MortgageMarketAnalysis

2022-06-16

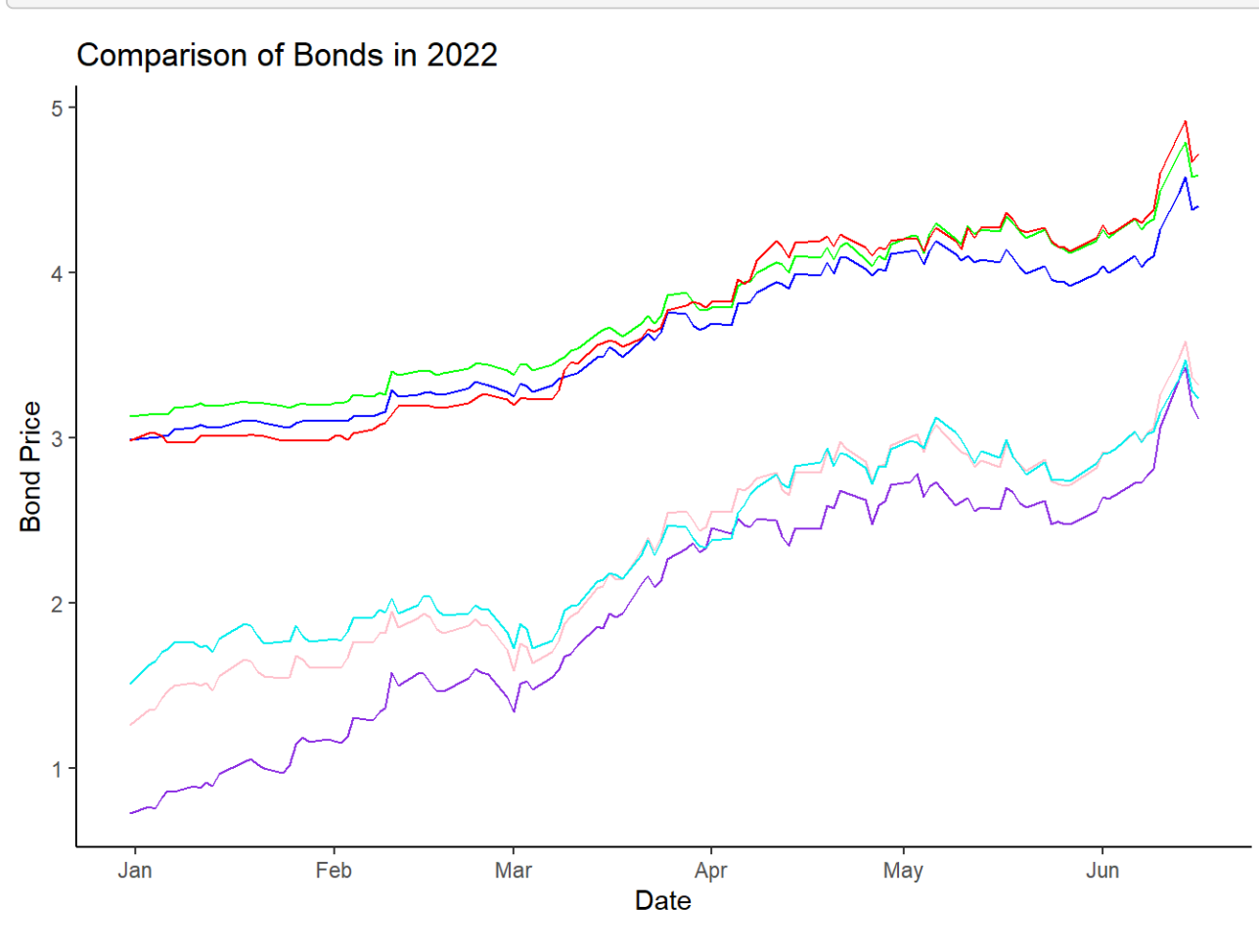
The goal of this analysis is to compare relative performance of mortgage bonds to US gov't bonds in order to see how the FOMC rate hike on 6/15/2022 and the abysmal housing report on 6/16/2022 are affecting the mortgage markets. In this analysis, I will import the data (Excel sheet attached), perform some basic exploratory data analysis, and then from the data analysis, decide on a feasible statistical method to extract insight into the current mortgage market.

First we begin by loading in the necessary data and libraries. Additionally, I filter out unneeded columns that are blank and limit the data to returns from 2022.

```r
# Clear the Workspace
rm(list=ls())

library(tidyverse)
library(readxl)
library(devtools)

rate_mbs_selloff <- read_excel("rate mbs selloff.xlsx")

#filter out NA columns by position
data1 <- rate_mbs_selloff %>% select(
  1,2,5,8,11,14,17
)

#only want 2022 data
data1 <- data1 %>% slice(1:116)#116

#check to make sure it works
View(data1)
```

After this is done, we can perform exploratory data analysis and plotting of our data. The main focus of our EDA will be to explore the correlation of the yields.

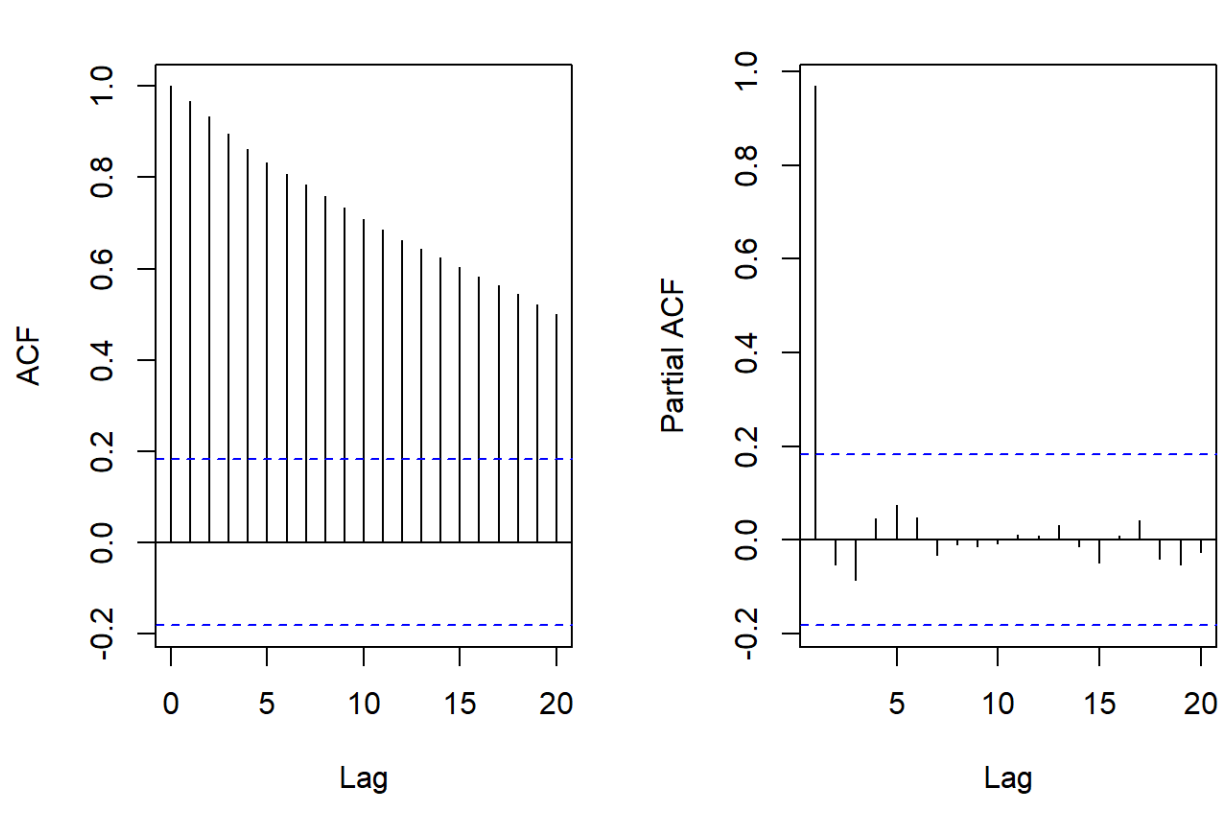
```r
#start EDA by plotting all line with ggplot2
library(ggplot2)

plot1 <- ggplot()+
  geom_line(data = data1,
            aes(x = Date, y = us_govt_2y),
            color='blueviolet')+

  geom_line(data = data1,
            aes(x = Date, y = us_govt_5y),
            color='pink')+

  geom_line(data = data1,
            aes(x = Date, y = us_govt_10y),
            color='cyan2')+

  geom_line(data = data1,
            aes(x = Date, y = fannie_4cpn),
            color='blue')+

  geom_line(data = data1,
            aes(x = Date, y = Mid_Yield_To_Convention),
            color='green')+

  geom_line(data = data1,
            aes(x = Date, y = fannie_5cpn),
            color='red')+

  theme_classic()+
  labs(title = 'Comparison of Bonds in 2022')+
  xlab('Date')+
  ylab('Bond Price')

plot1
```



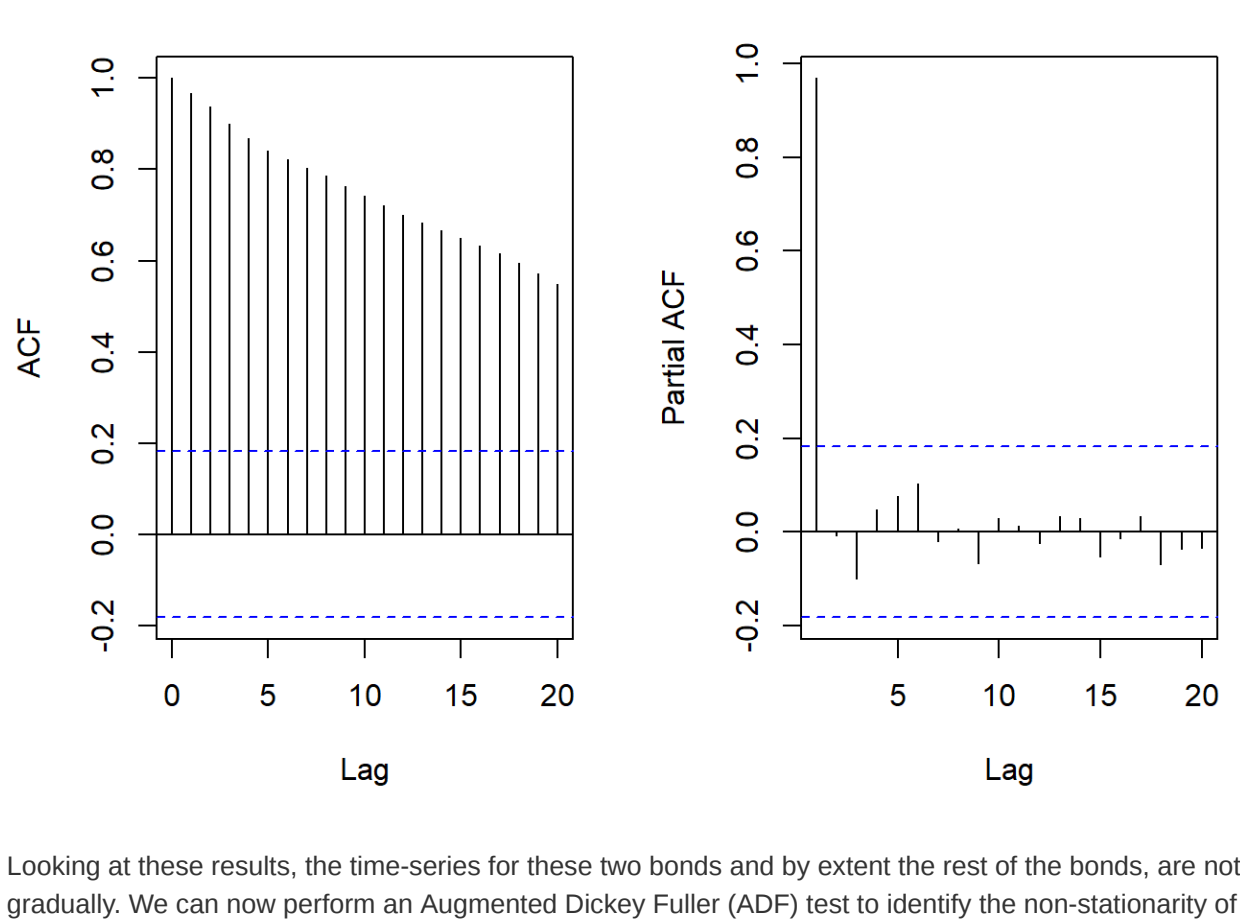
Looking at the plot above, it appears that the bonds might be correlated. In this case, principal component analysis will be a helpful statistical method to use, as it is a dimension-reduction tool that is used to reduce a large set of possibly correlated variables into a smaller set number of uncorrelated variables(called the principal components). These principal components will explain most of the variation of the large set of variables. The first principle component explains the maximum percentage of the total variance of interest change. The second principal component is designed to be linearly independent of the first and explains the maximum percentage of the remaining variance, and so on and so forth.

In order to conduct analysis on the data, we need to check to see if it is stationary, and if it is, we need to difference it.

```r
#remove dates to look at data
data4 <- data1 %>% select(
  2,3,4,5,6,7
)

#we can run acf and partial acf test on the 2 year Treasury Note and the Fannie 4% coupon to see if the data is non stationary
par(mfrow=c(1,2))
acf(data4$us_govt_2y)
pacf(data4$us_govt_2y)
```



```r
par(mfrow=c(1,2))
acf(data4$fannie_4cpn)
pacf(data4$fannie_4cpn)
```



Looking at these results, the time-series for these two bonds and by extent the rest of the bonds, are not stationary as the ACF decreases gradually. We can now perform an Augmented Dickey Fuller (ADF) test to identify the non-stationarity of the time-series. If the p-value from the test is less than some significance level, then we will fail to reject the null hypothesis and conclude that the time-series is non-stationary.

```r
library(tseries)
adf.test(data4$fannie_4cpn)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data4$fannie_4cpn
## Dickey-Fuller = -2.1884, Lag order = 4, p-value = 0.4984
## alternative hypothesis: stationary
```

The p-value comes out to 0.9484, meaning that we fail to reject the null hypothesis and conclude that the data is non-stationary. As the time-series shows a possible drift, let us check.

```r
library(urca)

test1 <- ur.df(data4$fannie_4cpn, type='drift',lags = 4)
summary(test1)
```

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression drift
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.133635 -0.032218  0.002744  0.028258  0.104783
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.021322   0.041220   0.517    0.606
## z.lag.1     -0.009307   0.011377  -0.818    0.415
## z.diff.lag1 -0.027299   0.090820  -0.301    0.764
## z.diff.lag2 -0.044436   0.089422  -0.497    0.620
## z.diff.lag3 -0.082354   0.083461  -0.987    0.326
## z.diff.lag4  0.100773   0.083646   1.205    0.231
##
## Residual standard error: 0.0485 on 105 degrees of freedom
## Multiple R-squared:  0.03452,    Adjusted R-squared:  -0.01146
## F-statistic: 0.7508 on 5 and 105 DF,  p-value: 0.5873
##
##
## Value of test-statistic is: -0.8181 2.9874
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau2 -3.46 -2.88 -2.57
## phi1  6.52  4.63  3.81
```

Upon running this Dickey-Fuller unit root with drift test, it can be seen that interpreting the test is time consuming to determine if there is drift involved in the data. Therefore, I used a function that can be found at https://gist.github.com/hankroark/968fc28b767f1e43b5a33b151b771bf9 to interpret the output.

```r
source("~/interp_urdf.R")

interp_urdf(test1, level = "5pct")
```

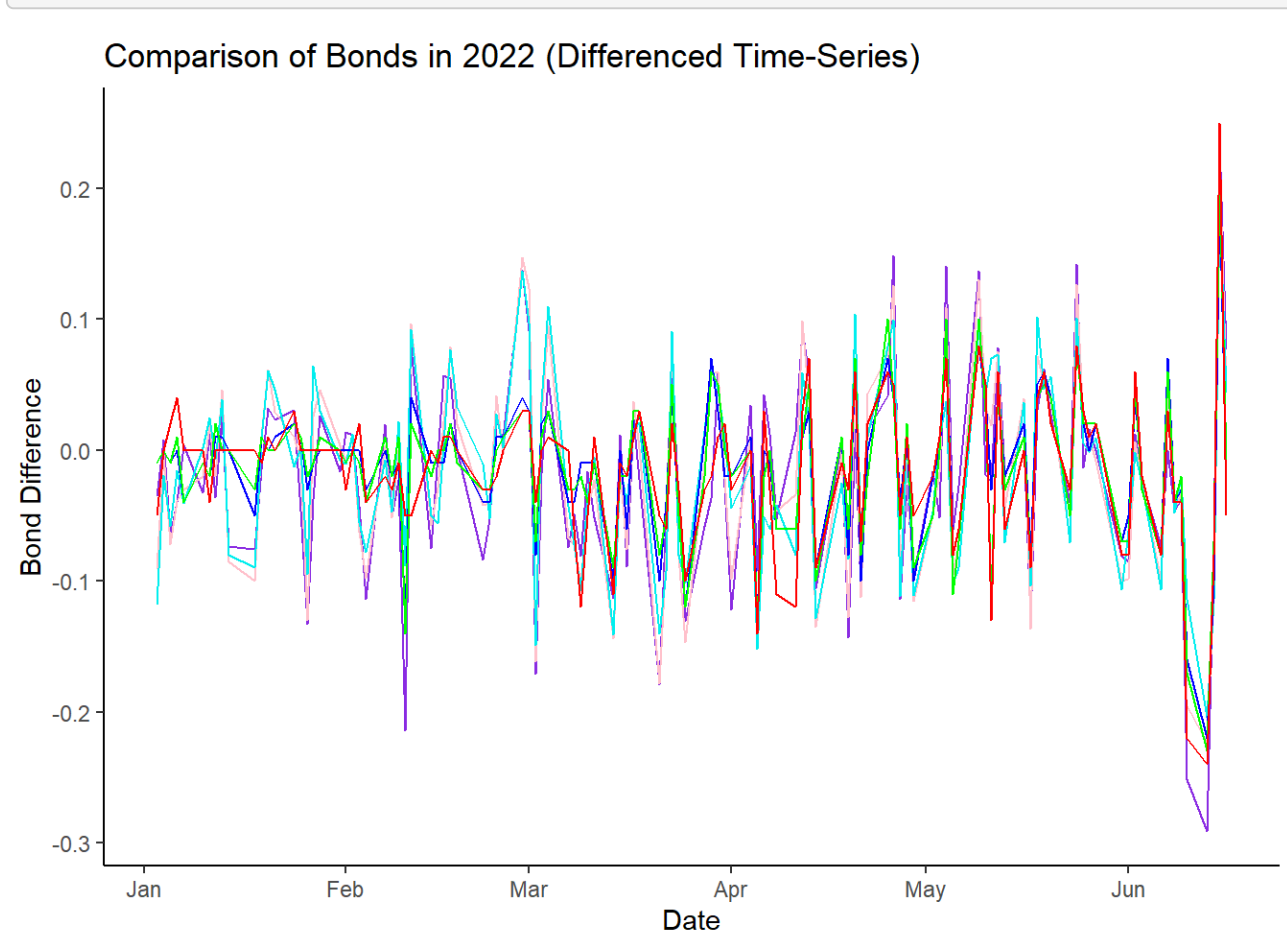
```
## ========================================================================
## At the 5pct level:
## The model is of type drift
## tau2: The first null hypothesis is not rejected, unit root is present
## phi1: The second null hypothesis is not rejected, unit root is present
##       and there is no drift.
## ========================================================================
```

As can be seen, this shows that the Fannie 4% coupon bond contains no drift.

Based on the analysis above, principle component analysis appears to be an useful model. This allows us to reduce the dimensionality of the different bonds and then analyze how the bond prices are functioning related to each other. The principal components generated by the model will be able to explain the vast majority of the yield curve while remaining uncorrelated to each other.

The graph above shows that the bonds shown above appear to be highly correlated. This is a problem if we would like to perform principal component analysis as PCA requires a stationary time-series, which means that a invariant of the market must be produced. According to Meucci(2005), taking the first difference will be sufficient. Therefore, we can establish that Rt,k = yield i,k - yield i-1,k as the yield return of maturity k at time t.

```r
#remove dates to difference the series
data3 <- data1 %>% select(
  2,3,4,5,6,7
)
data3 <- data.frame(data3)

#difference the series once
returns <- as.data.frame(lapply(data3, diff, lag=1))
returns <- data.frame(returns)

#remove last row of data1 as we have a differenced time-series now
data1 <- data1 %>% slice(1:115)#115

#add the date column back to the dataset
returns$Date <- data1$Date


#plot the data to check that it worked
plot2 <- ggplot()+
  geom_line(data = returns,
            aes(x = Date, y = us_govt_2y),
            color='blueviolet')+

  geom_line(data = returns,
            aes(x = Date, y = us_govt_5y),
            color='pink')+

  geom_line(data = returns,
            aes(x = Date, y = us_govt_10y),
            color='cyan2')+

  geom_line(data = returns,
            aes(x = Date, y = fannie_4cpn),
            color='blue')+

  geom_line(data = returns,
            aes(x = Date, y = Mid_Yield_To_Convention),
            color='green')+

  geom_line(data = returns,
            aes(x = Date, y = fannie_5cpn),
            color='red')+

  theme_classic()+
  labs(title = 'Comparison of Bonds in 2022 (Differenced Time-Series)')+
  xlab('Date')+
  ylab('Bond Difference')

plot2
```



We need to confirm the data is now stationary with an ADF test.

```r
adf.test(returns$fannie_4cpn)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  returns$fannie_4cpn
## Dickey-Fuller = -5.851, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

The p-value = 0.01, meaning we reject the null hypothesis that the data is non-stationary.

Next, we can check the yield return matrix for any kind of correlation. We expect to see a correlation in that if a bond with two year maturity starts to pay less, there should be an impact on a bond with 5 year maturity. The larger the difference in maturities, the lower the correlation should be. Additionally, it is good to look at the matrix to make sure there is some correlation, else PCA will not help reduce dimensionality.

```r
#remove dates to look at correlation
cor.returns <- returns %>% select(
  1,2,3,4,5,6
)

options(digits = 2) #better visualization
cor(cor.returns) #correlation matrix of yield returns
```

```
##                         us_govt_2y us_govt_5y us_govt_10y fannie_4cpn
## us_govt_2y                    1.00       0.93        0.82        0.84
## us_govt_5y                    0.93       1.00        0.95        0.88
## us_govt_10y                   0.82       0.95        1.00        0.85
## fannie_4cpn                   0.84       0.88        0.85        1.00
## Mid_Yield_To_Convention       0.82       0.85        0.81        0.97
## fannie_5cpn                   0.73       0.75        0.72        0.86
##                         Mid_Yield_To_Convention fannie_5cpn
## us_govt_2y                                 0.82        0.73
## us_govt_5y                                 0.85        0.75
## us_govt_10y                                0.81        0.72
## fannie_4cpn                                0.97        0.86
## Mid_Yield_To_Convention                    1.00        0.90
## fannie_5cpn                                0.90        1.00
```

As can be seen by a quick glance, this matrix is highly correlated. This confirms that Treasury notes and mortgage rates are highly correlated and will generally move in tandem with each other. This is an economically sound principle, as it has been observed that interest rates tend to become more synchronized in a distressed market environment, as the current market is slowly becoming.

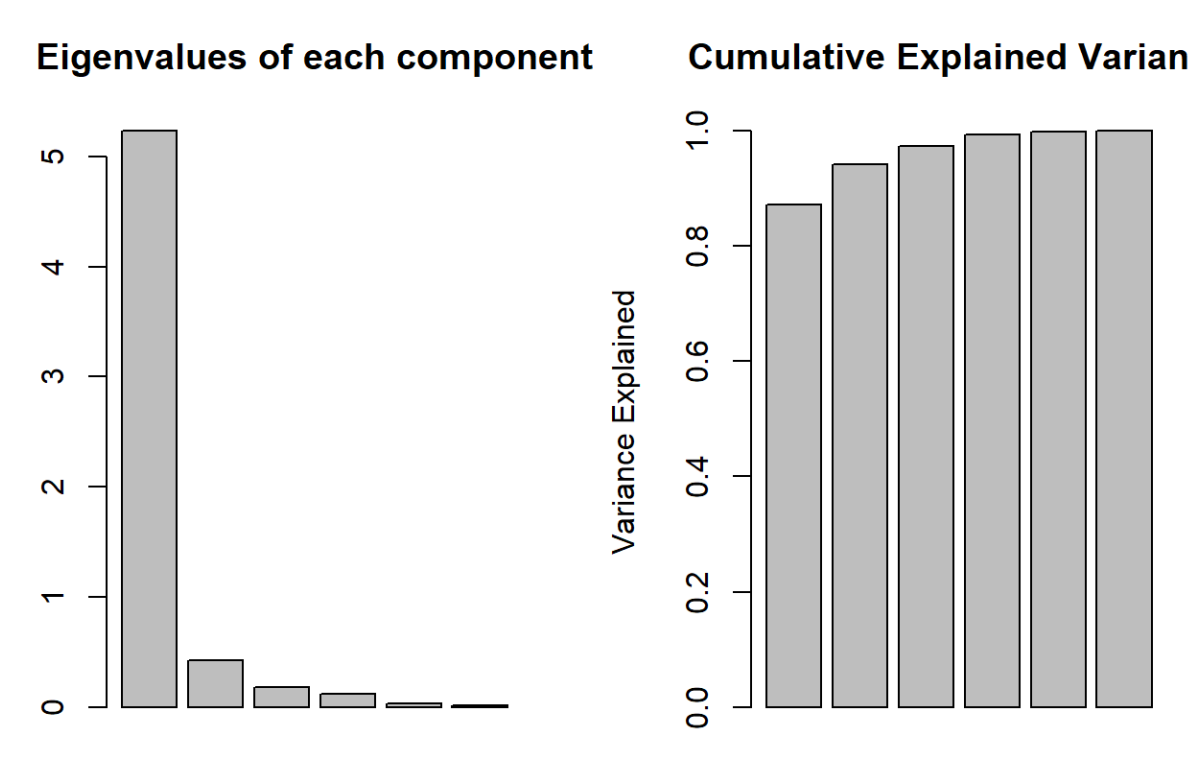
```r
model <- prcomp(cor.returns, scale = TRUE, center = TRUE)
summary(model)
```

```
## Importance of components:
##                          PC1   PC2    PC3    PC4     PC5     PC6
## Standard deviation     2.288 0.648 0.4279 0.3488 0.16655 0.11599
## Proportion of Variance 0.872 0.070 0.0305 0.0203 0.00462 0.00224
## Cumulative Proportion  0.872 0.942 0.9729 0.9931 0.99776 1.00000
```

```r
# first principal component sign flip so it's an upward shift in yield
model$x[,1] <- -model$x[,1]
# also have to flip sign for eigenvector
model$rotation[,1] <- -model$rotation[,1]

# how much variance is captured in first n eigenvectors?
variance1 <- cumsum(model$sdev^2)/sum(model$sdev^2)
print(variance1)
```

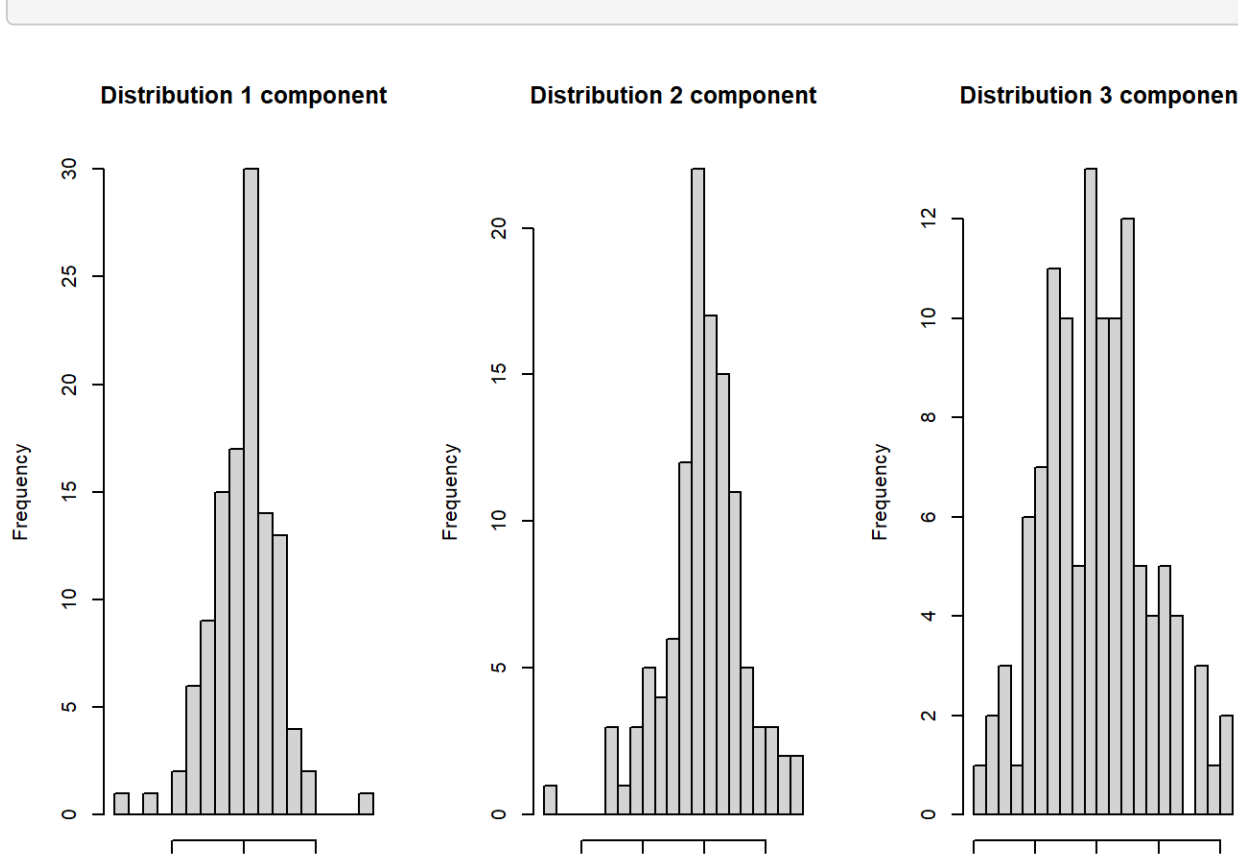
```
## [1] 0.87 0.94 0.97 0.99 1.00 1.00
```

```r
par(mfrow = c(1,2))
barplot(model$sdev^2, main = 'Eigenvalues of each component')
barplot(cumsum(model$sdev^2)/sum(model$sdev^2), main = 'Cumulative Explained Variance', ylab = 'Variance Explained')
```

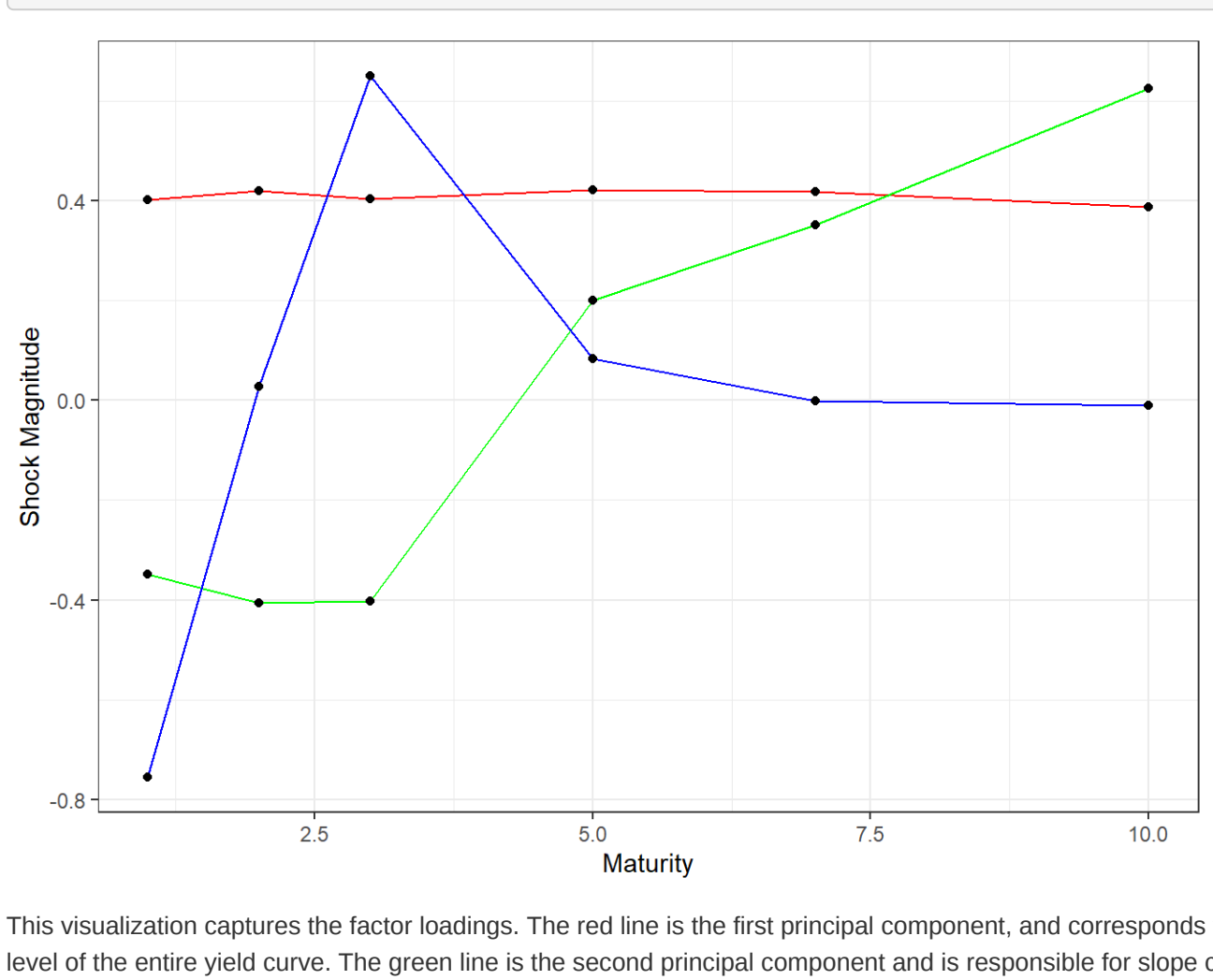


As can be seen from the Cumulative Explained Variance graph, the projection over three principal components can explain ~98% of the variance across all 6 contracts. Next we can look at the factor loadings to see how each one of them affect the returns of yields.

```r
par(mfrow = c(1,3))
hist(model$x[,1], breaks = 20, main = 'Distribution 1 component', xlab = paste('SD', round(model$sdev[1],2)))
hist(model$x[,2], breaks = 20, main = 'Distribution 2 component', xlab = paste('SD', round(model$sdev[2],2)))
hist(model$x[,3], breaks = 20, main = 'Distribution 3 component', xlab = paste('SD', round(model$sdev[3],2)))
```



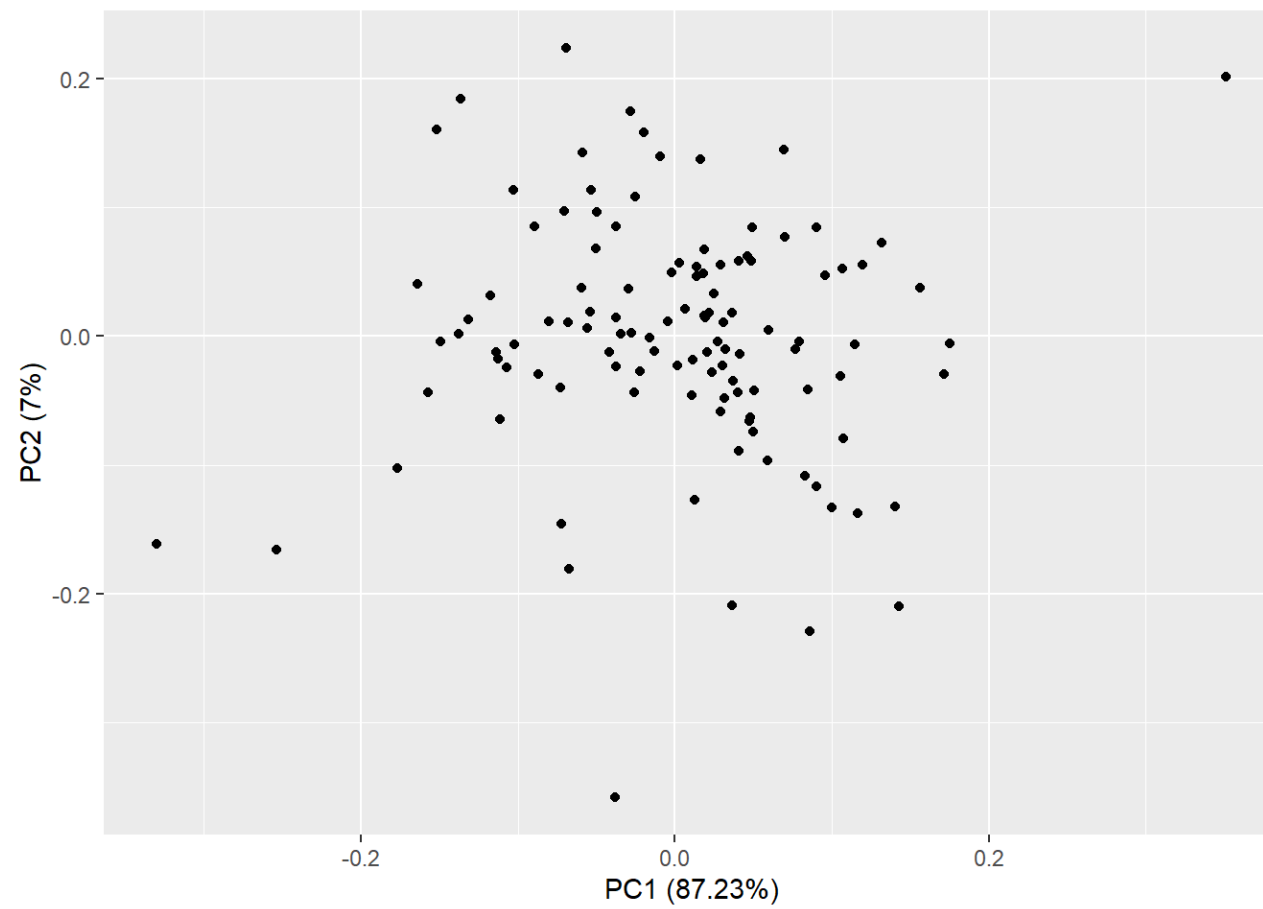
```r
# plot loading factors
library(ggplot2)
ggplot(data.frame(cbind(term=c(1,2,3,5,7,10),model$rotation[,1:3]))) +
  geom_line(aes(x=term,y=PC1),colour="red") + geom_point(aes(x=term,y=PC1)) +
  geom_line(aes(x=term,y=PC2),colour="green") + geom_point(aes(x=term,y=PC2)) +
  geom_line(aes(x=term,y=PC3),colour="blue") + geom_point(aes(x=term,y=PC3)) +
  xlab("Maturity") +
  ylab("Shock Magnitude") +
  theme_bw()
```



This visualization captures the factor loadings. The red line is the first principal component, and corresponds to a parallel move up and down in the level of the entire yield curve. The green line is the second principal component and is responsible for slope change and the third component displays the convexity.

We can plot the first two principal components as a scatterplot to see if there are any interesting features such as outliers or departures from normality.

```r
library(ggfortify)
pca.plot <- autoplot(model, data = cor.returns, color = "Group")
pca.plot
```



There are a few outliers, but the first two principal components are shown to behave within the limits of multivariate normality.

Now that it has been shown that the PCA is functioning correctly, we can interpret our findings. In order to interpret each component, we can compute the correlations between the original data and each principal component in order to see which PCs are most correlated with the bonds.

```r
#cor.returns is the differenced dataset
#model$x is the PCA data

#take the correlation matrix to interpret the correlation between of the original variables and the PCs
comps.scaled <- pca.plot$data[,c(1:12)]
cor(comps.scaled[,3:8], comps.scaled[,c(1:2,9)])
```

```
##                          PC1   PC2      PC3
## us_govt_2y              0.92 -0.23 -0.32326
## us_govt_5y              0.96 -0.26  0.01158
## us_govt_10y             0.92 -0.26  0.27783
## fannie_4cpn             0.97  0.13  0.03551
## Mid_Yield_To_Convention 0.95  0.23 -0.00052
## fannie_5cpn             0.88  0.40 -0.00464
```

This shows that the first principal analysis is positively correlated with the bonds, meaning that the bonds vary together and as one goes down, the others will decrease as well. This analysis allows us to conclude that the bonds are highly correlated to each other and as PCA analysis is shown to explain ~98% of the bond variance through three principal components, we can use the residuals to determine how the bonds are performing relative to each other.

The principal component analysis has served to confirm that mortgage rates and Treasury bills are positively correlated and in times of market distress, such as is being seen now, become even more positively correlated. From this, it seems reasonable to expect that in the current market, where Fed rate hikes are imminent and the recent housing market report was abysmal, mortgage rates will continue to rise alongside interest rates. In essence, the mortgage market currently hinges on inflation. The expectation in the market right now is that the Fed will keep increasing rates until it can be shown that inflation has peaked. This means that Treasury bonds and by extension, mortgage rates, will continue to increase until inflation is seen to have peaked. Because of the perceived market distress, the mortgage rates and Treasury notes will continue to become more positively correlated until the market can show it is not distressed (inflation will need to show it has peaked for this to happen). This sentiment is reflected in National Association of Realtors Chief Economist Lawrence Yun's comments when he said "Today's announcement by the Federal Reserve set a big increase in interest rates and means several more rounds of rate hikes are on the way in upcoming months … rental demand will strengthen along with rents. Only when consumer price inflation tops out and starts to fall will mortgage rates stabilize or even decline a bit."

The expectation after the recent Fed Rate hike and the underwhelming housing starts/permit number report this past week seems to be that the mortgage rates will keep climbing until inflation is shown to have reached a peak. Currently, the mortgage bonds and US Treasury Bonds are exhibiting an increasing positive correlation that shows that the mortgage market's performance is dependent upon inflation subsiding before mortgage rates can stabilize.

Resources

https://www.moodysanalytics.com/-/media/whitepaper/2014/2014-29-08-PCA-for-Yield-Curve-Modelling.pdf

https://insightr.wordpress.com/2017/04/14/american-bond-yields-and-principal-component-analysis/

https://arxiv.org/ftp/arxiv/papers/1911/1911.07288.pdf https://www.r-bloggers.com/2021/12/easy-interpretations-of-adf-test-in-r/

http://nakisa.org/bankr-useful-financial-r-snippets/principal-component-analysis/

https://research-doc.credit-suisse.com/docView?language=ENG&format=PDF&source_id=csplusresearchcp&document_id=1001969281&serialid=EVplkK6oNi2Oum067aSBs%2Bp%2F04%2F3pgbDBc%2B

http://www.bondeconomics.com/2018/12/primer-understanding-principal.html

https://online.stat.psu.edu/stat505/lesson/11/11.4