Kellogg School of Management
FINC 585-3: Asset Pricing
Spring 2024

Prof: Torben G. Andersen & Zhengyang Jiang

Assignment 3

Problem 1.

For this question, we are going to use the dataset `stocks_pca.zip` available on CANVAS. This dataset contains the daily price of the current constituents of the S&P 500 index from 2004 to 2023 available on CRSP. Stocks are identifies by their unique PERMNO. Using this data, calculate daily log-returns and drop stocks that contain any missing values in the sample.

Additionally, we are going to use the Fama-French 5 factors + Momentum (FF) dataset available on CANVAS under the `ffdaily.csv` file. This dataset contains the daily log-returns of the Fama-French 5 factors + Momentum from 2004 to 2023.

A) We start by making some simple Principal Component Analysis (PCA). Extract the first principal component of the log-returns. How does it look like on the time series? How much of the variance is explained by the first principal component?

B) Now extend this to up to 8 factors. How much does the explained variable increase with the addition of new factors? Using the Bayes Information Criteria, how many factors should we choose? How does your choice compare with using just the explained variance? Explain.

C) Pick the model with the greatest number of factors. Using rotation, how do the principal components correlate with the FF factors?

D) We now turn to using the Canonical Correlation Analysis (CCA) to analyze the relationship between the FF factors and the PCA factors. Using 1 to 5 components, calculate the canonical correlation between these two sets of factors. How does it change with the number of components?

E) Now obtain the canonical variables by rotating the PCA and FF factors according to the canonical loadings. How much correlation is there between the canonical variables?

Solution.

A) I extract the first principal component of the log-returns. The time series is shown in Figure 1. This factor can explain about 5% of the variation in the log-returns. We see that this factor picks up some volatility specially during the financial crises and the COVID-19 pandemic.

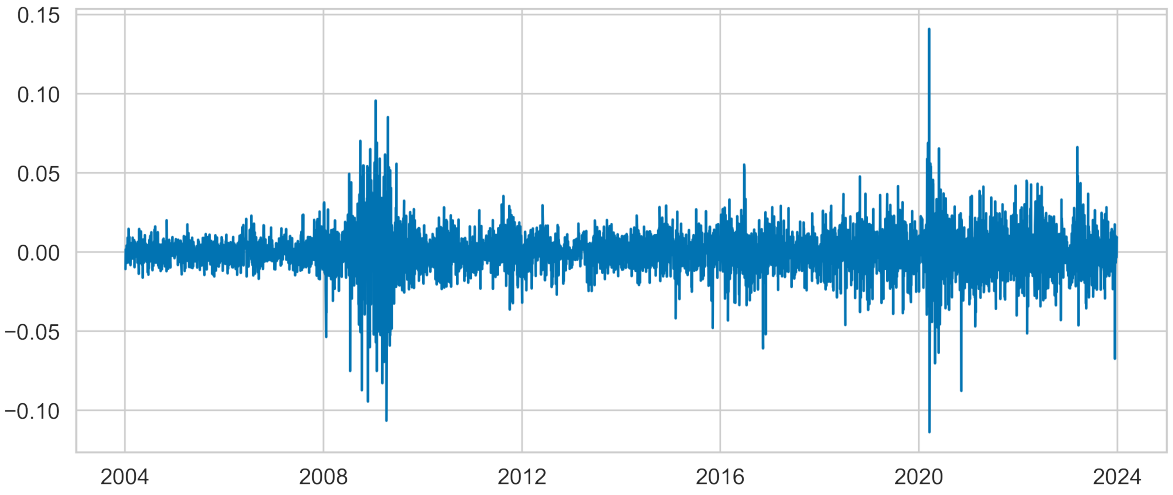

Figure 1: First Principal Component of Log-Returns

B) I iterate the PCA analysis using 1 to 8 factors. For each one of these, I calculate the explained variance and a BIC criteria using the formula

$$\text{BIC} = T \log(\widehat{\sigma}_\varepsilon^2) + k \log(T) \tag{1}$$

where T is the length of the time series, k the number of principal components and $\widehat{\sigma}_\varepsilon^2$ the estimated variance of the residuals. This estimate is based on Bishop and Nasrabadi (2006) and is available on the `scikit-learn` library. Figure 2 plots the explained variable and BIC as a function of the number of factors. As we know, increasing the number of factors will always increase the explained variance. However, by penalizing it with the number of parameters estimated (using BIC), we can obtain an optimal number of factors. In this case, the BIC suggest only extracting a single factor. In contrast, the explained variance will always suggest extracting more factors.

C) Using the model with 8 factors, I analyze how each one of them correlate with each of the FF factors. For this, I start by showing the correlation matrix between each PCA and FF factor. This is depicted in Figure 3. As we can see, the first PCA factor correlates the most with the FF. PC2 tends to pick up HML as well as some Momentum. PC4 and PC5 correlate negatively with the Market factor. PC6 seems to pick up some RMW. PC7 and PC8 don't show any significant correlation with the FF factors.

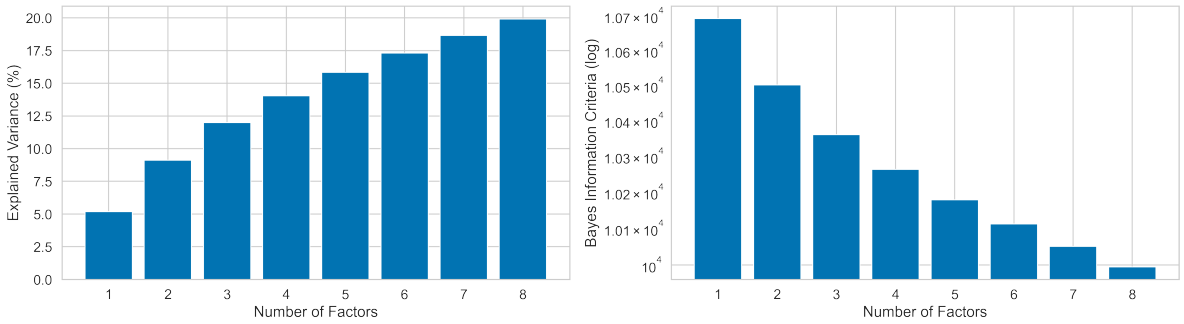

Figure 2: Explained Variance and BIC as a Function of the Number of Factors

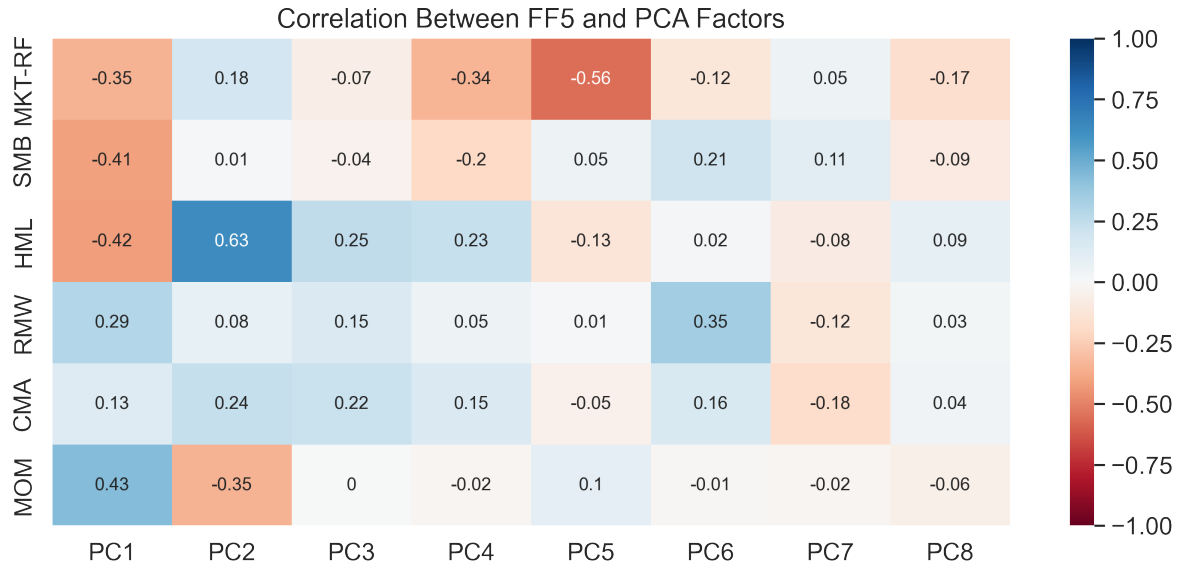

Figure 3: Correlation between PCA and FF factors

To better understand this relationship, I run an OLS regression with each of the FF factors as a dependent variable and the PCA as the explanatory variables. This tends to capture the linear relationship between these variables. Figure 4 shows the R-Squared of each of these regressions. According to this figure, the FF factors which mostly correlate with the PCA are Market and HML, but all of them present a significant relationship as implied by the F-test.

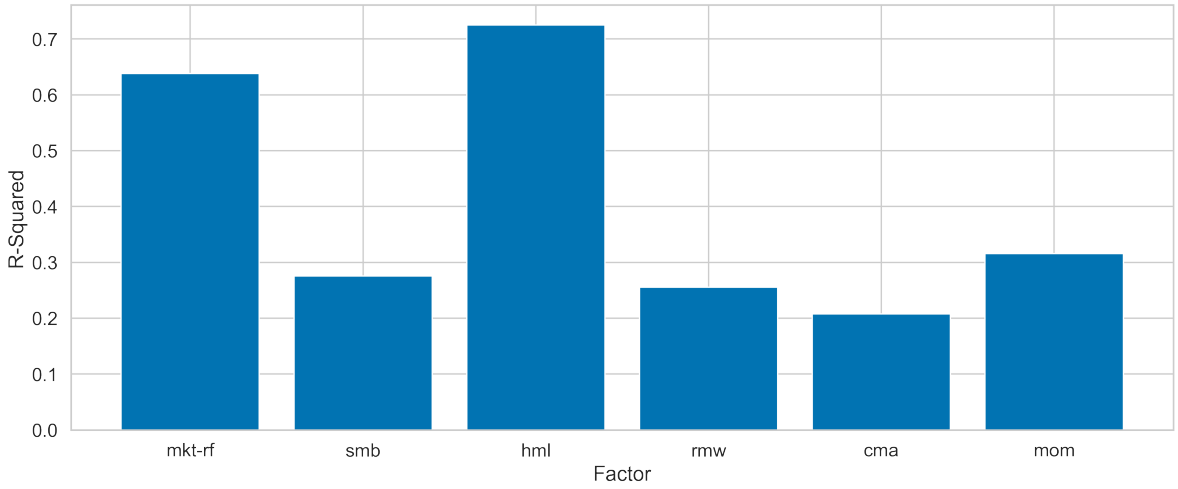

Figure 4: R-Squared of the OLS Regression between each FF factor and the Principal Components

D) I now turn to the Canonical Correlation Analysis (CCA) to analyze the relationship between the FF factors and the PCA factors. I calculate the canonical correlation between these two sets of factors using 1 to 5 components. Figure 5 shows the canon ical correlation (CCA Score) as a function of the number of components. As we can see, the CCA Score starts to decrease after the inclusion of the 3rd component, suggesting an optimal number.

E) Using the model with 3 components, I rotate both the PCA and FF factors according to the CCA loadings. This helps us understand the relationship between the two set of factors. Figure 6 shows the scatter plot of each pair of canonical variables. The correlation between the first 2 pairs appears to be stronger than the 3rd. This can also be seen on the correlation matrix in Figure 7, which shows a strong correlation between the first and second pairs of canonical variables and a weaker correlation with the third pair.

□

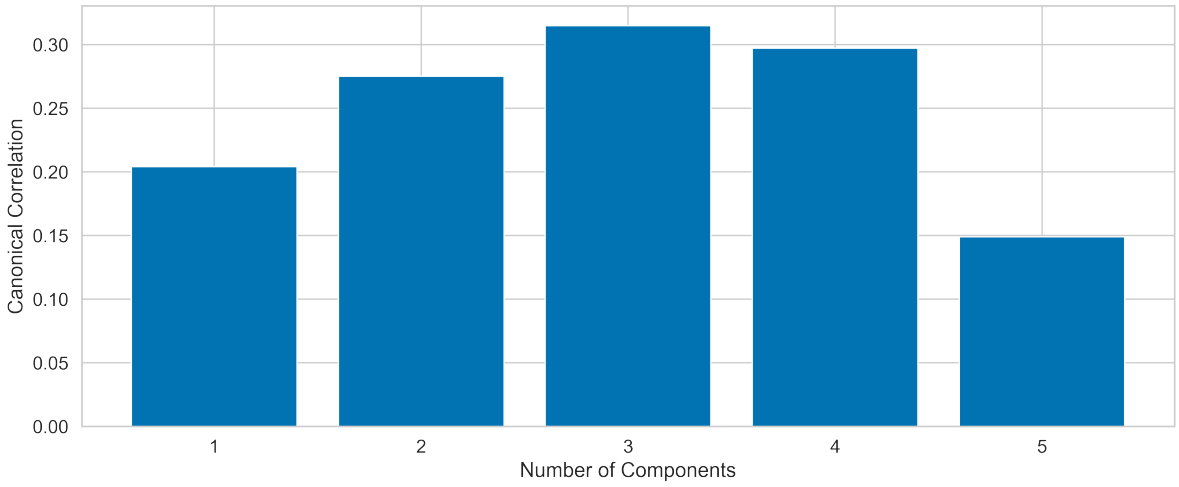

Figure 5: CCA Score as a Function of the Number of Components

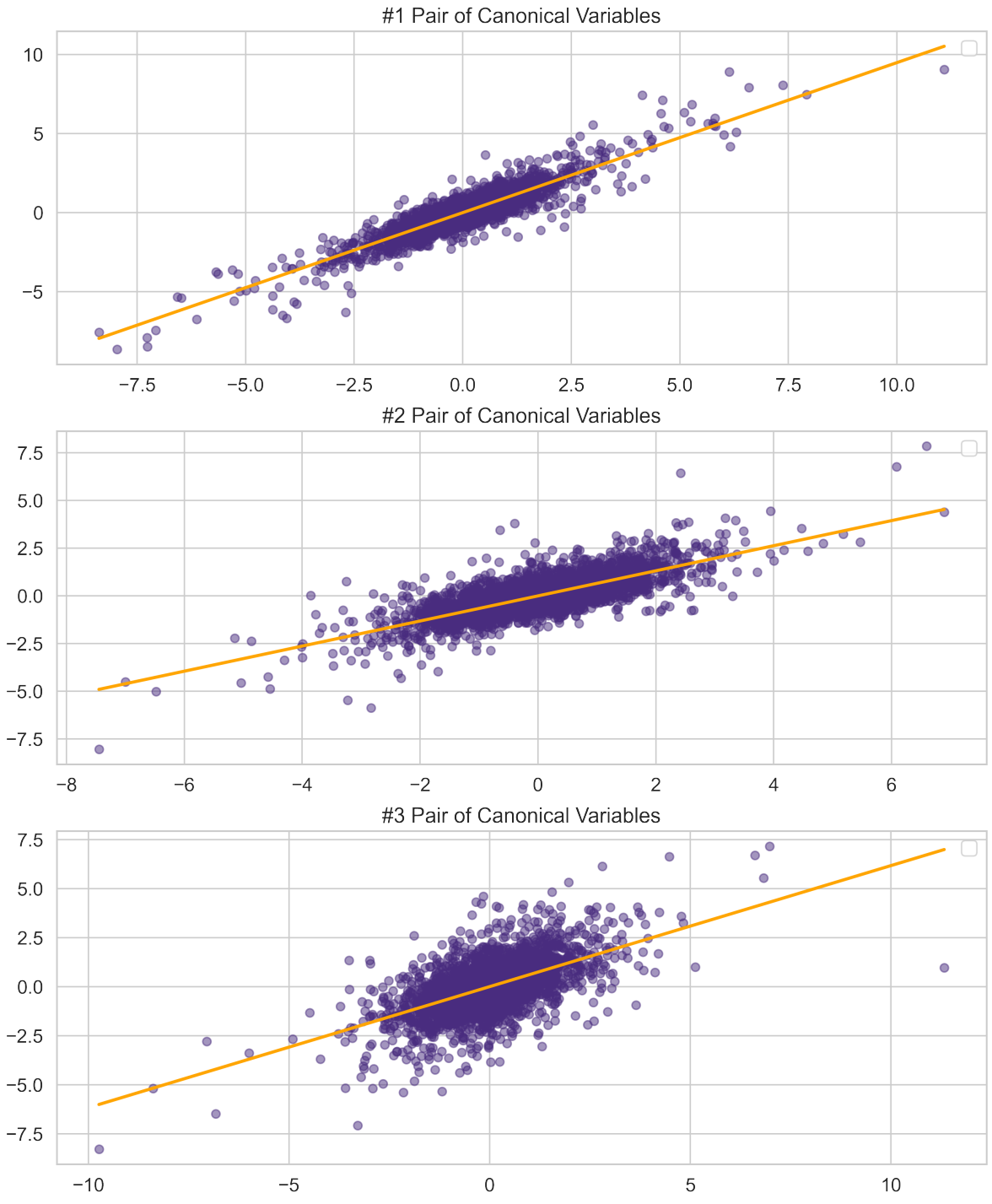

Figure 6: Relationship Between Pairs of Canonical Variables

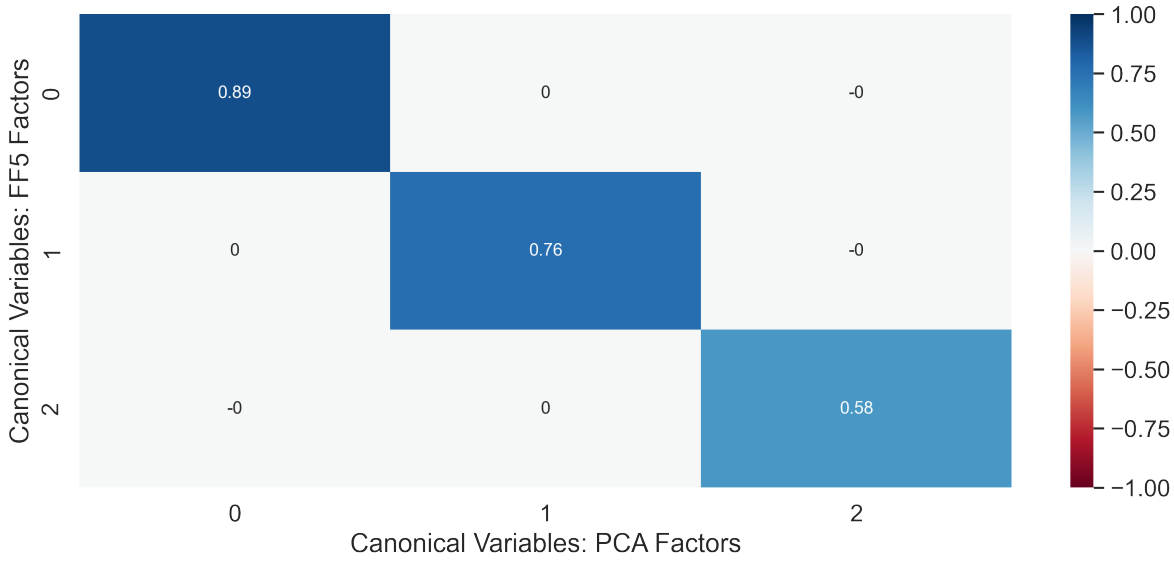

Figure 7: Correlation Between Canonical Variables

References

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.