University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.20 No.2 June 2025

Stock Market of Random Matrix Theory to High-Dimensional Statistics

Ali Hammoodi Ali [1,*]

[1] **Department of studies and planning, University of Thi-Qar**

ali.hammood@utq.edu.iq

https://doi.org/10.32792/utq/utj/vol20/2/8

Abstract

This dissertation investigates the application of Random Matrix Theory (RMT) to the analysis of stock market behavior in high-dimensional statistical settings. As financial markets generate increasingly large and complex datasets, traditional statistical tools often fail to capture the true underlying correlations between assets, especially when the number of variables exceeds the number of observations common scenario in modern finance. RMT offers a robust mathematical framework for distinguishing genuine information from random noise in large correlation matrices. By analyzing the eigenvalue spectrum of empirical correlation matrices derived from asset return data, this study identifies deviations from the theoretical predictions of RMT, which often correspond to meaningful market signals or latent factors. The research explores both theoretical and empirical dimensions. On the theoretical side, it examines the implications of the Marčenko–Pastur law and the behavior of eigenvalues in finite samples. On the empirical side, RMT-based filtering techniques apply to real-world financial datasets to enhance portfolio optimization, reduce estimation risk, and improve the stability of financial models under high-dimensional constraints. The findings demonstrate that RMT, when integrated with modern statistical learning techniques, provides a powerful approach for financial modeling, especially in contexts involving large asset universes and limited time series data. This study contributes to the growing body of literature that positions RMT as a cornerstone methodology in high-dimensional finance and paves the way for further interdisciplinary research at the intersection of statistical physics, econometrics, and machine learning.

Key words:(Random matrix theory (RMT) , Large data matrices ,High-dimensional statistics)

1. Introduction

The advent of big data in finance has introduced complexities that traditional statistical methodologies struggle to address. Financial datasets now encompass thousands of assets observed over limited time frames, leading to issues such as overfitting and multicollinearity. High-dimensional statistics, concerned with the analysis of where the number of variables p approaches or exceeds the sample size n, necessitates new analytical tools. Random Matrix Theory, with its roots in quantum physics, has shown promise in resolving such issues by offering probabilistic insights into the behavior of large random matrices. This research aims to harness the theoretical rigor and practical adaptability of RMT to enhance our understanding of stock market correlations and their implications.

1.1 Literature Review

Emerging in the 1950s, the theory of large-dimensional random matrices was initially developed to address specific questions in mathematical physics. Over time, it has grown into an independent and rapidly evolving branch of mathematics, with extensive applications in various scientific disciplines, including the study of disordered quantum systems, number theory, econometrics, and biological networks. By the 1990s, this theory had found significant applications in signal processing and communications. More recently, it has become increasingly relevant in the field of learning theory.

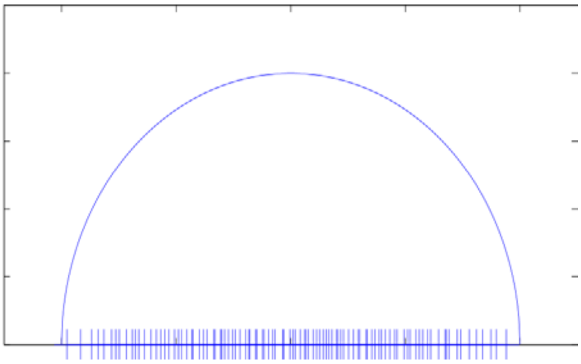
The foundational idea behind random matrix theory can be traced back to Eugene Wigner in the 1950s. While analyzing the energy spectra of heavy atomic nuclei, Wigner proposed modeling the (unknown) Hamiltonian operator—which governs the quantum states of such systems—as a large random matrix with a simple and generic structure. Specifically, he considered a real symmetric matrix $X = [x_{ij}]$ of dimension $N \times N$, where the entries x_{ij} are independent and identically distributed (i.i.d.) random variables, subject to symmetry, with zero mean and unit variance.

The energy levels of the nucleus correspond to the eigenvalues $\{\lambda_1, \dots, \lambda_N\}$ of the scaled matrix $N^{-1/2}X$. While the specific distribution of these eigenvalues depends on the distribution of the entries x_{ij}, Wigner demonstrated that in the limit as $N \to \infty$, the global spectral distribution becomes independent of the specific distribution of the entries—a phenomenon known as universality, analogous to the law of large numbers in probability theory.

To formalize this, Wigner introduced the empirical spectral distribution $\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i}$, which he proved converges almost surely to a deterministic probability distribution known as the *semicircle law*, supported on the interval $[-2,2]$ **[1].**

A decade later, Marchenko and Pastur introduced a second class of random matrix models, which have found even broader applications than Wigner's model, particularly in statistics and signal processing. The simplest model they examined is the *Gram matrix* $\Sigma = \frac{1}{n}XX^*$, where $X \in \mathbb{C}^{N\times n}$ is a random matrix with i.i.d. entries that are centered and have unit variance, and X^* denotes the conjugate transpose of X.

In the case where N is fixed and $n \to \infty$, the law of large numbers implies that Σ converges almost surely to the identity matrix I_N, and hence its spectral distribution μ_N converges almost surely to the Dirac delta measure at 1, denoted δ_1.



However, Marchenko and Pastur focused on a more intricate asymptotic regime where both N and n tend to infinity such that the ratio $N/n \to c > 0$. In this setting, although Σ still converges element-wise to I_N, its spectral behavior changes significantly. The empirical spectral distribution μ_N now converges almost surely to a new deterministic distribution—known as the *Marchenko–Pastur law*. This distribution is supported on the interval $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ within $[0, \infty)$, and it has an explicit analytical expression for its density function. **[2]**

Just as with Wigner's model, the convergence to the Marchenko–Pastur law is universal, meaning it holds irrespective of the precise distribution of the entries of X.

Since the pioneering works of Wigner and Marchenko–Pastur, a wide variety of random matrix models have been developed and analyzed, further enriching this vibrant and multidisciplinary field of study.

FIGURE 1 – Realization of eigenvalues (vertical lines) and limiting density (solid curve) for the Wigner model.N = 100.

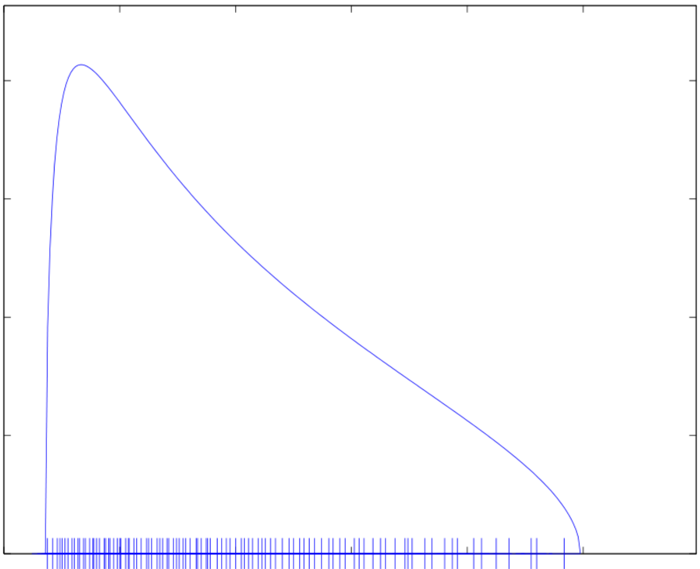


FIGURE 2 – Realization of eigenvalues (vertical lines) for a Marchenko-Pastur model, N = 100, n = 300. Limiting density (solid curve) for c = 1/3.

In general, **correlated**, **non-centered** random matrix models, as well as their various combinations, exhibit a key structural property: they possess **O(N²)** degrees of freedom. That is, the $N \times N$ matrix under study typically involves approximately $O(N^2)$ independent random variables.

Mathematically, we denote the spectral measure associated with the matrix as:$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i}$

where λ_i are the eigenvalues of the matrix, and δ_{λ_i} denotes the Dirac delta measure at λ_i.

The primary objective in the study of such models is to analyze the **tight convergence** (or weak convergence in distribution) of the spectral measure μ_N as $N \to \infty$. This analysis emphasizes the **global (macroscopic)** behavior of the spectrum, rather than the **local (microscopic)** fluctuations of individual eigenvalues λ_i.

Going further, Wigner was also interested in the microscopic behavior of the eigenvalues of a large random matrix. Because of the interest in the fine study of the separations between the states of the underlying quantum system, the determination of the law of the spacings between the eigenvalues is indeed a major question **[3]**.

During the 1960s, Gaudin, Mehta, and Dyson provided substantial support for the semicircle law, demonstrating that the typical spacing between eigenvalues of a Gaussian Wigner matrix within the interval [−2, 2] is on the order of 1/N. At this microscopic scale, they showed that the asymptotic distribution of eigenvalue fluctuations follows what is known as the sine kernel law.

Moreover, it was discovered in the early 1990s that the largest eigenvalue of a Gaussian random matrix exhibits typical spacing of order $N^{-2/3}$. At this refined scale, the asymptotic behavior of the fluctuations is described by the Tracy–Widom distribution, which characterizes the universal limiting law of the largest eigenvalues in such ensembles. Comparable results have also been derived for the Gaussian sample covariance matrix defined as $\frac{1}{n}XX^*$, where X consists of independent Gaussian entries. In this context, the largest eigenvalue similarly converges in distribution to the Tracy–Widom law under appropriate normalization.

Given these foundational results in the Gaussian setting, a natural question arose: Do these phenomena persist beyond Gaussian matrices? This led to the long-standing Wigner–Dyson–Mehta conjecture, which posited that the microscopic behavior of eigenvalues—just like their macroscopic behavior—is universal. That is, it depends only on the general class of the matrix model (e.g., Wigner, Marchenko–Pastur, or correlated ensembles) and not on the specific distribution of the entries.
This conjecture was rigorously proven in the latter half of the 2000s by Erdős, Schlein, Yau, and Yin. Parallel and complementary contributions were also made by Tao and Vu, further solidifying the universality of eigenvalue statistics in large-dimensional random matrix theory.

An important and application-relevant extension of random matrix theory involves finite-rank perturbations of large random matrices**.** Consider again the model $\Sigma = \frac{1}{n} XX^*$, and now suppose that $X \in \mathbb{C}^{N\times n}$ is of the form $X = A + W$, where W is a random matrix with independent and identically distributed (i.i.d.), centered entries of unit variance and finite fourth moment, and A is a deterministic matrix of rank one.

It is noteworthy that, in the absence of the deterministic component A, none of the eigenvalues of Σ deviates from the support of the Marchenko–Pastur distribution in the high-dimensional limit. Furthermore, the inclusion of A does not affect the **macroscopic behavior** of the eigenvalue distribution of Σ, owing to its low rank. That is, the spectral measure of Σ still converges to the Marchenko–Pastur law in the limit as $N, n \to \infty$.

However, a **phase transition phenomenon** may occur: an **outlier eigenvalue** can emerge outside the support of the limiting distribution. Specifically, when $\frac{1}{\sqrt{n}} \parallel A \parallel$ exceeds a critical threshold, the largest eigenvalue of Σ separates from the bulk and converges to a deterministic value that can be explicitly characterized (see Figure 4). Conversely, if the threshold is not surpassed, the largest eigenvalue remains at the right edge of the support of the Marchenko–Pastur law **[4].**

Up to this point, all random matrices considered in this discussion have been either Hermitian or symmetrical. However, an important branch of random matrix theory is devoted to the spectral analysis of non-Hermitian (or non-symmetric) matrices. This area of study emerged later than the corresponding investigations of Hermitian matrices.

In general, the eigenvalues of a square matrix exhibit greater sensitivity to perturbations compared to its singular values. Even a minor perturbation can cause significant changes in the eigenvalues, whereas the singular values tend to be more robust and undergo only limited variation. Random matrix theory reveals that for non-Hermitian matrices with $O(N^2)$ degrees of freedom, the spectral sensitivity to perturbations is significantly mitigated. That is, the random nature and high dimensionality of such matrices contribute to stabilizing their spectral characteristics.

A foundational model in the theory of non-Hermitian random matrices involves the matrix $N^{-1/2}X$, where $X = [x_{ij}]$ is a real or complex $N \times N$ matrix with independent and identically distributed (i.i.d.) entries, each centered and with unit variance. Since the eigenvalues of $N^{-1/2}X$ are generally complex, the associated spectral measure is supported over the complex plane $\mathbb{C}$.

A central result in this domain is that, regardless of the specific distribution of the entries x_{ij}, the empirical spectral measure of $N^{-1/2}X$ converges almost surely to the uniform distribution on the unit disk in $\mathbb{C}$. This phenomenon is known as the *circular law*, and its proof has evolved over several decades, involving significant developments in probability theory, functional analysis, and complex analysis **[5].**

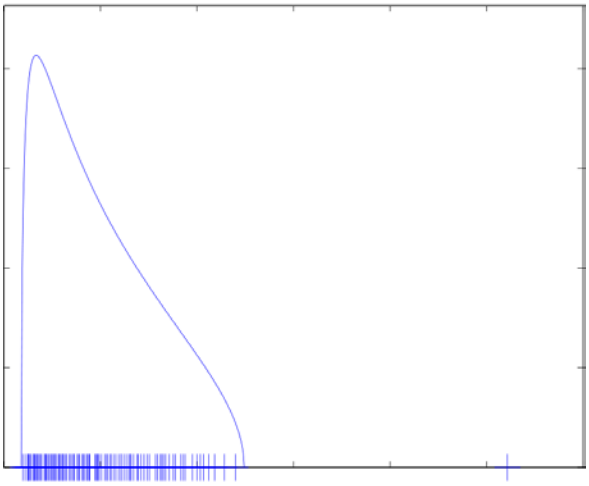


FIGURE 3–**The eigenvalues** corresponding to a realization of the matrix $n^{-1}(A+W)(A+W)^{*}$ are displayed alongside the **Marchenko–Pastur law**, for the parameters $N = 100$, $n = 300$, with **rank**$(A) = 1$, and $\| n^{-1/2}A \| = 2$. The figure clearly illustrates the .separated from the bulk of the spectrum ,**eigenvalue (isolated (outlier** appearance of an

In the 1960s, *Ginibre* established the circular law for the case of complex Gaussian random matrices. Two decades later, *Girko* introduced a more general framework based on the notion of logarithmic potential and the technique known as *Hermitization*, enabling the extension of this result to a broader class of distributions.

Since then, the theory has advanced to encompass more complex statistical models beyond the basic i.i.d. framework, including localized distributions and finite-rank perturbations, even in the non-Hermitian setting. The growing relevance of large random matrix theory in modern statistics stems from the high dimensionality of data in contemporary applications. This includes contexts such as high-dimensional feature spaces in machine learning, large-scale network inference, sensor arrays in distributed systems, and antenna arrays in fields like radio astronomy.

The relevance of large random matrix theory in statistical inference arises from the practical limitations encountered when the number of observations n in a statistical time series is not significantly larger than the series dimension N. This situation frequently occurs due to constraints on the observation period that are required to maintain the stationarity assumption. Under such conditions, the classical asymptotic framework—where N is fixed and $n \to \infty$—becomes inadequate. A more realistic setting considers both N and n tending to infinity at comparable rates.

To illustrate, consider a multivariate time series represented by a matrix $Y \in \mathbb{C}^{N\times n}$ of the form $Y = R^{1/2}X$, where $X \in \mathbb{C}^{N\times n}$ is a matrix with independent and identically distributed (i.i.d.) centered entries of unit variance, and R is a deterministic but unknown covariance matrix. In application areas such as antenna signal processing, econometrics, and others, a central objective is to develop inference procedures based on the eigenvalues or specific eigenspaces of the covariance matrix R.

Traditional inference methods, designed under the assumption N is fixed and $n \to \infty$, often fail to remain consistent under the high-dimensional asymptotic regime where $N, n \to \infty$ jointly. One of the notable achievements of random matrix theory is the development of inference algorithms that retain consistency under this more realistic asymptotic scenario.

Moreover, the scope of the theory extends beyond basic models such as $Y = R^{1/2}X$. Recent advancements have incorporated more sophisticated structures, including models for broadband antenna arrays and rational spectral representations characterized by systems of state equations **[6].**

Other research directions within the field have focused on inference for finite-rank perturbation models, particularly those of the form $W + A$ (as previously discussed), where the primary information is encoded in the low-rank matrix A and W represents a "noise" matrix. Additionally, significant attention has been given to the development of robust estimation algorithms, such as "M-estimators," which are designed to be resistant to impulsive noise. These approaches aim to enhance the reliability and accuracy of statistical inference in the presence of outliers or heavy-tailed disturbances in the data.

1.2 Data collection

For this study, 50 assets from the stock market were selected, with data spanning from September to December, collected at 5-minute intervals. In September, the data was retrieved from 10:00 AM to 5:05 PM, while in December, the final observation was made at 5:35 PM due to a change in the trading schedule to accommodate daylight saving time. As a result, each asset in September had 1,634 observations, while each asset in December had 1,656 observations. The assets were randomly selected, with the only criterion being that each asset must have at least 1,000 observations in each month.

Since the primary objective of this study is to analyze the correlation between assets during the trading day, only data from the market opening hours were considered. This ensures that the first value of each day, as provided by the Bloomberg system, is distinct from the last observation of the previous day, eliminating the risk of spurious correlation due to the absence of changes in the final log returns of each day. Furthermore, this approach allows for a more precise capture of potential relationships, particularly those stemming from high price fluctuations at the beginning of the trading day.

Given that assets do not necessarily trade at every 5-minute interval and there may be gaps in the data, an adjustment was made to fill any missing values. Specifically, if the missing data corresponds to the first observation of the month, the next available price was used to replace it. For gaps occurring in the middle or at the end of the month, the previous data point was used to complete the missing value.

2 Methodology

This study adopts a quantitative methodology, leveraging high-frequency and daily return data from a diverse pool of equities across major financial markets. The data preprocessing involves the calculation of standardized log returns and the construction of correlation matrices for the selected asset pools. These correlation matrices are central to the analysis, which utilizes eigenvalue decomposition, allowing for the assessment of eigenvalue distributions and their alignment with theoretical models from Random Matrix Theory (RMT), particularly the Marčenko-Pastur law. This law defines the asymptotic behavior of eigenvalue distributions for large random matrices under the null hypothesis of randomness.

Hypotheses

- **H1**: The eigenvalue distributions of the correlation matrices will converge to the Marčenko-Pastur law in the limit of large N and n, with the bulk of the eigenvalues being contained within the support of the Marčenko-Pastur distribution.
- **H2**: The presence of financial market correlations will lead to deviations from the Marčenko-Pastur distribution, with the emergence of outliers or isolated eigenvalues, indicating a structured correlation beyond randomness.

- **H3**: The correlation matrices derived from high-frequency return data will exhibit more pronounced non-random structures compared to those derived from daily returns.

Equations and Statistical Analysis

The analysis is framed within the following mathematical model:

1. **Standardized Log Returns**

The log returns $r_{i,t}$ for asset i at time t are defined as:

$$\boldsymbol{r_{i,t} = \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right)} \ldots\ldots\ldots\ldots\ldots(1)$$

where:

- $r_{i,t}$ is the log return of asset i at time t,
- $P_{i,t}$ is the price of asset i at time t,
- $P_{i,t-1}$ is the price of asset i at the previous time period.

2. Correlation Matrix

The correlation matrix C is computed as:

$$\boldsymbol{C = \frac{1}{n} X X^T} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

where:

- X is the matrix of standardized log returns with size $N \times n$ (N assets and n observations),
- C is the correlation matrix of the returns.

3. Eigenvalue Decomposition

The eigenvalues λ_i of the correlation matrix C are obtained by solving the following equation:

$$\boldsymbol{C v_i = \lambda_i v_i} \ldots\ldots\ldots\ldots\ldots(3)$$

where:

- v_i is the eigenvector corresponding to eigenvalue λ_i,
- λ_i are the eigenvalues of the correlation matrix C.

4. Marčenko-Pastur Law

For large N and n, the limiting distribution of eigenvalues for random matrices follows the Marčenko-Pastur law. The density function $\rho(\lambda)$ of the eigenvalues is given by:

$$\boldsymbol{\rho(\lambda)} = \frac{\mathbf{1}}{\boldsymbol{2\pi\sigma^2\lambda}}\sqrt{\boldsymbol{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}\text{............(4)}$$

where:

- $\lambda_+ = \left(1 + \sqrt{c}\right)^2$ and $\lambda_- = \left(1 - \sqrt{c}\right)^2$ are the upper and lower bounds of the spectrum,
- σ^2 is the variance of the returns (which can be set to 1 if returns are standardized),
- $c = \frac{n}{N}$ is the ratio of the number of assets to the number of observations.

5. Bootstrapping for Robustness Check

To assess the robustness of the eigenvalue distribution, a bootstrapping procedure can be employed. The procedure involves resampling with replacement from the returns data to generate multiple synthetic datasets. For each synthetic dataset, the correlation matrix and eigenvalues are recomputed. The confidence intervals for the eigenvalues are then estimated.

Let the bootstrap sample be denoted as $X^{(b)}$, where b indexes the bootstrap iteration. The bootstrapped correlation matrix is computed as:

$$\boldsymbol{C^{(b)}} = \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{X^{(b)}\left(X^{(b)}\right)^T}\text{.............(5)}$$

Then, the eigenvalues $\lambda_i^{(b)}$ for each bootstrap iteration are computed, and their distribution is used to form confidence intervals for the true eigenvalues.

6. Kernel Density Estimation (KDE)

To smooth the eigenvalue distribution and visualize the density, kernel density estimation (KDE) is applied to the eigenvalues. The KDE estimate $\hat{f}(\lambda)$ of the probability density function of the eigenvalues is given by:

$$\boldsymbol{\hat{f}(\lambda)} = \frac{\mathbf{1}}{\boldsymbol{Nh}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}\boldsymbol{K}\left(\frac{\boldsymbol{\lambda-\lambda_i}}{\boldsymbol{h}}\right)\text{.................(6)}$$

where:

- λ_i are the eigenvalues,
- $K(\cdot)$ is the kernel function (e.g., Gaussian kernel),
- h is the bandwidth parameter that controls the smoothness of the density estimate,
- N is the number of eigenvalues.

Additional tools such as kernel density estimation (KDE) and bootstrapping are employed to assess the robustness of the eigenvalue distributions, and a systematic filtering procedure is applied to eliminate noisy components, ensuring the refinement of the correlation structures. These steps contribute to verifying whether the empirical data adheres to theoretical predictions, providing insights into the underlying market dynamics. .**[7]**

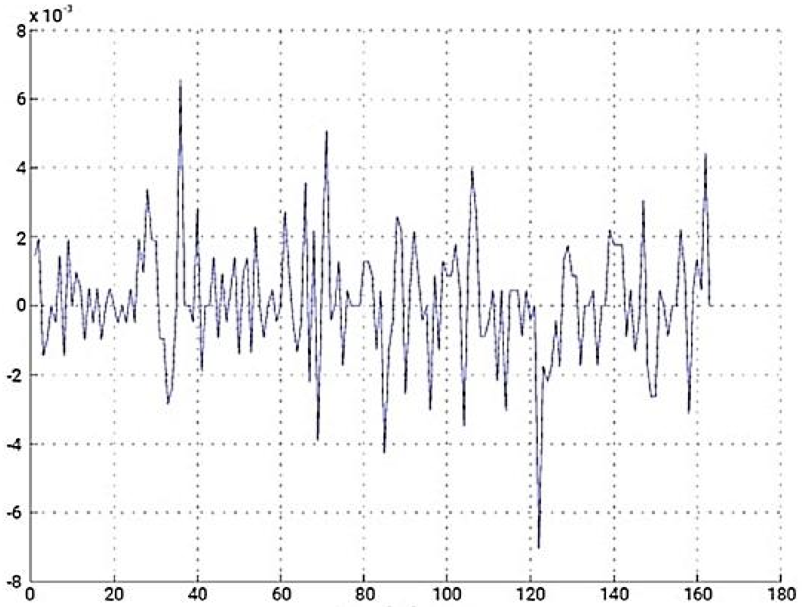


Figure 4 - Share price over the observed period of time

The equation you've provided represents the log return of an asset over a given time interval Δt. Here's the refined version of the equation with some added context:

Log Return for a Given Asset: For a given asset i, the return $R_i(t)$ over a time interval Δt is defined as the logarithmic difference between the price of the asset at time $t + \Delta t$ and the price at time t. Mathematically, it is expressed as:

$$\boldsymbol{R_i(t) = \ln\big(P_i(t + \Delta t)\big) - \ln\big(P_i(t)\big)} \qquad \textbf{(7)}$$

Where:

- $R_i(t)$ is the log return of asset i at time t,
- $P_i(t)$ is the price of asset i at time t,
- $P_i(t + \Delta t)$ is the price of asset i at time $t + \Delta t$,
- ln denotes the natural logarithm.

This equation gives the return of asset i over the time period from t to $t + \Delta t$, which is commonly used in financial models to calculate the change in asset price, accounting for compounding.

Figure 3 illustrates the time series of the price variation of the asset of the company Petrobrás, throughout the month of September. After applying equation 1, the time variation of the price of this asset for the referred month was obtained, represented in Figure 2.

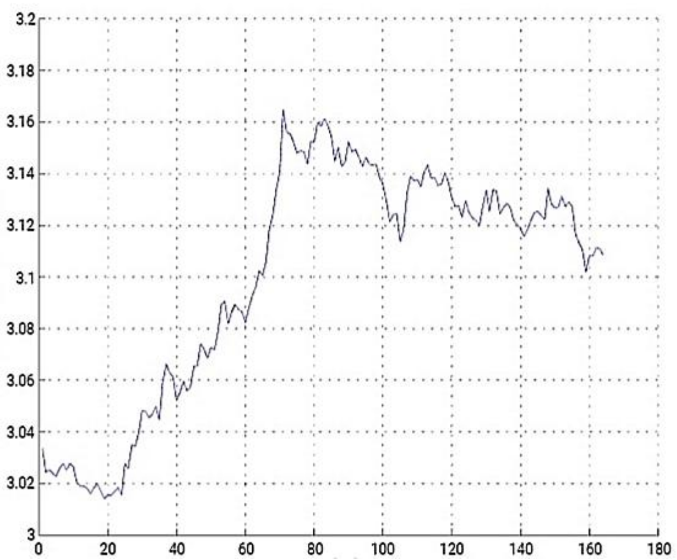


Figure 2. Variation in the price of shares over the observed period of time.

In probability theory and statistics, the correlation between two variables, in this case the returns for two stocks, is a measure of the strength and direction of the linear relationship between them. Figure 3 shows the return on the stock as a function of the return on the PETR3 stock for the same period of time.**[[8**

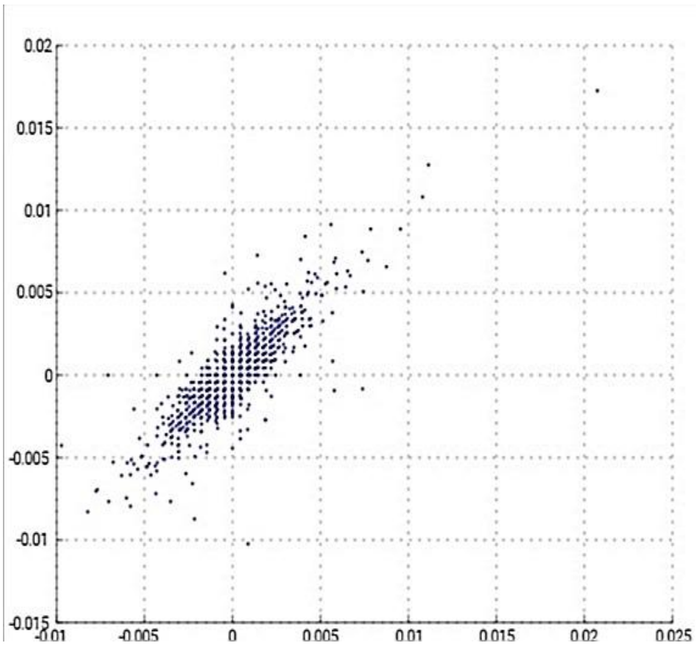


Figure 3. Dispersion of the variation in the price of stock as a function of the variation in the price of PETR3 stock.

It is evident from the scatter plot that the relationship between the returns of each pair of assets exhibits a significant linear characteristic. This is particularly expected in the case of two assets belonging to the same company. Such a relationship can be quantified using the **correlation coefficient** between the two variables, which is expressed by the following equation:

$$c_{ij} = \text{corr}\left(R_i(t), R_j(t)\right) = \frac{\mathbb{E}[(R_i(t) - \mu_i)(R_j(t) - \mu_j)]}{\sigma_i \cdot \sigma_j} \qquad (8)$$

Where:

- c_{ij} is the correlation coefficient between assets i and j,
- $R_i(t)$, $R_j(t)$ are the instantaneous returns of the assets at time t,
- μ_i, μ_j are the expected (mean) returns of each asset,
- σ_i, σ_j are the standard deviations of the returns,
- $\mathbb{E}[\cdot]$ denotes the expectation operator.

Due to its construction, the correlation coefficient is bounded within the interval $(-1,1)$, where:

- $c_{ij} = 1$ indicates a perfect positive linear correlation,
- $c_{ij} = -1$ indicates a perfect negative linear correlation,
- $c_{ij} = 0$ indicates no linear correlation between the variables.

In the case of the two **Petrobras** assets, a strong positive correlation was expected based on the scatter plot analysis. Empirically, this expectation is confirmed as the correlation coefficient between their returns is:

$$c_{ij} = 0.8453\text{…………………..……..(9)}$$

In practice, when dealing with a portfolio containing N assets, it is essential to compute the pairwise correlation coefficients for all combinations of asset pairs. A natural and effective way to represent these relationships is through the **correlation matrix**, defined as:

$$\mathbf{C}_{N\times N} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,N} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,N} \\ & & \ddots & \\ c_{N,1} & c_{N,2} & \cdots & c_{N,N} \end{bmatrix}\text{...............................(10)}$$

This matrix possesses the following properties:

- **Symmetry**: $c_{ij} = c_{ji}$,
- **Unit diagonal elements**: $c_{ii} = 1$ for all i, reflecting the perfect correlation of an asset with itself.

For a portfolio comprising N assets, the **correlation matrix C** $\in \mathbb{R}^{N\times N}$ is defined as a square matrix with N rows and N columns, where each element $c_{i,j}$ represents the Pearson correlation coefficient between the returns of asset i and asset j. This matrix is **symmetric**, which implies:

$$c_{i,j} = c_{j,i}, \quad \forall i,j \in \{1,2,\dots,N\} \text{……………………………..(11)}$$

Furthermore, the **main diagonal** of the matrix consists of unit values:

$$c_{i,i} = 1, \quad \forall i \text{……………………..(12)}$$

This arises from the fact that each asset is perfectly (i.e., maximally) correlated with itself.

Understanding and mapping the correlation structure among multiple assets is fundamental to modern portfolio theory, particularly in optimizing the trade-off between **risk and return**. However, empirical correlation coefficients are influenced by both **systematic (deterministic)** and **random (noise)** components. Disentangling these components is non-trivial and requires advanced statistical tools.

Spectral Analysis of the Correlation Matrix

A powerful method to analyze the internal structure of the correlation matrix and extract meaningful economic signals is through **eigenvalue decomposition**. The correlation matrix **C** can be decomposed as:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\top} \text{……………………..(13)}$$

Where:

- $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$ is a diagonal matrix containing the **eigenvalues** λ_i of **C**,
- $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N]$ is the matrix whose columns are the corresponding **eigenvectors** $\mathbf{v}_i$,
- $\mathbf{V}^{\top}$ denotes the transpose of **V**.

The eigenvalues and eigenvectors of the correlation matrix have widespread applications in various disciplines, including **statistics, physics**, and **financial economics**. In the context of portfolio management, they can be utilized to:

- Identify the most influential factors driving asset co-movements,
- Reduce portfolio dimensionality (via **Principal Component Analysis**),
- Distinguish between market-wide effects and asset-specific behavior.

This spectral decomposition enables a more nuanced understanding of systemic risk and hidden market structures that are not directly observable through raw correlation values alone. **[10]**

The **eigenvectors** of a square matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ are non-zero vectors $\mathbf{v}_k$ that satisfy the following linear transformation condition:

$$\mathbf{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k \qquad (14)$$

Where:

- $\mathbf{v}_k$ is the **eigenvector** associated with the transformation,
- $\lambda_k \in \mathbb{R}$ is the corresponding **eigenvalue**.

This relation implies that under the linear transformation defined by matrix **A**, the eigenvector $\mathbf{v}_k$ preserves its direction and is only scaled by the eigenvalue λ_k. This property is essential for the analysis of **stability and structure** in various systems—whether physical, biological, or economic.

The **eigenvalues** of a matrix are obtained by solving the **characteristic equation**:

$$\det(\mathbf{A} - \lambda_k \mathbf{I}) = 0 \qquad (15)$$

Where:

- $\det(\cdot)$ denotes the determinant,
- **I** is the identity matrix of the same dimension as **A**,
- λ_k are the roots of the characteristic polynomial and thus the eigenvalues of **A**.

When applied to the **correlation matrix C** of asset returns, eigenvalue decomposition becomes a powerful tool for isolating meaningful market dynamics. The correlation matrix inherently contains both **deterministic components**, which reflect genuine relationships among asset returns, and **random components**, which arise from data limitations and stochastic fluctuations.

There are **two main sources of randomness** in the eigenvalues of empirical correlation matrices:

1. **Non-stationarity of market conditions**: Financial correlations evolve over time, meaning that observed values may not represent stable, long-term relationships.
2. **Finite sample effects**: Since real-world datasets have limited length, estimations of correlation coefficients are contaminated by statistical noise, introducing spurious fluctuations.

Therefore, it is critical to **separate signal from noise** by identifying which eigenvalues reflect real economic structure and which are dominated by randomness. One prominent method to achieve this is the **Random Matrix Theory (RMT)**, which provides statistical benchmarks for distinguishing informative eigenvalues from those expected under a purely random correlation structure. .**[11]**

The central concept behind **Random Matrix Theory (RMT)** is the investigation of the statistical properties of matrices whose elements are randomly generated. This approach facilitates the derivation of **analytical bounds, probability distributions, and structural characteristics** for the eigenvalues of a generic random matrix **R**.

To apply RMT to a financial portfolio consisting of N assets, the method begins by generating N synthetic return time series, each with L observations. These synthetic series are sampled from a **normal distribution** with:

- Mean = 0
- Standard deviation = 1

To ensure comparability between the **random matrix** and the **empirical return matrix** derived from actual market data, each random series must be adjusted to match the statistical characteristics of its corresponding real-world asset. For an asset i with empirical mean μ_i and standard deviation σ_i, the **modified random return series** $r_m(t)$ is computed as:

$$r_m(t) = \mu_i + \sigma_i \cdot r(t) \qquad (16)$$

Where:

- $r(t)$ is the raw randomly generated return at time t,
- $r_m(t)$ is the adjusted random return used for matrix construction.

From the N modified random time series, one can compute the pairwise **correlation coefficients**, forming a square correlation matrix $\mathbf{H} \in \mathbb{R}^{N\times N}$, defined as:

$$\mathbf{H}_{N\times N} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N} \\ & & \ddots & \\ h_{N,1} & h_{N,2} & \cdots & h_{N,N} \end{bmatrix} \qquad (17)$$

This matrix is analogous to the empirical correlation matrix of actual asset returns, and it allows the application of spectral analysis techniques.

Once the random correlation matrix **H** is constructed, its **eigenvalues** λ and **eigenvectors** can be computed. The statistical distribution of the eigenvalues derived from such a random matrix is described by the **Marchenko–Pastur distribution**:

$$P_{\mathrm{RMT}}(\lambda) = \frac{Q}{2\pi} \cdot \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \quad \text{for } \lambda \in [\lambda_-, \lambda_+] \qquad (18)$$

Where:

- $Q = L/N$ is the **ratio** of time series length to the number of assets,
- λ_- and λ_+ are the **theoretical bounds** of the eigenvalue spectrum, defined as:

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \qquad (19)$$

These bounds represent the expected **eigenvalue limits** under purely random correlations. Any eigenvalue **outside** this range indicates the presence of **non-random (informative or structural)** market effects. **[12]**

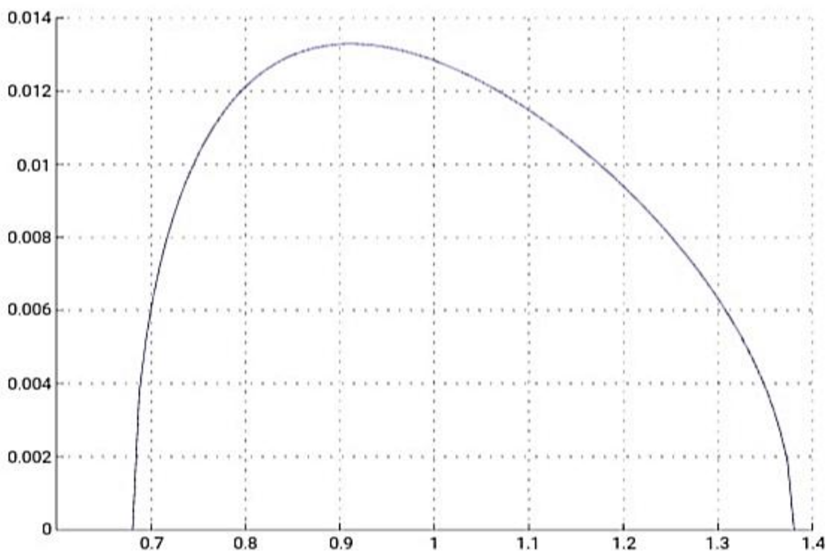


Figure 5. Theoretical distribution of eigenvalues of a random matrix.

Through this rule, it can be considered that eigenvalues outside this range should not be related to the random behavior of the correlation matrix but rather due to intrinsic characteristics of the system they represent. These limits were calculated considering the data from September and December and are available in Table 1.

Table 1. Theoretical bounds for the eigenvalues of a random correlation matrix.

Month	λ^+ (Upper Bound)	λ^- (Lower Bound)
September	1.3806	0.6807
December	1.3778	0.6826

These values are likely derived from the **Random Matrix Theory (RMT)** distribution bounds:

$$\lambda_{\pm} = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \text{………………………}(20)$$

Where $Q = \frac{L}{N}$, with:

- L: Number of observations in each time series
- N: Number of assets

These bounds are used to distinguish **eigenvalues containing random information** (within the interval $[\lambda^{-}, \lambda^{+}]$) from **eigenvalues potentially containing market (deterministic) information** (those outside this interval). That is:

- Eigenvalues **within** $[\lambda^{-}, \lambda^{+}]$ → likely **noise**
- Eigenvalues **outside** $[\lambda^{-}, \lambda^{+}]$ → potentially **informative**

:Observations

- The bounds are **very similar** across the two months, suggesting that the ratio $Q = \frac{L}{N}$ hasn't changed significantly.
- Small shifts in λ^{+} and λ^{-} may be due to slight changes in market conditions or in the length of the time series.

To ensure the **robustness and generalizability** of the findings obtained in this section, a total of **100 random correlation matrices** were simulated. This extensive simulation enabled validation of the relationships derived from random matrix theory, even when dealing with a **relatively large number of eigenvalues** (in this case, 5000). The observed results confirmed that the behavior of eigenvalues and eigenvectors from random matrices remains consistent under such conditions. **[13]**

The **scalar product (dot product)** is a fundamental operation between two vectors, which can be calculated using the following expression:

$$\vec{u} \cdot \vec{v} = \mid \vec{u} \mid \cdot \mid \vec{v} \mid \cdot \cos(\theta) \qquad (21)$$

Where:

- $\vec{u}$ and $\vec{v}$ are vectors in $\mathbb{R}^n$,
- $\mid \vec{u} \mid$ and $\mid \vec{v} \mid$ are the Euclidean norms (magnitudes) of the vectors,
- θ is the angle between them.

This operation can be interpreted geometrically as the **projection** of vector $\vec{u}$ onto vector $\vec{v}$. Consequently, the angle θ formed between these two vectors can be extracted by rearranging Equation (10), yielding:

$$\theta = \cos^{-1}\left(\frac{\vec{u}\cdot\vec{v}}{|\ \vec{u}\ |\cdot|\ \vec{v}\ |}\right) \qquad (22)$$

In the **special case** where the vectors point in the **same direction**, the angle $\theta = 0$, and the scalar product reaches its **maximum value**, equal to the product of the two magnitudes.

Based on this mathematical framework, one can evaluate the **temporal stability** of the eigenvectors of the **correlation matrix** by examining the angle between the eigenvectors computed at two different time points (e.g., across two consecutive months). If the **market structure** remains relatively unchanged, the **angle between corresponding eigenvectors** from one month to the next is expected to be **close to zero**.

This analytical method provides a **quantitative measure** of how the **direction and orientation** of the eigenvectors of the correlation matrix evolve over time. In particular, it serves as an indicator of the **stability or variability** in the **underlying economic or financial dynamics** represented by the correlation structure of asset returns. **[14]**

3 Results

In this section, the analysis will begin with the month of **September**, based on the extracted data as previously described. Utilizing the **100 simulated random correlation matrices**, **Figure 5** was generated to illustrate the distribution of eigenvalues derived from the random matrices. To facilitate a meaningful visual comparison, the scale of Figure 5 was adjusted to match that of **Figure 6**, which depicts the eigenvalue distribution obtained from the **empirical correlation matrix** constructed using actual return data. This comparison aims to identify deviations from the random matrix behavior, thereby isolating eigenvalues that potentially carry **economically meaningful information**.

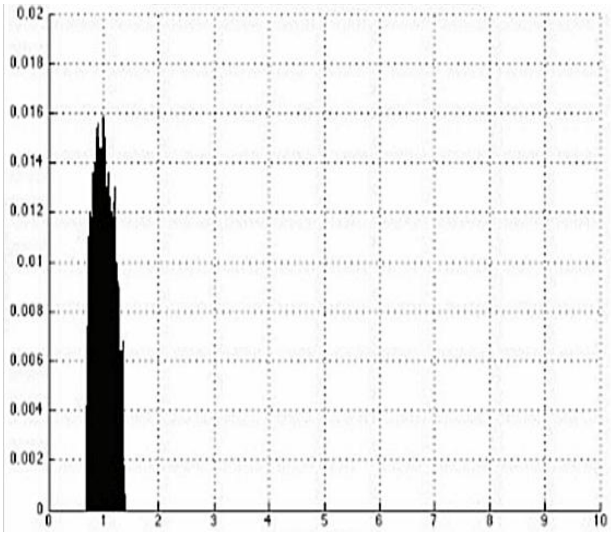
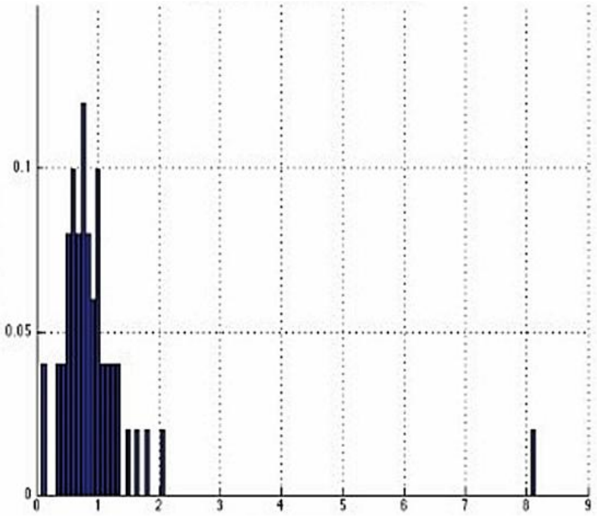


Figure 5. Distribution of eigenvalues Figure 6. Distribution of eigenvalues
of the simulated random matrix for September. .of September data



Following the methodology, the minimum (入) and maximum (1) values of the probability distribution function are given by equation 9, taking into account that in this study (month of September) there are 50 assets and 1634 observations for each, that is, n=50 and T=1635. Thus, we have the following relationship:**[15]**
λ- = 0.6807 e λ+ = 1.3806

These limits were then used to test the agreement of the eigenvalues of the real series with the Random Matrix Theory. When obtaining the largest eigenvalues for the real matrix, λ_{50} = 8.141.1 , λ_{49} = 2.060 and λ_{50} = 1.812, it is observed that none of them belong to the interval (they are all greater than the upper limit). Regarding the 3 smallest eigenvalues, they are: λ□ = 0.074, λ□ = 0.131 and λ_3 =0.352. These values are also all located below the lower limit found ,Another observation to be made is that the largest eigenvalue of the real matrix is about 5 times greater than the upper limit of the eigenvalues of the random matrix and also that 46% of the eigenvalues fall outside the range obtained, with 36% being lower than the minimum limit and 10% higher than the maximum limit.

Analyzing the largest eigenvalue of the September portfolio (λ_{50} = 8.14), we have the following eigenvector graph:

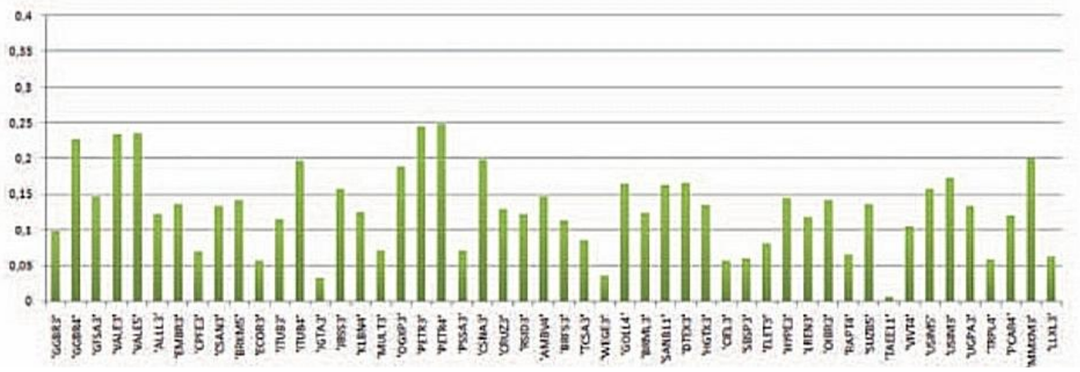


Figure 7. Eigenvectors of the largest eigenvalue of the September portfolio.

As can be seen, the portfolio representing the highest eigenvalue is purchased in almost all assets. This shows that the highest eigenvalue expresses a risk close to market risk. This eigenvector relates assets to the economic scenario. From it, it is possible to obtain an idea of how assets are influenced by shocks that impact the economy as a whole.

In this case, it can be observed that the company ""Transmissora Aliança de Energia Elétrica S.A." (TAEE11) is the least affected by economic shocks. From graph 1 it is possible to see that it is the company least correlated with the other assets. Thus, assuming that economic shocks tend to affect practically all sectors and all companies on the stock exchange and observing that TAEE11 has a low correlation with all other assets, it can be concluded that, regardless of the economic shock , its share prices will be the least affected.**[16]**

On the other hand, it can also be observed that Petrobras, represented in the portfolio by both PETR3 and , is the company that is most susceptible to market risk ,Now analyzing the eigenvectors resulting from the second and third largest eigenvalues (λ_{49} = 2.06 and λ_{48} = 1.81 respectively), we have the eigenvectors represented in figure 8 and figure 9, respectively.

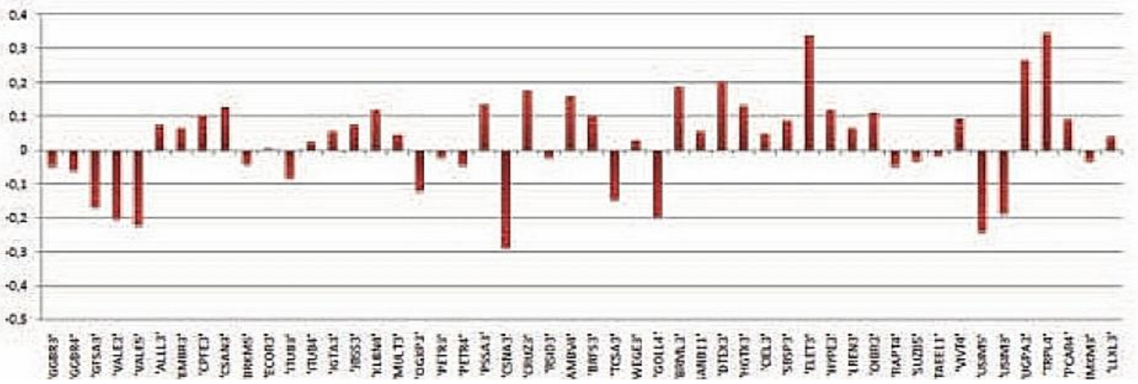


Figure 8. Eigenvectors of the second largest eigenvalue of the September portfolio.

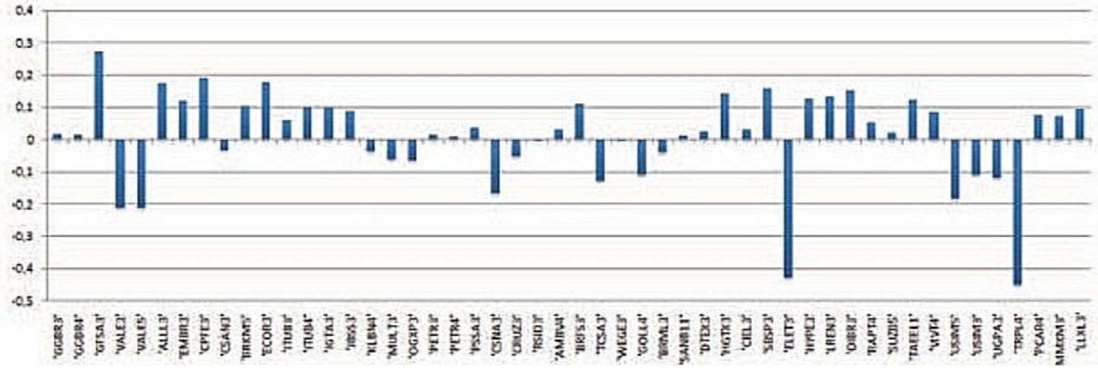


Figure 9. Eigenvectors of the third largest eigenvalue of the September portfolio.

As the eigenvalues decrease, the eigenvectors begin to represent portfolios with assets with similar correlations. For example, considering graph 3, it is possible to observe that the 4 most sold stocks are Vale (VALE3 and VALE5), Usiminas (USIM5) and Companhia Siderúrgica Nacional (CSNA3), which have a high correlation with each other and are all from the same sector, basic materials.

the random matrix test is performed for the American stock market, it is possible to separate eigenvalues that result in eigenvectors that represent portfolios of single segments. Below is the graph of this study showing the relationship between eigenvalues and segments of the American market **[17].**

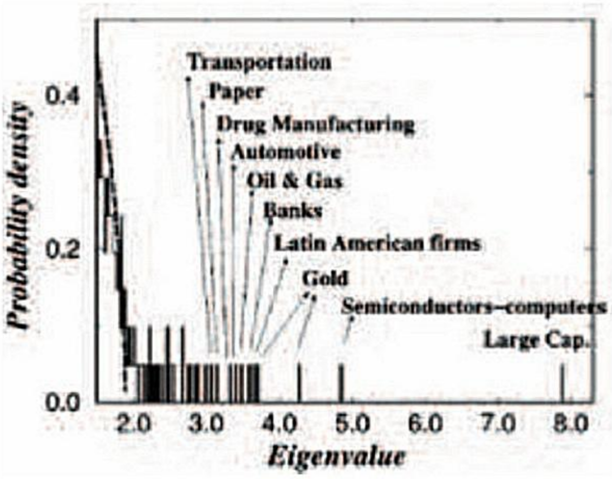


Figure 10. Eigenvalues and segments of the American market [2].

Comparing now the eigenvectors resulting from the three largest eigenvalues of the random matrix (λ = 1.30, λ = 1.32 and A = 1.36 respectively) we have the following graphs:

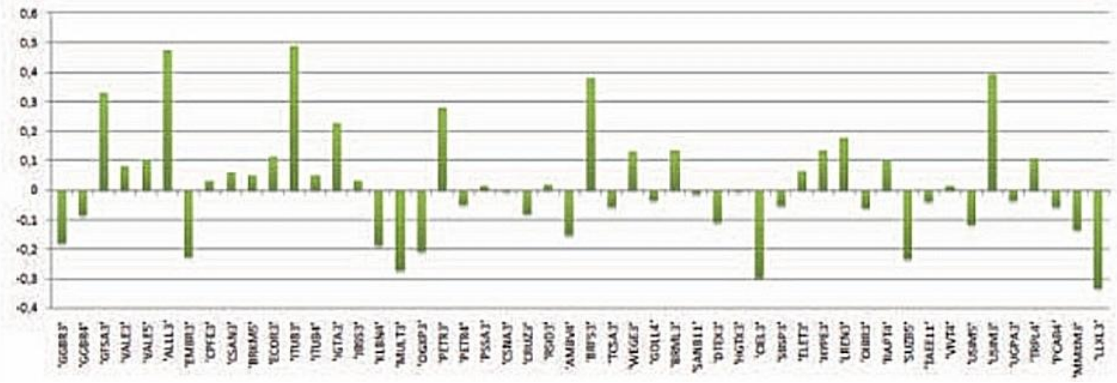


Figure 11. Eigenvectors of the largest eigenvalue of the random matrix portfolio.

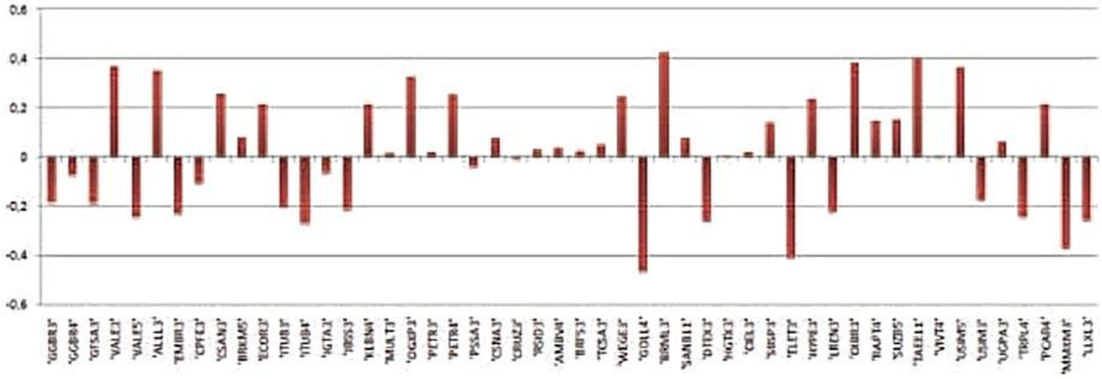


Figure 12. Eigenvectors of the second largest eigenvalue of the random matrix portfolio.

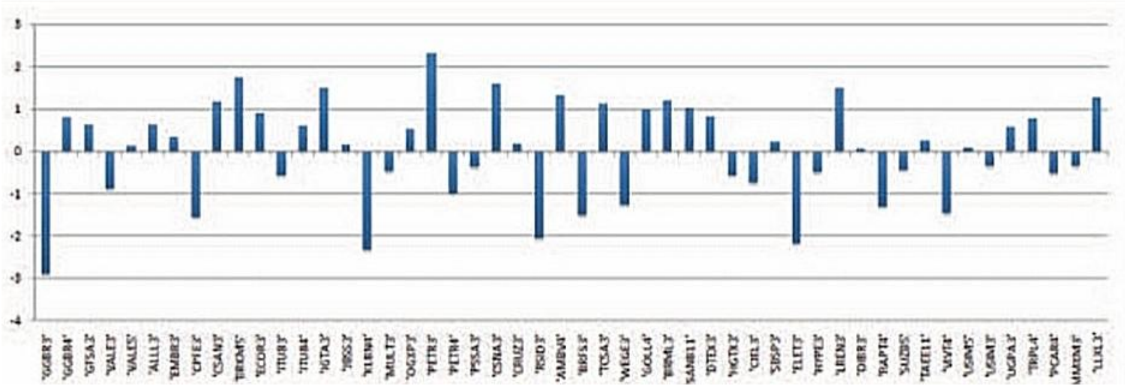


Figure 13. Eigenvectors of the third largest eigenvalue of the random matrix portfolio.

As can be seen, these eigenvectors are not capable of representing market risk. In fact, in Figure 12, the portfolio buys VALE3 and sells VALES, which would reduce the portfolio risk and mean that an economic shock would negatively affect VALE's common shares and positively affect its preferred shares, which is not consistent with the real situation. Furthermore, it is not possible to make any type of inference about the companies' sectors by observing only the eigenvectors Now, analyzing the graphs of the eigenvectors associated with the 3 smallest eigenvalues (λ□ = 0.07, λ_2 = 0.13 λ_3 = 0.35 respectively), we have figures 14, 15 and 16.**[[18**

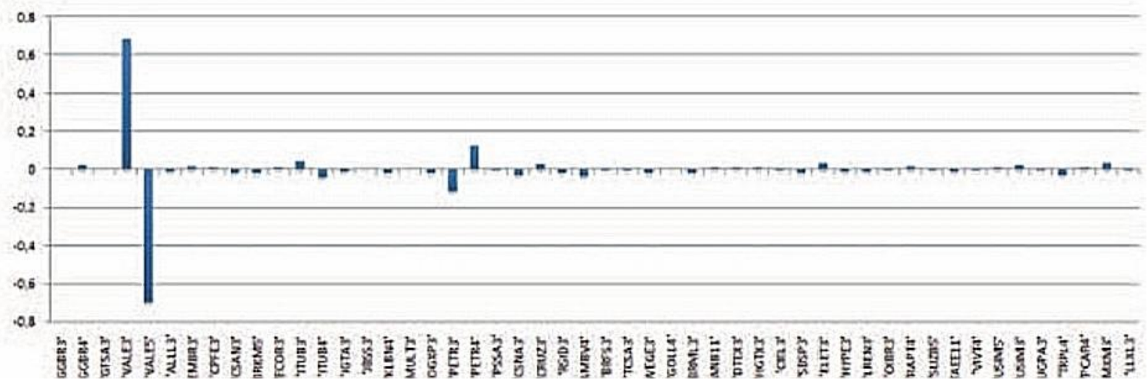


Figure 14. Eigenvectors of the smallest eigenvalue of the September portfolio.

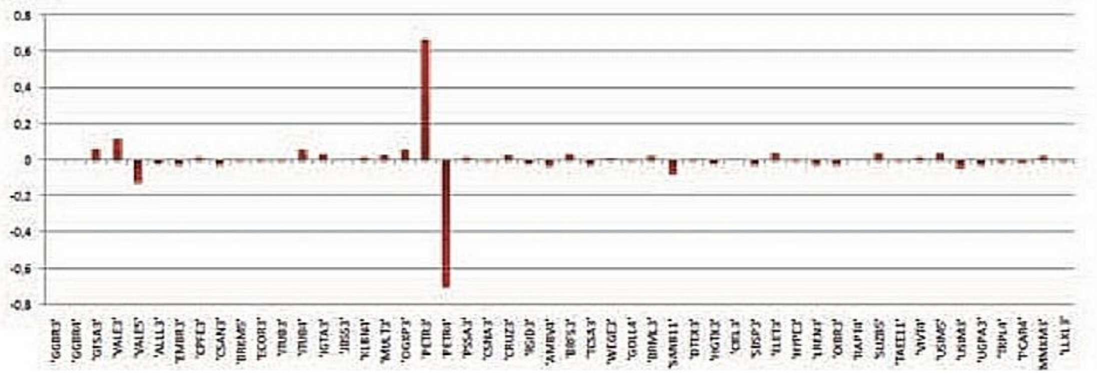


Figure 15. Eigenvectors of the second smallest eigenvalue of the September portfolio.

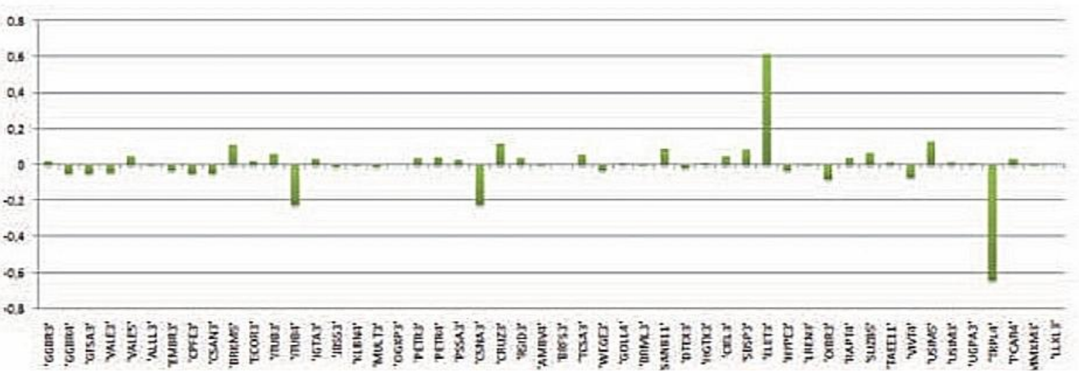


Figure 16. Eigenvectors of the third smallest eigenvalue of the September portfolio.

Portfolios constructed from the lowest eigenvalues are low-risk long-short portfolios. Figure 14, related to eigenvalue λ□ = 0.07, is a portfolio bought in VALE3 and sold in VALE5, which are two different types of VALE shares; graph 8, eigenvalue λ□ = 0.13, is a portfolio bought in PETR3 and sold in , which are also two different types of shares of the same company (Petrobrás); and finally, graph 9, eigenvalue λ3 = 0.35, represents a portfolio bought in ELET3 (Eletrobrás) and sold in TRPL4 (Transmissão Paulista), which, despite being different companies, are companies in the same sector of activity, public utility, and the performance of one company directly affects the performance of the other.**[19]**

Table 2 .The correlation matrix shows that these are portfolios of highly correlated assets, in fact, these are the most correlated assets in the sample of 50 shares.

	VALE3	VALE5	PETR3	PETRA	ELET3	TRPLA
VALE3	1	0.914	0.456	0.446	0.130	0.094
VALE5	0.914	1	0.444	0.462	0.143	0.073
PETR3	0.456	0.444	1	0.845	0.131	0.094
PETRA	0.446	0.462	0.845	1	0.143	0.087
ELET3	0.130	0.143	0.131	0.143	1	0.611
TRPLA	0.094	0.073	0.094	0.087	0.611	1

Analyzing the 3 smallest eigenvalues of the random matrix (λ = 0.70, λ = 0.72 and λ = 0.74 respectively), we have the following eigenvectors represented in figures 17, 18 and

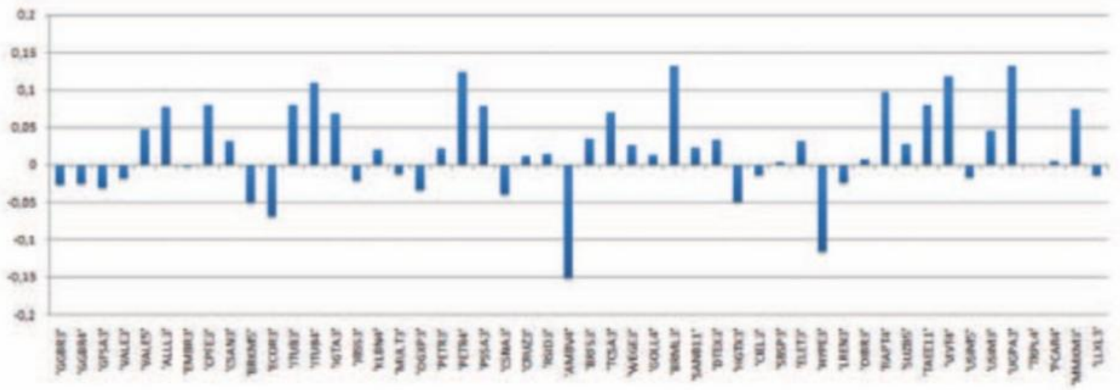


Figure 17. Eigenvectors of the smallest eigenvalue of the random matrix portfolio.

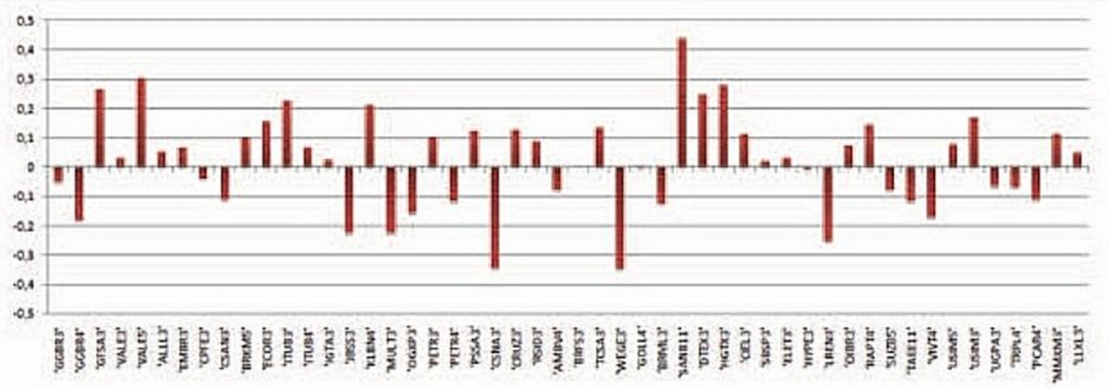


Figure 18. Eigenvectors of the second smallest eigenvalue of the random matrix portfolio.

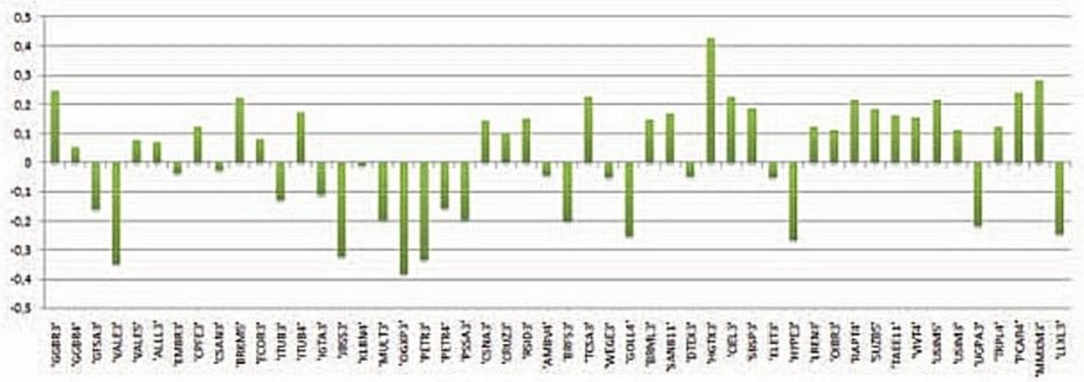
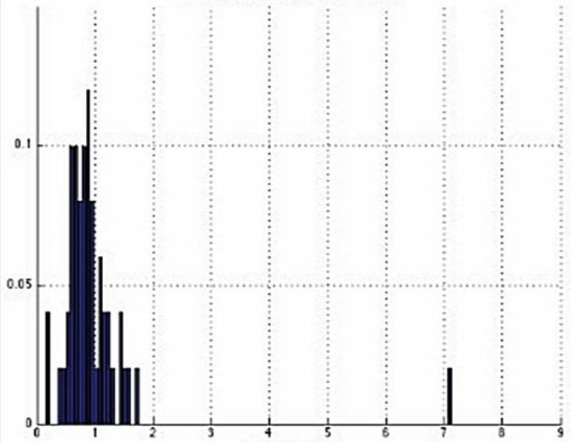


Figure 19. Eigenvectors of the third smallest eigenvalue of the random matrix portfolio.

As can be seen, these eigenvectors do not represent only the assets with the highest correlations and all of these portfolios have positions in practically all of the stocks in the study. Furthermore, it is practically impossible to perceive any difference between these portfolios and the portfolios constructed from the three highest eigenvalues of the random matrix **[20].**

Carrying out the same procedure as for the month of September, graph 20 of the eigenvalues of the random matrix was generated, and graph 21, representing the distribution of the eigenvalues of the real series

1



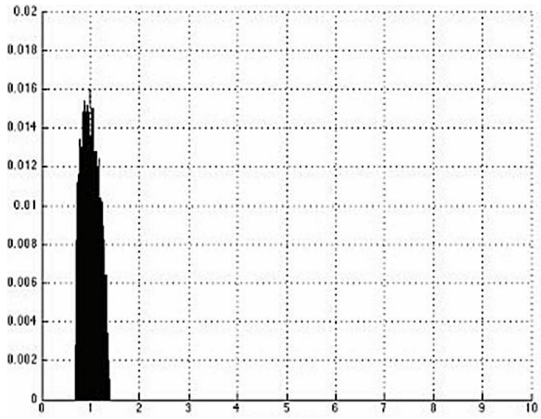


Figure 20. Distribution of eigenvalues
Figure 21. Distribution of eigenvalues
of the simulated random matrix
of December data.
for December.

The minimum (λ-) and maximum (λ+) values of the probability distribution function given by equation 9, taking into account that for the month of December the same 50 assets were used as for the month of September, but having 1657 observations for each, that is, n=50 and T=1635, are:

λ- = 0.6826 , λ+ = 1.3778

These values are very similar to those found in September. Also analyzing the agreement of the eigenvalues of the real series with the Random Matrix Theory, when obtaining the largest eigenvalues, λ_{50} = 7.123, λ_{49} = 1.712 and λ_{48} = 1.582, we see that none of them belong to the interval (they are all greater than the upper limit). Regarding the 3 smallest eigenvalues, they are:λ =0.149, λ, = 0.2181 and λl =0.376. These values are also all located below the lower limit found ,The same finding from September also applies here. The largest eigenvalue of the real matrix is more than 5 times larger than the upper limit of the eigenvalues of the random matrix. It is also observed that 40% of the eigenvalues are located outside the obtained range, with 28% below the minimum limit and 12% above the maximum limit **[21].**

4 iscussionD

A **correlation matrix** that captures the price variations of different assets provides valuable insight into the market relationships between companies. By utilizing such matrices, it becomes possible to simulate the returns of a portfolio given a price shock affecting one or more assets.

Consider the correlation matrix for three assets—**A**, **B**, and **C**—as follows:

$$M = \begin{bmatrix} 1 & 0.2 & 0.9 \\ 0.2 & 1 & -0.3 \\ 0.9 & -0.3 & 1 \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(23)$$

This matrix indicates the following relationships:

- A correlation of **0.2** between assets A and B means that, if asset A appreciates by 10%, asset B is expected to appreciate by **2%**.
- A correlation of **0.9** between assets A and C implies that, if asset A appreciates by 10%, asset C is likely to appreciate by **9%**.

- A correlation of **-0.3** between assets B and C suggests that an increase in asset B’s price is expected to lead to a decrease in asset C’s price, and vice versa.

For a portfolio consisting of **50% asset B** and **50% asset C**, assuming a 10% increase in asset A’s price, the hypothetical portfolio's return can be calculated as follows:

$$\text{Portfolio Return} = 0.5 \times 2\% + 0.5 \times 9\% = 5.5\%$$

This relationship is crucial for assessing the potential **spillover effects** of economic shocks. Specifically, if a significant price change occurs in one sector (represented by an asset), it is essential to understand how this shock can **contaminate** or influence other sectors of the economy, as reflected through the interrelationships between asset returns.

Accelerating Portfolio Shock Calculations Using Matrix Multiplication

To enhance the efficiency of calculating the final result of a price shock to a portfolio and to facilitate simulating shocks across different assets, the **correlation matrix** can be multiplied by a matrix representing the weights of the assets in the portfolio. This approach significantly reduces computational complexity and aids in quick simulations.

For instance, consider a portfolio comprising **50% of asset A**, **20% of asset B**, and **30% of asset C**. The weight matrix for this portfolio is represented as follows: **[22]**

$$P_{3,1} = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix} \text{…………………………(24)}$$

Multiplying this weight matrix by the **correlation matrix** of the assets, we obtain the following result:

$$r_{3,1} = \begin{bmatrix} 0.81 \\ 0.21 \\ 0.69 \end{bmatrix} \text{………………………..(25)}$$

This implies that:

- If **asset A** appreciates by 10%, the portfolio is expected to appreciate by **8.1%**.
- If **asset B** appreciates by 10%, the portfolio is expected to appreciate by **2.1%**.
- If **asset C** appreciates by 10%, the portfolio is expected to appreciate by **6.9%**.

To further simplify the calculation of the final result, price shocks can be represented in a matrix form and then multiplied by the previously obtained result matrix. For example, if a shock of **10%**, **15%**, and **-10%** occurs for assets A, B, and C, respectively, the shock matrix is:

$$z = \begin{bmatrix} 0.1 \\ 0.15 \\ -0.1 \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(26)$$

Multiplying the shock matrix by the portfolio impact matrix:

$$\begin{bmatrix} 0.1 \\ 0.15 \\ -0.1 \end{bmatrix} \cdot \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix} = 0.044 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(27)$$

This results in a **4.4%** appreciation of the portfolio, demonstrating how the combined shocks affect the overall portfolio return.

Eigenvectors and Portfolio Risk

In a specific scenario where the portfolio is represented by an **eigenvector**, the relationship can be expressed as follows:

$$HN_{NxN} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ & & \ddots & \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(28)$$

When this matrix is multiplied by the eigenvector (e.g., B_1, B_2, ..., B_n), it represents the weight distribution of the portfolio. The impact of a shock on any asset can be calculated using the relationship $\lambda \times \upsilon$, where:

- λ is the eigenvalue (a measure of risk), and
- υ is the eigenvector representing the portfolio weights.

This relationship suggests that the **eigenvalue** can be interpreted as a measure of risk: the higher the eigenvalue, the greater the **impact** of shocks on the portfolio's overall return. **[23]**

In this study, it is crucial to assess whether the **eigenvalues** and **eigenvectors** of the correlation matrices of returns exhibit temporal stability. Temporal stability implies that the patterns and relationships observed in historical data can be used to extrapolate and predict future market behavior. This section focuses on comparing the eigenvectors from the **September** and **December** correlation matrices, specifically comparing the maximum and minimum eigenvectors. The goal is to understand whether the market dynamics, as represented by these eigenvectors, remain consistent over time.

To investigate the temporal stability, comparisons will be made between the sets of **maximum eigenvalues** (λ48, λ49, λ50) and **minimum eigenvalues** (λ1, λ2, λ3) from the correlation matrices of returns observed in the months of **September** and **December**. These eigenvalues and their corresponding eigenvectors represent distinct aspects of the market's structure. The maximum eigenvalues typically correspond to the principal components of the data, capturing the largest variations in the returns, while the smallest eigenvalues correspond to the less significant components, reflecting noise or smaller-scale variations.

The comparisons will be conducted using two methods:

- **Graphical Representation**: The eigenvectors corresponding to the maximum and minimum eigenvalues will be visualized in graphical form. This allows for an intuitive comparison of their behavior across the two months. Any significant differences in the structure of the eigenvectors would indicate a lack of temporal stability.
- **Angle Between Eigenvectors**: As discussed in Section 3, the temporal stability of eigenvectors will be quantified by calculating the **angle** between the two eigenvectors being compared. The angle between two vectors provides a measure of their directional similarity. A smaller angle indicates that the eigenvectors are closely aligned, suggesting that the temporal stability is high, while a larger angle suggests greater variation between the eigenvectors over time.

The **maximum eigenvalues** (λ48, λ49, λ50) are typically associated with the dominant market factors or principal components. These eigenvalues are more likely to show stable patterns over time, as they represent significant underlying market behaviors or correlations between assets. On the other hand, the **minimum eigenvalues** (λ1, λ2, λ3) may exhibit more volatility, as they are often linked to smaller, noise-driven variations that could change more significantly across time periods. By examining both sets of eigenvectors and their corresponding eigenvalues, the study aims to determine whether the market's principal components and less significant fluctuations remain stable between **September** and **December**, providing insights into the potential for extrapolating historical data to predict future market behavior.

The largest eigenvalue λ_{50} has its eigenvectors presented in the graph illustrated in Figure 22. For this, the eigenvalue corresponding to the month of September was 8.141 and for the month of December it was 7.123. According to the interpretation that the largest eigenvalue represents the market portfolio, a reduction of approximately 12% was observed.

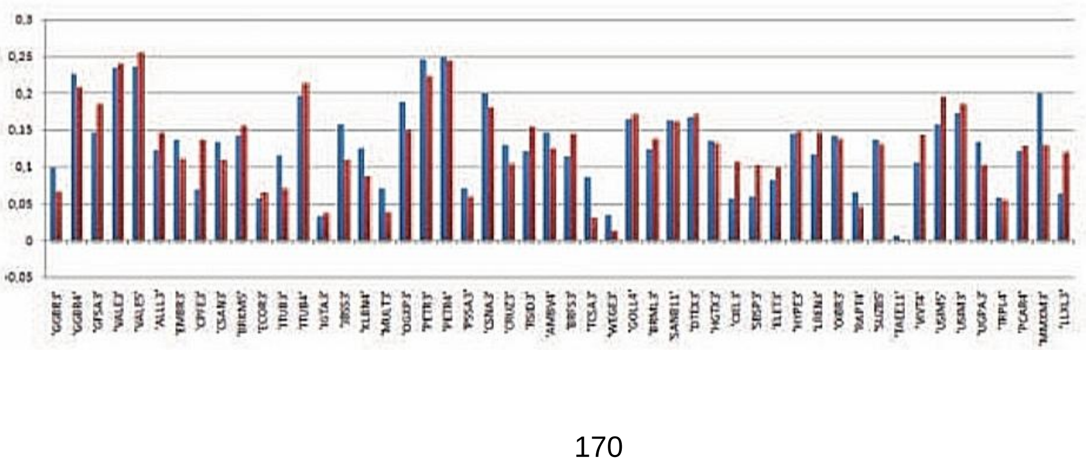


Figure 22. Eigenvectors corresponding to the eigenvalue 50 for the months of September and December.

From Figure 22, it can be seen that the market representative portfolio behaved in a similar way. When comparing these two months, it was possible to verify from equation 11 that the angle between these two vectors is approximately 12°, that is, these eigenvectors are close to alignment. It can be concluded that there were no significant variations in the market in this time interval to the point of impacting the portfolio and risk, for eigenvalue λ_{49}, a reduction in portfolio risk is observed similar to that which occurred for the highest eigenvalue. In this case, this parameter was equal to 2.060 for the month of September and 1.712 for the month of December, with these data corresponding to a reduction of 17% **[24].**

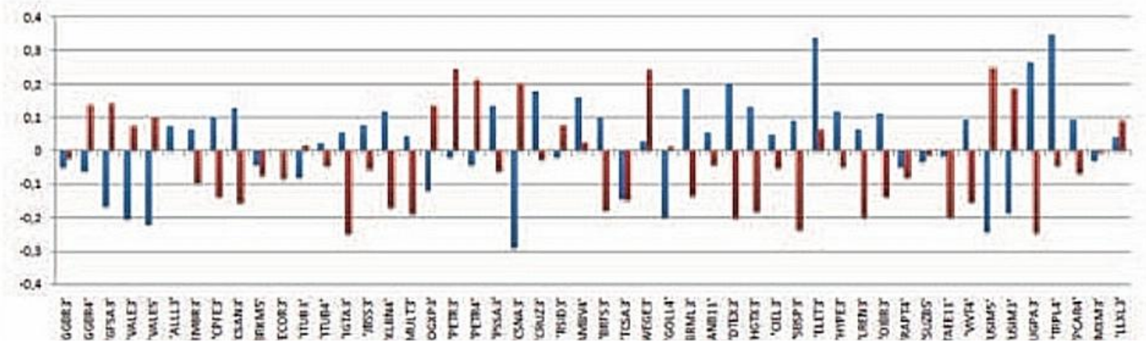


Figure 23. Eigenvectors corresponding to eigenvalue 49 for the months of September and December.

Figure 23 illustrates the eigenvectors corresponding to these eigenvalues for the two months of analysis. By inspecting the graph, it can be seen that there were not many inversions in the portfolio, although reductions in the sales of some assets are suggested, such as Vale and an increase in the purchase of Iguatemi shares. There are inversions in portfolio operations, of which Eletrobrás stands out.

The angle between the two eigenvectors was approximately 56°, indicating that there were significant changes in the set of operations in the portfolio. In other words, although the directions of the operations remained constant in the vast majority of assets, the weights of each of the operations were significantly changed.**[25]**

For eigenvalue λ_{48}, it is observed that this eigenvalue was equal to 1.812 in September, while for December its value was 1.582. In this case, a reduction of approximately13%.was observed

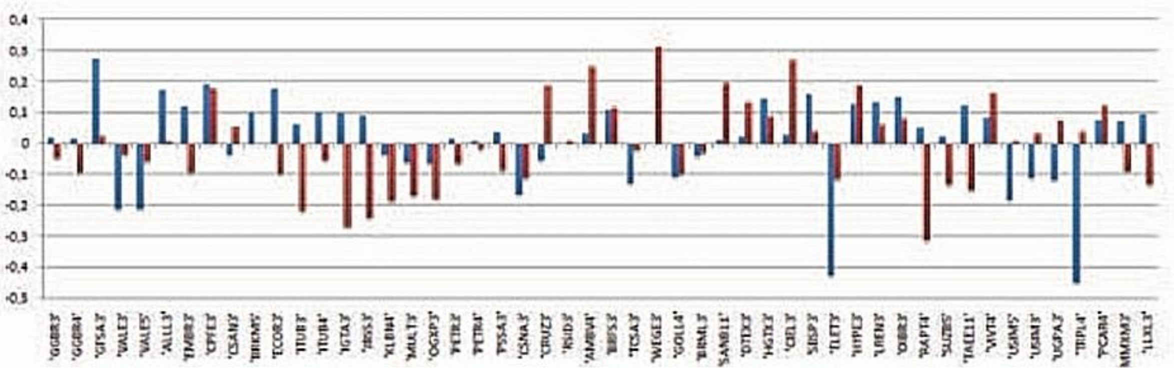


Figure 24. Eigenvectors corresponding to eigenvalue 48 for the months of September and December,

The graphical representation of these two eigenvectors is illustrated in Figure 24. In this case, there were significant changes in the composition of the portfolio, presenting the most unstable behavior among all the other eigenvectors studied to date. One of the hypotheses considered is that these eigenvalues are approaching the band that defines the region that behaves like eigenvalues of a random matrix and, due to this, presents greater volatility. In this case, the angle calculated between the two vectors was approximately 85°, revealing almost an orthogonality characteristic between them **[26].**

Next, analyses similar to those performed for the largest eigenvalues will be carried out, now considering the three smallest eigenvalues λ_1, λ_2 and λ_3. For these cases, the eigenvectors were grouped according to the distribution of their portfolio.

Taking these new groupings into account, the eigenvalues λ_2 for September and λ□ for December will be analyzed, with their values being 0.131 and 0.149. In this case, there was an increase in the risk associated with the portfolio in this period of 14%.

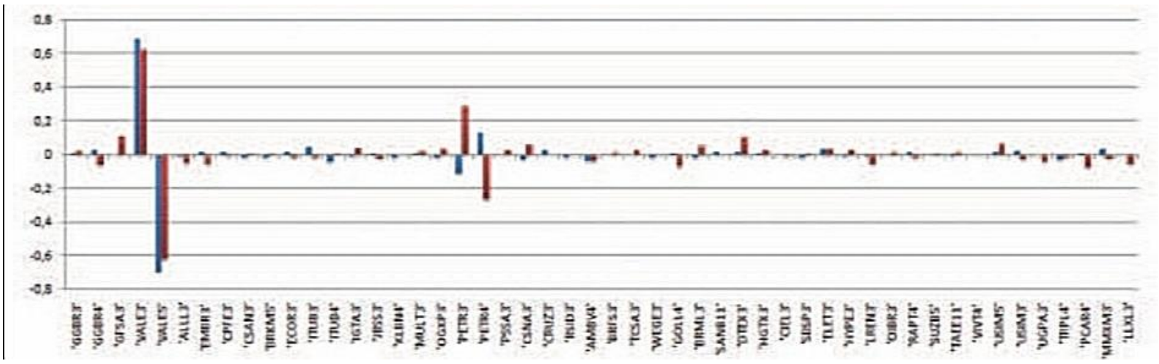


Figure 25. Eigenvectors corresponding to eigenvalue 2 and 1 for the months of September and December, respectively.

Figure 25 shows the graph of the eigenvectors related to these eigenvalues. It can be seen that for Petrobras assets (PETR3 and), behavior remained stable, with a small reduction in the purchase and sale of these, respectively. The biggest change occurred for Vale assets, with the December portfolio ceasing to buy the VALE3 asset and starting to buy the VALE5 asset. This instability may be due to the high correlation between these two assets. For these two eigenvectors, it is possible to observe a change of 36° in the angle between them, caused mainly by the inversion of the purchase/sale operations of Vale assets **[27].**

The second pair of eigenvalues to be analyzed will be λ□ for September and λ□ for December, which have values of 0.074 and 0.218. In this case, there was an increase in the associated risk of 94%, which may be a result of the increase in the correlation coefficient between the assets VALE5 and from 0.46 to 0.53.

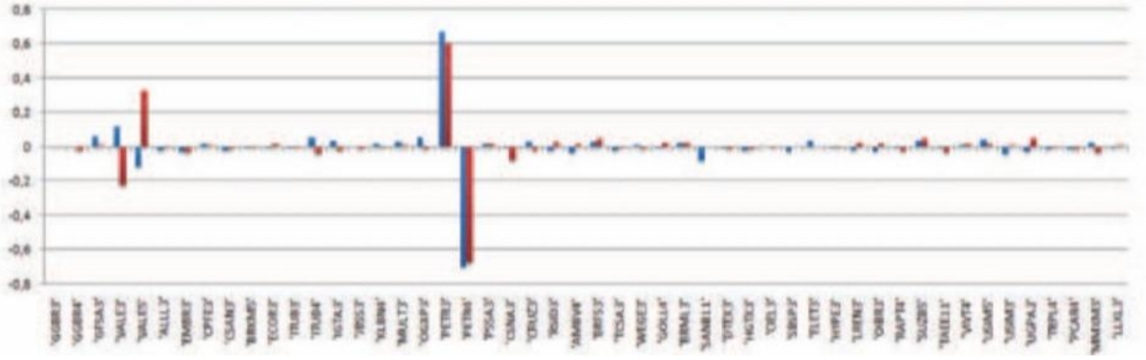


Figure 26. Eigenvectors corresponding to eigenvalue 1 and 2 for the months of September and December, respectively.

The eigenvector graph is illustrated in Figure 26. One can observe stability in relation to Vale's assets (VALE3 and VALE5), which represent the majority of the operations in this portfolio. In this case, one observes a behavior similar to the eigenvectors analyzed in Figure 25, with the difference that the inversion of operations occurred for Petrobras assets. The angle between the eigenvectors is 38°, and their misalignment was caused mainly by the inversion of operations for Petrobras assets.

The portfolios corresponding to eigenvalues 3 for the months of September and December presented the greatest instability among the cases analyzed for the lowest eigenvalues, with their values being equal to 0.352 and 0.376 respectively.

The graph illustrated in Figure 27 shows that in September the portfolio was mainly composed of assets from Eletrobrás and Transmissão Paulista de Energia. In this month, the correlation between these two assets was the third highest considering the entire portfolio. Although they are not assets from the same company, as mentioned previously, they have a high correlation, as they refer to the same electricity generation and transmission sector. For the month of December, it can be seen that the most correlated stocks are those related to the company Usiminas.

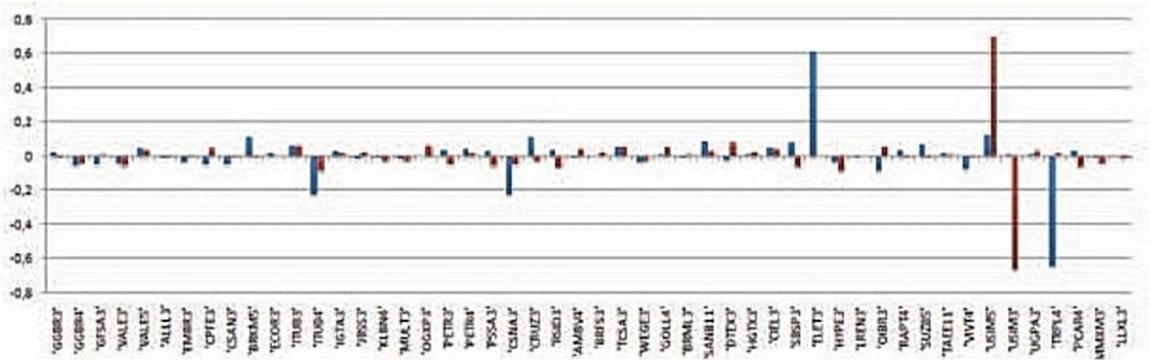


Figure 27. Eigenvectors corresponding to eigenvalues 3 for the months of September and December.

By calculating the angle, it is also possible to verify the situation of great instability, since its value is approximately 84", indicating a condition close to orthogonality for the two vectors [28].

5 Conclusion

The application of **Random Matrix Theory (RMT)** to high-dimensional statistics has significantly advanced the field of data analysis by providing essential tools for managing large and complex datasets. RMT has facilitated a deeper understanding of the spectral properties of large random matrices, leading to improvements in **covariance matrix estimation**, **principal component analysis (PCA)**, and **hypothesis testing** methods. These advancements are crucial in addressing the challenges posed by high-dimensional data, where traditional statistical techniques often struggle to deliver accurate results.

As the size and complexity of data continue to increase, the importance of RMT in high-dimensional statistics is expected to grow. Future research will likely explore its applications in new fields, such as **deep learning**, **network analysis**, and **signal processing**. By integrating RMT-based approaches with contemporary statistical and machine learning methods, researchers will be able to develop more efficient and reliable models, ultimately enhancing decision-making across various scientific and industrial domains.

The analysis of the largest eigenvalues revealed that the eigenvector associated with the largest eigenvalue effectively represents a portfolio that holds all assets in equal proportion, thus capturing the overall **market risk** and reflecting how assets are related to the broader economic environment. Additionally, as the eigenvalues decreased, the corresponding eigenvectors began to represent portfolios with assets that shared similar correlations. This study, which focused on 50 assets, suggests that with a broader set of stocks, eigenvalues could more clearly discriminate between specific market segments. In contrast, the smallest eigenvalues represented portfolios with minimal risk, aligning with **long-short strategies** involving companies within the same sector.

When extending this analysis to the month of **December**, no major changes were observed in the market relations. The properties of the eigenvalues (and consequently the eigenvectors) derived from RMT remained consistent, confirming that temporal stability exists in the short term. This stability supports the viability of using RMT for **market forecasting** and risk analysis, enabling its continued application in modeling future market behaviors.

ceReferen

1. *Altmeyer, R., Cialenco, L., & Pasemann, G. (2020). Parameter estimation for semilinear SPDEs from local measurements. arXiv preprint, arXiv:2004.14728.*
2. *Ambainis, A., Harrow, A. W., & Hastings, M. B. (2012). Random tensor theory: Extending random matrix theory to mixtures of random product states. Communications in Mathematical Physics, 310(1), 25-74.*
3. *Araya, H., & Tudor, C. A. (2021). Asymptotic expansion for the quadratic variations of the solution to the heat equation with additive white noise. Stochastics and Dynamics, 21(2).*
4. *Artstein, S., Ball, K. M., Barthe, R., & Nassor, A. (2004). On the convergence rate of the entropic central limit theorem. Probability Theory and Related Fields, 129(3), 381-390.*
5. *Ayache, A. (2020). Lower bound for local oscillations of Hermite processes. Stochastic Processes and their Applications, 130(8), 4593-4607.*
6. *Cialenco, I., & Kim, H.-J. (2022). Parameter estimation for discretely sampled stochastic heat equation driven by space-only noise. Stochastic Processes and their Applications, 143(5), 1-30.*
7. *Dhoyer, R., & Tudor, C. A. (2023). Limit behavior in high-dimensional regime for the Wishart tensors in Wiener chaos. Preprint.*
8. *Diez, Ch.-Ph., & Tudor, C. A. (2021). Limit behavior for Wishart matrices with Skorohod integrals. ALEA Lat. Am. J. Probab. Math. Stat., 18(2), 1625-1641.*
9. *Diez, Ch.-Ph., & Tudor, C. A. (2021). Non-central limit theorem for large Wishart matrices with Hermite entries. Journal of Stochastic Analysis, 2(1).*
10. *Es-Sebaiy, K., & Tudor, C. A. (2011). Noncentral limit theorem for cubic variation of a class of self-similar stochastic processes. Theory of Probability and Its Applications, 55(3), 411-431.*
11. *Fang, X., & Koike, Y. (2020). New error bounds in multivariate normal approximations via exchangeable pairs with applications to Wishart matrices and fourth moment theorems. Preprint.*
12. *Gamain, J., & Tudor, C. (2022). Exact variation and drift parameter estimation for the nonlinear fractional stochastic heat equation. Japanese Journal of Statistics and Data Science.*
13. *Gamain, J., & Tudor, C. (2022). Random matrices and the stochastic wave equation. Submitted.*
14. *Gamain, J., & Tudor, C. (2023). Limit behavior in high-dimensional regime for Wishart tensors with Rosenblatt entries. Submitted.*

15. *Garino, V., Nourdin, I., Nualart, D., & Salamat, M. (2021). Limit theorems for integral functionals of Hermite-driven processes. Bernoulli, 27(3), 1764-1788.*
16. *Gaudlitz, S., & Reiss, M. (2022). Estimation for the reaction term in semi-linear SPDEs under small diffusivity. arXiv preprint, arXiv:2203.10527.*
17. *Halconruy, H. (2021). Calcul de Malliavin et structures de Dirichlet pour des variables aléatoires indépendantes. Retrieved from https://www.theses.fr*
18. *Harnett, D., & Nualart, D. (2018). Central limit theorem for functionals of a generalized self-similar Gaussian process. Stochastic Processes and their Applications, 128(2), 404-425.*
19. *Hildebrandt, F., & Trabs, M. (2021). Parameter estimation for SPDEs based on discrete observations in time and space. Electronic Journal of Statistics, 15(1), 2716-2776.*
20. *Huang, J., Nualart, D., & Viitasaari, L. (2020). A central limit theorem for the stochastic heat equation. Stochastic Processes and their Applications, 130(12), 7170-7184.*
21. *Janák, J. (2021). Parameter estimation for stochastic wave equation based on observation window. Acta Applicandae Mathematicae, 172.*
22. *Jiang, T., & Xie, J. (2019). Limiting behavior of largest entry of random tensor constructed by high-dimensional data. Journal of Theoretical Probability, 33(1), 1-21.*
23. *Mikulincer, D. (2020). A CLT in Stein's distance for generalized Wishart matrices and higher-order tensors. International Mathematics Research Notices, 2022(10), 7839-7872.*
24. *Nourdin, I., & Pu, F. (2022). Gaussian fluctuation for Gaussian Wishart matrices of overall correlation. Statistics & Probability Letters, 181(2).*
25. *Nourdin, I., & Zheng, G. (2021). Asymptotic behavior of large Gaussian correlated Wishart matrices. Journal of Theoretical Probability, 35(3), 1-30.*
26. *Racz, M. Z., & Richey, J. (2018). A smooth transition from Wishart to GOE. Journal of Theoretical Probability, 32(2), 898-906.*
27. *Shevchenko, R., Slaoui, M., & Tudor, C. (2020). Generalized k-variations and Hurst parameter estimation for the fractional wave equation via Malliavin calculus. Journal of Statistical Planning and Inference, 207(3), 155-180.*
28. *Shi, X., Qiu, R., He, X., Ling, Z., Yang, H., & Chu, L. (2020). Early anomaly detection and localization in distribution networks: A data-driven approach. arXiv preprint, arXiv:1801.01669v4.*