NC STATE UNIVERSITY

PCA S&P 500 Stock Analysis

FIM 590-003 Machine Learning

December 8, 2024

Jinjia Peng

NCSU ID: 200604130

Project Goal

The purpose of this project is to apply Principal Component Analysis (PCA) to financial data from a selected group of stocks within the S&P 500 index. By analyzing the daily returns of these stocks over the past 2 years, the project aims to:

1. Identify patterns and correlations in the data.
2. Reduce the dimensionality of the dataset while retaining the maximum amount of variance.
3. Calculate and interpret the first three principal components to gain insights into the main drivers of variance in the dataset.

Selected Stocks

The analysis uses 10 stocks from the financial sector of S&P 500. Below is the list of the 10 choices:

Tickers:

['KKR', 'GS', 'MA', 'MCO', 'BAC', 'SCHW', 'CBOE', 'FDS', 'MKTX', 'PYPL']

Company Names:

KKR & Co. Inc.

Goldman Sachs Group, Inc.

Mastercard Incorporated

Moody's Corp.

Bank of America Corp.

Charles Schwab Corp.

Cboe Global Markets Inc.

FactSet Research Systems Inc.

MarketAxess Holdings Inc.

PayPal Holdings Inc.

Data Specifications

Time Interval: 2 year most recent data (filtered to the recent 400 trading days for each stock)
Source: Yahoo Finance
Data Used: Adjusted closing prices, incorporating dividend payments for accurate returns

Covariance Matrix

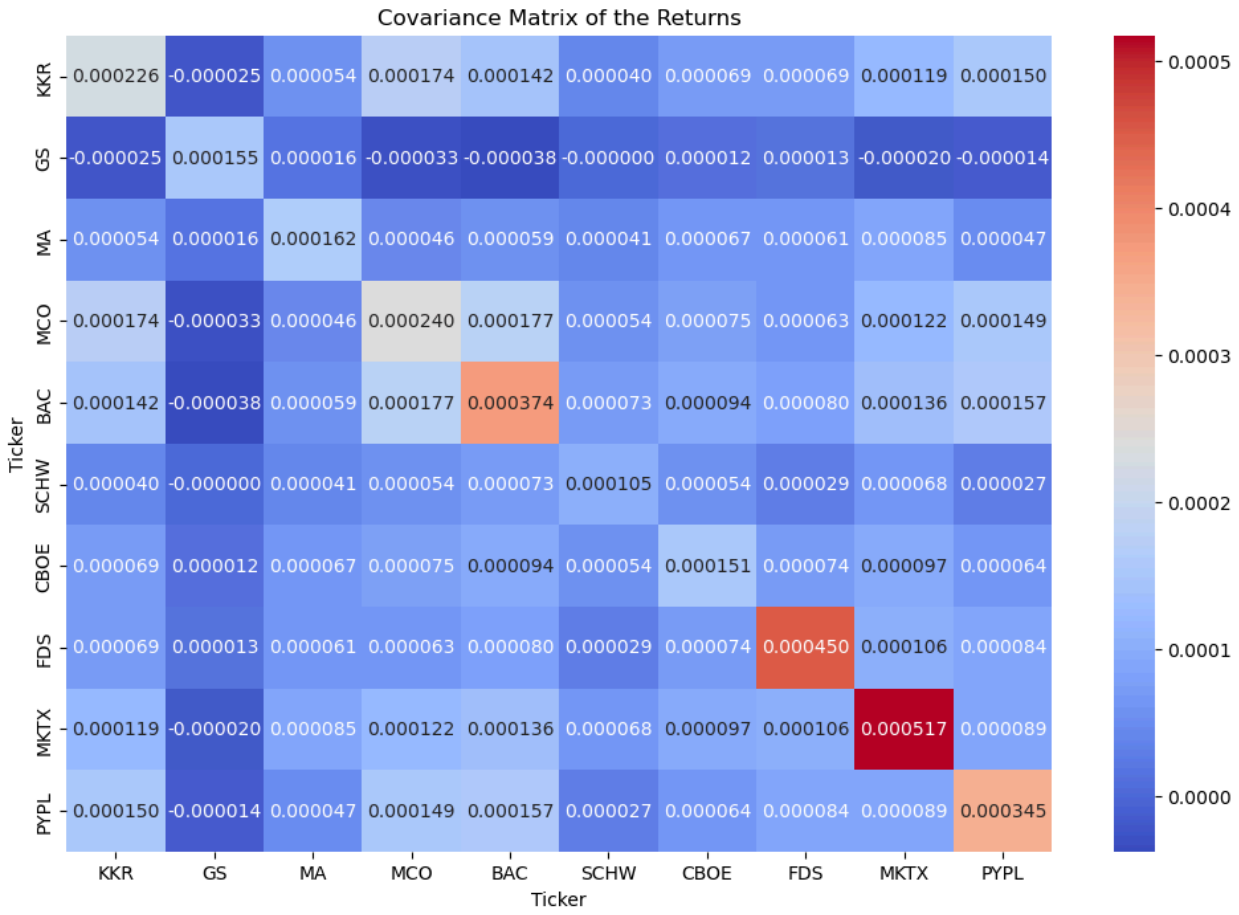
First, we need to calculate the daily returns matrix D, where:

$$Daily\ Return_{i,t} = \frac{Price_{i,t} - Price_{i,t-1}}{Price_{i,t-1}}$$

After updating the daily return matrix, the covariance matrix was computed based on daily returns calculated as the percentage change in adjusted closing prices. The formula used to calculate the covariance matrix:

$$\text{Covariance Matrix} = \frac{1}{n} DD^T$$

where D is the daily return matrix , and n is the number of days. The result is shown below:



The red indicates the highest covariance and the blue indicates the lowest variance. On the diagonal, it shows the variance of the return of each stock.

Principal Components

The first three principal components were calculated from the eigenvectors of the covariance matrix. The calculated eigenvalues and corresponding eigenvectors were sorted in descending order of eigenvalues.

Eigenvalues

The top 3 largest eigenvalues corresponding to **PC1**, **PC2**, and **PC3** are:

1. PC1 Eigenvalue: 1.07×10^{-3}
2. PC2 Eigenvalue: 4.08×10^{-4}
3. PC3 Eigenvalue: 3.76×10^{-4}

First Three Principal Component Vectors

Choosing the eigenvectors corresponding to the top 3 largest eigenvalues:

PC1: [-0.34406042 0.04632706 -0.17355513 -0.36609782 -0.43741247 -0.14028404 -0.221313 -0.31213921 -0.46817571 -0.37245017]

PC2: [-0.20781086 0.1342267 0.10131406 -0.26285611 -0.31974126 -0.00274053 0.05309337 0.70939572 0.38789561 -0.31729134]

PC3: [0.06246637 0.05501528 -0.02461567 0.04892298 0.07415518 -0.05852115 0.0014036 0.5937137 -0.74140266 0.28102016]

Variance Explained

Dividing the eigenvalue by the total value of the eigenvalues, obtain the variance explained by each PCA Component:

Variance Explained PC1: 39.23%

Variance Explained PC2: 14.97%

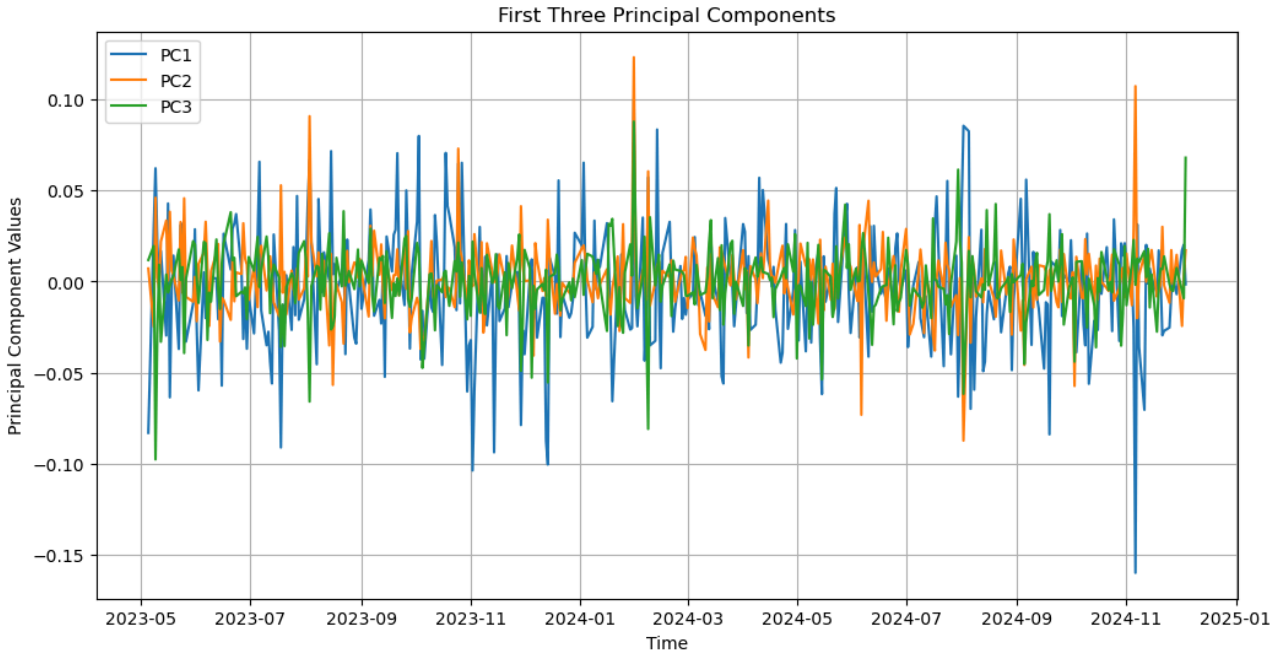
Variance Explained PC3: 13.81%

Total Variance Explained by the first PCA Components: 68.01%

Graphs

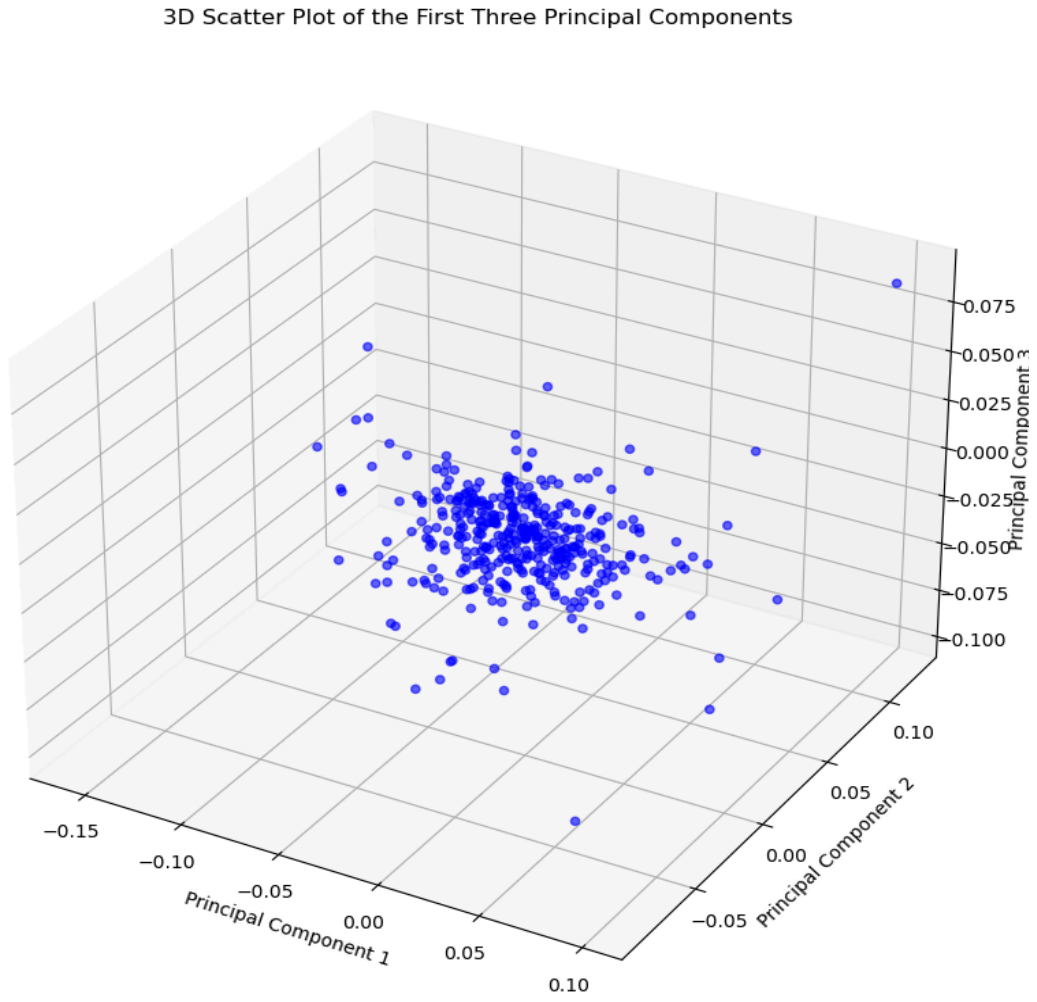
(a) Time-Series Graph

A line plot showing the temporal evolution of PC1, PC2, and PC3 over time.



(b) 3D Scatter Plot

A 3D scatter plot showing data distribution in the reduced space defined by PC1, PC2, and PC3.



Conclusion

This project successfully applied Principal Component Analysis (PCA) to the daily return data of 10 selected S&P 500 stocks over the past two years (400 trading days). By calculating the covariance matrix and performing eigenvalue decomposition, it identified the first three principal components and quantified their contribution to explaining the variance in the dataset.

The results showed that the First Principal Component (PC1) captured the largest portion of variance at 39.23%, followed by PC2 at 14.97% and PC3 at 13.81%. Together, these three components explained 68.01% of the total variance, highlighting their importance in understanding the underlying structure of the data.