Principal Component Analysis of Daily Returns for Selected S&P 500 Stocks

Pushkar Patil
200674803

Submission Date: 02 December 2025

1 Data Description

1.1 Stocks, Dates, and Data Source

- Stocks (tickers):
  - Palantir (PLTR), Microsoft (MSFT), Nvidia (NVDA), Amazon (AMZN), Apple (AAPL), Netflix (NFLX), Meta (META), Alphabet (GOOGL), Taiwan Semiconductor (TSM), Tesla (TSLA), Oracle (ORCL), Advanced Micro Devices (AMD), International Business Machines (IBM).
- Data frequency: Daily closing prices.
- Start date: 27 November 2023.
- End date: 27 November 2025.
- Data source: Adjusted Close prices downloaded from Yahoo Finance using the `yfinance` Python API.

1.2 Data Preprocessing

The following preprocessing steps are applied:

1. Download Adjusted Close prices for all 13 tickers over the common date range.
2. Align the data across tickers and drop any rows (dates) with missing values to obtain a complete panel.
3. Compute daily log returns for each stock:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right),$$

   where P_t is the Adjusted Close price on day t.
4. For covariance–based PCA, center each return series by subtracting its sample mean so that each column has mean zero.
5. For correlation–based PCA, further standardize each series by its standard deviation so each has variance one.

2 Covariance Matrix of Returns

Let R be the $T \times N$ matrix of daily log returns, where T is the number of days and $N = 13$ is the number of stocks. Let X be the centered return matrix:

$$X_{tj} = R_{tj} - \bar{R}_j,$$

where $\bar{R}_j$ is the sample mean of stock j.

The sample covariance matrix is defined as

$$\Sigma = \frac{1}{T-1} X^\top X.$$

In the code, this is computed with `np.cov(X.T)`. The covariance matrix for 13 x 13 was too big to be displayed here and is therefore exported to a csv.

3 Principal Components and Variance Explained

3.1 Eigen-Decomposition

Principal Component Analysis (PCA) is performed by eigen-decomposing the covariance matrix:

$$\Sigma v_i = \lambda_i v_i, \quad i = 1, \ldots, N,$$

where λ_i are the eigenvalues and v_i the corresponding eigenvectors (loading vectors). The eigenvalues are sorted in descending order:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N.$$

3.2 First Three Principal Components

The first three principal components are defined by the first three eigenvectors:

$$\text{PC}_1 = v_1, \quad \text{PC}_2 = v_2, \quad \text{PC}_3 = v_3.$$

Table 1: Covariance PCA: Variance Explained by First Three Principal Components

Component	Eigenvalue Ratio	% Variance Explained	Cumulative %
PC1	0.4678	46.78%	46.78%
PC2	0.1146	11.46%	58.24%
PC3	0.0996	9.96%	68.20%
Total (PC1–PC3)	—	68.20%	68.20%

The trace of the covariance matrix equals the sum of all eigenvalues:

$$\text{tr}(\Sigma) = 0.00937644 = \sum_{i=1}^{13} \lambda_i,$$

confirming numerical consistency.

Table 2: Correlation PCA: Variance Explained by First Three Principal Components

Component	Eigenvalue Ratio	% Variance Explained	Cumulative %
PC1	0.4439	44.39%	44.39%
PC2	0.0808	8.08%	52.47%
PC3	0.0718	7.18%	59.65%
Total (PC1–PC3)	—	59.64%	59.64%

The trace of the correlation matrix equals its dimension:

$$\mathrm{tr}(R) = 13 = \sum_{i=1}^{13} \lambda_i,$$

which is consistent with diagonal entries normalized to variance one.

3.3 Variance Explained by the First Three Components

The total variance of the data is the sum of all eigenvalues:

$$\mathrm{Var}_{\mathrm{total}} = \sum_{i=1}^{N} \lambda_i.$$

The proportion of variance explained by component i is

$$\mathrm{PVE}_i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}.$$

For the covariance PCA in this project, the first three eigenvalues explain approximately:

- PC1: $\mathrm{PVE}_1 \approx 46.8\%$,
- PC2: $\mathrm{PVE}_2 \approx 11.5\%$,
- PC3: $\mathrm{PVE}_3 \approx 10.0\%$,

for a combined total of about 68.2% of the total return variance.

For the correlation PCA in this project, the first three eigenvalues explain approximately:

- PC1: $\mathrm{PVE}_1 \approx 44.4\%$,
- PC2: $\mathrm{PVE}_2 \approx 8.1\%$,
- PC3: $\mathrm{PVE}_3 \approx 7.2\%$,

for a combined total of about 59.6% of the total return variance.

4 Covariance vs Correlation PCA

4.1 Why Use Both?

There are two natural choices of matrix for PCA on returns:

- **Covariance matrix** Σ:
  - Uses the original scales of each stock's returns.
  - Stocks with higher variance (more volatility) contribute more to the first components.
  - Answers the question: *what linear factors explain most of the actual risk (variance) in my portfolio?*
- **Correlation matrix** R:
  - Each series is standardized to mean 0 and variance 1.
  - All stocks contribute equally in terms of variance.
  - Answers the question: *what linear factors explain the pattern of co-movements, independent of scale?*

In this project I compute PCA in both ways:

1. Covariance PCA: eigen-decomposition of Σ based on centered returns X.
2. Correlation PCA: eigen-decomposition of the correlation matrix of standardized returns Z.

Comparing both helps separate the effect of volatility differences from the pure correlation structure of the stocks.

4.2 How They Differ Here

Empirically:

- In the covariance PCA, the first PC is more heavily influenced by very volatile names (e.g. PLTR and TSLA), because they dominate the raw variance.
- In the correlation PCA, the first PC has more even loadings across the 13 stocks, representing an average standardized "tech sector" factor.
- The variance explained by the first three PCs is slightly higher for the covariance PCA (about 68%) than for the correlation PCA (about 60%), reflecting the influence of high-volatility stocks.

5 Interpretation of the Results

Table 3: Loadings of First Three Principal Components (Covariance PCA)

Ticker	PC1	PC2	PC3
AAPL	0.1451	0.0329	0.1145
AMD	0.3782	-0.3334	0.1497
AMZN	0.2048	-0.0436	0.0519
GOOGL	0.1640	-0.0292	0.1214
IBM	0.0871	-0.0556	-0.0229
META	0.2228	-0.0918	-0.0025
MSFT	0.1443	-0.0623	0.0200
NFLX	0.1596	-0.0669	-0.0415
NVDA	0.3757	-0.3850	-0.0127
ORCL	0.2544	-0.2441	-0.0558
PLTR	0.4460	0.4545	-0.7553
TSLA	0.4204	0.6174	0.6088
TSM	0.2862	-0.2637	0.0083

Table 4: Loadings of First Three Principal Components (Correlation PCA)

Ticker	PC1	PC2	PC3
AAPL	0.2608	-0.4525	-0.1443
AMD	0.2972	0.1533	-0.1948
AMZN	0.3230	-0.2464	0.1527
GOOGL	0.2716	-0.3414	0.0323
IBM	0.1615	-0.0043	-0.9022
META	0.2994	-0.1623	0.0901
MSFT	0.3271	-0.0946	0.2376
NFLX	0.2416	0.0620	0.1445
NVDA	0.3201	0.3534	0.0693
ORCL	0.2337	0.4984	0.0821
PLTR	0.2570	0.0922	-0.0189
TSLA	0.2554	-0.2129	0.0672
TSM	0.3100	0.3582	-0.0329

5.1 Interpretation of the First Three Covariance PCs

From the covariance PCA:

- **PC1 (Covariance):** All loadings are positive and relatively large for high-volatility tech/-growth stocks (PLTR, TSLA, AMD, NVDA, etc.). This component behaves like a broad "tech/growth market factor". On days when PC1 has a large positive score, most stocks in the portfolio tend to have above-average positive returns; on large negative-score days, they tend to move down together. This factor alone explains almost half of the total variance.

- **PC2 (Covariance):** Loadings have mixed signs, with PLTR and TSLA on one side, and stocks like NVDA, ORCL, TSM and IBM on the other. This PC captures a contrast between more speculative/higher-beta names and relatively more established or defensive tech names. It explains a further $\approx 11.5\%$ of the variance.

- **PC3 (Covariance):** Again shows mixed signs and separates IBM and some stable names from PLTR/TSLA and others. PC3 represents another cross-sectional pattern of relative performance within the portfolio and adds about 10% of explained variance.

Together, the first three covariance PCs summarize the main sources of portfolio risk: one dominant broad factor plus two smaller, orthogonal style/stock-selection factors.

5.2 Interpretation of the First Three Correlation PCs

From the correlation PCA:

- **PC1 (Correlation):** Loadings are all positive and fairly similar in magnitude across stocks. This component is a clean "common tech sector" factor on standardized returns. It captures days when most tech stocks move in the same direction relative to their own volatility.

- **PC2 and PC3 (Correlation):** These components have more mixed signs and isolate relative differences between subgroups of stocks (e.g. IBM and some legacy/enterprise names versus high-growth names). They capture patterns of outperformance and underperformance within the sector rather than overall market moves.

5.3 Practical Significance

The first three principal components are significant because:

- They jointly explain a large fraction (around 60–70%) of the total variance in this 13-stock portfolio.

- PC1 represents the main systematic risk factor that drives most day-to-day variation in portfolio returns.

- PC2 and PC3 capture secondary but meaningful patterns of cross-sectional variation, useful for understanding diversification and for designing hedged or relative-value strategies.

6 Plot of the First Three Components

The time series of the first three covariance-based components are the scores

$$S = XV_{1:3},$$

where $V_{1:3}$ is the $N \times 3$ matrix of the first three eigenvectors. Each column of S is plotted against time on a single graph.

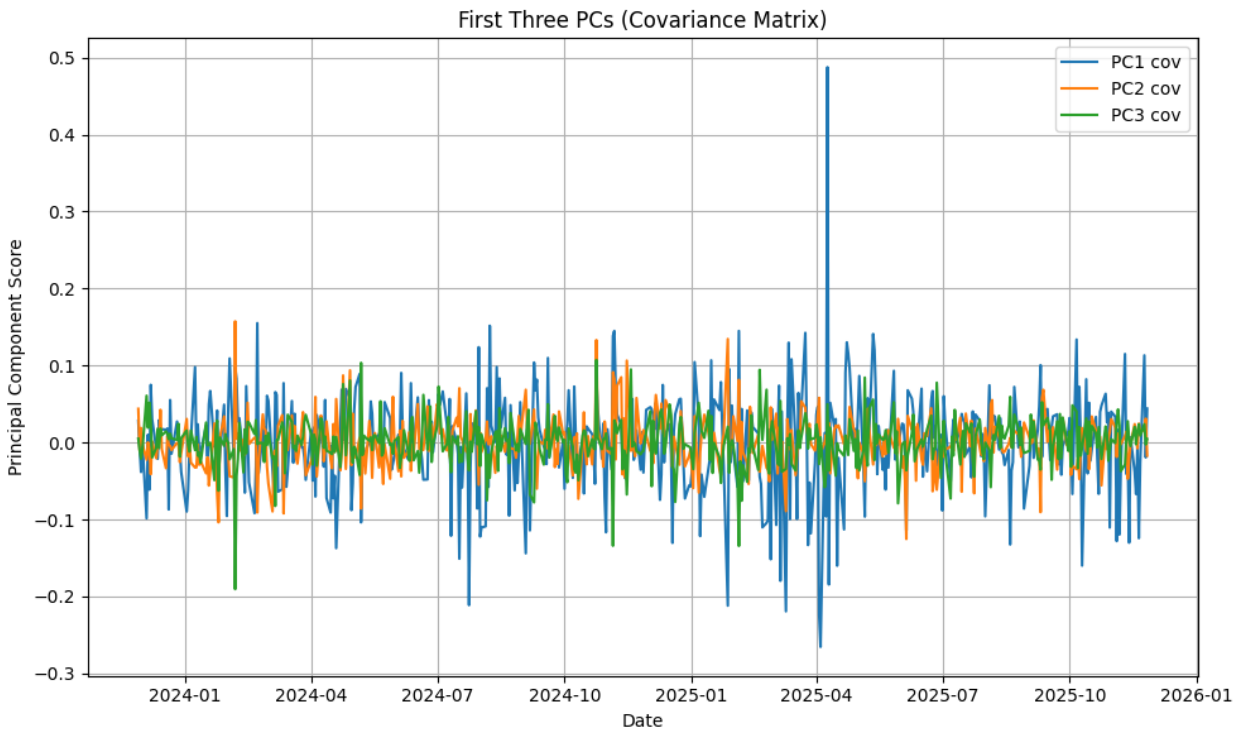

Figure 1: First Three Covariance PCA Component Scores

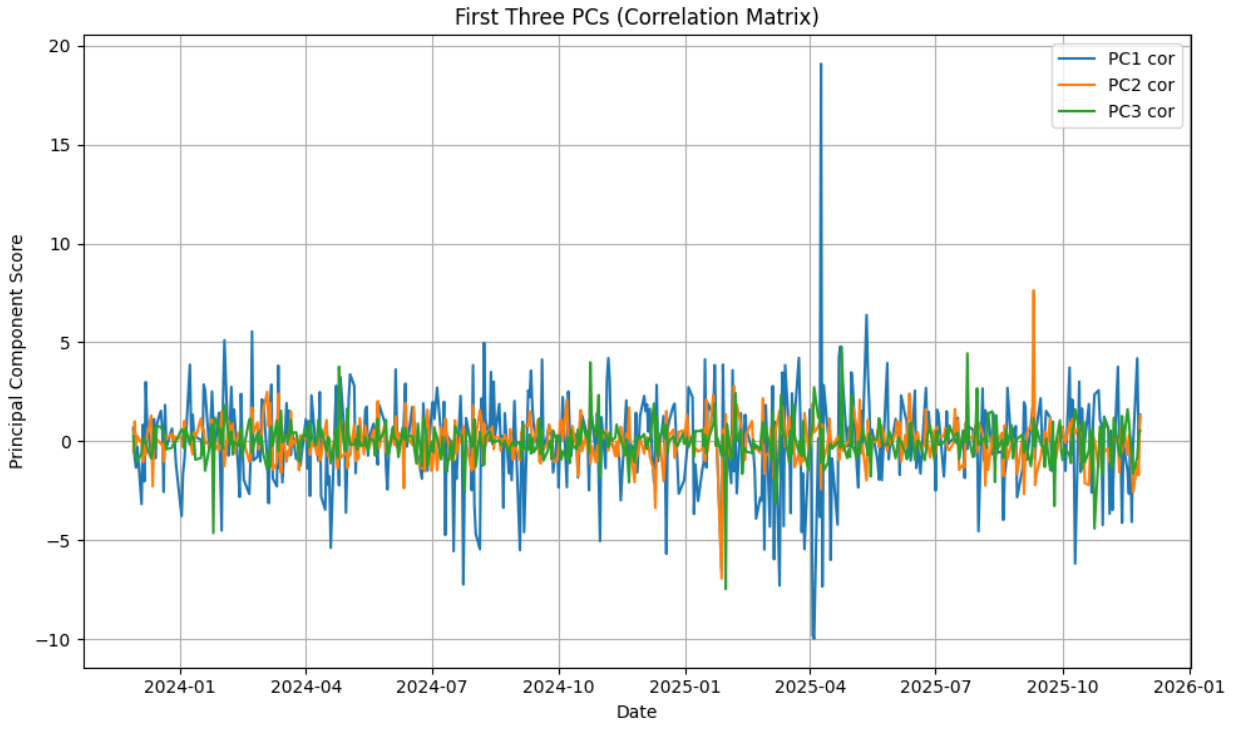

Figure 2: First Three Correlation PCA Component Scores

7 Pseudocode / Algorithm

```
Input: list of tickers, start_date, end_date

1. Download Adjusted Close prices for all tickers.
2. Drop any dates with missing prices.
3. Compute daily log returns: r_t = ln(P_t / P_{t-1}).
4. Form return matrix R (T x N).

Covariance PCA:
5. Center returns: X = R - column_means(R).
6. Compute covariance matrix: Sigma = cov(X).
7. Eigen-decomposition: (eig_vals, eig_vecs) = eigen(Sigma).
8. Sort eig_vals descending; reorder eig_vecs accordingly.
9. Extract first 3 PCs: PCk = eig_vecs[:, k-1], k = 1..3.
10. Compute variance ratios: pve[i] = eig_vals[i] / sum(eig_vals).
11. Compute scores: S_cov = X @ eig_vecs[:, :3].
12. Plot S_cov[:,0], S_cov[:,1], S_cov[:,2] vs. time.
(Optionally repeat steps 5 to 12 on standardized returns and correlation
matrix for correlation PCA.)

Output: covariance matrix, first three PCs, variance explained,
        and plots of the three components.
```