US Treasury Yields Dimension Reduction Analysis

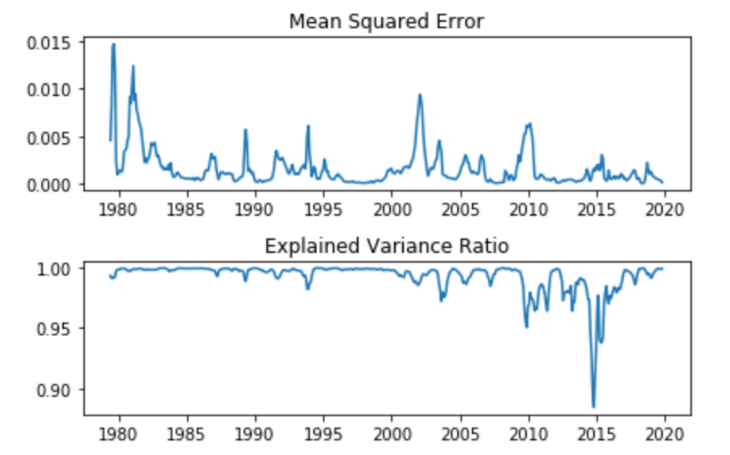
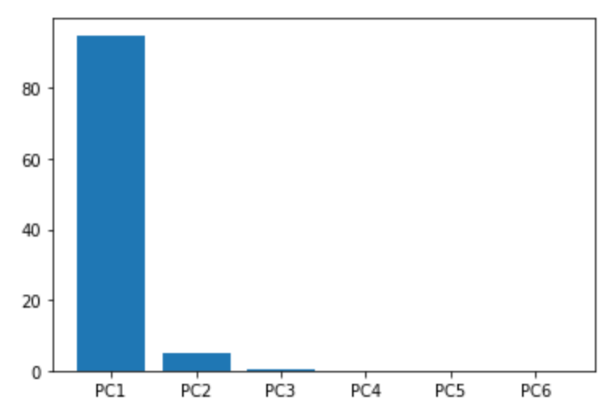
In this task, I choose Principal Component Analysis to implement dimension reduction on 6 US treasury yields ranging from 1yr to 10 yrs. Principal Component Analysis (PCA) is chosen for this task because it is easy to implement on Python with complete toolkits in Sklearn package. It captures ‘principal components’ and ‘factor loadings’ of our datasets to facilitate our analysis of historical yields.

Introduction of the model:

In order to measure out sample stability and capture the time-varying characteristic of yields data, I implement a moving-window PCA. Before implementing PCA, data is standardized in order to improve the performance. The model fits the previous 24 months yield data, generates principal components and uses those components to transform 12 months data after this time spot. This process keeps rolling from 1978 to 2018.

Number of Principal Components:

The following Figure 1 shows the explained variance of PCs generated by 2018-2019 yield data. Since the first 3 components nearly explain 100% of the data, we might be able to say 3 components are sufficient for our model. To test whether the first three PCs are stable during the entire time period, I recorded the explained variance of those 3 PCs in each moving window. Graph 1 plots the explained variance ratio curve. We can see that the first 3 PCs are stable for the most of the time except for some abnormal spots between 2014 – 2015. (Such abnormal performance will be further explained later) Figure 2 shows the mean square error between transformed 12-month data after each window and real historical data. Highest MSE is lower than 1.5%, which also supports that the first 3 components explain the most information of the dataset.



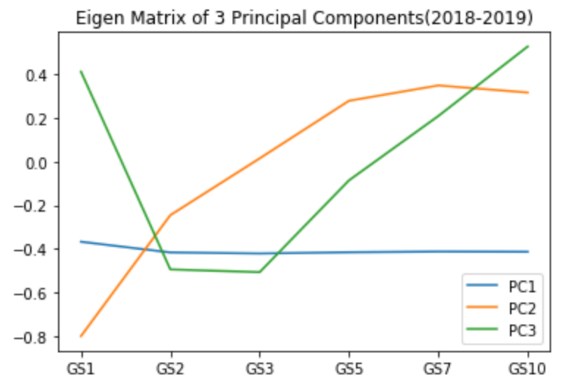
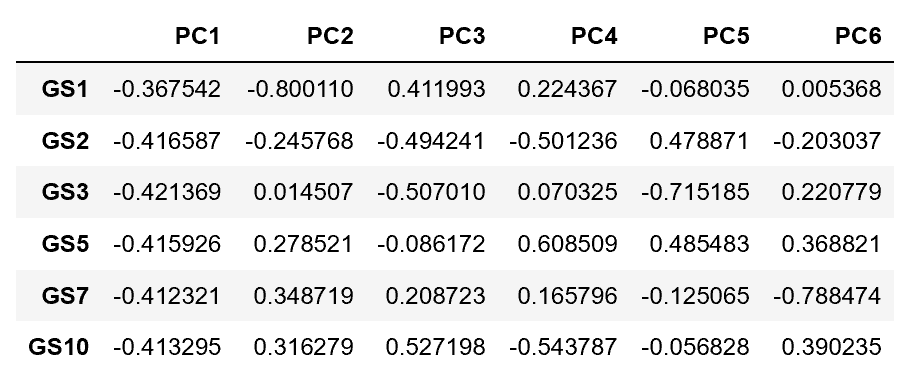
Figure

Figure

Analysis of Reduced Dimension:

Another question needs to answer is what kind of information those 3 components explain? Table 1 shows eigenvector sets of each PC and Figure 3 plots eigenvectors of the first 3 PCs.(both generated by 2018-2019 data) From PC1 in Table 1, we can see that this ‘factor’ poses similar effects to each treasury, which means if we add one factor, all Treasury rates will decrease. This corresponds to the general downward trend of Treasury rates this year. The second PC presents rising characteristics from short term to longer term rate, which corresponds to the slope: diminishingly increasing yield curve. And the third PC shows convex pattern, similar to curvature of bonds. Considering general characteristics of bond yield curve and relying on empirical experience, we might be able to infer that those 3 PCs represents ‘shifts’, ‘slope’ and ’curvature’ of Treasury.

Table

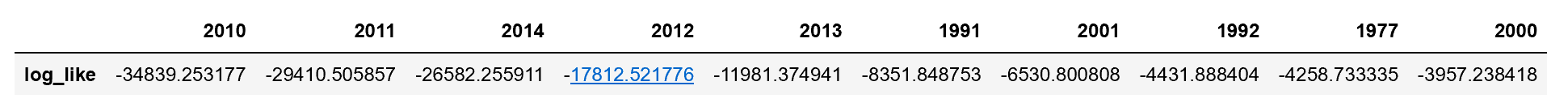


Figure

Abnormal Explained Variance during 2013-2014:

When testing the stability of those 3 PCs, I noticed that at some spots during 2013-2015, the first 3 PCs have terrible explained variance (even negative) of the data. One of possible explanation is that the first factor, ‘parallel shift’ can’t explain the movement. After 2008 financial crisis, the Fed implemented several rounds of quantitative easing by 2015 and fed rate remained at a very low level for several years. On the other hand, investors shared optimistic outlook of the long-term US economy, sold fixed income products and rushed into stock markets, pushing the longer term treasury yields higher and consequently caused a jump of longer term yield. To avoid economy overheating or in order to control the inflation, starting from 2015, the Fed raised rates and such phenomenon soon decayed. Therefore, during that time, short term and longer term have different movement and parallel shift can’t capture such pattern. After 2014, the first 3 PCs remain stable again.

Similarity Searching:

I implemented the rolling PCA again to find the year with the most similar ‘pattern’ with current year. Since this year we don’t have December’s data, I sliced the whole dataset by year and removed each year’s December’s data. Then, I Ran PCA model for each year and find the max log likelihood score among each year. Note that since we already knew 3 components are not sufficient to explain 2013 – 2015, I kept all components in the model in order to find similarity. The one with the lowest score(because it is negative log likelihood) means highest probability of fitting current year’s data. Based on Table 2, we can see that 2010, 2011, 1991 have very low scores and their movements are indeed similar to the current one. I would like to interpret 2012 – 2014 as noise since they don’t show clear pattern and movements are extremely abnormal that even keeping all components can’t help.

Table