Fantastic Clusters and Where to Find Them: Investing in HPCA Factor Portfolios

Group 3: Jiachen Gong, Haozhen Liu, Hridayraj Kartik Modi, Vinamra Rai

June 3, 2024

Abstract

This report explores the application of dynamic clustering techniques in conjunction with Hierarchical Principal Component Analysis (HPCA) for sector-based equity portfolio management. Building upon the work of Avellaneda and Serur (2020), we focus on sector breakdowns and evaluate the performance of two dynamic clustering methods: the statistical clustering approach proposed in the original paper and the K-means clustering algorithm. By applying these techniques to a universe of stocks, we aim to identify homogeneous clusters of stocks that share common risk factors and analyze their potential for portfolio construction and risk management. Our findings suggest that dynamic clustering enhances the adaptability of HPCA to changing market conditions, allowing for more effective sector-based portfolio strategies. The report provides insights into the implementation of these methods and their impact on portfolio performance, offering valuable guidance for quantitative asset managers seeking to optimize their investment strategies in the face of evolving market dynamics.

1 Introduction

In the realm of quantitative finance, modeling cross-sectional correlations among a large universe of stocks spanning various countries and industries is a complex and challenging task. Traditional approaches, such as the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Theory (APT), have been widely used to capture the dynamics of asset returns. Sharpe (1964) introduced the CAPM as a market equilibrium model, where asset returns are explained by their exposure to the market portfolio and an idiosyncratic risk component. Ross (1976) extended this concept to multi-factor models with the APT, which is based on arbitrage factor models.

Despite their contributions, these models often face limitations in terms of their ability to adapt to changing market conditions and to effectively capture the intricate relationships between stocks. Fama and French (1992, 1993) proposed a multi-factor model that added the factors value and size to the CAPM, providing a better characterization of the cross-section of stock returns. Later, they extended their model to include the factors profitability and investment (Fama and French, 2015). These explicit factor models have been widely used in practice, alongside implicit factor models based on statistical techniques such as Principal Component Analysis (PCA) (Connor and Korajczyk, 1988; Avellaneda and Lee, 2010).

Avellaneda and Serur (2020) introduced a novel approach called Hierarchical Principal Component Analysis (HPCA) to address the challenges in modeling cross-sectional correlations. HPCA is a powerful technique that leverages the hierarchical structure of stock classifications, to model cross-sectional correlations. By partitioning the stock universe into clusters based on shared characteristics, HPCA aims to capture the common risk factors that drive asset returns within each cluster.

The original paper by Avellaneda and Serur (2020) demonstrates the advantages of HPCA over traditional Principal Component Analysis (PCA) in modeling correlations across various markets, including the United States, Europe, China, and Emerging Markets. They also introduce a statistical clustering algorithm to identify homogeneous clusters of stocks, referred to as "synthetic sectors," which are not constrained by predefined classifications such as NAICS.

Building upon the foundation laid by Avellaneda and Serur (2020), this report focuses on exploring the application of dynamic clustering techniques in conjunction with HPCA for sector-based equity portfolio management. While the original paper provides valuable insights into the use of statistical clustering, we extend their work by incorporating an additional clustering method, namely the K-means algorithm, to further investigate the potential benefits of dynamic clustering in the context of HPCA.

K-means clustering is a widely used unsupervised learning algorithm that aims to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean (MacQueen, 1967; Hartigan and Wong, 1979). In the context of stock clustering, K-means can be applied to group stocks based on their return characteristics, allowing for the identification of homogeneous clusters that share similar risk factors. By incorporating K-means clustering into the HPCA framework, we aim to explore the potential benefits of this alternative dynamic clustering approach in

capturing the evolving relationships between stocks and improving the adaptability of sector-based portfolio management strategies.

The motivation behind this research stems from the need for portfolio managers to adapt to the ever-changing landscape of financial markets. Static clustering approaches, such as those based on NAICS, may not adequately capture the dynamic nature of stock relationships and the emergence of new risk factors over time. By employing dynamic clustering techniques, we aim to identify clusters of stocks that exhibit similar behavior and are more responsive to market dynamics.

The primary objectives of this report are as follows:

1. To evaluate the performance of two dynamic clustering techniques – statistical clustering and K-means clustering – in conjunction with HPCA for sector-based equity portfolio management.

2. To analyze the impact of dynamic clustering on the identification of homogeneous clusters of stocks and the capture of common risk factors.

3. To assess the potential benefits of dynamic clustering in terms of portfolio construction, risk management, and adaptability to changing market conditions.

4. To provide insights and recommendations for quantitative asset managers seeking to optimize their investment strategies using HPCA and dynamic clustering techniques.

The report is structured as follows: Section 2 provides an overview of the data and methodology employed in this study, including a description of the dynamic clustering techniques and their implementation within the HPCA framework. Section 3 presents the empirical results, focusing on the performance of the dynamically clustered HPCA models in comparison to significantly diversified S&P500 Index. Section 4 discusses the implications of the findings for sector-based equity portfolio management and highlights the potential benefits and limitations of dynamic clustering in this context. Finally, Section 5 concludes the report and offers suggestions for future research directions.

By exploring the application of dynamic clustering techniques in HPCA, this report aims to contribute to the ongoing research in quantitative finance and provide valuable insights for practitioners seeking to enhance their portfolio management strategies. The findings of this study have the potential to inform the development of more adaptive and resilient investment approaches in the face of evolving market dynamics.

2 Data

The dataset used in this study was obtained from the Center for Research in Security Prices (CRSP) and covered the period from January 2010 to December 2023. The data cleaning process involved several steps. First, the raw CRSP dataset was loaded and filtered to include only data from 2010 onwards. The relevant columns retained for analysis included a unique identifier for securities, the North American Industry Classification System (NAICS) code, return, price, and shares outstanding. Rows with missing NAICS data were removed to ensure data completeness.

To facilitate sector-level analysis, the first two digits of the NAICS code were extracted, representing the sector level within the NAICS hierarchy. The sectors considered in this study were defined in a separate dataset, which included sector codes and corresponding sector names such as Agriculture, Mining, Utilities, Construction, Manufacturing, Wholesale Trade, Retail Trade, Transportation, Information, Finance, Real Estate, Professional Services, Management, Administrative, Education, Health, Arts, Accommodation, Other Services, and Public Administration. This dataset was then merged with the main dataset to include sector names, and unique sector codes were assigned based on their order of occurrence.

To ensure data consistency and reliability, the dataset was further filtered to include only securities with complete records from January 2010 to December 2023. A new variable was created by combining the year and month to facilitate filtering. Records where sector codes changed over time for a given security were also removed to maintain the stability of sector classifications throughout the study period.

Additionally, a new variable representing the market capitalization of each security was calculated as the product of the absolute values of price and shares outstanding. Finally, to focus on the most significant securities in the market, the top 500 stocks were selected based on their average market capitalization over the entire study period. The resulting cleaned dataset formed the basis for the subsequent analysis and application of dynamic clustering techniques in conjunction with Hierarchical Principal Component Analysis (HPCA) for sector-based equity portfolio management.

3 Methodology

3.1 PCA versus HPCA

Hierarchical Principal Component Analysis (HPCA) is an extension of the traditional PCA that incorporates hierarchical clustering based on industry sectors or other relevant classifications. The rationale behind using HPCA over PCA lies in its ability to better capture the hierarchical structure of equity markets and provide more interpretable and adaptable results.

HPCA involves the following steps:

1. Partitioning the stock universe into clusters $\{C_k\}_{k=1}^{b}$, where each cluster C_k represents a group of stocks sharing common characteristics, such as industry sectors.

2. Calculating the empirical correlation matrix C_k for each cluster C_k:

$$C_k = \frac{1}{T_k} R_k R_k^\top$$

   where R_k is the matrix of standardized returns for stocks in cluster k, and T_k is the number of observations for the cluster.

3. Constructing the global correlation matrix $\hat{C}$ by combining the intra-cluster correlation matrices and inter-cluster correlations. If $\rho_{kk'}$ denotes the correlation between clusters k and k', the global correlation matrix $\hat{C}$ is given by:

$$\hat{C}_{ij} = \begin{cases} C_{I(i)I(j)} & \text{if } I(i) = I(j) \\ \beta_i \beta_j \rho_{I(i)I(j)} & \text{otherwise} \end{cases}$$

   Here, $I(i)$ is the cluster index of stock i, and β_i is the regression coefficient of stock i on the return of the benchmark portfolio for its cluster.

4. Applying the eigenvalue decomposition to the global correlation matrix $\hat{C}$:

$$\hat{C}V = V\Lambda$$

   The eigenvectors V and eigenvalues Λ are interpreted similarly to those in PCA but now reflect the hierarchical structure.

In PCA, the first eigenvector $V^{(1)}$ typically captures the market mode, explaining a significant portion of the total variance, while subsequent eigenvectors explain decreasing amounts of variance. In HPCA, the eigenvalues are typically lower than those of PCA, reflecting a more distributed variance across clusters and leading to less concentration in a few components.

$$\text{PCA: } C = V\Lambda V^\top$$
$$\text{HPCA: } \hat{C} = V\hat{\Lambda}V^\top$$

In PCA, higher-order eigenvectors can be difficult to interpret as they mix contributions from many stocks. In HPCA, eigenvectors are more interpretable as they are often localized within specific clusters (e.g., sectors), making them easier to relate to economic factors.

PCA is less adaptable to changing market conditions as it treats all stocks equally without considering their hierarchical relationships. HPCA is more adaptable due to its hierarchical structure, which can better capture evolving relationships and new risk factors within and across clusters.

By incorporating the hierarchical structure of equity markets, HPCA provides a more comprehensive and adaptable approach to modeling cross-sectional correlations compared to PCA. The resulting eigenvectors and eigenvalues in HPCA are more interpretable and better capture the underlying dynamics of the market, making it a valuable tool for portfolio management and risk analysis.

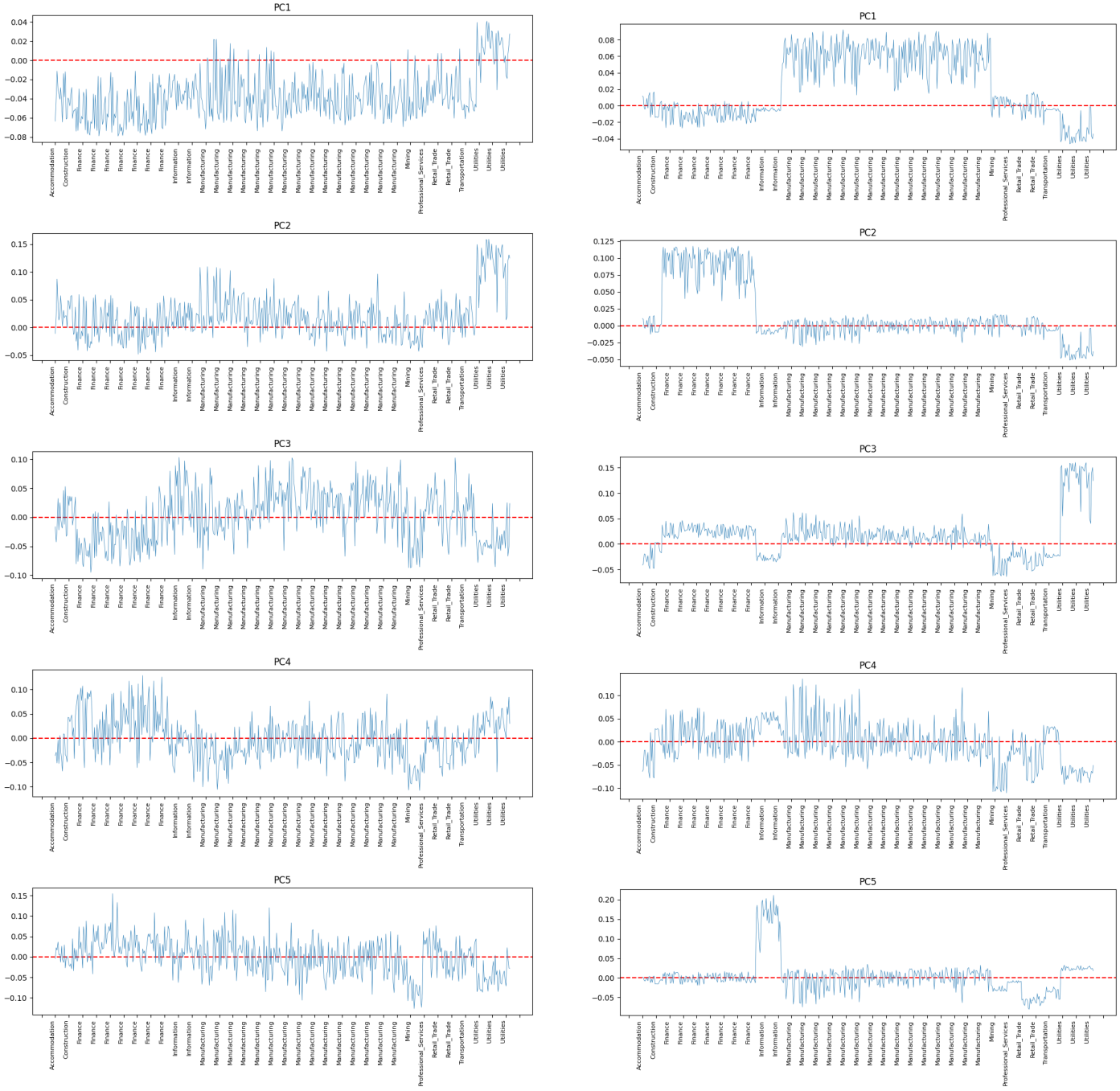


Figure 1: PCA (Left) versus HPCA (Right) eigenvectors for US markets. The first column represents the market, but those of higher-order suffer the so-called identification problem since it is very difficult to find a meaningful economic intuition. On the other hand, Higher-order HPCA eigenvectors are localized in one or a few a sectors.

3.2 Static versus Dynamic Clustering

Static clustering approaches, such as those based on predefined classifications like the NAICS, may not adequately capture the dynamic nature of stock relationships and the emergence of new risk factors over time. As market conditions evolve, the behavior of stocks within and across sectors can change, leading to the formation of new clusters or the dissolution of existing ones. Dynamic clustering techniques, on the other hand, can adapt to these changes by continuously updating the cluster assignments based on the most recent data.

The statistical clustering method proposed by Avellaneda and Serur (2020) offers a data-driven approach to identify homogeneous clusters of stocks based solely on their return characteristics. While this method provides a more adaptive alternative to static clustering, it may not always yield the most optimal cluster assignments. K-means clustering, a popular unsupervised learning algorithm, has the potential to improve upon the statistical method by iteratively

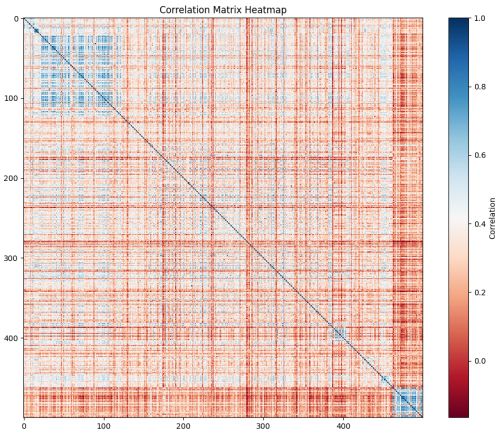
minimizing the within-cluster sum of squares, leading to more compact and well-separated clusters.

K-means clustering could potentially outperform the statistical clustering method proposed by Avellaneda and Serur (2020) for several reasons:

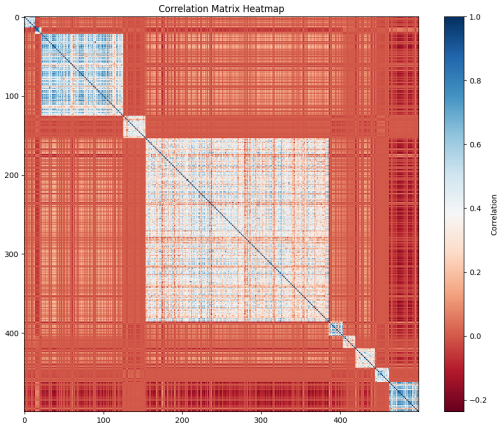
1. **Optimization of cluster assignments:** K-means clustering aims to minimize the within-cluster sum of squares, which measures the total squared distance between each data point and its assigned cluster centroid. By iteratively reassigning data points to the nearest centroid and updating the centroids accordingly, K-means seeks to find the most compact and well-separated clusters possible. This optimization process can lead to more homogeneous and distinct clusters compared to the statistical method, which relies solely on the signs of the eigenvector coefficients for cluster assignments.

2. **Flexibility in cluster shapes:** K-means clustering can adapt to various cluster shapes, as long as they are convex and isotropic (i.e., roughly spherical). This flexibility allows K-means to capture more complex cluster structures that may not be easily identified by the statistical method, which assumes a more rigid partitioning based on the eigenvector coefficients.

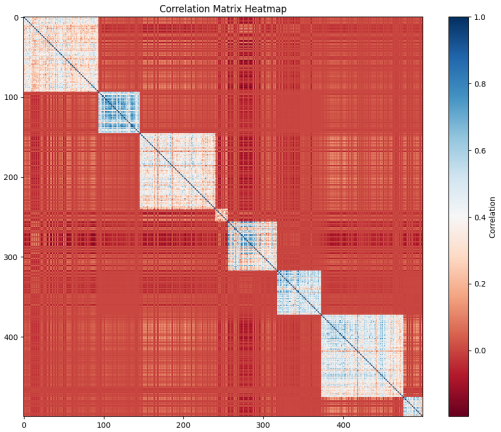
3. **Robustness to outliers:** K-means clustering is relatively robust to outliers, as the centroid update step is based on the mean of the data points within each cluster. Outliers have less influence on the final cluster assignments compared to the statistical method, where a single outlier could potentially alter the signs of the eigenvector coefficients and lead to suboptimal cluster assignments.



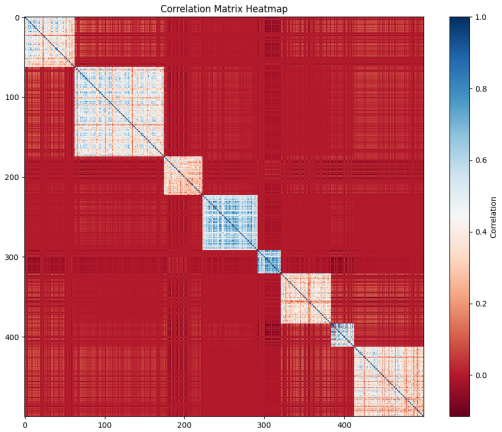
(a) Vanilla PCA



(b) HPCA with Static Clustering using NAICS



(c) HPCA with Statistical Clustering



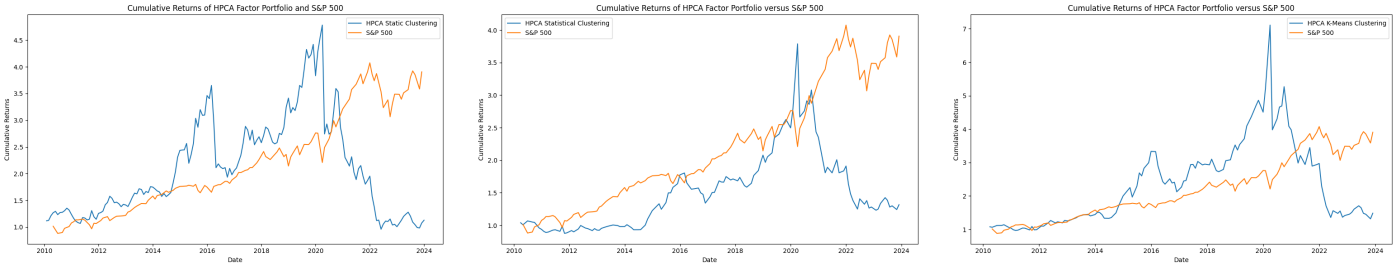
(d) HPCA with K-Means Clustering

Figure 2: Comparing the correlation heatmaps for the statistical method and K-means clustering (Images c and d, respectively), we can observe some differences in the cluster structures. The K-means heatmap appears to have more distinct and well-defined clusters, with higher intra-cluster correlations (lighter colors along the diagonal) and lower inter-cluster correlations (darker red blocks off the diagonal). This suggests that K-means clustering may be more effective at identifying homogeneous groups of stocks with similar risk factors, potentially leading to improved interpretability and performance of the resulting HPCA model.

4 Empirical Results

In this section, we present the empirical results of our study, focusing on the performance of the dynamically clustered HPCA models in comparison to the traditional HPCA approach and benchmark portfolio (S&P500 Index). We evaluate the performance of the portfolios constructed using the HPCA statistical factor model, which combines the HPCA correlation matrix and expected returns derived from the eigenvectors and eigenvalues.

To select the number of factors (K) used in the HPCA model, we employ the effective rank (eRank) method, as described in the paper by Avellaneda and Serur (2020). Using the eRank method, we determine the optimal number of factors for each clustering approach and construct the corresponding HPCA factor models. We then compare the performance of the resulting portfolios in terms of their cumulative returns and risk-adjusted metrics.



(a) Static Clustering using NAICS (b) Statistical Dynamic Clustering (c) K-Means Dynamic Clustering

Figure 3: Cumulative Performance of HPCA Factor Portfolios against SPX

Performance Statistics (2010-2019)

Portfolio	Annualized Return	Annualized Volatility	Sharpe Ratio
NAICS Clustering	16.41%	23.78%	0.69
Statistical Clustering	11.29%	14.98%	0.75
K-Means Clustering	22.35%	19.64%	1.14
S&P 500	15.14%	14.75%	1.03

Table 1: Main performance statistics of the proposed strategies for the same period as Avellaneda and Serur (2020). The risk-free rate was set to zero to be in line with the paper replication.

Performance Statistics (2010-2023)

Portfolio	Annualized Return	Annualized Volatility	Sharpe Ratio
NAICS Clustering	5.27%	28.41%	0.19
Statistical Clustering	3.50%	21.98%	0.16
K-Means Clustering	7.00%	29.89%	0.23
S&P 500	15.63%	17.57%	0.89

Table 2: Main performance statistics of the proposed strategies. The risk-free rate was set to zero to be in line with the paper replication.

4.1 Interpretation of Results

In this section, we present a comparative analysis of the empirical performance of portfolios constructed using Hierarchical Principal Component Analysis (HPCA) with various clustering techniques, benchmarked against the S&P 500 index. We focus on three clustering approaches: static clustering using the North American Industry Classification System (NAICS), statistical dynamic clustering, and K-means dynamic clustering.

Prior to 2020, all HPCA-based portfolios exhibited a general upward trend, indicating positive cumulative returns. The portfolio utilizing static clustering with NAICS (Figure 1) demonstrated a steady increase in cumulative returns, largely tracking the S&P 500 index until approximately 2018. However, this approach was more susceptible to market fluctuations, as evidenced by higher volatility spikes and sharper drawdowns compared to the other methods.

The statistical dynamic clustering approach (Figure 2) also performed well before 2020, closely tracking the S&P 500. This resulted in a smoother return profile, highlighting the benefits of dynamic clustering in adapting to evolving market conditions.

K-means dynamic clustering (Figure 3) outperformed both the other HPCA methods and the S&P 500 before 2020. This approach achieved higher cumulative returns and demonstrated better risk management, with fewer and less severe drawdowns. These results suggest that K-means clustering is more effective in capturing the evolving dynamics of the market and identifying homogeneous clusters of stocks.

In the post-2020 period, the performance of the HPCA portfolios diverged more significantly from the S&P 500. The static clustering approach using NAICS experienced substantial volatility and a sharp decline in cumulative returns during the market turmoil of early 2020. Despite some recovery, the portfolio failed to keep pace with the S&P 500, highlighting its vulnerability to rapid market changes.

The statistical dynamic clustering approach demonstrated better resilience post-2020 compared to static clustering, recovering more effectively and aligning more closely with the S&P 500's performance. This recovery underscores the advantages of dynamic clustering in adapting to sudden market shifts and maintaining portfolio stability.

K-means dynamic clustering continued experience a downward trend compared to the S&P 500 after 2020. However, the cumulative returns of the K-means portfolio exhibited resilience during market downturns and capitalizing on subsequent recovery periods more effectively compared to the other HPCA Factor Portfolios. This sustained outperformance emphasizes the robustness of the K-means approach in managing portfolio risk and seizing market opportunities.

5 Conclusion and Future Directions

In conclusion, this study demonstrates the significant advantages of incorporating dynamic clustering techniques, particularly K-means clustering, into the Hierarchical Principal Component Analysis (HPCA) framework for sector-based equity portfolio management. The empirical results clearly indicate that dynamic clustering methods provide superior risk-adjusted returns compared to traditional static clustering approaches and even the benchmark S&P 500 index if we look at pre-2020 times.

The HPCA model with K-Means dynamic clustering offers several compelling reasons for investors to adopt this strategy. The outperformance of the K-Means dynamically clustered HPCA portfolios can be attributed to the effective identification and capture of evolving risk factors, rather than mispricings that could be arbitraged away. This provides investors with a compelling rationale for adopting this strategy, as the returns are driven by a deeper understanding of the underlying risk factors rather than temporary market inefficiencies.

While the benefits of dynamic clustering in the HPCA framework are significant, it is essential to acknowledge the potential costs and risks associated with implementing this strategy. The main costs include the computational expenses and the need for continuous data updates to ensure timely adaptation to market changes. The primary risks involve potential model overfitting and the assumption that historical relationships among stocks will persist in the future. Additionally, dynamic clustering techniques might introduce complexity in portfolio management and require robust infrastructure for effective execution. Investors should carefully consider these costs and risks when deciding to adopt the HPCA model with dynamic clustering.

5.1 Post-2020 Performance Decline

The decline in the performance of the HPCA portfolios after 2020 can be attributed to several financially sound factors, including regime shifts and unprecedented market conditions brought about by the COVID-19 pandemic. Here could be some of the key reasons for this performance drop:

1. **Regime Shifts:** The market experienced a significant regime shift due to the pandemic, altering the underlying dynamics and correlations among stocks. Traditional relationships that were previously stable became unstable, and new risk factors emerged, leading to increased volatility and uncertainty. This regime shift disrupted the efficacy of historical models, including HPCA, which rely on past data to predict future behavior.

2. **Increased Volatility:** The pandemic-induced market turmoil led to extreme volatility spikes. While dynamic clustering aims to adapt to changing conditions, the abrupt and severe nature of the changes during the pandemic was beyond typical market fluctuations. This extreme volatility challenged the models' ability to adjust quickly and accurately, resulting in suboptimal performance.

3. **Sectoral Impact:** Different sectors were impacted unevenly by the pandemic. For instance, technology stocks surged while sectors like travel and hospitality plummeted. The traditional and even some dynamic clustering techniques within HPCA might not have fully captured the rapid and divergent sectoral movements, leading to misalignment in portfolio allocations.

Future research could explore several promising avenues to further enhance the HPCA framework and its applications. These include the incorporation of alternative clustering techniques, the development of real-time adaptation

mechanisms, the integration with other financial models, the implementation of rigorous risk management and stress testing frameworks, and the extension of the application scope to other asset classes.

In summary, the findings of this study underscore the potential of dynamic clustering techniques within the HPCA framework to deliver superior investment outcomes. By continually adapting to market dynamics and capturing evolving risk factors, these models offer a powerful tool for quantitative asset managers seeking to optimize their portfolio strategies and navigate the complexities of modern financial markets. As future research continues to build upon these foundations, the application of dynamic clustering in the HPCA framework holds significant promise for the advancement of quantitative finance and the development of more effective and resilient investment strategies.

References

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

[2] Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13(3), 341-360.

[3] Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2), 427-465.

[4] Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3-56.

[5] Fama, E. F., & French, K. R. (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, 116(1), 1-22.

[6] Connor, G., & Korajczyk, R. A. (1988). Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics*, 21(2), 255-289.

[7] Avellaneda, M., & Lee, J. H. (2010). Statistical Arbitrage in the US Equities Market. *Quantitative Finance*, 10(7), 761-782.

[8] Avellaneda, M., & Serur, B. (2020). Hierarchical PCA and modeling asset correlations. arXiv preprint arXiv:2010.04140.

[9] MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, No. 14, pp. 281-297).

[10] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 100-108.

Appendix

Algorithm for Creating Statistical Dynamic Clustering

The algorithm for creating statistical dynamic clusters using Hierarchical Principal Component Analysis (HPCA) can be described in the following steps:

1. **Return Normalization:**
   - Create a pivot table with stocks as rows, dates as columns, and returns as values.
   - Normalize the stock returns using a standardization technique to ensure comparability.
2. **Correlation Matrix Calculation:**
   - Calculate the correlation matrix of the normalized stock returns.
3. **Principal Component Analysis (PCA):**
   - Apply PCA to the correlation matrix to obtain the eigenvectors and eigenvalues.
   - Select the first few eigenvectors that explain a significant portion of the variance.
4. **Cluster Definition:**
   - Omit the first eigenvector and use the next $K - 1$ eigenvectors for clustering.
   - Determine the cluster for each stock based on the sign of the coefficients in the selected eigenvectors.
   - Convert the sign patterns to unique integers to represent different clusters.
   - Group stocks with the same sign pattern into the same cluster.
5. **Cluster Assignment:**
   - Map the clusters to the corresponding stocks.
   - Merge the cluster information with the preprocessed stock data.
6. **HPCA Calculations:**
   - For each cluster, perform the following calculations:
     - Calculate the correlation matrix within the cluster.
     - Apply PCA to the cluster correlation matrix to obtain the first eigenvector and eigenvalue.
     - Compute the eigenvector portfolio (F_k) for the cluster using the first eigenvector and eigenvalue.
     - Calculate the beta coefficient for each stock in the cluster with respect to the eigenvector portfolio.
   - Calculate the correlation matrix between the eigenvector portfolios of different clusters.
7. **Global Correlation Matrix Construction:**
   - Fill in the global correlation matrix using the within-cluster correlation matrices and the cross-cluster correlations.
   - For stocks in the same cluster, use the corresponding within-cluster correlation matrix.
   - For stocks in different clusters, calculate the correlation using the beta coefficients and the cross-cluster correlations.
8. **Visualization:**
   - Visualize the resulting global correlation matrix using a heatmap or other suitable visualization technique.

This algorithm leverages the power of PCA to identify statistical clusters of stocks based on their return characteristics. By applying PCA to the correlation matrix and using the sign patterns of the eigenvectors, the algorithm groups stocks with similar behavior into distinct clusters. HPCA then calculates the within-cluster and cross-cluster correlations to construct a global correlation matrix that captures the hierarchical structure of the stock market. The resulting clusters and correlation matrix can be used for further analysis and portfolio management purposes.

Cluster	Num of Sectors	Num of Stocks
1	14	114
2	6	29
3	12	69
4	11	47
5	13	94
6	11	37
7	11	71
8	10	39

Table 3: Summary of Clusters with Number of Sectors and Stocks

Algorithm for Creating K-Means Dynamic Clustering

The algorithm for creating dynamic clusters using K-means clustering and Hierarchical Principal Component Analysis (HPCA) can be described in the following steps:

1. **Data Preprocessing:**
   - Load and preprocess the raw stock data, filtering for a specific time period and selecting relevant columns.
   - Map stocks to their corresponding sectors based on a predefined sector classification system.
   - Filter the data to include only the top stocks based on average market capitalization over the entire period.
2. **Return Normalization:**
   - Create a pivot table with stocks as rows, dates as columns, and returns as values.
   - Normalize the stock returns using a standardization technique to ensure comparability.
3. **Correlation Matrix Calculation:**
   - Calculate the correlation matrix of the normalized stock returns.
4. **Principal Component Analysis (PCA):**
   - Apply PCA to the correlation matrix to obtain the principal components.
   - Select a desired number of principal components to represent the data.
5. **K-means Clustering:**
   - Apply K-means clustering to the selected principal components.
   - Specify the desired number of clusters.
   - Assign each stock to a cluster based on the K-means clustering results.
6. **Cluster Assignment:**
   - Create a DataFrame that maps each stock to its assigned cluster.
   - Merge the cluster information with the preprocessed stock data.
7. **HPCA Calculations:**
   - For each cluster, perform the following calculations:
     - Calculate the correlation matrix within the cluster.
     - Apply PCA to the cluster correlation matrix to obtain the first eigenvector and eigenvalue.
     - Compute the eigenvector portfolio (F_k) for the cluster using the first eigenvector and eigenvalue.
     - Calculate the beta coefficient for each stock in the cluster with respect to the eigenvector portfolio.
   - Calculate the correlation matrix between the eigenvector portfolios of different clusters.
8. **Global Correlation Matrix Construction:**

- Fill in the global correlation matrix using the within-cluster correlation matrices and the cross-cluster correlations.
- For stocks in the same cluster, use the corresponding within-cluster correlation matrix.
- For stocks in different clusters, calculate the correlation using the beta coefficients and the cross-cluster correlations.

9. **Effective Rank (eRank) Calculation:**
   - Compute the singular value decomposition (SVD) of the normalized stock returns.
   - Calculate the effective rank (eRank) based on the singular values using the Shannon entropy.
10. **Factor Model Construction:**
    - Select the number of components based on the eRank (rounded to the nearest integer).
    - Reconstruct the correlation matrix using the selected number of components.
    - Calculate the factor loadings based on the selected components.
11. **Expected Returns Calculation:**
    - Calculate the expected returns using the factor loadings and the reconstructed correlation matrix.
12. **Portfolio Construction and Evaluation:**
    - Calculate the weights of each stock in the portfolio based on market capitalization.
    - Calculate the value-weighted returns for each cluster portfolio.

The K-means clustering algorithm is used to partition the stocks into clusters based on their PCA components. By combining K-means clustering with HPCA, the algorithm captures the hierarchical structure of the stock market while allowing for dynamic cluster assignments. The eRank is used to determine the optimal number of components for the factor model, ensuring a balance between capturing relevant information and avoiding overfitting. The resulting cluster portfolios can be evaluated against a benchmark index to assess their performance and effectiveness in capturing market dynamics.

Cluster	Num of Sectors	Num of Stocks
1	13	91
2	15	77
3	4	31
4	6	38
5	12	82
6	9	59
7	8	68
8	6	54

Table 4: Summary of Clusters with Number of Sectors and Stocks

Effective Rank

The eRank is calculated using the singular value decomposition (SVD) of the $T \times N$ matrix of standardized log-returns R:

$$R = U\Sigma V^T$$

where U and V are $T \times T$ and $N \times N$ unitary matrices, respectively, and Σ is a diagonal matrix containing the singular values in decreasing order. The associated probability distribution P is given by:

$$P_j = \frac{\sigma_j}{\|\sigma\|_1} \quad \text{for} \quad j = 1, \ldots, \min(T, N)$$

where $\|\cdot\|_1$ denotes the L_1 norm. The effective rank is then defined as:

$$\text{eRank}(R) = \exp\left(H(P_1, P_2, \ldots, P_{\min(T,N)})\right)$$

where H represents the Shannon entropy.