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 (Luca and Mukul)

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. Traditional approaches often struggle with several limitations, including the assumption of linearity in relationships between variables,high-sensitivity outliers and estimation noise. RMT addresses these limitations by providing a robust framework for separating meaningful correlations from noise in large, complex datasets. 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. (2002) [1] 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. Project Motivation (Mukul)

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

[1]1172562, [2]1652398

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) **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

A. Problem Statement and Motivation of the study by Plerou et al. (Mukul)

Plerou et al. (2002) address the fundamental problem of modelling market dynamics by quantifying correlations in financial datasets. The primary motivations for the study were performing inference on the data and deepening the understanding of the underlying data-generating process. Recognising patterns in the stock market is crucial to improve predictions and facilitate the development of effective portfolios. However, due to the complexity, numerous hidden variables, and the non-linear nature of the underlying system, modelling the data-generating system is a challenging task.

Plerou et al. (2002) applied high-dimensional data analysis techniques, particularly Random Matrix Theory (RMT), to find the underlying structure of financial data. Their motivation for this approach was multifaceted. Traditional methods often struggle with high-dimensional datasets, leading to noisy estimates and potential overfitting. Financial markets exhibit complex interactions that are difficult to capture with conventional correlation analysis. RMT had previously successfully explained complex seemingly random systems, such as the energy levels of complex nuclei, suggesting its potential applicability to financial markets. Moreover, this approach separates genuine correlations from random noise, a crucial step in financial data analysis. RMT provides a theoretical framework to identify statistically significant deviations from randomness, potentially revealing meaningful market structures.

By applying RMT to financial data, Plerou et al. sought to overcome the limitations of traditional methods and uncover hidden patterns and structures in the market. Their work laid the groundwork for a more sophisticated understanding of market dynamics, particularly in the context of high-dimensional data. They analysed the 'deviating eigenvectors' from the RMT and their temporal stability. They attempted to interpret these deviations and their practical implications for optimal portfolio construction. The first stage of our study focuses on analysing the methods and results presented in Plerou et al. to understand how RMT can be effectively applied to investigate correlation patterns in complex, high-dimensional stock return data. This analysis serves as a foundation for understanding the application of RMT to financial markets, allowing us to evaluate the approach's strengths and limitations critically, and helping to identify potential improvements or adaptations for our specific study. These methods form the basis for the initial analysis of the post- and pre-COVID datasets in the second stage of our study (Section IV). By thoroughly examining their approach, we ensure a more robust and informed application of these techniques to our data.

This methodological foundation is pivotal as we seek to reveal how market dynamics have shifted in response to the unprecedented global shock of the COVID-19 pandemic. Our analysis aims to leverage RMT to uncover subtle changes in market structure and correlation patterns that may have emerged during this extraordinary period, potentially providing valuable insights into the resilience and adaptability of financial markets in the face of global crises.

B. The datasets (Luca)

The paper by Plerou et al. involves two datasets. Both 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 Automated Quotation (NASDAQ). The first dataset is the Trades and Quotes (TAQ) database, which documents all transactions for all major securities listed in all three stock exchanges. The times series (starting on January 3, 1994, with the length of 2 years) of prices of the 1000 largest stocks by market capitalization were analyzed. A total of $N = 6448$ records are in the time series analysis. Each record is a 30-min return of $M = 1000$ US stocks for the 2 years 1994-1995. From the same dataset, $N = 6448$ records of $M = 881$ 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 the 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. Plerou et al. analyzed daily returns for the stocks that survived for the 35 years 1962–1996 and extracted $N = 8685$ records of 1-day returns for $M = 422$ stocks.

C. Rationale for High-Dimensional Data Analysis (Luca and Mukul)

The sample covariance matrix (SCM) Σ_M of a multidimensional data matrix X = $\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$ can be estimated by $\hat{\Sigma}_M$ via:

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

where N is the number of samples, M is the number of features or stocks, and X_j is the j^{th} column of X or the stocks at time $t = j$. This convergence implies for a finite N, there is bound to be estimation noise in the $\hat{\Sigma}$, due to the finite length of the time series available. However, using a long-time series is not a valid solution for the financial data due to the non-stationarity of the cross-correlations between different market sectors. Moreover, 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 analyses. 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 (LLN) is a fundamental theorem that exemplifies the concept of universality in probability theory. Universality refers to the phenomenon where large-scale systems exhibit common patterns or behaviours, regardless of the specific details of the individual components. In the context of the LLN, Universality manifests in the convergence of the sample mean to the expected value for a wide range of probability distributions. This convergence often follows a Gaussian distribution, as described by the Central Limit Theorem (CLT), regardless of the underlying distribution of the individual random variables. However, Plerou et al. (2002) found that financial market data often deviates from these classical statistical expectations. Their analysis revealed that stock market correlations exhibit patterns and structures that cannot be fully explained by standard statistical models, necessitating more advanced analytical approaches such as Random Matrix Theory.

- **Dimensionality**: The study examines the daily returns of 1000 stocks over 2 years, resulting in a comparable number of features (1000 stocks) and observations (approximately 500 trading days). In other words, both $N \to \infty$ and $M \to \infty$, while their ratio M/N remains fixed. This scenario, where the number of stocks (M) and the number of observations (N) are of similar magnitude and both large, is precisely the setting where Random Matrix Theory (RMT) becomes applicable and useful for analyzing the correlation structure.
- **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 another CLT requirement as shown in Plerou et al. (2002). In the context of RMT, these fat-tailed distributions can significantly impact the eigenvalue spectrum of the correlation matrix, leading to deviations from the Marčenko-Pastur distribution, particularly in the tails of the eigenvalue spectrum.
- **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

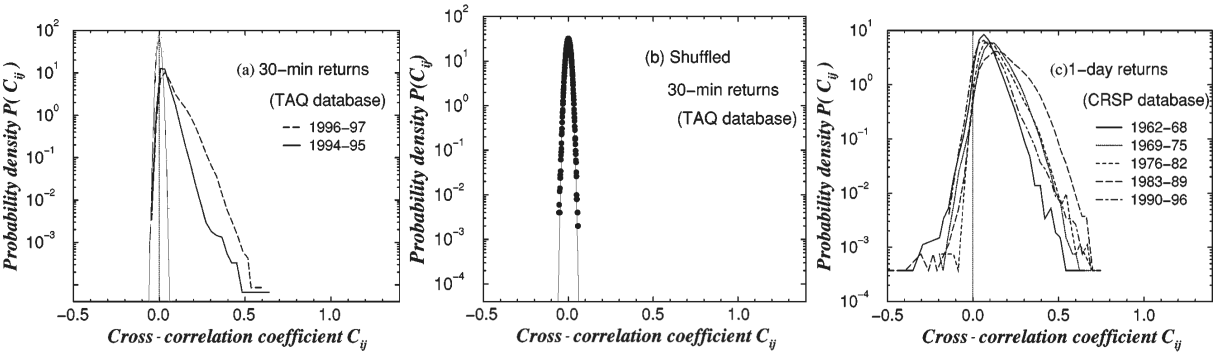


Fig. 1: Cross-correlation coefficient distributions from the Plerou et al.

dimensionality increases, necessitating the use of RMT to separate meaningful correlations from noise-induced artifacts.

Even though CLT cannot be used, universality is also shown under a different framework: the random matrix theory. The underlying idea is to study the SCM under the conditions:

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

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 $X_j \sim \mathcal{N}(0, I_M)$ - which can be interpreted as noise - then the sample covariance matrix eigenvalues' frequencies will follow a known distribution, the Marčenko-Pastur (MP) Law. Hence, if we do detect deviations from 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 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 (Mukul)

The analysis in Plerou et al. primarily employs correlation matrices and random matrix theory (RMT). The study follows a structured approach to analyse the financial data:

1) Statistics of the Correlation Coefficients: Plerou et al. analyses the financial dataset using normalised log returns. The normalisation is necessary to account for individual stocks' different volatility levels.

The log returns of a stock S_i over Δt are calculated by:

$$G_i(t) \equiv \ln S_i(t + \Delta t) - \ln S_i(t) \tag{1}$$

The normalised return is defined as

$$g_i(t) = \frac{G_i(t) - \langle G_i \rangle}{\sigma_i} \tag{2}$$

where $\sigma_i = \sqrt{\langle G_i^2 \rangle - \langle G_i \rangle^2}$ is the standard deviation of G_i, and $\langle \dots \rangle$ denotes the time average over the time period studied.

The cross-correlation C matrix is obtained with elements $C_{i,j}$ by

$$C_{i,j} = \langle g_i(t) g_j(t) \rangle \tag{3}$$

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. Plerou et al. analyse the distribution of C in Section III of their study, revealing several key insights into market behaviour.

The distribution $P(C_{ij})$ of the correlation matrix elements exhibits asymmetry and centres around a positive mean value, indicating a prevalence of positively correlated behaviour in stock markets. The distribution $P(C_{ij})$ was compared against a control correlation matrix - R constructed from mutually uncorrelated time series, $A : M \times N$, of similar size to the data.

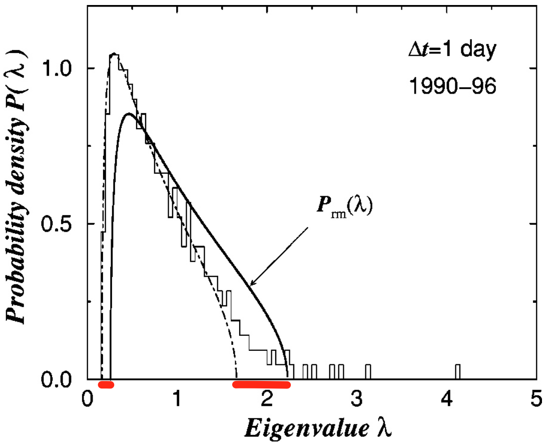


Fig. 2: $P(\lambda)$ for C, overlayed with P_{rm} (solid) and $P_{fit}(lambda)$ (dashed), Eq. 5 fitted on $P(\lambda)$ for $\lambda_{\pm}$. The red region corresponds to eigenvalues, λ^*, between P_{fit} and P_{rm}. Figure taken from Plerou et al.

$$R = \frac{1}{N} A A^T, A_{ij} \sim \mathcal{N}(0, 1) \tag{4}$$

The negative correlation components found in $P(C_{ij})$ were consistent within the zero mean Gaussian curves of the control $P(R_{ij})$, suggesting these negative correlations may be due to randomness.

The null model was constructed by randomly shuffling the 30-minute return database. The cross-correlation distribution of the null model was found to align with the $P(R_{ij})$ in their study, as seen in Fig. 1b. Thus, this alignment provides strong evidence that the randomisation process effectively removes any temporal correlations present in the original data. It establishes a reliable baseline for comparison, allowing researchers to distinguish between genuine correlations and those arising from statistical noise. This comparison is crucial for identifying significant deviations from randomness in the empirical data, which may indicate meaningful market structures or inefficiencies. Furthermore, the agreement between the null model and the Gaussian control validates the methodology of applying RMT used in the study, reinforcing the robustness of the subsequent analyses and conclusions drawn from the comparison with the empirical distribution.

Moreover, the average correlation $\langle C_{ij} \rangle$ exhibits significant temporal variation, with distinct correlation patterns emerging across different periods. Plerou et al. observed that epochs characterised by high market volatility, such as 1983-1989 and notably the 1987 market crash, corresponded directly to more pronounced cross-correlations. This volatility-correlation relationship manifests in Fig. 1c, where the cross-correlation distribution during these turbulent periods displays a markedly larger mean and wider spread. Such findings underscore the dynamic nature of market correlations and suggest a strong interplay between market turbulence and the degree of interdependence among stocks.

2) Eigenvalue Distribution of the Correlation Matrix: The analysis of the eigenvalue distribution of the correlation matrix C reveals important aspects of the correlation structure among stock returns. The study by Plerou et al. demonstrates that the empirical distribution of eigenvalues $P(\lambda)$ largely conforms to the Marchenko-Pastur (MP) law, as described by Eq. 5. This law provides a theoretical prediction for the distribution of eigenvalues in a random matrix, serving as a benchmark for identifying non-random features in empirical data.

The correlation matrix C is decomposed using eigendecomposition: $C = U \Sigma U^T$, where U is the matrix of eigenvectors and Σ is the diagonal matrix of eigenvalues. According to Random Matrix Theory (RMT), when both M and $N \to \infty$ with their ratio $\eta = \frac{M}{N} < 1$, the probability density function of the eigenvalues λ is given by:

$$P_{rm}(\lambda) = \frac{1}{2\pi\eta} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{5}$$

for λ within the bounds $\lambda_- \leq \lambda_i \leq \lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalues predicted by the MP law. These bounds are known exactly and are given by:

$$\lambda_{\pm} = 1 + \eta \pm 2\sqrt{\eta} \tag{6}$$

When comparing the empirical eigenvalue distribution $P(\lambda)$ with the theoretical distribution $P_{rm}(\lambda)$, it was found that a significant portion of the eigenvalues falls within these bounds, forming what is referred to as the ”bulk.” This bulk represents random correlations that are consistent with noise. However, some eigenvalues lie outside this bulk, with the largest eigenvalue being approximately 25 times greater than λ_+. This significant deviation suggests that these outlying eigenvalues capture genuine information about correlations among stocks, beyond what would be expected from random noise.

To further substantiate these findings, Plerou et al. compared these results with those obtained from uncorrelated time series, R and a shuffled null model. Both control cases conformed well to the MP law, reinforcing that deviations observed in empirical data are due to actual correlations present among stocks rather than random fluctuations.

For 1-day return data, while a bulk of eigenvalues was observed, its shape did not precisely conform to P_{rm} for the given η. The researchers fitted Eq. 5 to this data to determine bounds, yet they did not fully explore the implications of this discrepancy. The eigenvalues λ^* between these empirical, $P_{fit}(\lambda)$ and theoretical bounds $P_{rm}(\lambda)$ require further investigation to understand their significance, which will be addressed in Section IV. These eigenvalues could correspond to noise or may signify correlations between the stocks and may be explained by the Tracy Widom distribution [4], distribution of the λ_+. However, this hypothesis is yet to be tested and confirmed.

Despite the differences in two λ_+ and λ_+^{fitted}, we still observe a clear separation of the deviating eigenvalues from the bulk histogram after λ_+ in Fig. 2. From the perspective of RMT, the key difference between the 30-minute and 1-day datasets is the number of samples and stocks. From observation during workshops, we know that as M and N increase the $P(\lambda)$ conforms to $P_{rm}(\lambda)$ for a given η. This behaviour is also observed in Fig. 3 and 4 in the study by Plerou et al., where the 30-minute dataset has greater M and N compared to the 1-day dataset and the bulk of $P(\lambda)$ of this dataset aligns well with the $P_{rm}(\lambda)$.

3) Statistics of Eigenvectors:

a) ***Distribution of Eigenvectors****:* The analysis of eigenvector distributions provides further insights into the correlation structure, complementing the findings from the eigenvalue analysis. Plerou et al. examined the distribution of eigenvector components to identify deviations from RMT predictions.

Under RMT, the distribution of eigenvector entries should conform to a Normal distribution:

$$\rho_m(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{7}$$

where u represents the components of the eigenvector.

To quantify deviations from this theoretical distribution, Plerou et al. employed the kurtosis as a measure. For a zero-mean Gaussian distribution, the kurtosis is 3. The study used deviations from this value to identify significant non-Gaussian behaviour, although a specific threshold for significance was not explicitly stated.

The results aligned with expectations derived from the eigenvalue analysis:

1) Eigenvectors corresponding to eigenvalues within the bulk of the distribution showed close agreement with the zero-mean Gaussian distribution predicted by RMT. These eigenvectors had kurtosis values close to 3, consistent with Eq. 7.
2) Eigenvectors associated with eigenvalues deviating from the MP law exhibited non-Gaussian distributions. Specifically, the eigenvectors corresponding to the 20 highest eigenvalues had kurtosis values significantly different from 3, indicating a departure from RMT predictions.
3) The eigenvector corresponding to the largest eigenvalue displayed a particularly notable non-Gaussian behaviour, with all its components being non-negative. This characteristic is reminiscent of the ”market mode” concept introduced in the lectures of ELN90094, suggesting the presence of an underlying common trend or bias affecting all significant components of the market.

These findings provide strong evidence that the deviations observed in the eigenvalue distribution are indeed reflective of genuine correlations in the market, rather than statistical artefacts. The non-Gaussian nature of eigenvectors associated with the largest eigenvalues, particularly the all-positive components of the leading eigenvector, suggests the presence of market-wide factors influencing stock correlations. The consistency between the eigenvalue and eigenvector analyses reinforces the utility of RMT in distinguishing between random noise and meaningful market structures. However, the lack of a specified threshold for significant kurtosis deviation highlights that we must adopt a similar test such as the Jarque-Bera test [5] in our analysis during Section IV.

b) ***Interpretation of the Largest Eigenvalue and Corresponding Eigenvector****:* The analysis of the largest eigenvalue and its corresponding eigenvector provides crucial insights into the collective behaviour of the market. Plerou et al. hypothesised that this eigenvalue-eigenvector pair describes the market mode or the collective response of the market. To test this hypothesis, they constructed a portfolio using the components of the largest eigenvector and compared its return, denoted as G^{1000}, against the S&P 500 index return, G_{SP}, a standard measure of US stock market performance.

The portfolio return G^{1000} is defined as:

$$G^{1000} \equiv \sum_{j=1}^{1000} u_j^{1000} G_j(t) = u^{1000^T} \cdot G \tag{8}$$

where u_j^{1000} are the components of the largest eigenvector and $G_j(t)$ are the returns of individual stocks.

The comparison revealed a high correlation of 0.85 between G^{1000} and G_{SP}, with low variance, strongly suggesting that G^{1000} indeed captures the underlying influences affecting all significant components of the market. Importantly, this relationship held true for $\Delta t = 1$ day, implying that this behaviour is consistent across different time scales.

Furthermore, the study found that market volatility is strongly linked to the magnitude of the largest eigenvalue. This connection, combined with the interpretation of eigenvalues as variances in inverse Principal Component Analysis (PCA), provides a comprehensive picture: the largest eigenvalue of C over a time period indicates market volatility, while its corresponding eigenvector describes the market's behaviour. We hypothesise that similar patterns will be observed in the COVID-19 datasets analysed in Section IV.

Plerou et al. also outlined a method to remove the influence of G^{1000} from the remaining eigenvectors and reconstruct the cross-correlation matrix. This process reduced the average of C and the occurrence of high correlation values, suggesting that the filtered C could be used to identify market sectors that move together, independent of the overall market trend.

Interestingly, this analysis was made possible using a simple linear model to express the return G_i:

$$G_i(t) = \alpha_i + \beta_i M(t) + \epsilon_i(t) \tag{9}$$

where $M(t)$ is the market mode (approximated by G^{1000}), $\epsilon_i(t)$ are the residuals, and α_i and β_i are free parameters for the i^{th} stock. This model allows for the determination of stock-specific parameters, providing a framework for understanding individual stock behaviour in the context of overall market movements. The new cross-correlation matrix is then constructed using the residuals, $\epsilon_i(t)$.

*c) **Inverse Participation Ratio (IPR):*** Plerou et al. introduced the Inverse Participation Ratio (IPR) as a metric to quantify the number of components that contribute significantly to each eigenvector of the correlation matrix. The IPR for eigenvector k is defined as:

$$I^k = \sum_{m=1}^{M} [u_m^k]^4 \tag{10}$$

where $u_m^k, m = 1, ..., M$ are the components of eigenvector u^k. This metric provides insights into the localisation of correlations within the market structure. The IPR can be intuitively understood through two extreme cases: if all M components contribute equally, the IPR approaches $1/M$, while if only one component contributes, the IPR equals 1. This range allows for a nuanced interpretation of eigenvector composition across the spectrum of eigenvalues.

In their analysis across different time scales, Plerou et al. observed consistent patterns in the IPR behaviour. Eigenvectors corresponding to eigenvalues within the bulk of the distribution showed similar IPR values, aligning with RMT predictions. However, significant deviations in IPR were observed for eigenvectors associated with eigenvalues outside the bulk, indicating non-random structures in the correlation matrix. Notably, the eigenvector corresponding to the largest eigenvalue consistently exhibited the smallest IPR, close to $1/M$, across all time scales examined. This suggests that the largest eigenvector represents market-wide forces acting simultaneously across the entire stock market, with an approximately uniform distribution of its components. Conversely, eigenvectors associated with the smallest eigenvalues typically displayed very large IPR values. This characteristic suggests that these eigenvectors might encode information about smaller groups or pairs of stocks that have a disproportionate impact on market variance, despite their association with small eigenvalues.

Plerou et al. drew connections between their findings and the Anderson localisation theory, noting the presence of large IPR values in certain eigenvectors. This observation led to comparisons with random band matrices, which are often encountered in localisation phenomena. The analogy to random band matrices provides insights into the structure of the true correlation matrix, suggesting that significant correlations might be concentrated within a band around the diagonal of the matrix, with correlations diminishing as one moves away from this diagonal. The consistent behaviour of IPR across different time scales reinforces the robustness of this metric in capturing fundamental aspects of market structure, independent of the specific time horizon of the analysis. This consistency enhances the utility of IPR as a tool for analysing market correlations and identifying significant deviations from random behaviour.

*d) **Right Side of the Bulk:*** After removing the effect of the market mode represented by u^{1000}, Plerou et al. observed that eigenvectors from u^{999} to u^{990} exhibited components that contributed significantly to specific economic sectors. By examining the ten largest components of these eigenvectors associated with eigenvalues deviating from the bulk on the right side, they were able to identify and group stocks into distinct market groups. They only focused on the 10 largest eigenvalues as these meaningful groups become harder to find as you go closer to the bulk.

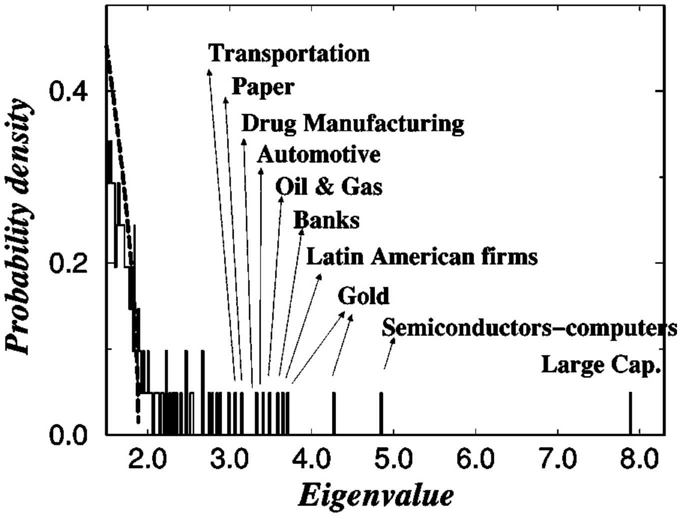


Fig. 3: Groups of stocks identified for the eigenvectors corresponding to the eigenvalues that deviate from the RMT upper bound. Plerou et al. Figure 13.

A notable finding was that individual eigenvectors corresponded to distinct industries within the market, suggesting that the correlation structure captured by these eigenvectors reflects sector-specific dynamics. This observation provides valuable insights into the hierarchical structure of market correlations, with the market mode representing the broadest level of correlation, followed by sector-specific correlations captured by subsequent eigenvectors. Plerou et al. discovered that not only industry classifications determined the stock correlations but also geographical factors. An example is the grouping of Latin American firms. This finding suggests that geography-related economic factors can lead to significant stock correlations, potentially rivalling or complementing industry-based relationships.

To validate their findings, the researchers applied various filtering methods beyond the regression approach described in Equation 9. The consistency of the identified groups across different methodologies reinforces the robustness of their results and suggests that the observed sector and geographical correlations are fundamental features of the market structure rather than artefacts of a particular analytical approach. These insights into sector-specific and geographical correlations provide a more nuanced understanding of market dynamics. They highlight the complex interplay between industry-specific factors, regional economic conditions, and broader market trends in shaping stock correlations. Such understanding can be valuable for portfolio diversification strategies, risk management, and the development of more sophisticated market models.

However, it is important to note that this analysis was conducted only on the 30-minute dataset, limiting the ability to conclude how these sector relationships might change during periods of market crisis, such as the 1987 market crash. This limitation presents a challenge when attempting to compare the results of similar analyses conducted in Section IV, as we lack a baseline for understanding how market sectors might behave across different types of crises. Therefore, our analysis may only be limited to the impact of COVID-19 on the stock market in the United States. The identification of these correlation structures also raises intriguing questions about their stability over time and their behaviour during different types of market stress. Future research could explore how these sector and geographical correlations evolve during various market conditions, potentially providing insights into the propagation of shocks through the financial system and the changing nature of market risks.

e) Left Side of the Bulk: Plerou et al. conducted an in-depth analysis of the eigenvectors corresponding to the smallest eigenvalues, which revealed intriguing properties of highly correlated stocks in the correlation matrix C. Their investigation combined empirical observations with theoretical analogies to explain the observed phenomena.

The researchers found that the eigenvectors associated with the smallest eigenvalues exhibited a distinctive characteristic: their largest contributing components often corresponded to pairs of highly correlated stocks. To elucidate this phenomenon, Plerou et al. employed an analogy using a 2×2 correlation matrix. This simplified model demonstrated that the largest contributing components of the smallest eigenvectors tend to have negative signs. This arises from the nature of eigenvector composition in such cases - the smaller eigenvector represents an antisymmetric linear combination of the basis vectors, while the larger eigenvector represents a symmetric combination. Furthermore, the study established a monotonic relationship between the cross-correlation coefficient and the eigenvalue magnitude. Specifically, they showed that as the cross-correlation coefficient between two stocks increases, the corresponding smaller eigenvalue decreases. This theoretical prediction aligned well with their empirical findings. A notable empirical example supporting their hypothesis involved the stocks of Texas

Instruments (TXN) and Micron Technology (MU). These two stocks exhibited the largest correlation coefficient in the dataset and, correspondingly, formed the largest components of the eigenvector u^1, associated with the smallest eigenvalue. Importantly, Plerou et al. also observed that TXN and MU were the largest components of u^{998}, one of the eigenvectors corresponding to the largest eigenvalues. This dual appearance in both extremes of the eigenvalue spectrum provided strong empirical evidence for their theoretical framework.

This analysis of the left side of the bulk offers valuable insights into the structure of stock correlations. It demonstrates that the smallest eigenvalues and their corresponding eigenvectors contain crucial information about the most strongly correlated pairs of stocks in the market. Such information can be particularly useful for identifying potential arbitrage opportunities, understanding market microstructure, or detecting anomalies in stock behaviour. The consistency between the theoretical predictions derived from the 2×2 matrix analogy and the empirical observations in the full market dataset underscores the robustness of this approach. It suggests that even in complex, high-dimensional financial systems, certain fundamental principles of linear algebra and correlation structure hold true and can provide meaningful insights into market dynamics.

These findings have potential implications for portfolio management, risk assessment, and market modelling. By identifying pairs of highly correlated stocks through eigenvalue analysis, investors and analysts can gain a deeper understanding of the market structure and potentially improve their strategies for diversification and risk management. Moreover, this approach offers a novel method for detecting and quantifying strong pairwise correlations in large, complex datasets, which could have applications beyond financial markets to other fields dealing with high-dimensional correlation structures.

4) Stability of Eigenvectors in Time: Plerou et al. investigated the temporal stability of correlations in the stock market by constructing and analysing an overlap matrix. This matrix is defined as:

$$O_{i,j}(t,\tau) = \sum_{k=1}^{n} D_{ik}(t) D_{jk}(t+\tau) \tag{11}$$

where D is a $p \times n$ matrix constructed from the p largest eigenvalues that deviate from the upper bound λ_+ of the Marchenko-Pastur distribution. This approach allows for a quantitative assessment of how correlation structures evolve over time. The study revealed that eigenvectors associated with the largest eigenvalues, specifically u^{1000}, u^{999}, and u^{998}, exhibit remarkable stability in their correlations over extended periods. This stability suggests that these eigenvectors capture fundamental, persistent structures in the market. As the time lag τ increases, the researchers observed two key phenomena:

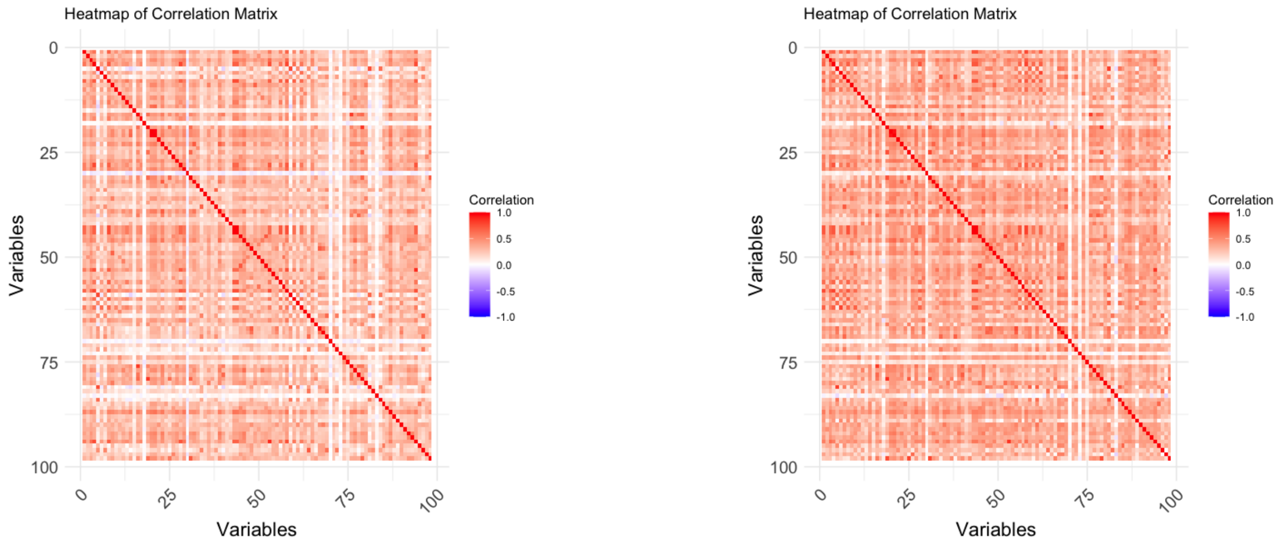
1) Correlations between different eigenvectors, which are theoretically orthogonal at $\tau = 0$, begin to deviate from zero.
2) The auto-correlation of individual eigenvectors gradually decreases.

Importantly, Plerou et al. found that the eigenvectors corresponding to eigenvalues deviating from the random matrix bulk maintain their stability for considerable periods. However, they noted a clear trend: the degree of stability diminishes as the associated eigenvalues approach the upper bound λ_+ of the bulk distribution. This pattern of decreasing stability held consistently across both 30-minute and daily return data, indicating that it is a robust feature of market dynamics across different time scales. The consistency across time scales reinforces the significance of these findings and suggests that they capture fundamental aspects of market behaviour.

The observed behaviour has important implications for understanding market structure and dynamics. While the deviating eigenvectors indeed encode significant correlations, the time-dependent nature of these correlations becomes more pronounced for eigenvalues closer to the bulk. This pattern suggests a hierarchy of stability in market correlations, with the most significant market-wide factors (represented by the top eigenvectors) being more persistent and the more nuanced or sector-specific correlations (represented by eigenvectors closer to the bulk) exhibiting higher temporal variability. These findings provide valuable insights into theoretical models of market behaviour and practical applications in finance. They suggest that while some correlation structures in the market are highly stable and may form a reliable basis for long-term analysis, others are more dynamic and require more frequent reassessment. The methodology employed by Plerou et al. in this analysis offers a powerful tool for quantifying and visualising the temporal evolution of market correlations. It provides a framework for distinguishing between persistent market structures and more transient correlations, which could be particularly valuable in understanding market behaviour during periods of significant change or stress.

IV. Second stage: Dataset introduction and aim (Mukul)

This study utilises 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 (1st Jan 2017 to 9th Jan 2020) to post-COVID (10th Jan 2020 to 31st Dec 2022).



(a) Pre-COVID Correlation Matrix

(b) Post-COVID Correlation Matrix

Fig. 4: Sample Correlation Matrix for Pre-COVID and Post-COVID Periods

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 time frame 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) Distinguish true correlation patterns from spurious relationships arising from high-dimensional statistical noise.
2) Quantify the level of statistical noise present in the estimated correlation values, providing a measure of uncertainty in our findings.
3) Uncover structured correlation patterns or networks that persist despite the presence of noise, revealing robust inter-stock dependencies.
4) Investigate the temporal stability of estimated correlation matrices and their properties, including eigenvalues and eigenvectors, to understand how stock relationships evolved during the pandemic.
5) Provide meaningful interpretations of the analysis results, offering insights into:
   - The complexity and nature of dependencies among stock returns
   - How these dependencies changed from the pre-COVID to post-COVID periods
   - The practical implications for investors and policymakers in understanding market dynamics during crises

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. Second stage: 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. We utilise the methodology introduced in Section III to analyse and contrast the underlying structure in these datasets. 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 resilience patterns, and evaluate the effectiveness of various policy responses implemented during this period.

A. Statistics of the correlation coefficients (Mukul)

The examination of raw correlation matrices for both pre-COVID and post-COVID periods provides initial insights into market dynamics and the pandemic's impact. Fig. 4 illustrates the sample correlation matrices for these periods, revealing considerable estimate noise in both datasets. The ratios of variables to samples are remarkably similar: $\eta = \frac{98}{759} = 0.129$ for the pre-COVID period and $\eta = \frac{98}{750} = 0.13$ for the post-COVID period. While these

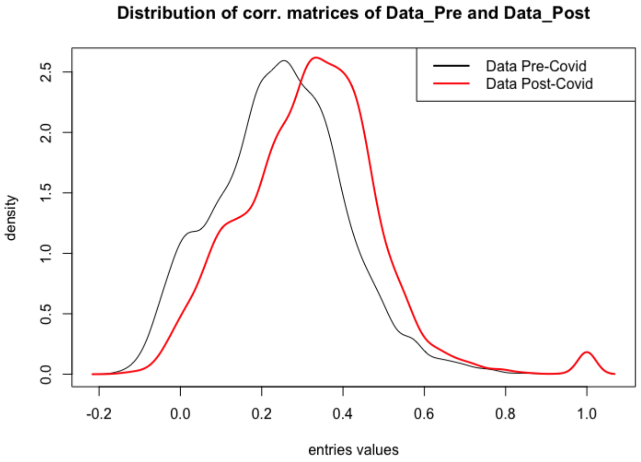


Fig. 5: $P(C_{ij})$ of the Pre- and Post-COVID Dataset.

matrices may contain underlying structures, they are not immediately apparent through visual inspection alone, necessitating a more in-depth analysis of the matrices' eigenvalues and eigenvectors to uncover these hidden patterns. The presence of noise in correlation matrix estimation is a common phenomenon in financial markets. As Plerou et al. demonstrated, RMT can be applied to differentiate between noise and meaningful information. The ratios observed in our data fall within the range of RMT applicability, suggesting that we can potentially identify genuine correlations amid the noise in both periods. This ability to extract meaningful information from noisy data is crucial for understanding market dynamics and making informed financial decisions. The structure of these correlation matrices has significant implications for portfolio management and risk assessment. Markowitz (1952) [6] established that understanding the correlations between assets is fundamental for efficient portfolio construction. The noise present in our matrices suggests that naive diversification strategies based on raw correlations might be suboptimal, potentially leading to underestimation of portfolio risk. This context underscores the importance of sophisticated analysis techniques to uncover true market structures and dependencies.

Despite the visual similarity of the correlation heatmaps, a closer examination of the correlation distribution $P(C_{ij})$ in Fig. 5 reveals notable differences between the pre-COVID and post-COVID periods. The average $P(C_{ij})$ for the post-COVID dataset is significantly higher compared to the pre-COVID period. This observation aligns with the findings of Campbell et al. (2001) [7] and Plerou et al., who noted that during periods of market calm, correlations between stocks tend to be lower. The higher correlations observed in the post-COVID period are reminiscent of the increased volatility observed in the 1987 crash dataset analysed by Plerou et al., suggesting a common pattern in market behaviour during times of crisis. This shift in correlation structure may indicate more complex underlying behaviour in response to the COVID-19 pandemic. Baker et al. (2020) [8] found that the stock market's reaction to COVID-19 was unprecedented in speed and severity compared to previous pandemics, suggesting a significant shift in market dynamics. While the visual differences in our correlation heatmaps are subtle, the changes in correlation distribution point to potentially profound alterations in market behaviour and interdependencies.

The observed increase in average correlations in post-COVID has significant implications for risk management and portfolio diversification. Higher correlations among assets typically reduce the benefits of diversification, as assets tend to move more in tandem, leading to increased portfolio volatility and potentially higher risks for investors relying on traditional diversification strategies. This phenomenon aligns with findings from previous market disruptions, such as the 1987 crash, suggesting that certain aspects of market behaviour during crises may be universal. Understanding these patterns could be valuable for developing more robust models of market dynamics during extreme events, potentially improving our ability to predict and manage market risks in future crises. In the subsequent sections, we delve deeper into analysing eigenvalues and eigenvectors, uncovering more nuanced structures within these correlation matrices. This further analysis will provide a comprehensive understanding of how the COVID-19 pandemic has altered market dynamics, potentially revealing sector-specific impacts and changes in the underlying factors driving stock returns. Such insights are crucial for adapting investment strategies and risk management approaches to the post-pandemic financial landscape.

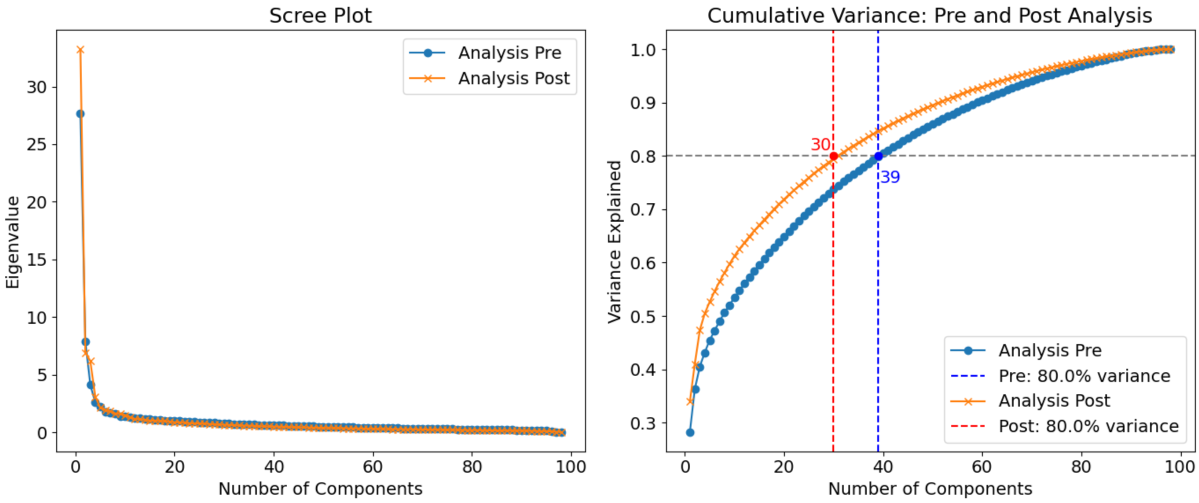


Fig. 6: Scree and Cumulative Variance Plots for Pre-COVID and Post-COVID Periods

B. *Eigenvalue distribution of the correlation matrix (Mukul)*

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.

a) ***Scree Plots****:* To analyze the eigenvalue distributions for both datasets, we utilized scree plots. A scree plot is a graphical representation that displays eigenvalues in descending order, plotted on a Cartesian plane. Each point on the plot corresponds to an eigenvalue, with the x-coordinate representing its position in the sequence and the y-coordinate indicating its magnitude. The primary objective of a scree plot is to help determine the number of significant components to retain by identifying where the curve begins to level off, known as the "elbow." This point suggests that additional components contribute little additional variance and can be excluded. The cumulative variance plots show the number of components required to capture the total variance in the principal components.

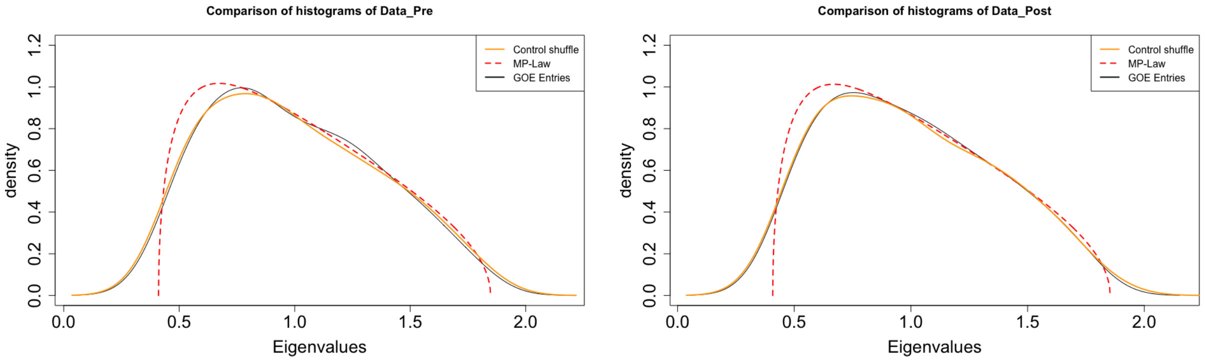
In our analysis, the scree plot for both datasets in Fig. 6 revealed an exponential decay pattern, which is characteristic of financial market data, as noted by Laloux et al. (1999) [9]. This pattern indicates that a few principal components capture most of the variance, while subsequent components contribute progressively less. Although the post-COVID scree plot is almost identical to that of the pre-COVID period, the cumulative variance plot shows that significantly fewer principal components are needed to explain the same amount of variance post-COVID. Specifically, to explain 80% of the variance, 39 components are required in the pre-COVID dataset, but only 30 in the post-COVID dataset. This consolidation of variance into fewer components post-COVID suggests increased market efficiency or more pronounced underlying trends driving stock movements during this period. The concentration of variance may also imply heightened market volatility, as fewer factors dominate market behaviour. This aligns with observations from previous market disruptions, where increased correlations among assets reduced diversification benefits and led to greater volatility.

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 MP Law.

c) ***Eigenvalue Distribution****:* In our analysis, we compared the eigenvalue distributions of pre-COVID and post-COVID datasets against the Marchenko-Pastur (MP) law, similar to the approach by Plerou et al. in Section III-D.2. We employed a Gaussian Orthogonal Ensemble (GOE) and a control shuffle null model, generating 500 realizations where each stock's data was randomly shuffled over time to eliminate temporal dependencies. Fig. 7 demonstrates that both the null models and GOE closely adhere to the MP law for both datasets. This alignment allows us to establish effective thresholds for distinguishing meaningful correlations from noise, as detailed in Section III.

The largest eigenvalue in the pre-COVID dataset is approximately 15 times larger than the theoretical upper bound $\lambda_+ \approx 1.85$, while it is about 18 times larger in the post-COVID dataset. This observation mirrors findings by Plerou et al., indicating that during crises, the largest eigenvalue becomes more pronounced, suggesting periods of heightened volatility. In Figure 8, we exclude the largest eigenvalue to focus on those near the bulk.

Notably, the bulk of the eigenvalue distribution in our data does not exactly match the theoretical bulk predicted by the MP law, similar to observations by Plerou et al. for the 1-day return dataset as shown in Fig. 2. Plerou et al. suggested fitting the MP law with supports as free parameters; however, they did not discuss the implications of



(a) Shuffled Pre-COVID Eigenvalue Distribution

(b) Shuffled Post-COVID Eigenvalue Distribution

Fig. 7: Comparison of shuffled Pre and Post-COVID Eigenvalue Distributions to MP-Law and GOE

eigenvalues lying between the fitted and limiting MP law curves. Fig. 9 reveals that for both pre- and post-COVID periods, a significant number of eigenvalues fall between these two curves. In upcoming sections, we demonstrate that fitting the MP law provides a more accurate estimate of noise using the IPR.

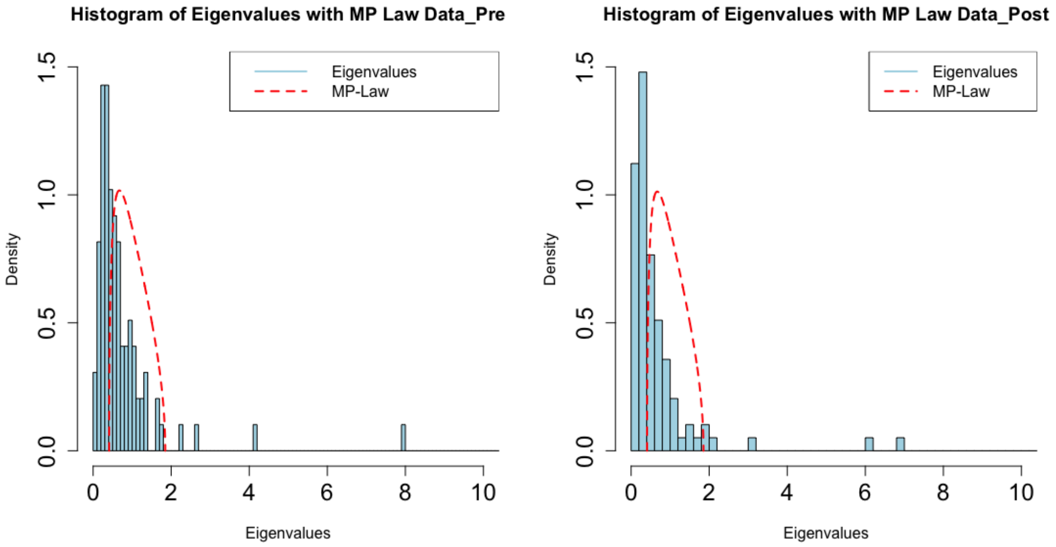


Fig. 8: $P(\lambda)$ for both Pre-and Post-COVID dataset. The largest eigenvalue is not shown in these plots.

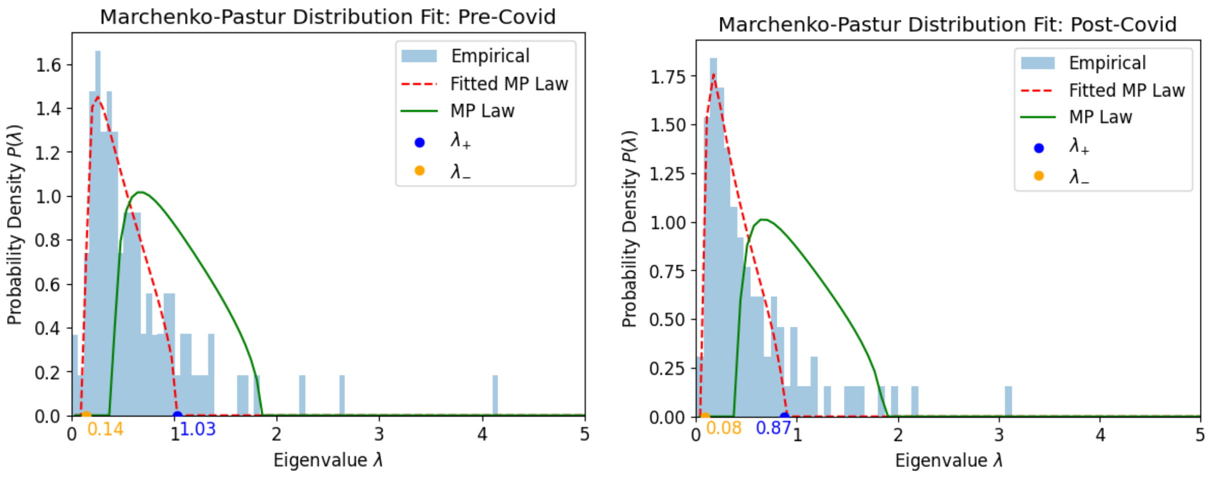


Fig. 9: MP law fitted to Pre and Post-COVID Eigenvalue Distributions

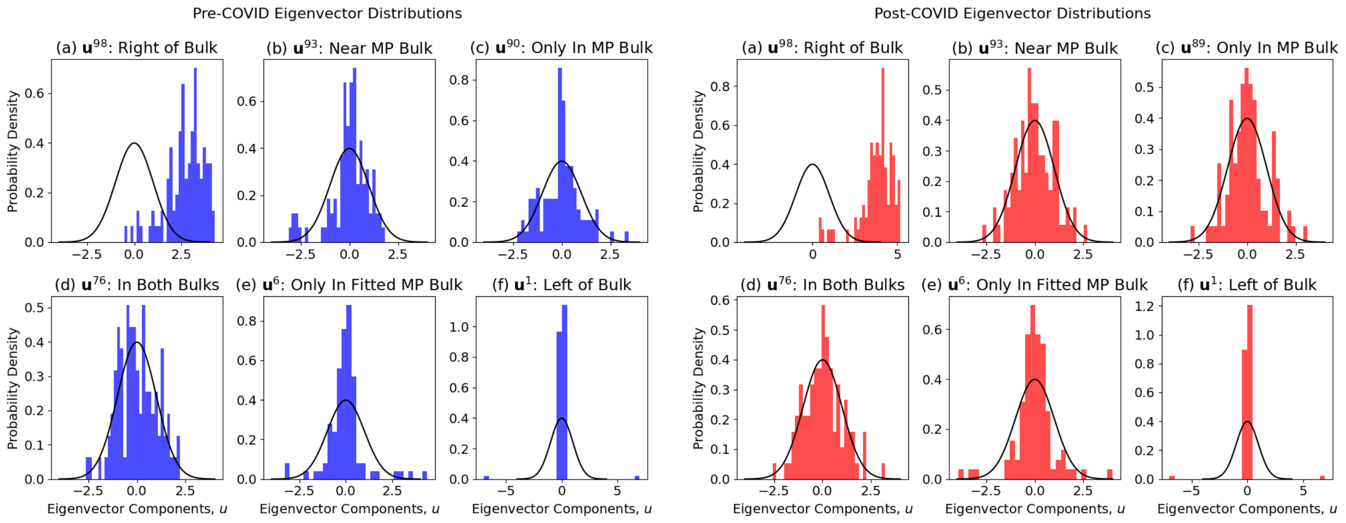


Fig. 10: Eigenvector Distributions in key regions corresponding to the bulk.

C. *Statistics of Eigenvectors (Mukul)*

1) Distribution of Eigenvector Components: In our analysis of eigenvector components, we focus on five key regions of the eigenvalue distribution: the right side of the fitted MP law, only in the fitted MP bulk, in the bulk region of both MP and fitted MP law, only in the MP bulk, and the left side of the MP law bulk. By examining these regions, we gain insights into the noise characteristics of each eigenvalue-eigenvector pair.

Fig. 10 provides valuable insights into these regions of interest. In panel (a), we observe that the largest eigenvector predominantly comprises positive components, with nearly all components contributing in both pre-COVID and post-COVID datasets. This behavior is consistent with findings by Plerou et al. and is characteristic of the market mode, where a single factor influences a broad range of stocks. As we move towards the right of the MP bulk, panel (b) reveals that eigenvectors begin to exhibit more noisy components. This transition becomes more pronounced in panel (c), where eigenvectors within the MP bulk appear purely Gaussian, suggesting that this region aligns with theoretical expectations for random noise. Panel (d) shows eigenvectors inside both bulks as purely zero-mean Gaussian, indistinguishable from noise. The region within only the fitted MP bulk is particularly intriguing; as seen in panel (e), these distributions are not purely Gaussian but lack characteristics typical of the left side of the bulk. This region indicates small groups of stocks that correlate together. On the left side of the bulk, shown in panel (f), we observe pairs of stocks exhibiting isolated behaviour with opposite signs in their eigenvectors. As Plerou et al. pointed out these decoupled stocks often correspond to stocks with correlation coefficients much larger than the average correlation coefficient, effectively decoupling them.

These observations hold true for both pre-COVID and post-COVID periods and are consistent with results reported by Plerou et al. This consistency underscores the robustness of our findings across different market conditions and highlights the utility of examining eigenvector distributions to discern underlying market dynamics. However, to effectively distinguish noisy eigenvectors from non-noisy ones, we require more robust methods. Therefore, we turn to the IPR and Gaussianity tests in the following section.

2) Number of Significant Participants in an Eigenvector: To assess the number of significant contributors within an eigenvector, we apply the Inverse Participation Ratio (IPR) as utilized by Plerou et al. This metric leverages the structure emerging from Random Matrix Theory (RMT), where smaller eigenvalues tend to have large IPR values, indicating fewer significant components. In contrast, larger eigenvalues exhibit very small IPR values, suggesting more distributed contributions with some convergence to a baseline within the bulk.

In addition to the IPR, we incorporate Gaussianity tests such as the Kolmogorov-Smirnov (K-S) test [10] and the Jarque-Bera (JB) test [5]. The K-S test evaluates the goodness of fit between a sample distribution and a reference Gaussian distribution, providing insights into how closely the eigenvector components follow a normal distribution. The JB test assesses whether sample data have skewness and kurtosis matching those of a normal distribution, offering a complementary perspective by focusing on these specific distributional properties.

These tests focus on different aspects of data distribution. The K-S test is sensitive to differences in both the center and tails of the distribution, while the JB test focuses specifically on skewness and kurtosis. A situation where the K-S test might indicate non-normality while the JB test suggests normality could occur if the sample has deviations in its overall shape that affect the fit to a normal distribution but still maintains skewness and kurtosis values close to those expected under normality. This discrepancy arises because the K-S test considers the entire distribution, while the JB test only considers skewness and kurtosis, potentially overlooking other

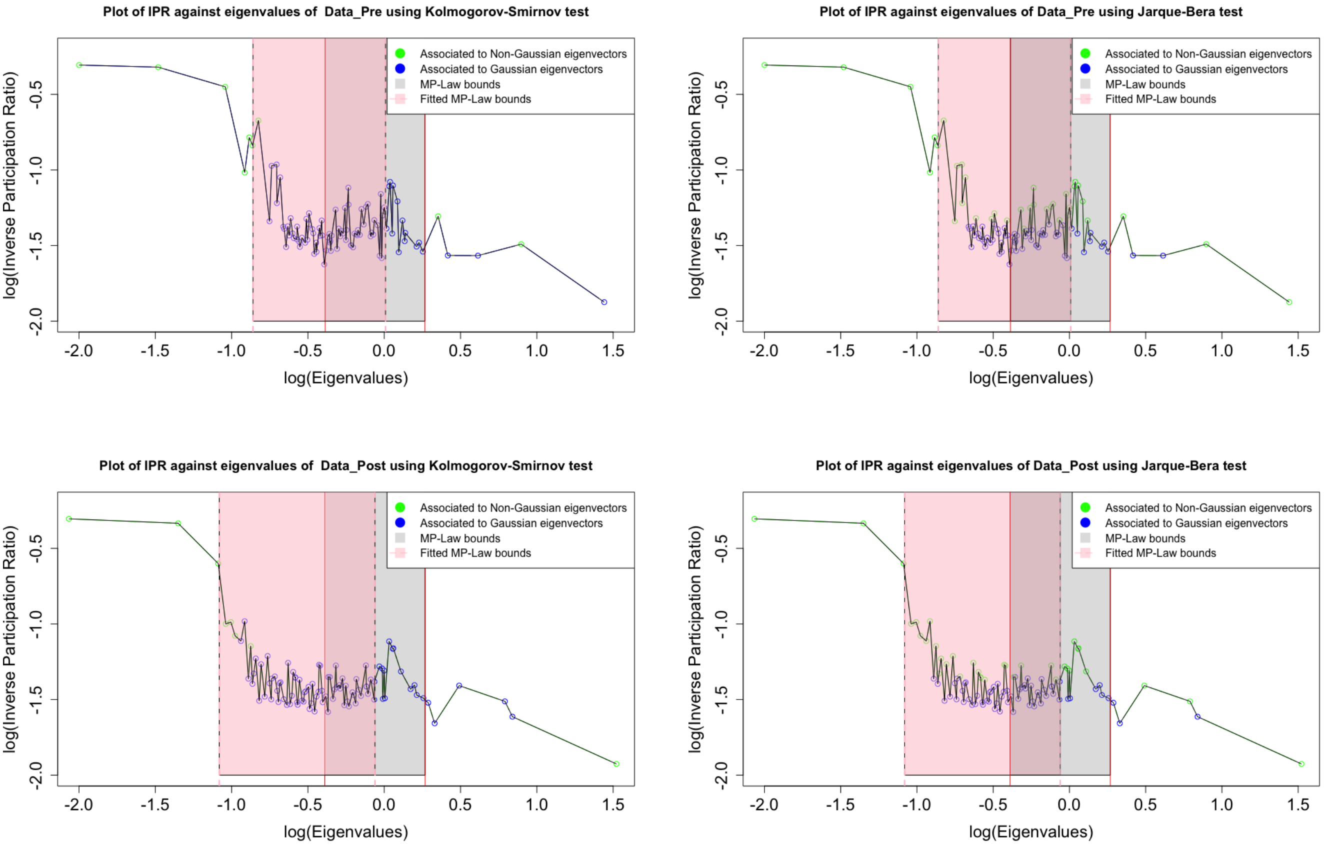


Fig. 11: Plots of Inverse Participation Ratio against eigenvalues for pre-COVID and post-COVID datasets using Kolmogorov-Smirnov and Jarque-Bera tests. The plots highlight regions associated with Gaussian and non-Gaussian eigenvectors, delineated by MP-Law and fitted MP-Law bounds.

deviations. This situation is common for eigenvector distributions, leading to varying results as seen in Fig. 11. The JB test often labels eigenvectors within the bulk as non-Gaussian due to discontinuous peaks separate from the common central zero mean unit variance distributions. However, it performs well for large eigenvectors where many components contribute but may include noisy regions, where the K-S test often fails. Conversely, the K-S test excels on the left side of the bulk, distinguishing well between entries corresponding to noisy parts.

Our analysis confirms that it is insufficient to consider only the supports of the limiting MP law as bounds for distinguishing between noisy and non-noisy components. Similarly, relying solely on the supports of the fitted MP law is inadequate. As illustrated in Fig. 11, effective bounds must encompass both limiting and fitted MP laws, specifically λ_-^{fitted} and λ_+. These bounds generally define the complete bulk well and allow for clear distinction between noisy components.

3) Interpretation of deviating eigenvectors (right of the bulk): To interpret and extract relevant groups from the largest eigenvectors, we follow the procedure outlined by Plerou et al. This involves removing the influence of the largest eigenvector u^{98} by reconstructing the correlation matrix C using the residuals $\epsilon_i(t)$ from Equation 9. This removal significantly shifts the distribution $P(C_{ij})$ to the left for both pre-COVID and post-COVID datasets, as illustrated in Fig. 13. This shift supports Plerou et al.'s claim that a substantial degree of correlations in C can be attributed to the influence of the largest eigenvalue. We also observe a clearer distinction of the eigenvalues outside the bulk in Fig. 13. Another clear observation is that the support of the fitted MP law is now much closer to the limiting MP law, suggesting a convergence towards a more accurate representation of noise and genuine correlations.

We also note that in this adjusted C, the mean and standard deviation of $P(C_{ij})$ are 0.00281 and 0.183 for the post-COVID dataset, and 0.00796 and 0.1628 for the pre-COVID dataset, respectively. These statistics imply a reduction in overall correlation strength post-COVID, with a slight increase in variability. This change suggests that while correlations remain significant, they are more dispersed, potentially reflecting increased market uncertainty and diversification in stock behaviours during the pandemic. This refined analysis enhances our understanding of market dynamics by isolating genuine correlations from noise, allowing for more precise modelling and prediction of market behaviour.

After this adjustment, we identify the largest significant eigenvectors using λ_+ as a threshold. The five most contributing components of these eigenvectors are listed in Table I for the pre-COVID period and Table II for the post-COVID period.

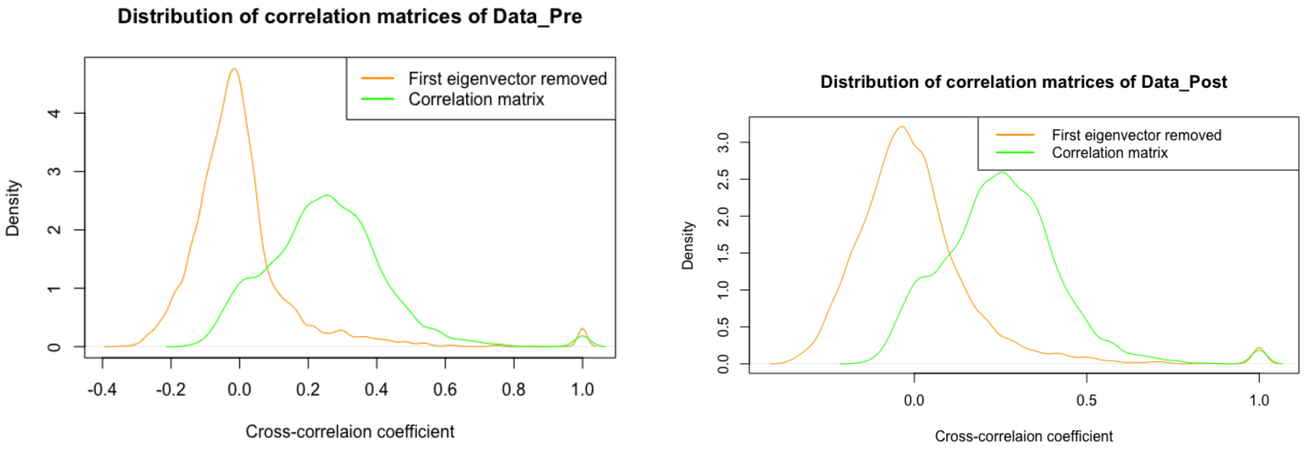


Fig. 12: $P(C_{ij})$ after removing the influence of the largest eigenvector.

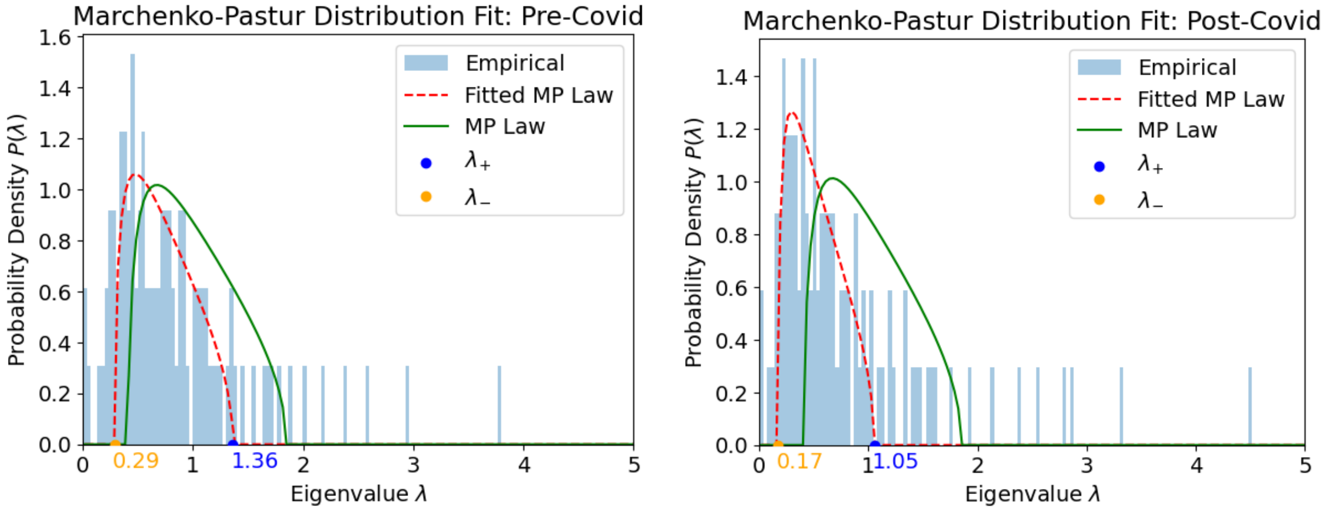


Fig. 13: $P(\lambda)$ after removing the influence of the largest eigenvector.

Similar to Plerou et al., we find that these groups often correspond to specific industries or closely related sectors. This clustering reflects inherent market structures where stocks within the same industry exhibit similar behavior due to shared economic drivers and sector-specific factors. Upon examining the shifts in industry groupings from pre-COVID to post-COVID, we observe notable changes:

- **Utilities and Essential Services**: In the post-COVID dataset, there is a noticeable increase in the presence of utility companies such as American Electric Power and NextEra Energy in eigenvector u^{97}. This shift reflects a heightened focus on essential services during the pandemic.
- **Technology and Financial Services**: The post-COVID table shows a stronger emphasis on technology and financial services. Companies like Adobe, Salesforce, and Visa appear more prominently in eigenvector u^{96}, indicating an increased reliance on digital solutions and financial transactions during the pandemic.
- **Healthcare**: There is a significant increase in the presence of healthcare companies in the post-COVID dataset. For instance, eigenvector u^{94} is made up of companies that were key players in medicine and vaccine development during the pandemic including Eli Lilly and Company and AstraZeneca. This shift highlights the critical role of pharmaceutical companies in responding to COVID-19 and their impact on market dynamics. While these companies did appear in the pre-COVID dataset, their impact was not as significant.
- **Energy Sector**: While energy companies like Shell and BP remain present in both datasets, their dominance is slightly reduced post-COVID as other sectors gain prominence.

Overall, these shifts from energy dominance pre-COVID to a more diversified post-COVID focus illustrate how market dynamics have evolved due to the pandemic. This transition may be driven by changes in consumer behaviour, technological advancements, and shifts in energy demand. By analyzing these shifts in eigenvector composition, we gain insights into evolving market dynamics, allowing for more informed investment strategies that

TABLE I: Largest five components of the eigenvectors u^{97} up to u^{88} for Pre-COVID dataset.

Symbol	Company Name	Sector
	u^{97}	
SHEL	Shell plc	Energy
BP	BP p.l.c.	Energy
COP	ConocoPhillips	Energy
CVX	Chevron Corporation	Energy
XOM	Exxon Mobil Corporation	Energy
	u^{96}	
NVDA	NVIDIA Corporation	Technology
AMZN	Amazon.com, Inc.	Consumer Cyclical
SO	The Southern Company	Utilities
GOOGL	Alphabet Inc.	Communication Services
META	Meta Platforms, Inc.	Communication Services
	u^{95}	
O	Realty Income Corporation	Real Estate
LOW	Lowe's Companies, Inc.	Consumer Cyclical
WELL	Welltower Inc.	Real Estate
SPG	Simon Property Group, Inc.	Real Estate
BHP	BHP Group Limited	Basic Materials
	u^{94}	
ETN	Eaton Corporation plc	Industrials
HON	Honeywell International Inc.	Industrials
MA	Mastercard Incorporated	Financial Services
BRK-B	Berkshire Hathaway Inc.	Financial Services
DIS	The Walt Disney Company	Communication Services
	u^{93}	
ABBV	AbbVie Inc.	Healthcare
HD	The Home Depot, Inc.	Consumer Cyclical
APD	Air Products and Chemicals, Inc.	Basic Materials
LOW	Lowe's Companies, Inc.	Consumer Cyclical
SRE	Sempra	Utilities
	u^{92}	
ETN	Eaton Corporation plc	Industrials
CMCSA	Comcast Corporation	Communication Services
LIN	Linde plc	Basic Materials
T	AT&T Inc.	Communication Services
APD	Air Products and Chemicals, Inc.	Basic Materials
	u^{91}	
HD	The Home Depot, Inc.	Consumer Cyclical
LOW	Lowe's Companies, Inc.	Consumer Cyclical
TJX	The TJX Companies, Inc.	Consumer Cyclical
ECL	Ecolab Inc.	Basic Materials
BUD	Anheuser-Busch InBev SA/NV	Consumer Defensive
	u^{90}	
DTEGY	Deutsche Telekom AG	Communication Services
BUD	Anheuser-Busch InBev SA/NV	Consumer Defensive
RHHBY	Roche Holding AG	Healthcare
XOM	Exxon Mobil Corporation	Energy
AZN	AstraZeneca PLC	Healthcare
	u^{89}	
ABBV	AbbVie Inc.	Healthcare
MRK	Merck & Co., Inc.	Healthcare
LLY	Eli Lilly and Company	Healthcare
AZN	AstraZeneca PLC	Healthcare
RHHBY	Roche Holding AG	Healthcare
	u^{88}	
BAC-PE	Bank of America Corporation	Financial Services
PLD	Prologis, Inc.	Real Estate
CTA-PB	EIDP, Inc.	Basic Materials
V	Visa Inc.	Financial Services
MA	Mastercard Incorporated	Financial Services

consider both historical trends and current market conditions. Furthermore, these insights can guide policymakers in crafting robust economic policies that support sectors pivotal to recovery and growth during crises.

4) Interpretation of Smallest Eigenvalues and Eigenvectors (Left of the Bulk): The analysis of the smallest eigenvalues and their corresponding eigenvectors provides insights into the most isolated or unique behaviors within the market. Table III and Table IV reveal significant shifts in industry prominence between the pre-COVID and post-COVID periods.

In the pre-COVID period, there were 10 significant eigenvectors on the left side of the bulk, compared to only 7 in the post-COVID dataset. This reduction suggests a consolidation of isolated market behaviors, possibly due to increased market integration during the pandemic. The presence of fewer isolated eigenvectors post-COVID may indicate that sectors have become more interconnected, reflecting broader market trends influenced by global events.

Interestingly, for u^1 in both datasets, the C_{ij} values were less than the mean, and all eigenvector components were negative. This could indicate that these stocks are inversely correlated with broader market trends, suggesting a hedging behavior or unique market positioning. This might also be an artifact of removing the influence of u^{98} from all eigenvectors, which could impact how these correlations are perceived.

The tables show that many of these eigenvectors correspond to highly correlated pairs found in the largest eigenvectors. For example, in the pre-COVID dataset, financial services like Visa and Mastercard appear prominently, indicating strong internal correlations within this sector. Post-COVID, there is a shift towards energy companies such as Shell and BP, reflecting changes in energy demand and market dynamics during the pandemic. Notably, the post-COVID period's significant eigenvectors no longer include groups like basic materials from pre-COVID u^7, real estate in u^8 and u^9, and consumer drinks in u^{10}. This suggests a shift away from these sectors towards a focus on essential services like energy. Though the COVID-19 may not be considered the sole reason for these changes as events like wars and complex geo-political dynamics changes of the energy industry during the same perioud might also be impacting this.

These differences highlight how COVID-19 impacted various industries. The increased presence of energy companies post-COVID suggests a shift in focus towards essential services and resources necessary during crises. In contrast, the pre-COVID data's emphasis on financial services reflects a period of economic stability and growth.

Linking back to Plerou et al.'s observations, these findings underscore how external shocks can significantly alter correlation structures within specific sectors. By examining these shifts, we gain valuable insights into evolving market dynamics and can develop more informed investment strategies that consider both historical trends and current conditions.

TABLE II: Largest five participants of the eigenvectors u^{97} up to u^{89} for Post-COVID dataset.

Symbol	**Company Name**	**Sector**
	u^{97}	
AEP	American Electric Power Company, Inc.	Utilities
NEE	NextEra Energy, Inc.	Utilities
SO	The Southern Company	Utilities
PEP	PepsiCo, Inc.	Consumer Defensive
O	Realty Income Corporation	Real Estate
	u^{96}	
ADBE	Adobe Inc.	Technology
MA	Mastercard Incorporated	Financial Services
CVX	Chevron Corporation	Energy
CRM	Salesforce, Inc.	Technology
V	Visa Inc.	Financial Services
	u^{95}	
BP	BP p.l.c.	Energy
SHEL	Shell plc	Energy
BHP	BHP Group Limited	Basic Materials
RIO	Rio Tinto Group	Basic Materials
BRK-B	Berkshire Hathaway Inc.	Financial Services
	u^{94}	
LLY	Eli Lilly and Company	Healthcare
MRK	Merck & Co., Inc.	Healthcare
NVO	Novo Nordisk A/S	Healthcare
AZN	AstraZeneca PLC	Healthcare
JNJ	Johnson & Johnson	Healthcare
	u^{93}	
COST	Costco Wholesale Corporation	Consumer Defensive
ETN	Eaton Corporation plc	Industrials
CAT	Caterpillar Inc.	Industrials
DE	Deere & Company	Industrials
WMT	Walmart Inc.	Consumer Defensive
	u^{92}	
ACN	Accenture plc	Technology
COP	ConocoPhillips	Energy
CVX	Chevron Corporation	Energy
XOM	Exxon Mobil Corporation	Energy
APD	Air Products and Chemicals, Inc.	Basic Materials
	u^{91}	
SPG	Simon Property Group, Inc.	Real Estate
O	Realty Income Corporation	Real Estate
PLD	Prologis, Inc.	Real Estate
MDLZ	Mondelez International, Inc.	Consumer Defensive
LLY	Eli Lilly and Company	Healthcare
	u^{90}	
HD	The Home Depot, Inc.	Consumer Cyclical
LOW	Lowe's Companies, Inc.	Consumer Cyclical
BRK-B	Berkshire Hathaway Inc.	Financial Services
BRK-A	Berkshire Hathaway Inc.	Financial Services
TJX	The TJX Companies, Inc.	Consumer Cyclical
	u^{89}	
APD	Air Products and Chemicals, Inc.	Basic Materials
LIN	Linde plc	Basic Materials
LMT	Lockheed Martin Corporation	Industrials
RTX	RTX Corporation	Industrials
ECL	Ecolab Inc.	Basic Materials

TABLE III: Largest two participants of the eigenvectors u^1 up to u^{10} for Pre-COVID. $Mean(C_{ij}) = 0.00796$

Symbol	Company Name	Sector
	u^1	0.006
CTA-PB	EIDP, Inc.	Basic Materials
AMD	Advanced Micro Devices, Inc.	Technology
	u^2	0.971
GOOGL	Alphabet Inc.	Communication Services
GOOG	Alphabet Inc.	Communication Services
	u^3	0.921
BRK-A	Berkshire Hathaway Inc.	Financial Services
BRK-B	Berkshire Hathaway Inc.	Financial Services
	u^4	0.766
NEE	NextEra Energy, Inc.	Utilities
AEP	American Electric Power Company, Inc.	Utilities
	u^5	0.732
V	Visa Inc.	Financial Services
MA	Mastercard Incorporated	Financial Services
	u^6	0.754
BP	BP p.l.c.	Energy
SHEL	Shell plc	Energy
	u^7	0.718
BHP	BHP Group Limited	Basic Materials
RIO	Rio Tinto Group	Basic Materials
	u^8	0.68
WELL	Welltower Inc.	Real Estate
O	Realty Income Corporation	Real Estate
	u^9	0.398
O	Realty Income Corporation	Real Estate
SO	The Southern Company	Utilities
	u^{10}	0.642
KO	The Coca-Cola Company	Consumer Defensive
PEP	PepsiCo, Inc.	Consumer Defensive

TABLE IV: Largest two participants of the eigenvectors u^1 up to u^7 for Post-COVID. $Mean(C_{ij}) = 0.00281$

Symbol	Company Name	Sector
	u^1	-0.032
PLDGP	Prologis, Inc.	Real Estate
NVDA	NVIDIA Corporation	Technology
	u^2	0.979
GOOG	Alphabet Inc.	Communication Services
GOOGL	Alphabet Inc.	Communication Services
	u^3	0.867
BRK-A	Berkshire Hathaway Inc.	Financial Services
BRK-B	Berkshire Hathaway Inc.	Financial Services
	u^4	0.819
SHEL	Shell plc	Energy
BP	BP p.l.c.	Energy
	u^5	-0.068
HD	The Home Depot, Inc.	Consumer Cyclical
SO	The Southern Company	Utilities
	u^6	0.716
COP	ConocoPhillips	Energy
XOM	Exxon Mobil Corporation	Energy
	u^7	0.709
COP	ConocoPhillips	Energy
CVX	Chevron Corporation	Energy

D. Stability of eigenvectors in time (Luca)

We studied the overlap matrices $\mathbf{O}$ using daily returns. The number of records was $L_{\textbf{pre}} = 759$ for the pre-COVID dataset and $L_{\textbf{post}} = 750$ for the post-COVID dataset. We used a moving window of length $W = 100$ records and conducted five experiments on each dataset, each time increasing τ, the number of days by which to slide the window. The step sizes were $\tau = n\frac{L}{5}$, with $n = \{1, 2, 3, 4, 5\}$. For each time period, we first identified the eigenvectors of the window. In the pre-COVID dataset, there were five such eigenvectors, while in the post-COVID dataset, there were six. We retained only the eigenvectors associated with eigenvalues greater than λ_+. We then calculated the overlap matrix $\mathbf{O}(t_0, n\frac{L}{5})$ between the eigenvectors for $t = 0$ and $t = \tau$, as described in Section 4: First Stage.

In Fig. 14, the overlap matrices are visualized as heatmaps. In both Fig. 14a and 14b, the heatmap on the top-left corner represents $\tau = 20$, and the one to its right $\tau = 40$; the heatmaps below these corresponds to $\tau = 60$ and $\tau = 80$, respectively, with the last one showing $\tau = 100$.

In Fig. 14a, the first phenomenon we noticed was an overall increase in the color intensity of the off-diagonal elements of the heatmaps as τ increased– the dot products approached values closer to ± 1. This behavior, also noted by Plerou et al., reflects changes over time in the sectors represented by the eigenvectors (i.e., decreased stability). Although eigenvectors are expected to remain perpendicular, shifts in market conditions across time lags cause them to deviate, losing this perpendicularity. For smaller values of (up to $\tau = 40$), the projections of the second and third eigenvectors have values distinctly different from both zero and the off-diagonal elements, indicating a high level of stability in these eigenvectors' contributions to market variability. However, when $\tau > 40$, the projections along the diagonal become more similar to the off-diagonal values, signaling a decrease in eigenvector stability. These results align with Plerou et al.'s findings, showing that eigenvector stability diminishes as τ increases, particularly for eigenvectors associated with eigenvalues close to λ_+.

Although Fig. 14b also shows a decline in stability as τ, it is evident that post-COVID returns exhibit more eigenvector stability than pre-COVID returns. At $\tau = 20$, he first four eigenvectors display positive projections (especially strong in the first three), along with nearly perpendicular orientations (i.e., dot products close to zero) between them. This pattern persists at $\tau = 40$ but decays quickly in the following heatmaps, as in the pre-COVID case. However, the increase in off-diagonal values is less pronounced in Fig. 14a, especially for the first eigenvector, which not only remains stable but also retains perpendicularity with the others.

Our analysis confirmed two key phenomena as τ increases:

1) Correlations between different eigenvectors, which are theoretically orthogonal at $\tau = 0$, begin to deviate from zero.
2) The auto-correlation of individual eigenvectors gradually decreases.

Note that the stability of eigenvectors does not necessarily indicate periods of high or low volatility. In a highly volatile market, the correlation matrix often exhibits high values; however, eigenvectors are more concerned with economic sectors or geographical locations. Thus, it is possible to have a volatile market with stable eigenvectors, where the sources (i.e., economic sectors or geographical regions) of volatility (price fluctuations) are consistent over time. Conversely, in periods of low volatility, the eigenvectors could be unstable, with the diagonal elements of the overlap matrix approaching zero. In other words, the sources of price variability shift over time, but the overall price fluctuations remain relatively low.

In Tables I and II, we examined the five largest components of eigenvectors on the right side of the bulk, noting that each eigenvector often represents companies within the same economic sector, similar to the findings of Plerou et al. We also considered whether the ranking of stock market indexes associated with the largest entries changes over time and by how much. Do the same indexes dominate a particular eigenvector across all time lags?

To address this question, we analyzed the persistence of the indexes of the largest components of the eigenvectors by tracking changes in the ranking of their top 10 entries over time. We employed the same sliding window approach as before, with a different similarity metric. For each eigenvector (five in the pre-COVID period and six in the post-COVID period), we identified the positions of the top 10 entries (by absolute value) and compared them across successive windows. A change was recorded whenever the top entries' positions differed between consecutive time windows. The number of changes per eigenvector per cycle was documented and visualized using a boxplot.

The analysis was repeated with the same step widths $\tau = \{20, 40, 60, 80, 100\}$, allowing us to observe how the results vary at different time resolutions. Finally, multiple boxplots (Fig. 15) were generated for each value of τ in both the Pre and Post-COIVD to visualize stability and assess how fluctuations in market conditions impact the composition of these leading eigenvectors.

The measure used to compare the change between two eigenvectors was a varivation of the Szymkiewicz–Simpson coefficient, also known as the Overlap coefficient. The Overlap coefficient is a common measure of similarity between two sets, A and B, and takes values in the interval $[0, 1]$, where 1 means that A is a subset of B (or vice versa), and 0 means that $|A \cap B| = 0$.

$$\text{Overlap coefficient} = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{12}$$

In our case, the sizes of $|A|$ and $|B|$ were the identical, which is why a modified form of the Overlap coefficient was used.

Let p_t be an eigenvector at time t, and $p_{t+\tau}$ be the corresponding eigenvector in the time-shifted window at $t + \tau$. Let E_t be the set of the stocks indexes associated with the j largest entries (by absolute value) in p_t, and let $E_{t+\tau}$ be defined similarly for $p_{t+\tau}$. Then, the Overlap coefficient, given that $|E_t| = |E_{t+\tau}|$ is the cardinality of their intersection divided by j:

$$\frac{|E_t \cap E_{t+\tau}|}{j} \tag{13}$$

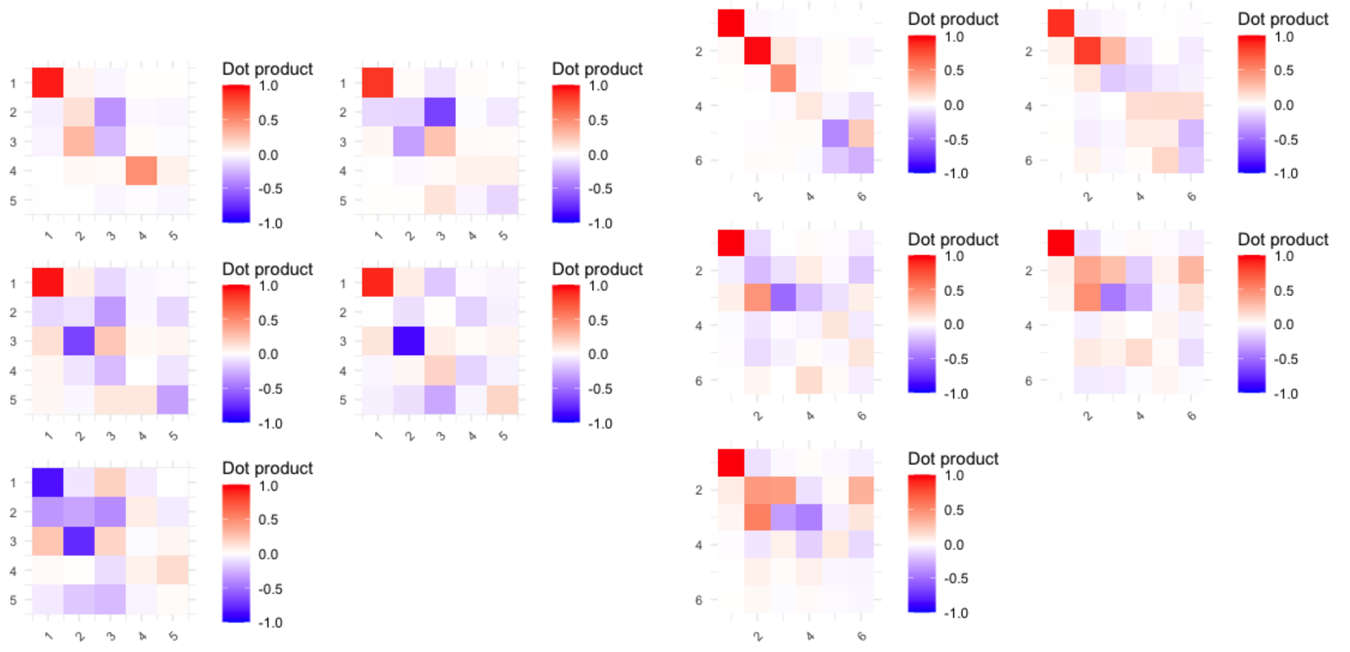
In our analysis, we introduced a measure called the *change coefficient* (CC):

$$\text{CC} = 1 - \text{Overlap coefficient} \tag{14}$$

We will now discuss Fig. 15, focusing on individual values of τ .

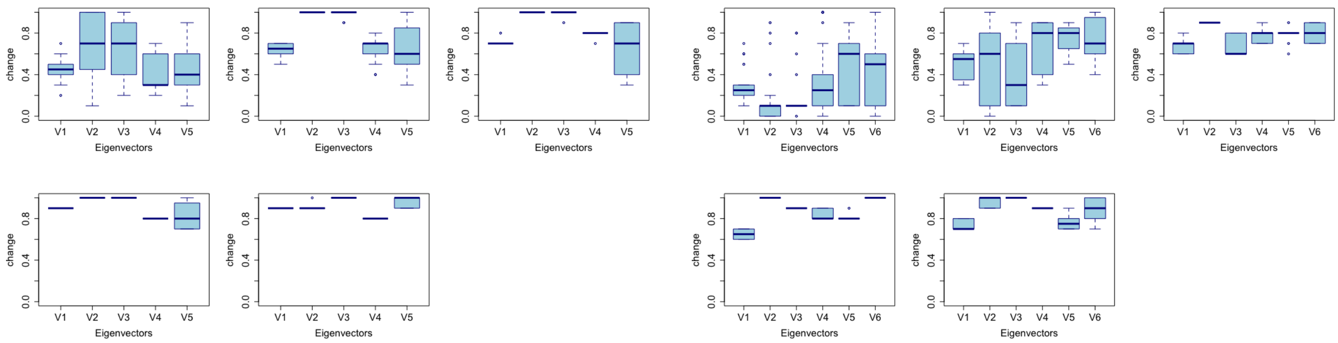
For $\tau = 20$, the pre-COVID data shows higher CC values compared to the post-COVID data. This observation aligns with the respective heatmaps, where the main diagonal in the pre-COVID heatmap has values closer to 0, indicating more unstable eigenvectors. However, eigenvectors 4 and 5 have similar CC values across both datasets. This may suggest that the first three post-COVID eigenvectors capture stable patterns absent in the pre-COVID data, while eigenvectors 4 and 5 capture behaviors common to both datasets.

For $\tau = 40$, the pre-COVID dataset still exhibits higher CC values than the post-COVID dataset, as confirmed by their respective heatmaps, where the main diagonal in pre-COVID has values closer to 0, indicating less stability. Overall, both datasets show an increase in CC at this step size. A notable difference is seen in eigenvectors 3 and 4 between the datasets: in pre-COVID, they rarely retain the same stock market indexes across time windows, with CC values nearly equal to 1. In post-COVID, they display significant CC variability, with values even lower than those of the first eigenvector; particularly, eigenvector 3 has a mean CC lower than the first eigenvector. This finding supports the idea that high volatility (post-COVID) can coexist with stable eigenvectors. The broader CC distribution for the second and third eigenvectors could indicate that sectors undergo periodic changes, allowing some cycles of stability—hence lower CC values—and others of pronounced shifts in composition, resulting in higher CC values. For the first eigenvector, it seems that only a subset of the 10 tracked indexes remain stable across cycles, while others vary (though this hypothesis cannot be confirmed solely by the boxplots). Eigenvectors 4 and 5 show similar CC values across datasets.



(a) Pre-COVID heatmaps of the overlap matrices $\mathbf{O}_{\mathbf{Pre}}(t, \tau)$ (b) Post-COVID heatmaps of the overlap matrices $\mathbf{O}_{\mathbf{Post}}(t, \tau)$

Fig. 14: Heatmaps of the overlay matrix



(a) Pre-COVID, first 5 eigenvectors

(b) Post-COVID, first 6 eigenvectors

Fig. 15: Boxplots of the values of CC

For $\tau = \{60, 80, 100\}$, both datasets exhibit a gradual CC increase, which is expected as eigenvector stability typically decreases with longer time lags. There are several noteworthy observations in these boxplots. First, the average CC of the first eigenvector in post-COVID is slightly lower than in pre-COVID and shows a narrower distribution. Additionally, the average CC in post-COVID remains between 0.6 and 0.8, suggesting that 2 to 4 indexes are retained in each time shift, while in pre-COVID, the average is above 0.9 for $\tau = 80,$ and 100. This indicates that the first post-COVID eigenvector is more consistent in describing certain indexes than the pre-COVID eigenvector. Secondly, nearly all pre-COVID eigenvectors exhibit consistently high CC values, indicating that their first 10 stock market indexes change continuously.

Several key insights arise from comparing the heatmaps with the boxplots. Firstly, heatmap trends do not directly correspond to those in the boxplots; high or low dot product values can occur with any CC value and vice versa. For example, the dot product for the first eigenvector is close to 1 across both datasets, yet the CC values span almost the entire interval [0,1]. Secondly, any eigenvector to the right of the bulk can have a lower CC value than eigenvectors associated with larger eigenvalues. For instance, the fifth eigenvector in post-COVID at $\tau = 80$ and $\tau = 100$ consistently has lower CC values than the second, third, and fourth eigenvectors in the same boxplots. In the pre-COVID dataset at $\tau = 80$ and at $\tau = 100$, the fourth eigenvector has lower CC values than the others; however, the corresponding heatmaps display dot products close to 0. This finding does not contradict Plerou's results on time stability, as he did not use the CC metric, which focuses solely on the first 10 entries. Nonetheless, within the set of eigenvectors to the right of the bulk, higher eigenvalues do not always correspond to lower CC values. This suggests that eigenvectors explaining most market variability via specific sectors might be less consistent in sector selection than eigenvectors with lower eigenvalues.

1) Tracking the movement of the first 10 entries of the first eigenvector: Through the boxplots in Fig. 15, we observed that many entries of the eigenvectors change across time frames. A natural leading question is: where do they go? To answer this, we devised a new experiment. The experiment was conducted on both datasets. The idea is to track the first 10 entries of the first eigenvector across time frames. We chose to use $\tau = 20$ with the usual time window $W = 100$. We began by identifying the top 10 largest entries of the first eigenvector in

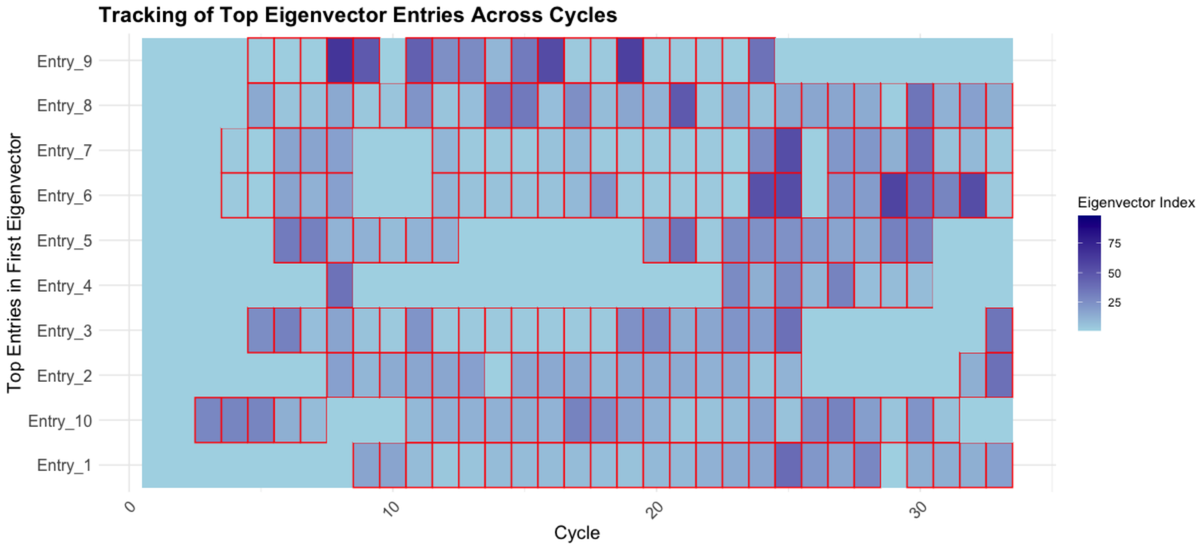


Fig. 16: Tracking of pre-COVID top eigenvector entries across cycles

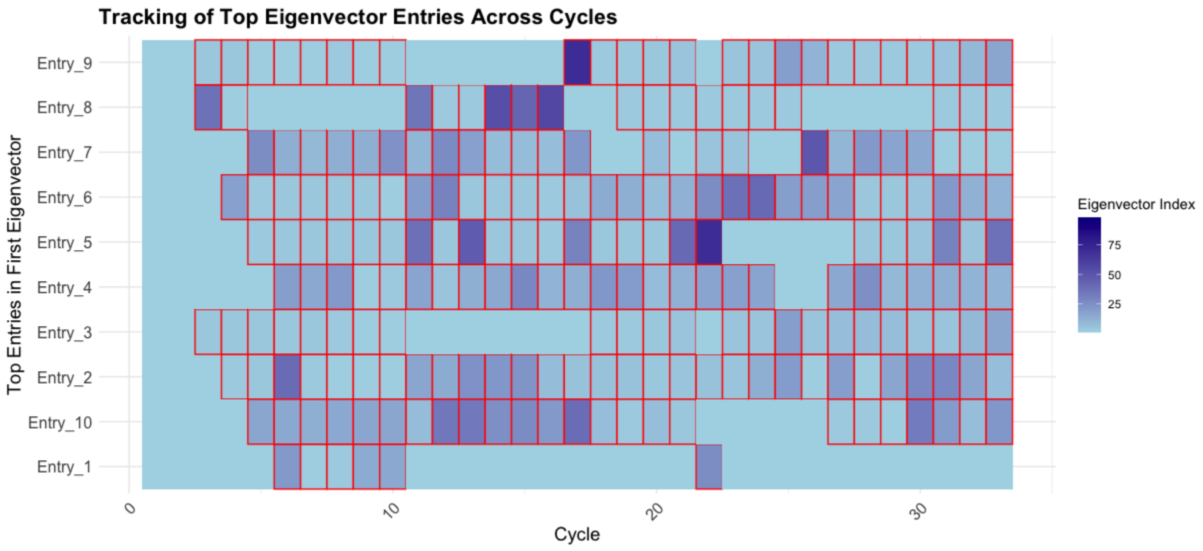


Fig. 17: Tracking of post-COVID top eigenvector entries across cycles

the initial time window, storing each entry's position as a separate target to track. For each time window (or cycle), we checked whether the same entry retained its prominence within the top 10 entries of the corresponding eigenvector. If the entry remained in the top 10, we recorded the eigenvector's index in a tracking matrix. If not, the algorithm searched through all eigenvectors in the current cycle to find the first one where the tracked entry reappeared in the top 10 positions. The resulting output is a matrix where each row corresponds to one of the initially tracked entries, and each column represents a time window.

The results are presented in Figures 16 and 17. The entries are arranged along the rows, and each column represents a cycle of the time window (a new value of τ). The colour of the tiles represents the eigenvector associated with that entry (row) for that cycle (column). Dark blue corresponds to eigenvectors associated with very small eigenvalues, while lighter blues indicate eigenvectors associated with larger eigenvalues. The borders of the tiles are red if they represent an eigenvector other than the first.

In the pre-COVID dataset (Figure 16), the indexes tend to cluster around the eigenvectors associated with eigenvalues on the right side of the bulk. In contrast, in the post-COVID dataset, the 10 tracked indexes "wander" more around eigenvectors within the bulk or even on the left side (note the dark blues in Figure 17). These results warrant further exploration by examining additional eigenvectors simultaneously and at different time resolutions. If these patterns are confirmed, they could have significant implications for the analysis of high-dimensional financial data. We know that post-COVID data exhibit high volatility, and the analysis suggests that this volatility may cause the left side of the bulk to influence certain market behaviours in conjunction with the eigenvectors on the right side. Here, we refer to instances of certain stocks shifting from the right side to the left side and back.

Secondly, many of the indexes frequently correspond to eigenvectors other than the first for most cycles. Thirdly, the hypothesis drawn from the boxplot analysis at $\tau = 40$ regarding some eigenvector indexes undergoing periodic cyclical changes while others experience shifts seems to be confirmed.

However, we should keep in mind that the association of these indexes with the first eigenvector is dependent on the initial choice of t_0. If we were to start the experiment at $t_0 = 20$, we would observe different results after only one cycle.

E. Market Structure Analysis (Mukul)

The next question that naturally occurs is how these identified sectors or individual stocks are related. While the earlier analysis of eigenvalues and eigenvectors gave us a comprehensive picture of different industries that naturally arise from the correlation matrix structure, it does not tell us how these might be interrelated.

In this section, we combine the ideas of robust covariance estimators and partial correlations to visualize the relationships between stocks.

1) Linear Shrinkage Estimator: Ledoit-Wolf Estimator: In the assignments of ELEN90094, we explored the linear shrinkage estimator, which gives a more robust estimate of the covariance matrix. The shrinkage strategy combines the sample correlation matrix with a more stable, structured estimator like the identity matrix or a target correlation matrix.

$$C_{\text{shrunk}} = (1-\alpha)\hat{C} + \alpha\frac{\text{Tr}\hat{C}}{M}\text{Id} \tag{15}$$

Where α is computed by the method proposed by Ledoit et al. (2004) [11] which minimises the mean squared error between the estimated and the real covariance matrix. We utilise C_{shrunk} for the following steps.

2) Clustering: For clustering, we employed the Affinity Propagation algorithm [12]. Unlike methods such as K-means, Affinity Propagation does not require specifying the number of clusters in advance. It identifies exemplars among data points and forms clusters through message passing between points. Using a correlation matrix, each entry represents a similarity measure, and Affinity Propagation treats these similarities as input to determine cluster centres or exemplars. The algorithm iteratively refines these exemplars based on maximising net similarity within clusters.

3) 2D embedding: To display these partial correlations on a 2D plot, we apply locally linear embedding (LLE) to the correlation matrix. LLe assumes that while the data may lie on a complex, non-linear manifold globally, each small neighbourhood around a data point can be approximated linearly. Within each neighbourhood, LLE performs a form of ”local principal component analysis(PCA)” by finding the weights that reconstruct each data point as a linear combination of its weights that reconstruct each data point as a linear combination of its neighbours. After calculating these local weights across all points, LLE aligns these local structures globally by finding a low-dimensional embedding where the local neighbourhoods and their weighted relationships are preserved as closely as possible. This approach results in a non-linear embedding that captures the manifold structure by ”stitching” together the local linear representations, essentially connecting the local PCAs into a cohesive lower-dimensional map.

This approach works better as PCA finds a global linear subspace that best explains the overall variance in the data, which may miss non-linear patterns if the data lies on a curved manifold. However, LLEh finds a global non-linear representation by focusing on preserving the neighbourhood relationships, making it more effective at capturing curved or complex structures.

4) Partial Correlations: Partial correlations measure the direct relationship between two variables while controlling for the influence of others. They are calculated using the precision matrix. A precision matrix, $\Theta = C^{-1}$, is the inverse of a covariance matrix and reveals conditional dependencies between variables. Unlike a correlation matrix, which shows pairwise relationships without accounting for other variables, a precision matrix highlights direct connections by removing indirect influences. That is, if $\Theta_{ij} = 0$, then the stocks i and j are conditionally independent given all other variables. This deeper insight into stock relationships helps identify key players within clusters and understand their roles in market dynamics.

The partial correlation between two variables i and j is given by:

$$\rho_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \tag{16}$$

For visualisation purposes, we consider where the partial correlations are above a certain threshold, in our case $|\rho_{ij}| > 0.14$.

5) Visualisation: To visualise, we use the embeddings as the 2D coordinates, the node colour represents the cluster group and the weights are the partial correlation values. To make it easier to visualise, we include the sectors of the stocks instead of the company names. The pre-COVID and post-COVID visualisations are in Fig. 18 and Fig. 19.

Both pre and post-COVID visualisations show triangular structures, with tech-related and energy-related industries forming two out of three vertices of the triangle. The third vertex however differs, the pre-COVID market has a strong focus on the real estate group, while it’s the consumer defensive and utility group in the post-COVID market. In the post-COVID market, The real estate group does have some presence in this third vertex, however, the overall cluster of real estate is spread out inside the triangle. Another key observation is that in the pre-COVID market have most of the industries are along the edges, with the Tech group being the central component that has links to all other industries. In the post-COVID market, we observe this structure is no longer there. The post-COVID market is more segregated.

In both pre-COVID while everything seemed interlinked, the post-COVID structure eluded interesting relationships between the companies. For example there is are strong partial correlation between the basic material industry giants such as BHP, Rio Tinto, and Freeport-McMoran Inc. which was not surprising. Moreover, an interesting link between the energy group and tech group was found which when looked into revealed a service provider and customer link between Microsoft and Petrobras respectively. Another similar strong link was found between the technology sector and healthcare (the cyan link in Fig. 19); with Cisco as the service provider to healthcare company Novo Nordisk.

These relationships were not apparent from the groups formed through eigenvector analysis in earlier sections. This underscores the importance of integrating multiple analytical techniques with those developed by Plerou et al. to achieve a comprehensive understanding of market dynamics. While PCA-based approaches, particularly

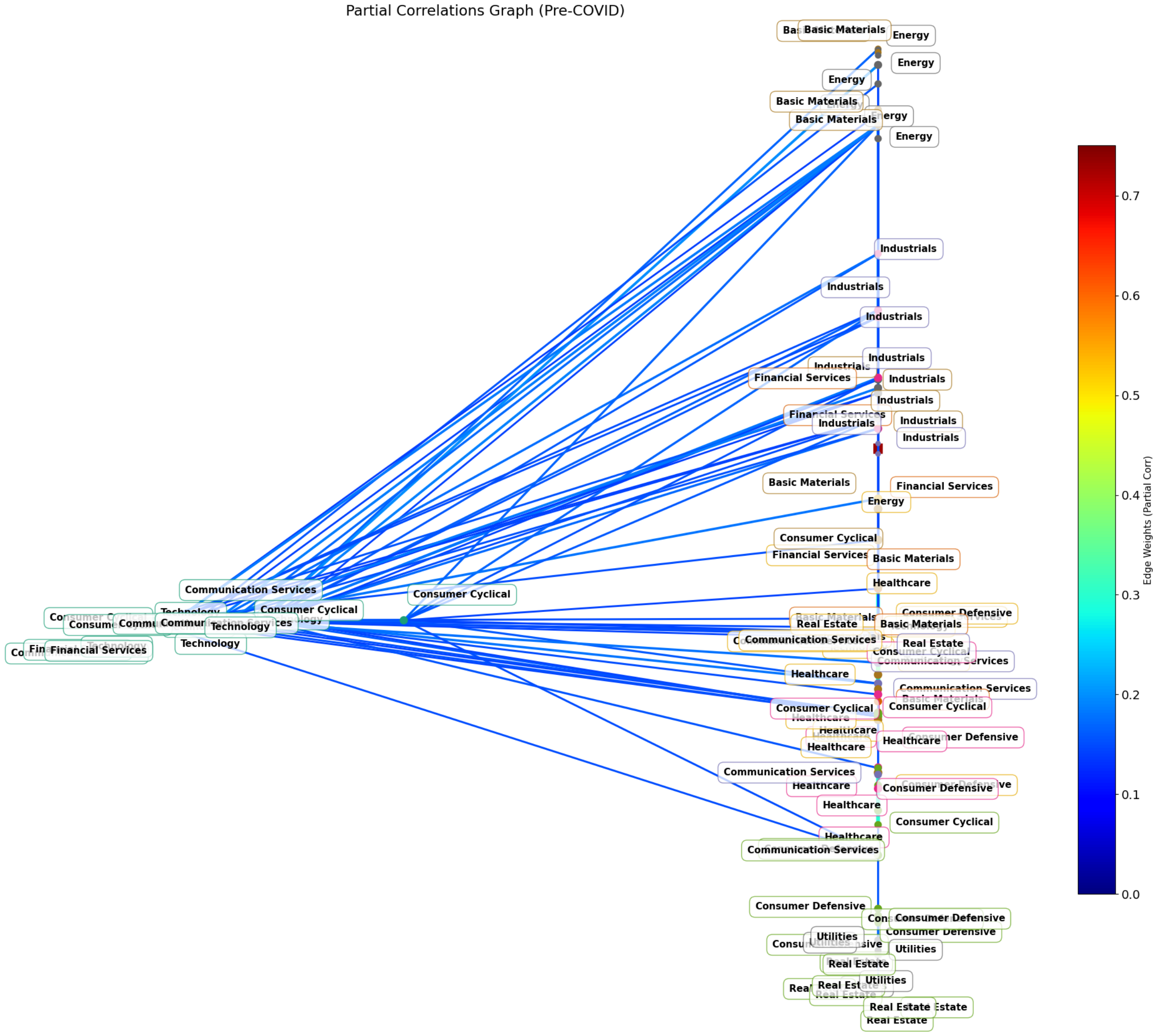


Fig. 18: Pre-COVID Partial Correlations

eigenvectors of the principal components, provide a broad view of overall relationships, partial correlations reveal smaller inter-dependencies that are crucial for a deeper insight. Both methods benefit from the application of RMT, which enhances the robustness of correlation matrix estimations. By leveraging these complementary techniques, we can better capture the complex interconnections within financial markets.

VI. Conclusion (Mukul and Luca)

This study has provided a comprehensive analysis of the U.S. stock market before and after the onset of the COVID-19 pandemic, employing random matrix theory and other statistical techniques. Our findings reveal a significant shift in market dynamics and correlation structures, offering valuable insights into the pandemic's impact on the financial markets.

By applying the techniques used by Plerou et al. in their 2002 study, we observed similar patterns in the stock market during a highly volatile period. The post-COVID period saw a notable increase in average correlations, consistent with previous studies on market behaviour during crises. The eigenvalue distribution analysis revealed that fewer principal components were required to explain as much variance as in the post-COVID period, indicating a consolidation of market drivers. This finding suggests that the pandemic may have simplified the market structure, with fewer factors dominating overall market movements. Our analysis of eigenvector components revealed significant shifts in sector importance and correlations:

- Increased prominence of utilities and essential services in the post-COVID period
- Greater emphasis on technology and financial services sectors
- Heightened importance of healthcare companies, particularly those involved in vaccine development
- Reduced dominance of the energy sector

These shifts reflect changing economic priorities and consumer behaviours in response to the pandemic.

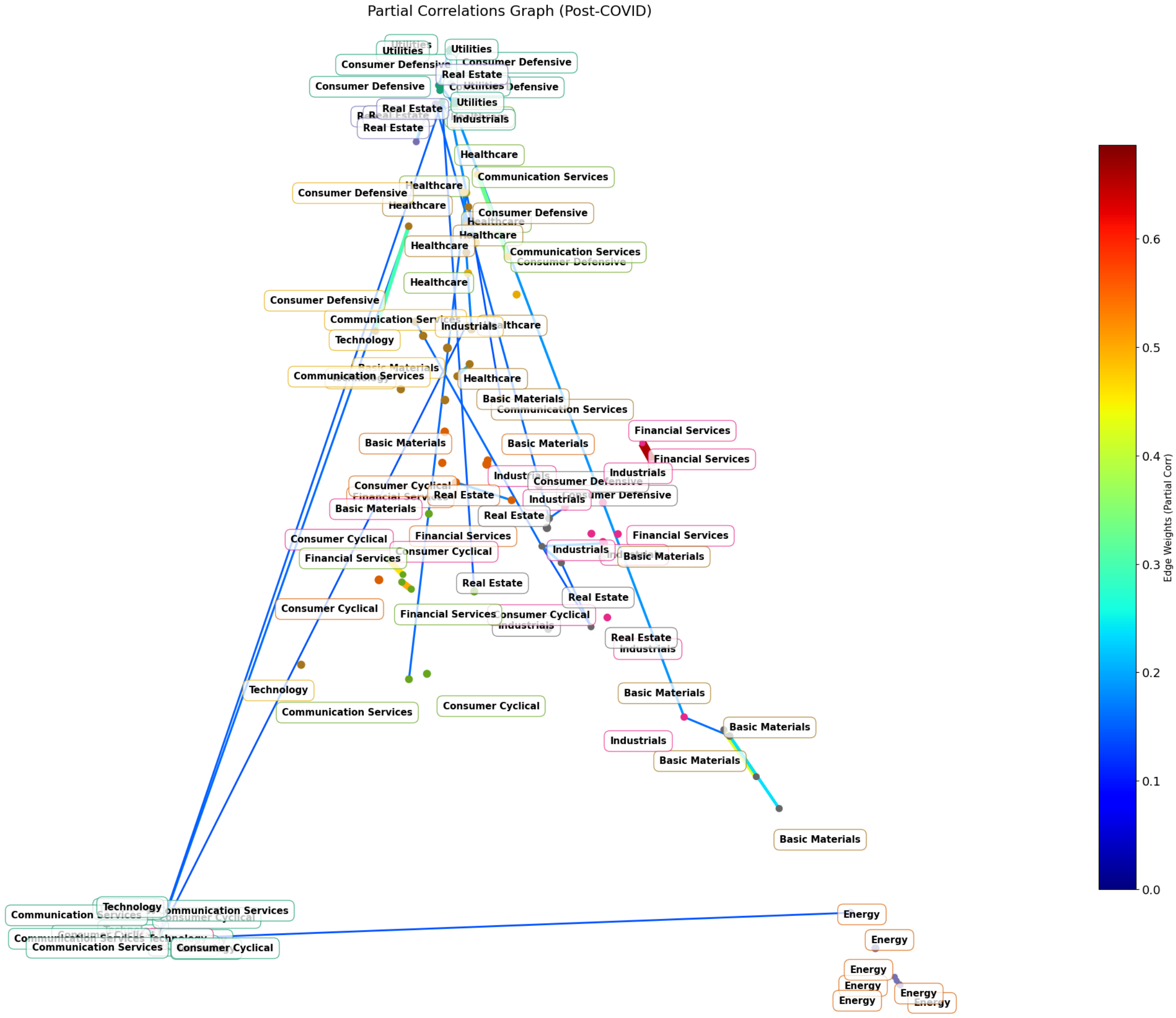


Fig. 19: Post-COVID Partial Correlations

The partial correlation analysis and visualization revealed a transition from a more interconnected pre-COVID market structure to a more segregated post-COVID structure. This change suggests that while overall correlations increased, the direct relationships between sectors became more distinct and isolated.

Our study successfully applied and extended the methods introduced by Plerou et al., effectively distinguishing ”true” correlations from statistical noise by comparing empirical eigenvalue distributions to theoretical predictions. Analyzing the bulk of the eigenvalue distribution allowed us to quantify the statistical noise in our correlation matrices. Using Gaussanity tests with inverse participation ratios, we also showed that in the finite M and N scenario, we need to consider both bounds of the bulks given by the limiting MP law and the fitted MP law to the data. Eigenvector analysis revealed persistent correlation patterns among stocks, often aligning with industry sectors and cross-sector relationships. Examining eigenvector stability over time offered valuable insights into the evolving nature of market dependencies during the pandemic. Our findings corroborate those of Plerou et al., showing that an increase in the time lag reduces the stability of eigenvectors over time while increasing their correlations. Post-COVID eigenvectors more stable in both their overlap matrices and ranking of their top 10 indexes than the pre-COVID eigenvectors. Furthermore, the ranking analysis yielded useful insights into index tracking for the leading eigenvectors, indicating how these indexes may consistently appear as prominent entries in eigenvectors within the bulk or on its left side.

The findings of this study have several important implications for investors, policymakers, and researchers. The increased market correlations and shifts in sector dynamics highlighted the need for investors to prefer more sophisticated diversification strategies, and traditional portfolio allocation methods may need to be reevaluated in light of these structural changes. The policymakers might need to implement targeted policy interventions that consider sector-specific impacts to mitigate the potential for rapid transmission of shocks across the market. For researchers, our study demonstrates the value of combining RMT with other advanced statistical techniques for a more comprehensive understanding of market dynamics.

While our study provides valuable insights, the specific time frame and geographic focus limit its scope. Future research could extend this analysis to compare market behaviors across different global markets, investigate longer-term impacts of the pandemic on market structures, compare with market mode descriptors like S&P 500, develop predictive models based on the observed structural changes, utilise more robust covariance estimators, and explore the application of these techniques to other types of financial crises or market shocks.

In conclusion, this study clarifies how global crises like the COVID-19 pandemic can fundamentally alter financial market structures. By leveraging advanced analytical techniques, we have uncovered subtle yet significant changes in market dynamics that have important implications for investment strategies, risk management, and financial policy. As markets continue to evolve in response to global events, the methodologies employed in this study offer a robust framework for ongoing analysis and adaptation to changing financial landscapes.

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] C. A. Tracy and H. Widom, "Level-spacing distributions and the Airy kernel," *Communications in Mathematical Physics*, vol. 159, no. 1, pp. 151–174, Jan. 1994. [Online]. Available: https://doi.org/10.1007/BF02100489

[5] C. M. Jarque and A. K. Bera, "A test for normality of observations and regression residuals," *International Statistical Review / Revue Internationale de Statistique*, vol. 55, no. 2, pp. 163–172, 1987. [Online]. Available: http://www.jstor.org/stable/1403192

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

[7] 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

[8] 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

[9] 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

[10] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951. [Online]. Available: http://www.jstor.org/stable/2280095

[11] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X03000964

[12] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1136800

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: Current version written by Mukul, Preliminary report section done by Luca.
4) Second Stage and Analysis: Luca and Mukul, Preliminary report section done by Mukul
5) Code: Techniques from Part 1 were applied by Luca using R and Mukul using Python. The source code with the readme file has been provided. Mukul wrote the source code for the Market Structure Analysis section, which is also included in the Python files.

Section by Section Split

Digitally Signed by:

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