Yijun(Aaron) Shen
11.18.2022
J.P. Morgan

# Implement

**Step 1: Data Retrieving**

The project is related to retrieving data from FRED using the FRED API and then reducing the dimensions of the fixed income dataset with six features. Initially, the "fred" package on python needs to be imported in order to directly retrieve data from the website. I used the "Fred( ).series" function to first request the datasets from FRED individually, and then get the value of the datasets. All six datasets are time series data, and because they belong to different periods, we must set the time range starting at 1976-06-01 and then merge them to one dataset. The first step after obtaining the dataset should be converged from the original dataset with price to dataset with monthly return. Therefore, I used the "df.pct_change().dropna()"function to calculate the monthly return of the dataset and drop any missing values in the dataset. From there, we are ready for further model construction.

| date | GS1 | GS2 | GS3 | GS5 | GS7 | GS10 |
|---|---|---|---|---|---|---|
| 1976-07-01 | -0.049080 | -0.029745 | -0.025992 | -0.015769 | -0.006452 | -0.003817 |
| 1976-08-01 | -0.032258 | -0.032117 | -0.036517 | -0.024032 | -0.015584 | -0.007663 |
| 1976-09-01 | -0.026667 | -0.031674 | -0.029155 | -0.024624 | -0.022427 | -0.023166 |
| 1976-10-01 | -0.058219 | -0.068536 | -0.063063 | -0.053296 | -0.033738 | -0.023715 |
| 1976-11-01 | -0.038182 | -0.028428 | -0.024038 | -0.034074 | -0.041899 | -0.016194 |

*Table 1: Monthly Return for U.S. Treasury Bonds*

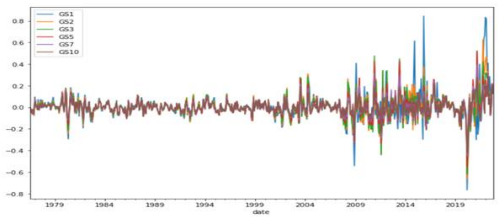


*Figure 1: Monthly Return for U.S. Treasury Bonds*

**Step 2: Dimensional Reduction**

I conducted dimensional reduction to the monthly return dataset in order to transform the predictors so it will be easier for us in the future to fit different models. For the dimensional reduction method, I chose the principal component analysis (PCA). PCA is a popular technic that uses linear way to combine original dimensions in order to generate new dimensions. The theory behind dimensional reduction is that for the original predictor $X_1$ to $X_p$ , there are basically $P$ predictors with $P$ linear combinations. For those linear combinations, we have $P+1$ coefficients from $\beta_1$ to $\beta_P$. When we use dimension reduction methods, we create predictors $Z_1$ to $Z_m$ where $P > m$ and Zm $= \sum_{j=1}^{p} \phi_{jm} X_j$. In our case, each new dimension created by PCA is the combination of the six original dimensions, and each new dimension is orthogonal to others. With PCA, we are able to have fewer dimensions while maintaining the major variations of the original bonds monthly return dataset. Another reason to use PCA is that all the bond returns that we have are correlated with each other, and they can be influenced by some market factors. Since the PCA calculates a new projection of the original dataset, normalization is needed before we conduct the PCA process. When applying the PCA process, I first checked the auto solver of PCA, which reduces six dimensions of the original dataset into five dimensions. However, the last two dimensions are not as significant, because the amount of variations that they explain are extremely small, while the first three dimensions can explain over 99% of the datasets. Therefore, I changed the process to produce only three dimensions. The three dimensions that we have represent the projections of our original monthly return dataset, and those three dimensions can explain over 99% variations of the original dataset.

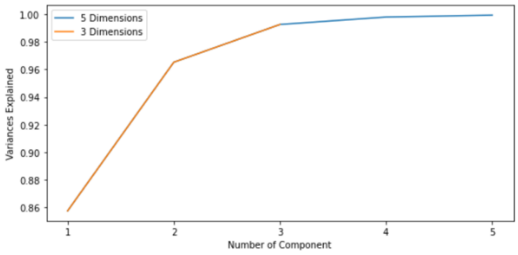


*Figure 2: Variances Explained of 5 Dimensions VS. 3 Dimensions*

The next step is to check the stability of the three dimensions. I have conducted an augmented Dickey–Fuller test (ADF) to check if our new dataset is a stationary time series. The null hypothesis is that the time series data is non-stationary, and the alternative hypothesis is that the time series data is stationary. Table 2 presents the result of the test.

```
First_D ADF statistic Value:  -4.16008485357575
First_D P-Value:  0.0007692565429625551
Second_D ADF statistic Value:  -6.036087877915685
Second_D P-Value:  1.3809966685692844e-07
Third_D ADF statistic Value:  -5.570061441423882
Third_D P-Value:  1.4743983175050937e-06
```

*Table 2: Results of the ADF Test*

From the result, all p-values for the first dimension, the second dimension and the third dimension of our data are less than 0.05. Therefore, we are able to reject the null hypothesis that the time series PCA data is not stationary. In the long term, our dimensional reduction is stationary, so it is stable.

**Conclusion:**

From the PCA model, I transformed the dataset from six dimensions to three components. The first component PC1 is a fitting line in the six-dimensional variable space that best approximates the dataset in the least square sense, and it basically goes through the average points of all six dimensions, so it is able to explain about 85% variation of the original dataset. Since we want to better model the systematic variation of the dataset, the second component PC2 and the third component PC3 are needed. They are orthogonal to PC1 and also pass through the average point, which improves the approximation of the original dataset. With those three components, over 99% variation of the original dataset is explained, and the information can be more easily visualized and analyzed for further research. Stability is a key factor for prediction performance. To test the stability of the PCA data, I have done the ADF hypothesis test which can prove that the dataset is stationary and stable. Stationary dataset allows us to conduct models such as ARMA and SARIMA in the future.