Investigation of methods for resolving statistical noise and understanding correlation structure in high dimensional data

ELEN90094: Large Data Methods & Applications

Mukul Chodhary[1] Luca Di Cola[2]

I. Introduction

Financial markets are complex, dynamical systems characterised by intricate interconnections and hidden variables that challenge traditional modelling approaches. The complexity of these markets originates from various factors: nonlinearity, volatility, the influence of multiple macro and microeconomic forces, global events, human behaviour, and the inherent noise in financial data. Indeed, the fundamental forces driving daily market transactions often remain elusive. Despite these challenges, a comprehensive understanding of market behaviour is crucial for creating well-performing portfolios, developing effective financial tools, and informing economic policy.

The COVID-19 pandemic presents a unique opportunity to study how global financial markets respond to an unprecedented, widespread shock. This extraordinary event triggered simultaneous supply and demand disruptions across various sectors and regions, providing a rare chance to examine market adaptation and evolution in the face of a global crisis.

This study aims to investigate the financial market during the global crisis of COVID-19, comparing and analysing market dynamics in the pre- and post-pandemic periods. By examining the different sectors embedded in the correlation matrices, we seek to identify shifts in market structure, correlation patterns, and systemic risks. Our analysis spans from pre-COVID (1st Jan 2017 to 9th Jan 2020) to post-COVID (10th Jan 2020 to 31st Dec 2022), allowing for a comprehensive examination of market evolution through the crisis.

To navigate the complexity of financial market data, we employ Random Matrix Theory (RMT), drawing inspiration from the seminal work of Plerou et al. (2002) [1]. RMT offers a powerful tool for uncovering hidden structures in complex systems, providing advantages over traditional correlation analysis methods. This approach allows us to distinguish between genuine market signals and random noise, potentially revealing subtle changes in market dynamics that might otherwise remain hidden. The pandemic is a suitable period to apply the theory used by Pleoru et al. as unusual phenomena are expected in the market. By looking at the correlations in the datasets through the lenses of Random Matrix Theory, we expect to locate indexes in the market that react similarly to the pandemic by forming sectors of highly correlated stocks.

The universality of RMT is a fundamental aspect of its applicability to financial markets. First developed in nuclear physics, RMT has shown remarkable universality across various complex systems, including financial markets [1], telecommunications [2], and biological networks [3]. This universality stems from the fact that many large, complex systems exhibit similar statistical properties in their correlation structures, regardless of the specific details of the system. In the context of financial markets, this universality allows us to apply RMT techniques to different market conditions, periods, and even across different national markets, providing a consistent framework for analysis.

This study is structured in two stages. The first stage reviews the methods used by Plerou et al. (2002) [1], discussing the underlying theory and analytical tools to prepare the reader for the second stage. The second stage applies these techniques to two financial datasets: pre-COVID and post-COVID stock market returns. This comparative analysis aims to reveal how the pandemic has altered market dynamics and interdependencies.

II. Motivation

The motivation for this study stems from the need to better understand financial market dynamics during periods of significant stress, such as the COVID-19 pandemic, and the potential of Random Matrix Theory (RMT) to provide insights into these complex systems. Our approach is exploratory, aiming to uncover patterns and relationships rather than test specific hypotheses or make predictions. RMT offers unique advantages in analysing complex financial systems, particularly its ability to distinguish between genuine correlations and random noise, as demonstrated by Plerou et al. (2001) [1]. This feature is especially valuable when studying market behaviour during unprecedented events. The universality of RMT, which has been successfully applied across various fields, provides a robust framework for comparing market behaviours across different time periods and even different markets, potentially extending our findings to other economic systems. By applying RMT to the COVID-19 crisis, we aim to uncover potential patterns or structures that may not be apparent through traditional analysis methods, allowing us to remain open to unexpected findings and potentially identify new areas for future research.

Financially, exploring markets during crises is crucial for several reasons:

[1]1172562, [2]1652398

1) **Pattern Recognition:** By examining market behaviour before, during, and after the crisis, we may identify patterns that could serve as early warning signals for future market stress. This exploratory approach could reveal subtle shifts in market dynamics that precede larger, more obvious changes.
2) **Investments:** Understanding how correlations between different market sectors change during crises could aid in developing more resilient portfolio diversification strategies. This could help in mitigating losses during market downturns and identifying opportunities for recovery.
3) **Market Movers:** Large institutional players can use the insights from how market dynamics evolve to navigate turbulent market conditions better, potentially reducing the likelihood of actions that could exacerbate market instability.
4) **Policy Implications:** While not directly addressing policy issues, our exploratory analysis may uncover market behaviours that could inform future policy discussions. This could be particularly relevant for policymakers and regulators in developing strategies to enhance market resilience.
5) **Market Resilience:** By exploring market reactions to the COVID-19 shock, we hope to contribute to the broader understanding of how financial systems respond to and recover from significant disruptions. This knowledge could be valuable in efforts to make markets more robust to global crises.

III. First Stage: Methods analysis

In this study's first stage, we analyse the methods presented in Plerou et al. (2002) [1] to better understand how random matrix theory (RMT) can be better used to investigate correlation patterns in complex, high-dimensional stock return data. relate first stage back to financial dataset and also talk about findings of the study.

A. Problem Description and Motivation

The fundamental problem addressed in the study is the modelling of market dynamics through the quantification of correlations in financial datasets. The primary motivation is to perform inference on the data and gain a deeper understanding of the underlying data-generating process. Recognizing patterns in the stock market is crucial as it enables better predictions and facilitates the development of effective portfolios. However due to the complexity, numerous hidden variables and the non-linear nature of the underlying system it is a challenging task to model the data generating system. The aim of the study in Plerou et al. (2002) is to apply the high-dimensional data analysis techniques in particular RMT to find the underlying structure in the data, as it was previously performed to understand the energy levels of complex nuclei.

B. The datasets

In the paper by Plerou et al. (2002), two datasets are involved. Both of them cover the three major US stock exchanges: the New York Stock Exchange (NYSE), the American Stock Exchange (AMEX), and the National Association of Securities Dealers Autometed Quotation (NASDAQ). The first dataset is the Trades and Quotes (TAQ) database, that documents all transactions for all major securities listed in all the three stock exchanges. The times series (starting in January 3, 1994 with length of 2 years) of prices of the 1000 largest stocks by market capitalization were analyzed. A total of $L = 6448$ records are in the time series analysis. Each record is a 30-min return of $N = 1000$ US stocks for the 2-yr period 1994-1995. From the same dataset, $L = 6448$ records of $N = 881$ stocks from the period 1996-1997 are also used. The second dataset is the Center for Research in Security Prices (CRSP) database. The CRSP stock files cover common stocks listed on NYSE beginning in 1925, the AMEX beginning in 1962, and the NASDAQ beginning in 1972. The files provide complete historical descriptive information and market data including comprehensive distribution information, high, low, and closing prices, trading volumes, shares outstanding, and total returns. We analyze daily returns for the stocks that survive for the 35-yr period 1962–1996 and extract $L = 8685$ records of 1-day returns for $N = 422$ stocks.

C. Rationale for High-Dimensional Data Analysis

The covariance matrix Σ_M of a multidimensional data matrix $\mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$ can be estimated by $\hat{\Sigma}_M$ via:

$$\hat{\Sigma}_M = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top \xrightarrow{a.s.} \Sigma_M \quad \text{as} \quad N \to \infty,$$

where N is the number of samples and M is the number of features. This convergence assumes that $N \gg M$, meaning that the number of samples is much larger than the number of features.

However, when N and M are comparable in size, the estimator $\hat{\Sigma}_M$ produces unreliable estimates. Estimating correlations is vital for data analysis, as it is a basic measure to predict patterns and to make subsequent analysis. This scenario where N

and M are comparable in size motivated the birth of Random Matrix Theory and is the playground of the study by Plerou et al. (2002).

The Law of Large Numbers is a framework to exploit the characteristic of Universality. Universality is the property of a large scale system (in this case large data quantities) to show patterns regardless of the behaviour of its constituents. In the LLN, universality shows into the Gaussian behaviour of the mean despite the very different distributions that X can assume. However, the stock market data analysed by Plerou et al. (2002) does not show the large scale patterns expected by classical statistics.

- **Dimensionality**: The study examines daily returns of 1000 stocks over a 2-year period, resulting in a comparable number of features (1000 stocks) and observations (approximately 500 trading days). In other words, both $N \to \infty$ and $M \to \infty$
- **Non-stationarity**: Financial markets are dynamic systems where the data-generating process varies with time due to changing economic conditions, market sentiments, and external events. This non-stationarity violates the second CLT requirement of identically distributed variables.
- **Fat-tailed distributions**: Stock returns often exhibit leptokurtic (fat-tailed) distributions, which may have infinite variance, violating the third CLT requirement [1].
- **Complex correlations**: The study aims to uncover genuine correlations among stocks, which are obscured by statistical noise in high-dimensional settings. Standard correlation measures become unreliable as dimensionality increases.

Even though the CLT cannot be used, universality shows also under a different framework: Random Matrix Theory. The underlying idea is to study the sample covariance matrix (SCM) under the condition:

$$M/N = \eta \quad \text{where} \quad 0 < \eta < 1$$

That is the ratio between variables and observations is fixed. What universal behaviour do we see when both of them grow large? Consider the eigenvalues of the sample covariance matrix. If the data matrix is composed only of i.i.d. data $X_i \sim \mathcal{N}(0, I_M)$ - which can be interpreted as noise - then the sample covariance matrix eigenvalues' frequencies will follow a known distribution, the Marčenko-Pastur Law. Hence, if we do not detect this distribution in the SCM, we can hypothesise the presence of information in the data.

RMT is employed by Plerou et al. (2002) to:

- provide a theoretical framework for understanding the eigenvalue distribution of correlation matrices in high-dimensional settings.
- separate of genuine correlations from noise-induced spurious correlations.
- avoid assumptions about the underlying distribution of returns (approach is non-parametric).
- handle the non-stationarity of financial data by focusing on equal-time correlations.

By using these advanced methods, the study aims to extract meaningful information about stock market structure and dynamics that standard statistical approaches would misinterpret.

D. Mathematical and Data Analysis Approaches

The analysis primarily employs correlation matrices, singular value decomposition (SVD), and random matrix theory (RMT).

1) Correlation Matrix: The correlation matrix serves as the starting point of the analysis. It is a symmetric matrix with 1s along the diagonal and values between -1 and 1 off the diagonal, quantifying linear relationships between features. The correlation matrix C is obtained by:

$$C = \frac{1}{m} A A^T, A_{i,j} \sim \mathcal{N}(0,1) \tag{1}$$

where A is the standardized dataset. The challenge lies in distinguishing meaningful correlations from noise within C.

The study examines the 30-minute returns from the TAQ database for the 2-yrs periods 1994-1995 and 1996-1997. The two correlation matrices (C_{94-95} and C_{96-97}) that are calculated show some information without much further analysis. The majority of their off-diagonal entries and their average are positive; with positive correlation being more present than negative correlations. That statement means that, on average, the datasets' values tend to increase or decrease altogether. We will see that this behaviour is exemplified by the first eigenvector (i.e. corresponding to the largest eigenvalue) of the matrix. Secondly, the values of C_{96-97} are generally higher than those of C_{94-95}, which could imply that the distribution of C depends on time.

THE PAPER HIGH CORRELATION PERIODS ARE ASSOCIATED TO HIGH VOLATILITY. WE CAN CHECK HOW THE PORTFOLIO ALLOCATION CHANGES IN THESE PERIODS

WE CAN ALSO SEE IF WITH HIGH VOLATILITY THE NUMBER OF SIGNIFICANT EIGENVALUES INCREASES

2) The expected distribution of noise: To filter out noise, C is decomposed using the eigendecomposition: $C = U\Sigma U^T$, where U is the matrix of eigenvectors and Σ is the matrix of eigenvalues. The eigenvectors encoded are independent from each other and along with the eigenvalues they encode combinations of variables that show significant variance. RMT provides a theoretical framework for understanding the eigenvalue distribution of a random cross-correlation matrix of mutually uncorrelated variables with mean 0 and unit variance. When both M and N approach infinity and their ratio $\eta = \frac{M}{N} > 1$, the probability density function of the eigenvalues λ is given by:

$$P_m(\eta) = \frac{\eta}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{2}$$

for λ within the bounds $\lambda_- \leq \lambda_i \leq \lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalues of the matrix respectively.

WE COULD CHECK WHAT HAPPENS TO THE DISTRIBUTION WHEN WE CHANGE THE FOLLOWING: THE VARIANCE, THE MEAN, INTRODUCE CORRELATION.

3) Eigenvector Analysis: Under RMT, the distribution of eigenvector entries should conform to a Normal distribution:

$$\rho_m(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{3}$$

The eigenvectors that are associated with the eigenvalues in the bulk of their distribution are indeed distributed as $\rho_m(u)$. The statistic used to measure this similarity is the kurtosis. All the eigenvalues in the bulk had a kurtosis close to 3; the eigenvectors associated to the 20 highest eigenvalues had a kurtosis significantly different from 3. In particular, the first eigenvector has a uniform distribution.

Since the eigenvectors have unit length, they can be interpreted as portfolio allocations: each entry is the percentage of budget to allocate to a given index. Negative entries imply shorting an asset, positive buying it. In the study, the return of a portfolio is given by:

$$G^k(t) = \sum_{j=1}^{1000} u_j^k G_j(t)$$

IT WOULD BE INTERESTING TO STUDY THE MATRIX OF DIFFERENCES WITH TIME.

4) Inverse Participation Ratio: The Inverse Participation Ratio (IPR) quantifies the reciprocal of the number of components that contribute significantly to the covariance matrix. For eigenvector k, it is defined as:

$$I^k = \sum_{i=1}^{n} [u_m^k]^4 \tag{4}$$

where $u_m^k, m = 1, ..., 1000$ are the components of eigenvector u_m^k. A vector with identical components $u_m^k = \frac{1}{\sqrt{N}}$ has a $I^k = \frac{1}{N}$, whereas a vector with one component $u_m^k = 1$ and the remainder 0 has $I^k = 1$. The IPR is the main tool to interpret the eigenvectors, however the first eigenvector u_k^{1000} dims the effect of the subsequent 10 eigenvectors. To extrapolate the effect of u_k^{1000} from the matrix, we recreate the correlation matrix C by expressing the data $G_i(t) = \alpha_i + \beta_i M(t) + \epsilon_i(t)$ where the component $M(t)$ is u_k^{1000} and the error components $\epsilon_i(t)$ account for all the information not expressed by u_k^{1000}, hence also the subsequent 10 eigenvectors. Taking a look at the 10 eigenvectors that originate from the so computed C matrix, we notice that their main entries belong to specific economic sectors.

DO WE SEE ANYTHING INTERESTING BY ORDERING THE SIGNIFICANT EIGENVECTORS BY THEIR IPR INSTEAD OF THEIR LAMBDAS?

The eigenvectors

5) Overlap Matrix: To assess the stability of correlations over time, an overlap matrix is constructed:

$$O_{i,j}(t, \tau) = \sum_{i=1}^{n} D_{ik}(t) D_{jk}(t + \tau) \tag{5}$$

where D is a $p \times n$ matrix constructed from the p largest eigenvalues that deviate from the upper bound λ_+.

E. Interpretation of Results

1) Largest Eigenvector: The largest eigenvector, with an IPR close to $\frac{1}{n}$, represents forces acting simultaneously across the entire stock market.

2) Significant Participants: The IPR quantifies the localization of correlations expressed by each eigenvector. A value close to 1 indicates high localization, while a lower value suggests low localization.

3) First 10 Eigenvectors: After removing the effect of u^{1000}, eigenvectors from u^{999} to u^{990} show contributing components belonging to specific sectors or industries.

4) Small Eigenvalues and Eigenvectors: Eigenvectors corresponding to the smallest eigenvalues reveal contributing entries of highly correlated indexes in C.

5) Stability Analysis: Eigenvectors with the largest eigenvalues ($u^{1000}, u^{999}, u^{998}$) maintain stable correlations over time. As the time frame widens, correlations between different eigenvectors deviate from 0, and their auto-correlation decreases.

These results provide valuable insights into the structure and dynamics of the stock market, enabling more informed decision-making in portfolio management and market analysis.

IV. Second stage

A. Dataset introduction and motivation

This study utilizes a comprehensive dataset of daily stock returns to investigate the impact of the COVID-19 pandemic on the U.S. financial markets. The dataset comprises daily log-returns for 98 stocks, strategically divided into two periods: pre-COVID (2017-01-01 to 2020-01-09) and post-COVID (2020-01-10 to 2022-12-31). This temporal segmentation allows for a rigorous comparative analysis of stock market behaviour before and after the onset of the global health crisis.

The pre-COVID period, spanning 759 days, serves as a baseline for normal market conditions. In contrast, the post-COVID period, covering 750 days captures the market's response to the pandemic and its ongoing effects. This balanced timeframe enables us to conduct robust statistical analyses and draw meaningful conclusions about the pandemic's impact on various sectors and individual stocks.

By examining the daily log-returns, this work aims to:

1) Quantify the differential impact of COVID-19 across various industries
2) Identify patterns of market resilience or vulnerability during crisis periods
3) Evaluate the effectiveness of policy responses in stabilizing financial markets
4) Contribute to the broader understanding of how external shocks affect stock market dynamics

This dataset provides a unique opportunity to empirically assess the financial ramifications of a global health crisis, offering valuable insights for policymakers, investors, and academics studying market behaviour under extreme conditions.

V. Analysis

This section presents a comparative analysis of pre-COVID and post-COVID datasets, focusing on the differential impact across industries and broader implications for stock market dynamics. To analyse and contrast the underlying structure in these datasets, we utilise the methodology introduced in Section III. Our analysis pipeline consists of four main steps:

1) Describing and examining the correlation matrix
2) Filtering and reconstructing the correlation matrix
3) Identifying sectors of interest and interpreting the results
4) Repeating Steps 1-3 for different correlation matrix estimation techniques

This structured approach allows us to thoroughly investigate changes in market behaviour and interdependencies before and after the onset of the COVID-19 pandemic. By following these steps, we can uncover subtle shifts in market dynamics, identify emerging patterns of resilience, and evaluate the effectiveness of various policy responses implemented during this period.

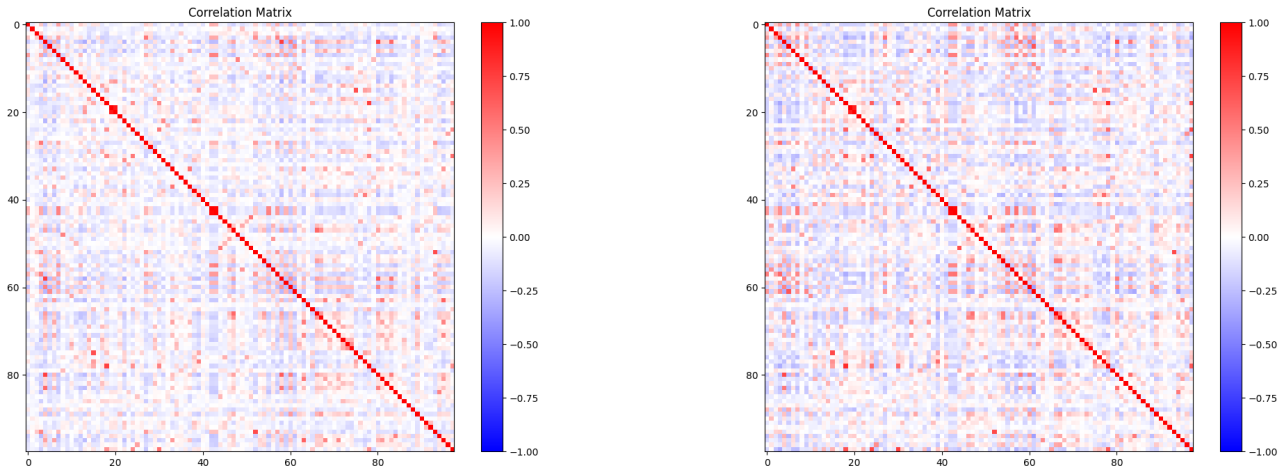
A. Raw Correlation Matrix Analysis

The analysis of raw correlation matrices for both pre-COVID and post-COVID periods reveals some initial insights into market dynamics and the pandemic's impact. Figure 1 presents the sample correlation matrices for both periods.

Both matrices exhibit considerable noise, with similar ratios of variables to samples: $\eta = \frac{98}{759} = 0.129$ for the pre-COVID period and $\eta = \frac{98}{750} = 0.13$ for the post-COVID period. While some underlying structure may exist, it is not immediately evident visually in both; filtering and reordering of the matrix is required to make these structures appear.

The presence of noise in correlation matrix estimation is not unusual in financial markets. As noted by Plerou et al. (2002) [1], random matrix theory (RMT) can be applied to filter this noise and extract meaningful information. The ratios we observe are within the range where RMT is applicable, suggesting that we can potentially identify genuine correlations amid the noise in both periods. The structure of these correlation matrices has significant implications for portfolio management and risk assessment. As Markowitz (1952) [4] demonstrated, understanding the correlations between assets is crucial for efficient portfolio construction. The noise in our matrices suggests that naive diversification strategies based on raw correlations might be suboptimal, potentially leading to underestimation of portfolio risk.

The lack of clear visual structure in the pre-COVID correlation matrix might indicate a relatively stable market period, where sector-specific or macroeconomic factors do not dominate the correlation structure. This aligns with the findings of Campbell et al. (2001) [5], who observed that in periods of market calm, correlations between stocks tend to be lower and less structured. However, comparing pre- and post-COVID matrices provides initial insights into the pandemic's impact on market correlations. A clear structure is present near the diagonal in both matrices, suggesting some commonality in



(a) Pre-COVID Correlation Matrix

(b) Post-COVID Correlation Matrix

Fig. 1: Sample Correlation Matrix for Pre-COVID and Post-COVID Periods

market dynamics both pre and post-COVID pandemic. We also observe that the post-COVID correlation matrix heatmap is a lot more active than the pre-COVID correlation matrix which may suggest more complex underlying behaviour. Baker et al. (2020) found that the stock market's reaction to COVID-19 was unprecedented in its speed and severity compared to previous pandemics, suggesting a significant shift in market dynamics [6]. This shift is likely reflected in our correlation matrices, particularly in the post-COVID dataset, although the visual differences are subtle and require further analysis to fully understand the changes in market behavior and interdependencies.

B. Eigenvalue Decomposition and Filtering

1) Eigenvalue Decomposition and Initial Analysis: We performed eigenvalue decomposition on the correlation matrices of both pre-COVID and post-COVID datasets. This process is crucial for understanding the underlying structure of the data and identifying the most significant factors driving market movements.

Our null model considers a matrix that is close to the identity matrix but includes small random perturbations. Mathematically, such a matrix can be written as:

$$M = I_m + \epsilon A$$

where:

- I_m is the $m \times m$ identity matrix,
- A is a random matrix (often with independent and identically distributed entries, usually drawn from a Gaussian distribution),
- ϵ is a small parameter that controls the strength of the perturbation.

The matrix M distribution is rotationally invariant. This means that if we apply a random rotation to the matrix, its statistical properties remain unchanged. Since all the eigenvectors are unit length, all the entries are subject to the constraint: $\sum_{i=1}^{m} v_i^2 = 1$, where v_i are the entries (components) of the eigenvector. If we consider each component v_i as a random variable, the previous constraint means that the entries v_i cannot be completely independent (since they must satisfy the normalization condition). However, given a long enough vector, the constraint action on each individual entry gets milder, making the entries behaviour closer to i.i.d. random variables. Why specifically the Gaussian behaviour is adopted by these r.v. can be understood by means of high-dimensional geometry. In high-dimensional space, most of the "volume" of the unit sphere is concentrated near its equator (due to the geometry of high dimensions). Hence, most of the random vectors drawn from the uniform distribution on the sphere will have their components v_i concentrated around zero, with small Gaussian fluctuations. Therefore, $v_i \sim \mathcal{N}\left(0, \frac{1}{n}\right)$.

This invariance leads to a very important consequence: when PCA is applied to M, the eigenvectors are distributed uniformly on the surface of a high-dimensional sphere (the so-called "uniform measure on the sphere") and their entries are distributed like a Gaussian.

Whenever M contains a signal strong enough to produce eigenvalues outside of the bulk of the MP distribution, the corresponding eigenvectors entries will deviate from the Gaussian distribution. Here, "strong enough" corresponds to $l > 1 + \sqrt{\beta}$, where l is the variance of the eigenvector and β is the ratio of variables to samples. This threshold is called

phase transition. The eigenvalue corresponding to l will take value $\hat{l}_\nu$, where:

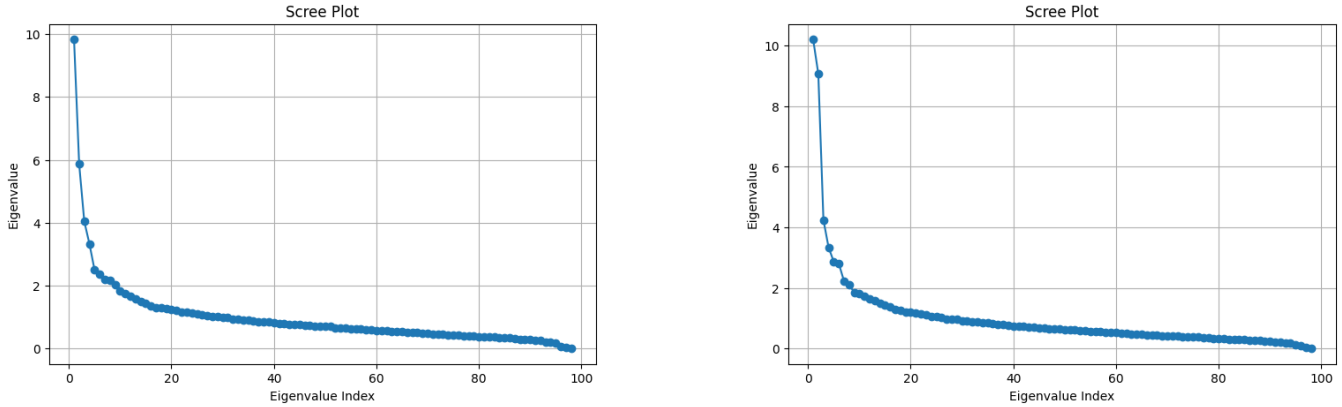
$$\hat{l}_\nu \xrightarrow{a.s.} \begin{cases} \rho_\nu > b_\beta & \text{if } l_\nu > 1 + \sqrt{\beta} \\ b_\beta & \text{if } l_\nu \leqslant 1 + \sqrt{\beta} \end{cases}$$

and

$$\rho_\nu = l_\nu + \gamma \frac{l_\nu}{l_\nu - 1}$$

a) Scree Plots: For both datasets, we visualized the eigenvalues distributions via scree plots. A scree plot plots the eigenvalues in decreasing order on a cartesian plane where each point is an eigenvalue. The x coordinate is the position of the eigenvalue in the sequence of eigenvalues; the y coordinate is the value of the eigenvalue.

The pre-COVID scree plot revealed an exponential decay pattern, characteristic of financial market data as noted by Laloux et al. (1999) [7]. Even if the post-COVID scree plot is almost identical to the pre-COVID, the post-COVID scree plot has more eigenvalues greater than 2, suggesting more significant principal components which may also imply more complex dynamics than the pre-COVID period.



(a) Pre-COVID Scree Plot

(b) Post-COVID Scree Plot

Fig. 2: Scree Plots for Pre-COVID and Post-COVID Periods

b) Comparison to Gaussian model and Null Model (MP law): To distinguish between eigenvalues representing genuine correlations and those potentially due to noise, we compared both eigenvalue distributions to the Marchenko-Pastur (MP) law (Section III). The MP law provides a theoretical distribution for eigenvalues of random correlation matrices, describing how we would expect eigenvalues to be distributed if they were to come from M realizations of N random variables, all with mean 0 and variance σ^2. He suggested to investigate the poor match of the curves and the nature of the eigenvalues on the left of the bulk

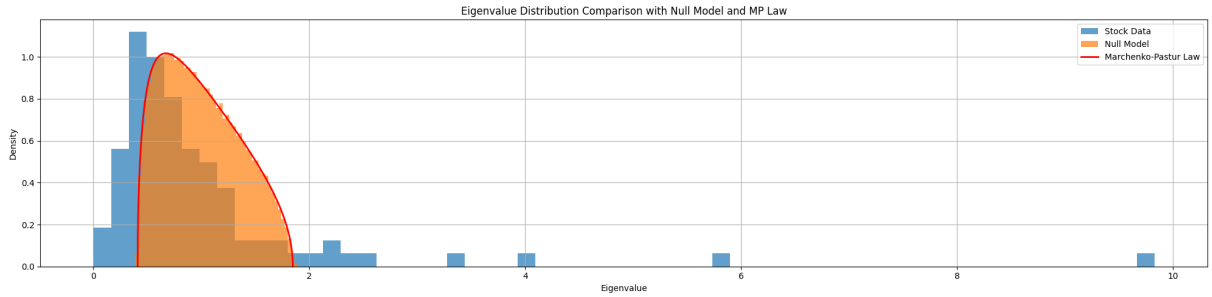
To generate a null model, we adopted the approach of Plerou et al. (2001) [1], performing 500 realizations where each stock was randomly shuffled over time to eliminate temporal dependencies. Figure 3 illustrates that the resulting null models for both pre- and post-COVID datasets closely adhere to the MP law. This alignment enables us to establish effective thresholds for eigenvalues, as detailed in Section III, facilitating the distinction between meaningful correlations and noise in our analysis.

Additionally, we compared the eigenvalue distributions against those of a covariance matrix derived from Gaussian data of similar length. This Gaussian model comparison serves as another tool to identify potentially "non-noisy" eigenvalues, complementing the insights gained from the MP law comparison.

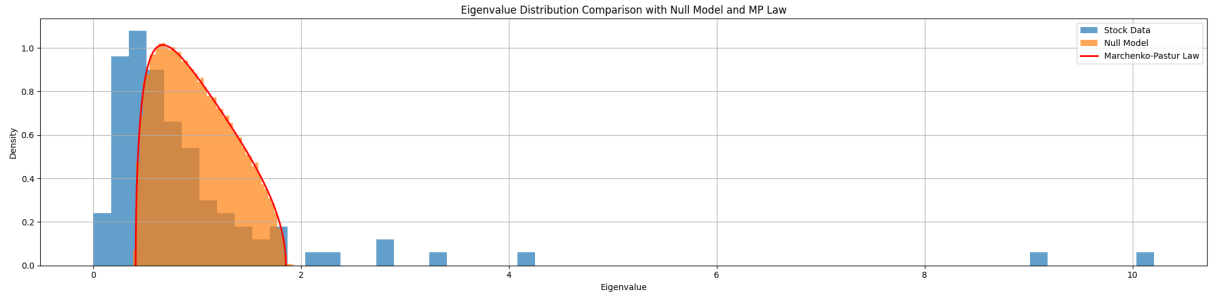
It is noteworthy to observe that the bulk of the eigenvalue distribution of the actual data does not coincide exactly with the theoretical bulk of the null and Gaussian model. Therefore, the principal components near the upper support of the null model need to be considered carefully and would require much more careful filtering. These eigenvalues could be significant in understanding micro-patterns in the pre and post-COVID economy.

2) Eigenvalue Selection: The number of significant eigenvalues was determined by comparing the eigenvalues of the correlation matrix to a threshold derived from a null model. Specifically, the threshold was set as the first quantile of the null model's eigenvalue distribution. Eigenvalues greater than this threshold were considered significant, as they represent deviations from random noise, suggesting meaningful structure in the data. The count of these eigenvalues indicated the number of significant components to retain. This was then used to select the corresponding top eigenvectors for further analysis, focusing on the key factors driving the data's variability.

For the pre-COVID dataset, we retained the first 11 eigenvalues. For the post-COVID dataset, we retained [number] eigenvalues.

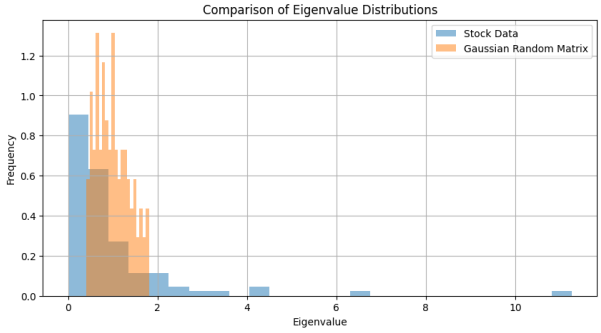


(a) Pre-COVID Eigenvalue Distribution

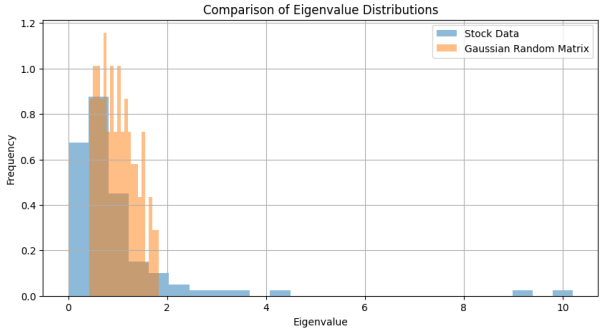


(b) Post-COVID Eigenvalue Distribution

Fig. 3: Comparison of Pre-COVID and Post-COVID Eigenvalue Distributions to the null model



(a) Pre-COVID comparison with gaussian model



(b) Post-COVID comparison with gaussian model

Fig. 4: Comparison of Pre-COVID and Post-COVID Eigenvalue Distributions to the gaussian model

If the post-COVID dataset retains more eigenvalues than the pre-COVID dataset, this suggests that the financial system has become more complex or that the data exhibits more significant sources of variance. Some potential implications include:

- Increased Market Volatility: Post-COVID financial markets have been characterized by greater volatility due to uncertainty surrounding economic recovery, policy changes, supply chain disruptions, and global shifts. More eigenvalues may indicate that more factors (i.e., eigenvectors) are needed to explain the variability in the financial system.
- Interpretation: The financial system may have become less predictable, and multiple new sources of risk or variability have emerged post-COVID.
- Emergence of New Financial Factors: The pandemic caused significant shifts in economic activity, leading to the emergence of new factors that influence financial markets. For instance, industries such as technology or healthcare may have gained prominence, while travel and hospitality might have experienced significant downturns. This structural shift can be reflected in the retained eigenvalues, indicating that new dimensions of variability (new factors) are now important in explaining the data.
- Interpretation: New factors or sectors might have risen in prominence, reflecting changing economic conditions or investor behavior in the post-COVID landscape.
- Increased Noise or Uncertainty: The post-COVID dataset may also retain more eigenvalues because of increased noise in the data. The pandemic introduced numerous sources of uncertainty, such as government interventions, changing consumer behavior, and unpredictable global events. These factors might contribute to a larger number of eigenvalues

being retained, even if some are capturing noise rather than true signal.

- Interpretation: Some of the additional retained eigenvalues may reflect an increase in noise or short-term fluctuations rather than long-term structural changes.

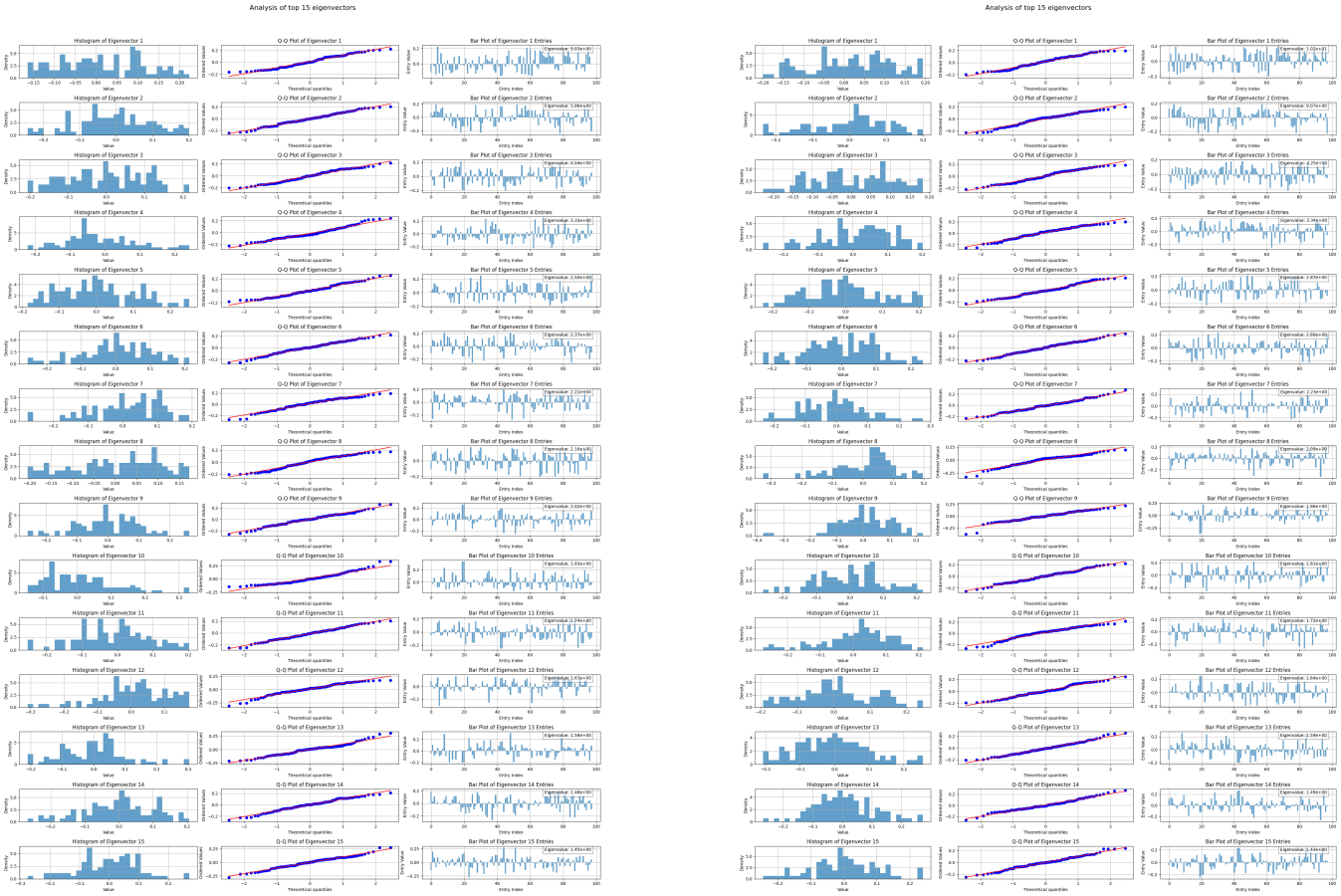
3) Eigenvector Analysis: We analyzed the eigenvectors corresponding to the largest eigenvalues for both datasets using histograms, Q-Q plots, and bar plots. These visualizations help identify non-Gaussian behavior and highlight specific stocks or sectors that contribute significantly to each eigenvector.

If the frequencies of the entries of the eigenvectors are distributed according to a Gaussian distribution, then most likely the eigenvectors are not carrying information/signals. This statement will be confirmed by a meaningful interpretation of the entries.

a) Precovid: Some eigenvectors (numbers 5, 6, 9, 13, 14, and 15) are distributed similarly to a Gaussian.

b) Postcovid: The eigenvectors from 5 to 15 are distributed similarly to a Gaussian. [Discuss gausian distribution of the entries in post covid and not pre covid, and their different significant number of eigenvalues]

[Describe any notable differences in eigenvector characteristics between pre-COVID and post-COVID datasets]



(a) Pre-COVID Eigenvector Analysis

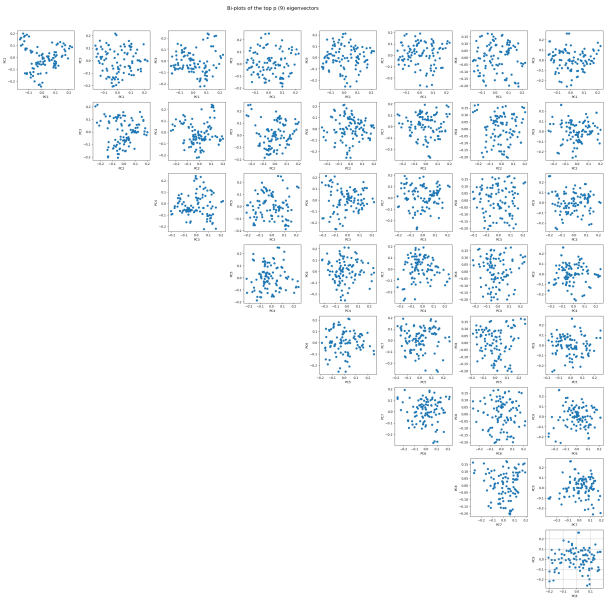
(b) Post-COVID Eigenvector Analysis

Fig. 5: Eigenvector Analysis for Pre-COVID and Post-COVID Periods

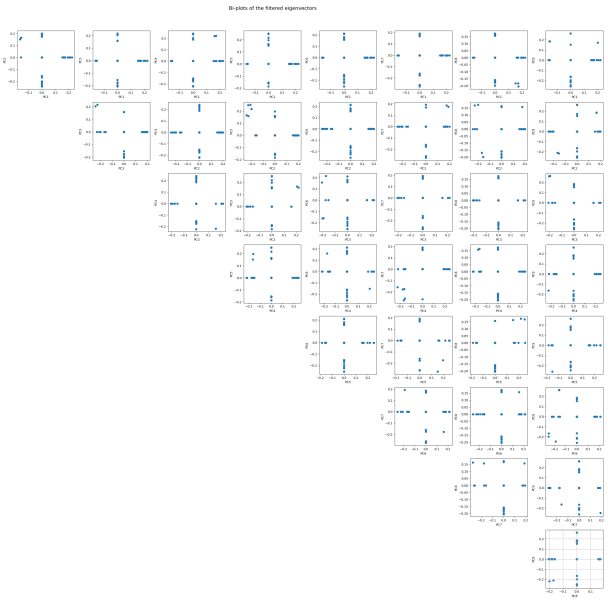
4) Eigenvector Trimming: The eigenvectors were trimmed using a threshold-based filtering approach. First, the most significant eigenvectors were selected based on the corresponding eigenvalues exceeding a cutoff derived from the null model. To remove noise, a confidence threshold was calculated by multiplying the standard deviation of a mid-range eigenvector by a scalar confidence factor (1.8). Any elements of the eigenvectors whose absolute values were below this threshold were set to zero, effectively reducing small, noisy contributions. This process ensured that only the most meaningful components of the eigenvectors were retained for further analysis, such as bi-plots and spectral clustering. This approach is similar to that suggested by Bun et al. (2017) [8].

5) Bi-plots Before and After Trimming: We created bi-plots of the eigenvectors both before and after trimming for each dataset. These plots help visualize the relationships between different stocks or sectors in the two-dimensional space defined by pairs of eigenvectors.

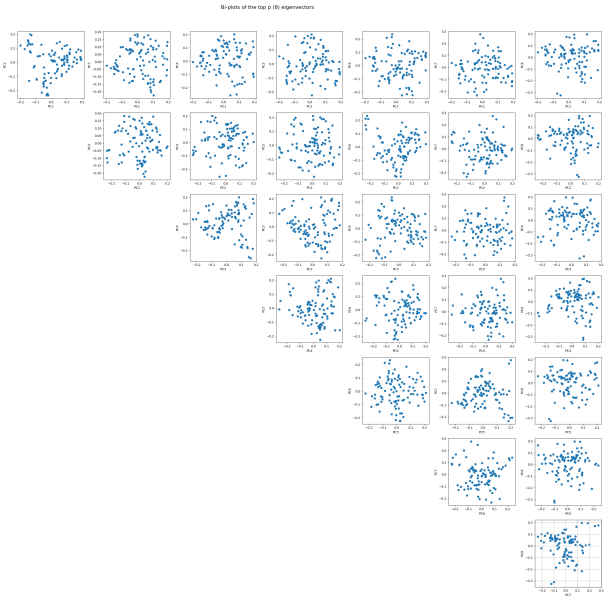
[Discuss any significant differences in bi-plots between pre-COVID and post-COVID datasets, and changes due to trimming]



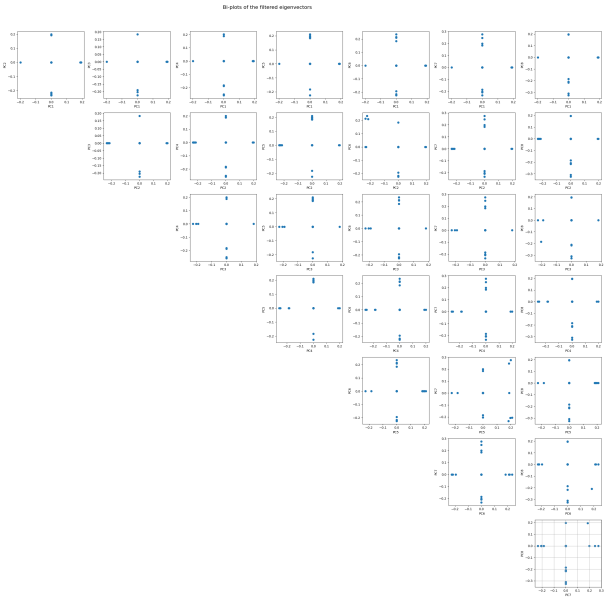
(a) Pre-COVID Bi-plot (Unfiltered)



(b) Pre-COVID Bi-plot (Filtered)



(c) Post-COVID Bi-plot (Unfiltered)



(d) Post-COVID Bi-plot (Filtered)

Fig. 6: Comparison of Pre-COVID and Post-COVID Bi-plots Before and After Trimming

6) Implications of Eigenvalue Analysis: The eigenvalue analysis reveals important aspects of market structure in both pre-COVID and post-COVID periods. As Bouchaud and Potters (2003) [9] explain, eigenvalues exceeding the MP upper bound likely correspond to genuine correlations rather than noise. The factors represented by these significant eigenvalues may represent key drivers of market movements, such as broad economic trends, sector-specific influences, or global events. [Discuss any changes in the number or nature of significant factors between pre-COVID and post-COVID periods]

From a practical perspective, identifying these factors is crucial for risk management and factor investing strategies. For instance, the number and nature of these factors can inform the construction of factor models, as described by Fama and French (1993) [10]. Our findings suggest [discuss implications for factor models and risk decomposition strategies, noting any changes due to COVID-19].

The comparison between pre- and post-COVID eigenvalue structures reveals shifts in market dynamics and the potential emergence of new factors influencing stock returns. This aligns with the findings of Mazur et al. (2021), who observed significant changes in market behavior during the March 2020 stock market crash [11].

[Elaborate on specific changes observed and their potential implications for market structure and dynamics]

C. Eigenvalue Decomposition and Filtering

1) Pre-COVID Dataset: The pre-COVID scree-plot shows an exponential decay, making it challenging to determine the number of eigenvalues to retain. Using the Marchenko-Pastur (MP) law and a confidence interval of 1.5σ, we retain the first 11 eigenvalues.

The exponential decay observed in the scree-plot is characteristic of financial market data, as noted by Laloux et al. (1999) [7]. This decay pattern suggests that while there are dominant factors influencing market movements, their influence diminishes rapidly, leading to a long tail of less significant factors.

The decision to retain the first 11 eigenvalues based on the MP law has important implications for understanding market structure. As Bouchaud and Potters (2003) [9] explain, eigenvalues exceeding the MP upper bound likely correspond to genuine correlations rather than noise. In our case, these 11 factors may represent key drivers of market movements, such as broad economic trends, sector-specific influences, or global events.

From a practical perspective, identifying these factors is crucial for risk management and factor investing strategies. For instance, the number and nature of these factors can inform the construction of factor models, as described by Fama and French (1993) [10]. Our finding of 11 significant factors suggests a more complex market structure than captured by traditional three or five-factor models, potentially offering opportunities for more nuanced risk decomposition and alpha generation strategies.

2) Post-COVID Dataset: [Insert analysis of post-COVID eigenvalue decomposition]

Comparing the eigenvalue structures between pre- and post-COVID periods reveals shifts in market dynamics and the potential emergence of new factors influencing stock returns. This aligns with the findings of Mazur et al. (2021), who observed significant changes in market behavior during the March 2020 stock market crash [11]. They found that certain industries, such as natural gas, food, healthcare, and software stocks, earned significant positive returns, while industries like petroleum, real estate, entertainment, and hospitality experienced considerably negative returns.

D. Eigenvector Analysis

1) Pre-COVID Dataset: Prior to trimming, pre-COVID eigenvectors show limited informativeness. After trimming using a 1.5σ confidence interval, non-Gaussian behavior becomes evident in Q-Q plots, and bar plots highlight specific stocks.

The initial limited informativeness of the eigenvectors aligns with the observations of Plerou et al. (2002) [1], who noted that raw eigenvectors in financial data often contain a mix of signal and noise. The trimming procedure we apply is similar to the approach suggested by Bun et al. (2017) [8], which aims to denoise the eigenvectors and reveal underlying structure.

The emergence of non-Gaussian behavior in the trimmed eigenvectors is particularly significant. As noted by Cont (2001) [12], non-Gaussian features in financial data often indicate the presence of important market phenomena that deviate from standard assumptions of normality. In our case, this non-Gaussian behavior might represent sector-specific dynamics, the influence of large market players, or other systematic factors affecting stock returns.

The bar plots highlighting specific stocks in each eigenvector provide valuable insights into market structure. As demonstrated by Plerou et al. (2001) [1], the composition of significant eigenvectors often reveals economic sectors or other meaningful groupings of stocks. Our analysis may uncover similar patterns, potentially identifying:

1) Dominant market sectors
2) Groups of stocks that tend to move together
3) Stocks that are particularly influential in driving market movements

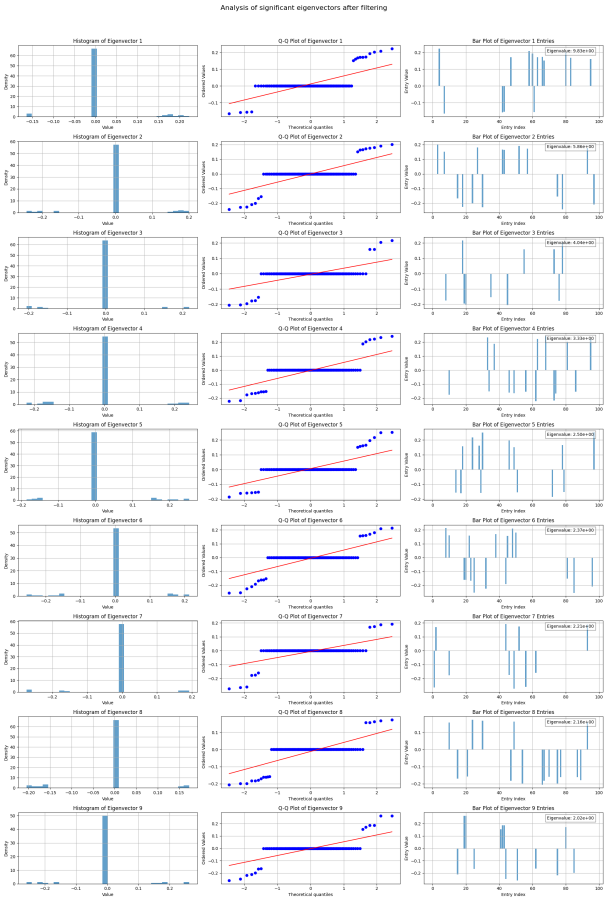
These findings have important implications for portfolio management and risk assessment. For instance:

- Sector identification can inform sector rotation strategies or sector-specific risk management approaches.
- Understanding which stocks tend to move together can help in designing diversification strategies that go beyond simple correlation-based approaches.
- Identifying influential stocks can be valuable for market monitoring and potentially for predicting broader market movements.

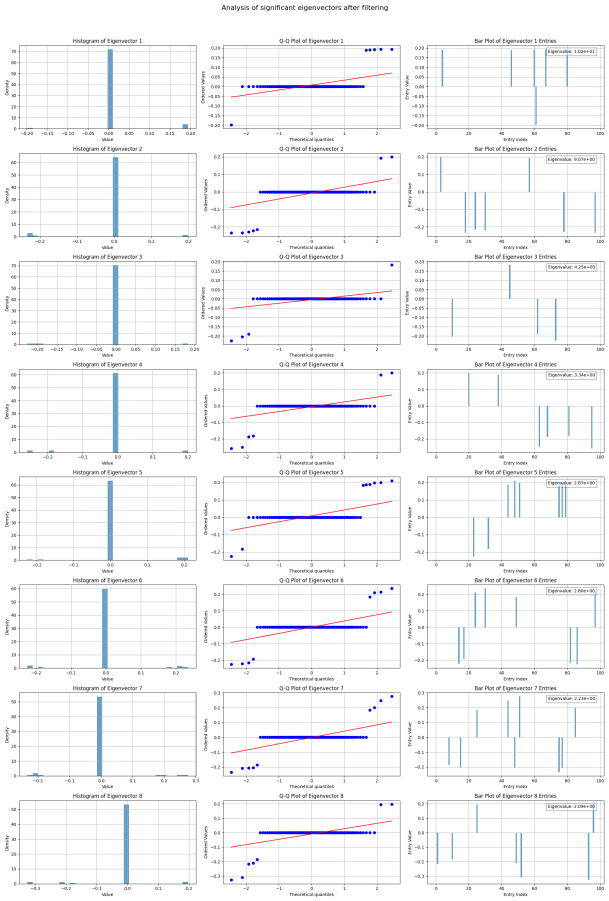
Moreover, the pre-COVID eigenvector structure serves as a baseline for comparison with the post-COVID period, potentially revealing how the pandemic has altered fundamental market dynamics and relationships between stocks.

2) Post-COVID Dataset: [Insert analysis of post-COVID eigenvectors]

The comparison of pre- and post-COVID eigenvectors provides insights into changing relationships between stocks and sectors, potentially identifying industries most impacted by the pandemic. Ding et al. (2021) found that firms with stronger pre-2020 finances, less exposure to COVID-19 through global supply chains, and more CSR activities experienced milder stock price reactions to COVID-19 market shocks [13]. This suggests that our eigenvector analysis might reveal similar patterns of resilience among certain stocks or sectors.



(a) Pre-COVID Eigenvector Analysis



(b) Post-COVID Eigenvector Analysis

Fig. 7: Eigenvector Analysis for Pre-COVID and Post-COVID Periods after filtering

E. Cluster Identification and Sector Analysis

This section focuses on quantifying the differential impact of COVID-19 across various industries and identifying patterns of market resilience or vulnerability. Baek et al. (2020) conducted an industry-level analysis of COVID-19's impact on stock market volatility [14]. They found that industries related to necessities and COVID-19 substitutes, such as food and staples retailing, household products, and telecommunications, showed greater resilience. In contrast, industries like energy equipment and services, consumer finance, and airlines were most impacted.

Our analysis aims to corroborate these findings and potentially identify additional patterns:

- Sectors most affected by the pandemic (e.g., travel, hospitality)
- Industries showing resilience (e.g., technology, healthcare)
- Emerging correlations between previously unrelated sectors

Various studies have shown differential impacts of COVID-19 across industries. For instance, Mazur et al. (2021) found that natural gas, food, healthcare, and software stocks earned significant positive returns during the March 2020 stock market crash, while petroleum, real estate, entertainment, and hospitality experienced considerably negative returns [11]. Similarly, Baek et al. (2020) observed that industries related to necessities and COVID-19 substitutes showed greater resilience, while sectors like energy equipment and services, consumer finance, and airlines were more severely impacted [14].

[Reorder using the clusters, change the stock index to stock label, reflect the findings with the ones in literature]

Discuss the patterns emerging form the reordered corr matrix

table

F. Correlation Matrix Reconstruction

[reordering using clusters and reconstructions using top p eigenvalue and vector pair]

1) Pre-COVID Reconstruction: The reconstructed pre-COVID correlation matrix reveals areas of uncorrelated stocks, which may represent sectors with independent behavior or unexplained market phenomena.

Cluster 0		
Symbol	**Company Name**	**Sector**
Energy		
BP	BP p.l.c.	Energy
COP	ConocoPhillips	Energy
CVX	Chevron Corporation	Energy
ENB	Enbridge Inc.	Energy
PBR	Petróleo Brasileiro S.A. - Petrobras	Energy
SHEL	Shell plc	Energy
XOM	Exxon Mobil Corporation	Energy
Other Sectors		
ACN	Accenture plc	Technology
ORCL	Oracle Corporation	Technology
BUD	Anheuser-Busch InBev SA/NV	Consumer Defensive
COST	Costco Wholesale Corporation	Consumer Defensive
MA	Mastercard Incorporated	Financial Services
V	Visa Inc.	Financial Services
UPS	United Parcel Service, Inc.	Industrials

Cluster 1		
Symbol	**Company Name**	**Sector**
Financial Services		
BAC	Bank of America Corporation	Financial Services
BRK-A	Berkshire Hathaway Inc.	Financial Services
BRK-B	Berkshire Hathaway Inc.	Financial Services
Industrials		
CAT	Caterpillar Inc.	Industrials
DE	Deere & Company	Industrials
ETN	Eaton Corporation plc	Industrials
GE	GE Aerospace	Industrials
HON	Honeywell International Inc.	Industrials
RTX	RTX Corporation	Industrials
UNP	Union Pacific Corporation	Industrials
Other Sectors		
CRH	CRH plc	Basic Materials
EQIX	Equinix, Inc.	Real Estate

Cluster 2		
Symbol	**Company Name**	**Sector**
Basic Materials		
APD	Air Products and Chemicals, Inc.	Basic Materials
ECL	Ecolab Inc.	Basic Materials
LIN	Linde plc	Basic Materials
SHW	The Sherwin-Williams Company	Basic Materials
Consumer Cyclical		
HD	The Home Depot, Inc.	Consumer Cyclical
LOW	Lowe's Companies, Inc.	Consumer Cyclical

Cluster 3		
Symbol	**Company Name**	**Sector**
Technology		
AAPL	Apple Inc.	Technology
ADBE	Adobe Inc.	Technology
AMD	Advanced Micro Devices, Inc.	Technology
AVGO	Broadcom Inc.	Technology
CRM	Salesforce, Inc.	Technology
MSFT	Microsoft Corporation	Technology
NVDA	NVIDIA Corporation	Technology
Consumer Cyclical		
AMZN	Amazon.com, Inc.	Consumer Cyclical
BABA	Alibaba Group Holding Limited	Consumer Cyclical
TSLA	Tesla, Inc.	Consumer Cyclical
Communication Services		
GOOG	Alphabet Inc.	Communication Services
GOOGL	Alphabet Inc.	Communication Services
META	Meta Platforms, Inc.	Communication Services
NFLX	Netflix, Inc.	Communication Services
T	AT&T Inc.	Communication Services
Other Sectors		
FMX	Fomento Económico Mexicano, S.A.B. de C.V.	Consumer Defensive
LMT	Lockheed Martin Corporation	Industrials

Cluster 4		
Symbol	**Company Name**	**Sector**
Healthcare		
ABBV	AbbVie Inc.	Healthcare
AZN	AstraZeneca PLC	Healthcare
JNJ	Johnson & Johnson	Healthcare
LLY	Eli Lilly and Company	Healthcare
MRK	Merck & Co., Inc.	Healthcare
NVO	Novo Nordisk A/S	Healthcare
RHHBF	Roche Holding AG	Healthcare
RHHBY	Roche Holding AG	Healthcare
UNH	UnitedHealth Group Incorporated	Healthcare
Other Sectors		
BA	The Boeing Company	Industrials
BKNG	Booking Holdings Inc.	Consumer Cyclical
TJX	The TJX Companies, Inc.	Consumer Cyclical
DIS	The Walt Disney Company	Communication Services
SPG	Simon Property Group, Inc.	Real Estate

Cluster 5		
Symbol	**Company Name**	**Sector**
Financial Services		
BAC-PE	Bank of America Corporation	Financial Services
BML-PG	Bank of America Corporation	Financial Services
Real Estate		
PLDGP	Prologis, Inc.	Real Estate
SPG-PJ	Simon Property Group, Inc.	Real Estate
Other Sectors		
CTA-PB	EIDP, Inc.	Basic Materials
DTEGY	Deutsche Telekom AG	Communication Services
RHHVF	Roche Holding AG	Healthcare
TM	Toyota Motor Corporation	Consumer Cyclical

Cluster 6		
Symbol	**Company Name**	**Sector**
Utilities		
AEP	American Electric Power Company, Inc.	Utilities
NEE	NextEra Energy, Inc.	Utilities
SO	The Southern Company	Utilities
SRE	Sempra	Utilities
Real Estate		
AMT	American Tower Corporation	Real Estate
DLR	Digital Realty Trust, Inc.	Real Estate
O	Realty Income Corporation	Real Estate
PLD	Prologis, Inc.	Real Estate
PSA	Public Storage	Real Estate
WELL	Welltower Inc.	Real Estate

Cluster 7		
Symbol	**Company Name**	**Sector**
Basic Materials		
BHP	BHP Group Limited	Basic Materials
FCX	Freeport-McMoRan Inc.	Basic Materials
RIO	Rio Tinto Group	Basic Materials
SCCO	Southern Copper Corporation	Basic Materials
Communication Services		
CMCSA	Comcast Corporation	Communication Services
TMUS	T-Mobile US, Inc.	Communication Services
VZ	Verizon Communications Inc.	Communication Services
Consumer Defensive		
KO	The Coca-Cola Company	Consumer Defensive
MDLZ	Mondelez International, Inc.	Consumer Defensive
PEP	PepsiCo, Inc.	Consumer Defensive
PG	The Procter & Gamble Company	Consumer Defensive
PM	Philip Morris International Inc.	Consumer Defensive
UL	Unilever PLC	Consumer Defensive
WMT	Walmart Inc.	Consumer Defensive
Other Sectors		
MCD	McDonald's Corporation	Consumer Cyclical
CSCO	Cisco Systems, Inc.	Technology
NGG	National Grid plc	Utilities

TABLE I: Clusters of Post-COVID Stock Market Data

Cluster 0		
Symbol	**Company Name**	**Sector**
Energy		
BP	BP p.l.c.	Energy
COP	ConocoPhillips	Energy
CVX	Chevron Corporation	Energy
ENB	Enbridge Inc.	Energy
PBR	Petróleo Brasileiro S.A.	Energy
SHEL	Shell plc	Energy
XOM	Exxon Mobil Corporation	Energy
Basic Materials		
BHP	BHP Group Limited	Basic Materials
RIO	Rio Tinto Group	Basic Materials

Cluster 1		
Symbol	**Company Name**	**Sector**
Technology		
ACN	Accenture plc	Technology
CSCO	Cisco Systems, Inc.	Technology
ORCL	Oracle Corporation	Technology
Financial Services		
BAC-PE	Bank of America Corporation	Financial Services
BML-PG	Bank of America Corporation	Financial Services
Healthcare		
RHHBF	Roche Holding AG	Healthcare
RHHBY	Roche Holding AG	Healthcare
RHHVF	Roche Holding AG	Healthcare
Other Sectors		
CTA-PB	EIDP, Inc.	Basic Materials
DTEGY	Deutsche Telekom AG	Communication Services
PLDGP	Prologis, Inc.	Real Estate
SPG-PJ	Simon Property Group, Inc.	Real Estate
TM	Toyota Motor Corporation	Consumer Cyclical

Cluster 2		
Symbol	**Company Name**	**Sector**
Consumer Defensive		
BUD	Anheuser-Busch InBev SA/NV	Consumer Defensive
KO	The Coca-Cola Company	Consumer Defensive
MDLZ	Mondelez International, Inc.	Consumer Defensive
PEP	PepsiCo, Inc.	Consumer Defensive
PG	The Procter & Gamble Company	Consumer Defensive
PM	Philip Morris International Inc.	Consumer Defensive
UL	Unilever PLC	Consumer Defensive
Utilities		
AEP	American Electric Power Company	Utilities
NEE	NextEra Energy, Inc.	Utilities
NGG	National Grid plc	Utilities
SO	The Southern Company	Utilities
SRE	Sempra	Utilities
Other Sectors		
CAT	Caterpillar Inc.	Industrials
FCX	Freeport-McMoRan Inc.	Basic Materials
SCCO	Southern Copper Corporation	Basic Materials
MCD	McDonald's Corporation	Consumer Cyclical

Cluster 3		
Symbol	**Company Name**	**Sector**
Consumer Cyclical		
HD	The Home Depot, Inc.	Consumer Cyclical
LOW	Lowe's Companies, Inc.	Consumer Cyclical
TJX	The TJX Companies, Inc.	Consumer Cyclical
Consumer Defensive		
COST	Costco Wholesale Corporation	Consumer Defensive

Cluster 4		
Symbol	**Company Name**	**Sector**
Basic Materials		
APD	Air Products and Chemicals, Inc.	Basic Materials
ECL	Ecolab Inc.	Basic Materials
LIN	Linde plc	Basic Materials
SHW	The Sherwin-Williams Company	Basic Materials

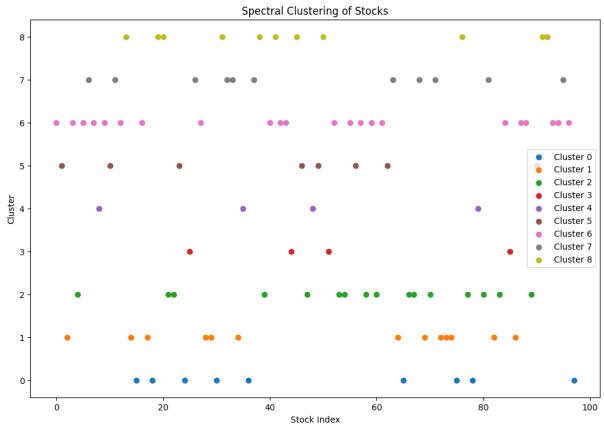
Cluster 5		
Symbol	**Company Name**	**Sector**
Healthcare		
ABBV	AbbVie Inc.	Healthcare
AZN	AstraZeneca PLC	Healthcare
JNJ	Johnson & Johnson	Healthcare
LLY	Eli Lilly and Company	Healthcare
MRK	Merck & Co., Inc.	Healthcare
NVO	Novo Nordisk A/S	Healthcare
UNH	UnitedHealth Group Incorporated	Healthcare
Communication Services		
CMCSA	Comcast Corporation	Communication Services

Cluster 6		
Symbol	**Company Name**	**Sector**
Technology		
AAPL	Apple Inc.	Technology
ADBE	Adobe Inc.	Technology
AMD	Advanced Micro Devices, Inc.	Technology
AVGO	Broadcom Inc.	Technology
CRM	Salesforce, Inc.	Technology
MSFT	Microsoft Corporation	Technology
NVDA	NVIDIA Corporation	Technology
Communication Services		
GOOG	Alphabet Inc.	Communication Services
GOOGL	Alphabet Inc.	Communication Services
META	Meta Platforms, Inc.	Communication Services
NFLX	Netflix, Inc.	Communication Services
T	AT&T Inc.	Communication Services
TMUS	T-Mobile US, Inc.	Communication Services
VZ	Verizon Communications Inc.	Communication Services
Consumer Cyclical		
AMZN	Amazon.com, Inc.	Consumer Cyclical
BABA	Alibaba Group Holding Limited	Consumer Cyclical
BKNG	Booking Holdings Inc.	Consumer Cyclical
TSLA	Tesla, Inc.	Consumer Cyclical
Other Sectors		
FMX	Fomento Económico Mexicano	Consumer Defensive
WMT	Walmart Inc.	Consumer Defensive
MA	Mastercard Incorporated	Financial Services
V	Visa Inc.	Financial Services

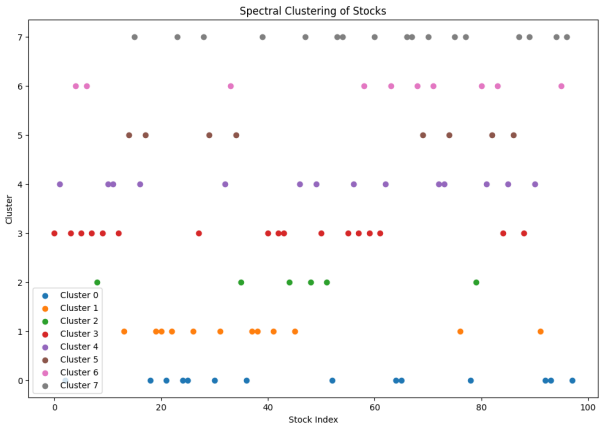
Cluster 7		
Symbol	**Company Name**	**Sector**
Real Estate		
AMT	American Tower Corporation	Real Estate
DLR	Digital Realty Trust, Inc.	Real Estate
EQIX	Equinix, Inc.	Real Estate
O	Realty Income Corporation	Real Estate
PLD	Prologis, Inc.	Real Estate
PSA	Public Storage	Real Estate
SPG	Simon Property Group, Inc.	Real Estate
WELL	Welltower Inc.	Real Estate
Other Sectors		
BA	The Boeing Company	Industrials
CRH	CRH plc	Basic Materials
DIS	The Walt Disney Company	Communication Services

Cluster 8		
Symbol	**Company Name**	**Sector**
Industrials		
DE	Deere & Company	Industrials
ETN	Eaton Corporation plc	Industrials
GE	GE Aerospace	Industrials
HON	Honeywell International Inc.	Industrials
LMT	Lockheed Martin Corporation	Industrials
RTX	RTX Corporation	Industrials
UNP	Union Pacific Corporation	Industrials
UPS	United Parcel Service, Inc.	Industrials
Financial Services		
BAC	Bank of America Corporation	Financial Services
BRK-A	Berkshire Hathaway Inc.	Financial Services
BRK-B	Berkshire Hathaway Inc.	Financial Services

TABLE II: Clusters of Pre-COVID Stock Market Data

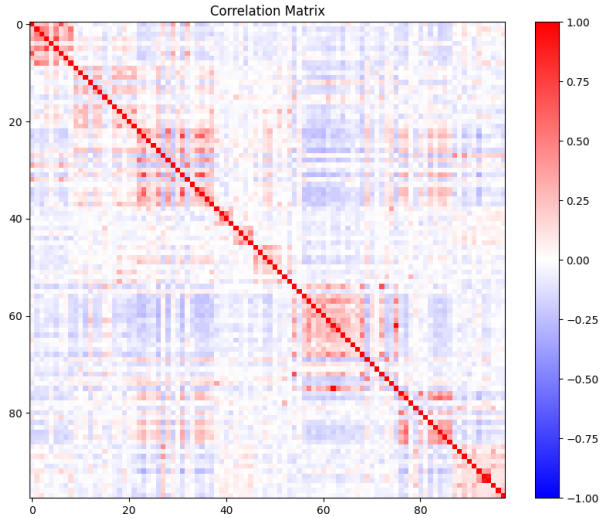


(a) Pre-COVID Cluster

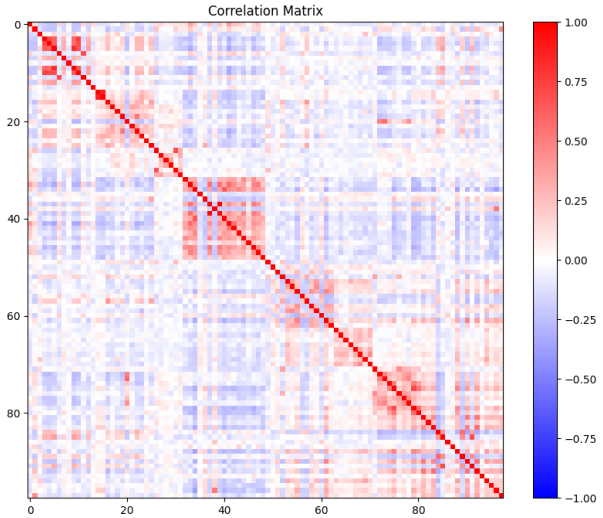


(b) Post-COVID Clusters

Fig. 8: Spectral Clustering from the top p eigenvectors



(a) Pre-COVID Reordered Correlation Matrix



(b) Post-COVID Reordered Correlation Matrix

Fig. 9: Reordered Correlation Matrix

2) Post-COVID Reconstruction: [Insert analysis of reconstructed post-COVID correlation matrix]

Comparing the reconstructed matrices allows us to evaluate the effectiveness of policy responses in stabilizing financial markets. Berger and Demirgüç-Kunt (2021) emphasize that the COVID-19 pandemic was the ”most unanticipated large and widespread exogenous economic shock of all time” [15]. They review various policy responses and their impacts on the banking sector, including government guarantees, regulatory forbearance, and central bank liquidity support. Our reconstructed matrices may reflect the impact of these policy interventions, potentially showing altered correlations in sectors that received targeted support.

G. Recovering Eigenvectors before the threshold

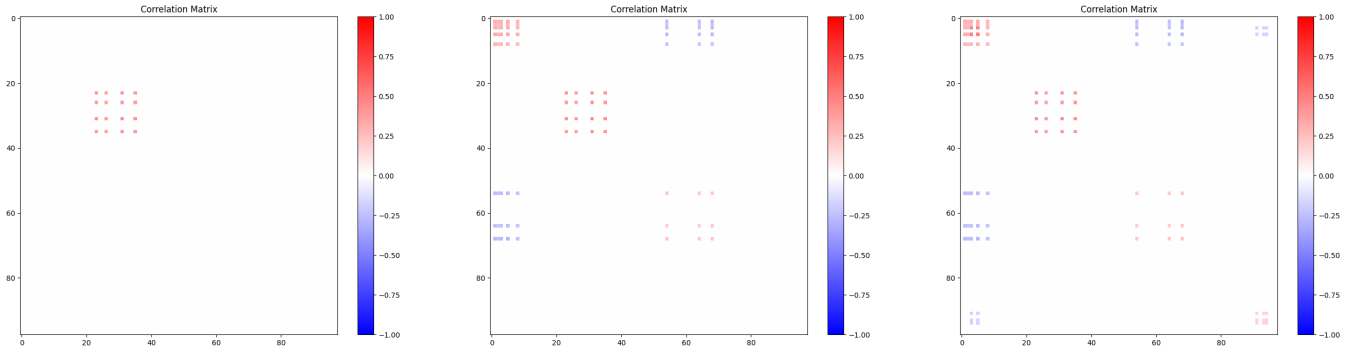
TODO: different thresholding method to recover these and discussing the information they provide

H. Robust Covariance estimation techniques

summarise the results if we use a robust technique and how that shows

I. Temporal Stability Analysis

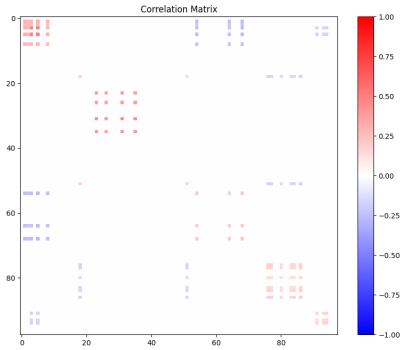
[TODO: moving window approach]



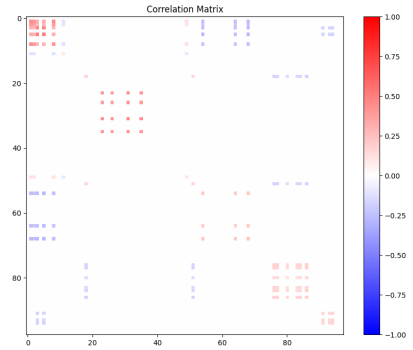
(a) 1 Eigenvalue

(b) 2 Eigenvalues

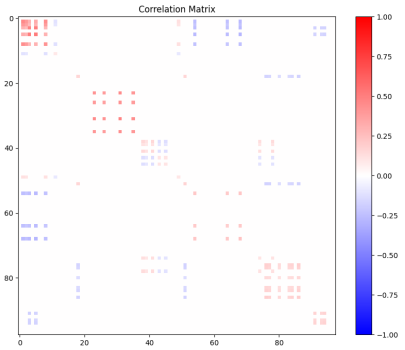
(c) 3 Eigenvalues



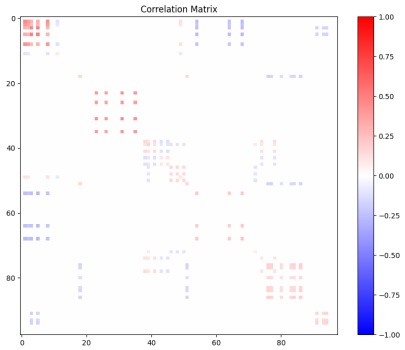
(d) 4 Eigenvalues



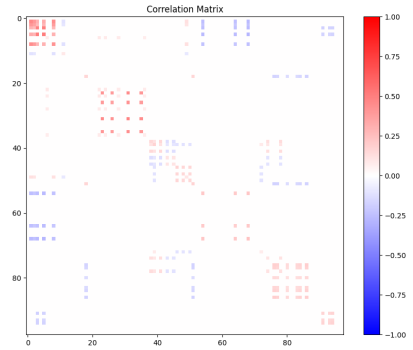
(e) 5 Eigenvalues



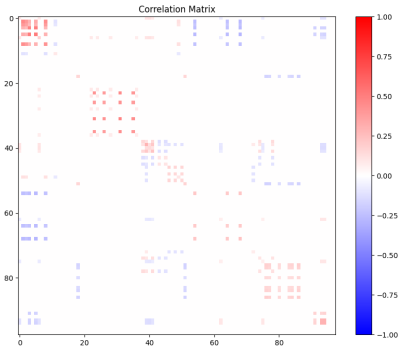
(f) 6 Eigenvalues



(g) 7 Eigenvalues



(h) 8 Eigenvalues



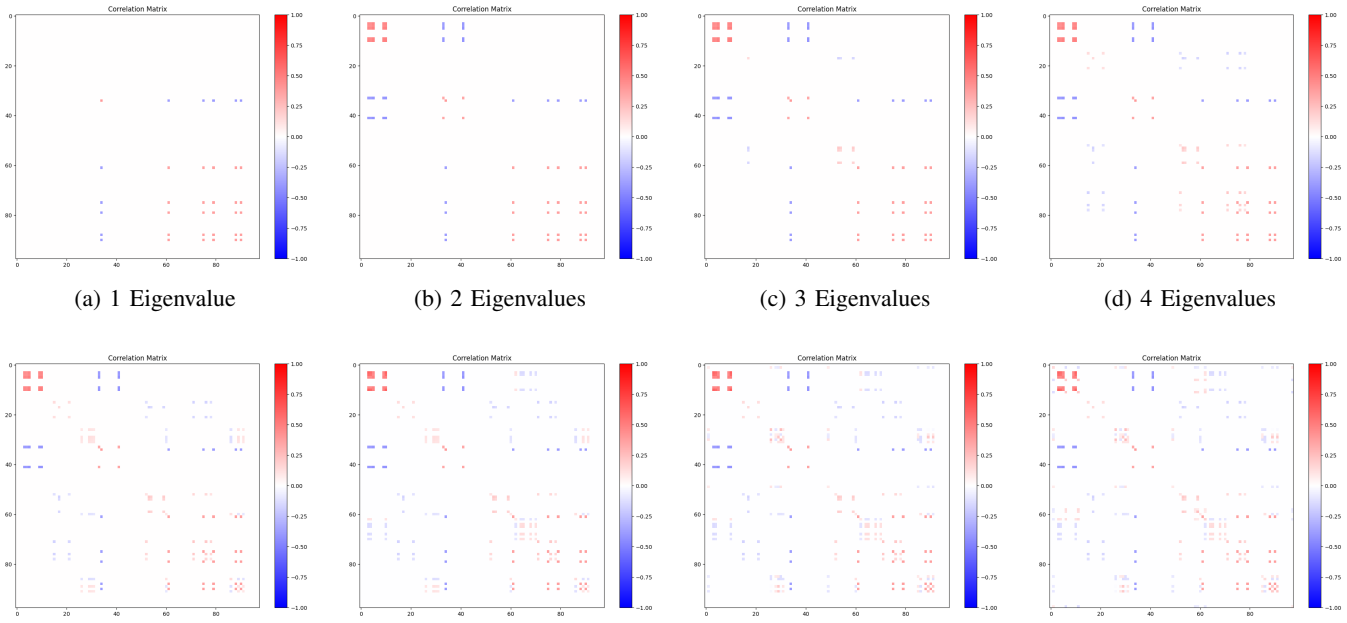
(i) 9 Eigenvalues

Fig. 10: Correlation Matrix Reconstructions using Different Numbers of Eigenvalues (Pre-COVID)

VI. Discussion

This section synthesizes the findings from previous analyses to address the broader understanding of how external shocks affect stock market dynamics:

1) **Differential Impact**: We quantify the varying effects of COVID-19 across industries, building on the work of Baek et al. (2020) [14]. Their study found that COVID-19 related uncertainty negatively impacted returns across all industries and generally led to higher volatility. Our analysis aims to provide a more granular view of these impacts across global industries.
2) **Market Resilience**: By comparing pre- and post-COVID clusters and correlation structures, we identify patterns of market resilience or vulnerability. This extends the findings of Ding et al. (2021) [13], who identified firm-specific characteristics associated with resilience to COVID-19 market shocks.
3) **Policy Effectiveness**: Changes in correlation structures and cluster compositions between the two periods provide insights into the effectiveness of policy responses. Berger and Demirgüç-Kunt (2021) [15] found that policymakers in richer and more populous countries were significantly more responsive and took more policy measures. Our analysis may reveal whether these interventions had observable effects on market structures.
4) **Market Dynamics**: We discuss how the identified changes contribute to the broader understanding of external shock



(a) 1 Eigenvalue (b) 2 Eigenvalues (c) 3 Eigenvalues (d) 4 Eigenvalues

(e) 5 Eigenvalues (f) 6 Eigenvalues (g) 7 Eigenvalues (h) 8 Eigenvalues

Fig. 11: Correlation Matrix Reconstructions using Different Numbers of Eigenvalues (Post-COVID)

impacts on stock market dynamics. This builds on the work of Goodell (2020) [16], who outlined several areas for future research on COVID-19's impact on finance, including changes in consumer behavior, supply chain reorganization, and the role of information.

[Relate it back to the data analysis in previous part]

Our analysis also considers the findings of Ashraf (2020) [17], who showed that the effects of COVID-19 on financial markets differ between developed and developing nations. We examine whether our results support the conclusion that economic criteria are the most essential transmission channel in developed nations, while social criteria play a more prominent role in developing nations.

This comprehensive analysis provides insights into the pandemic's impact on various industries, market resilience, policy effectiveness, and overall stock market dynamics, contributing to the growing body of literature on COVID-19's effects on global financial markets.

VII. Conclusion

REFERENCES

[1] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, "Random matrix approach to cross correlations in financial data," *Phys. Rev. E*, vol. 65, p. 066126, Jun 2002. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.65.066126

[2] R. Couillet and M. Deb11:46 PM 7/10/2024bah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[3] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chakraborty, "Coordinate linkage of hiv evolution reveals regions of immunological vulnerability," *Proceedings of the National Academy of Sciences*, vol. 108, no. 28, pp. 11 530–11 535, 2011. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1105315108

[4] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952. [Online]. Available: http://www.jstor.org/stable/2975974

[5] J. Y. Campbell, M. Lettau, B. G. Malkiel, and Y. Xu, "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk," *Journal of Finance*, vol. 56, no. 1, pp. 1–43, February 2001. [Online]. Available: https://ideas.repec.org/a/bla/jfinan/v56y2001i1p1-43.html

[6] S. R. Baker, N. Bloom, S. J. Davis, K. Kost, M. Sammon, and T. Viratyosin, "The Unprecedented Stock Market Reaction to COVID-19," *The Review of Asset Pricing Studies*, vol. 10, no. 4, pp. 742–758, 07 2020. [Online]. Available: https://doi.org/10.1093/rapstu/raaa008

[7] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, "Noise dressing of financial correlation matrices," *Physical Review Letters*, vol. 83, no. 7, p. 1467–1470, Aug. 1999. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.83.1467

[8] J. Bun, J.-P. Bouchaud, and M. Potters, "Cleaning large correlation matrices: Tools from random matrix theory," *Physics Reports*, vol. 666, p. 1–109, Jan. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.physrep.2016.10.005

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

[10] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304405X93900235

[11] M. Mazur, M. Dang, and M. Vega, "Covid-19 and the march 2020 stock market crash. evidence from sp1500," *Finance Research Letters*, vol. 38, p. 101690, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1544612320306668

[12] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, vol. 1, no. 2, p. 223, mar 2001. [Online]. Available: https://dx.doi.org/10.1088/1469-7688/1/2/304

[13] W. Ding, R. Levine, C. Lin, and W. Xie, "Corporate immunity to the covid-19 pandemic," *Journal of Financial Economics*, vol. 141, no. 2, pp. 802–830, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304405X21000957

[14] S. Baek, S. K. Mohanty, and M. Glambosky, "Covid-19 and stock market volatility: An industry level analysis," *Finance Research Letters*, vol. 37, p. 101748, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1544612320311843

[15] A. N. Berger and A. Demirgüç-Kunt, "Banking research in the time of covid-19," *Journal of Financial Stability*, vol. 57, p. 100939, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157230892100098X

[16] J. W. Goodell, "Covid-19 and finance: Agendas for future research," *Finance Research Letters*, vol. 35, p. 101512, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1544612320303974

[17] B. N. Ashraf, "Stock markets' reaction to covid-19: Cases or fatalities?" *Research in International Business and Finance*, vol. 54, p. 101249, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0275531920304141

STATEMENT OF CONTRIBUTION OF TEAM MEMBERS

This work is a joint effort by Mukul Chodhary and Luca Di Cola. We can divide the sections of work into two main parts, sections done individually and together.

As of this submission, the current split is as follows

1) Introduction: Both
2) Motivation: Mukul
3) First Stage: Luca
4) Second Stage and Analysis: Mukul

A. Future Plan

In the following weeks we plan to expand on the project as follows:

1) First stage (Mukul):
   a) Expand on the current work and relate the implication of the theory back to the stock market.
   b) Expand on the results and interpretation from the study and discuss.
   c) Make the current section a lot more readable and easy to follow and relevant to the overall project.
2) Second stage (Mukul and Luca):
   a) Provide a brief overview of methods used such as bi-plots, QQ-plots, spectral clustering
   b) Finish analysing the preliminary results.
   c) Find the shifts in sectors from before to after and interpret them, use literature to verify if similar trends were seen in other studies.
   d) Apply moving window approach to understand the changes in temporal behaviour from before and after.
   e) If possible relate the analysis to policies, which may be difficult as this is only for one country.
   f) Analyse the pcs outside the right support of the bulk
   g) Analyse the pcs close to the left support of the bulk
   h) reconstruct the correlation matrix using robust methods/ other estimation techniques.
3) Discussion (Mukul and Luca):
   Discuss the analysis section broadly and interpret the findings, relate everything back to real work and the impact of COVID.
4) Conclusion (Mukul and Luca):
   Conclude with overall insights and reflection
5) Proof-read and re-organise (Mukul and Luca)

The report also has coloured comments outlining how we plan to expand on the report.

Digitally Signed by:

Mukul Chodhary (1172562), Luca Di Cola (1652398)