# PCA for Portfolio Risk: Evidence from 20 US Equities

By Prince Grant Stalen P. Oncada

September 7, 2025

## Abstract

This study applies Principal Component Analysis (PCA) to daily returns of 20 large-cap US equities from 2015 to 2025 to investigate how a small number of latent factors capture the majority of portfolio risk. We analyze correlations, variance explained, scree plots, loadings, and factor returns to evaluate the effectiveness of dimensionality reduction. Results show that the first three principal components account for nearly 65% of total variance: PC1 represents a broad market factor, PC2 distinguishes technology from industrial and energy exposures, and PC3 reflects idiosyncratic risk dominated by Tesla and Nvidia. A covariance reconstruction test confirms that three to five components are sufficient to approximate the full covariance matrix with minimal error. Applying a Varimax rotation further enhances interpretability by producing sector-aligned factors. These findings demonstrate how PCA can uncover hidden risk drivers, reduce noise in risk estimation, and provide a foundation for factor-based portfolio modeling and optimization.


## 1. Introduction

Financial markets consist of portfolios containing many correlated assets. Price movements in technology, financial, and energy stocks often occur together, making it challenging to disentangle the true sources of portfolio risk. Principal Component Analysis (PCA) provides a systematic method to reduce such complexity by extracting a small number of uncorrelated "factors" that explain most of the variance in returns.

In quantitative finance, PCA is widely applied for three main purposes:

- Risk modeling: compressing large covariance matrices into lower-rank approximations that are more stable and less noisy.
- Factor analysis: uncovering latent drivers such as broad market risk, sector tilts, or single-stock dominance.
- Stress testing: evaluating portfolio exposure to the most influential risk factors.

This project applies PCA to a basket of 20 large-cap US equities across multiple sectors to address three central questions:

1. How much of the overall variance can be explained by the first few components?
2. Do the extracted principal components correspond to interpretable economic features, such as sectoral distinctions?
3. Does applying a rotation method improve factor interpretability without sacrificing variance explained?

By answering these questions, we aim to demonstrate PCA's dual role as both a statistical tool for dimensionality reduction and a practical technique for portfolio risk management.

## 2. Data

We collected daily closing prices for 20 large-cap US equities across major sectors using Yahoo Finance, covering the period of January 2015 to January 2025 (~2,500 trading days). This ten-year horizon ensures sufficient observations for stable estimation of covariances and principal components.

The sample includes a diversified cross-section of sectors:

- Technology/Communication: AAPL, MSFT, AMZN, GOOGL, META, NVDA, NFLX
- Financials/Payments: JPM, GS, V, MA
- Industrials/Energy: BA, GE, CAT, XOM, CVX
- Consumer/Others: DIS, IBM, WMT, TSLA

Daily returns were computed as percentage changes in closing prices:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Where P*t* is the close on day *t*.

To ensure comparability across assets with different levels of volatility, returns were standardized (z-scored) prior to analysis. This transformation aligns with PCA assumptions, giving each stock equal weight in the factor extraction process.

## 3. Methodology

We applied PCA to the standardized return matrix of 20 US equities. The procedure consisted of the following steps, each designed to progressively reveal the latent structure of portfolio risk:

## 3.1 Correlation Matrix

We first computed the pairwise correlation matrix of daily returns. The heatmap visualization provided a preliminary view of clustering behavior (e.g., technology stocks co-moving strongly, energy stocks forming their own block), motivating the need for dimensionality reduction.

## 3.2 Principal Component Extraction

PCA was then fitted to the standardized return matrix. Eigenvalues and explained variance ratios were obtained to quantify the proportion of total variance captured by each component. These results were visualized using:

- A bar plot of variance explained by each component.
- A cumulative variance curve showing how variance accumulates as more PCs are added.

## 3.3 Scree Plot and Retention Criteria

Eigenvalues were plotted in descending order to produce a scree plot. Two criteria guided component retention:

- The elbow rule, where the curve flattens after the first few components.
- The Kaiser criterion, which retains components with eigenvalues greater than one (indicating explanatory power exceeding that of a single standardized variable).

## 3.4 Factor Loadings (Heatmap)

We computed factor loadings (the correlation between each stock and each principal component). These were analyzed via:

- A heatmap of the first five PCs to identify dominant stock-factor relationships.
- A scatterplot of PC1 vs PC2 loadings, which positioned each stock according to its contribution to the two most important factors. This step enabled economic interpretations, e.g., whether PCs aligned with sector themes.

## 3.5 Factor Loadings (Scatter: PC1 vs PC2)

To aid interpretation, we also plotted stocks in a two-dimensional loading space defined by PC1 and PC2. This scatterplot allowed us to detect sector-based groupings and to assess the economic meaning of the first two components.

### 3.6 Factor Returns (Scores)

The original returns were projected into the PC space, producing a time series of factor returns (also called scores). These synthetic factor portfolios were plotted to visualize how the latent drivers evolved through time, particularly during market stress events.

### 3.7 Covariance Reconstruction

Finally, we evaluated how well a reduced-rank PCA model approximates the full covariance matrix of returns. Low-rank approximations were constructed for $k = 1,\ldots, 20$, and the discrepancy from the full covariance matrix was measured using the Frobenius norm. This analysis quantified the trade-off between dimensionality reduction and accuracy in capturing portfolio risk structure.

### 3.8 Varimax Rotation

To enhance interpretability, we applied an orthogonal Varimax rotation to the first three PCs. This procedure redistributes loadings across factors, producing sharper distinctions between groups of stocks (e.g., separating technology/growth from industrials/energy). Importantly, Varimax rotation does not change the total variance explained by the components; it only alters the orientation of the factors to make them more economically interpretable.

## 4. Results

### 4.1 Correlation Structure

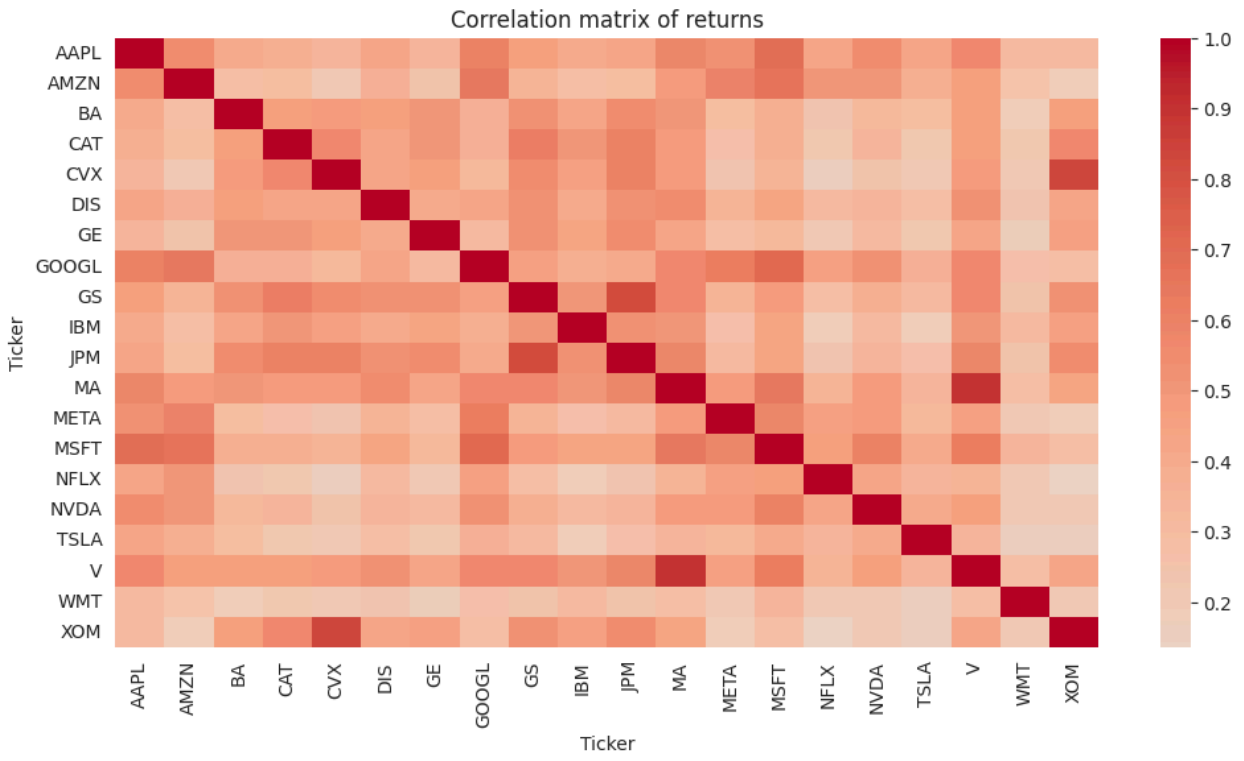


*Figure 1: Correlation matrix of daily returns for 20 US equities (2015–2025). Red indicates strong positive correlation; lighter colors indicate weaker relationships.*

To establish the degree of dependence among assets, we computed the correlation matrix of daily returns across the 20 equities. The heatmap in Figure 1 displays pairwise correlation coefficients, ranging from -1 (perfect negative) to + 1 (perfect positive).

Overall, the results reveal strong positive correlation across most equities, consistent with the idea that large-cap US stocks share broad market exposure. Technology companies such as Apple (AAPL), Microsoft (MSFT), Nvidia (NVDA), and Tesla (TSLA) exhibited particularly high correlations with one another, reflecting their common sensitivity to growth and innovation cycles. Energy firms (ExxonMobil, Chevron) formed another cluster, while consumer and industrial stocks (Walmart, Caterpillar, Boeing, GE) showed weaker but still positive correlations.

These clusters suggest that despite having 20 distinct securities, much of the return variation is shared. This redundancy motivates the use of PCA: rather than modeling all pairwise correlations, PCA compresses the dataset into a small number of orthogonal components that capture the bulk of the co-movements.

## 4.2 Variance Explained by Principal Components

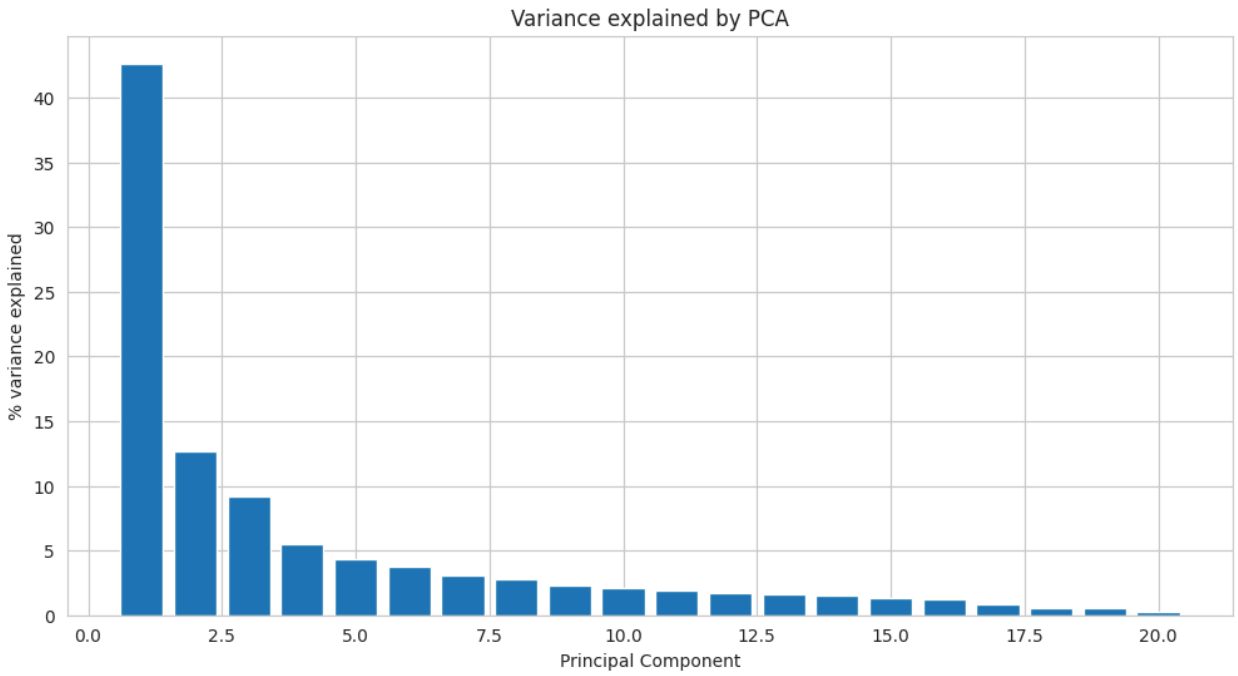


*Figure 2: Variance explained by each principal component. The first three PCs dominate, together accounting for nearly two-thirds of total variance.*

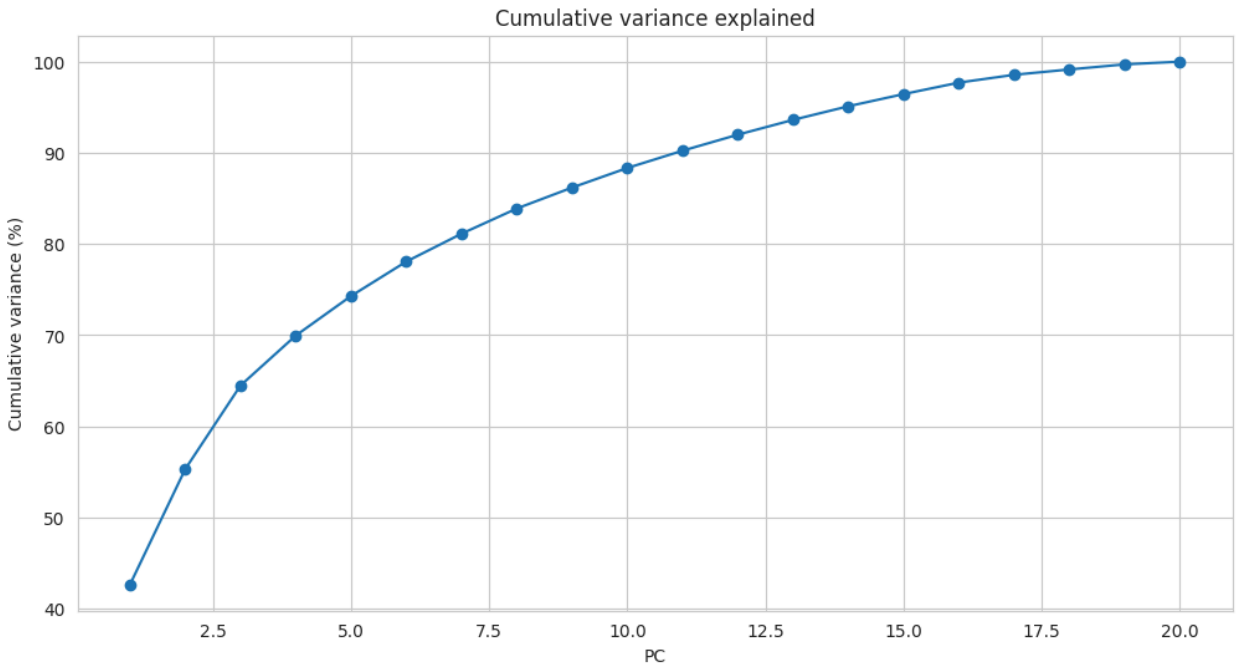


*Figure 3: Cumulative variance explained. The first five PCs capture nearly 80% of variance, after which additional PCs contribute little.*

After fitting PCA on the standardized returns of the 20 equities, we examined how much variance each component captured. Figure 2 shows the proportion of variance explained by each principal component, while Figure 3 plots the cumulative variance explained.

The results highlight the dominance of the first few components:

- PC1 alone explained approximately 43% of the total variance, acting as a broad market factor that drives co-movement across nearly all stocks.
- PC2 explained about 13%, and PC3 about 9%. Together, the top three components captured nearly 65% of the total variance.
- Adding PC4 and PC5 increased the cumulative variance explained to roughly 75-80%. Beyond this point, each additional component contributed only marginally (less than 5% each).

The cumulative curve confirms that a small subset of factors suffices to summarize the majority of portfolio risk. Instead of modelling all 20 correlated assets, risk can be effectively summarized by 3–5 principal components without significant loss of information.

## 4.3 Scree Plot

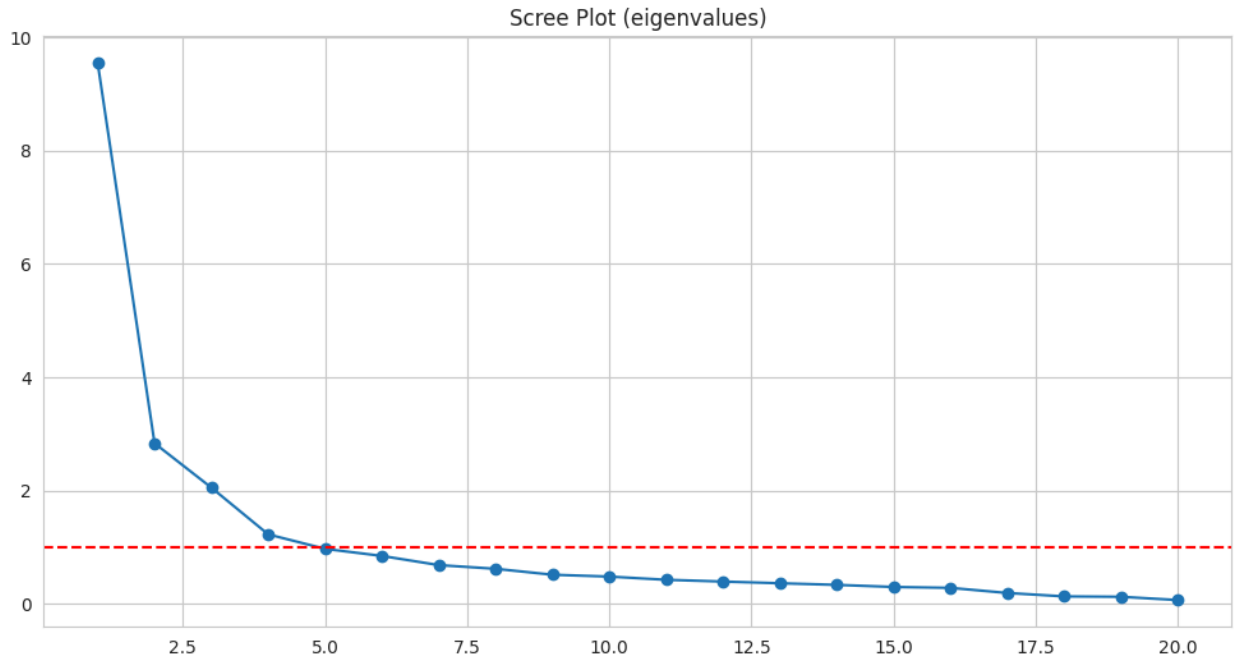


*Figure 4: Scree plot of eigenvalues for 20 principal components. The first three components stand out strongly, while later components fall below the Kaiser threshold, indicating diminishing importance.*

To further assess dimensionality, we examined the eigenvalues of each principal component using a scree plot (Figure 4). The eigenvalues represent the amount of variance explained by

each component in the original scale. A red dashed line at y = 1 marks the Kaiser criterion, which recommends retaining components with eigenvalues greater than one.

The scree plot reveals a steep decline after the first three components, with eigenvalues falling below one from the fifth component onward. Specifically:

- PC1 had an eigenvalue near 9.6, far above the cutoff.
- PC2 and PC3 also exceeded the threshold, with eigenvalues around 2.8 and 2.0, respectively.
- By PC5, eigenvalues dropped below 1, suggesting limited incremental explanatory power.

The "elbow" of the curve occurs around PC3–PC4, where the slope flattens and subsequent components contribute little variance. Both the elbow rule and the Kaiser criterion point to retaining the first three to four components as sufficient for summarizing the dataset.

## 4.4 Factor Loadings

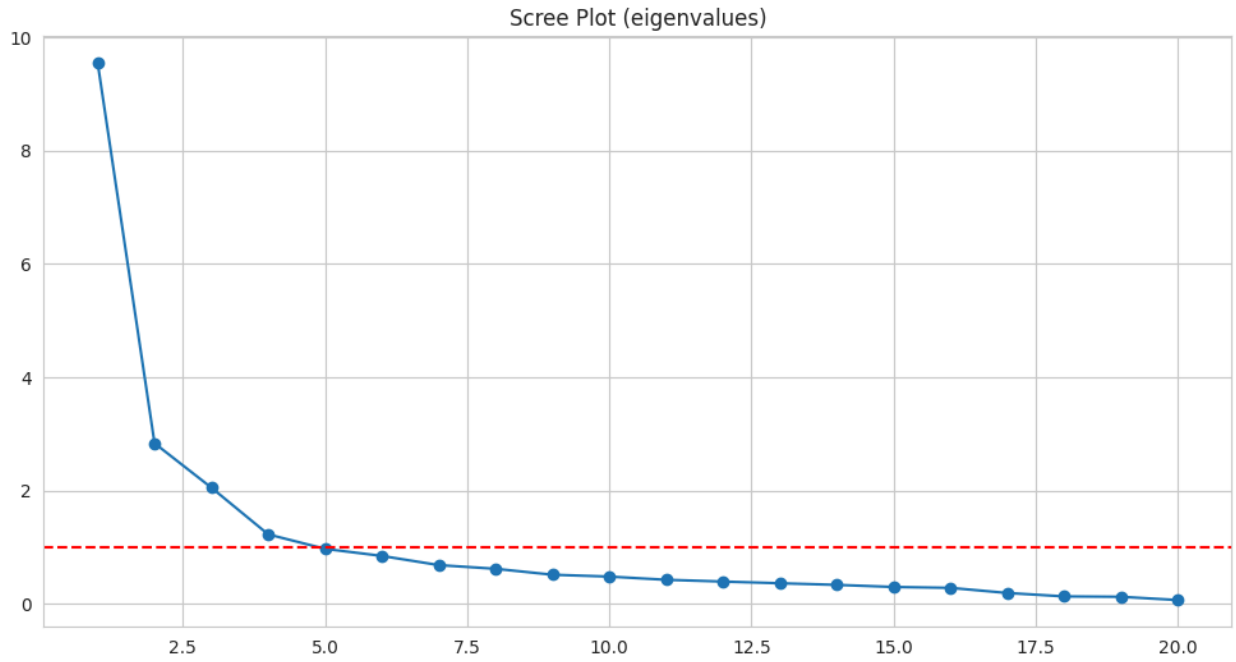


*Figure 5: Factor loadings for the first five principal components. Red indicates positive weights; blue indicates negative weights. Tesla and Nvidia dominate PC3, while PC1 captures broad market co-movement across most equities.*

The factor loadings matrix (Figure 5) shows how each of the 20 equities contributes to the first five principal components. Each value represents the weight, or correlation, between an

individual stock's return and principal component. Positive loadings indicate that the stock moves in the same direction as the factor, while negative loadings imply opposite movement.

Several patterns emerge:

- PC1 displayed broadly positive loadings across nearly all equities, with the largest contributions from Tesla (0.37), Nvidia (0.36), Netflix (0.26), and Boeing (0.26). This confirms PC1 as a general market factor, driving co-movement across the portfolio.
- PC2 highlighted a split between technology and industrial/energy names. Tesla (0.52) and Amazon (0.16) loaded positively, while Boeing (-0.30), Exxon (-0.29), and IBM (-0.16) loaded negatively, suggesting a tech-versus-industrials factor.
- PC3 was heavily dominated by Tesla (0.76) and Nvidia (0.72), with strong negative loadings for Netflix (-0.31) and Meta (-0.28). This component reflects an idiosyncratic high-growth factor linked to volatility in specific technology leaders.
- PC4 showed concentrated effects, with Nvidia (0.72) and Netflix (-0.66) as key drivers.
- PC5 captured small patterns, including positive contributions from Netflix (0.54) and Nvidia (0.45), alongside negative loadings from Meta (-0.47).

Overall, the loadings confirm that the first few components have clear economic interpretations. PC1 reflects broad market exposure, while PCs 2–3 uncover sectoral contrasts and idiosyncratic growth risks. This demonstrates PCA's ability to transform a noisy correlation matrix into interpretable factors.

## 4.5 Factor Returns

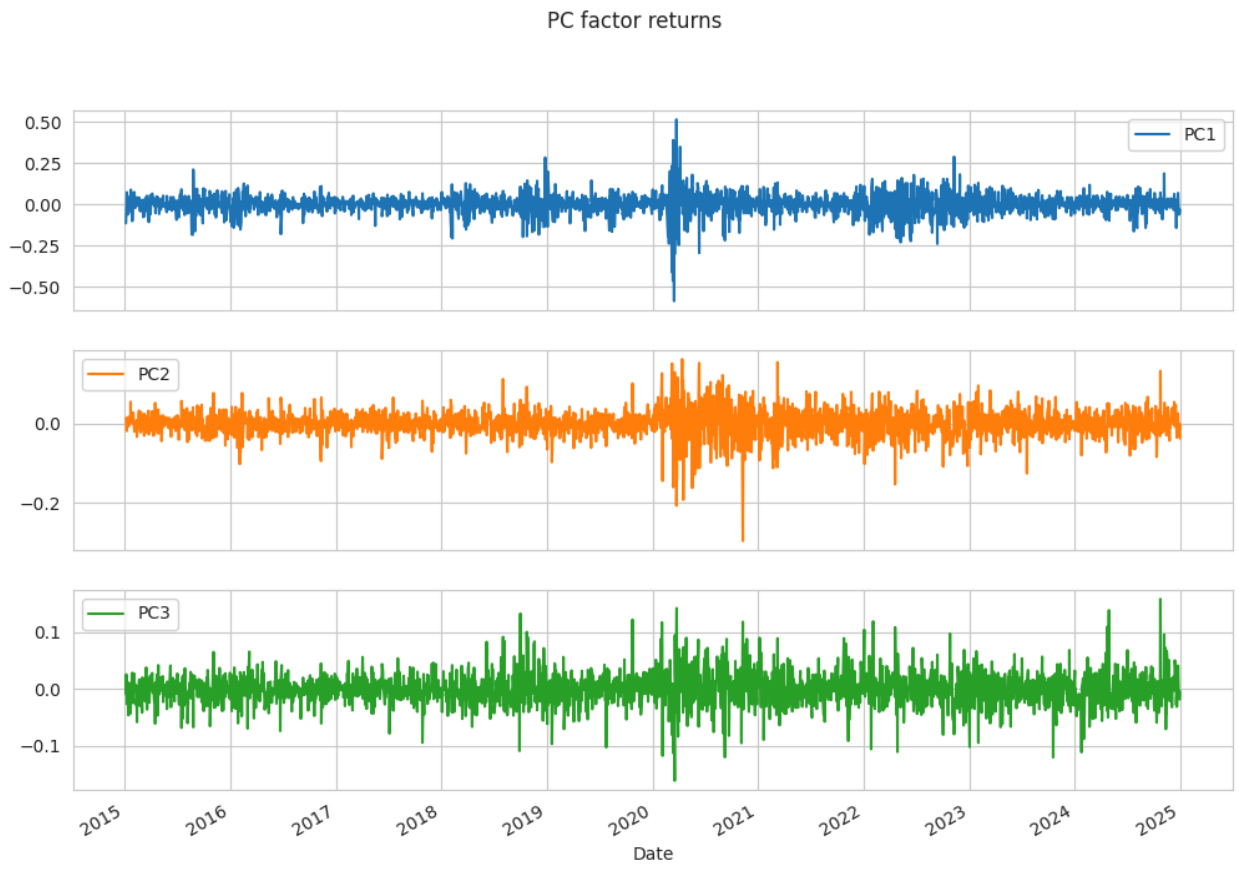


*Figure 6: Time series of factor returns for the first three principal components. PC1 tracks broad market risk, PC2 reflects sector contrasts, and PC3 captures idiosyncratic technology-driven volatility.*

Projecting the daily return onto the principal component axes yielded a new time series of factor returns (Figure 6). These factor returns represent the behavior of the latent drivers identified by PCA:

- PC1 factor returns closely track overall market movements. Large swings, such as during the COVID-19 market shock in early 2020, are clearly visible, reflecting its role as a broad systematic risk factor.
- PC2 factor returns exhibited more volatility than PC1, especially in sector-rotation periods, indicating that this component captures contrasting dynamics between technology and industrial/energy sectors.
- PC3 factor returns were lower in magnitude but more idiosyncratic, with spikes linked to volatility in high-growth stocks such as Tesla and Nvidia. This suggests PC3 reflects stock-specific risk concentrated in a few names rather than broad market effects.

Together, the factor return series demonstrates that PCA not only reduces dimensionality but also creates interpretable synthetic factors. These factors can be analyzed like tradeable portfolios or "hidden indices," offering valuable insight for risk monitoring and strategy development.

## 4.6 PC1 vs PC2 Loadings

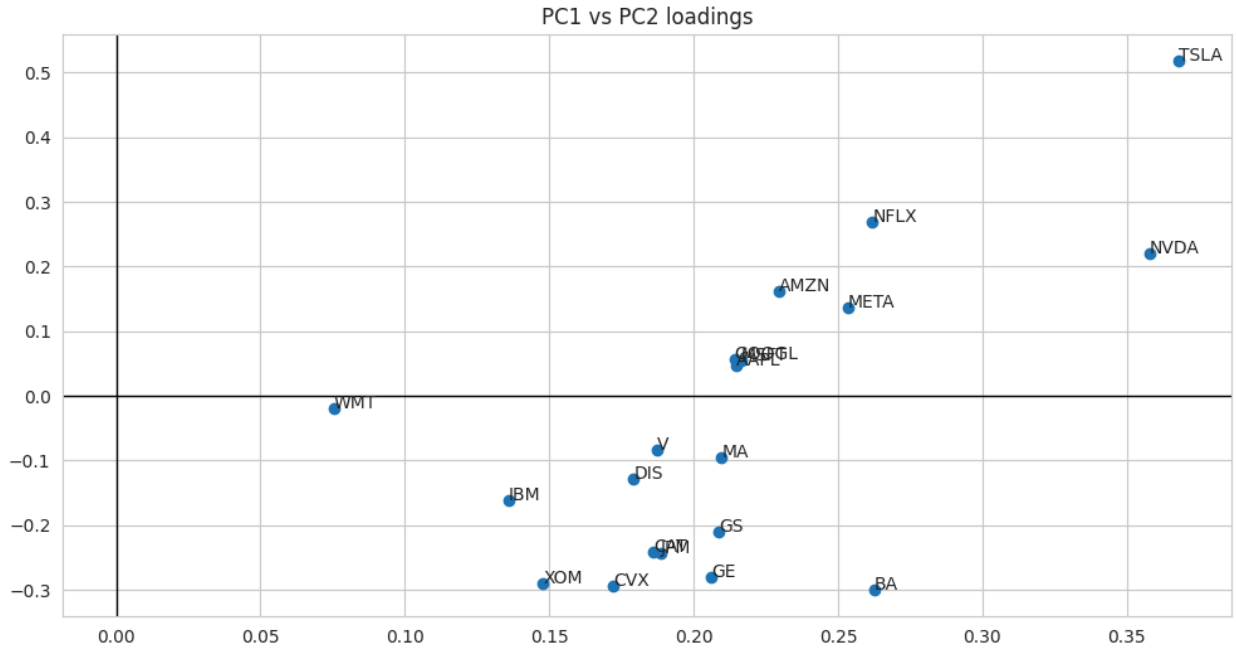


*Figure 7: Scatterplot of PC1 vs PC2 loadings. Growth-oriented technology stocks cluster in the top-right, while industrial and energy stocks cluster in the bottom-left, illustrating sectoral contrasts uncovered by PCA.*

To further interpret the principal components, we plotted each stock's loadings on the first two components in a two-dimensional scatterplot (Figure 7). The x-axis corresponds to loadings on PC1, while the y-axis corresponds to PC2. Each point represents a stock, positioned according to its contribution to the two dominant factors.

The scatter reveals clear clustering:

- Top-right quadrant: High loadings on both PC1 and PC2 were concentrated among technology and growth-oriented stocks such as Tesla, Nvidia, Netflix, Amazon, and Meta. These names are strongly associated with the broad market factor (PC1) while also driving sector-specific contrast captured by PC2.
- Bottom-left quadrant: Negative loadings on both PC1 and PC2 included ExxonMobil, Chevron, IBM, and Boeing, indicating lower correlation with the market factor and stronger association with industrial/energy exposures.

- Center cluster: Several diversified names (e.g., Apple, Google, Visa, Mastercard) fell near the origin, reflecting balanced contributions without extreme tilt toward any single component.

This visualization complements the loadings heatmap by showing how equities group into interpretable factors. PC1 represents a broad market driver, while PC2 highlights the contrast between high-growth technology and traditional industrial/energy sectors.

## 4.7 Covariance Reconstruction

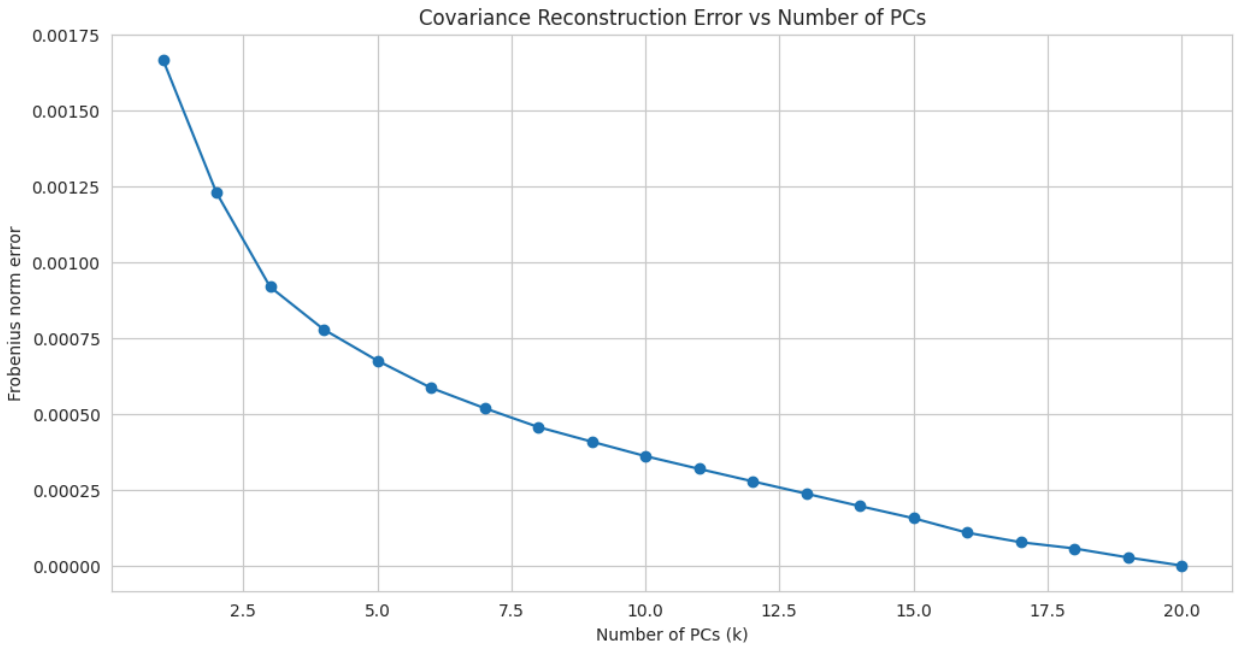


*Figure 8: Frobenius norm reconstruction error of the covariance matrix as a function of the number of principal components. Error declines rapidly with the first few components, flattening beyond five PCs.*

To evaluate how effectively a reduced set of principal components can capture the covariance structure of the portfolio, we compared the full 20x20 covariance matrix with low-rank approximations constructed from the top *k* principal components. The discrepancy between the true and approximated covariance matrices was measured using the Frobenius norm.

Figure 8 plots the reconstruction error as a function of *k,* the number of retained components. The error decreases sharply as more components are added:

- With one component, the error is relatively large, as the approximation relies solely on the broad market factor.

- With three components, the error drops below 0.001, consistent with earlier findings that the top three PCs capture ~65% of variance.
- By five components, the error falls further, with diminishing improvements beyond that point.
- As *k* approaches 20, the error approaches zero, since the approximation converges to the full covariance matrix.

This result confirms that the covariance structure can be accurately summarized by a handful of principal components. For practical applications in risk modeling and portfolio optimization, using the top 3–5 factors provides a strong balance between parsimony and accuracy, while avoiding the noise amplification that arises from modeling all 20 correlated assets.

## 4.8 Varimax Rotation

While unrotated principal components maximize variance explained, they often lack clear economic interpretation because many assets load moderately on multiple components. To address this, we applied a Varimax rotation to the first three PCs. This orthogonal transformation redistributes factor loadings while preserving total variance, making each rotated component easier to interpret.

This orthogonality test confirmed that the rotation matrix was valid, with a negligible error of $||T'T - I||_F \approx 1 \times 10^{-12}$. Total variance was preserved (6.43 unrotated vs 6.43 rotated).

### 4.8.1 Rotated Loadings

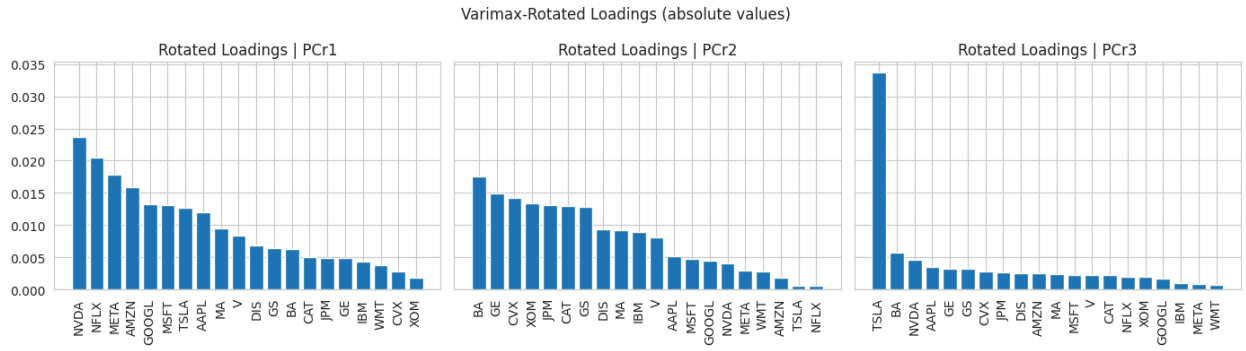


*Figure 9: Varimax-rotated loadings. Tech/growth names dominate PCr1, industrial/energy dominate PCr2, while Tesla dominates PCr3.*

The rotated loadings (Figure 9) revealed sharper distinctions between groups of assets compared to the unrotated factors.

- PCr1 was dominated by Nvidia, Netflix, Meta, Tesla, and Amazon, clearly representing a technology/growth factor.
- PCr2 loaded heavily on Boeing, GE, Exxon, Chevron, and JPMorgan, reflecting an industrial and energy factor.
- PCr3 was driven almost entirely by Tesla, highlighting its unique volatility and outsized influence on portfolio risk.

### 4.8.2 Rotated Scatter

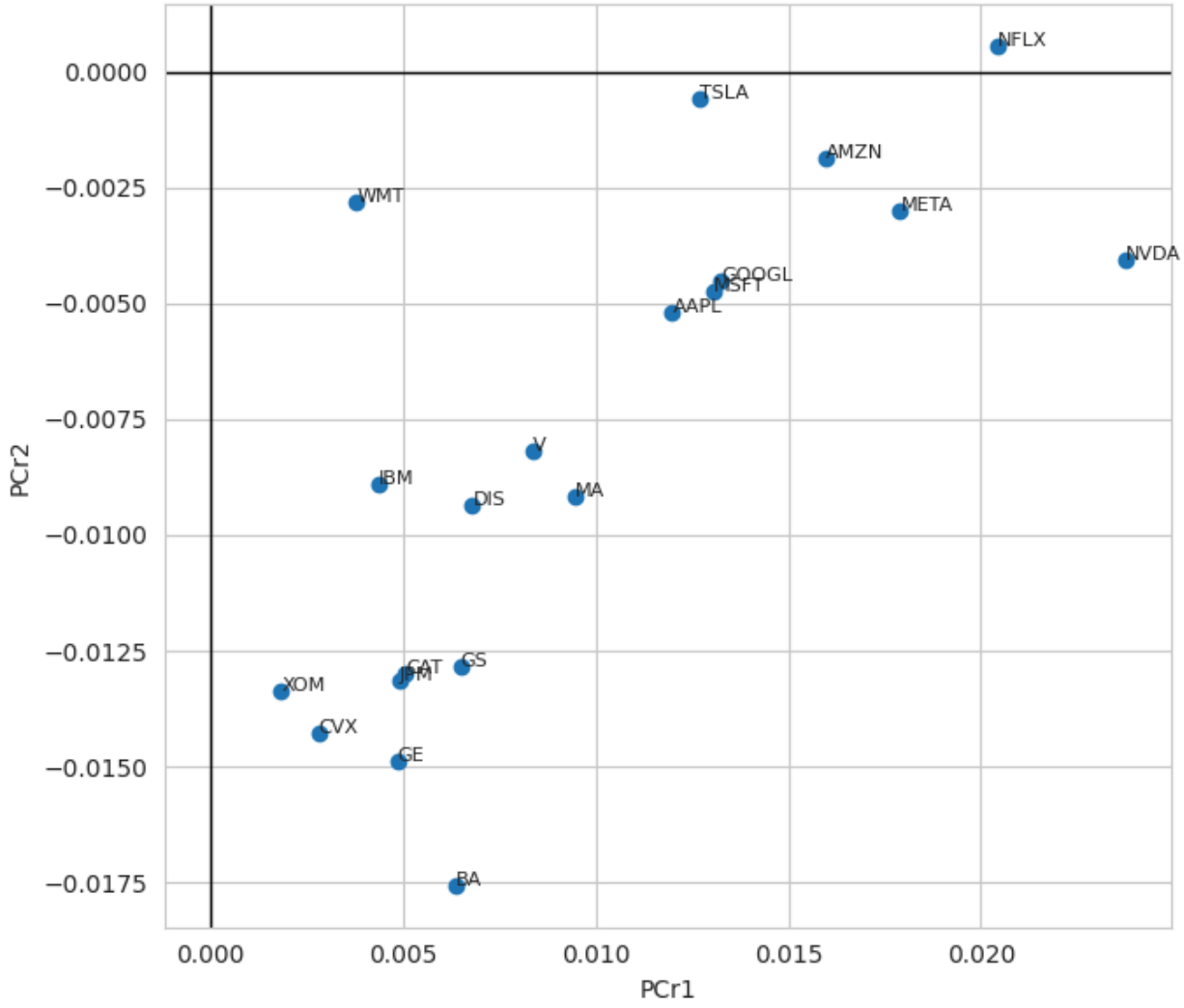


*Figure 10: Scatterplot of rotated PC1 vs PC2 loadings. Rotated factors show clearer sector clustering compared to unrotated PCs.*

The scatterplot of rotated loadings (Figure 10) provided further confirmation of these groupings. After rotation, the sectoral clusters became more distinct than in the unrotated scatter.

Technology names clustered together in the upper-right region, while energy and industrials concentrated in the lower-left.

### 4.8.3 Rotated Factor Returns

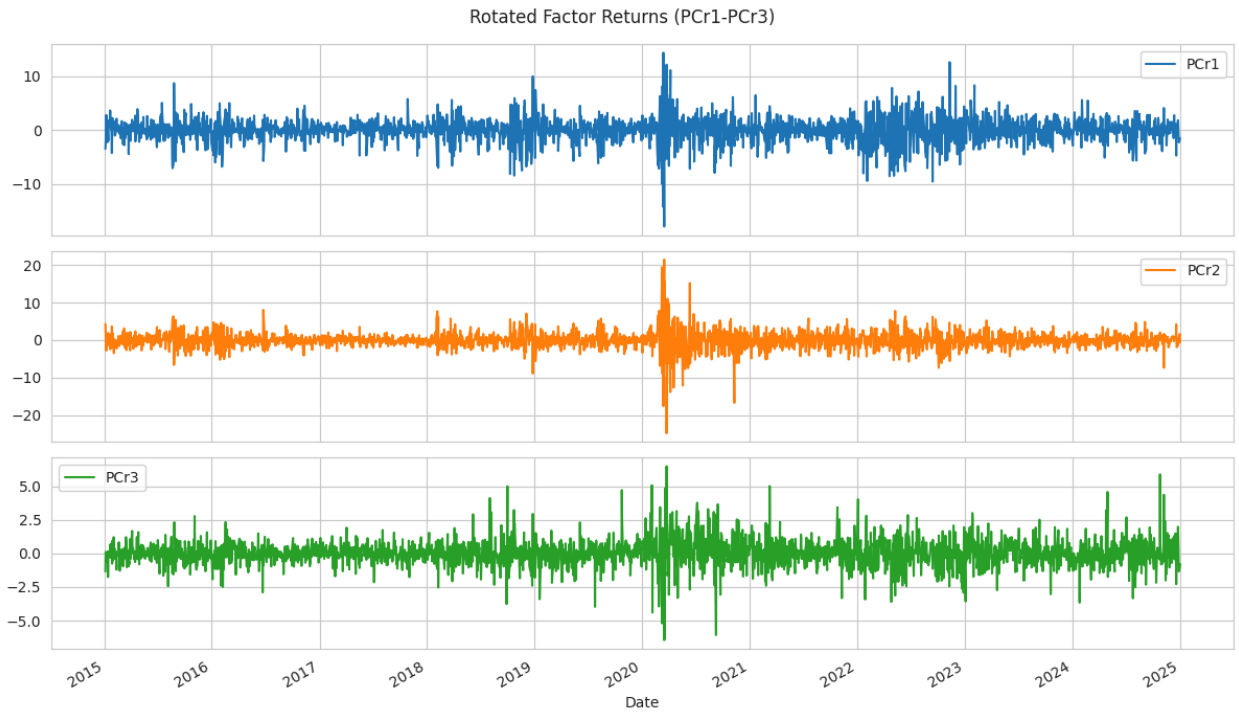


*Figure 11: Time series of rotated factor returns (PCr1–PCr3). Rotation aligns each factor with economically intuitive themes, aiding practical interpretation in portfolio risk analysis.*

Finally, the rotated factor return series (Figure 11) illustrated how these clearer groupings translate into interpretable time dynamics.

- PCr1 returns fluctuated with technology sector cycles.
- PCr2 returns reflected shocks to industrial and energy, including macroeconomics and commodity-driven volatility.
- PCr3 returns captured idiosyncratic Tesla-driven risk, while visible spikes around known Tesla events.

## 5. Discussion

This study demonstrated how Principal Component Analysis (PCA) reduces the complexity of equity portfolios by extracting a small set of latent factors that capture most of the variance in

returns. The findings provide several important insights for both quantitative research and practical risk management.

First, the results confirm the dominance of a broad market factor. PC1 alone explained over 40% of variance and loaded positively on nearly all equities, showing that large-cap US stocks are driven by common macroeconomic forces. This supports the long-standing view that systematic risk cannot be diversified away, as it permeates across sectors.

Second, the analysis uncovered sectoral contrasts that are less obvious in raw correlation matrices. PC2 distinguished between technology/growth names (Tesla, Nvidia, Netflix, Amazon) and industrial/energy stocks (Boeing, Exxon, Chevron). This aligns with economic intuition: while both groups are tied to the overall market, they often respond differently to interest rates, commodity shocks, or business cycles.

Third, the presence of idiosyncratic factors such as Tesla underscores how individual firms can dominate risk at the portfolio level. PC3 was largely a Tesla-specific component, showing that single companies with extreme volatility may behave as independent "factors." For portfolio managers, this highlights the importance of independent factors and identifying concentrated exposures that traditional diversification may overlook.

The Varimax rotation enhanced interpretability without sacrificing variance explained. By rotating the factor axes, the loadings aligned more clearly with economic themes: one factor for tech/growth, one for industrials/energy, and one dominated by Tesla. This illustrated a broader lesson: while raw PCA is useful for noise reduction, rotation is often necessary to bridge the gap between statistical output and economic meaning.

Finally, the covariance reconstruction test confirmed that a low-rank factor model can approximate the full risk structure with minimal error. Using just three to five components captured most of the covariance dynamics, demonstrating that dimensionality reduction not only simplifies interpretation but also provides stable inputs for portfolio optimization.

Together, these findings reinforce PCA's role as both an exploratory and practical tool. For quants, it provides a foundation for factor models, stress testing, and portfolio optimization. For practitioners, it translates into clearer insights about which economic forces matter most — and how hidden concentrations can drive portfolio risk.

## 6. Conclusion

This project applied Principal Component Analysis (PCA) to daily returns of 20 large-cap US equities (2015-2025) to investigate how dimensionality reduction can uncover hidden drivers of portfolio risk.

The results showed that:

- Three components captured nearly 65% of total variance, with PC1 representing the broad market factor, PC2 highlighting a technology-versus-industrial contrast, and PC3 dominated by Tesla.
- Varimax rotation improved interpretability, clearly separating factors into intuitive economic themes (tech/growth, industrials/energy, and Tesla-specific risk).
- Covariance reconstruction confirmed that using only 3–5 components approximates the full covariance matrix with minimal error, balancing simplicity and accuracy.

These findings reinforce PCA as a powerful tool in quantitative finance. It not only compresses high-dimensional risk structures into a handful of orthogonal factors, but — when combined with rotation — also yields interpretable components aligned with economic reality.

Looking ahead, the insights gained here provide a foundation for more advanced applications:

- Portfolio optimization using PCA-reduced covariance matrices (Project 13).
- Factor-based trading strategies, where rotated factors serve as synthetic tradeable indices.
- Risk monitoring, where single-name exposures like Tesla are recognized as independent risk drivers.

In summary, PCA bridges the gap between statistics and market intuition, offering a rigorous yet practical lens for understanding portfolio risk.

# 7. Appendix

## 7.1 Data Source

- Provider: Yahoo Finance
- Frequency: Daily closing prices
- Period: January 2015 – January 2025 (~2,500 trading days)
- Universe (20 US large caps):
    - *Technology/Communication*: AAPL, MSFT, AMZN, GOOGL, META, NVDA, NFLX
    - *Financials/Payments*: JPM, GS, V, MA
    - *Industrials/Energy*: BA, GE, CAT, XOM, CVX

- *Consumer/Other*: DIS, IBM, WMT, TSLA

### 7.2 Return Transformation

- Daily returns were computed as percentage changes:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

- Returns standardized (z-scored) before PCA to ensure comparability across assets.

### 7.3 Methods Summary

- Correlation Matrix: pairwise Pearson correlations, visualized with heatmap.
- PCA Extraction: eigenvalues, explained variance ratios, and cumulative variance.
- Scree Plot: eigenvalues in descending order; elbow rule and Kaiser criterion.
- Factor Loadings: examined via heatmap (PC1–PC5) and scatterplot (PC1 vs PC2).
- Factor Returns: projections of returns into PC space (scores).
- Covariance Reconstruction: approximation error measured with Frobenius norm across $k = 1…20$.
- Varimax Rotation: orthogonal rotation applied to top 3 PCs for interpretability.

### 7.4 Python Packages

- Data Handling: pandas, numpy
- Visualization: matplotlib, seaborn
- Analysis: scikit-learn (PCA), custom varimax function
- Data Retrieval: yfinance

### 7.5 Exported Figures

- corr_matrix.png — correlation heatmap
- pca_variance.png — variance explained by PCs
- pca_cumvar.png — cumulative variance explained
- scree_plot.png — scree plot of eigenvalues
- pca_loadings.png — heatmap of loadings (PC1–PC5)
- pc_scatter.png — PC1 vs PC2 loadings scatterplot

- pc_timeseries.png — factor returns (PC1–PC3)
- cov_recon_error.png — covariance reconstruction error vs number of PCs
- pca_varimax_rotated_loadings.png — rotated loadings barplots
- pca_varimax_pc1_pc2_scatter.png — rotated PC1 vs PC2 loadings scatterplot
- pca_varimax_rotated_scores.png — rotated factor returns