writeup_doc

Common Factors in NYSE Stock Returns (2005–2019): PCA and Factor Analysis with Industry Insights

Executive Summary

We analyze monthly returns for 91 NYSE-listed stocks (January 2005–December 2019) along with the S&P 500 market return. Principal Component Analysis (PCA) and Factor Analysis (FA) both reveal that the first latent dimension represents a broad **market factor**. After orienting the sign, the first principal component correlates strongly with the market return (+0.95).

Industry-level patterns are pronounced. **Finance, Insurance, and Real Estate (H)** and **Construction (C)** show the largest absolute loadings on the first factor, while **Retail (G)** and **Services (I)** load more weakly. Factor Analysis confirms that Finance has the lowest mean uniqueness (~0.47), indicating a higher proportion of systematic (non-diversifiable) variance, whereas Retail and Services are primarily driven by idiosyncratic risk.

For investors seeking exposure aligned with overall market movements, Finance-sector stocks are most suitable. We specifically recommend **H88608, H90004, H89994, H88291, and H83218**. A complementary clustering analysis at the industry level provides further support for these findings.

Data and Preparation

We analyze *SampleK.csv* (91 stocks, labeled by industry + PERMNO) and *Market.csv* (S&P 500 returns). Dates were aligned to monthly frequency, with no missing data. Returns were standardized for comparability. Exploratory summaries and correlations showed strong cross-stock dependence, supporting the use of dimension reduction.

Methods

1. **Principal Component Analysis (PCA)**

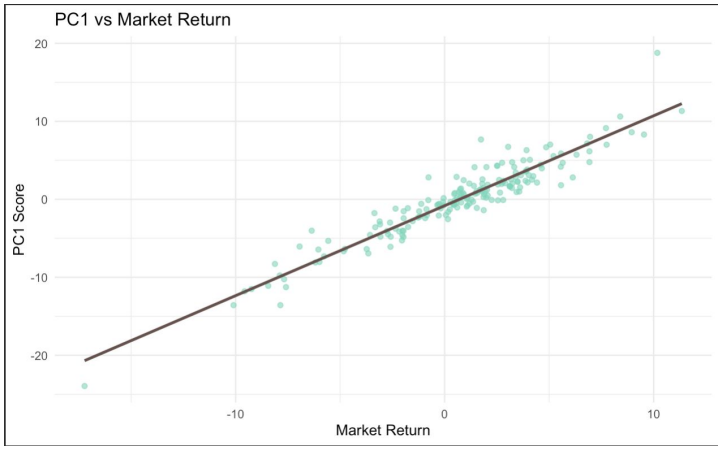
PCA was applied to the standardized stock return matrix to identify latent components. Both the scree plot and the Kaiser criterion were examined. Setting a lower bound for uniqueness at 0.05, the Kaiser rule also suggested retaining two components, consistent with the scree plot, as capturing meaningful common variation. PC1 was oriented to correlate positively with the market return, and industry-average loadings were computed to compare sector exposures.

2. **Factor Analysis (FA)**
   FA was estimated via maximum likelihood with varimax rotation, guided by PCA and model fit. A uniqueness bound of 0.05 ensured convergence. Industry averages of loadings and uniqueness distinguished systematic from idiosyncratic risk.

3. **Stock Selection**
   Stocks were ranked by PC1 loading magnitude and market correlation. The top five stocks were chosen as most sensitive to market movements.

Results

1.The **first principal component (PC1)** is **highly aligned with the S&P500 market return**, with a **correlation of 0.95**. This means that **around 90% of the variation in PC1 can be explained by market movements**. The **scatterplot of PC1 scores against market returns** shows a **clear positive linear relationship**, confirming that **PC1 effectively represents the market factor**. Industries such as **Finance** and **Construction** load most strongly on **PC1**, reflecting their **high systematic exposure**, while sectors like **Retail** and **Services** retain **higher uniqueness**, indicating a **larger idiosyncratic component**.



2. **Construction (C)** exhibits the **highest loading on the first latent factor**, demonstrating that it is the industry most sensitive to **broad market-wide variation**, with **Finance, Insurance and Real Estate (H)** and **Transportation (E)** also showing substantial exposure. However, the analysis shifts when examining the **second latent factor**, where **Finance, Insurance and Real Estate (H)** records the **strongest loading**, followed by **Wholesale Trade (F)** and **Mining (B)**. This indicates that while **Construction dominates the primary, market-like factor**, **Finance is the key driver of the secondary dimension of systematic variation**, likely reflecting **credit- or sector-specific financial influences** distinct from the **general market**. **This highlights the importance of considering multiple latent factors in factor analysis**, as **different industries emerge as dominant depending on the dimension of systematic variation under review.**

Mean Factor Loadings by Industry

Industry	Mean F1	Mean F2
B - Mining	0.349	0.250
C - Construction	0.489	0.249
D - Manufacturing	0.333	0.166
E - Transport & Utilities	0.424	0.211
F - Wholesale Trade	0.350	0.288
G - Retail Trade	0.273	0.215
H - Finance/Insurance/RE	0.435	0.367
I - Services	0.350	0.124

3. From the factor loadings, it is evident that **all industry groups have positive loadings on the first latent factor**, meaning they generally **move in the same direction as the dominant market factor**, with **Construction (C)** and **Finance, Insurance and Real Estate (H)** showing particularly strong alignment. On the **second factor**, however, there is more variation: while **Finance (H)**, **Wholesale Trade (F)**, and **Mining (B)** load positively, some industries such as **Retail (G)** and **Services (I)** have relatively **small or weaker loadings**, suggesting **lower sensitivity or idiosyncratic movement**. Overall, the **heterogeneity across industry groups** indicates that while most sectors are jointly exposed to **broad market forces**, **certain industries retain more unique or sector-specific risk components**, highlighting the importance of considering **multiple factors** to capture differences in **systematic versus idiosyncratic variation**.

4. From the uniqueness results, **Finance, Insurance and Real Estate (H)** shows the **lowest mean (0.51) and median (0.46) uniqueness**, indicating that a larger portion of its variation is explained by common factors rather than idiosyncratic risk. This implies that **Finance is the industry most exposed to systematic risk**, as its returns move more closely with market-wide influences that cannot be diversified away. By contrast, industries such as **Manufacturing (D)**, **Retail (G)**, and **Services (I)** display the **highest uniqueness values (above 0.82)**, suggesting that their variation is more idiosyncratic and less tied to systematic market factors. Overall, the results highlight that **systematic risk is concentrated in the Finance sector**, while other industries retain a greater degree of diversifiable, industry-specific variation.

Industry-level Uniqueness (Mean and Median)

Industry	Mean uniqueness	Median uniqueness
H - Finance/Insurance/RE	0.508	0.460
C - Construction	0.686	0.751
E - Transport & Utilities	0.725	0.707
F - Wholesale Trade	0.779	0.763
B - Mining	0.815	0.815
I - Services	0.824	0.848
G - Retail Trade	0.831	0.853
D - Manufacturing	0.836	0.859

5. If the investor's goal is to hold stocks that **move closely with the market**, the selection should be based on **high correlations with the market (cor_Market)** and **large PC1 loadings**, since PC1 effectively represents the market factor. Among the available stocks in **Finance, Insurance and Real Estate (H)**, the five strongest candidates are:

- **H90004** (cor_Market = 0.97, PC1 = 0.1854)
- **H88608** (cor_Market = 0.92, PC1 = 0.1905)
- **H88291** (cor_Market = 0.93, PC1 = 0.1750)
- **H89994** (cor_Market = 0.88, PC1 = 0.1760)
- **H83218** (cor_Market = 0.78, PC1 = 0.1545)

These stocks show the **highest alignment with the market factor**, making them the most suitable for an investor seeking exposure to systematic risk. Notably, **H90004** and **H88608**

stand out as the strongest choices, given their combination of **high factor loadings** and **very strong correlation with market returns**.

Top 5 Stocks by Absolute PC1 Loading and Market Correlation

StockID	Industry	PC1	cor_Market	abs_PC1
H88608	H	0.191	0.921	0.191
H90004	H	0.185	0.969	0.185
H89994	H	0.176	0.881	0.176
H88291	H	0.175	0.927	0.175
H83218	H	0.154	0.779	0.154

6. The biplot of industry-average factor loadings highlights clear differences in systematic exposure across sectors. **Construction (C)** has the strongest loading on **Factor 1**, indicating it is most sensitive to broad market-wide movements. In contrast, **Finance, Insurance and Real Estate (H)** shows the highest loading on **Factor 2**, suggesting it is driven by an additional source of systematic variation, likely tied to financial or credit-related shocks. **Wholesale Trade (F)** and **Mining (B)** also load relatively strongly on Factor 2, whereas **Services (I)** and **Manufacturing (D)** exhibit weaker secondary loadings, reflecting greater idiosyncratic components. The **closeness of vectors** for some industries, such as **Construction (C) and Transportation (E)**, indicates similar response patterns to the latent factors. Overall, the plot shows that while all industries move broadly with the market (positive loadings), **heterogeneity in factor exposures** reveals that different sectors are influenced to varying degrees by market-wide versus sector-specific dynamics.