Untitled

YiTao Hu, Charles Rambo, Jin(jane) Huangfu, Junyu(Kevin) Wu

27/02/2020

```
#import data and libraries
library(readr)
library(dplyr)
library(DataAnalytics)
Industry_Port_rtn= read_csv("48_Industry_Portfolios.CSV",
     col_types = cols(FabPr = col_skip(),
         Gold = col_skip(), Guns = col_skip(),
        Hlth = col_skip(), Soda = col_skip()))
FF3Factors=read_csv("F-F_Research_Data_Factors.CSV")
FF25Port=read_csv("25_Portfolios_5x5.CSV")
```

First,we need to preprocess data,compute the risk preium of each industry portfolio

```
#compute risk-premium
Industry_Port_rtn=Industry_Port_rtn-FF3Factors$RF
Industry_Port_rtn=Industry_Port_rtn/100
FF3Factors=FF3Factors/100
FF25Port=FF25Port/100
```

1

Perform PCA

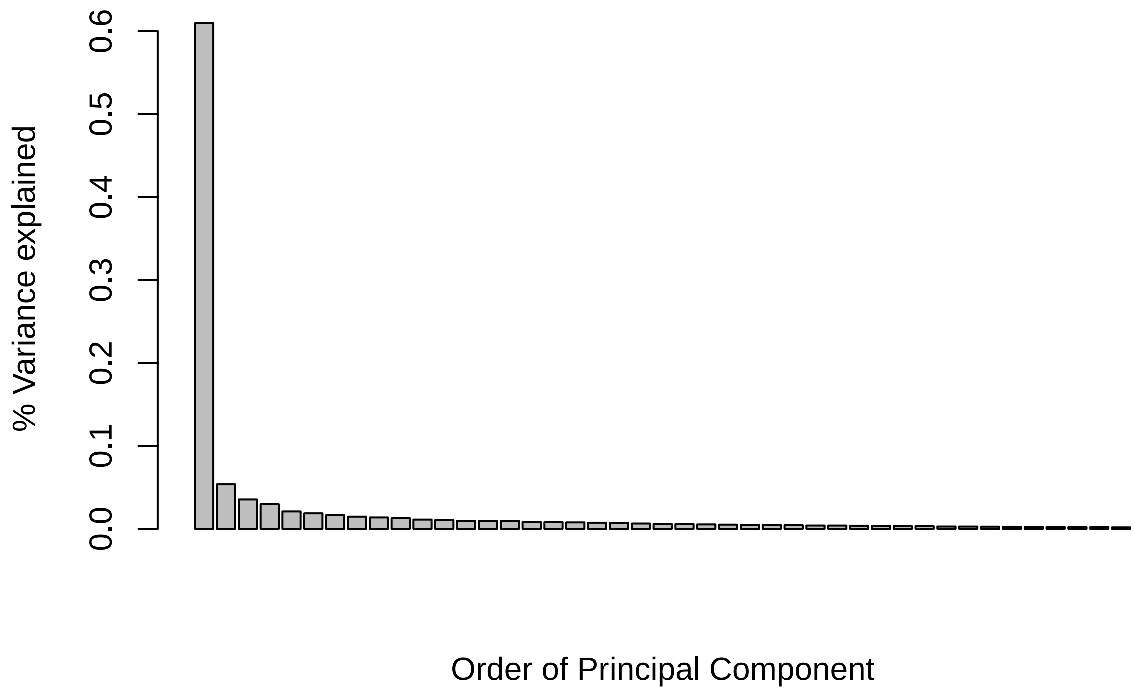
```
PCA_model=prcomp(Industry_Port_rtn[-1])
```

To compute percentage of variance explained by each principal component, we use the following formula.

$$Pct(Ci) = \frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i}$$

where N is the number of total component. λ_i is the ith eigenvalue of the var-cov matrix.

```
eigen_vals=PCA_model$sdev^2
pct_var=eigen_vals/sum(eigen_vals)
#plot the bar chart of pct variance explained by each PC
barplot(pct_var,xlab = 'Order of Principal Component',ylab = '% Variance explained')
```



2

a. The variance first three PCs can explain

```r
sum(pct_var[1:3])
```

```
## [1] 0.698805
```

b. To compute the principal component, we have to use the following formula:

$$\vec{y_t} = U_{reduce}\vec{r_t}$$

where $\vec{r_t}$ is the industry risk preium at time t, $\vec{y_t}$ is first k PCs' value at time t and U_{reduce} is a matrix stacked by first k orthonormal eigen-vectors.

```r
U=PCA_model$rotation
U_reduce=U[,1:3]
#compute the series of first 3 PC
PC3=as.matrix(Industry_Port_rtn[-1])%*%U_reduce
```

Here, we compute the mean and SD of the first 3 PC

```r
descStat(PC3)
```

```
##       Mean Median    SD   IQR SE Mean 95% CI-L 95% CI-U NMissing
## PC1  0.038  0.053 0.326 0.388   0.013    0.013    0.062        0
## PC2  0.002  0.007 0.097 0.110   0.004   -0.005    0.009        0
## PC3 -0.005 -0.004 0.079 0.073   0.003   -0.011    0.001        0
## Number of Observations =  672
```

Here, as we can see, the 3 PC has 0 correlation because they are orthogonized during the SVD.

```r
cor(PC3)
```

```
##               PC1           PC2           PC3
## PC1  1.000000e+00  7.355087e-15 -5.623563e-16
```

```
## PC2  7.355087e-15  1.000000e+00 -2.425297e-17
## PC3 -5.623563e-16 -2.425297e-17  1.000000e+00
```

c. Because the factor loadings $\vec{\beta_i}$ should just be the row vectors of the reduced factorized matrix $U_{reduced}$ for each industry portfolio, the predictd returns for all industries can be computed from the formula below:

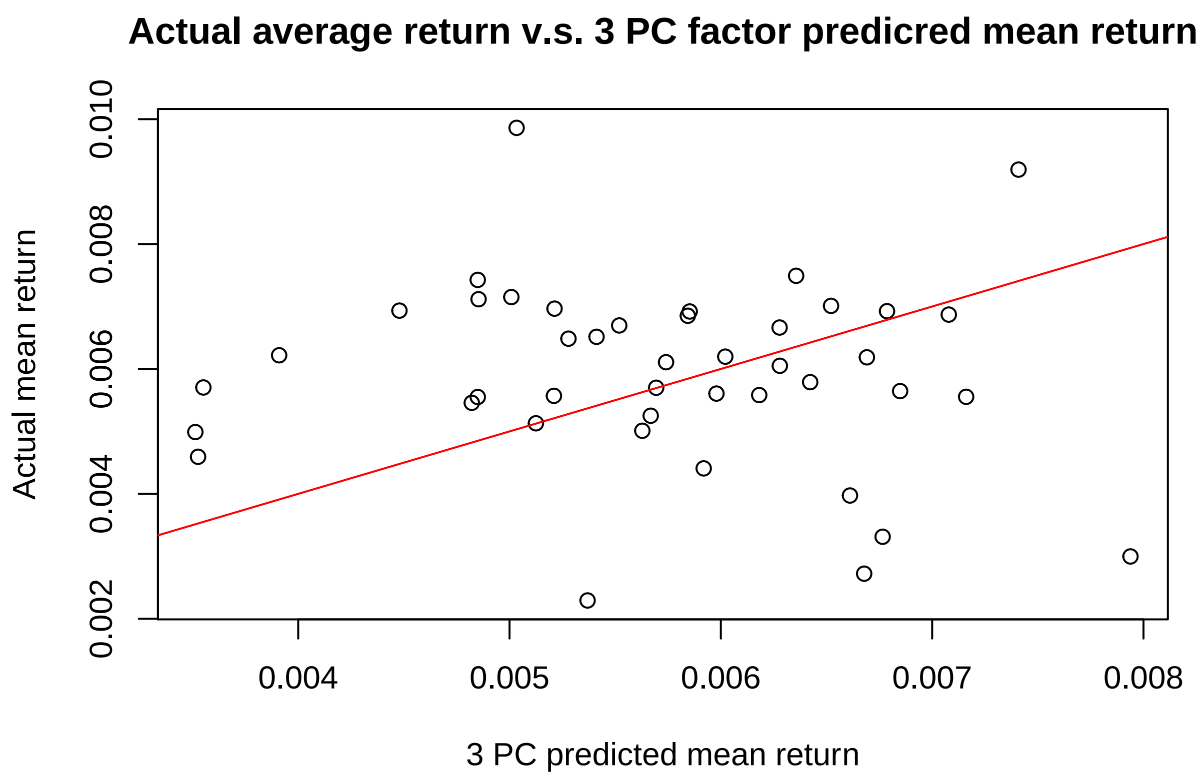
$$\hat{R} = YU_{reduce}^T$$

where $\hat{R}$ is a 672 (timestep) by 43 (industry) matrix, and Y is a 672 by 3 (factors) matrix and U_{reduce} is a 43 (industry) by 3 (factors) matrix

```
R_predicted=PC3%*%t(U_reduce)
```

Then we can plot actual sample returns over expected sample returns

```
plot(x = colMeans(R_predicted),y = colMeans(Industry_Port_rtn[-1]),type = 'p',main = 'Actual average re
abline(c(0,1),col='red')
```



d.Then we can compute the implied cross-section R^2, which should be close to the percentage of variance explained by first three princial components.

$$R^2_{cross-section} = 1 - \frac{Var(\hat{R})}{Var(R^{\bar{act}})}$$

```
R_sq=1-mean(rowSums((Industry_Port_rtn[,-1]-R_predicted[,])^2))/mean(rowSums(Industry_Port_rtn[,-1]^2))
R_sq
```

```
## [1] 0.7006915
```

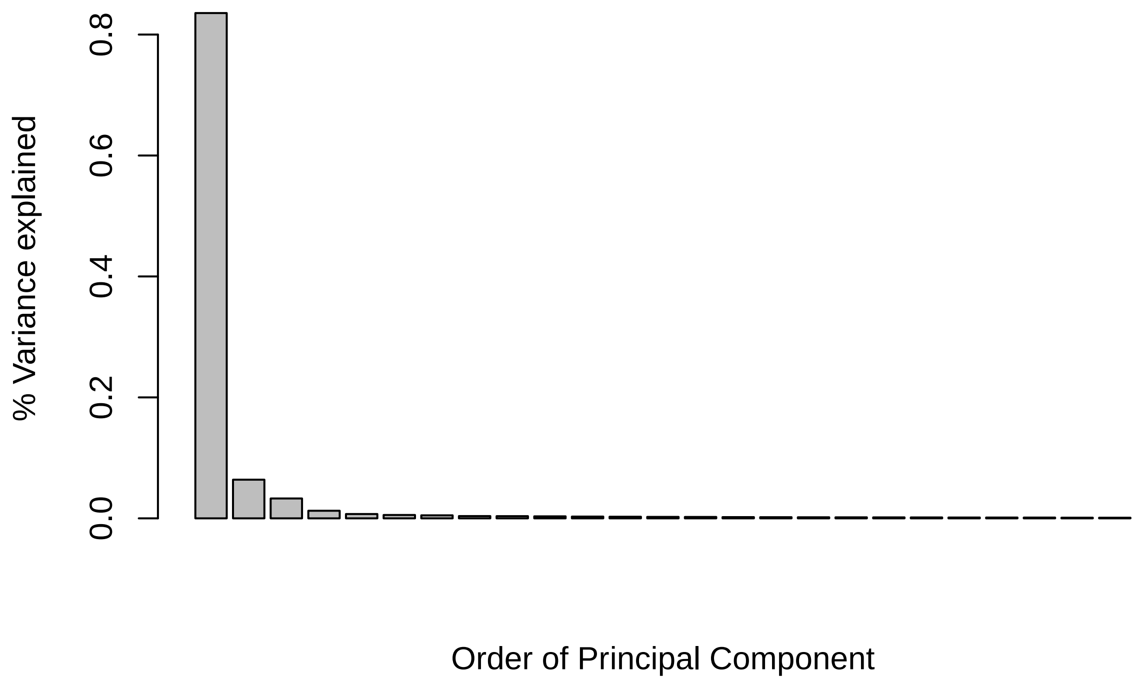
3

a. Perform PCA on the FF 25 portfolio

```r
PCA_FF25Port=prcomp(FF25Port[,-1])
```

Plot the percentage variance explained by each principal component

```r
eigen_vals=PCA_FF25Port$sdev^2
pct_var=eigen_vals/sum(eigen_vals)
#plot the bar chart of pct variance explained by each PC
barplot(pct_var,xlab = 'Order of Principal Component',ylab = '% Variance explained')
```



b. We can see the cumulative explanatory power of increasing factors, and then decide how many factors we want.

```r
cum_pct_var=cumsum(pct_var)
names(cum_pct_var)=seq(1,length(cum_pct_var))
cum_pct_var
```

```
##         1         2         3         4         5         6         7
## 0.8355389 0.8994655 0.9323277 0.9448999 0.9520442 0.9575955 0.9625748
##         8         9        10        11        12        13        14
## 0.9663161 0.9699263 0.9732397 0.9761392 0.9787953 0.9811856 0.9834872
##        15        16        17        18        19        20        21
## 0.9855210 0.9874244 0.9891655 0.9908557 0.9924409 0.9939592 0.9953252
##        22        23        24        25
## 0.9966314 0.9978930 0.9989713 1.0000000
```

From the table above, we can see if we want 95% explanatory power, we need to keep 5 factors.