The Random Matrix-based informative content of correlation matrices in stock markets

Laura Molero González, Roy Cerqueti, Raffaele Mattera, J.E. Trinidad Segovia

The Random Matrix-based informative content of correlation matrices in stock markets

Laura Molero González [1,2], Roy Cerqueti [2,3], Raffaele Mattera [4], Juan E. Trinidad Segovia[5]

[1,5] Department of Economics and Business, University of Almería, 04120, Almería, Spain.
[2] Department of Social and Economic Sciences, Sapienza University of Rome, 00185, Rome, Italy.
[3] GRANEM, University of Angers, 49100, Angers, France.
[4] Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", 81100, Caserta, Italy.

Email: `lmg172@ual.es` (L. Molero Gonzalez); `roy.cerqueti@uniroma1.it` (R. Cerqueti); `raffaele.mattera@unicampania.it` (R. Mattera); `jetrini@ual.es` (J. E. Trinidad Segovia)

Abstract

Studying and comprehending the eigenvalue distribution of the correlation matrices of stock returns is a powerful tool to delve into the complex structure of financial markets. This paper deals with the analysis of the role of eigenvalues and their associated eigenvectors of correlation matrices within the context of financial markets. We exploit the meaningfulness of Random Matrix Theory with the specific aspect of the Marchenko-Pastur distribution law to separate noise from true signal but with a special focus on giving an interpretation of what mean these signals in the financial context. We empirically show that the highest eigenvalue serves as a proxy of market spillover. Furthermore, based on an analysis of portfolio betas, we prove that the eigenvector associated with this eigenvalue is the market portfolio. These analyses of portfolio betas also reveal that the second and third-highest eigenvalues, and their associated eigenvectors, result in some cases of counter-behavior that makes them suitable to be a safe haven during high-volatility periods. The analysis is performed on a set of indices coming from developed and emerging countries over a time period ranging from 2015 to 2024.

Keywords. Econophysics; Random Matrix Theory; Marchenko-Pastur distribution; Eigenvalue distribution; Financial Markets; Spillover.

1 Introduction

It is well known that financial markets exhibit a complex structure, which is why non-equilibrium statistical physics has been widely used to study them (Puertas et al., 2020). For example, market price distributions often exhibit long tails similar to many other natural systems such as river networks (Tarboton et al., 1988), the structure of rocks (Kruhl, 2013) or the aggregation of mesoscopic particles (Weitz and Oliveria, 1984). Models imported from statistical physics (Stanley et al., 1996, 1999) have exploited these similarities. The power laws that have become so popular for studying price fluctuations are reminiscent of critical phenomena (Gabaix et al., 2003; Lux and Alfarano, 2016). The concept of entropy has been used as a measure of information for financial time series (Abril-Bermúdez et al., 2024) by connecting it to the degree of market efficiency (Wang, 2022; Molero González et al., 2024; Shternshis et al., 2022). Ising's model (Krawiecki et al., 2002) or linear response theory (Puertas et al., 2021, 2023) has been used to study market dynamics. The concept of memory, introduced by the hydrologist H. Hurst (Hurst, 1951), has also established itself as a very powerful indicator when studying the efficiency of markets (Dimitrova et al., 2019; Gómez-Águila et al., 2022; Fernández-Martínez et al., 2020).

In this way, Random Matrix Theory (RMT) is presented as a powerful tool that makes it possible to identify non-random properties, which represent deviations from the universal RMT predictions (Sharifi et al., 2004). The analysis and understanding of the eigenvalue distribution of stock returns provide powerful insights into the complex structure of financial markets (Achitouv, 2024). By comparing empirical correlation matrices **C** with those predicted by RMT, it is possible to identify significant deviations that highlight genuine market correlations versus pure noise (Laloux et al., 1999; Plerou et al., 2002). In other words, deviations from the purely random matrix distribution might suggest the presence of true information (signal). One of the foundational results in RMT is the Wigner (1958)'s Semicircle law, which aims to describe the limiting distribution of eigenvalues of large symmetric random matrices with independent entries (Mehta, 2004). The results from this study played a crucial role in understanding the spectral behavior of large random matrices and served as preceding to more refined results: the Marchenko and Pastur (1967) distribution law.

Within this framework, the Marchenko-Pastur (MP) distribution specifically describes the asymptotic behavior of the eigenvalue spectrum of empirical covariance matrices, when the dimensionality of the system (p) grows proportionally to the number of observations (n) (Molero-González et al., 2023). The MP distribution defines an upper and lower bound (λ_+ and λ_-, respectively) so that all eigenvalues $\lambda \in [\lambda_-, \lambda_+]$ are said to have their origin in random behavior. All eigenvalues outside the boundaries are, thus, due to signal. In this way, the MP law becomes a tool to distinguish between noise and signal in the understanding of financial markets dynamics.

In this line, Achitouv (2024) use complex network analysis to simulate the noise and the market components, proving that eigenvalues of the matrix above the upper bound of the MP distribution can be interpreted as collective modes behavior, while those under the bound are usually considered as noise. Jiang (2024) state that when there exists certain anomalous eigenvalues that do not follow the predicted RMT distributions it is an indication of the existence of an underlying non-random signal inside the data. Achitouv et al. (2024) analyze stock return correlations through the lens of RMT to distinguish the underlying signal from spurious correlations. They introduce a stochastic field theory model to establish a detection threshold from signal present in the limit where eigenvalues are within the continuous spectrum. Wàtorek et al. (2024) use the MP law to separate noise from signal in the NFT market.

In the present study, we follow the line about using the MP distribution law to separate noise from signal in financial markets but especially focusing on clearly interpreting the signals. The findings of this study are several. Firstly, we explore the empirical link between the highest eigenvalue of the correlation matrix and the spillover, showing a perfect overlap between the two variables. This proves that the highest eigenvalue can be used as a proxy for market spillover. Interestingly, we confirm for this eigenvalue some relevant outcomes which have been already found in the reference literature on the interpretation of it as the market factor (see e.g., Plerou et al. (2002) and Molero-González et al. (2023)). As a further investigation, we discuss the connection between spillover and the second eigenvalue obtaining a counter behavior of these quantities. By looking at the betas of the second-highest eigenvalue with respect to the market factor, we find a substantial lack of statistical significance except for a few cases, where we obtain that the eigenvector associated with the second eigenvalue is a defensive component for the risk of the market. In the same context, we found a lack of statistical significance also for the case of the third-highest eigenvalue and quite all the markets.

The rest of the paper is organized as follows. Section 2 gives an overview of relevant RMT theoretical aspects and introduces the MP distribution. Section 3 presents the dataset explored in the empirical experiments and the investigation methodology. Section 5 offers the results of the analysis, together with a related discussion. The last section concludes and outlines the routes for future research.

2 The Random Matrix Theory: an overview of relevant theoretical aspects

Random Matrix Theory tries to describe the statistics of the eigenvalues of random matrices, often on the edge of large dimensions (Pharasi et al., 2019). It was with Wishart (1928) that the issue first came up. The author proposed that the white noise time series correlation matrix was a suitable precedent for correlation matrices. Later on, Cartan (1935) proposed the classical random matrices ensembles.

This theory was introduced in physics by Wigner (1958, 1959), based on the assumption that the interactions between the nuclear constituents were so complex that they could be modeled as random fluctuations in the framework of his dispersion theory of the matrix[1] R (Wigner, 1993; Pharasi et al., 2019).

The author sought a purely statistical description of a set of a given number n of energy levels which could give rise to properties such as their empirical distribution and the distribution of the spacings. Furthermore, he formulated the hypothesis that the local statistical behavior of the energy levels (or eigenvalues) is adequately modeled by that of the eigenvalues of a random matrix, the so-called Wigner matrix (Wigner, 1958, 1967). Thus, the approximation consists of replacing the Hermitian operator H by a finite random matrix H_n, of size $n \times n$ (Johnstone, 2006). RMT is, thus, presented as a new type of statistical mechanics in which, instead of having a set of states governed by the same Hamiltonian, you have a set of Hamiltonians governed by the same symmetry.

Random Matrices are grouped into ensembles due to the similarity of their characteristics: GOE (Gaussian Orthogonal Ensemble), GUE (Gaussian Unitary Ensemble), and GSE (Gaussian Symplectic Ensemble). Each of them arises naturally in different physical and mathematical settings; particularly, in the study of energy levels of quantum systems. The classification is made on the basis of the symmetry properties of the matrix entries and the type of invariance under specific transformations. In this way, GOE consists of real symmetric matrices, GUE is made of complex Hermetian matrices and GSE of quaternionic self-dual matrices. One of the most significant characteristics of these ensembles is the behavior of the eigenvalue spectrum. As the size of the random matrix tends to infinity, the eigenvalues follow a certain distribution, depending on the ensemble to which they belong. However, there is a universal probability distribution such that the density of the eigenvalues of any Wigner matrix converges to it, regardless of the detailed structure of the matrix elements.

Specifically, let $H_n = (A + A^T)/2$ be a symmetric matrix of size $n \times n$, where the elements of matrix A are independent and identically distributed (i.i.d.) random variables following a standard Gaussian distribution $N(0, 1)$ when $n \to \infty$ and when $n \to \infty$ the spectral density function of H_n converges to the Wigner (1958)'s Semicircle Law (Livan et al., 2018), mathematically expressed in equation (1):

$$\rho(x) = \frac{1}{\pi}\sqrt{2 - x^2}. \tag{1}$$

This result plays a key role in understanding the spectral behaviour of large random matrices and serves as precursor to more refined results, such as the Marchenko and Pastur (1967) distribution law. In this study, we focus on the RMT and its interactions with important areas of what is known in statistics as Multivariate Analysis. Multivariate Analysis deals with observations of more than one variable when there is (or may be) some dependence between them. The most basic phenomenon known is the correlation.

The downscaling of financial data is a topic that has attracted the attention of researchers (Abdi and Williams, 2010). One technique widely used for this purpose is Principal Component Analysis (PCA) (see Ghorbani and Chong (2020); Yang et al. (2020); Dhingra et al. (2024) and Bilokon and Finkelstein (2021)). In statistics, it is common to assume a stochastic model in terms of random variables, whose distributions contain unknown parameters; which, in this case, would be the correlation matrix and its principal components (Johnstone, 2006). However, quantitative analysis requires more specific assumptions about the process of data generation. The simplest and most conventional model assumes that the p-dimensional random vector X follows a p-variate Gaussian distribution $N_p(\mu, \Sigma)$, with mean μ and covariance matrix Σ, and with a probability density function for X given by equation (2):

$$f(X) = |\sqrt{2\pi\Sigma}| \exp\left(-\frac{1}{2}(X - \mu)^T \Sigma^{-1}(X - \mu)\right) \tag{2}$$

The observed sample consists of n independent draws $X_1, \ldots, X_i, \ldots, X_n$ of $\mathbf{X} \sim N_p(\mu, \Sigma)$, collected in a data matrix $X = [X_1, \ldots, X_i, \ldots, X_n]$ of size $p \times n$. Reasonably, we set $n \geq p$ to guarantee a large enough sample. Focusing on correlations and without losing much generality, we conveniently set $\mu = 0$. Thus, let X be a matrix of size $p \times n$, with independent and identically distributed vector entries $X_i \sim N_p(0, \Sigma), i = 1, ..., n$; $\mathbf{W} = \frac{1}{n}\mathbf{X}\mathbf{X}'$ is said to have a Wishart p-variate distribution of n degrees of

[1] The matrix R is used in scattering theory to model interactions between particles by means of matrix elements that have a random behavior due to the intrinsic complexity of the system.

freedom $W_p(n, \Sigma)$. In other words, the joint distribution $f(w_{11}, ..., w_{pp})$ of the $\frac{1}{p}(p+1)$ elements of $\mathbf{W}$ is given by the expression (3) (Wishart, 1928; Peña, 1987):

$$f(w_{11}, \ldots, w_{pp}) = c|\mathbf{\Sigma}|^{-1/2}|\mathbf{W}|^{(n-p-1)/2} \exp\{-\frac{1}{2}Tr\mathbf{W}\}, \tag{3}$$

where c is a normalization constant and the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$ is supposed to be positive definite.The self-decomposition of the Wishart matrix is directly related to PCA and to the problem of determining the number of eigenvectors that best characterize the dynamics of the underlying data (Johnstone, 2006). A well-known fact nowadays in high-dimensional statistics is that, when $p, n \to \infty$, so that $\frac{p}{n} \to q \in (0, \infty)$, the eigenvalue distribution of the matrix $\mathbf{W} = \frac{1}{n}\mathbf{X}\mathbf{X}'$ converges to the distribution of Marchenko and Pastur (1967) shown on expression (4):

$$\rho(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi q \lambda} \tag{4}$$

where it is assumed that $\mathbf{\Sigma} = \mathrm{I}$. The quantities λ_+ and λ_- are the maximum and minimum expected eigenvalues, respectively, and are defined as $\lambda_\pm = (1 \pm \sqrt{q})^2$. This means that the eigenvalues $\lambda \in [\lambda_-, \lambda_+]$ have their origin in random behaviour (Molero-González et al., 2023). However, eigenvalues larger than this upper range λ_+ are, with high probability, attributed to the signal and not to random noise.

The MP distribution was discovered just over a decade after Wigner's Law, but it comes from a similar motivation. Instead of focusing on the distribution of the eigenvalues of a symmetric matrix, it focuses on the distribution of the eigenvalues of the correlation matrix of a rectangular matrix of size $p \times n$ given (Little, 2020). The key point of the MP distribution law is its universality: any distribution with independent and identically distributed entries will yield the same asymptotic eigenvalue distribution (Lu and Murayama, 2014).

RMT and the MP distribution have allowed a very precise study of large sample covariance matrices and also the design of estimators that are consistent in the Large Dimensional Limit (LDL), that is, when $\frac{p}{n} \to \infty$. The estimation of correlation matrices is a very old problem in multivariate statistics. The cleaning of large correlation matrices using the MP distribution is a topic that has attracted much attention in the literature (Bun et al., 2017). The study of empirical correlation matrices is even more important when speaking about financial assets since an important aspect of risk management is the estimation of the correlations between the price movements of assets (Laloux et al., 1999).

3 Data

This section is devoted to the description of the dataset and the methodological details of the analysis.To achieve the objectives of this study, the adjusted closing daily prices of eight developed stock markets and four emerging markets have been taken. Therefore, we consider a set of $k = 12$ different markets. After filtering the data to correct information gaps, the logarithmic returns have been obtained. Table 1 presents the countries, indeces and number of stocks of each stock market. The study period goes from January, 2015 to December, 2024. Table 2 shows the main descriptive statistics of the dataset used in the present study. Although the study period is the same for all markets, the total number of trading days varies slightly for each market. We denote the number of trading days for a generic market by n. The higher expected return is found in the Indian market, with a 0.063%, followed by S&P500 and Dow Jones markets, with expected returns of 0.039% and 0.038%, respectively. The lowest expected return corresponds to the Mexican market, with a 0.019%. European markets move around an expected return of 0.03%. The highest variability is found in the Merval market, with $\sigma_{\mathrm{MERVAL}} = 0.02464$. The European and North American markets are all around a standard deviation value of 0.012. The market with a lower variability is Mexico, with $\sigma_{\mathrm{MEXICO}} = 0.00974$. The lowest minimum return is found in the stock market of Argentina. For the European markets, returns did not fall more than -17.67%. For the North American markets, this number remained at -13.79%. Mexico is the market with a higher minimum return: -7.14%. Somehow in parallel, the highest maximum return is also found in Merval stock market. European markets had maximum return rates of 9%. For the North American markets this maximum was 10%.

Country	Index	Number of Stocks (p)
Germany	DAX40	40
United Kingdom	FTSE 100	100
Spain	IBEX35	35
Italy	FTSE MIB	40
Japan	NIKKEI	225
United States	Dow Jones	30
United States	S&P500	500
Mexico	IPC Mexico	35
Argentina	Merval	23
Brazil	Ibovespa	90
India	BSE Sensex	30

Table 1: Country, index, and number of stocks of the different stocks markets analyzed in the present study.

As it can be seen from Table 2 the mean and the median do not fit for any of the markets analyzed, meaning that the distribution of the data is not symmetric. All markets have a negative skewness, i.e., the distribution is skewed to the left. This indicates that extreme low returns are more frequent than extreme high returns, pulling the mean below the median and highlighting potential downside risk.

	CAC40	DAX40	FTSE100	IBEX35	FTSE MIB	NIKKEI	DOW JONES	S&P500	IPC MEXICO	MERVAL	IBOVESPA	INDIA
n	2569	2542	2525	2558	2543	2466	2515	2515	2514	2440	2487	2464
Mean	0.00031	0.00030	0.00027	0.00027	0.00031	0.00030	0.00038	0.00039	0.00019	0.00246	0.00023	0.00063
Std	0.01142	0.01190	0.01057	0.01183	0.01334	0.01263	0.01093	0.01176	0.00974	0.02462	0.01607	0.01027
Min	-0.13636	-0.13686	-0.11512	-0.16221	-0.17667	-0.13276	-0.12700	-0.13788	-0.07137	-0.44491	-0.18884	-0.14674
25%	-0.00448	-0.00475	-0.00457	-0.00527	-0.00568	-0.00557	-0.00387	-0.00440	-0.00480	-0.00927	-0.00758	-0.00414
50%	0.00070	0.00072	0.00058	0.00080	0.00114	0.00064	0.00066	0.00074	0.00015	0.00275	0.00068	0.00117
75%	0.00581	0.00622	0.00551	0.00657	0.00720	0.00686	0.00519	0.00616	0.00554	0.01498	0.00904	0.00577
Max	0.08464	0.09404	0.09005	0.09322	0.08590	0.08771	0.10065	0.10813	0.04440	0.20424	0.12654	0.08858
Skewness	-0.98507	-0.81342	-0.73358	-1.62700	-1.67527	-0.54170	-0.80287	-1.02302	-0.68256	-2.57101	-1.46959	-1.65321

Table 2: Descriptive Statistics. The value n represents the total number of the observation days of each market.

4 Methodology

This section contains the main methodological instruments used in the analysis. We present the methods in separate sections. We present the case of a correlation matrix $\mathbf{C}$ for the daily logarithmic returns of p stocks for a given financial market.

4.1 Eigenvalue analysis for signal and noise separation

To achieve the goals of our study, a sliding window experiment has been designed. For each market, we start by constructing windows of $n = 252$ days with a sliding step Δ. The selection of n corresponds to one year of trading days. However, S&P500 and Nikkei markets have a small number of stocks to be considered, i.e., $p = 500$ for the S&P500 and $p = 225$ for Nikkei. Therefore, taking $n = 252$ leads to high-dimensionality problems. To overcome them, we modify the selection of n for these two markets by taking $n = 1000$ days for S&P500 and $n = 500$ for Nikkei. The step of the shift is set to $\Delta = 5$ days, that is one week of trading days. On each window and for each market, the procedure followed to get the eigenvalues and classify them into signal or noise was the one shown below:

- First, we obtain the eigenvalues of $\mathbf{C}$ and their associated eigenvectors.
- Second, following expression (4), the MP time-invariant distribution is constructed.
- Third, we compare the empirical eigenvalue density distribution of $\mathbf{C}$ with the theoretical density distribution MP found in the previous step.
- Fourth, we identify all the eigenvalues greater than the upper bound of the MP distribution λ_+ and focus our study on them. Indeed, they are considered to be not random, i.e., they are signals.

By implementing the procedure described above, we realize that the fourth eigenvalue is almost never above the upper bound of the MP distribution. Therefore, we focus our attention on the first, second, and third-highest eigenvalues.

Next, we aim to exploit the informative power of the significant eigenvalues. Many previous studies suggest that the eigenvectors associated with the first largest eigenvalue are related to the market portfolio. In what follows, we study how the largest eigenvalues are associated with measures of connectedness in the market. To this aim, related the estimated eigenvalues with a spillover measure.

4.2 Spillover estimation

Since Diebold and Yilmaz (2009, 2012) there has been an increasing interest in measuring the magnitude of spillovers. Nowadays, the Diebold and Yilmaz (2012) approach is the most widely used in empirical research, as they proposed a relatively simple framework to compute the spillover index on a rolling-window basis for a set of p time series. The obtained spillover index is then interpreted as a measure of connecdtness. In the context of the stock market, we consider the p stocks included in a market and compute its returns spillover index.

To be more precise, the spillover index is computed from the Generalized Forecast Error Variance Decomposition (GFEVD), obtained from the following VAR(l)

$$X_i = \sum_{m=1}^{l} \mathbf{\Phi}_m X_{i-m} + \varepsilon_i \quad i = 2, \dots, n \text{ and } l = 1, \dots, i-1. \tag{5}$$

where $\mathbf{\Phi}_m$ is a $p \times p$ parameter matrix considering the m-th lag and $\varepsilon_i \sim (0, \Sigma_\varepsilon)$ is a vector of i.i.d. error terms. The GFEVD (Pesaran and Shin, 1998) indicates the fraction of the S ($s = 1, \dots, S$) step-ahead error variance of an j-th time series due to the shocks of another j'-th time series and it is computed as follows

$$\theta_{jj'}(S) = \frac{\sigma_{jj}^{-1} \sum_{s=0}^{S-1} \left(\iota_j' A_s \Sigma_\varepsilon \iota_{j'}\right)^2}{\sum_{s=0}^{S-1} \left(\iota_{j'}' A_s \Sigma_\varepsilon A_s' \iota_j\right)}, \tag{6}$$

where Σ_ε is the variance matrix for error vector, σ_{jj} is the standard derivation of the error term of the j-th equation of the process (5) and ι_j is the select vector which takes value of one for the j-th element and zero otherwise. Before computing the total amount of spillover, all the elements $\theta_{jj'}(S)$ are row-normalized so that $\sum_{j=1}^{n} \tilde{\theta}_{jj'}(S) = 1$, thus obtaining

$$\tilde{\theta}_{jj'}(S) = \frac{\theta_{jj'}(S)}{\sum_{h,k=1}^{n} \theta_{hk}(S)}. \tag{7}$$

Finally, we can define the Total Spillover Index (TSI) as the contribution of spillovers across the underlying stocks (7) to the total forecast error variance, that is,

$$\mathrm{TSI}(S) = \frac{1}{n} \sum_{\substack{j,j'=1 \\ j \neq j'}}^{n} \tilde{\theta}_{jj'}(S) \times 100. \tag{8}$$

Therefore, the total spillover is given by the sum of all spillovers computed by (7), but excluding the spillover of each j-th stock with itself. We can construct a dynamic TSI by considering a standard rolling window approach.

4.3 Market factor analysis

In Molero-González et al. (2023) the authors applied the Onatski (2008, 2009) test to determine the number of statistically significant factors in explaining the cross-section of stock returns. The Onatski test exploits the Tracy and Widom (1994) distribution in the context of factor models to analyze the eigenvalues of the correlation matrix of returns and determine the number of significant factors. The main idea is that the first highest eigenvalues correspond to common factors and should be separated from the eigenvalues that are noise (Onatski, 2008). In this way, using the properties of the Tracy-Windom distribution, this test evaluates the gaps (differences) between consecutive eigenvalues to identify statistically significant factors. In practical terms, when the number of factors k is correct, a very pronounced gap is observed after them. Using this methodology, the paper concludes that, for a confidence level of 99%, only one significant factor results to be statistically significant: the market factor. When the confidence level drops to 95% and 90%, up to three factors are identified.

This research expands the study of the information contained in the eigenvalues of the correlation matrix of returns, but delving into the dynamics of the highest eigenvalues, e.g. those corresponding to common factors. In Subsection 4.1 the eigenvalues were classified into signal and noise. This Subsection delves in the understanding of the role of the highest eigenvalues within the financial field. Concretely, we concentrate our efforts on the interpretation of these eigenvalues in the light of the Single Index Model of Sharpe (1964). To this aim, we computed the β_j with respect to the market factor (R_{mrkt}) of the eigenvectors associated to the highest eigenvalues. Below an in-depth explanation of this procedure is presented.

Also in this case we exploit the sliding window experiment presented in Subsection 4.1. For each window and market, we compute the eigenvalues of the correlation matrix of returns and take the first, second, and third-highest ones and their associated eigenvectors. We then take these eigenvectors as vectors of weights of a portfolio and normalize them to sum one. We computed the expected returns of the obtained portfolios for each window, jointly with the expected return of the market they belong to.

Having the expected return of the portfolios belonging to the first, second, and third-highest eigenvalues (we denote them with R_j in both cases, to avoid a cumbersome notation) and the expected return of each market (R_{mrkt}), we regress both variables by OLS by selecting one eigenvalue at each step as shown in equation (9):

$$R_j = \alpha_j + \beta_j \cdot R_{\text{mrkt}} + \epsilon_j, \tag{9}$$

where α_j and β_j are parameters to be calibrated and ϵ_j is a random noise with zero mean and form an i.i.d. process. We notice that the different β_j can also calculated as in equation (10):

$$\beta_j = \frac{\text{Cov}(R_j, R_{\text{mrkt}})}{\sigma^2(R_{\text{mrkt}})}. \tag{10}$$

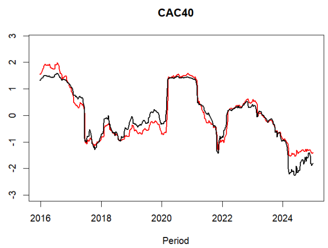
Furthermore, for the market portfolio we have that $R_j = R_{\text{mrkt}}$ and

$$\beta_m = \frac{\text{Cov}(R_{\text{mrkt}}, R_{\text{mrkt}})}{\sigma^2(R_{\text{mrkt}})} = \frac{\sigma^2(R_{\text{mrkt}})}{\sigma^2(R_{\text{mrkt}})} = 1. \tag{11}$$

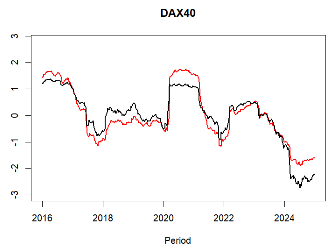
Thus, the beta of the market β_m always equals 1. Hence, the distance between β_j and 1 explains the analogy between the considered eigenvalue for stock p_k in market i and the market factor for i. We can preannounce that for the first highest eigenvalue we will obtain β_j always equals to 1, while for the second eigenvalue β_j is almost different from 1 (see the next section).

5 Results and discussion

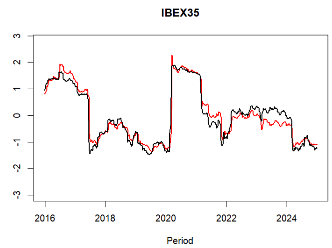
The results obtained in this study are presented below. We start by analyzing the (possibly existing) empirical link between the dynamics of the market spillover and the dynamics of the highest eigenvalues of the correlation matrix of returns for each of the markets studied. For this aim, we adopt the rolling-



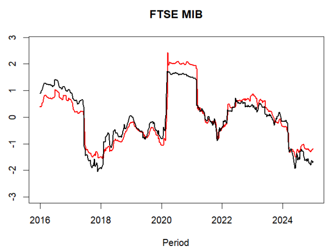
(a) CAC40



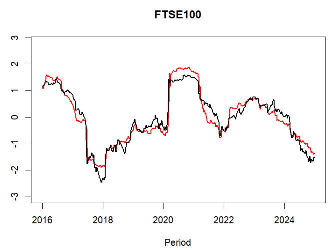
(b) DAX40



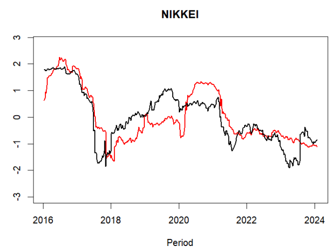
(c) IBEX35



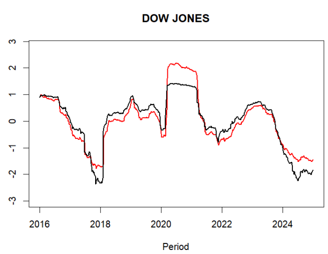
(d) FTSE MIB



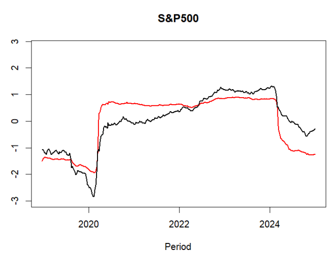
(e) FTSE100



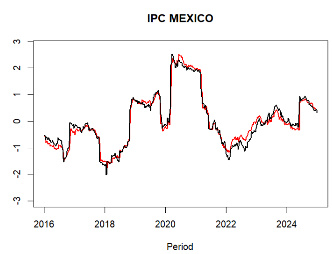
(f) NIKKEI



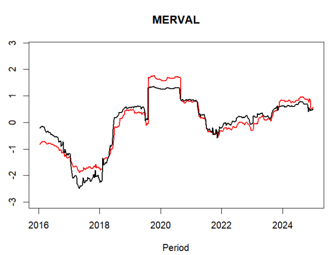
(g) Dow Jones



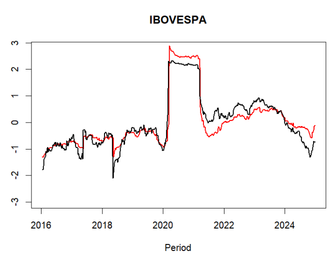
(h) S&P500



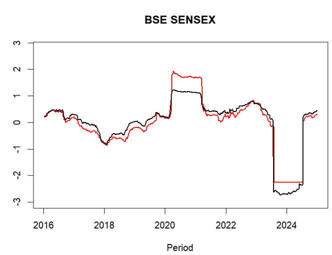
(i) IPC Mexico



(j) Merval



(k) Ibovespa



(l) BSE Sensex

Figure 1: Evolution over time of the dynamics of the market spillover computed on the returns (black line) against the dynamics of the highest eigenvalue of the correlation matrix of returns (red line) of the markets comprised in this study. Both variables are standardized, to allow comparisons on a common scale and avoid scaling problems in the analysis.

window procedure explained in Section 5 and compute the value of the largest eigenvalue over time. Figure 1 shows the temporal evolution of the market spillover for returns, represented by the black line, against that of the highest eigenvalue over time, represented by the red line. The results include both groups of markets, that is, developed and emerging markets. As can be seen, there exists an almost perfect overlap between the market spillover and the highest eigenvalue.

The observed similarity between the largest eigenvalue and spillover indices highlights the capacity of correlation matrices to encode systemic risk beyond pairwise correlations. This allows market participants to use the largest eigenvalue as a rapid and intuitive measure of market interconnectedness, bypassing the need for complex multivariate models. This finding offers significant advantages for financial risk assessment, portfolio optimization, and systemic risk monitoring. Moreover, while prior research recognized the link between high correlations and a dominant largest eigenvalue, our findings explicitly demonstrate this property for a variety of stock markets, establishing the spectral properties of correlation matrices as a powerful, alternative tool for gauging market interconnectedness.

We then investigate the significance of the largest eigenvalue from a portfolio perspective. We indeed remind that, following the classical Sharpe model (see Sharpe (1964)), the beta of the market portfolio should be equal to 1. Therefore, we find evidence supporting this theory if the portfolio resulting from the eigenvectors associated with the largest eigenvalue shows a beta equal to 1. Tables 3 and 4 show the results for all the considered markets. According to Table 4, the β_j is equal to 1 for all markets and, as shown in Table 4, it is statistically significant at 99%, confirming that the highest eigenvalue should be interpreted as the market factor of Sharpe, being its associated eigenvector the market portfolio.

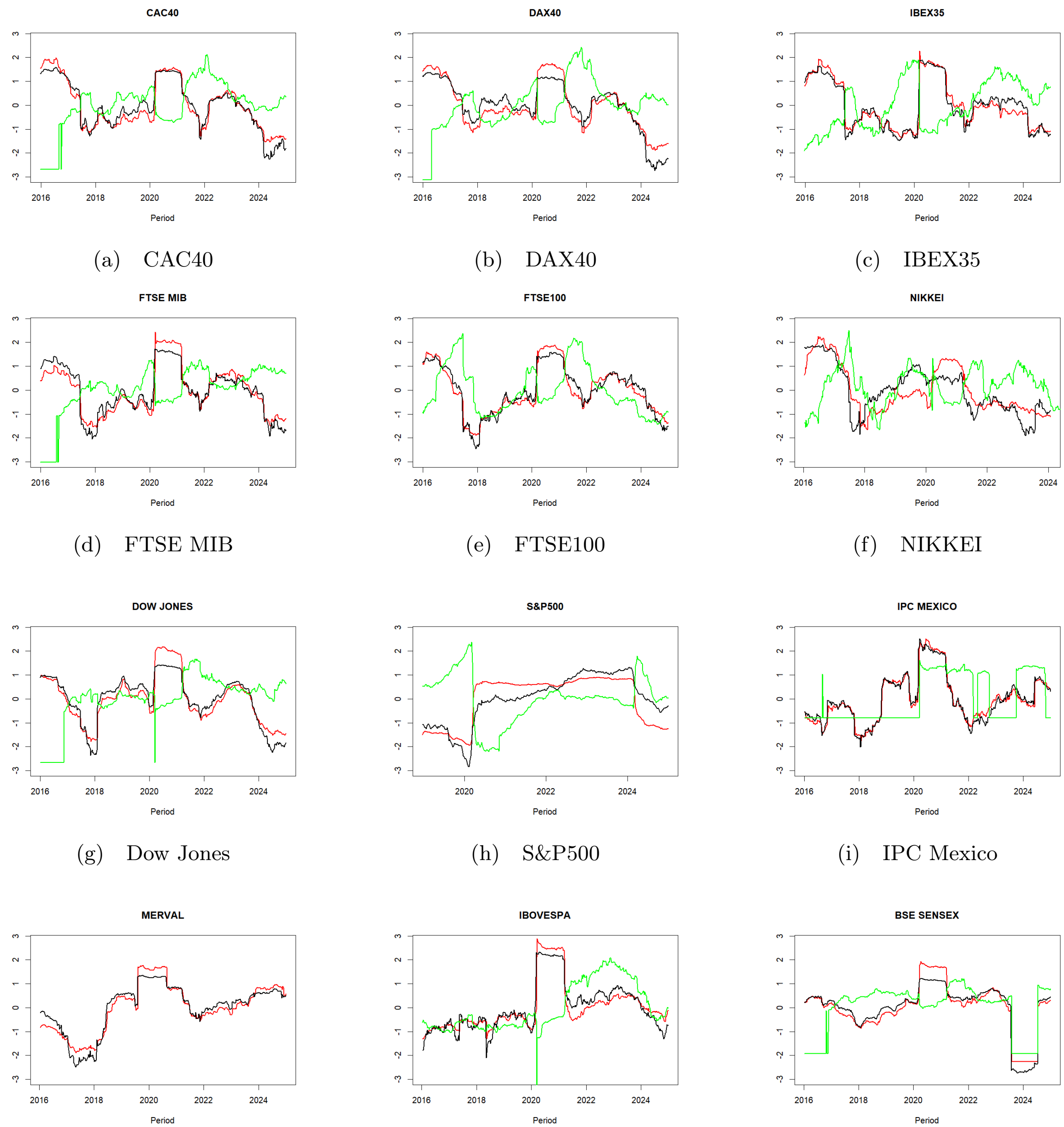
We then analyze the dynamics of the eigenvalues of the correlation matrix of returns of the different markets that were outside the MP upper bound, that is, those that can be considered as signals rather than noise. In practice, we continued the study by analyzing the second-highest eigenvalue. Figure 2 shows the dynamics of the first (red line) and second highest (green line) eigenvalues of the correlation matrix of returns against the dynamics of the market spillover (black line) computed on the returns. It is worth mentioning that for the case of the Merval market, just one signal is identified for all periods, the reason why it has no green line plotted. It is striking that the dynamics of the second-highest eigenvalue move in the opposite direction as the market spillover and, thus, as the highest eigenvalue. This happens for all markets, with some exceptions in some periods of Ibovespa and BSE Sensex markets, where both variables move in the same direction by the end of the study period analyzed. This almost perfect counter behavior leads to the thinking that this second-highest eigenvalue, and its corresponding eigenvector, should be a defensive risk component, which should be reflected by its portfolio beta. While Table 3 supports this result, Table 4 shows that the betas in the second-highest eigenvalue in model (9) are not statistically significant, except for the cases of NIKKEI and Dow Jones. For these markets, the values of the beta are remarkably negative. This confirms that they move in the opposite direction to the market and, therefore, to the first eigenvalue. Therefore, these portfolios are inversely related to the market, making them a safe haven for periods of high volatility.

Finally, we consider the empirical results for the third largest eigenvalue and its associated eigenvector. The results of the beta study are listed in the last columns of both Tables 3 and 4. For the third largest eigenvalue, the results appear to be not statistically significant, with the exceptions of NIKKEI at 99%, BSE Sensex at 95%, For these significant outcomes results are mixed, with a negative value for NIKKEI and FTSE100 and positive one for BSE Sensex.

6 Conclusions

The study and comprehension of the eigenvalue distribution of the correlation matrices of stock returns is presented as a powerful tool to explore the complex structure of financial markets. RMT is presented as a versatile methodological instrument that allows the identification of non-random properties, represented by the deviations from its universal predictions. Through comparison of empirical correlation matrices **C** with those predicted by RMT, it is possible to identify significant deviations that highlight genuine market correlations versus noise. This means that all deviations from the purely random matrix distribution suggest the presence of true information (signal). We use the MP distribution law to separate noise from signal in financial markets, with a special focus on giving a clear interpretation of the signals detected. To this end, eight developed markets and four emerging markets were analyzed. We took the daily logarithmic returns of stocks for the period ranging from January 2015 to December 2024. The research was carried out through the design of a sliding window experiment.

We began by exploring the link between the highest eigenvalue of the correlation matrix of returns



(a) CAC40 (b) DAX40 (c) IBEX35

(d) FTSE MIB (e) FTSE100 (f) NIKKEI

(g) Dow Jones (h) S&P500 (i) IPC Mexico

(j) Merval (k) Ibovespa (l) BSE Sensex

Figure 2: Evolution over time of the dynamics of the market spillover computed on the returns (black line) against the dynamics of the highest eigenvalue (red line) and the second highest eigenvalue (green line) of the correlation matrix of returns of the markets comprised in this study. All variables were standardized, to allow comparisons on a common scale and avoid scaling problems in the analysis.

and the market spillover. The results show a perfect overlap between the dynamics over time of both variables for all markets and periods considered. In this way, it is shown that the highest eigenvalue can be used as a proxy for market spillover. In addition, since this eigenvalue is always outside the MP distribution upper bound, it can be stated that the market spillover is never random for the market.

Furthermore, we delved into the connection between the market spillover and the second-highest

Market	β_{j1}	β_{j2}	β_{j3}
CAC40	1	-4.1	0.2
DAX40	1	-1.5	0.6
IBEX35	1	-11	-0.5
FTSE MIB	1	-1.7	-8.4
FTSE100	1	-8.2	-12.2
NIKKEI	1	-3.6	-3.3
Dow Jones	1	-5.8	-17.4
S&P500	1	-2.4	-1.8
IPC Mexico	1	-0.9	-0.5
Merval	1	-	-
Ibovespa	1	-0.4	3.7
BSE Sensex	1	-0.9	12.5

Table 3: Values of the beta of the portfolios (eigenvectors) of the three highest eigenvalues outside the boundary of MP distribution for the different markets analyzed. The last two columns for Meraval are empty because for this market we always found a single relevant eigenvalue.

		β_{1j}	β_{2j}	β_{3j}
CAC40	R^2	0,8017	0,0009	0,0000
	p-value	0,0000***	0,5000	0,9900
DAX40	R^2	0,8580	0,0028	0,0002
	p-value	0,0000***	0,2500	0,7900
FTSE100	R^2	0,6998	0,0003	0,0073
	p-value	0,0000***	0,6926	0,0687*
IBEX35	R^2	0,8512	0,0017	0,0004
	p-value	0,0000***	0,3741	0,6704
FTSE MIB	R^2	0,9534	0,0020	0,0005
	p-value	0,0000***	0,3347	0,6223
NIKKEI	R^2	0,5902	0,0175	0,0361
	p-value	0,0000***	0,0052***	0,0001***
Dow Jones	R^2	0,9460	0,1935	0,0039
	p-value	0,0000***	0,0000***	0,1839
S&P500	R^2	0,8083	0,0065	0,0001
	p-value	0,0000***	0,5873	0,8171
IPC Mexico	R^2	0,9186	0,0012	0,0003
	p-value	0,0000***	0,4643	0,7156
Merval	R^2	0,8892	-	-
	p-value	0,0000***	-	-
Ibovespa	R^2	0,7386	0,0001	0,0002
	p-value	0,0000***	0,8688	0,7598
BSE Sensex	R^2	0,8824	0,0002	0,0089
	p-value	0,0000***	0,7606	0,0469**

Table 4: Results of the coefficient of determination (R^2) and p-values from the linear regressions between the market return of each stock market and the first three highest eigenvalues. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The last two columns for Meraval are empty because for this market we always found a single relevant eigenvalue.

eigenvalue. In this case, we saw that this second eigenvalue exhibits a counter behavior when analyzing its dynamics with respect to the dynamics of the market spillover and the highest eigenvalue. The study of the betas revealed that for the highest eigenvalue β_j equals always one and it is statistically significant at 99%, confirming that it is the market portfolio, and, therefore, its associated eigenvalue should be the market factor of Sharpe. We also confirm the previous findings of Plerou et al. (2002) and Molero-González et al. (2023) concerning the interpretation of it as a market factor. Regarding the eigenvector associated with the second eigenvalue, the study of the beta manifests the previously identified counter-behavior by a negative beta for all markets, even if in a substantial lack of statistical significance. Some exceptions show that the portfolios built based on the eigenvector associated with the second-highest eigenvalue can be seen as a safe haven for investment during high-volatility periods.

The future line of research focuses on the study of specific economic sectors, in order to see if this counter-behavior is statistically significant. In this case, it can be interesting to assess if this outcome is the result of diversification or if it is something more general, that can appear in the financial field under some conditions.

Acknowledgements

Laura Molero González is supported by the funding of the grant PID2021-127836NB-I00 from the Ministerio Español de Ciencia e Innovación and FEDER.
Juan E. Trinidad Segovia is supported by the funding of the grant PID2021-127836NB-I00 from the Ministerio Español de Ciencia e Innovación and FEDER and PPIT-UAL, Junta de Andalucía-ERDF 2021–2027. Objective RSO1.1. Programme: 54.A.

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Abril-Bermúdez, F. S., Trinidad-Segovia, J. E., Sánchez-Granero, M. A., and Quimbay-Herrera, C. J. (2024). Multifractality approach of a generalized Shannon index in financial time series. *PLoS ONE*, 19(6):e0303252.

Achitouv, D. (2024). Inferring financial stock returns correlation from complex network analysis. *arXiv preprint*.

Achitouv, I., Lahoche, V., and Ousmane Samary, D. (2024). Signal inference in financial stock return correlations through phase-ordering kinetics in the quenched regime. *arXiv preprint*.

Bilokon, P. and Finkelstein, D. (2021). Iterated and exponentially weighted moving principal component analysis. *arXiv preprint arXiv:2108.13072*.

Bun, J., Bouchaud, J.-P., and Potters, M. (2017). Cleaning Large Correlation Matrices: Tools from Random Matrix Theory. *Physics Reports*, 666:1–109.

Cartan, È. (1935). *Domaines Fondamentaux et Structures Symétriques des Espaces Homogènes de Type Symétrique*. Hermann.

Dhingra, V., Sharma, A., and Gupta, S. K. (2024). Sectoral portfolio optimization by judicious selection of financial ratios via PCA. *Optimization and Engineering*, 25(3):1431–1468.

Diebold, F. X. and Yilmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal*, 119(534):158–171.

Diebold, F. X. and Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, 28(1):57–66.

Dimitrova, V., Fernández-Martínez, M., Sánchez-Granero, M., and Trinidad Segovia, J. (2019). Some comments on Bitcoin market (in) efficiency. *PLoS ONE*, 14(7):e0219243.

Fernández-Martínez, M., Sánchez-Granero, M., Casado Belmonte, M., and Trinidad Segovia, J. (2020). A note on power-law cross-correlated processes. *Chaos, Solitons & Fractals*, 138:109914.

Gabaix, X., Gopikrishnan, P., Plerou, V., and Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937):267–270.

Ghorbani, M. and Chong, E. K. (2020). Stock Price Prediction using Principle Components. *PLoS ONE*, 15(3):e0230124.

Gómez-Águila, A., Trinidad-Segovia, J., and Sánchez-Granero, M. (2022). Improvement in Hurst exponent estimation and its application to financial markets. *Financial Innovation*, 8(1):86.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799.

Jiang, N. (2024). Denoising Data with Random Matrix Theory. *Journal of Applied Mathematics and Physics*, 12(11):3902–3911.

Johnstone, I. M. (2006). High Dimensional Analysis of Principal Components: A Survey. *arXiv:math/0611589v1*.

Krawiecki, A., Hołyst, J., and Helbing, D. (2002). Volatility clustering and scaling for financial time series due to attractor bubbling. *Physical Review Letters*, 89(15):158701.

Kruhl, J. H. (2013). Fractal-geometry techniques in the quantification of complex rock structures: a special view on scaling regimes, inhomogeneity and anisotropy. *Journal of Structural Geology*, 46:2–21.

Laloux, L., Cizeau, P., Bouchaud, J.-P., and Potters, M. (1999). Noise Dressing of Financial Correlation Matrices. *Physical Review Letters*, 83(7):1467.

Little, E. (2020). The Marchenko-Pastur Distribution: From Understanding to Application. *arXiv preprint*.

Livan, G., Mehta, M. L., and Prats, D. (2018). *Introduction to Random Matrices: Theory and Practice*. Cambridge University Press.

Lu, X. and Murayama, H. (2014). Universal Asymptotic Eigenvalue Distribution of Large N Random Matrices—A Direct Diagrammatic Proof to Marchenko-Pastur Law. *arXiv preprint*.

Lux, T. and Alfarano, S. (2016). Financial power laws: Empirical evidence, models, and mechanisms. *Chaos, Solitons & Fractals*, 88:3–18.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of Eigenvalues for Some Sets of Random Matrices. *Mat. Sb.(NS)*, 114(4):507–536.

Mehta, M. L. (2004). *Random Matrices*. Elsevier.

Molero González, L., Abril Bermúdez, F. S., Trinidad Segovia, J. E., and Yousaf, I. (2024). Connecting Market Efficiency and Asset Pricing Models: Insights from an Entropy-Based Approach. *Available at SSRN 5041644*.

Molero-González, L., Trinidad-Segovia, J. E., Sánchez-Granero, M. A., and García-Medina, A. (2023). Market Beta is Not Dead: An Approach from Random Matrix Theory. *Finance Research Letters*, 55:103816.

Onatski, A. (2008). Determining the Number of Factors from Empirical Data. *The Review of Economics and Statistics*, 92(4):1004–1014.

Onatski, A. (2009). Testing for Weak Instruments in Linear IV Regression. *Econometrica*, 77(5):1013–1045.

Pesaran, H. H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58(1):17–29.

Peña, D. (1987). Identifying Informative Variables in Large Datasets. *Journal of the American Statistical Association*, 82(398):486–502.

Pharasi, H. K., Sharma, K., Chakraborti, A., and Seligman, T. H. (2019). Complex Market Dynamics in the Light of Random Matrix Theory. In *New Perspectives and Challenges in Econophysics and Sociophysics*, pages 13–34. Springer.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., and Stanley, H. E. (2002). Random Matrix Approach to Cross Correlations in Financial Data. *Physical Review E*, 65(6):066126.

Puertas, A. M., Clara-Rahola, J., Sánchez-Granero, M. A., de Las Nieves, F. J., and Trinidad-Segovia, J. E. (2023). A new look at financial markets efficiency from linear response theory. *Finance Research Letters*, 51:103455.

Puertas, A. M., Sánchez-Granero, M. A., Clara-Rahola, J., Trinidad-Segovia, J. E., and de Las Nieves, F. J. (2020). Stock markets: A view from soft matter. *Physical Review E*, 101(3):032307.

Puertas, A. M., Trinidad-Segovia, J. E., Sánchez-Granero, M. A., Clara-Rahora, J., and de Las Nieves, F. J. (2021). Linear response theory in stock markets. *Scientific Reports*, 11(1):23076.

Sharifi, S., Crane, M., Shamaie, A., and Ruskin, H. (2004). Random Matrix Theory for Portfolio Optimization: A Stability Approach. *Physica A: Statistical Mechanics and its Applications*, 335(3-4):629–643.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 19(3):425–442.

Shternshis, A., Mazzarisi, P., and Marmi, S. (2022). Measuring market efficiency: The Shannon entropy of high-frequency financial time series. *Chaos, Solitons & Fractals*, 162:112403.

Stanley, H. E., Afanasyev, V., Amaral, L. A. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Leschhorn, H., Maass, P., Mantegna, R. N., Peng, C.-K., et al. (1996). Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Physica A: Statistical Mechanics and its Applications*, 224(1-2):302–321.

Stanley, H. E., Amaral, L. A. N., Canning, D., Gopikrishnan, P., Lee, Y., and Liu, Y. (1999). Econophysics: Can physicists contribute to the science of economics? *Physica A: Statistical Mechanics and its Applications*, 269(1):156–169.

Tarboton, D. G., Bras, R. L., and Rodriguez-Iturbe, I. (1988). The fractal nature of river networks. *Water resources research*, 24(8):1317–1322.

Tracy, C. A. and Widom, H. (1994). Level-Spacing Distributions and the Airy Kernel. *Communications in Mathematical Physics*, 159:151–174.

Wang, X. (2022). Efficient markets are more connected: An entropy-based analysis of the energy, industrial metal and financial markets. *Energy Economics*, 111:106067.

Weitz, D. and Oliveria, M. (1984). Fractal structures formed by kinetic aggregation of aqueous gold colloids. *Physical Review Letters*, 52(16):1433.

Wigner, E. P. (1958). On the Distribution of the Roots of Certain Symmetric Matrices. *Mathematics of Computation*.

Wigner, E. P. (1959). *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*. Academic Press, New York.

Wigner, E. P. (1967). Random Matrices in Physics. *SIAM Review*, 9(1):1–23.

Wigner, E. P. (1993). On a class of analytic functions from the quantum theory of collisions. In *The Collected Works of Eugene Paul Wigner: Part A: The Scientific Papers*, pages 409–440. Springer.

Wishart, J. (1928). The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika*, 20A(1):32–52.

Wàtorek, M., Szydło, P., Kwapień, J., and Drożdż, S. (2024). Correlations Versus Noise in the NFT Market. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(7).

Yang, S., Florescu, I., and Islam, M. T. (2020). Principal Component Analysis and Factor Analysis for Feature Selection in Credit Rating. *arXiv preprint*.