of sampling in sales data is fixed at three months, we have more flexibility in selecting sampling frequencies for prices, ranging from seconds to daily intervals. The choice of sampling frequency for computing (log) returns allows us to explore different statistical characteristics, a well-known fact in the literature [24]. This raises the question of whether the correlation structure depends on the chosen sampling frequency.

In this section, we aim to demonstrate that the homogeneity or disparity in structure within our CompuStat prices data is not influenced by the particular forced sampling frequency used. Instead, it is an inherent characteristic of the data sample itself.

To investigate this, we collected daily data for over 5000 stocks spanning from January 2, 1970, to August 24, 2018. We then computed the eigendecomposition and related statistics for different time intervals when calculating the log return. Specifically, we calculated log returns by subtracting the "adjusted close" price (i.e. the closing price after adjustments for all applicable splits and dividend distributions) with intervals of 1, 7, 30, and 90 days, respectively. The sectorial representation of the second most informative eigenvector (excluding the market mode) for different sampling frequencies is depicted in Figure 4.1. Remarkably, despite some minor differences, the structure remains relatively consistent across various sampling frequencies.

This analysis provides evidence to support the legitimacy of using quarterly frequency for



(a) Daily

(b) Weekly

(c) Monthly

(d) Quarterly

Figure 4.1: Sectorial representation of the second top eigenvector for a different sampling of the returns. (i) Daily returns, (ii) weekly returns, (iii) monthly returns and (iv) quarterly returns. Similar sectorial structures are observed

computing correlations and modes in our treatment of CompuStat sales and price data. The observed stability of the structure across different sampling frequencies suggests that our findings are not contingent on a specific frequency choice but reflect an inherent property of the data itself.

4.1.1. Spectral Distance Metric

To quantitatively examine the change in economic sectors (eigenvectors of the correlation matrix) when altering the sampling frequency, we employed the Spectral distance metric. The core idea was to assess the overlap between the eigenspaces generated by different correlation matrices obtained from various sampling frequencies. We assumed that if the same information is present when computing the correlation matrix for the returns, the resulting eigenspaces should exhibit a similar structure, leading to substantial overlap. To study this overlap, we employed the distance metric introduced by R. Allez and J.P. Bouchaud in their article [1].

> The idea is then to study the stability of a whole subspace V_0 spanned by the top n eigenmodes $|v_1\rangle, |v_2\rangle, \ldots |v_n\rangle$ with respect to the new subspace $V_0^{'}$ spanned by the top ten eigenvectors $\left|v_1^{'}\right\rangle, \left|v_2^{'}\right\rangle, \ldots \left|v_n^{'}\right\rangle$ of the new correlation matrix obtained, for example, downsampling quarterly returns to monthly returns.
>
> To quantify the overlap between these two spaces, we constructed the $n \times n$ rectangular matrix of overlaps $\mathbf{G}$ with entries defined as:
>
> $$G_{ij} := \left\langle v_i \mid v_j^{'} \right\rangle . \tag{4.1}$$
>
> The n non-zero singular values $1 \geqslant s_1 \geqslant s_2 \geqslant \ldots \geqslant s_n \geqslant 0$ of $\mathbf{G}$ provide complete information about the overlap between the two spaces. For instance, the largest singular value s_1 indicates that there exists a certain linear combination of the n perturbed eigenvectors with a scalar product s_1 when compared to a certain linear combination of the n unperturbed eigenvectors. If $s_1 = 1$, it implies that the initial subspace is entirely spanned by the perturbed subspace. Conversely, if $s_1 \ll 1$, it suggests that the initial and perturbed eigenspaces are nearly orthogonal to one another, as even the largest possible overlap between any linear combination of the original and new eigenvectors is very small. A suitable measure of the overlap distance $D\left(V_0, V_1\right)$ between the two spaces V_0 and V_1 is given by the average of the logarithm

of the singular values:

$$D\left(V_0, V_1\right) := -\frac{\sum_i \ln s_i}{n}$$

but alternative measures, such as $1 - \sum_i s_i/n$, can be considered as well.

In our treatment, we, therefore, compute the overlap distance, from now on called **spectral distance**, between the subspaces generated by the top 10 eigenvectors of the correlation matrix. We decided to include only the first 10 eigenvectors in the computation because we wanted to compare only the most significative informative mode in the comparison since we are less interested in the others, and we want to be careful to exclude noisy, bulk information from the former distance.

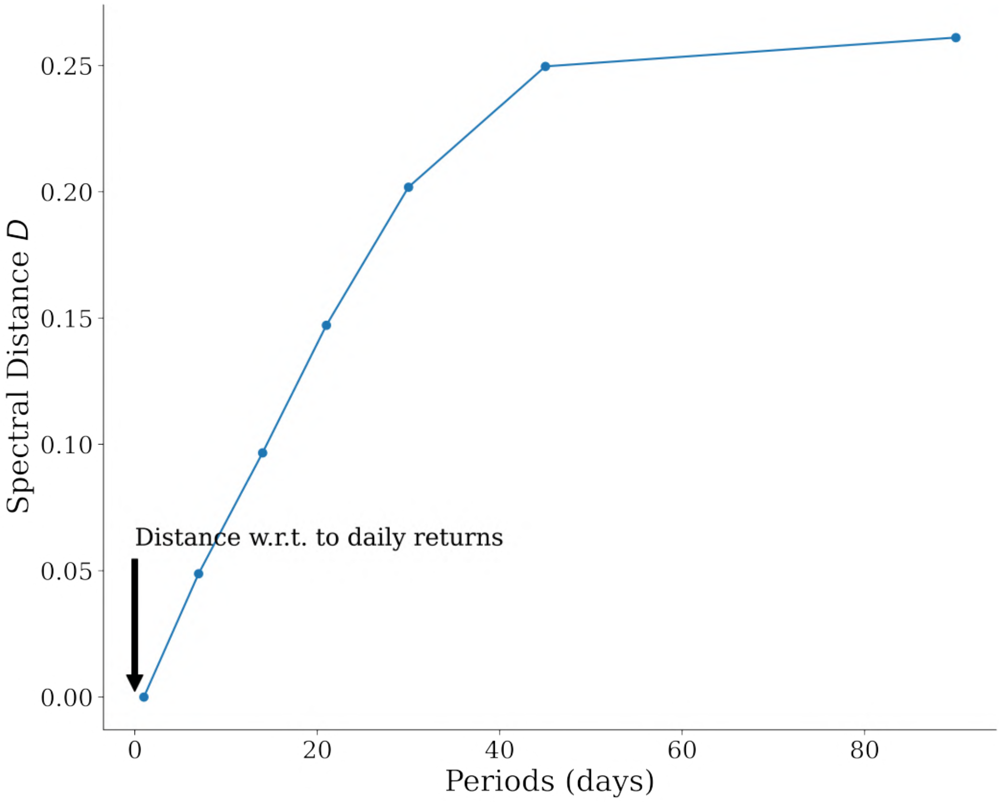


Figure 4.2: The spectral distance metric is computed for different periods concerning the daily returns eigenspace. For all the distances we took into account only the subspace composed by the top 10 eigenvectors.

In Figure 4.2, we present the distances between the first subspace associated with eigenvectors from the daily period and the subspaces linked to one week, two weeks, three weeks, monthly, one month and a half, and one-quarter periods, respectively. As expected, the distance is equal to 0 in the first case, indicating a complete overlap between the two eigenspaces.

Furthermore, we observe that the distance, although not exactly 0 as anticipated, remains remarkably stable and shows minimal variation when upsampling the returns for different periods. This finding supports our initial intuition, as a significant difference in the sectorial structure of the data due to changes in sampling would result in distinct eigenspaces, leading to non-overlapping subspaces.

The spectral distance metric provides further evidence that our previous assumptions hold true. The relatively stable distance between the different periods reinforces the idea that the sectorial structure of the data remains consistent regardless of the sampling frequency. Any major variation in the structure would be reflected in non-overlapping eigenspaces, which is not observed in our analysis.

4.2. Generative process for sales and prices

The second question we aim to explore is whether the sales and prices series share a common origin from the same generative process. To achieve this, we will develop specialized methods to create new synthetic time series for price returns while preserving certain characteristics of the original time series. Subsequently, we will compute the spectral distance between the generated time series and the original price time series.

The primary goal is to generate new time series for price returns that simulate possible different scenarios of sales, scenarios that have not been observed in reality. We will repeat this process multiple times, creating an ensemble of synthetic time series, and then measure the distance of this ensemble from the original data. This will result in a distribution of distances, which can be compared to the distance between the sales and the prices.

We assume that if the sales are indeed just a different realization of the prices, meaning they arise from the same generative process, we would expect that the distances between the sales and prices, after introducing a reasonable amount of randomness through the synthetic time series, should not significantly deviate from this distribution. In other words, the similarity between the sales and prices should be reflected in the proximity of their respective distances to the distribution of synthetic ensemble distances.

To achieve this, synthetic time series were created following two different methods:

1. **Sampling**: The first approach involves computing the distribution of the mode signals. Subsequently, new mode signals are sampled from this distribution, and a synthetic time series for the returns is obtained by multiplying these mode signals with the inverse of the eigenvectors matrix. This operation is the inverse transfor-

mation used to convert the returns into the eigenvectors basis.

2. **Shuffling** The second method involves shuffling the distribution of the mode signals. The new synthetic returns are then obtained by performing the inverse base transformation, which allows us to revert the shuffled mode signals back to the original basis, thus obtaining the synthetic returns.

In both scenarios, we recreate the synthetic time series by introducing an element of randomness into the realization of the mode signals. By doing so, we eliminate the potential autocorrelation effect in the mode signals while retaining the existing distribution of values.

The intention behind generating these synthetic time series is to ensure that they display a comparable correlation structure and retain certain statistical properties of the original data. However, it is important to acknowledge that these synthetic series may not fully capture specific patterns or intricate details that are not explained by the major modes of correlation represented by the eigenvectors. Nevertheless, our observations indicate that the correlation matrix between the generated time series closely resembles that of the original data.

Quantifying the extent of disparity between the synthetic series and the original data is not a straightforward task. An unanswered question is how much valuable information lies in the autocorrelation among mode signals. This aspect remains unresolved and merits further investigation.

4.2.1. Sampling from the distribution of the mode signals

The initial approach involves generating synthetic time series through a two-step process. Firstly, we calculate the distribution of values for the mode signals, and secondly, we sample new mode signals from this distribution. Using these newly sampled mode signals, we reconstruct the synthetic values of the returns by applying the inverse-base transformation, which allows us to transition back to the eigenvectors basis.

To provide a more comprehensive understanding of the method we explain it in a linear algebra fashion, furnishing more details and clarify all the procedures of the method.

1. **Returns time series**: Suppose we have a returns time series represented as a matrix, denoted by $\mathbf{X}$. Each column of $\mathbf{X}$ represents the returns of a specific asset or variable, and each row represents a different time period. So, $\mathbf{X}$ has dimensions $N \times M$, where N is the number of periods and M is the number of assets/variables.

2. **Sample correlation matrix**: We compute the sample correlation matrix, denoted by $\mathbf{C}$, by calculating the correlation coefficients between the returns of all pairs of assets. $\mathbf{C}$ is an $M \times M$ square matrix, where each entry $C(i,j)$ represents the correlation between the ith and jth assets.

3. **Eigendecomposition**: We perform the eigendecomposition of the correlation matrix $\mathbf{C}$, which gives us a set of eigenvectors and eigenvalues. Let's denote the eigenvectors matrix as $\mathbf{V}$ and the eigenvalues vector as λ. $\mathbf{V}$ is an $M \times M$ matrix, where each column represents an eigenvector, and λ is an M-dimensional vector containing the eigenvalues.

4. **Projection onto eigenspace**:We project the returns time series $\mathbf{X}$ onto the eigenspace spanned by the eigenvectors $\mathbf{V}$. This projection yields a new matrix $\mathbf{Y}$, which represents the coordinates of the returns time series in the eigenspace basis. $\mathbf{Y}$ is obtained by multiplying $\mathbf{X}$ with $\mathbf{V}$, i.e., $\mathbf{Y} = \mathbf{X} * \mathbf{V}$.

5. **Distribution of mode signals**: For each mode (eigenvalue-eigenvector pair), we sample the distribution of the values of the projection matrix $\mathbf{Y}$, i.e., we find the distribution of each column, i.e. each mode signal. Afterward, we reproduce each column sampling N values from the former distribution. This creates an element of randomness in the mode signals while maintaining the overall structure of the modes. Let's denote the sampled projection matrix as $\mathbf{Y_{sampled}}$.

6. **Generate synthetic time series**: To generate synthetic time series, we make the inverse operation by multiplying the sampled projection matrix $\mathbf{Y_{sampled}}$ to the inverse of the eigenvectors matrix $\mathbf{V}$. Let's denote the reconstructed matrix as $\mathbf{X_{synthetic}}$, which has the same dimensions as the original returns time series X. We, then, perform the matrix multiplication: $\mathbf{X_{synthetic}} = \mathbf{Y_{shuffled}} * \mathbf{V}^{-\mathbf{1}}$. [1]

7. **Recompute correlation matrix**: Finally, we recompute the correlation matrix for the synthetic time series $\mathbf{X_{synthetic}}$, obtaining a new correlation matrix $\mathbf{C_{synthetic}}$ from which we can compute the new eigenspace. The last one will be needed to compute the spectral distance.

To show the results of the method in practice we plot the returns of the company *'ASA Gold and Precious Metals'* and the corresponding synthetic returns obtained using the method in Figure 4.3.

Figure 4.4 displays the plot of the first mode (market mode) along with its corresponding

[1] Note that the eigenvectors are orthogonal, which allows us to reconstruct the synthetic time series by multiplying the shuffled projection matrix by the transposed eigenvectors. This assumes that the eigenvectors form an orthonormal basis.

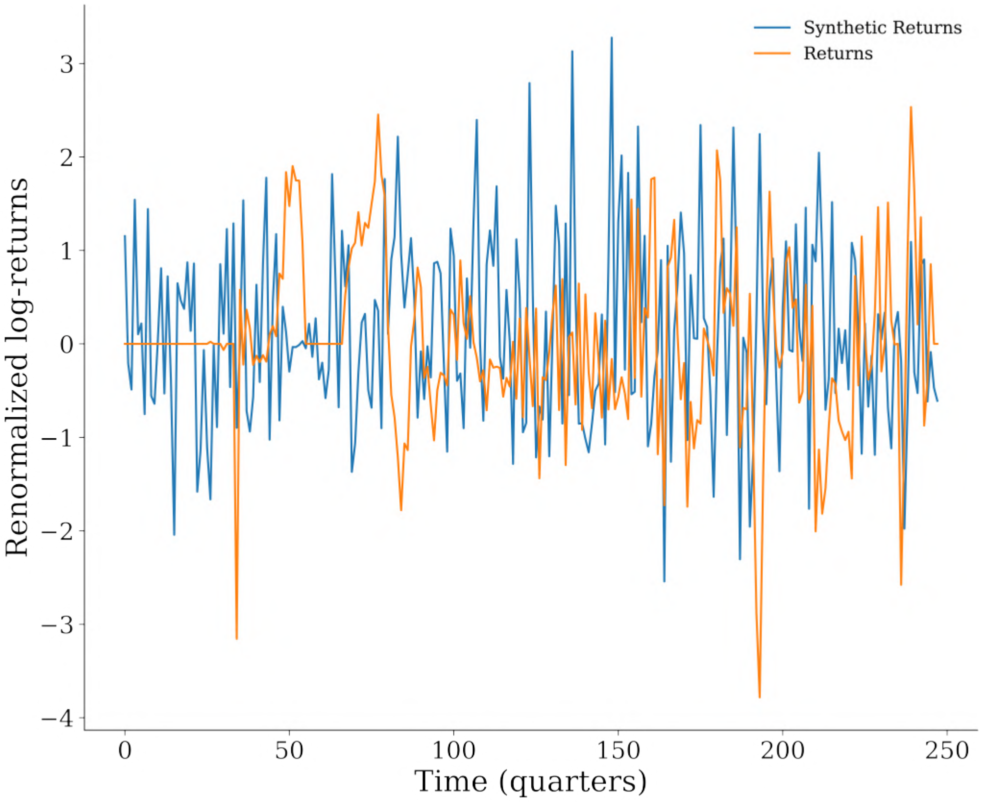
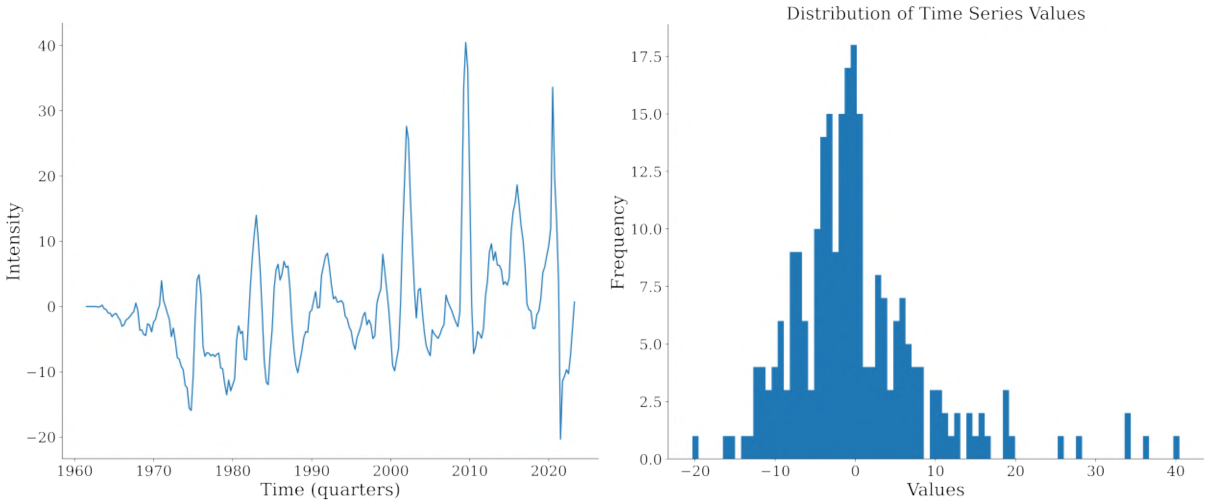


Figure 4.3: Synthetic returns vs returns of *'ASA Gold and Precious Metals'* (picked at random) over time, generated using the *Mode Signals' sampling method.*

distribution of values. We will utilize this distribution to sample new synthetic values for the mode. Similar distributions are obtained for the subsequent modes, and these will be employed to sample the subsequent synthetic mode signals.

Regarding the CompuStat data, the outcomes of our analysis have been rather unexpected. After generating 200 synthetic time series for the sales returns, we computed the distance between these synthetic series and the actual sales sample. This allowed us to compare the distribution of distances with the sole empirical distance we possess, which is the distance between sales and prices in the top eigenvectors subspace.

Surprisingly, the resulting spectral distance appears to deviate significantly from the bulk distribution, as illustrated in Figure 4.5. Although the distance falls far from the distribution, it is relatively small, indicating a substantial overlap between the two returns in the eigenspace sense. It is important to note that a distance of approximately $D \approx 0$ implies a high degree of overlap between the two eigenspaces. This result is quite controversial and deserves further analysis. This means that the two returns, in the eigenspace sense,



(a) Market mode (b) Market mode distribution

Figure 4.4: (i) Market mode and (ii) distribution of the market mode values from which we sample the synthetic mode signals.

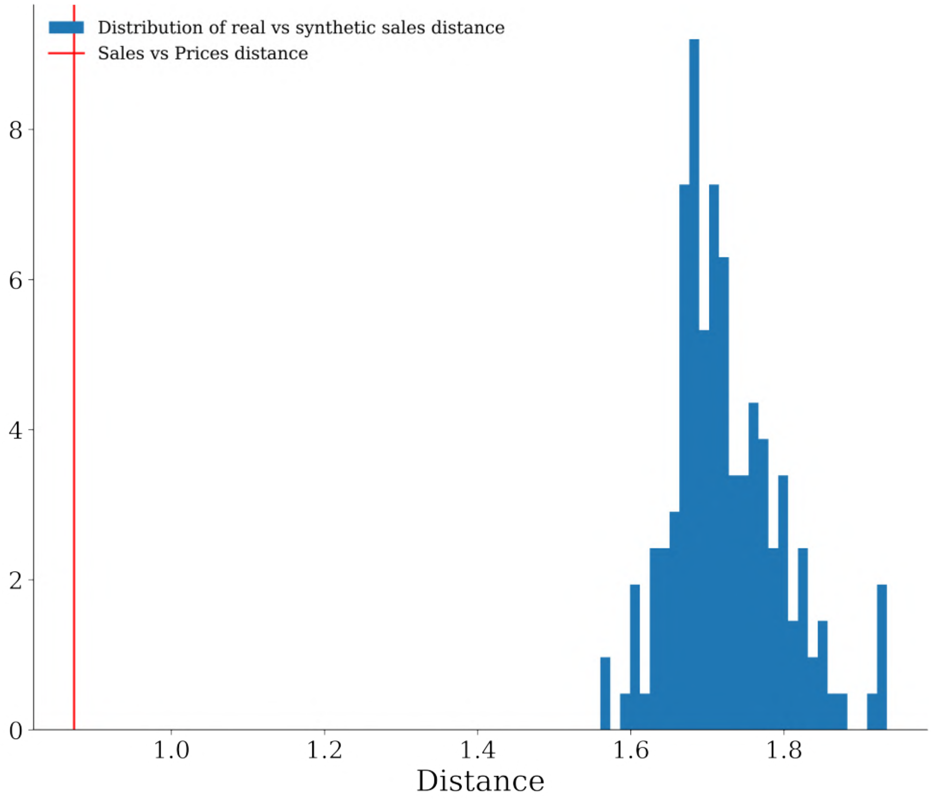


Figure 4.5: Distance between sales and synthetic sales returns, compared with sales-prices distance using the Spectral distance metric, considering the subspace generated by the top 10 eigenvectors.

overlap quite a lot, even more than the synthetic returns with themselves [2].

4.2.2. Shuffling the mode signals

The second method involves re-shuffling the components of the mode signals in our data. This introduces randomness into the process of generating synthetic time series. The overall procedure is quite similar to the first method. For the sake of completeness, let's outline the main steps here:

1. **Returns time series**: Suppose we have a returns time series represented as a matrix, denoted by $\mathbf{X}$. Each column of $\mathbf{X}$ represents the returns of a specific asset or variable, and each row represents a different time period. So, $\mathbf{X}$ has dimensions NxM, where N is the number of time periods and M is the number of assets/variables.

2. **Sample correlation matrix**: We compute the sample correlation matrix, denoted by $\mathbf{C}$, by calculating the correlation coefficients between the returns of all pairs of assets. $\mathbf{C}$ is an MxM square matrix, where each entry $C(i, j)$ represents the correlation between the ith and jth assets.

3. **Eigendecomposition**: We perform the eigendecomposition of the correlation matrix $\mathbf{C}$, which gives us a set of eigenvectors and eigenvalues. Let's denote the eigenvectors matrix as $\mathbf{V}$ and the eigenvalues vector as λ. $\mathbf{V}$ is an MxM matrix, where each column represents an eigenvector, and λ is an M-dimensional vector containing the eigenvalues.

4. **Projection onto eigenspace**: We project the returns time series $\mathbf{X}$ onto the eigenspace spanned by the eigenvectors $\mathbf{V}$. This projection yields a new matrix $\mathbf{Y}$, which represents the coordinates of the returns time series in the eigenspace basis. $\mathbf{Y}$ is obtained by multiplying $\mathbf{X}$ with $\mathbf{V}$, i.e., $\mathbf{Y} = \mathbf{X} * \mathbf{V}$.

5. **Shuffle projection components**: For each mode (eigenvalue-eigenvector pair), we shuffle the components of the projection matrix $\mathbf{Y}$, i.e., we change the order of the rows randomly. This creates an element of randomness in the mode signals while maintaining the overall structure of the modes. Let's denote the shuffled projection matrix as $\mathbf{Y}_{\mathbf{shuffled}}$.

6. **Generate synthetic time series**: To generate synthetic time series, we make

[2] We note that even using eigenspaces that include more eigenvectors we obtain similar results, following our initial intuition that it's not important to include more than the first, say, ten eigenvectors. By the way, this fact doesn't surprise us, since most of the information, as we have repeated hundreds of times is contained in the first handful of eigenvectors.

the inverse operation by multiplying the shuffled projection matrix $\mathbf{Y_{shuffled}}$ to the inverse of the eigenvectors matrix $\mathbf{V}$. Let's denote the reconstructed matrix as $\mathbf{X_{synthetic}}$, which has the same dimensions as the original returns time series X. We, then, perform the matrix multiplication: $\mathbf{X_{synthetic}} = \mathbf{Y_{shuffled}} * \mathbf{V^{-1}}$.

7. **Recompute correlation matrix**: Finally, we recompute the correlation matrix for the synthetic time series $\mathbf{X_{synthetic}}$, obtaining a new correlation matrix $\mathbf{C_{synthetic}}$ from which we can compute the new eigenspace. The last one will be needed to compute the spectral distance.

To keep the thesis concise, we will present only the final results obtained in this case, without including all the intermediate steps. Specifically, we will display the final distribution to draw our conclusions and summarize the key insights inferred from these results.

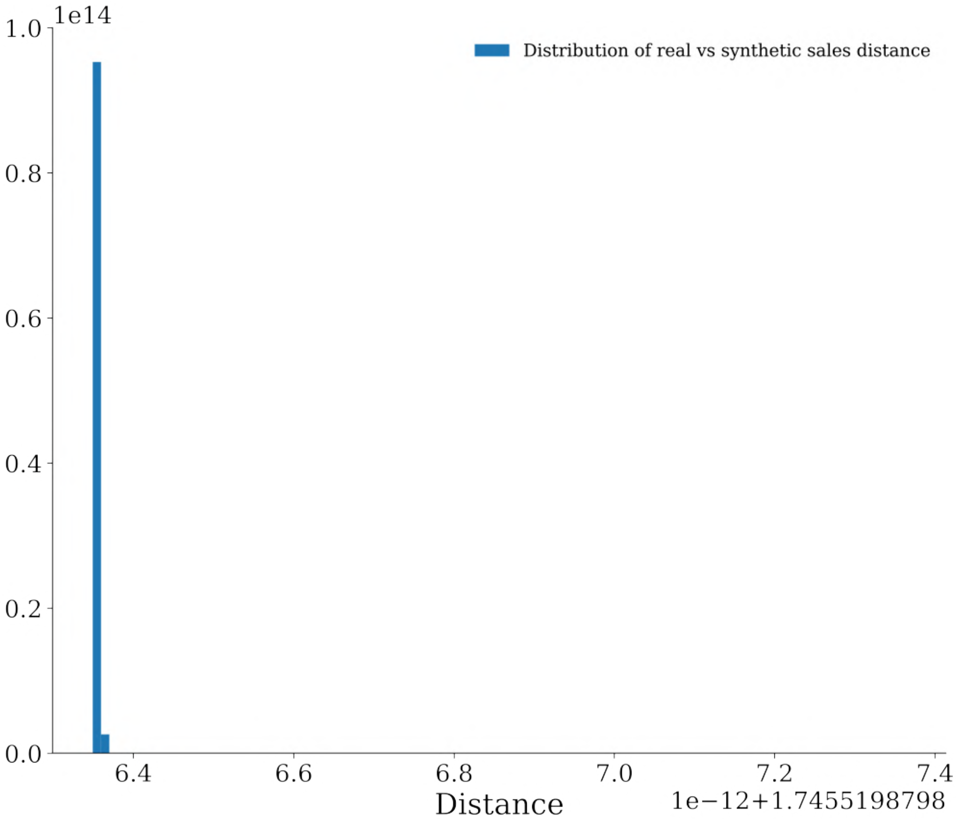


Figure 4.6: Distance between sales and synthetic sales returns, compared with sales-prices distance using the spectral distance metric, considering the subspace generated by the top 10 eigenvectors.

The results obtained from both methods were initially expected to be quite similar due to their similarity. However, as evident in Figure 4.6, we observe a striking contrast in the distribution. In this case, the distribution of distances is significantly confined close to 0, indicating that the distance between the original sales and prices (approximately ≈ 0.3) deviates considerably from this distribution. This confirms our hypothesis that the two samples cannot be attributed to the same generative process.

Unsatisfied with the initial findings, we attempted to reproduce the results by shuffling in blocks instead of shuffling all the rows, which preserved some level of autocorrelation in the mode signals. Remarkably, the results remained comparable for different block sizes.

A plausible explanation for this phenomenon is that by conserving more information about the original data, particularly preserving some autocorrelation structure and retaining the same values obtained for the signal modes instead of resampling them from the distribution, the spectral distance diminishes to zero. This suggests that the inherent structure and patterns in the mode signals hold more information than anticipated, and the two original samples differ to such an extent that they cannot be explained by the same generative process when considered together.

It's important to note that this idea is merely an attempt to illustrate this result, and we remain open to exploring new ideas to test the underlying generative process further. We acknowledge that these findings do not yet serve as conclusive evidence and further research and investigation are required to strengthen our understanding of this phenomenon.

5 | Lead-lag effects

In this chapter, we intend to examine the lead-lag effects observed among different companies. The goal is to provide a coherent perspective on the complete temporal correlation dynamics within a network of interconnected variables, public companies in our case. How can we approach this general problem? In other words, how can we gain insights into the network dynamics at different points in time? How do changes in one node simultaneously and differentially influence other nodes? Our primary focus is to address these questions, along with several related ones.

As illustrated in preceding chapters, economic systems are composed of numerous units that interact with each other in various ways and on varying scales. As a result, the dynamics within these systems are intricate. Employing appropriate statistical methodologies becomes imperative for grasping the holistic dynamics of the economic system. We have already explored different methods to study the pairwise relationships between assets, trying to estimate the strength of mutual interaction.

The majority of these research studies primarily focus on analyzing simultaneous relationships between the returns of the assets. This usually comes from the importance of the efficient market hypothesis (EMH) in the literature [26]. In a few words, the hypothesis asserts that the existing market price of any traded asset encapsulates all relevant information, and any attempt to foresee its future development would lack profitability unless increasing the associated risk [12]. The idea is that these results are only valid when the time it takes for news or information to spread across the financial market is relatively short compared to the period during which the logarithmic returns are being calculated. Therefore, if accepted, this hypothesis is a double-edged sword. On one hand, it validates the rationale behind considering log returns on a daily or longer time scale. On the other hand, it implies that investigating lead-lag effects on low-frequency data is unlikely to bring actionable insights. Nevertheless, a substantial body of literature delves into the EMH, and as anticipated, not all scholars unquestionably embrace this hypothesis. Without delving into more details, we provide a few references for a more in-depth exploration of the topic [15, 41].

Being formal, we can extend the formalism precedently used and define the lagged correlation matrices, namely

$$C_{ij}(\tau) = \mathbb{E}\left[\widetilde{g}_i(t)\widetilde{g}_j(t+\tau)\right] \tag{5.1}$$

Where $\widetilde{g}_i(t)$ is still the rescaled growth rate. When $\tau = 0$ we recover the well-known correlation matrix. This notation designates the initial index as the *leader* and the second index as the *lagger*.

It's important to note that, in most cases, lagged covariances do not exhibit commutative properties, in contrast to the usual zero-lag correlation matrix.

$$C_{ij}(\tau) \neq C_{ji}(\tau)$$

In the general case where $\tau \neq 0$ we can generalize what we have done with the same-time correlation matrix by employing the **singular values decomposition (SVD)**. However, as we will soon discover, this approach has limitations in capturing the lagged correlations present within the multivariate time series.

Apart from a few exceptions the correlation among different times has not been extensively examined in the literature [23, 27, 37]. Prior research on lead-lag relationships has predominantly centered on the dynamics of stock prices or market indexes, often neglecting various other types of metrics such as inventories or sales, and their comparisons with stock prices. This chapter enters the picture by trying to give a more exhaust explanation of where the conventional method based on studying the lead-lag correlation matrix falls short, highlighting its limitations, and proposing alternative methodologies taken from climate science. Bridging this gap is a challenging task. Even though we are still uncertain of the existence of *arbitrage*, we believe that a broader view of the complex lagged time dynamics can be helpful for at least two reasons: firstly, it can help market players in the decision-making process and secondly, it can aid in identifying significant variables for factor models.

The chapter is structured into two sections: The first section delves into the analysis of the lagged correlation matrix, exploring the singular value decomposition of this matrix and explaining its peculiar interpretation in this context. Moving on to the most significant lag times, we aim to address the following questions: Can we identify specific lags τ that optimize the lead-lag relationship? Can this knowledge be used for forecasting firm sales growth? Finally, the second section provides an exploration of the application of HPCA (Hilbert PCA), a technique borrowed from climate science, to economic systems.

5.1. Lagged correlation matrix

The lagged correlation matrix serves to quantify the connection between the returns of different companies considered at different times. In particular, each element of the matrix quantifies the intensity of the lagged correlation of two companies, with one of the stocks lagging the other by a certain time lag τ. When a statistically significant correlation emerges, it suggests that the earlier returns of one entity might be impacting the other.

We computed these lagged correlations across all pairs within the dataset and condensed them into a corresponding $N \times N$ correlation matrix. Since time intervals for our prices and sales data are known to be misaligned in many cases, a noteworthy number of missing values (or NaN) terms are evident within the lagged correlation matrix. We replaced them with zeros. As anticipated, the lagged correlation matrix presents asymmetry due to the dissimilarities in the influence of lag correlations exerted in each direction. For example, the fact that a company leads another one by 3 months does not imply the opposite (this fact is well represented by the non-commutativity property of the lagged correlation matrix).

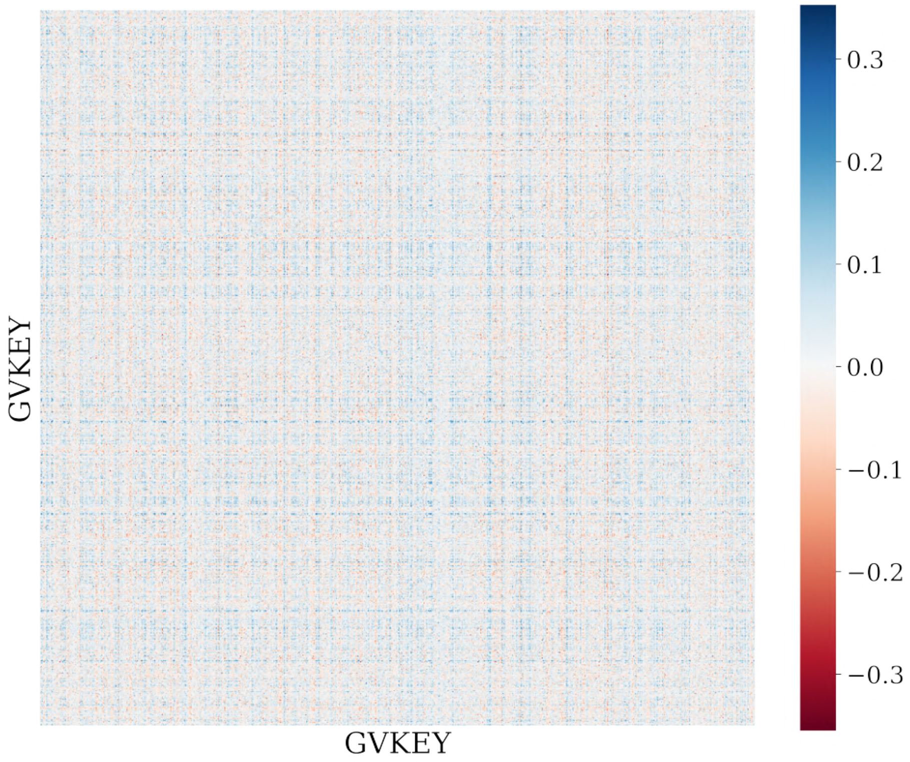


Figure 5.1: Lagged correlation matrix for prices (leading) and sales (lagging) using $\tau = 6$ months.

In Figure 5.1 we observe, as an example, the results obtained for the lagged correlation matrix with lagged time $\tau = 6$ months between, respectively, prices and sales for the 248 companies in the Computstat dataset. In this case, the *leading variable* is prices and *lagging variable* is sales. However, as we said before, it's quite hard to extract meaningful collective information about the dynamics of the system from this matrix. Indeed, to get more insights we should compute the matrix for different τ and see how it changes. This task is computationally expensive.

To uncover the signal, it is worth computing the Singular Value Decomposition for this lagged correlation matrix. As we will see this has a precise interpretation when studying different interrelated quantities, such as prices and sales.

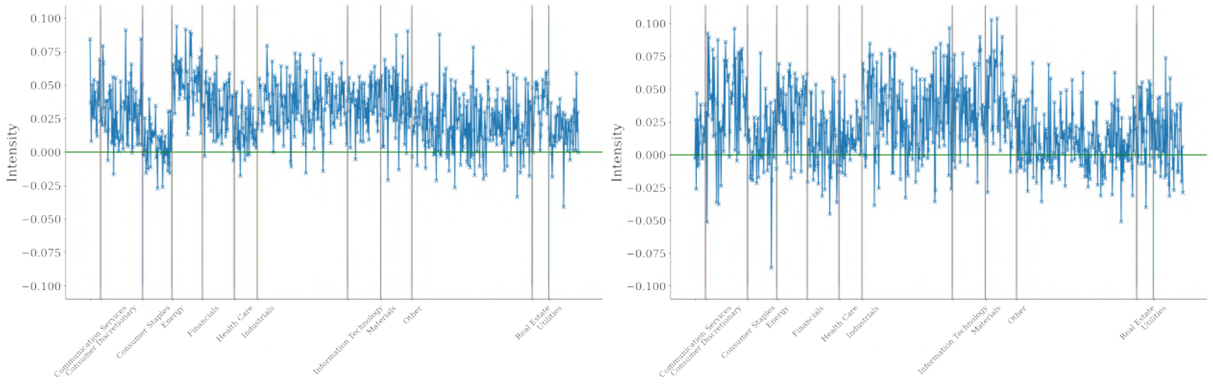
Singular Value Decomposition (SVD) The SVD is suitable to address the question of whether the lagged matrix contains preferential information directions. SVD is a factorization that generalizes the eigen-decomposition (PCA) to matrices that are not symmetrical and definite-positive. The Singular Value Decomposition (SVD) is a matrix factorization technique that decomposes a matrix $\mathbf{X}$ into three matrices: $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}^T$, where $\mathbf{X}$ is an $m \times n$ matrix, $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with non-negative real numbers called singular values on the diagonal, and $\mathbf{V}$ is an $n \times n$ orthogonal matrix.

Mathematically, the SVD of a matrix $\mathbf{X}$ is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

When working with a lagged correlation matrix, a nice theoretical interpretation can be applied to both the left and right singular vectors. The left singular vectors can be understood as representing optimal linear combinations of the initial variables – in this context, stock prices - chosen to effectively capture the underlying signal present in the first singular value. On the other hand, the right singular vectors represent the linear combinations of the lagged variables – in this context, representing the sales of the same companies. These combinations are designed to best explain the lagged correlation associated with a specific time lag τ. In summary, the two linear combinations form new balanced variables for which the lag is maximum. The first linear combination represents the *leading* variable (or set of firms) while the second linear combination represents the *lagging* variable.

In Figure 5.2 we can observe the first left and right singular vectors for the six months lagged correlation matrix computed before.



(a) First left singular vector

(b) First right singular vector

Figure 5.2: First left (i) and right (ii) singular vectors for the lagged correlation matrix for prices (leading) and sales (lagging) using $\tau = 6$ months. The variables are represented in a sectorial order

The subsequent singular value decomposition, as explained before, can also shed light on the lead-lag relationship between prices and sales of specific firms. In other words, using the former interpretation it is easy to see if the lagged correlation is due purely to same-firm correlations, i.e. a firm that lags with itself and not with others. If this relationship exists, indicating that the primary driver of the correlation between sales and returns is the interaction within the same firm, with negligible cross-firm effects, we would expect significant overlap between the left and right singular vectors corresponding to the largest singular value. However, in our case, this is just partially true. Even though the two patterns are noisy and do not perfectly overlap, we can still notice that some sectorial clustering patterns hold. We want to systematize this by plotting the scatter plot of the first singular vectors.

As we can observe in the scatting plot shown in Figure 5.3, computing the linear regression of these two variables, a linear dependence emerges, meaning that there's a linear relationship between the two singular vectors. In conclusion, even though sales and price dynamics are different, it is still possible to notice lead-lag interactions within same-firm prices and sales. Wherein same-firm prices anticipate sales.

For what concerns the lagged correlation matrix, many questions about lead-lag relationships remain unsolved. How can we understand in an economic system which companies, or in general variables, are leading and which companies are lagging? To be honest, we could consider the companies for which the left (and right) singular vector intensity is nonnegligible, but this procedure can be fuzzy.

While this question cannot be easily answered within this framework, we can still infer

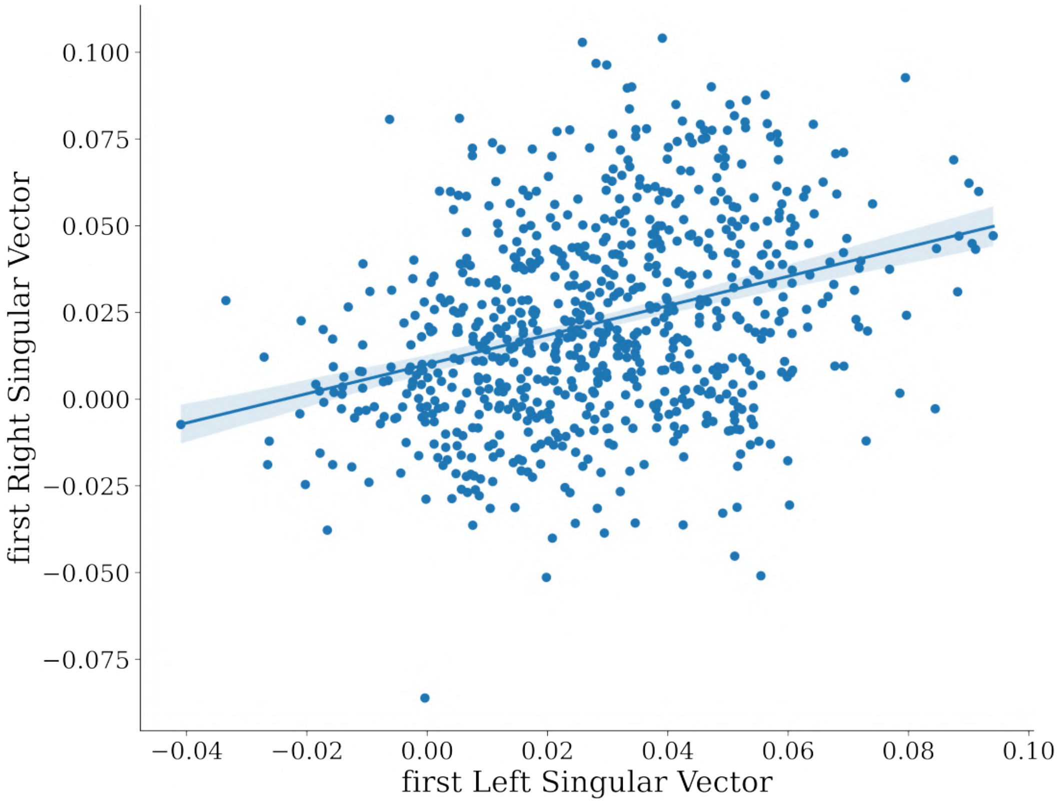


Figure 5.3: Scatter plot of the first left and right singular vectors. The plotted line represents the linear regression operated between the two variables.

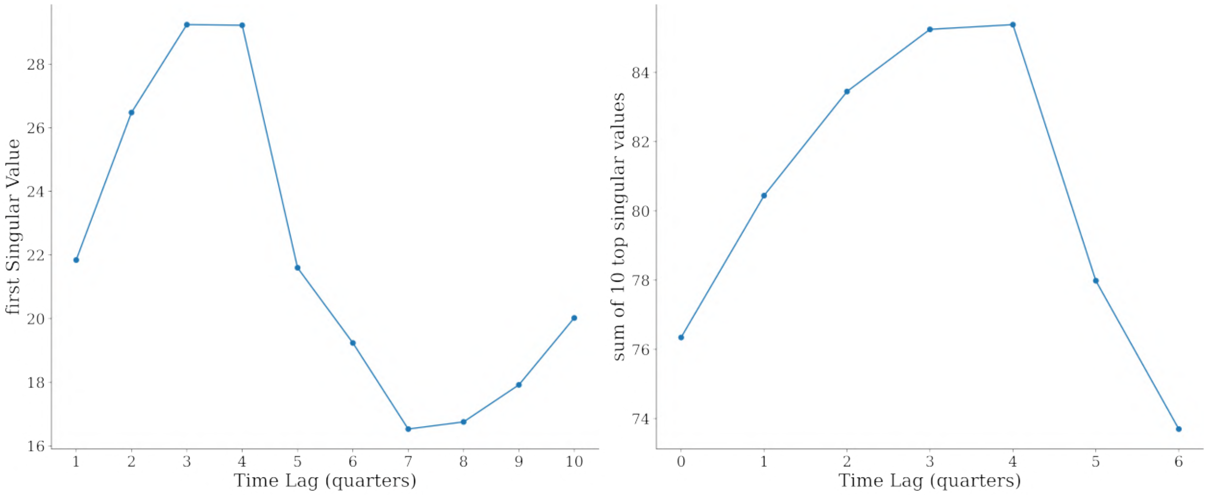
more about what are the lagged times for which the lead-lag effects are more visible, i.e. we can find relevant values of τ for the system, that deserve deeper examinations.

5.1.1. How to spot time lags that maximize lead-lag relationships?

In the following, we will approach the problem of identifying the lags for which the lead-lag effects in the dataset are maximized. We will distinguish between two different approaches. The first one is based on studying how the singular values change varying the time lag. Secondly, we focus just on the same-firm lead-lag effects ignoring the cross-firm effects. We, then, study the variation of the averaged autocorrelation function at different lag times.

Singular Values as a measure of lead-lag intensity with lag τ

The SVD can be used to study the maximal lead-lag relationship between different assets. To do this, we can look at the largest singular value, $s_{max(\tau)}$, or the sum of the first, say, 10 singular values, and study how it behaves for different lag times τ. The largest singular value represents the strongest correlation between the assets, i.e. a large value of s means that there's a lagged correlation pattern between the variables and the value of τ at which it occurs gives us an idea of a representative time lag for the collective motion.



(a) Top singular value

(b) Sum of 10 top singular values

Figure 5.4: (i) Top singular value and (ii) sum of 10 top singular values of the lagged correlation matrix between prices and sales computed for different time lags τ

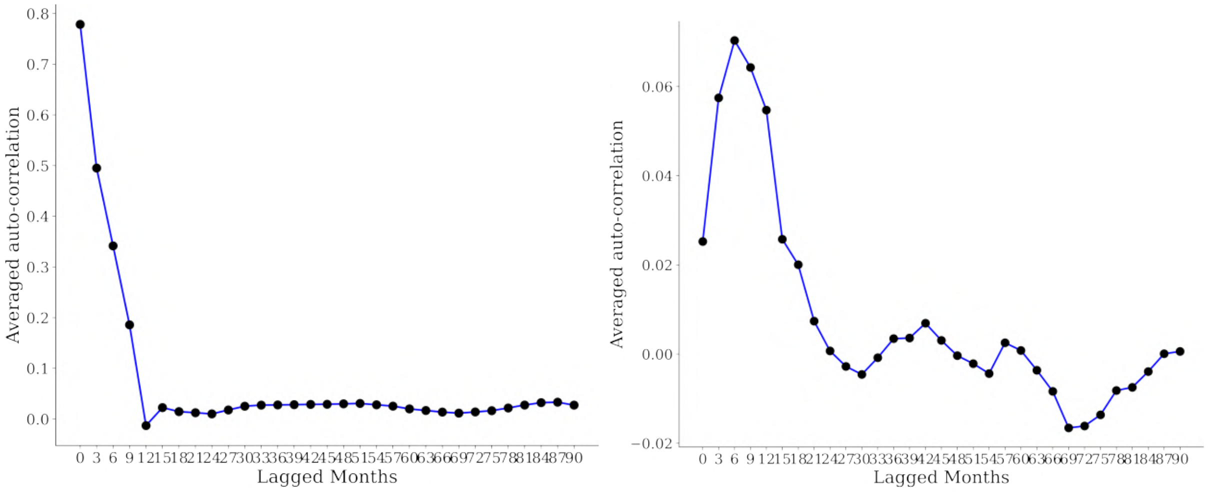
In Figure 5.4 we show how the top singular value and the sum of the 10 top singular values, respectively, behave for the lagged correlation matrix between prices and sales, varying the time lag τ. We expect it to be maximal when the collective lead-lag correlation is maximum. The fact that the data are sampled quarterly, as mentioned in the preceding chapters, also in this case, does not help since we have a grained picture of the system dynamic.

We can easily see that there's not much difference in simply considering the top singular value instead of the 10 largest singular values. In both cases, there's a visible peak between 9 and 12 months of lag time. This result takes into account both the same firm lagged correlations and the cross-firm effects. This quantity takes into account, possibly, complex linear combinations of firms that are correlated to other linear combinations of the same firms considered at a lagged time τ. This can result, in principle, in correlation patterns that include different firms (to be more precise, same firms with different intensities). The averaged auto-correlation function (studied in the next paragraph), instead, considers only same-firm lead-lag effects ignoring the cross-firm effects, that are not taken directly into account when looking for the optimal τ.

Averaged auto-correlation function as a measure of lead-lag intensity

The averaged auto-correlation function between two variables $x_i(t)$ and $y_i(t)$ is defined as $E[x_i(t)y_i(t+T)]_{i,t}$ where the average runs over i and t. We expect it to be a good measure for the same-firm system correlation given a certain lag time T. Indeed, for each company it computes the lagged correlation coefficient between a single variable, say returns, or two variables, say prices and sales, and it computes the average over all the companies. In this procedure, cross-correlation is not taken into account. We try now to compute the averaged auto-correlation function in two different cases:

- The averaged auto-correlation function between prices $x_i(t)$ and sales $g_i(t)$, i.e. $E[x_i(t)g_i(t+T)]_{i,t}$.
- The averaged auto-correlation function between prices $x_i(t)$ and prices (or the opposite) $x_i(t)$, i.e. $E[x_i(t)x_i(t+T)]_{i,t}$.



(a) Averaged auto-correlation among sales for different values of τ

(b) Averaged auto-correlation between prices and sales for different values of τ

Figure 5.5: The figure shows the averaged auto-correlation for different values of τ in the case of (i) sales-sales and (ii) prices-sales. It is worth noticing that a peak appears for $\tau = 6$ months remarking that a clear same firm prices-sales lead-lag relation exists at that lagged time.

In figure 5.5 we show the results. As expected in the first case, the auto-correlation decays quite fast, indicating that between the same firm sales, there's no lagged information. Essentially, this implies that the current sales value doesn't have a predictive effect on the future sales value of the same company. Instead, there's a clear lagged relation between prices and sales within the same firm. Indeed a clear peak emerges for $\tau = 6$ months, indicating a notable same-firm lead-lag relation. By the way, this fact does not contradict the Efficient Market Hypothesis, since there's no clear *arbitrage* opportunity coming from this information[1]. Apart from that, we are satisfied to have found, through statistical methods, this interesting same-firm structure between prices and sales.

5.2. Hilbert Principal Component Analysis: nonconventional approach from climate science

Finally, we intend to explore the connections between the former approach and other lead-lag studies employed in fields such as climate science. Specifically, we aim to compare and contrast our methodologies with a technique called Hilbert Complex Principal Component Analysis (HPCA). In this section, we will present a novel approach called Hilbert Principal Components, which is nothing but the extension of the conventional PCA to a complex space, operated through a Hilbert transform of the original multivariate time series. CHPCA enables us to extract significant comovements with a time lead/delay in the data. After having described the method, its upsides and limitations, we apply the former to our prices and sales data.

This method was originally developed in the context of climate science [11, 34, 55] and first applied to economics in [5, 6] to untangle temporal comovements among variables observed in the macro-economic systems. [62]

We believe that the HPCA approach is beneficial for, at least, two reasons:

- The results are easily interpreted on the complex plane corresponding to each eigenmode,
- It's a compact framework for exploratory analyses. It can used to generate hypotheses on possible causalities among a huge number of variables.

Indeed, when dealing with real-world multivariate time series we have to keep into account both information on the behavior of individual time series and the inter-correlations between the single units. In this section, we are interested in untangling the intercorrelations between price, and sales.

As a result of direct and indirect causal relationships, we observe comovement in the data set involving time lead and delay. Following the main goal of the chapter, we aim to set up a methodology suitable for detecting relationships with time delay.

[1] Had the situation been inverse (sales leading prices) we would have become all rich.

5.2.1. The method

Let us suppose that we have N different time series $x_i(t)$, $i = 1, \cdots, N$ of length T, $t = 1, \cdots, T$. We first derive complex time series $\omega_i(t)$ from $x_i(t)$ obtaining the analytical signal,

$$\omega_i(t) = x_i(t) + iy_i(t),$$

where the imaginary part $y_i(t)$ is Hilbert transform of $x_i(t)$ defined by

$$y_i(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x_i(z)}{t-z}\,\mathrm{d}z,$$

where integration over z is Cauchy's principal integration.

The relative complex correlation matrix $\tilde{\boldsymbol{C}}$ is:

$$\tilde{\boldsymbol{C}} = \frac{1}{T}\boldsymbol{\Omega}\boldsymbol{\Omega}^{\dagger},$$

where $\boldsymbol{\Omega}$ represents the $N \times T$ time series matrix and $\boldsymbol{\Omega}^{\dagger}$ is Hermite conjugate of $\boldsymbol{\Omega}$.

We, now, solve the eigenvalue problem for $\tilde{\boldsymbol{C}}$. The larger the eigenvalue, the more important the presence of the eigenmode. The correlation matrix $\tilde{\boldsymbol{C}}$ can now be expressed in terms of its eigenvalues and eigenvectors as

$$\tilde{\boldsymbol{C}} = \sum_{\ell=1}^{N} \lambda_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^{\dagger},$$

where λ_ℓ and $\boldsymbol{v}_\ell$ are the ℓ-th eigenvalue and its associated eigenvector, respectively, and, as always, we order the eigenvalues in descending order, that is, $\lambda_1 > \lambda_2 > \cdots > \lambda_N$.

Since $\tilde{\boldsymbol{C}}$ is a Hermitian matrix:

- Eigenvalues are real and definite positive,
- Eigenvectors are complex.

We can now perform a change of basis of the multivariate time series using the eigenvectors $\boldsymbol{\alpha}_\ell$ as an orthonormal complete basis set:

$$\boldsymbol{\omega}(t) = \sum_{i=1}^{N} \omega_i(t)\boldsymbol{e}_i = \sum_{\ell=1}^{N} a_\ell(t)\boldsymbol{\alpha}_\ell,$$

with,

$$a_\ell(t) = \boldsymbol{\alpha}_\ell^\dagger \cdot \boldsymbol{\omega}(t).$$

As in the conventional PCA case, we call the coefficient $a_\ell(t)$ *mode signal* of the ℓ-th eigenmode. The mode signals represent the temporal behavior of the eigenmodes.

A few interesting properties concerning the eigenvalues λ_l can be easily proven, such as:

$$\sum_{l=1}^{N} \lambda_l = N \tag{5.2}$$

$$\sum_{t=1}^{T} |a_\ell(t)|^2 = \lambda_l \tag{5.3}$$

5.2.2. Playing with a toy-model

To better understand the approach, we will apply it through a straightforward analytical example. Consider a scenario where we have two cosine functions characterized by the same frequency ω but varying phases ϕ_1 and ϕ_2. This situation can be easily solved using analytical methods. In the subsequent steps, we will calculate the analytical signal, the correlation matrix, and perform the eigendecomposition analytically.

HPCA for analytical toy-models *Let us consider two variables. The first variable $s_1(t)$ is:* $s_1(t) = \cos(\omega t)$

$$\hat{s}_1(t) = \cos\left(\omega t - \frac{\pi}{2}\right) = \sin(\omega t), \tag{5.4}$$

$$s_1^a(t) = s(t) + j\hat{s}(t) = \cos(\omega t) + j\sin(\omega t) = e^{j\omega t} \tag{5.5}$$

For what concern the second variable $s_2(t)$:

$$s_2(t) = \cos(\omega t + \theta) = \frac{1}{2}\left(e^{j(\omega t+\theta)} + e^{-j(\omega t+\theta)}\right)$$

$$s_2^a(t) = \begin{cases} e^{j(\omega t+\theta)} & = e^{j|\omega|t} \cdot e^{j\theta}, \quad \text{if } \omega > 0, \\ e^{-j(\omega t+\theta)} & = e^{j|\omega|t} \cdot e^{-j\theta}, \quad \text{if } \omega < 0. \end{cases}$$

Now, we can compute the correlation by taking the expectation value of the product of s_1 and the complex conjugate of s_2, denoted as s_2^*.

$$C_{1,2} = \langle s_1(t)s_2^*(t)\rangle$$

Similarly, we can compute the correlation for s_1s_1 and s_2s_2 by substituting the appropriate expressions. To find the eigenvectors and eigenvalues, we solve the equation:

$$(\mathbf{C} - \lambda \mathbf{I})\mathbf{x} = 0$$

where $\mathbf{C}$ is the correlation matrix, λ is the eigenvalue, $\mathbf{I}$ is the identity matrix, and $\mathbf{x}$ is the eigenvector. First, let's find the eigenvalues:

$$|\mathbf{C} - \lambda \mathbf{I}| = \begin{vmatrix} 1-\lambda & e^{-j\theta} \\ e^{j\theta} & 1-\lambda \end{vmatrix}$$

Simplifying, we find two possible eigenvalues: 1)

$$\lambda_1 = 2,$$

2)

$$\lambda_2 = 0.$$

Next, let's find the eigenvectors associated with each eigenvalue. The eigenvector for $\lambda_1 = 2$ is:

$$\mathbf{x}_1 = \begin{bmatrix} x_1 \\ e^{-j\theta} x_1 \end{bmatrix}, \textit{ where } x_1 \neq 0 \textit{ and } e^{j\theta} \neq 1$$

While, for $\lambda_2 = 0$ we get:

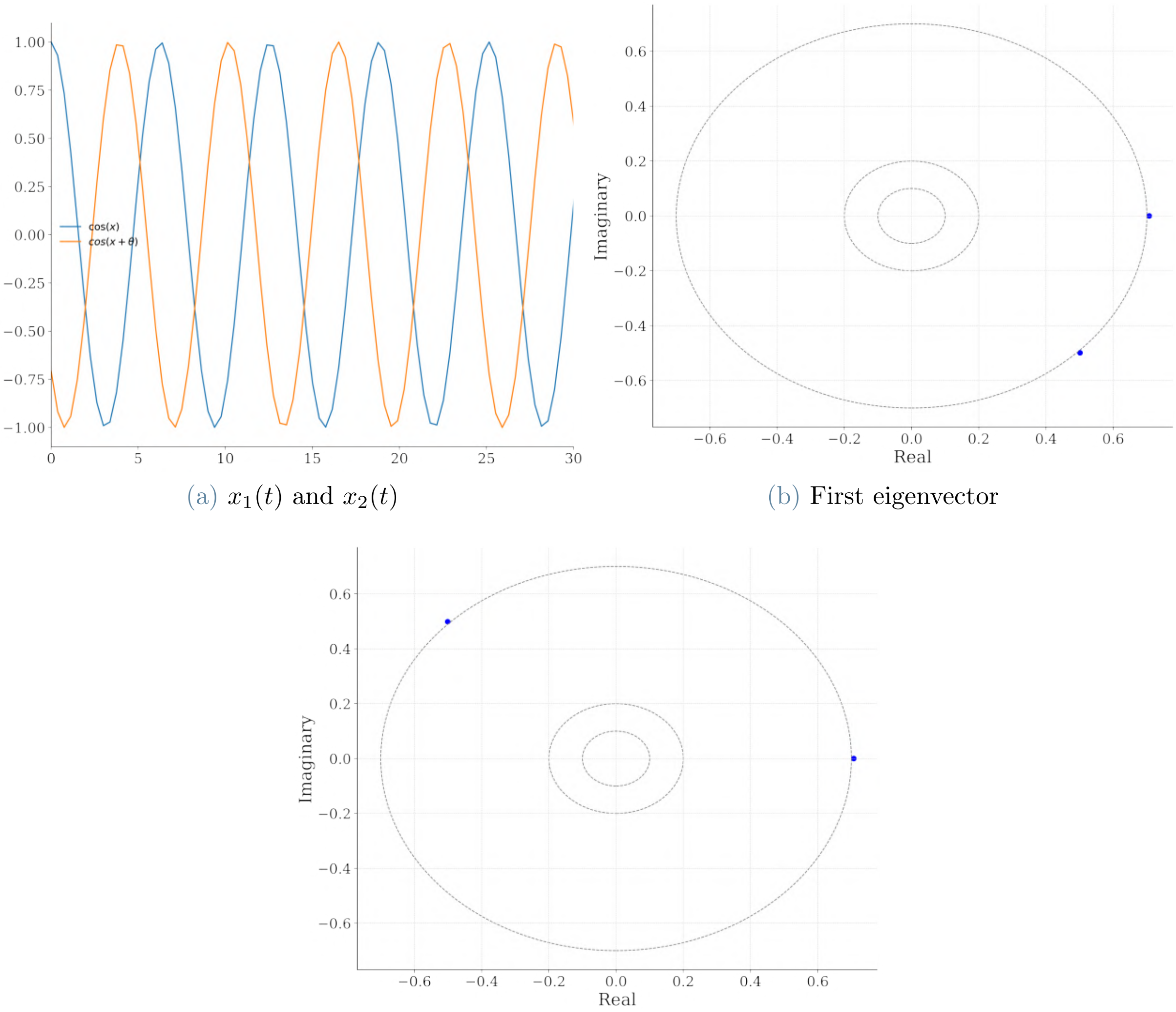
$$\mathbf{x}_2 = \begin{bmatrix} x_1 \\ -e^{-j\theta} x_1 \end{bmatrix}, \textit{ where } x_1 \neq 0 \textit{ and } e^{j\theta} \neq -1$$

As demonstrated, the complex phase of both eigenvectors directly reflects the introduced phase shift. To enhance the visualization, we executed the procedure computationally and plotted the outcomes.

The results are shown in Figure 5.6. Both eigenvectors give us a clear representation of the phase shift of the second variable. Indeed, in our case, where have taken $\theta = 3/4\pi$, the phase shift is well incorporated in the second eigenvector. The representation in the complex plane clearly shows the phase relation between the two cosines. As we can see in the plot of the second eigenvector the lead-lag relationship appears clearly.

We can extend these results to a series of shifted harmonic functions. Rather than restricting ourselves to only two functions, we expand our analysis to include a collection of seven distinct shifted cosine functions.

Figure 5.7 showcases the outcomes of this endeavor. Once again, the lead-lag relationship with precise phase shifts is evident. We have, indeed, examined a phase shift of $\theta_i = i \times \frac{1}{7}\pi$ for $i = 1, \ldots, 7$. Both the first and second eigenvectors reveal evenly spaced phases, demonstrating a consistent and direct relationship between the phase shifts and the phases in the complex plane, maintaining a 1:1 ratio. This correspondence can be confirmed by visualizing the phase distribution in (iv).

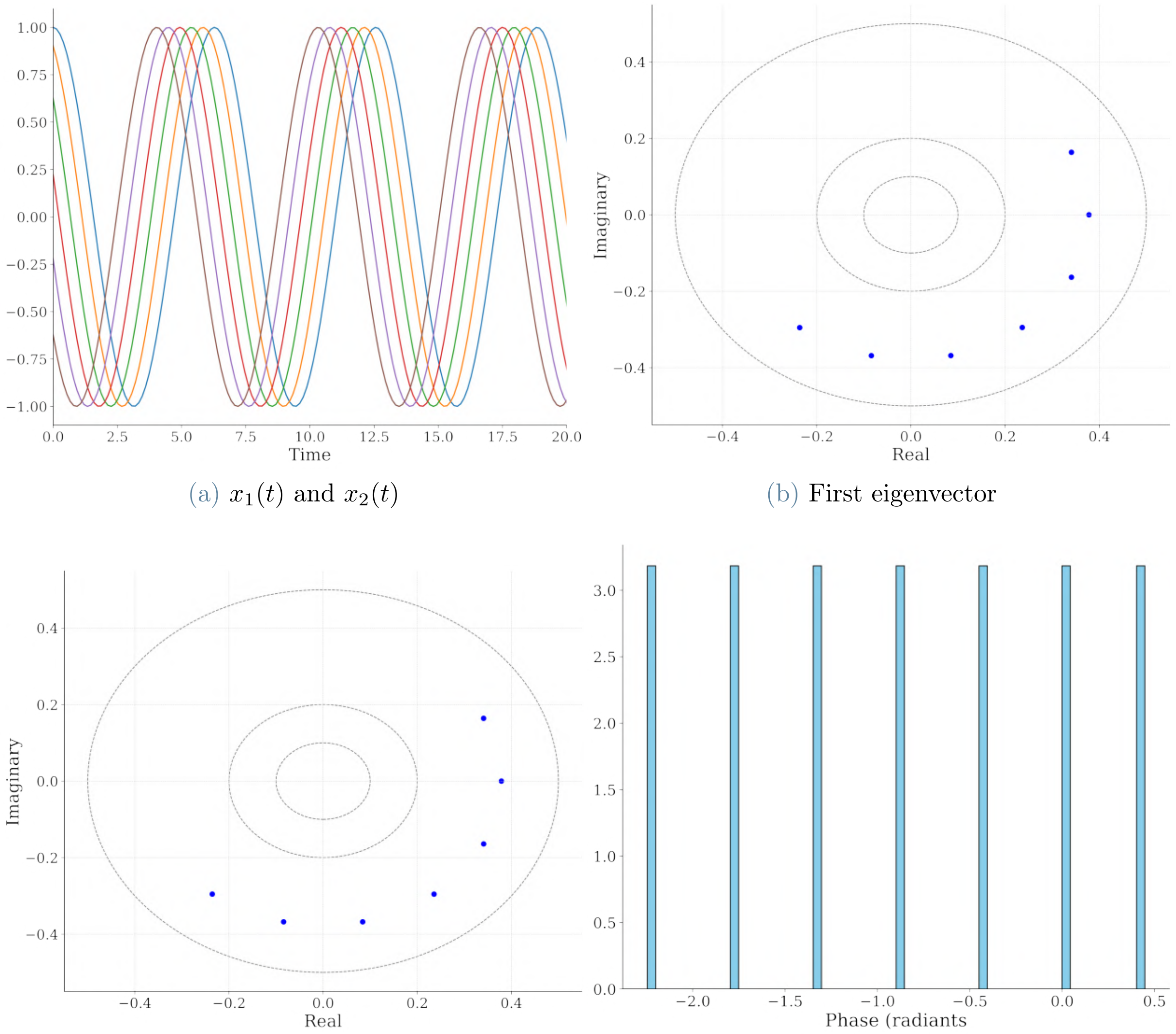


(a) $x_1(t)$ and $x_2(t)$

(b) First eigenvector

(c) Second eigenvector

Figure 5.6: The figure shows (i) the two cosines $x_1(t)$ and $x_2(t)$with a phase shift $\theta = 3/4\pi$ we applied the HPCA to. (ii)The first (ii) and second (iii) complex eigenvectors are visualized in the complex plane.



(a) $x_1(t)$ and $x_2(t)$

(b) First eigenvector

(c) Second eigenvector

(d) Second eigenvector

Figure 5.7: The figure shows (i) the 7 cosines series $x_1(t), \ldots, x_7(t)$ with a phase shift $\theta_i = i \times 1/7\pi$ we applied the HPCA to. (ii)The first (ii) and second (iii) complex eigenvectors are visualized in the complex plane. (iv) Shows the distribution of the phases for the first eigenvector

5.2.3. Results with price and sales data

In this last section, we show the results obtained by applying the HPCA (or complex PCA) to our sales and price data.

How to read the complex plane? Given a complex eigenvector ε_i expressed in polar components $\varepsilon_i = R\exp(i\theta)$ and its visualization in the complex plane, i.e. the plot of its components on the complex plane we see that a lagging relationship between two stocks is visually represented by the phase difference.

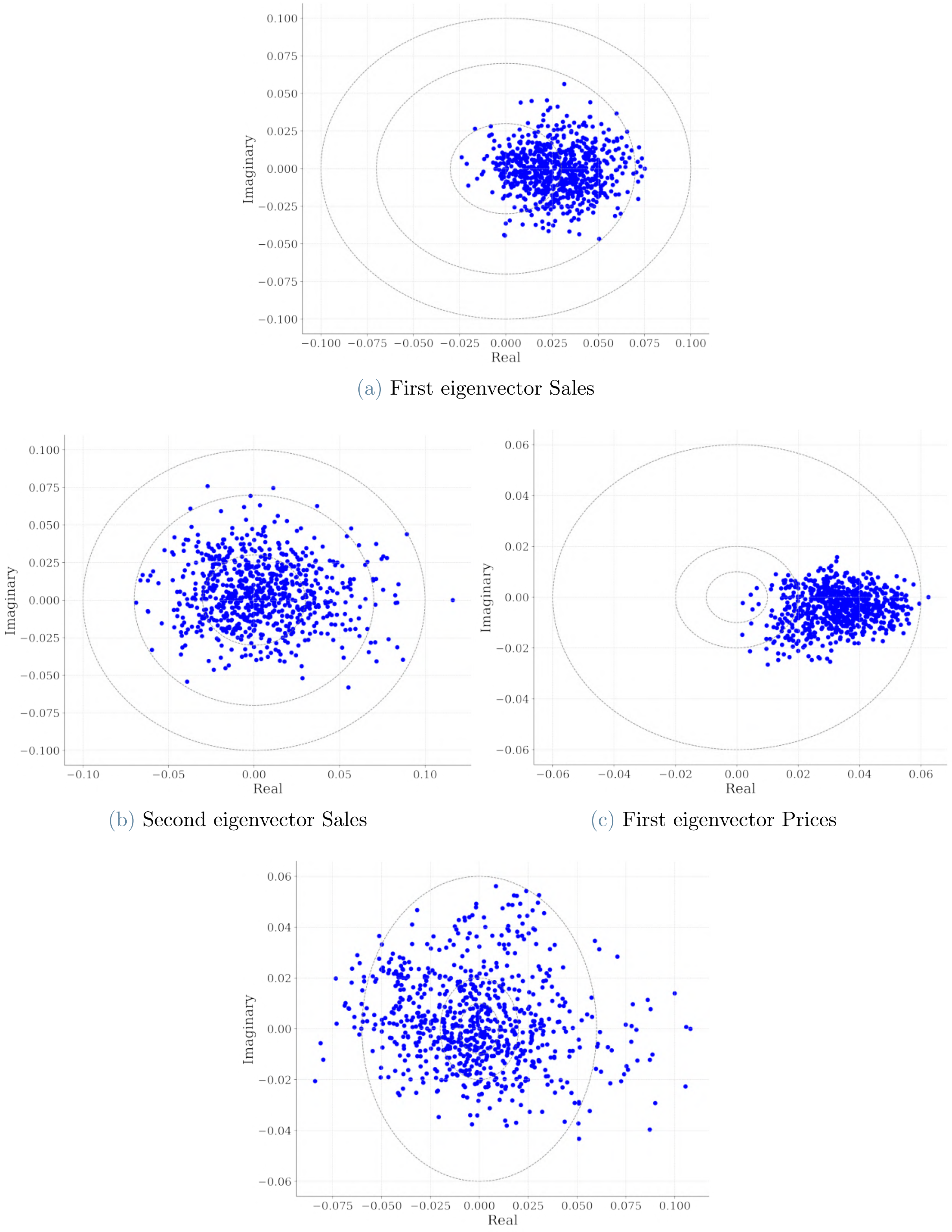
- Stocks exhibiting larger module values R tend to take on a leading role within the system. A larger module suggests that the stock has a stronger influence and tends to lead other stocks within the system.
- Stocks with a larger phase angle signify a pronounced lagging behavior, wherein one stock follows the other with a delay. The magnitude of this phase angle serves as an indicator of the extent of lag between the stocks.

Analyzing the time-delayed co-movements between economic variables we focused our interest on the first and second eigenvectors since we believe that most of the usable information is contained there.

In Figure 5.8 we show the top two eigenvectors for sales and price data. We can see that most of the components in the first eigenvector are condensed close to the origin, on the right part of the complex plane indicating that the leading and lagging relationships are concentrated on small phases. In our representation time development corresponds to the clockwise direction. This means that companies (i.e. components of the eigenvector) that lead are generally on the edges going clockwise direction and have a non-negligible module.

We recognized that this representation in our case with hundreds of companies it's not that informative. Since, from the beginning, we are interested in obtaining sectorial relations out of the data, we decided to plot again the first sales eigenvector but coloring each company with respect to the sector. As we can see in Figure 5.9 a pattern seems to emerge. Even though the large number of companies makes it difficult to read the figure, we see that certain specific sectors seem, on average, to separate from others.

To provide further validation for our hypothesis, we plotted the phase distribution of the first eigenvector, divided by sector. As demonstrated in Figure 5.10, the phase distributions exhibit partial overlaps for certain sectors, while distinct disparities are evident in others. To quantify and formalize this observation, we opted for the **one-way ANOVA test** as a statistical analysis.



(a) First eigenvector Sales

(b) Second eigenvector Sales

(c) First eigenvector Prices

(d) Second eigenvector Prices

Figure 5.8: The figure shows the first and second complex eigenvectors, respectively, for sales (i and ii) and prices (i and ii).

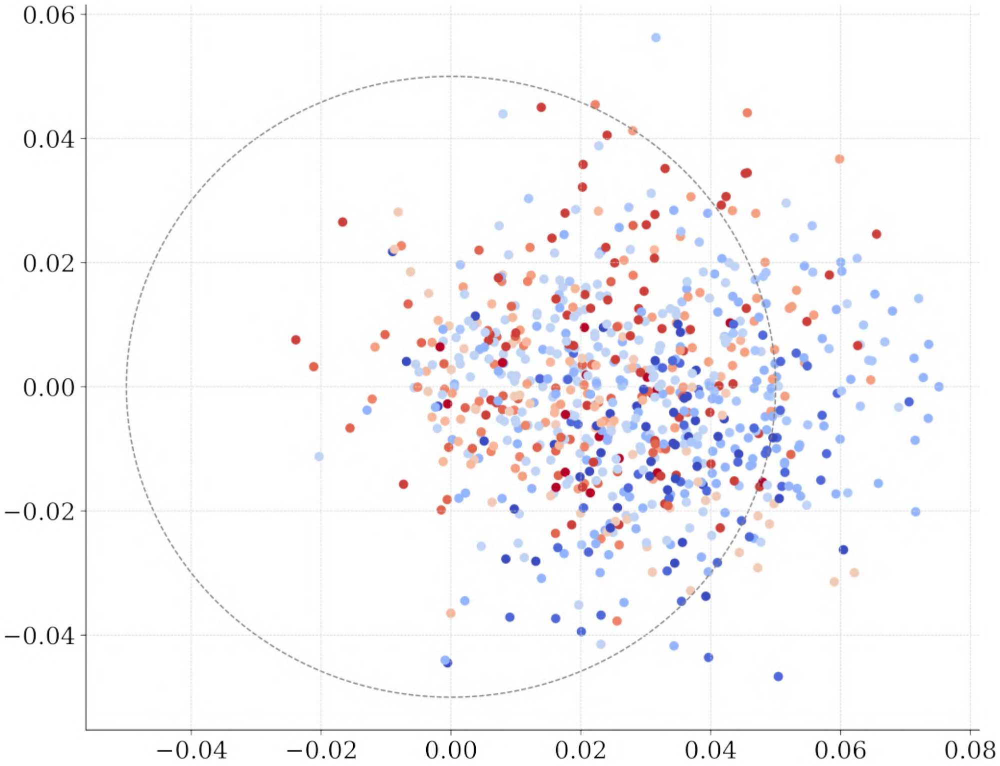
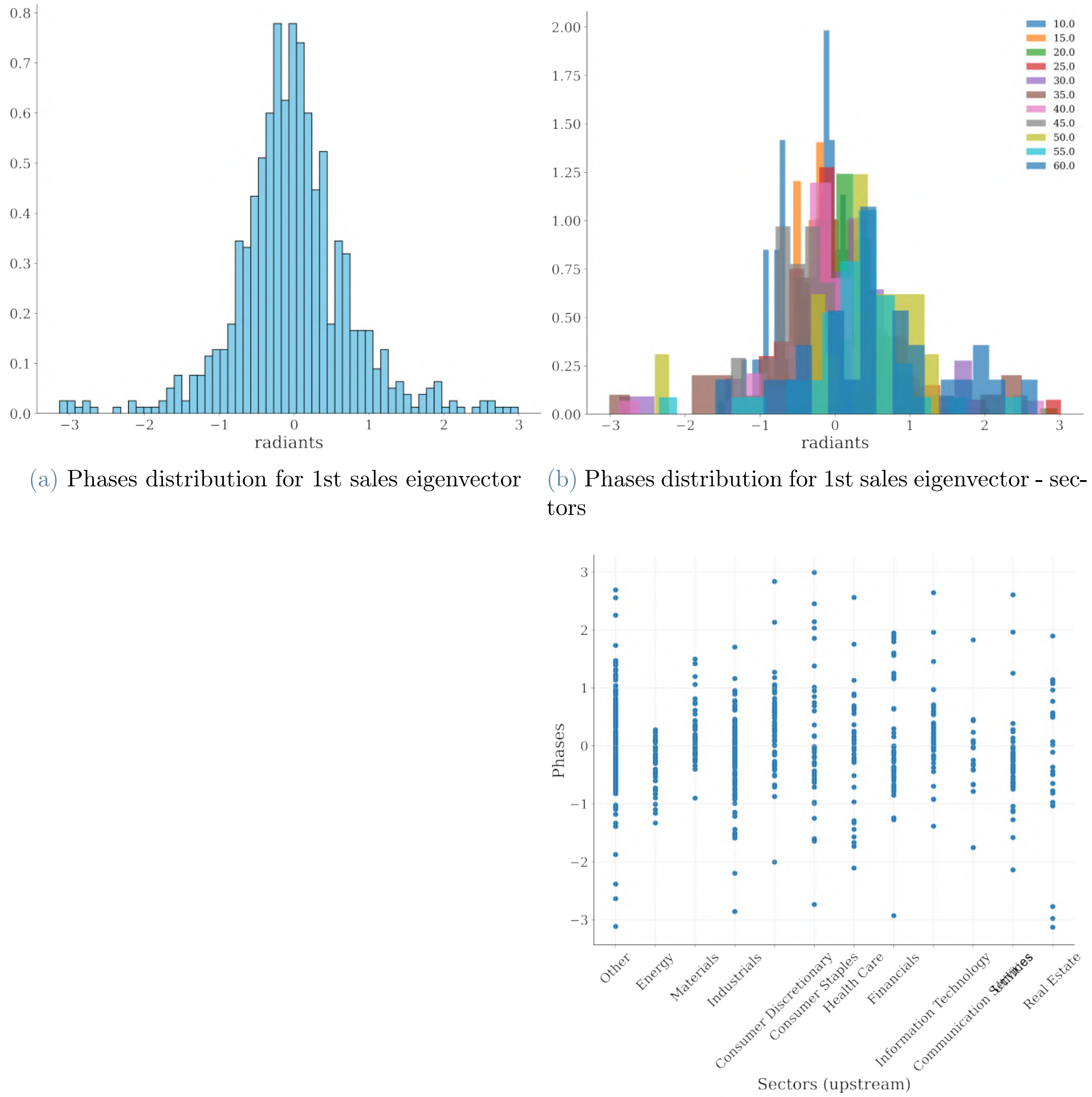


Figure 5.9: The figure shows the first eigenvector on the complex plane for sales data. Each component is colored by sector.

The test returns a p-value, which is a probability. The p-value represents the probability of obtaining the observed differences between the groups by chance alone, assuming the null hypothesis is true (i.e., assuming no significant differences between the group means). A low p-value indicates that the observed differences are unlikely to have occurred by chance, supporting the rejection of the null hypothesis and suggesting that there are statistically significant differences between the groups. In our case, we obtained a p-value $p = 5.9e-5$. This very low probability suggests that the samples are likely from different distributions, indicating a real sectorial difference in the phases.

Specifically, we observe a distinctive trend where the "energy sector" assumes a leading role over other industrial sectors. To illustrate this pattern, we've organized the sectors according to the Global Industry Classification Standard (GICS), following an upstream sequence. It becomes evident, as shown in Figure 5.10c that both the energy and utilities sectors demonstrate leading positions compared to the rest, while the financial and IT sectors appear relatively neutral in this context.

This intuition can be corroborated by applying the very same procedure to the second sales complex eigenvector. The results are shown in Figure 5.11. We can see that the

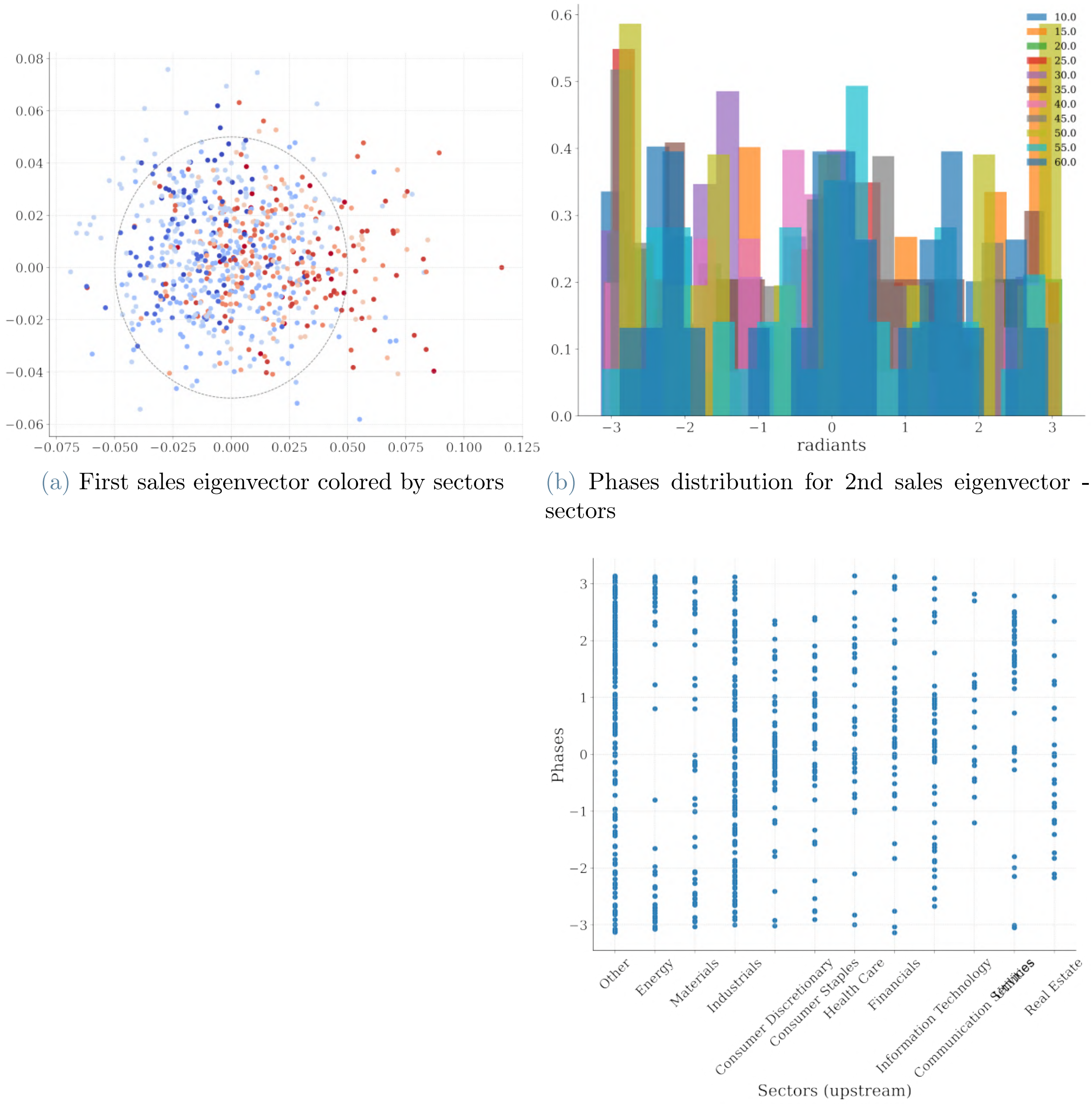


(a) Phases distribution for 1st sales eigenvector

(b) Phases distribution for 1st sales eigenvector - sectors

(c) Scatter plot of phases distribution by sector

Figure 5.10: The figure shows the phase distribution for the first sales complex eigenvector (i) and dividing by sector (ii). In (iii) the scatter plot of the distribution divided by phases is shown.



(a) First sales eigenvector colored by sectors

(b) Phases distribution for 2nd sales eigenvector - sectors

(c) Scatter plot of phases distribution by sector

Figure 5.11: The figure shows the (i) second sales complex eigenvector, (ii) its phases distribution divided by sector and (iii) the scatter plot of the distribution divided by phases.

components are more spread out in the complex plane and that there's an even clearer sectorial distinction. Being aware that the second mode describes another oscillation pattern in the data, we can see that a clear lead-lag relationship emerges for what concerns the energy sector and utilities, as we can see in 5.11c. The reason why energy and utilities might be correlated could be due to various factors. These sectors are often closely related as utilities, such as water, electricity, and natural gas, are essential services that rely on energy sources for their operation. Fluctuations in energy prices, availability, or technological advancements can impact the cost and supply of utilities.

Conclusions and future developments

In this research, we explored the entangled dynamics of companies in the supply chain, focusing on economic growth rates. To achieve this we analyzed the price and sale returns and their reciprocal correlations. The main goal was to find meaningful sectorial correlations in these multivariate datasets while filtering out noisy correlations.

Using the Marchenko-Pastur distribution we have been able to isolate the “signal” from the “noise”, pinpointing the genuine relationships between the companies. This correlation structure has been further examined through a deep investigation of the principal modes. Clear sectoralization patterns in their components emerged. When analyzed, these mode signals appear to exhibit strong nonlinearity features. Hence, we proved that these time series can be attributed to a non-linear dynamical process, using surrogate data methods for validation. We then compared the mode signals, i.e. the time evolution of each mode, with market indicators to analyze their sensitivity to market events. As a result, visible correlations emerged.

We then took to examining the disparity between sales and price returns. A major limitation in studying sales arises from their fixed frequency of sampling, as public companies report sales every quarter. As a fact, the price mode demonstrates significant differences compared to the sales mode. Initially, we hypothesized that the disparity in structure between the time series could be attributed to the different frequencies of sampling. We proved in our research that this is not the case, since the same structure persists even when downsampling the price data at various frequencies. The next question we sought to answer was whether the two processes could hypothetically originate from the same generative process. Our novel finding reveals that the difference between two "draws" of sales and prices, assuming they originated from the same "true" generative process plus noise, is inconsistent with two draws from the same distribution. In other words, the two quantities are not merely different realizations of the same distribution process. Although we observed some correlations between the two, we discovered that they originate from distinct generative processes. Nevertheless, we recognize that the question requires a more in-depth analysis.

Following the analysis of the generative process behind price and sales, we focused on analyzing the lead-lag effects in our data. In this network of interconnected variables, we applied several methods to explore the complete temporal correlation dynamics. Delving into the analysis of the lagged correlation matrix, through a singular component decomposition and the averaged auto-correlation function, we spot several time-lags that optimize the overall lead-lag relationship among the series. Even though we obtained meaningful information about the most significant lags in the systems we were still not satisfied. One of the downsides of the SVD approach is that it's quite hard to visualize which are the components involved in such a lead-lag relationship and ultimately generate a hypothesis on possible causalities among a huge number of variables. We were interested in easily representing which are the leading companies and which are the lagging ones. This can be achieved through the HPCA method. Indeed, this method, borrowed from climate science, proves to be a strong framework for this task. The results are easily interpretable on the complex plane corresponding to each eigenmode allowing us to identify distinctive trends indicating real sectorial differences in the lagging phases.

In conclusion, based on our discoveries, future research should focus on better understanding why sales and price returns seem connected, even though they come from different sources. We want to figure out what makes them related and how outside factors affect this connection. By doing this, we can learn more about how supply chain dynamics work.

Expanding the scope of the research, an important avenue for future exploration involves the development of fine-grained macroeconomic models, particularly Agent-Based Models (ABMs), that can accurately replicate detailed economic time series data. To be considered truly realistic, firm-level ABMs must eventually simulate firm-level time series data, if not precisely, then at least by capturing certain aggregated properties. Among these characteristics, one critical aspect to consider is the correlation between time series data. For instance, how do the sales of companies within the same industrial sector compare to one another? What dynamics emerge in the sales of firms that have established commercial relationships?

In the end, we believe that with a granular understanding of the interplay between firm-level economic variables, such as sales and price returns, and their correlations, we can contribute to the development of more sophisticated and accurate macroeconomic models. These models will not only improve predictive capabilities but also provide valuable insights into the complex interactions that drive economic growth and stability.

Bibliography

[1] R. Allez and J.-P. Bouchaud. Eigenvector dynamics: general theory and some applications. *Physical Review E*, 86(4):046202, 2012.

[2] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices.* Number 118. Cambridge university press, 2010.

[3] P. W. Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972.

[4] T. W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

[5] H. Aoyama, Y. Fujiwara, Y. Ikeda, H. Iyetomi, and W. Souma. *Econophysics and companies: statistical life and death in complex business networks.* Cambridge University Press, 2010.

[6] H. Aoyama, Y. Fujiwara, and Y. Ikeda. *Macro-econophysics: new studies on economic networks and synchronization.* Cambridge University Press, 2017.

[7] L. Arnold. Deterministic version of wigner's semicircle law for the distribution of matrix eigenvalues. *Linear algebra and its applications*, 13(3):185–199, 1976.

[8] G. B. Arous and A. Guionnet. The spectrum of heavy tailed random matrices. *Communications in Mathematical Physics*, 278(3):715–751, 2008.

[9] Z. Bai, B. Miao, and J.-F. Yao. Convergence rates of spectral distributions of large sample covariance matrices. *SIAM journal on matrix analysis and applications*, 25 (1):105–127, 2003.

[10] Z. Bai, H. Liu, and W.-K. Wong. Making markowitz's portfolio optimization theory practically useful. *Available at SSRN 900972*, 2016.

[11] T. Barnett. Interaction of the monsoon and pacific trade wind system at interannual time scales part i: The equatorial zone. *Monthly Weather Review*, 111(4):756–773, 1983.

[12] L. Basnarkov, V. Stojkoski, Z. Utkovski, and L. Kocarev. Lead–lag relationships in foreign exchange markets. *Physica A: Statistical Mechanics and its Applications*, 539: 122986, 2020.

[13] E. Bogomolny and M. Sieber. Power-law random banded matrices and ultrametric matrices: Eigenvector distribution in the intermediate regime. *Physical Review E*, 98 (4):042116, 2018.

[14] J. Bouchaud, L. Laloux, M. A. Miceli, and M. Potters. Large dimension forecasting models and random singular value spectra. *The European Physical Journal B*, 55: 201–207, 2007.

[15] J.-P. Bouchaud. Economics needs a scientific revolution. *Nature*, 455(7217):1181–1181, 2008.

[16] J.-P. Bouchaud and M. Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management.* Cambridge university press, 2003.

[17] J.-P. Bouchaud and M. Potters. Financial applications of random matrix theory: a short review. *arXiv preprint arXiv:0910.1205*, 2009.

[18] J.-P. Bouchaud and M. Potters. Financial applications of random matrix theory: a short review. 2015.

[19] M. J. Bowick and É. Brézin. Universal scaling of the tail of the density of eigenvalues in random matrix models. *Physics Letters B*, 268(1):21–28, 1991.

[20] E. Brezin, C. Itzykson, G. Parisi, and J.-B. Zuber. Planar diagrams. *Communications in Mathematical Physics*, 59:35–51, 1978.

[21] J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

[22] Z. Burda, A. Gorlich, A. Jarosz, and J. Jurkiewicz. Signal and noise in correlation matrix. *Physica A: Statistical Mechanics and its Applications*, 343:295–310, 2004.

[23] F. De Jong and T. Nijman. High frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance*, 4(2-3):259–277, 1997.

[24] T. Di Matteo. Multi-scaling in finance. *Quantitative finance*, 7(1):21–36, 2007.

[25] E. F. Fama and K. R. French. The capital asset pricing model: Theory and evidence. *Journal of economic perspectives*, 18(3):25–46, 2004.

[26] J. D. Farmer and A. W. Lo. Frontiers of finance: Evolution and efficient markets. *Proceedings of the National Academy of Sciences*, 96(18):9991–9992, 1999.

[27] C. Floros and D. Vougas. Lead-lag relationship between futures and spot markets in greece: 1999-2001. *International Research Journal of Finance and Economics*, (7): 168–174, 2007.

[28] Y. V. Fyodorov and A. D. Mirlin. Analytical derivation of the scaling law for the inverse participation ratio in quasi-one-dimensional disordered systems. *Physical review letters*, 69(7):1093, 1992.

[29] P. Grassberger. Grassberger-procaccia algorithm. *Scholarpedia*, 2(5):3043, 2007.

[30] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: nonlinear phenomena*, 9(1-2):189–208, 1983.

[31] A. Groth and M. Ghil. Synchronization of world economic activity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(12), 2017.

[32] T. Guhr, A. Muller-Groeling, and H. A. Weidenmuller. Random-matrix theories in quantum physics: common concepts. *Physics Reports*, 299(4-6):189–425, 1998.

[33] M. Haas and C. Pigorsch. Financial economics, fat-tailed distributions. *Encyclopedia of Complexity and Systems Science*, 4(1):3404–3435, 2009.

[34] J. D. Horel. Complex principal component analysis: Theory and examples. *Journal of Applied Meteorology and Climatology*, 23(12):1660–1673, 1984.

[35] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[36] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

[37] M. G. Kavussanos, I. D. Visvikis, and P. D. Alexakis. The lead-lag relationship between cash and stock index futures in a new market. *European Financial Management*, 14(5):1007–1025, 2008.

[38] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Physical review letters*, 83(7):1467, 1999.

[39] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3 (03):391–397, 2000.

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

[41] F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in nonlinear dynamics & econometrics*, 8(3), 2004.

[42] R. N. Mantegna and H. E. Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

[43] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[44] M. L. Mehta. *Random matrices*. Elsevier, 2004.

[45] J. A. Mingo and A. Nica. Annular noncrossing permutations and partitions, and second-order asymptotics for random matrices. *International Mathematics Research Notices*, 2004(28):1413–1460, 2004.

[46] A. D. Mirlin and Y. V. Fyodorov. The statistics of eigenvector components of random band matrices: analytical results. *Journal of physics A: mathematical and general*, 26(12):L551, 1993.

[47] C. Nadal and S. N. Majumdar. A simple derivation of the tracy–widom distribution of the maximal eigenvalue of a gaussian unitary random matrix. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(04):P04001, 2011.

[48] L. Noakes. The takens embedding theorem. *International Journal of Bifurcation and Chaos*, 1(04):867–872, 1991.

[49] R. Oppermann and F. Wegner. Disordered system with n orbitals per site: 1/n expansion. *Zeitschrift fur Physik B Condensed Matter*, 34(4):327–348, 1979.

[50] T. Panagiotidis, D. Chappell, et al. Using the correlation dimension to detect nonlinear dynamics. Technical report, 2004.

[51] S. Peche. *Universality of local eigenvalue statistics for random sample covariance matrices*. PhD thesis, Verlag nicht ermittelbar, 2003.

[52] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Physical review letters*, 83(7):1471, 1999.

[53] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65(6):066126, 2002.

[54] M. Potters, J.-P. Bouchaud, and L. Laloux. Financial applications of random matrix theory: Old laces and new pieces. *arXiv preprint physics/0507111*, 2005.

[55] E. M. Rasmusson, P. A. Arkin, W.-Y. Chen, and J. B. Jalickee. Biennial variations in surface temperature over the united states as revealed by singular decomposition. *Monthly weather review*, 109(3):587–598, 1981.

[56] P. Schmidt. *Econometrics*. CRC Press, 2020.

[57] J. W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

[58] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States, 1956.

[59] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1-4):77–94, 1992.

[60] C. A. Tracy and H. Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159:151–174, 1994.

[61] G. Viswanathan, U. Fulco, M. Lyra, and M. Serva. The origin of fat-tailed distributions in financial time series. *Physica A: Statistical Mechanics and its Applications*, 329(1-2):273–280, 2003.

[62] I. Vodenska, H. Aoyama, Y. Fujiwara, H. Iyetomi, and Y. Arai. Interdependencies and causalities in coupled financial networks. *PloS one*, 11(3):e0150994, 2016.

[63] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.

[64] M. Woodford. Learning to believe in sunspots. *Econometrica: Journal of the Econometric Society*, pages 277–307, 1990.

[65] Y. Q. Yin. Limiting spectral distribution for a class of random matrices. *Journal of multivariate analysis*, 20(1):50–68, 1986.

A | Appendix: Stieltjes transform for the Wishart matrix

Let us first show how to get the derivation of (1.17) in the context of the Coulomb Gas analogy. This formula creates a link between $\mathfrak{g}(z)$ and V'.

Throughout this discussion, we set $\beta = 1$. First, we introduce the normalized trace of the resolvent $\mathfrak{g}(z)$ in (1.16) by multiplying both sides of the equation by $N^{-1}(z-\nu_i)^{-1}$ and summing over all i. This leads to:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{V'(\nu_i)}{z-\nu_i} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1;j\neq i}^{N}\frac{1}{(z-\nu_i)(\nu_i-\nu_j)}. \tag{A.1}$$

It is important to note that this equation is valid as an analytical function for $z \in \mathbb{C}\backslash\text{Supp}[\rho_{\mathbf{M}}]$. Next, we manipulate the left-hand side (LHS) of the equation using algebraic operations:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{V'(\nu_i)}{z-\nu_i} = V'(z)\mathfrak{g}(z) - \frac{1}{N}\sum_{i=1}^{N}\frac{V'(z)-V'(\nu_i)}{z-\nu_i},$$

and for the right-hand side (RHS), we have:

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1;j\neq i}^{N}\frac{1}{(z-\nu_i)(\nu_i-\nu_j)} \equiv \frac{1}{2}\left[\mathfrak{g}^2(z) + \frac{1}{N}\mathfrak{g}'(z)\right].$$

By combining these two equations, we arrive at the saddle-point equation (A.1) in the form:

$$\frac{1}{2}\left[\mathfrak{g}^2(z) + \frac{1}{N}\mathfrak{g}'(z)\right] = V'(z)\mathfrak{g}(z) - \frac{1}{N}\sum_{i=1}^{N}\frac{V'(z)-V'(\nu_i)}{z-\nu_i}.$$

Since our focus is on the large N limit, we need to solve the following quadratic equation for $\mathfrak{g}(z)$:

$$\mathfrak{g}^2(z) - 2V'(z)\mathfrak{g}(z) + \frac{2}{N}\sum_{i=1}^{N}\frac{V'(z)-V'(\nu_i)}{z-\nu_i} = 0. \tag{A.2}$$

The most challenging term is the last one, as it involves an implicit sum. To simplify matters, let's consider the case where $V'(z)$ is a polynomial of degree $d > 0$. The extension to Laurent polynomials, which include negative powers, is straightforward. For a polynomial $V'(z)$, we find that:

$$P(z) \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{V'(z) - V'(\nu_i)}{z - \nu_i}$$

is also a polynomial of degree $d-1$. The coefficients of $P(z)$ can be determined later using normalization constraints or by matching certain moments. Consequently, the solution to Eq. (A.2) takes the form:

$$\mathfrak{g}(z) = V'(z) \pm \sqrt{V'(z)^2 - 2P(z)}.$$

In the case of a one-cut framework (where ρ has a unique compact support), the expression above can be further simplified, specifically when $d \geqslant 1$, resulting in:

$$\mathfrak{g}(z) = V'(z) \pm Q(z)\sqrt{(z-\nu_+)(z-\nu_-)},$$

where ν_- and ν_+ denote the edges of supp$[\rho]$, and $Q(z)$ is a polynomial of degree $d-1$. This simplification leads us to (1.17).

Let us now prove another necessary preliminary result. We will prove an interesting general property for the Stieltjes transform. We want to explore the asymptotic expansion of $\mathfrak{g}(z)$ as z becomes large (outside the support of ρ).

Expanding $\mathfrak{g}(z)$ in powers of z^{-1} gives us:

$$\mathfrak{g}(z) \underset{z\to\infty}{=} \frac{1}{z}\int \rho(u) \sum_{k=0}^{\infty} \left(\frac{u}{z}\right)^k \mathrm{d}u.$$

At the leading order, in agreement with the property $\lim_{|z|\to\infty} z\mathfrak{g}(z) = 1$ mentioned earlier, we have:

$$\mathfrak{g}(z) \sim \frac{1}{z}\int \rho(u)\mathrm{d}u \equiv \frac{1}{z},$$

where the last equality holds because the empirical spectral density (ESD) is normalized to unity.

The remaining terms of the expansion are also of particular interest. We observe that:

$$\mathfrak{g}(z) \underset{z\to\infty}{=} \frac{1}{z} + \frac{1}{N}\sum_{k=1}^{\infty} \frac{\mathrm{Tr}\mathbf{M}^k}{z^{k+1}} \equiv \frac{1}{z} + \sum_{k=1}^{\infty} \frac{\varphi(\mathbf{M}^k)}{z^{k+1}}, \tag{A.3}$$

where we introduce the k-th moment of the ESD as $\varphi(\mathbf{M}^k) := \frac{1}{N}\mathrm{Tr}\mathbf{M}^k$. This equation reveals the connection between the Stieltjes transform and the moment-generating function of the random matrix $\mathbf{M}$. It emphasizes the fact that the Stieltjes transform contains complete information about the eigenvalue density. Conversely, if one can measure the moments of the eigenvalue distribution, it is possible to reconstruct a parametric eigenvalue density function that accurately matches the empirical data. This property makes the Stieltjes transform particularly useful for statistical inference purposes. We will sometimes use the abbreviation $\varphi(\mathbf{M}^k) \equiv \varphi_k$ when there is no confusion about the matrix under consideration.

We will now derive the Stieltjes transform (1.21) using the BIPZ formalism introduced in Eq. (A.1). As mentioned earlier, the Stieltjes transform (A.1) for the isotropic Wishart matrix can be expressed as:

$$\mathfrak{g}(z) = \frac{1}{2q}\left[1 - \frac{1-q}{z}\right] - \frac{c}{z}\sqrt{z-\nu_+}\sqrt{z-\nu_-}, \tag{A.4}$$

where the constants c, ν_+, and ν_- need to be determined.

To achieve this, we refer to (A.3), which tells us the behavior of $\mathfrak{g}(z)$ as $|z| \to \infty$:

$$\mathfrak{g}(z) = \frac{1}{z} + \frac{\varphi(\mathbf{M})}{z^2} + \mathcal{O}(z^{-3}). \tag{A.5}$$

On the other hand, if we take the limit $z \to \infty$ in (A.4), we find:

$$\mathfrak{g}(z) = \frac{1}{2q}\left[1 - \frac{1-q}{z}\right] - c\left[1 - \frac{\nu_+ + \nu_-}{2z} - \frac{(\nu_+ - \nu_-)^2}{8z^2}\right] + \mathcal{O}(z^{-3}). \tag{A.6}$$

By comparing this equation to (A.5), we can fix the value of c by observing that the leading order term satisfies:

$$\frac{1}{2q} - c = 0,$$

since $\mathfrak{g}(z)$ behaves as $\mathcal{O}(z^{-1})$ for very large z. Consequently, we obtain:

$$c = \frac{1}{2q}. \tag{A.7}$$

Moving on to the $\mathcal{O}(z^{-1})$ term, we have:

$$1 = -\frac{(1-q)}{2q} + \frac{\nu_+ + \nu_-}{4q}, \tag{A.8}$$

which implies:

$$\nu_+ = 2(1+q) - \nu_-. \tag{A.9}$$

Finally, we determine the last constant using the condition at $\mathcal{O}(z^{-2})$:

$$\varphi(\mathbf{M}) = \frac{(\nu_+ - \nu_-)^2}{16q}, \tag{A.10}$$

which can be rearranged as:

$$\nu_- = \nu_+ - 4\sqrt{q\varphi(\mathbf{M})} = (1+q) - 2\sqrt{q} = (1-\sqrt{q})^2, \tag{A.11}$$

where we utilized (A.9) and $\varphi(\mathbf{M}) = 1$ in the third step. Consequently, we deduce from (A.9) that $\nu_+ = (1+\sqrt{q})^2$. Combining the equations (A.7), (A.9), and (A.11), we obtain the desired result (1.21).

List of Figures