Fantastic Clusters and Where to Find Them:

Investing in HPCA Factor Portfolios

UCLA Anderson School of Management

Introduction

Background

CAPM and Arbitrage Pricing Theory(APT) have been foundational, but face **limitations** in the ability to adapt to changing market conditions and capture the intricate relationships between stocks.

Fama and French enhanced these with multi-factor models including size, value, profitability, and investment factors.

Hierarchical PCA

Avellaneda and Serur (2020) introduced Hierarchical Principal Component Analysis (HPCA) to better model cross-sectional correlations by utilizing stock hierarchical structures.

Clustering

Our project extends statistical clustering from the paper by incorporating **K-means clustering** with HPCA, aiming to improve sector-based equity portfolio management by identifying homogeneous clusters of stocks with similar risk factors.

Overview

Dataset

Basic of PCA and HPCA Algorithm

Static vs Dynamic Clustering

Statistical, NAICS, K-means Clustering

Trading Strategy/Factor Models

Interpretation of Results/Recommendations

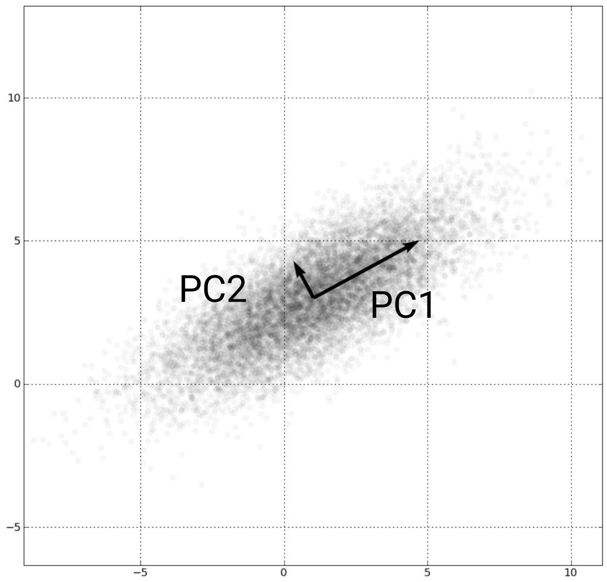
Conclusion

UCLA Anderson School of Management

Dataset

Step 1	Load Data	Load Data from CRSP
Step 2	Data Cleaning	Columns: Unique Identifiers, NAICS codes, return, price, shares; Remove incomplete record
Step 3	Sector Prep	Extract sector-level codes from NAICS, Merge back to the main dataset, Assign sector codes
Step 4	Data Filtering	Include data from Jan 2010 to Dec 2023, Create combined year-month variable, Remove the records with inconsistent sector
Step 5	Calculation	Calculate market capitalization by multiplying price and shares
Step 6	Data Selection and Output	Select top 500 stocks by the avg mkt cap, Output the clean data

Introduction to PCA



PCA Analysis of a 2D normal Distribution

Key Idea: Find orthogonal vectors that best represent the variance

When performing PCA, the first principal component of a set of variables is the derived variable formed as a linear combination of the original variables that explains **the most variance**. The second principal component explains the most variance in what is left once the effect of the first component is removed, and we may proceed through iterations until all the variance is explained.

The principal components are eigenvectors of the data covariance matrix.

Applying PCA

Data

T x N Matrix

Assume that there are N assets over T periods, their returns can fit into a T x N Matrix

Correlation Matrix

N x N Matrix

For each asset, using the records for the past T periods, we can come up with a N x N correlation matrix.

Decomposition

U x Σ x V’

Applying Singular value decomposition to our correlation matrix, we can get U and V’ which are rectangular matrix (N x N) and a rectangular diagonal matrix λ (N x N).

Result

N x 1, λ

Each λ in matrix Σ is a eigenvalue and the corresponding column vector of matrix U is the eigenvector.

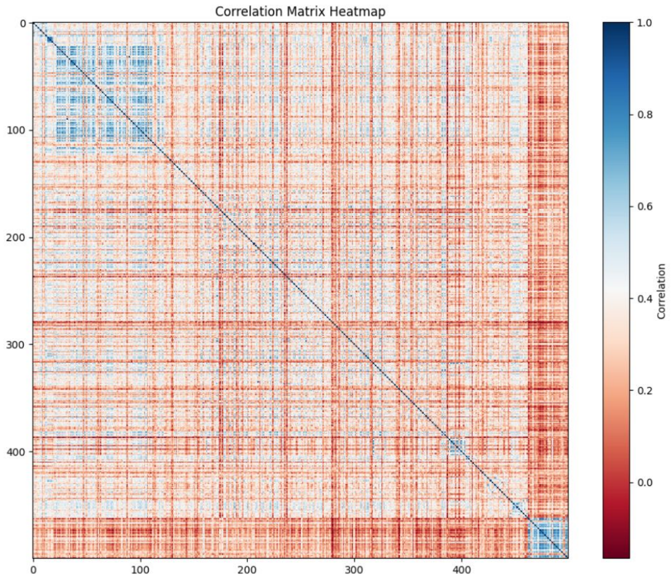
$$\underset{m\times n}{M} = \underset{m\times m}{U}\ \underset{m\times n}{S}\ \underset{n\times n}{V^T}$$

Orthogonal (“Rotate”)

Diagonal (“Stretch”)

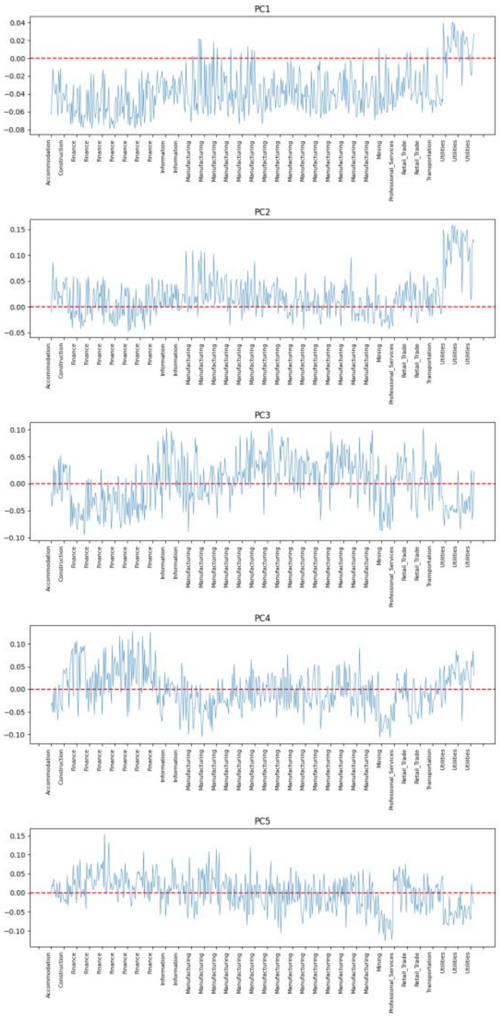
Orthogonal (“Rotate”)

PCA Results



Upon examination of the PCA methodology, it becomes evident that the first principal component (PC1) accounts for the majority of the variance observed within the market data. Consequently, **PC1 can be construed as a representation of the market portfolio.**

Nevertheless, subsequent principal components lack an explicit economic interpretation, which raises the question of **whether alternative methodologies exist to render the PCA results intelligible beyond PC1.**



Hierarchical PCA (HPCA)

Define Clusters

The example uses NAICS industry code to divide stocks into different sectors.

Correlation Matrix Within Clusters

For stocks within a cluster, calculated their correlation matrix according to the procedure as normal PCA

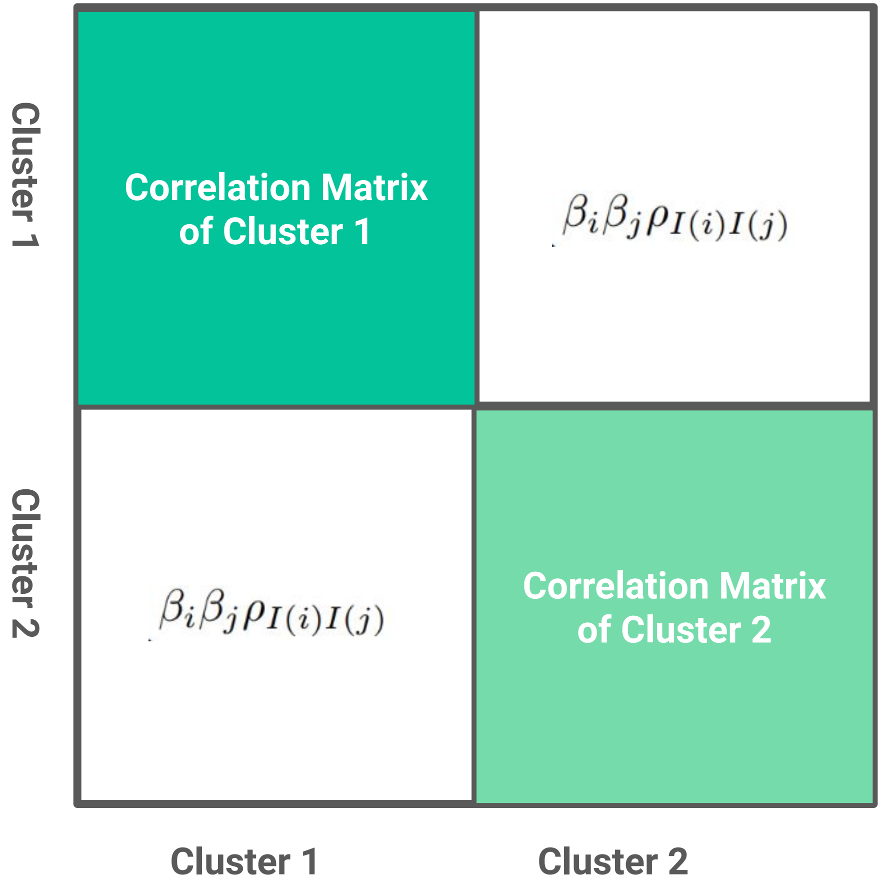
Cluster Benchmark

For each cluster, apply PCA and use PC1 as its benchmark. For stocks in the cluster, calculate their β with the benchmark. For different clusters, calculate their correlation ρ

Cross Terms

If two stocks are not belonging to the same cluster, their correlation is calculated as $\beta_i \beta_j \rho_{I(i)I(j)}$

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



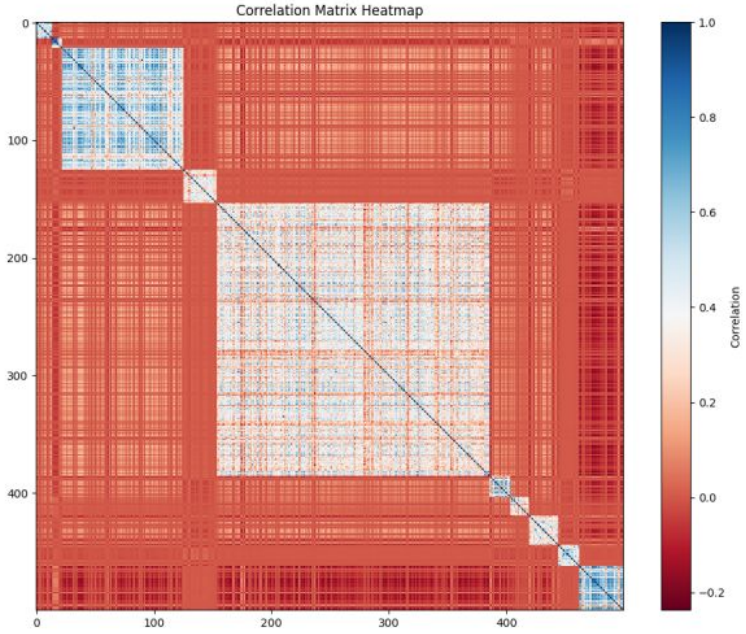
$\beta_i \; \beta_j$

The beta of stock i, j with its sector's benchmark

$\rho_{I(i)I(j)}$

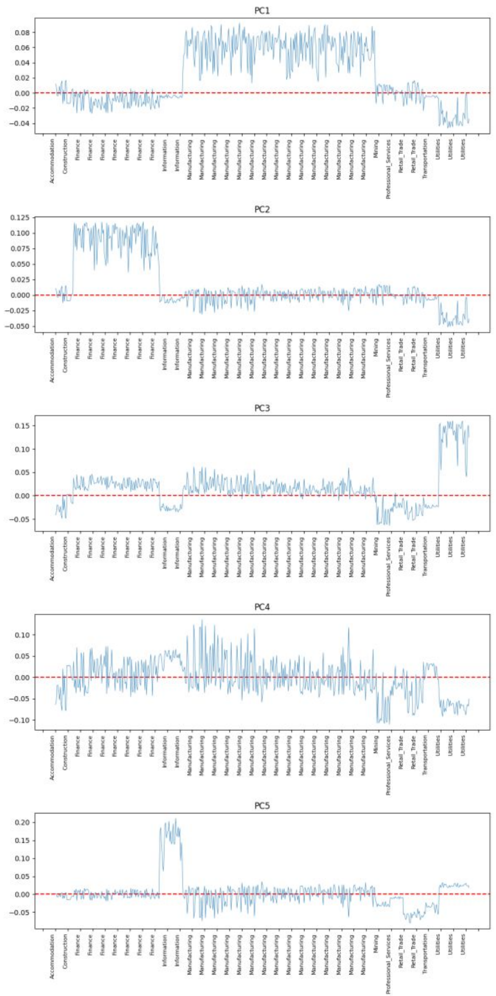
The correlation between stock i's cluster benchmark and stock j's cluster benchmark

Hierarchical PCA (HPCA)



The examination of the correlation matrix reveals a **more discernible structure** characterized by a pronounced correlation among stocks within the same cluster, contrasted with a markedly lower correlation when stocks are not grouped together.

Furthermore, for the first five PCs, there is a **noticeable aggregation of PCs within distinct sectors**. This observation suggests a more economically rational delineation compared to the results obtained through standard Principal Component Analysis (PCA).



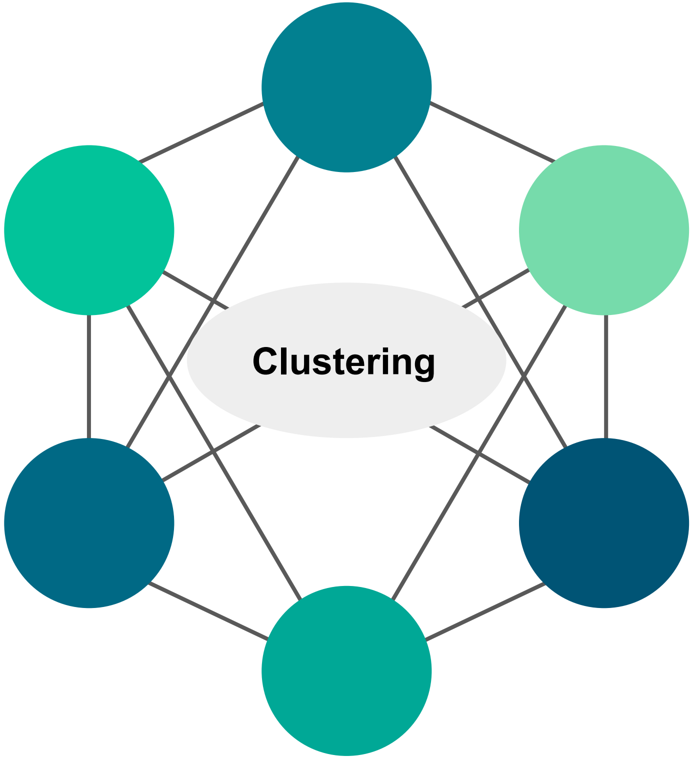
Dynamic Vs Static Clustering

Static Clustering

Uses static or fixed clustering over a period of time using NAICS or GICS

Static clustering approaches do 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

Clustering changes with time dependant on prevailing market conditions

Here we cluster using dynamic market factors rather than countries or sectors such as statistical factors which can change over time

One such factor can be the stocks exposure to climate risk and clustering factors with similar ESG scores together

Issues with Static Clustering

Diversification Faults

Many investment portfolios base their mandates on diversifying their allocations among sectors, sub-sectors, countries, etc., to avoid high and undesirable idiosyncratic risk but isn't the only factor

For example, when interest rates rise sharply, capital-intensive companies are negatively affected and diversification vanishes

Failure In Sector Rotation

Trading strategies, such as the so-called sector/country rotation may also been affected for the same reasons.

Securities that belong to a specific sector/country can change their behavior sharply under the changes of a market regime and the strategy that worked ex-ante may stop working overnight.

Statistical Clustering

To account for hidden risk factors, we have adopted a statistical technique which dynamically adapts to changes in market conditions over time, making it suitable for managing trading portfolios.

We define the number of clusters, which depends on the number of K eigenvectors, without specifying any other parameters or hyper-parameters.

The algorithm constructs new features in the space that retain the behavior of each component based on linear combinations of its main characteristics, leading to statistical clusters of similar-behaved securities.

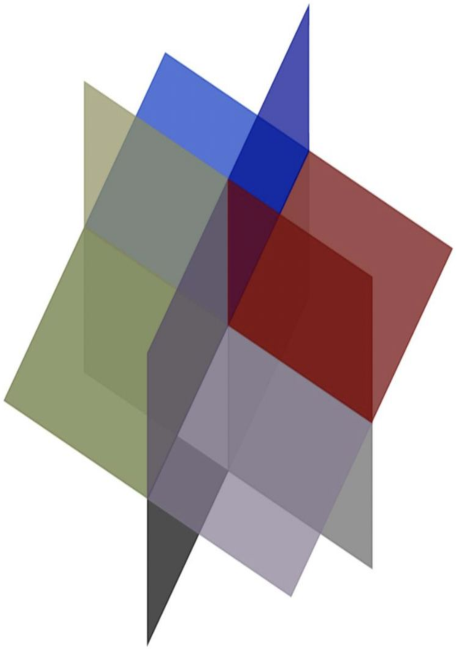


Figure 17: The space is divided into different quadrants (clusters) to which each asset belongs based on the sign of the eigenvectors.

UCLA Anderson School of Management

K-Means Clustering

It is 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.

1. Minimising distance from centroid

   Optimization of cluster assignments

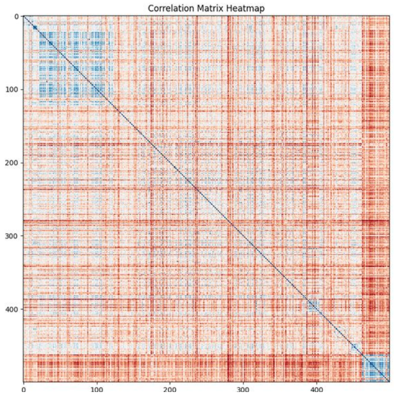
2. Capture more complex cluster structures

   Flexibility in cluster shapes

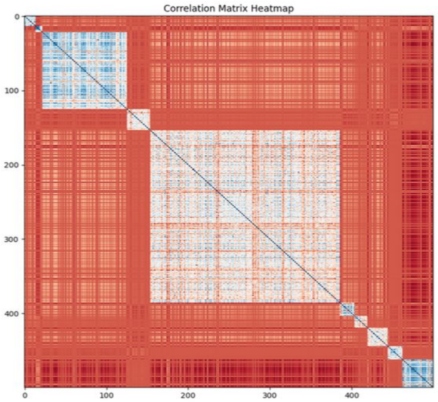
3. Outliers have less influence on the final cluster

   Robustness to outliers

CORRELATION MATRICES COMPARISON

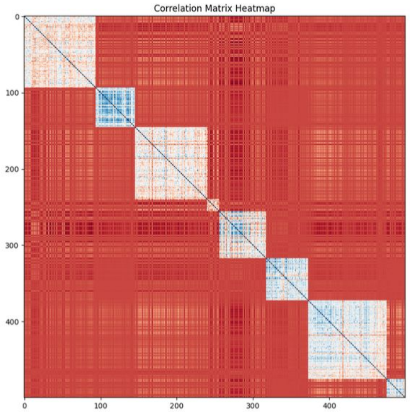


VANILLA PCA

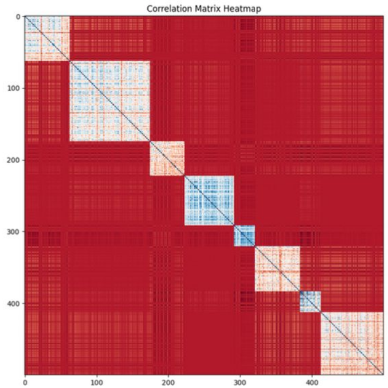


HPCA CLUSTERING USING NAICS CLUSTERS

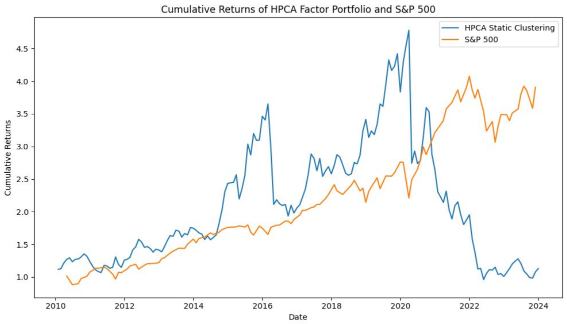
HPCA CLUSTERING USING STATISTICAL CLUSTERS

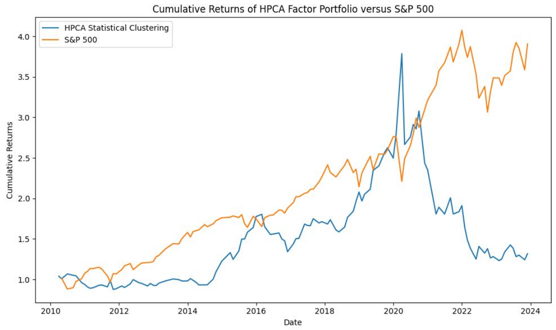


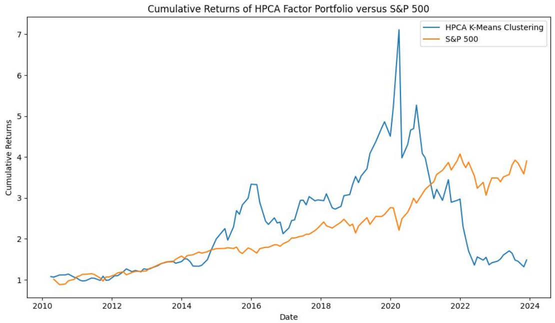
HPCA CLUSTERING USING K_MEANS CLUSTERS



Investment Strategy Performance







Static Clustering

steady returns before 2020 but struggled with higher volatility and drawdowns especially post 2020

Statistical Clustering

closely tracked the S&P 500 before 2020, relatively better resilience post-2020 compared to static clustering

K-Means Clustering

outperformed other methods before 2020 with higher returns and showing better resilience during market downturns

Post-2020 Performance Decline

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

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

1

Regime Shift

The pandemic caused significant regime shifts, altering stock dynamics and correlations

2

Increased Volatility

The market turmoil led to extreme volatility spikes, challenging the models' ability to adapt quickly

3

Sectoral Impact

Different sectors were unevenly impacted, causing misalignment in portfolio allocations

Conclusion and Future Directions

Conclusions

Dynamic Clustering Advantage

K-means clustering, significantly enhance the HPCA framework for sector-based equity portfolio management

Outperformance Justification

Superior performance due to effective identification and capture of evolving risk factors rather than market mispricings

Costs and Risks

Computational expenses, model overfitting risk, and the assumption of persistent historical relationship among stocks

Future Directions

Alternative Clustering Techniques

Incorporating other clustering techniques to further enhance the HPCA framework

Real-time Adaptation

Develop real-time adaptation mechanisms for market cap to ensure timely responses to market changes

Risk Management

Implement rigorous risk management and stress testing frameworks, and extend the application of HPCA to other asset classes