Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Zirui Song December 7, 2024

Unveiling Systematic Risks:

PCA Analysis to High-Frequency and Low-Frequency Factor Data

by

Zirui Song

Ruoxuan Xiong
Adviser

Quantitative Science

Ruoxuan Xiong

Adviser

William Giles Mann

Committee Member

Matthew R. Lyle

Committee Member

2024

Unveiling Systematic Risks:

PCA Analysis to High-Frequency and Low-Frequency Factor Data

By

Zirui Song

Ruoxuan Xiong
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Science

2024

Abstract

Unveiling Systematic Risks: PCA Analysis to High-Frequency and Low-Frequency Factor Data
By Zirui Song

Systematic risks have long been analyzed through factor models. This paper applies quadratic Principal Component Analysis (PCA), as proposed by Pelger (2019), to extract stable systematic risk factors from high-frequency data and compares them with factors identified from low-frequency data. The findings suggest that industry factors like technology, financials, energy, and industrials consistently dominate the explained variance, highlighting their central role in systematic risk. While for the low-frequency data, accounting-related factors emerge as the primary drivers of variance. However, during portfolio optimization, the factors with the highest optimized weights are notably different from the factors in the leading principal components. These weights are more firm-specific and diversified across various categories, emphasizing the importance of idiosyncratic drivers in practical investment strategies. This study contributes to the literature by highlighting the divergence between systematic risk decomposition and optimized portfolio construction, offering valuable insights for risk management and investment strategies. It underscores the importance of integrating firm-specific factors with broader systematic components to enhance portfolio performance and suggests future research directions in dynamic modeling and event-driven analysis.

Unveiling Systematic Risks:

PCA Analysis to High-Frequency and Low-Frequency Factor Data

By

Zirui Song

Ruoxuan Xiong
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Science

2024

Acknowledgements

Special thanks to Professor Ruoxuan Xiong, who guided me throughout the research process, and to Professor William Mann and Professor Matthew Lyle for their constructive feedback, which significantly enriched my work.

Table of Contents

List of Figures

1 Introduction

Factor analysis employs a small number of composite factors to explain a significant portion of co-movements in data. It has been extensively used to explore financial and macroeconomic issues (Bai and Ng 2008; Stock and Watson 2006; Ludvigson and Ng 2009). There are three main types of factor models for asset returns: macro-economic factors, fundamental factors, and statistical factors Zivot and Wang (2003). Macroeconomic factor models rely on measurable economic indicators, such as interest rates and inflation, to capture broad, shared influences on asset returns. In contrast, fundamental factor models utilize observable characteristics specific to firms or assets, such as company size, dividend yield, or industry classification, to identify shared drivers of asset returns. Meanwhile, statistical factor models approach common factors as unobservable or latent, inferring them from the data. The most commonly utilized method for statistical factor analysis is the Principal Component Analysis (PCA).

Traditional PCA relies on the covariance matrix to capture variance and relationships among variables within a dataset. Historically, studies have applied constant factor loadings, disregarding time variations. Given that financial and macroeconomic data often span long periods, using a constant loading model is inadequate for modeling individual stock returns over extended time horizons (Lettau and Pelger, 2020). Over time, modified versions of PCA have been developed to address time variations and nuances in the data. For instance, Kelly et al. (2017), and Fan et al. (2016), incorporated subject-specific conditioning information into factor model estimates. Pelger and Xiong (2020) introduced a kernel method to PCA, allowing it to capture state variations. Lettau and Pelger (2020) developed a risk-premium PCA by including a penalty term to account for pricing errors. More recently, Pelger (2019) proposed an alternative PCA that explains variation without the need for deliberately selecting conditional characteristics, using a quadratic covariance matrix instead.

Understanding systematic risks is crucial for gaining insights into financial markets and asset pricing. Systematic risk is the inherent risk that affects the entire market or a broad sector, often driven by macroeconomic factors such as inflation, interest rates, or geopolitical events. This risk cannot be mitigated through diversification alone, only through hedging or by using the correct asset allocation strategy. There are three commonly employed approaches to identify factors contributing to systematic risk: the Capital Asset Pricing Model (CAPM), the Fama-French three-factor model, and Arbitrage Pricing Theory (APT). CAPM uses a single factor, market risk, to predict asset returns. The Fama-French three-factor model expands on CAPM by incorporating two additional factors: size and value, allowing for the study of systematic risk through the effects of company size, valuation, and market

risk. APT, in contrast, is a multi-factor linear model that posits asset returns are influenced by a variety of macroeconomic factors. Factor analysis is often motivated by the theory underpinning APT.

Boudoukh et al. (2024) investigates how systematic risk evolves during global crises, using high-frequency data and PCA. The study finds that PC1 reflects broad market movements with stable composition, while PC2 acts as a crisis factor with significant shifts during systemic stress, underscoring the diminished effectiveness of diversification and hedging strategies in such periods. Pelger (2019)'s quadratic covariance PCA helps to identify the systematic risks from high-frequency data. Quadratic covariance matrix has bounded eigenvalues for idiosyncratic risks and unbounded eigenvalues for systematic risks. This dynamic factor model using high-frequency data found four stable continuous systematic factors, which can be well-approximated by a market, oil, finance, and electricity portfolio, and one stable jump market factor. Accounting related factors' relationship with systematic risks have also been intensively studied. Bowman (1979) suggests that systematic risk increases with leverage due to heightened exposure to market variability. Campbell et al. (2009) found that stocks sharing similar accounting attributes, such as profitability or leverage, exhibit correlated risks primarily because these attributes are linked to broader economic or market-wide factors.

This paper largely builds upon the foundational work of Pelger (2019) and Pelger (2018), by offering new insights into the stability and economic relevance of systematic risk factors derived from high- and low-frequency data over an extended time horizon. Specifically, it examines the relationship between these factors and asset pricing while considering the influence of significant economic events such as the global financial crisis and the COVID-19 pandemic. In addition, this study evaluates the practical implications of these factors for portfolio optimization, providing a more comprehensive perspective on their utility.

The analysis reveals that, for high-frequency data, the first three principal components collectively explain over 50% of the total variance and exhibit remarkable stability on both yearly and monthly scales. Industry-related factors, including technology, financials, energy, and industrials, consistently dominate the variance explained, highlighting their central role in systematic risk. In contrast, for low-frequency data, accounting-related factors emerge as the primary contributors to explained variance, with the top three components demonstrating greater temporal consistency. However, during portfolio optimization, the factors with the highest optimized weights diverge from the principal components that systematically explain risk, favoring company-specific attributes distributed across a wider range of factors.

This paper is organized as follows. The first section provides an overview of the models employed and the datasets analyzed. This is followed by an in-depth examination of the empirical results, focusing on the stability, economic interpretation of the principal com-

ponents, and portfolio optimization. The final section discusses the broader implications of these findings for asset pricing and portfolio construction, emphasizing the complex interplay between systematic and idiosyncratic factors.

2 Methodology

2.1 PCA Overview

In traditional factor analysis, the covariance matrix $\widehat{\mathbf{\Sigma}}_X$ is used to study co-movements in a dataset, capturing the relationships between different variables in the vector $\mathbf{X}$. The covariance matrix is defined as:

$$\widehat{\mathbf{\Sigma}}_X = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{X}(t) - \overline{\mathbf{X}})(\mathbf{X}(t) - \overline{\mathbf{X}})^\top,$$

where $\mathbf{X}(t) \in \mathbb{R}^n$ is the observed data vector at time t, and $\overline{\mathbf{X}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}(t)$ is the mean vector of the process over time T. This matrix $\widehat{\mathbf{\Sigma}}_X \in \mathbb{R}^{n\times n}$ encapsulates the variances and covariances among all elements of $\mathbf{X}(t)$.

Covariance-based PCA was chosen for analyzing low-frequency data due to its simplicity and effectiveness in capturing primary patterns in stable, linear relationships. Specifically, covariance PCA aligns well with the more stable and less noisy nature of low-frequency datasets, such as monthly or quarterly returns. It is computationally simpler and focuses directly on the variance structure, making it efficient and interpretable for uncovering primary drivers of variation in low-frequency data.

However, with high-frequency data, this covariance matrix becomes difficult to estimate accurately with fixed T. In high-frequency econometrics, Pelger (2019) proposed replacing the covariance matrix with the quadratic covariation process, which is suitable for continuous-time semimartingales.

Let $\mathbf{X}(t)$ represent a semimartingale observed at high frequency. The quadratic covariation process captures the sum of squared increments of $\mathbf{X}(t)$. Partitioning $[0, T]$ into M subintervals of size $\Delta_M = \frac{T}{M}$, with $M \to \infty$ and $\Delta_M \to 0$, the quadratic covariation is given by:

$$\sum_{j=1}^{M}(\mathbf{X}(t_{j+1}) - \mathbf{X}(t_j))(\mathbf{X}(t_{j+1}) - \mathbf{X}(t_j))^\top \xrightarrow{P} [\mathbf{X}, \mathbf{X}]_T,$$

where $[\mathbf{X}, \mathbf{X}]_T$ is the quadratic variation matrix of $\mathbf{X}(t)$ over $[0, T]$, capturing the high-frequency co-movements of the system.

The estimation process uses M observations over $[0, T]$, with $M \to \infty$ and $N \to \infty$ while T is fixed. Define the observed increments as $\mathbf{X}_j = \mathbf{X}_{t_{j+1}} - \mathbf{X}_{t_j}$. For the factors and errors, similar notation applies: $\mathbf{F}_j = \mathbf{F}_{t_{j+1}} - \mathbf{F}_{t_j}$ and $\mathbf{e}_j = \mathbf{e}_{t_{j+1}} - \mathbf{e}_{t_j}$.

The system's dynamics in matrix form are:

$$\mathbf{X}(t) = \mathbf{\Lambda}^\top \mathbf{F}(t) + \mathbf{e}(t),$$

where $\mathbf{\Lambda} \in \mathbb{R}^{K \times N}$ is the factor loading matrix, $\mathbf{F}(t) \in \mathbb{R}^K$ is the vector of systematic risk factors, and $\mathbf{e}(t) \in \mathbb{R}^N$ represents idiosyncratic noise. Over all observations, this becomes:

$$\mathbf{X}_{(M \times N)} = \mathbf{F}_{(M \times K)} \mathbf{\Lambda}^\top + \mathbf{e}_{(M \times N)},$$

We estimate $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{F}}$, where $\mathbf{V}_{NM}$ is the matrix of the K largest eigenvalues of $\frac{\mathbf{X}^\top \mathbf{X}}{N}$. $\widehat{\mathbf{\Lambda}}$ consists of the eigenvectors of $\mathbf{V}_{NM}$, and $\widehat{\mathbf{F}} = \frac{\mathbf{X}\widehat{\mathbf{\Lambda}}}{N}$.

2.2 Comparison between Factors

To compare two sets of factors, $\mathbf{F}$ and $\mathbf{G}$, I use the generalized correlation. Generalized correlation isolates the degree to which individual subspaces align, projecting out previously explained subspaces and iterating through all dimensions. This measure allows for a more detailed understanding of how many factors are shared between the two sets and whether certain factors can fully replicate the others. Specifically, generalized correlation calculates the maximum achievable correlation between linear combinations of the factors from $\mathbf{F}$ and $\mathbf{G}$. The generalized correlation matrix $\mathbf{C}$ is defined as:

$$\mathbf{C} = \left(\mathbf{F}^\top \mathbf{F}\right)^{-1/2} \mathbf{F}^\top \mathbf{G} \left(\mathbf{G}^\top \mathbf{G}\right)^{-1/2}.$$

The entries of $\mathbf{C}$ represent the pairwise correlations between the factor spaces of $\mathbf{F}$ and $\mathbf{G}$ after normalizing each set of factors. The generalized correlation between $\mathbf{F}$ and $\mathbf{G}$ can be expressed as:

$$\text{Generalized Correlation} = \sqrt{\sum_{k=1}^{K} \sigma_k^2},$$

where σ_k are the singular values of $\mathbf{C}$. This measure provides a single value that indicates the overall alignment between the two factor sets, with values closer to 1 indicating stronger alignment.

2.3 Mean-Variance Portfolio Optimization

Systematic risk refers to the inherent market-wide risk that cannot be mitigated through diversification, as it arises from factors like economic downturns, geopolitical events, or major financial crises. Portfolio optimization techniques, such as mean-variance optimization proposed by Harry Markowitz, aim to construct portfolios that balance expected returns against systematic risk. By analyzing assets' co-movements, the approach emphasizes diversification to allocate capital effectively, reducing overall exposure to systematic risk while maintaining desired returns. In this context, the expected portfolio return in matrix form is often represented as:

$$\mathbb{E}(\mathbf{R}_p) = \mathbf{w}^\top \boldsymbol{\mu},$$

where: $\mathbf{w} \in \mathbb{R}^n$ is the vector of asset weights in the portfolio, $\boldsymbol{\mu} \in \mathbb{R}^n$ is the vector of expected returns for each asset.

The portfolio variance, which represents the risk, is given by:

$$\sigma_p^2 = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w},$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is the covariance matrix of asset returns, capturing the co-movement between assets.

The objective of mean-variance optimization can be formulated as a trade-off between maximizing return and minimizing risk. This trade-off is controlled by a parameter λ and expressed as:

$$\min_{\mathbf{w}} \quad \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} - \lambda \, \mathbf{w}^\top \boldsymbol{\mu},$$

where λ is a parameter that determines the emphasis on return versus risk in the optimization.

To avoid over-fitting and ensure robustness, I added a L_2 regularization (ridge) term. This regularization penalizes large portfolio weights, effectively shrinking them towards zero. The regularization term is defined as:

$$\text{Penalty} = \alpha \|\mathbf{w}\|^2 = \alpha \sum_{i=1}^{n} w_i^2,$$

where α is a regularization parameter that controls the strength of the penalty.

The regularized objective function for portfolio optimization then becomes:

$$\min_{\mathbf{w}} \quad \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} - \lambda \, \mathbf{w}^\top \boldsymbol{\mu} + \alpha \|\mathbf{w}\|^2.$$

This formulation allows for balancing the trade-off between achieving high returns, managing risk, and avoiding overfitting through regularization. The solution to this optimization yields an "efficient frontier," representing portfolios that offer the highest expected return for each level of risk.

3 Dataset

High frequency dataset allows to estimate a time-varying factor model without any time variation Pelger (2019). This paper utilizes the dataset featured in Aleti (2022)'s working paper, 'The High-Frequency Factor Zoo.' This dataset comprises 272 high-frequency factor portfolios, including 218 characteristic-sorted factor portfolios drawn from existing literature (Chen and Zimmermann 2021; JENSEN et al. 2023), 6 factor portfolios identified by Fama and French (2015), and 48 industry portfolios. The portfolios are constructed using the stock prices of all common stocks listed on the NYSE, NASDAQ, and NYSEMKT, covering the period from January 1996 to December 2020. High-frequency price data are obtained from the NYSE Trade and Quote Database (TAQ) on WRDS. The data is cleaned using the standard procedures established by Barndorff-Nielsen et al. (2009) and sampled at a 5-minute frequency during the market hours of 09:30 to 16:00. The high frequency factor portfolios are constructed using monthly rebalancing and value weights. The final data used in this paper contains 1-minute percent returns including the adjusted overnight return for 272 portfolios.

Low frequency dataset used in this paper is a monthly open-source asset pricing data compiled by Chen and Zimmermann (2021). This dataset contains over 319 firm level characteristics from prior meta-studies and academic papers. The characteristics were drawn from sources such as Hou et al. (2018), McLEAN and PONTIFF (2016), and Green et al. (2017), along with additions from Harvey et al. (2015). The characteristics cover a wide range of predictors related to stock returns, including accounting data, analyst forecasts, and corporate events. Comparing to the high-frequency dataset, these portfolios put less weights on industry related factors. The dataset ranges from July 1951 to December 2023 and is updated monthly.

4 Empirical Results

4.1 Latent Factor Estimation

4.1.1 High Frequency Data

Results from applying quadratic PCA to the high-frequency dataset largely align with those of Pelger (2019). The proportion of variance explained by each category across the top five principal components (PC1 through PC5) collectively capture around 57% of the total variance in the dataset. The marginal increment of explained variance decreases after PC5. A closer examination of the variance composition, as shown in Figure 1, reveals that industry factors consistently explain a significant portion of the variance within each of these principal components. This dominance of industry-related variance suggests that industry exposure is a key driver of asset movements and co-movements in this high-frequency setting. Additionally, low risk, leverage, momentum, and profitability factors emerge as other influential categories across different principal components, each contributing meaningfully to the variance explained. For instance, "Low Risk" dominates in PC1, underscoring its importance in capturing risk-sensitive variations across assets, particularly in high-frequency settings where stability and volatility are key considerations. Similarly, "Value" plays a significant role in PC2 and PC5, suggesting its importance in identifying long-term structural drivers in asset pricing. Additional contributions from categories such as "Size" and "Profitability" in PC3 reflect the influence of firm-level characteristics on overall variance. This distribution suggests that, beyond industry-specific impacts, factors related to capital structure, market sentiment, and company performance also shape the data's underlying structure.

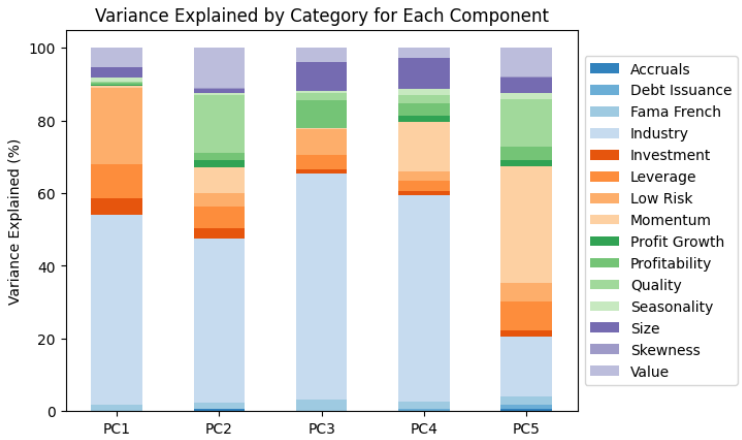

Figure 1: **Explained Variance Ratio by Category**. Skewness captures the asymmetry of return distributions, where positive skewness suggests potential for extreme positive returns, while negative skewness indicates a higher likelihood of severe losses. Low risk refers to strategies that aim to identify assets with lower levels of volatility or variability in returns

A deeper analysis of the signals within each principal component, as shown in Figure 2, emphasizes the critical role of the top-ranked signals in capturing variance. For instance, the steep initial rise in cumulative variance across all components demonstrates that a handful of dominant signals—such as "chips" and "comps" in PC1 or "banks" and "coal" in PC2—play a disproportionately large role in defining the structure of the component. As we progress down the signal ranking, the cumulative variance curve flattens, highlighting the diminishing marginal contributions of lower-ranked signals. This pattern suggests that while the leading signals encapsulate the essential characteristics of each principal component, the residual variance is distributed across numerous minor signals, which may capture sector-specific nuances or noise.

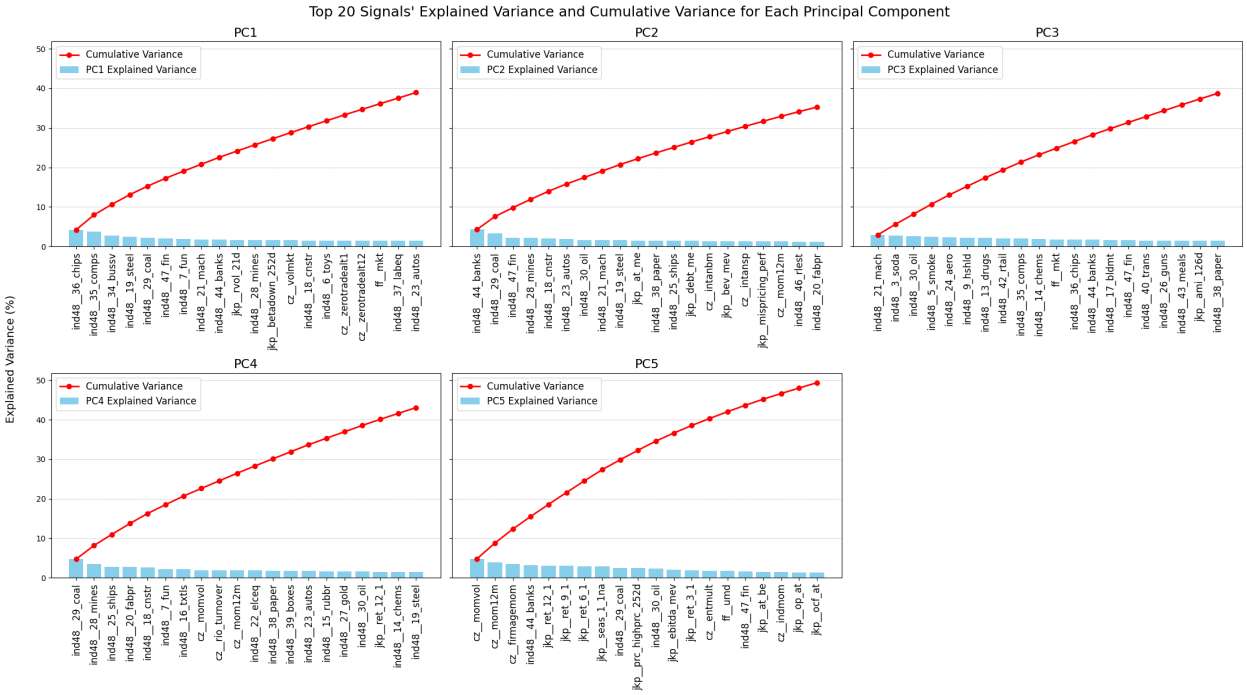

Figure 2: **Top 20 Signals' Explained Variance and Cumulative Variance for Each Principal Component.** This figure highlights the top 20 signals contributing to each principal component (PC1 to PC5) based on their explained variance. The bars represent the individual explained variance for each signal, while the red line shows the cumulative variance across the ranked signals.

Further contextualizing these findings, the analysis shown in Figure 3 highlights the top 10 dominant industries that explain a significant portion of variance within each component. Different principal components capture varying industry-specific influences, with technology, financials, energy, and industrials recurring as key drivers.

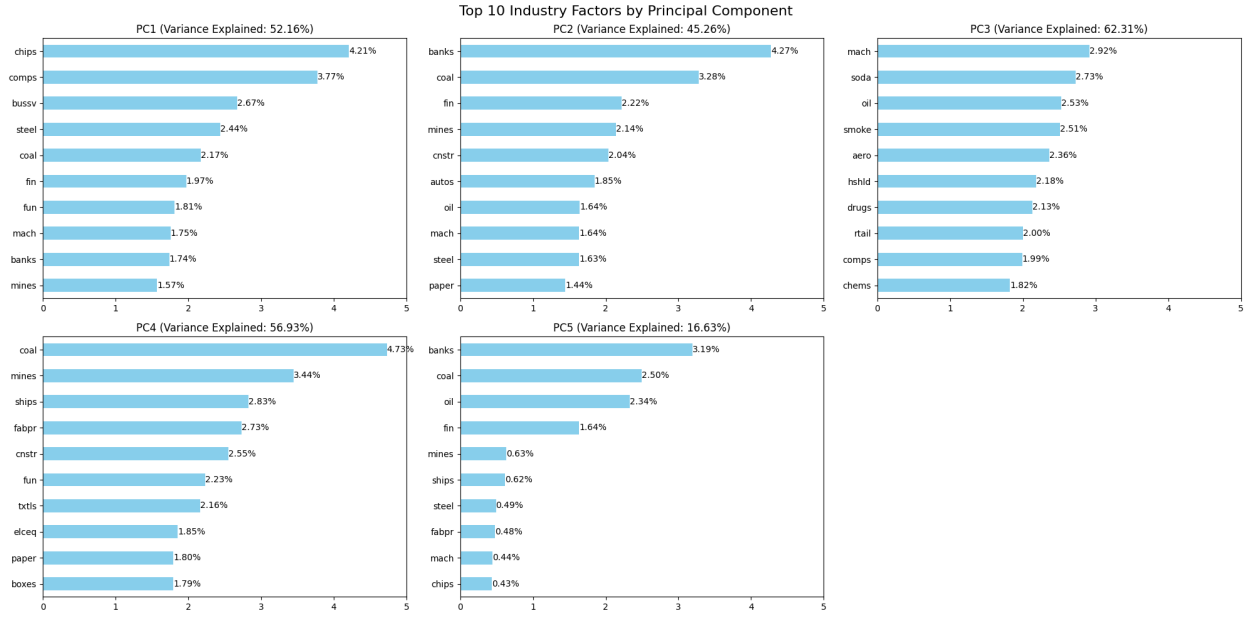

Figure 3: **Top 10 Industry Factors by Principal Component.** This figure displays the top 10 industry factors contributing to each principal component based on their explained variance ratios

For PC1, in which industry factors explain 52.16% of the variance explained by PC1, the chips (semiconductor) and computers industries lead, followed by business services and steel, indicating that technology and basic materials play a prominent role in this component. Industry factors in PC2 explain 45.26% of the variance, with the banks and coal industries at the top, suggesting an emphasis on financial and energy-related sectors in this component. Moving to PC3, which industry factors account for 62.31% of the variance, the primary factors include machinery, soda, and oil, pointing toward industrial goods and consumer products as influential industries. Industry factors in PC4 explains 56.93% of the variance, with coal and mines as the leading industries, followed by ships and fabricated products, emphasizing sectors related to natural resources and heavy manufacturing. Finally, industry factors in PC5 explains 16.63% of the variance and is primarily influenced by banks, coal, and oil, with financials and energy again showing significant importance, though with a lower overall variance explained compared to other components.

Building on the detailed breakdown of signals across the top five principal components, it is also crucial to examine how these components evolve over time. I independently estimated the factor structure within each year and observed how it evolves over time. The cumulative explained variance ratio for each year follows a similar pattern, with most variance being captured by the first few principal components, and diminishing marginal improvements after approximately the fifth component. This trend suggests that the main structure of the data is concentrated within the initial components.

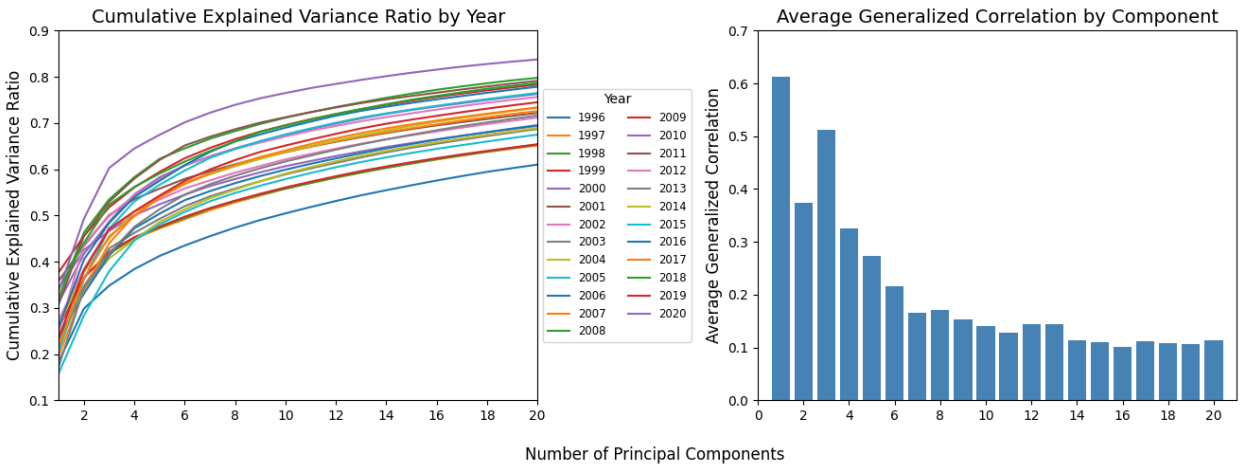

Figure 4: **Annual Cumulative Explained Variance and Component-wise Average Generalized Correlation.** The left plot shows the cumulative explained variance ratio by year, indicating the proportion of variance captured as additional principal components are added. The right plot illustrates the average generalized correlation by component.

To further assess the stability of these components, I applied a generalized correlation analysis across years. The results shown in the right plot of Figure 4 indicate that the first three principal components are the most stable over time, with consistently higher pairwise correlations between years. In contrast, the subsequent components display greater variability, suggesting they may capture more transient or noise-like patterns. This behavior highlights the robustness of the primary components in representing fundamental structures in the data, while the remaining components are more sensitive to year-specific fluctuations and potentially reflect less stable information.

On a monthly basis, as shown in Figure 5 the results follow a similar pattern as observed on the yearly scale: the top three principal components exhibit greater stability compared to the remaining components. However, the generalized correlation values are generally lower, reflecting increased noise and volatility in the monthly data. This decrease in correlation suggests that shorter time frames introduce more transient fluctuations, which affect the stability of the components. Consequently, while the leading components remain relatively consistent, the minor components capture more of the short-term noise, resulting in less stable correlations across months.

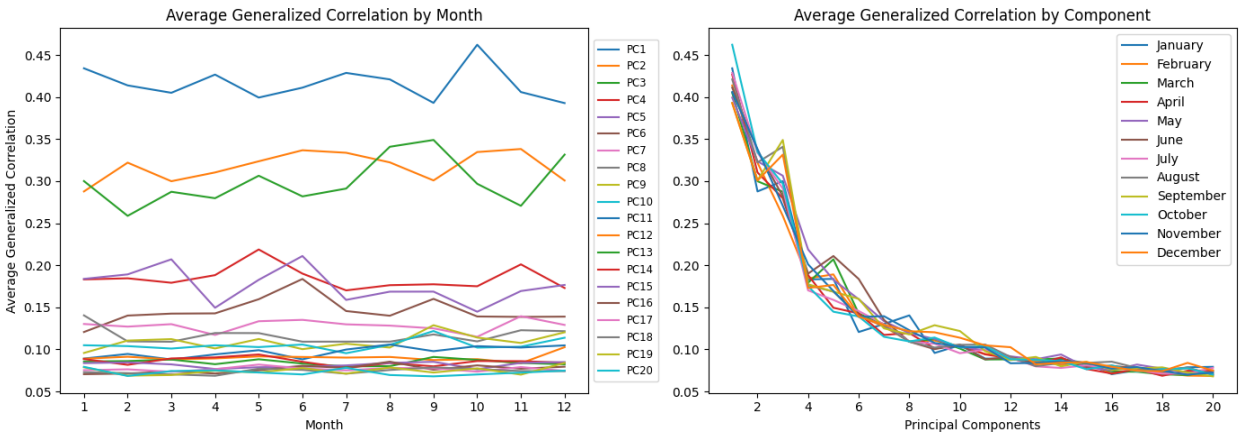

Figure 5: Monthly Average Generalized Correlation.

Previous studies indicate that the first principal component, often the most significant in terms of systematic impact, typically aligns closely with the market factor. To test this hypothesis, I gathered SPY trading data at the millisecond level from WRDS, covering a period starting in 2015. The data was then aggregated to the minute level, allowing me to compute and normalize returns on a minute-by-minute basis.

Upon analyzing the generalized correlation between the initial principal components and SPY returns, I found that the first principal component had a surprisingly low correlation of just 0.27 with SPY returns. This weaker correlation may result from the heightened noise inherent in high-frequency (minute-level) data compared to daily data, where market patterns are more discernible. Interestingly, the second and third principal components exhibited much stronger correlations, at 0.86 and 0.80, respectively. Given that these three components collectively capture over 50% of the total variance, it suggests that they jointly represent substantial market influence, despite the first component's lower correlation.

To further investigate, I examined the correlation between SPY returns and all 272 factors in the dataset. Interestingly, factors with correlations exceeding 80% were predominantly industry-related, with the exception of the Fama-French market factor, which held the highest correlation. As illustrated in Figure 6, for instance, factors representing the finance and service sectors—both of which load heavily on the first principal components—demonstrated correlations above 85% with SPY returns. This suggests that, although the first principal component alone may not fully capture the market factor, the combined influence of the top three components, especially those linked to significant industry factors, effectively encapsulates the market's impact. Overall, the primary principal components, particularly those with strong industry associations, appear to represent the market factor within this high-frequency dataset effectively.

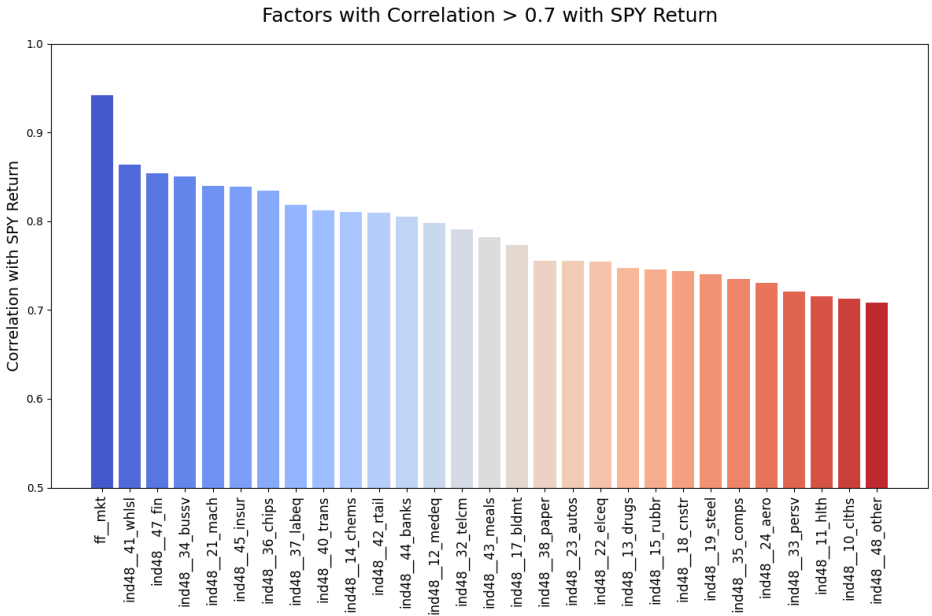

Figure 6: **Factors with Strong Correlation to SPY Returns.** This figure highlights the factors exhibiting a correlation greater than 0.7 with SPY returns. The Fama-French market factor (ff_mkt) demonstrates the strongest correlation, followed by industry-specific factors such as wholesale, finance, and business services.

4.1.2 Low Frequency Data

Applying covariance PCA model to the lower frequency dataset produces a much stronger set of factors. The first principal component explains over 90% of the variance. While the marginal increase in the explained variance ratio decreases as more principal components are added, the first five principal components together account for over 97% of the variation. To test whether the first principal component corresponds to the market factor, I retrieved monthly market factor data from Kenneth R. French's website and calculated the correlation. A correlation coefficient of 0.9 indicates that this principal component closely reflects the market factor.

However, since the low-frequency data places emphasize less on industry factors, the factors with the highest loadings in each principal component are primarily related to company specific factors such as financial or accounting as shown in Figure 7. In PC1, the "Accounting" category dominates, contributing the largest share of the explained variance, followed by smaller but notable contributions from "Price," "Analyst," and "Trading." This distribution aligns with the interpretation that PC1 encapsulates a broad, systematic factor that aggregates contributions from multiple categories rather than being dominated by a single, narrow driver. In contrast, subsequent PCs—PC2 through PC5—exhibit a more diversified structure. For instance, "Analyst" and "Trading" factors significantly influence PC2 and

PC3, while PC5 is primarily driven by "13F" data. This shift in category dominance across PCs highlights the differing roles of these components: while PC1 captures the overarching systematic variance, the other PCs reflect category-specific dynamics or less prominent drivers of variation.

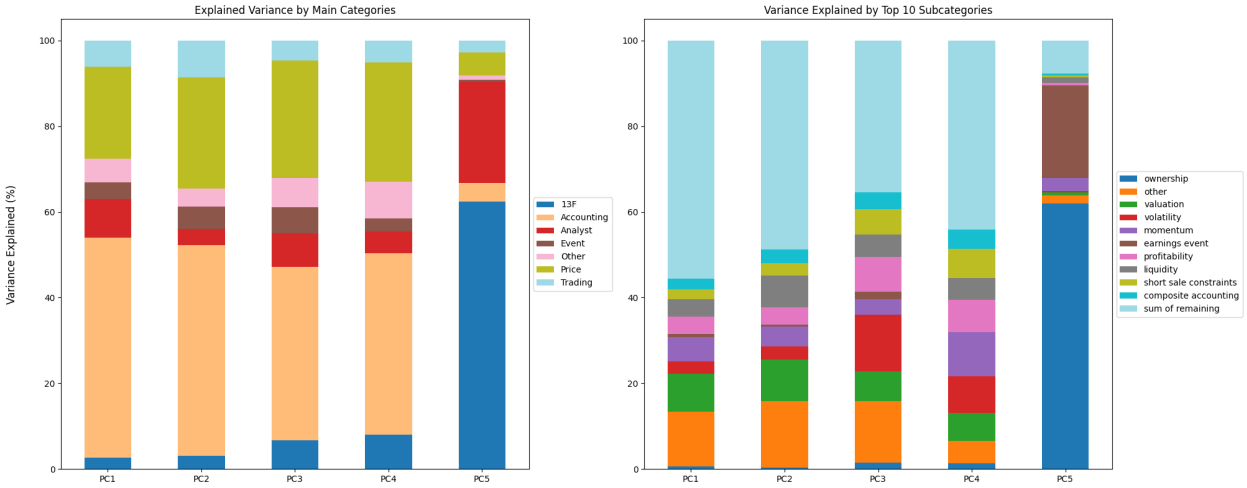

Figure 7: **Explained Variance Ratio by Main and Subcategories.**13F represents institutional ownership metrics derived from SEC Form 13F filings, capturing insights into institutional activity and market constraints such as short sale limitations. Analyst encompasses signals derived from financial analysts' activities, including earnings forecasts, recommendation changes, and revisions, reflecting their influence on market sentiment and stock performance.

The cumulative explained variance of the top 20 signals, as shown in Figure 8, further supports this observation. Notably, while PC1 captures over 90% of the total variance across the entire dataset, the top 20 signals within PC1 collectively explain only around 10% of the variance for this component. In contrast, the cumulative explained variance of the top 20 signals is significantly higher for other PCs, such as PC2, PC3, and PC4, where these signals contribute a more substantial proportion of their respective variances. Together, these findings emphasize that PC1 serves as a general market factor, while the subsequent PCs are more reflective of distinct, category-specific behaviors within the dataset.

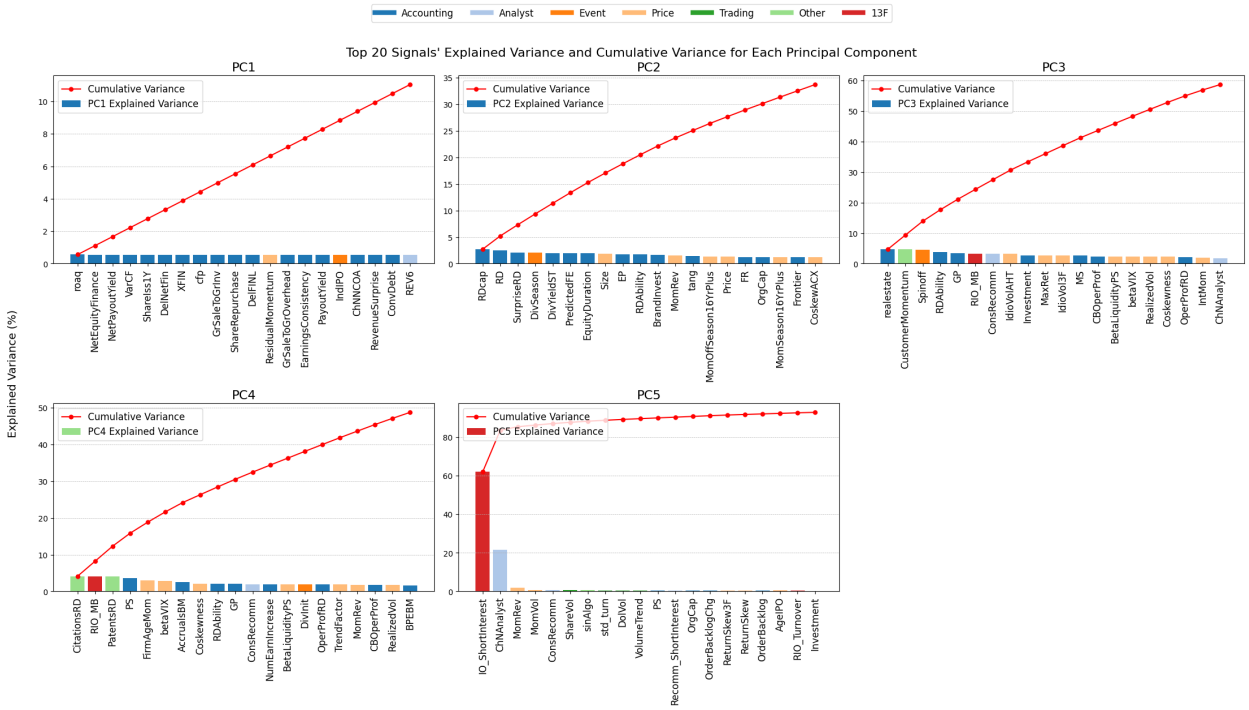

Figure 8: **Distribution of the top 20 signals' explained variance for each principal component (PC1-PC5)**, showing the cumulative contribution (red line) and individual contributions (bars) of key signals, categorized by sector (e.g., Accounting, Analyst, Price, etc.), with PC1 displaying the smallest cumulative variance for the top signals compared to the higher PCs.

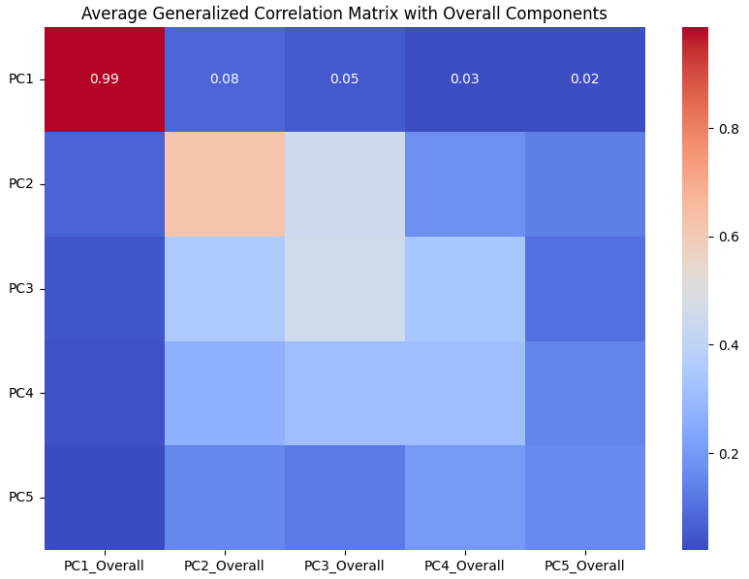

Figure 9: **Generalized Correlation Matrix.** The heatmap displays the average generalized correlation matrix between PCs for each year (y-axis) and PCs for all years (x-axis).

The generalized correlation matrix shown in Figure 9 reveals a clear pattern of stability

across principal components over year. The high diagonal value for PC1 (0.986) shows that this component is consistently aligned with the overall structure, indicating that PC1 captures a dominant, stable source of variance across years. For other components, such as PC2 and PC3, the diagonal values are somewhat lower (0.626 and 0.456, respectively), suggesting moderate stability but also potential variation in the factors they capture year to year. The declining values down the diagonal, particularly for PCs 4 and 5, reflect less consistent alignment with the overall structure, implying these components capture more transient or less dominant patterns in the data. This result highlights the persistent influence of PC1 and the more variable roles of subsequent components, which may be sensitive to specific temporal factors or less central to the dataset's overall variance.

4.2 State-varying Factors

4.2.1 High Frequency Data

I augmented the high frequency dataset with binary recession indicators sourced from FRED, where '1' denotes a recession and '0' denotes a non-recession period. Dissecting the data into recessionary and expansion periods revealed variations in the factor loadings for each principal component, which also differed across recessions (2001, 2008, and 2020).

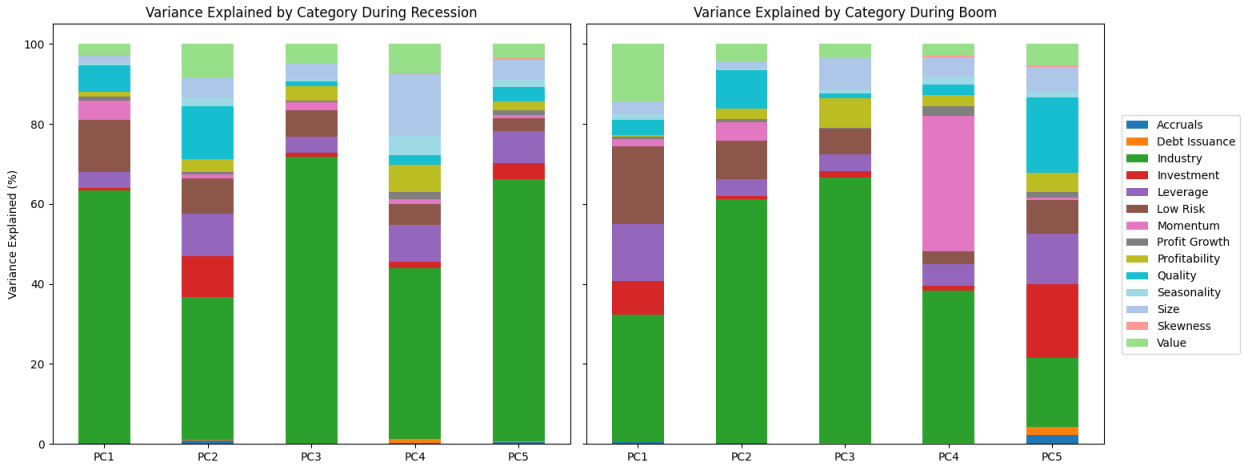

Figure 10: **Variance explained by main categories during recession (left) and boom (right) periods**

During recessions, as shown in the left plot in Figure 10, the "Industry" category emerges as the dominant contributor to variance across all five principal components. This underscores the outsized role that sector-specific dynamics play in shaping market behavior during economic downturns, where structural characteristics and performance variability between

industries become more pronounced. Recessions tend to affect sectors unevenly, with industries such as banking, utilities, and mining often serving as anchors of stability due to their structural importance and critical functions in the economy. This is evident in Figure 11, which reveals that top signals contributing to variance during recessions are predominantly industry-focused, including essential sectors like utilities and resource-driven industries like mining and energy. Additionally, categories such as "Quality," "Low Risk," "Leverage," and "Value" play significant roles, reflecting the market's focus on defensive strategies and risk mitigation during periods of economic distress. The emphasis on "Quality" highlights the importance of firms with strong operational efficiency and consistent performance, which are viewed as safer investments during uncertain times. The prominence of "Low Risk" factors demonstrates the market's preference for stability and resilience, aligning with the defensive positioning characteristic of recessionary environments, while "Value" emphasizes on undervalued opportunities relative to their intrinsic worth.

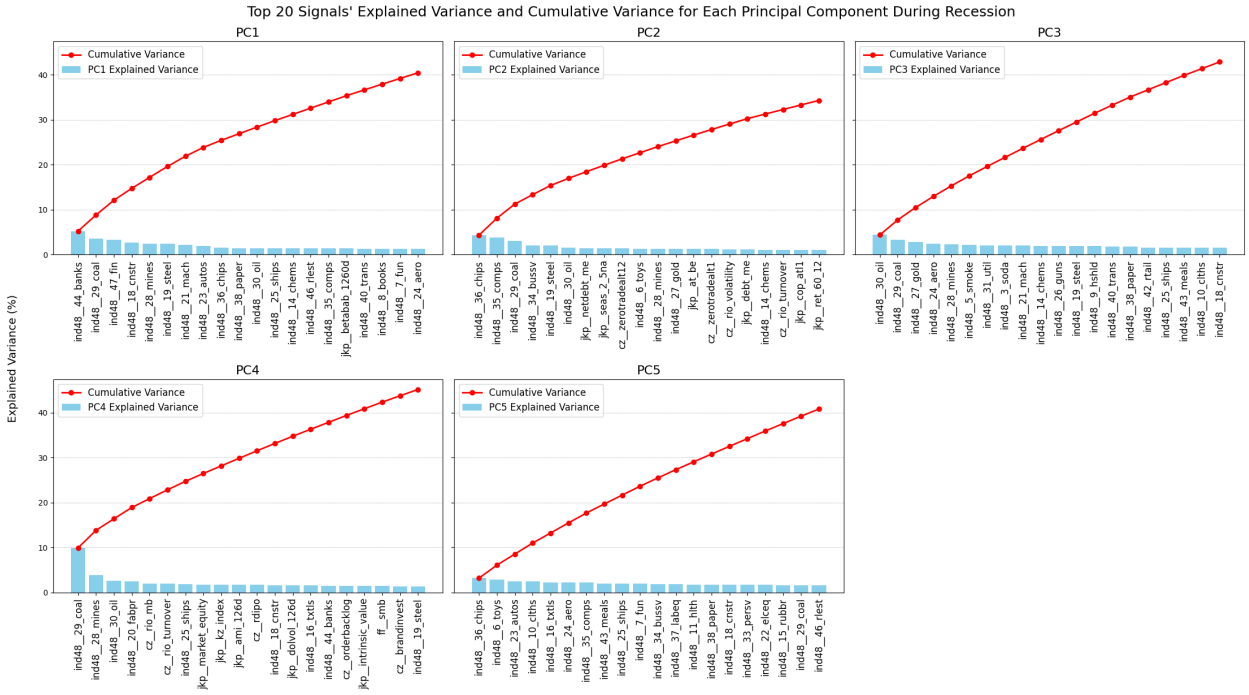

Figure 11: **Top 20 Signals' Explained Variance within Each Component During Recession**

During boom periods, "Industry" remains a primary driver of variance across all principal components, mirroring its importance during recessionary periods. However, "Leverage" gains greater prominence, particularly in PC1, reflecting heightened debt utilization and risk-taking as hallmarks of economic growth. Additionally, factors such as "Momentum," "Profitability," and "Value" grow in significance, signaling a market focus on sustained up-

ward trends, strong earnings quality, and undervalued opportunities. The heightened emphasis on "Value" underscores investors' expansion-oriented behavior, while the growing role of "Profitability" highlights the importance of robust earnings and operational efficiency in a thriving economy. Together, these shifts reflect a more diversified market focus during boom periods, where investors balance risk-taking with growth-oriented strategies, all while maintaining an underlying emphasis on sector-specific dynamics. Moreover, as shown in Figure 12, the explained variance for signals within PC1 is more distributed and diverse compared to recession periods, indicating that investors are broadening their focus across a wider range of factors rather than relying on a few dominant signals. This diversification marks a transition from the defensive concentration typical of recessions to a more growth-driven, opportunistic approach during periods of economic expansion.

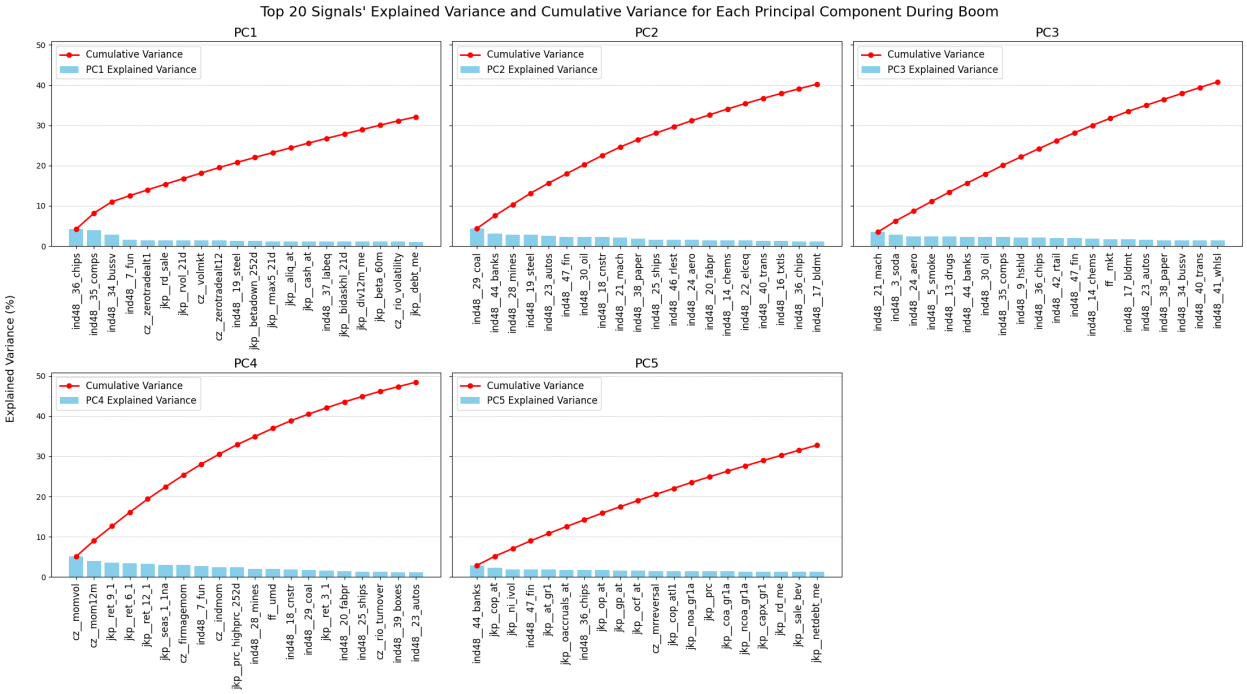

Figure 12: **Top 20 Signals' Explained Variance within Each Component During Boom**

Taking a closer look at the three recessions (2001, 2008, and 2020), the sum of the diagonal increases as the number of principal components included in the analysis rises. The first three components maintain the highest generalized correlation across all three recessions, whereas the fourth and fifth components consistently exhibit much lower correlations. This indicates that the first three components are relatively stable and capture the core economic factors that persist across different recessions, while the fourth and fifth components are more sensitive to idiosyncratic characteristics of each individual recession, playing a more prominent role in explaining the unique aspects of each period.

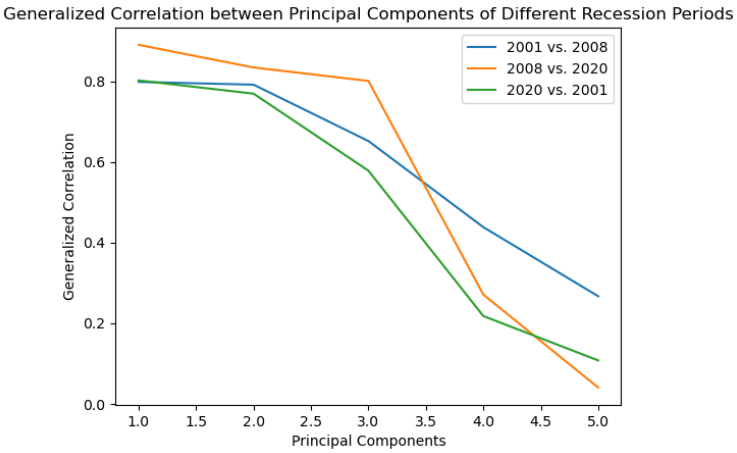

Figure 13: **Generalized Correlation between Principal Components of Different Recessions**

4.2.2 Low Frequency Data

Similar analysis on the low-frequency dataset highlights that the generalized correlation between recession and boom loadings is predominantly strong only for PC1, which represents the market factor. This high correlation (close to 1) indicates that the market factor's influence remains relatively stable and consistent across economic cycles, underscoring its resilience as a primary driver of variance in both boom and recession periods. Interestingly, as shown in Figure 14, there are cross-matching correlations between other components as well, though at a lower strength. Specifically, the second principal component during recession correlates most strongly with the third principal component in boom, while the third principal component in recession aligns with the second principal component in boom. This suggests that certain underlying economic signals or patterns may shift in order of importance or manifestation between recession and boom periods, reflecting a reordering of factors depending on economic context.

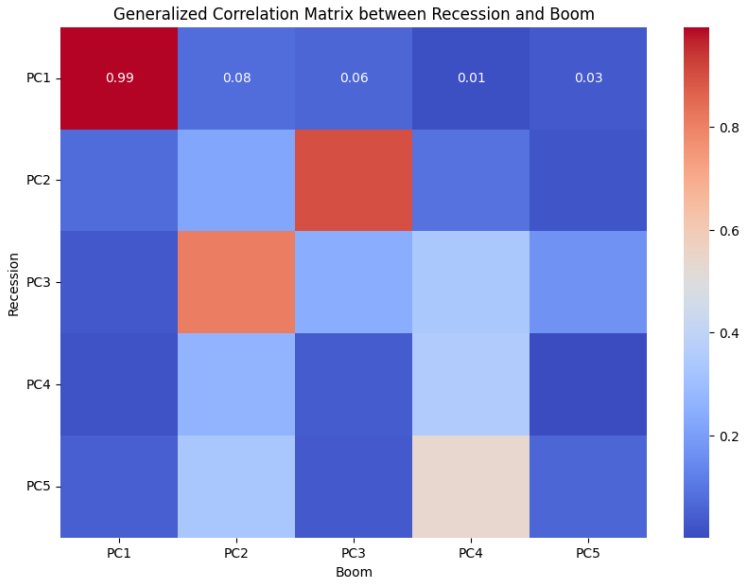

Figure 14: **Generalized Correlation Matrix between Recession and Boom**

Breaking down the variance explained by main and subcategories reveals that accounting-related factors consistently play a dominant role in explaining the top principal components across both recession and boom periods. However, the specific subcategories within these main categories vary significantly between recession and boom phases, reflecting a shift in economic signals and priorities depending on the broader economic conditions. During recession periods, subcategories related to liquidity, payout indicators, and risk management gain prominence, while in boom periods, factors such as valuation and momentum become more influential. This differentiation underscores the adaptability of key financial and economic drivers to align with the prevailing economic landscape.

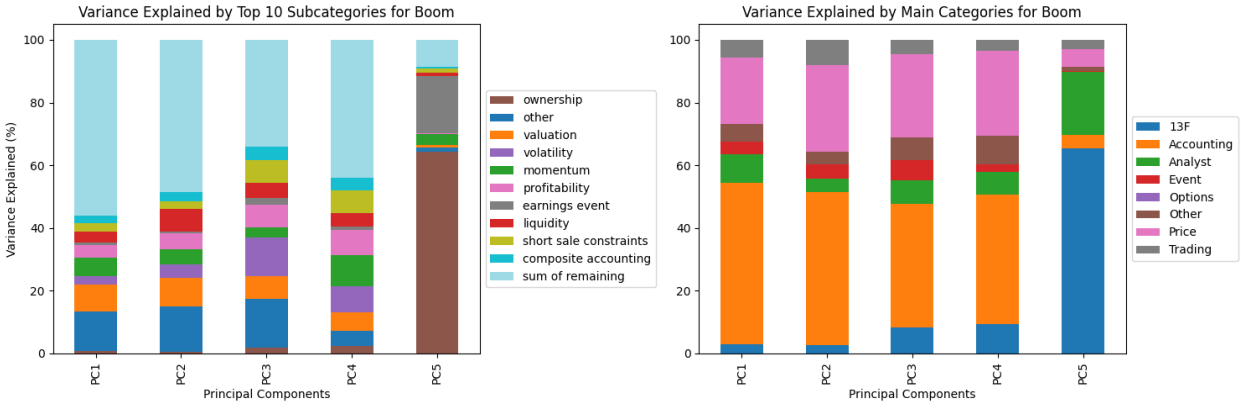

Figure 15: **Variance Explained by Sub and Main Categories for Boom**

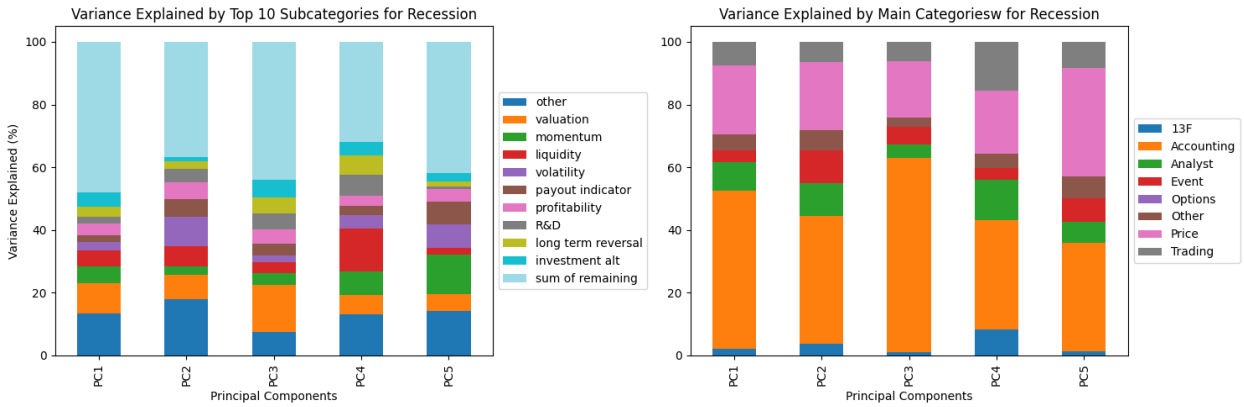

Figure 16: **Variance Explained by Sub and Main Categories for Recession**

4.3 Portfolio Optimization

I applied the principal components estimated from the quadratic PCA model (for high frequency data) to optimize portfolios using mean-variance optimization on a rolling window basis.

I began with a 3-year training period, followed by 1 year for validation and another year for testing. Eigenvalues and factor loadings were derived from the training set and subsequently applied to both the validation and testing sets. Using the optimization model, I optimized the mean-variance portfolio weights on the validation set and applied the resulting optimal weights to the test set. However, this setup yielded no improvement in the testing Sharpe ratio. To address this, I experimented with a shorter time frame, comprising 1 year of training, 1 month of validation, and 1 month of testing. Despite this adjustment, the results still failed to show a substantial improvement, likely due to excessive noise in the high-frequency data.

To minimize noise, I aggregated the data into 5-minute intervals and reran the optimization process. The highest out-of-sample Sharpe ratio was achieved using 15 principal components. Upon further analysis, I observed that the first five components, which collectively explain approximately 60% of the variance in the data, did not correspond to the highest weights identified through optimization. This aligns with the findings of Pelger (2018), which suggest that while the initial principal components primarily capture systematic risk, subsequent components are better at identifying pricing errors, making them particularly useful for constructing portfolios with enhanced Sharpe ratios.

Moreover, the generalized correlation between corresponding components across rolling windows shows that the correlations for the later components are significantly lower than those for the first three components. This indicates that the later principal components are

less stable over time, potentially reflecting transient or noise-like patterns, while the first few components represent more consistent and systemic influences within the data.

The analysis of optimized weights for principal components over time reveals distinct behaviors during different economic conditions. While the weights for principal components vary across periods, the weights assigned to specific signals, as shown in Figure 17, demonstrate a much clearer and more prominent trend. This trend illustrates that a wide range of factors collectively contributed to shifts in risk exposures and portfolio allocations. Notably, the spike and subsequent plummet of signal weights in the first quarter of 2005 may be driven by the Federal Reserve's tightening monetary policy, as interest rates were raised to combat inflation during the post-recession recovery. Rising energy prices and a booming housing market also contributed to changes in asset pricing dynamics, prompting portfolio adjustments. In 4Q2017, another significant shift happening, several significant events may influence the weight allocation. The U.S. passed the Tax Cuts and Jobs Act, which included corporate tax rate reductions, fueling optimism in equity markets. Additionally, the Federal Reserve raised interest rates, signaling confidence in economic growth. Globally, rising oil prices due to OPEC-led production cuts and geopolitical tensions influenced commodity and energy sectors, while strong corporate earnings bolstered market sentiment. Other sharp movements in weights can be observed around the 2008 financial crisis, early 2016 (a period marked by concerns over global economic slowdown), and the 2020 COVID-19-induced recession, each reflecting heightened market sensitivity during these times.

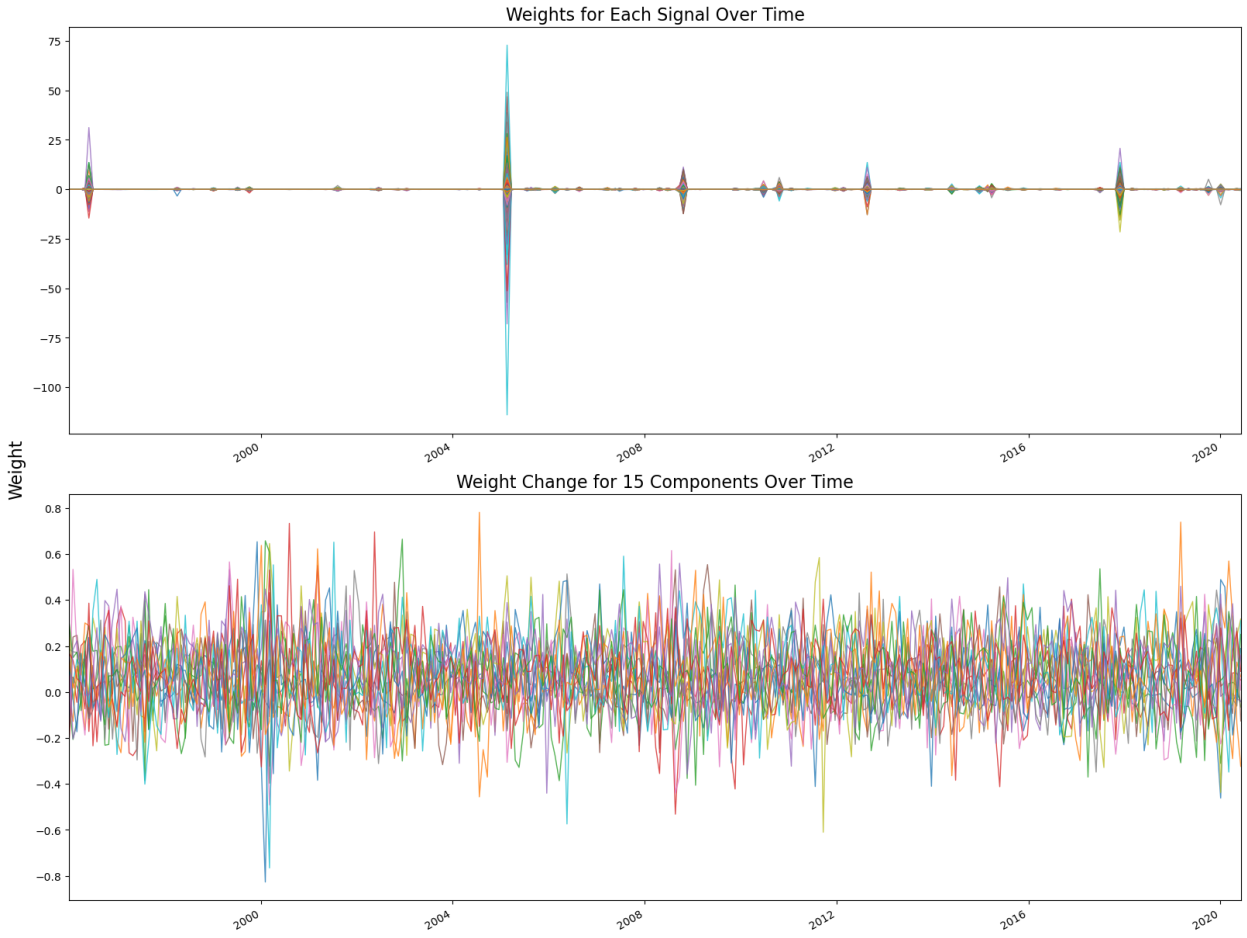

Figure 17: **Optimized Weight Change for 15 Principal Components and 272 Specific Signals Over Time**

Figure 18 highlights a clear relationship between heightened market volatility, as indicated by the spikes in the VIX index, and increased activity in portfolio allocations. Periods of elevated VIX values, which signify market uncertainty and risk aversion, correspond to noticeable fluctuations and peaks in signal weights. This suggests that during volatile market conditions, portfolio optimization strategies react dynamically, potentially reallocating weights across assets to manage risk or capture opportunities. The alignment between VIX spikes and changes in signal weights demonstrates how market sentiment and uncertainty drive more pronounced adjustments in asset allocation.

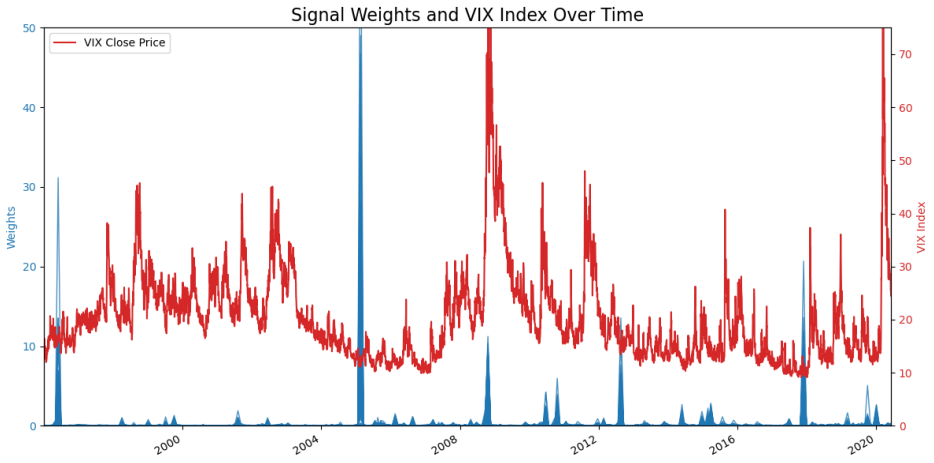

Figure 18: **Relationship between Optimized Signal Weights and VIX Index**

Breaking these components down into individual signals reveals a notable shift, with company-specific factors gaining prominence and industry-related factors assuming a less significant role. These top features also capture the major fluctuations over time. Figure 19 provides a clear depiction of the top 20 long and bottom 20 short signals based on their average weights, with classifications represented by distinct colors. Notably, there are significantly less industry signals comparing to what we got from the dominant PCs. Signals with positive average weights (long positions) are spread across classifications like "Leverage," "Quality," "Size," and "Seasonality," reflecting a targeted focus on firm-specific characteristics and cyclic trends. The prominence of "Leverage" signals a market preference for firms strategically using debt to enhance returns, while "Quality" captures a tilt toward companies with strong fundamentals and reliable earnings. "Size" and "Seasonality" classifications indicate a nuanced approach to capturing opportunities tied to firm-specific attributes and cyclic trends.

Conversely, the bottom 20 signals with negative average weights (short positions) include a diverse set of classifications such as "Value," "Momentum," "Industry," and "Profitability." These short positions suggest a strategic underweighting of factors perceived as overvalued, volatile, or less favorable in the current market environment. The inclusion of "Value" in short positions suggests caution toward undervalued stocks that may lack growth catalysts, while shorting "Momentum" highlights the potential risk of reversals in assets with strong recent performance. The presence of "Industry" in this category, though less dominant compared to long positions, suggests an intentional reduction in exposure to sectors expected to underperform or over-exposed to risks.

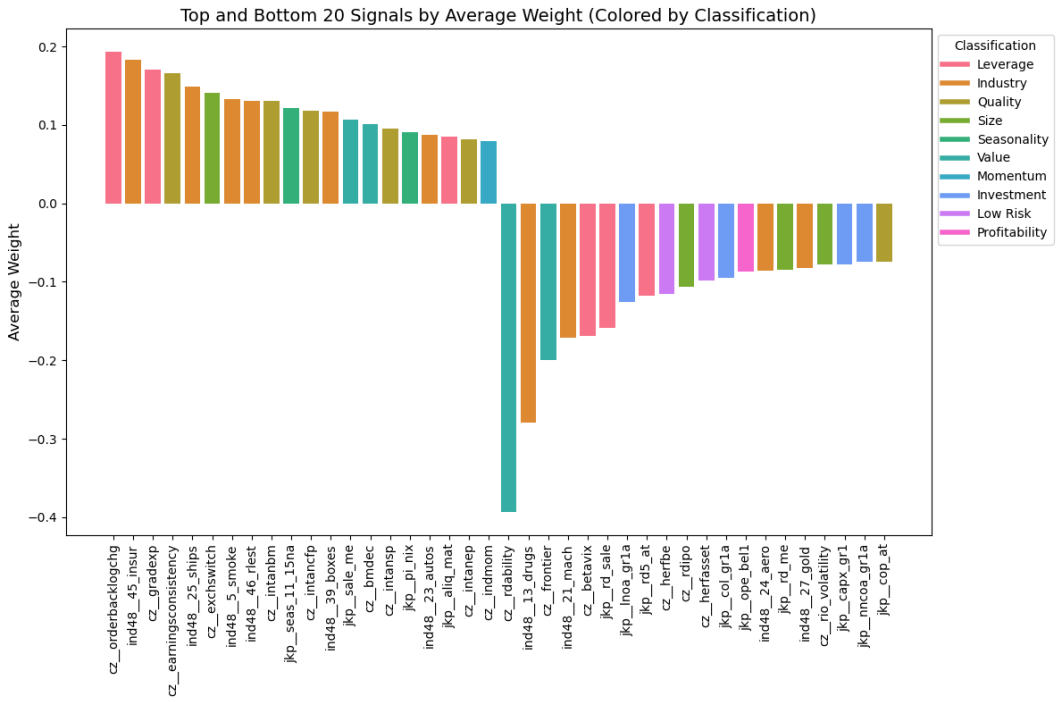

Figure 19: **Top and Bottom 20 Signals by Average Weight (Colored by Classification)** This bar chart illustrates the top 20 (long positions) and bottom 20 (short positions) signals based on their average weights, categorized by their classifications.

The optimized weights for signals across recession and non-recession intervals reveal unique patterns in market sensitivity to economic and financial factors. During non-recession periods, as shown in Figure 20, there are frequent and significant fluctuations in signal weights, suggesting a high degree of market responsiveness to economic indicators, possibly driven by market corrections, speculative trading, and investor reactions to minor shifts in financial or economic conditions. These fluctuations could be attributed to market corrections as investors adjust their portfolios in response to perceived risks, policy changes, or shifts in growth expectations. The non-recession intervals demonstrate a persistent rebalancing of signal weights, which may reflect the volatility that arises during economic expansions when markets are more susceptible to overvaluations and speculative corrections.

In contrast, during recession intervals, illustrated in Figure 21, the signal weights exhibit volatility patterns that correspond more directly with the economic downturns themselves. For example, the 2008–2009 recession, triggered by the housing crisis, shows pronounced spikes and troughs in weights across signals, indicating that financial signals were highly reactive to the crisis. This heightened volatility in weights during 2008–2009 reflects the market's struggle to find stability as it grappled with unprecedented levels of economic distress, liquidity crises, and systemic risk. In comparison, the 2001 and 2020 recessions, although still volatile, exhibit less dramatic swings, suggesting that while the recessions

impacted the market, they did not evoke the same level of prolonged instability across signals as seen in 2008–2009.

The differences in signal weight fluctuations between recession and non-recession periods underscore how portfolio weights are depending on the broader economic context. In non-recession periods, the frequent corrections may indicate that investors need to continually recalibrating their portfolio in response to policy changes, earnings reports, economic data releases, and market sentiments. However, during recessions, the adjustments in weights become more synchronized as markets collectively respond to downturn-induced stresses, with the degree of fluctuation dependent on the nature of the recession.

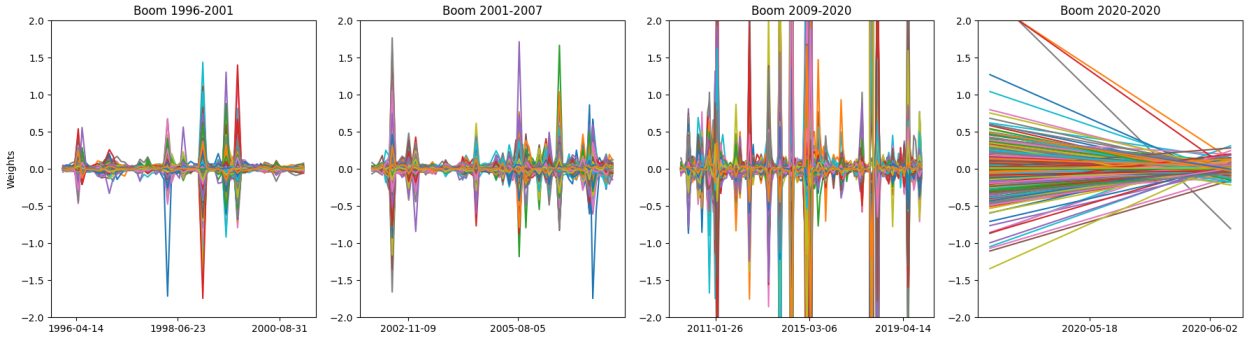

Figure 20: **Optimized Signal Weight Changes During Boom**

Figure 21: **Optimized Signal Weight Changes During Recession**

Taking a closer look at the signal composition, Figure 22 highlights the top long and short signals during both recession and non-recession periods, categorized by their classifications. Although the leading two signals in each period remain industry-related, the distribution of other significant signals reflects a broader range of categories. This suggests that firm-specific or cross-sectional factors, such as leverage, profitability, and quality, play a crucial role in explaining variations during different economic conditions.

For the top 20 long portfolios, the recession period shows a greater presence of signals related to leverage and value, alongside industry-based factors. In contrast, the non-recession

period incorporates more signals associated with profitability, momentum, and seasonal effects, indicating a shift in the driving forces behind performance during economic expansions. In the bottom 20 short portfolios, during recessions, signals tied to size, profitability, and value are more frequently observed, reflecting the defensive nature of these characteristics in downturns. During non-recession periods, a wider variety of classifications, including momentum and growth, become prominent, suggesting an opportunistic approach in portfolio construction.

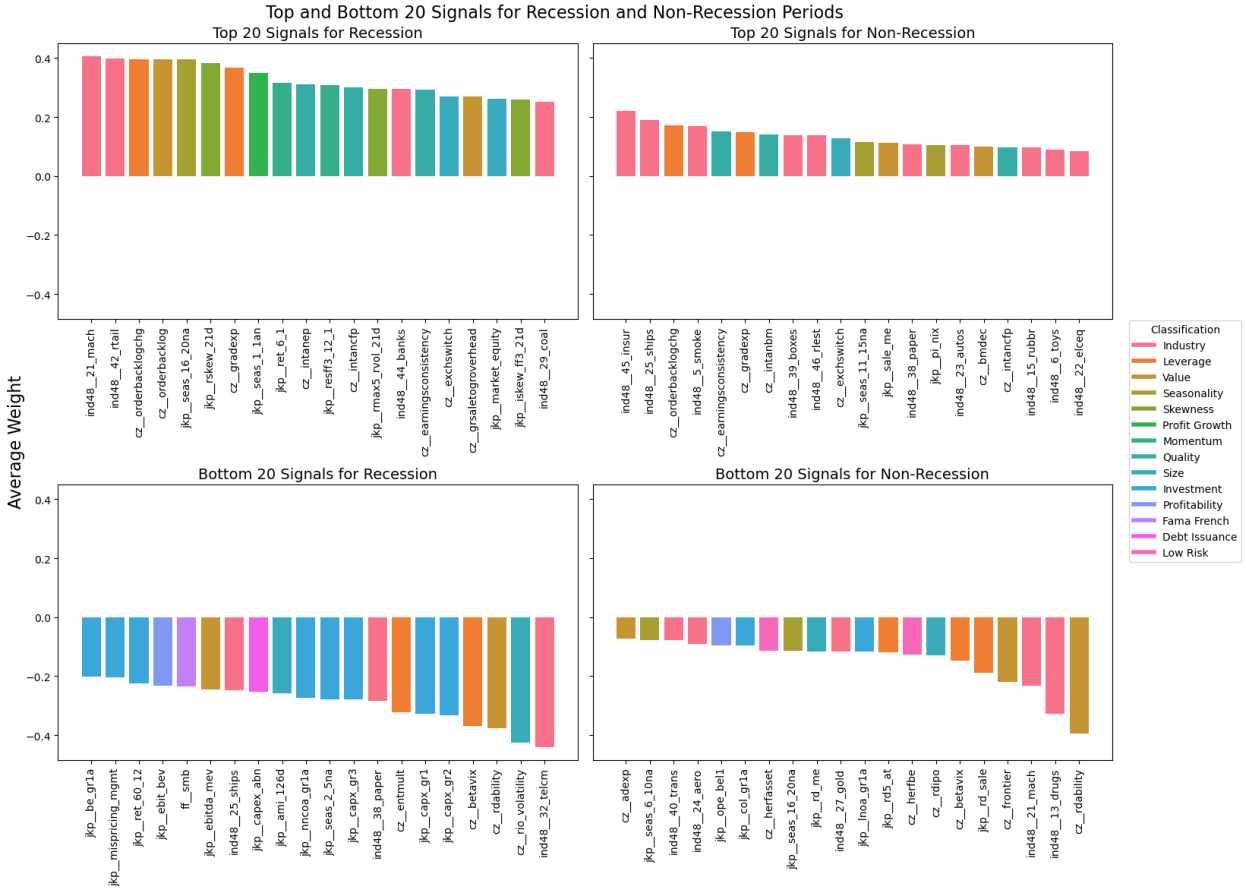

Figure 22: **Top and Bottom 20 Signals for Recession and Non-Recession Periods (Colored by Classification)**

5 Conclusion

This study extends the understanding of systematic risk factors in financial markets by analyzing their stability, economic meaning, and practical implications in both high-frequency and low-frequency data settings. Using Principal Component Analysis (PCA), the results highlight a clear distinction between the systematic risk captured by principal components

and the signals that are most relevant in portfolio optimization.

For high-frequency data, the first three principal components consistently capture over 50% of the total variance, with industry factors—particularly technology, financials, energy, and industrials—emerging as the dominant contributors. This underscores the pivotal role of sector-level dynamics in shaping systematic risk at higher frequencies. However, during portfolio optimization on high frequency data, factors with the highest weights are more firm-specific, diverging from the systematic components and emphasizing the importance of idiosyncratic drivers in practical investment strategies. Conversely, low-frequency data reveals that accounting-related factors are the primary drivers of variance, with the top three principal components demonstrating strong temporal stability.

These findings have several implications. First, the stability of the leading principal components over time reinforces their utility for systematic risk modeling and their potential as benchmarks for tracking macroeconomic and sector-specific risks. Second, the divergence between systematic components and optimized portfolio weights suggests that portfolio managers may need to balance systematic risk considerations with firm-specific attributes to achieve better performance. This trade-off highlights the nuanced relationship between risk decomposition and return optimization, suggesting that traditional factor models may need to be complemented by methods that capture more granular, company-level variations.

6 Further Discussion

Future research could build on these findings by delving deeper into nonlinear approaches, such as neural networks or advanced machine learning models, for latent factor extraction. These methods have the potential to uncover intricate patterns and pricing anomalies that linear models may overlook. Additionally, studying the interplay between high-frequency and low-frequency factors could provide new insights into how these dimensions collectively influence asset pricing and portfolio optimization. A focused analysis of high-volatility events, characterized by significant spikes in the VIX index, could further illuminate their effects on the stability and structure of latent factor spaces.

Practical applications of these findings are also critical for advancing asset management. Integrating systematic and idiosyncratic factors into dynamic portfolio allocation models could improve their robustness during volatile market conditions. Bridging theoretical insights with practical tools will enhance the relevance of factor-based models in real-world investment scenarios, paving the way for more effective strategies in navigating complex financial landscapes.

References

Aleti, S. (2022, January). The high-frequency factor zoo. Available at SSRN: `https://ssrn.com/abstract=4021620` or `http://dx.doi.org/10.2139/ssrn.4021620`.

Bai, J. and S. Ng (2008). Large dimensional factor analysis. *Foundations and Trends(R) in Econometrics 3*(2), 89–163.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009). Realized kernels in practice: trades and quotes. *The Econometrics Journal 12*(3), C1–C32.

Boudoukh, J., Y. Liu, T. J. Moskowitz, and M. P. Richardson (2024, July). Identifying shocks to systematic risk in times of crisis. Working Paper 32693, National Bureau of Economic Research.

Bowman, R. G. (1979). The theoretical relationship between systematic risk and financial (accounting) variables. *The Journal of Finance 34*(3), 617–630.

Campbell, J. Y., C. Polk, and T. Vuolteenaho (2009, 05). Growth or glamour? fundamentals and systematic risk in stock returns. *The Review of Financial Studies 23*(1), 305–344.

Chen, A. Y. and T. Zimmermann (2021, May). Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*. Posted: 12 Jun 2020; Last revised: 17 Jun 2021.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *The Annals of Statistics 44*(1), 219–254.

Green, J., J. R. M. Hand, and X. F. Zhang (2017, 03). The characteristics that provide independent information about average u.s. monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Harvey, C. R., Y. Liu, and H. Zhu (2015, 10). ... and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Hou, K., C. Xue, and L. Zhang (2018, 12). Replicating anomalies. *The Review of Financial Studies 33*(5), 2019–2133.

JENSEN, T. I., B. KELLY, and L. H. PEDERSEN (2023). Is there a replication crisis in finance? *The Journal of Finance 78*(5), 2465–2518.

Kelly, B. T., S. Pruitt, and Y. Su (2017). Characteristics are covariances: A factor model of stock returns. *SSRN Electronic Journal*.

Lettau, M. and M. Pelger (2020). Estimating latent asset-pricing factors. *Journal of Econometrics 218*(1), 1–31.

Ludvigson, S. C. and S. Ng (2009, July). A Factor Analysis of Bond Risk Premia. NBER Working Papers 15188, National Bureau of Economic Research, Inc.

McLEAN, R. D. and J. PONTIFF (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–31.

Pelger, M. (2018). Large-dimensional factor modeling based on high-frequency observations. *SSRN Electronic Journal*.

Pelger, M. (2019, May 12). Understanding systematic risk: A high-frequency approach. *Journal of Finance, Forthcoming*. Available at SSRN: `https://ssrn.com/abstract=2647040` or `http://dx.doi.org/10.2139/ssrn.2647040`.

Pelger, M. and R. Xiong (2020). State-varying factor models of large dimensions. *SSRN Electronic Journal*. Last revised: October 15, 2020.

Stock, J. H. and M. Watson (2006). Forecasting with many predictors. Volume 1, Chapter 10, pp. 515–554. Elsevier.

Zivot, E. and J. Wang (2003). *Factor Models for Asset Returns*, pp. 543–589. New York, NY: Springer New York.