Empirical Methods in Finance
Homework 5: Solution

Prof. Lars A. Lochstoer
TAs: Chady Gemayel and Mahyar Kargar

March 4, 2019

Principal Component Analysis

Download the 48 industry portfolio data (monthly) from Kenneth French's web site. Use the data from 1960 through 2015. Use the value-weighted returns. You may drop the industries that have missing values and are reported as -99.99. Also, download the 3 Fama-French factors from his web site. Use the monthly risk-free rate series provided by French in the same FF factor dataset to compute excess returns on these 48 portfolios.

1. Get the eigenvalues for the sample variance-covariance matrix of the excess returns to the 48 industries. Plot the fraction of variance explained by each eigenvalue in a bar plot.

   Suggested solution:
   We first subtract risk free rate from industry returns to get the excess returns. Then we get the variance-covariance matrix of the industry excess returns and perform eigen decomposition of this variance-covariance matrix to get the eigenvalues (λs) and eigenvectors. The fraction of total variance explained by the i-th eigenvalue is $\lambda_i / \sum_j \lambda_j$. Figure 1 plots the fraction of variance explained by each eigenvalue.

2. Choose the 3 first (largest) principal components.

   (a) How much of the total variance do these 3 factors explain?
   Suggested solution:

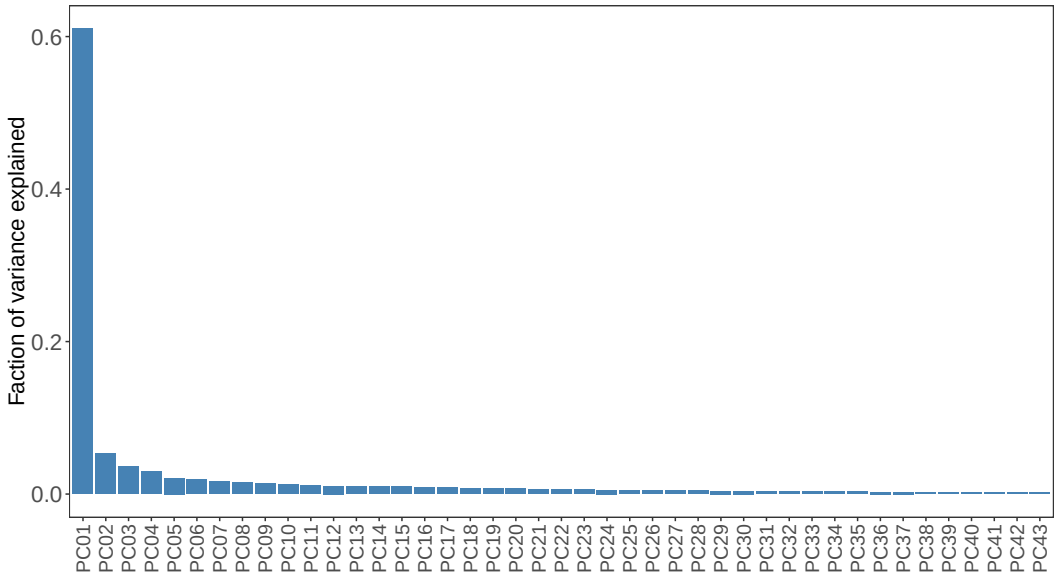

Figure 1: Fraction of variance explained by each eigenvalue

```
# using eigenvalues
sum(eigs$values[1:3])/sum(eigs$values)

## [1] 0.6994112

# another way: PCA
PCA <- prcomp(ind_exret_nodate)
imp <- summary(PCA)$importance[2,]
sum(imp[1:3])/sum(imp)

## [1] 0.699417
```

The first 3 PCs explain 69.94% of the total variance of the industry excess returns.

(b) Give the mean sample return to these 3 factor portfolios, their standard deviation, and correlation.

Suggested solution:

The i-th principal component is given by: $y_t = \mathbf{e}_i' \mathbf{r}_t$, where $\mathbf{e}_i$ is the i-th eigenvector and $\mathbf{r}_t$ is the time-series of the industry excess returns. So, we get the loadings for each factor from the corresponding eigenvector and then calculate the mean return, variance and correlation of the first 3 PCs as shown in Table 1. Note that the variance of the i-th PC is equal to the i-th eigenvalue λ_i and the correlations are almost 0. For the mean returns, I also have t-stats in the brackets. We notice that the mean returns for the second and third factor are not significantly different

Table 1: Summary statistics for the first three factors

Mean			Variance			Correlation		
PC1	PC2	PC3	σ_1^2	σ_2^2	σ_3^2	ρ_{12}	ρ_{13}	ρ_{23}
3.78 (3.01)	0.22 (0.59)	−0.51 (−1.69)	1061.78	93.55	62.06	≈ 0	≈ 0	≈ 0

from zero.

```
# calculate the first 3 PCs
PC1 <- ind_exret_nodate %*% PCA$rotation[,1]
PC2 <- ind_exret_nodate %*% PCA$rotation[,2]
PC3 <- ind_exret_nodate %*% PCA$rotation[,3]
PC1_3 <- cbind(PC1,PC2,PC3)
# covaraince matrix of PC1-3
cov(PC1_3)

##               [,1]         [,2]          [,3]
## [1,]  1.061779e+03 2.179900e-12 -9.508183e-14
## [2,]  2.179900e-12 9.354744e+01  4.909308e-15
## [3,] -9.508183e-14 4.909308e-15  6.206336e+01

# sample mean of PC1-3
colMeans(PC1_3)

## [1]  3.7786699  0.2156594 -0.5141413

# std errors
std_err <- function(x) sd(x)/sqrt(length(x))
std_errs <- apply(PC1_3, 2, std_err)

# t_stats
(t_stat <- colMeans(PC1_3)/std_errs)

## [1]  3.0061215  0.5780122 -1.6918018
```

Figure 2 plots how each industry loads on the first three PCs. These loadings are the first three eigenvectors of the covariance matrix of industry excess returns. In general, it is not easy to interpret the principal components. However, from Figure 2, it is easy to see that PC1 represent a common factor that drives all

industry excess returns. We can thus interpret PC1 as the market factor. Unlike the zero-coupon yields we went over in the TA session, PC2 and PC3 are harder to interpret here.

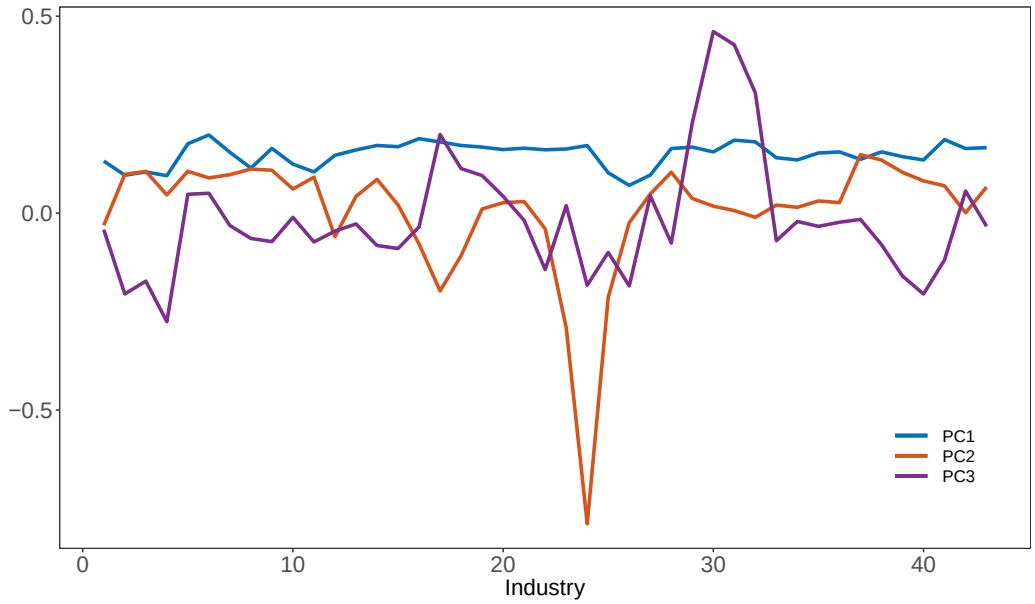

Figure 2: Industry loadings on the first 3 PCs

(c) Consider a multi-factor model of returns using these three factors as pricing factors. Plot the predicted return from this model for all the industries versus the realized average industry returns. Add a 45 degree line to this plot (as in the lecture note). [Recall you get factor loadings (betas) from the eigenvectors. Or, if you like, you can run the time-series regression of each industry's return on the 3 factors. The result is the same.]

Suggested solution:

The first three eigenvectors are the betas for the three PC factors. We multiply the betas by their corresponding factors to get the predicted return:

$$r_t = \beta_1 \, \mathrm{PC1}_t + \beta_2 \, \mathrm{PC2}_t + \beta_3 \, \mathrm{PC3}_t$$

Figure 3 plots the predicted against the realized industry excess returns. Since most of the dots are not lining up on the 45 degree line, this model does a poor job explaining industry excess returns.

(d) Give the implied cross-sectional R^2 of the plot in c).

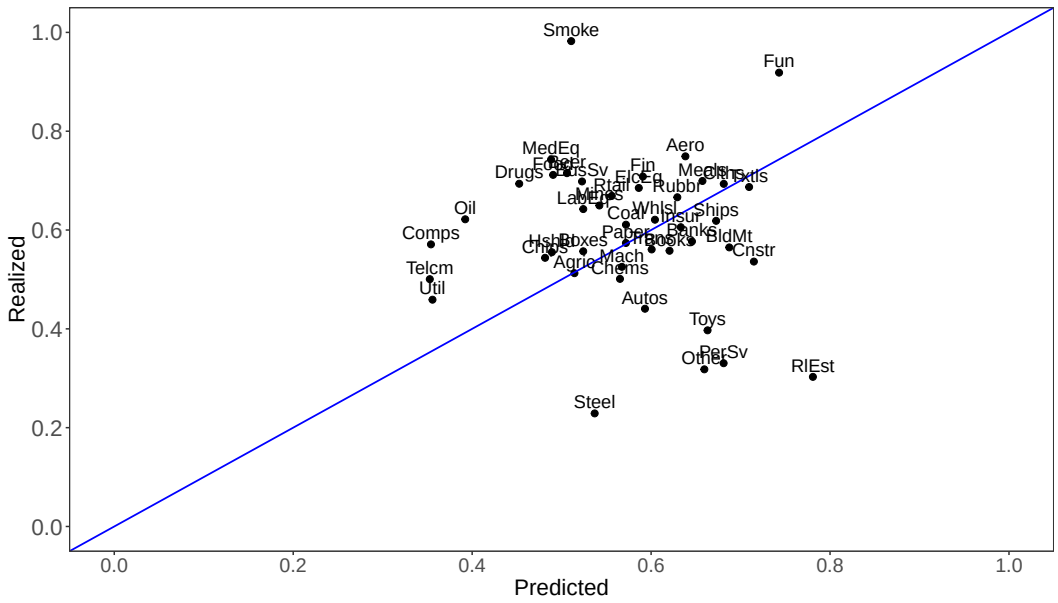

Figure 3: Predicted vs. realized average industry excess returns

Suggested solution:

The implied cross-sectional R^2 is -0.57. This tells us that the three largest PCs explaining the variation in industry returns are most likely not priced (lambda = 0) and pricing errors from this model are quite large. Also recall from Table 1 the mean returns of the second and third factors are not significantly different from zero. This is also evident from the poor cross-sectional fit in Figure 3.

3. Now, download the 25 FF portfolios, same sample period.

   (a) Get the eigenvalues for the sample variance-covariance matrix of the excess returns to these 25 F-F portfolios. Plot the fraction of variance explained by each eigenvalue in a bar plot.

   Suggested solution:

   We use the value-weighted 25 FF portfolios and repeat the same exercise as before. Figure 4 shows the fraction of variance that is explained by each PC.

   (b) Given (a), how many factors does do you reckon you need to explain average returns to the 25 F-F portfolios?

   Suggested solution:

   Observing Figure 4, we can see that first three PCs should be able to explain

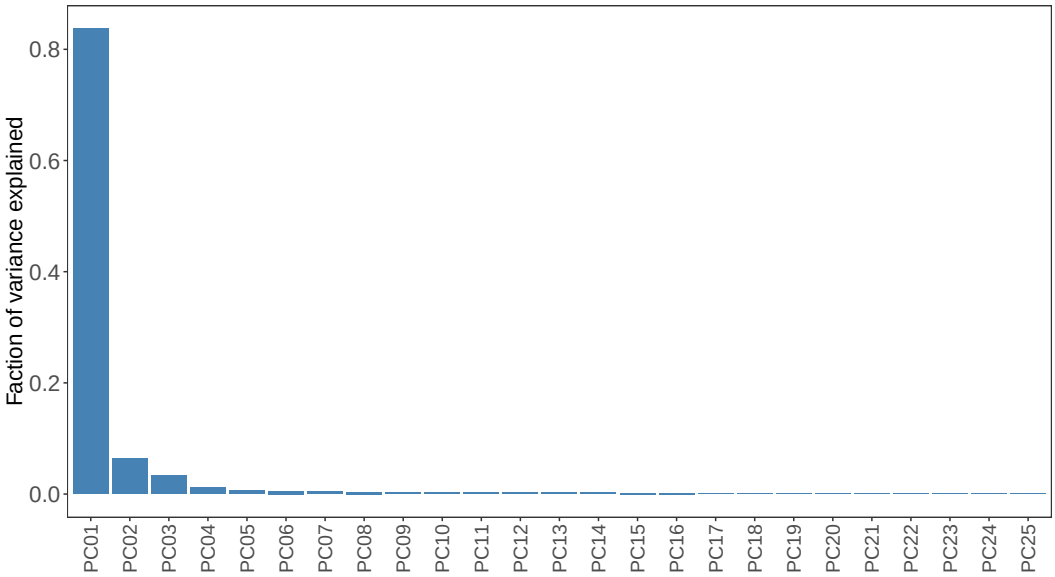

Figure 4: Fraction of variance explained by each eigenvalue

Table 2: Cumulative proportion of variance explained by the fist 3 PCs for 25 FF portfolio

PC1	PC2	PC3
83.67%	90.01%	93.29%

most of the variation in data. In fact, Table 2 reports the cumulative proportion of variance explained by the first 3 factors. The first 3 PCs explain more than 93% of the excess returns on these portfolios. Also, since these portfolios are sorted based on size and book-to-market, from Fama and French (1992, 1993), we know that three factors (MktRf, SMB, and HML) can explain the cross-section of size/book-to-market portfolios. So, I would choose 3 factors to explain average returns to the 25 FF portfolios.

```
# Cumulative variance explained by each PC
imp_FF25 <- summary(PCA_FF25)$importance[2,]
imp_FF25[1]/sum(imp_FF25)

##     PC1
## 0.83668

sum(imp_FF25[1:2])/sum(imp_FF25)
```

```
## [1] 0.90017

sum(imp_FF25[1:3])/sum(imp_FF25)

## [1] 0.93288
```

Figure 5 shows the predicted against realized excess returns for the 25 FF portfolios using first three PCs as pricing factors. The cross-sectional R^2 is 0.59. So we get a much better cross-sectional fit than the 3-factor model for the industry portfolios above.

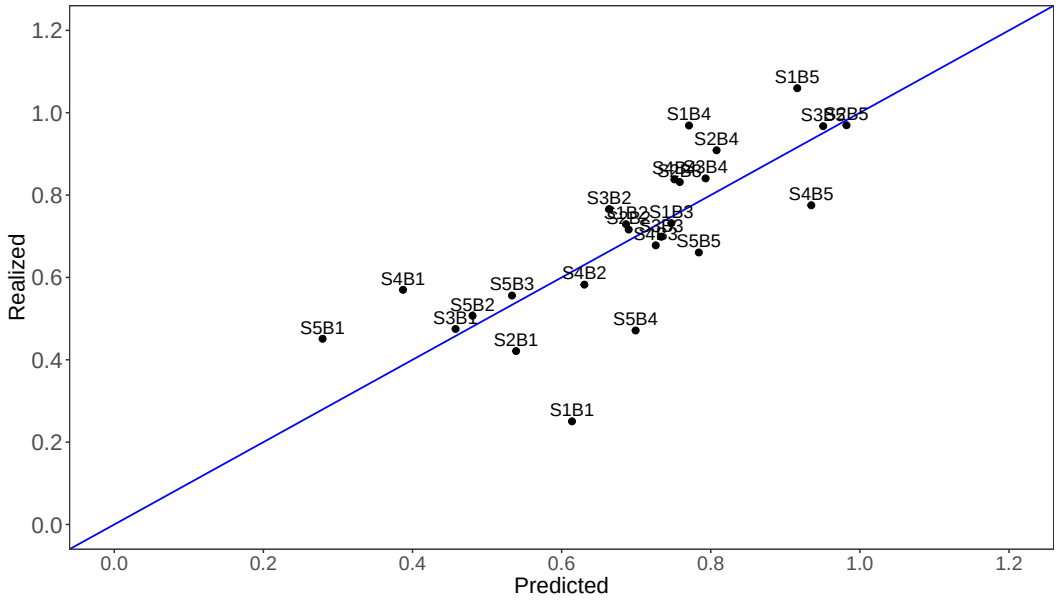

Figure 5: Predicted vs. realized average FF25 excess returns

Suggested R Code

```
#################
# Code for HW 5 #
#################
library(data.table)
library(ggplot2)
library(zoo)
library(lubridate)
library(sandwich)

rm(list=ls())
options(max.print=999999)

FF <- fread("./F-F_Research_Data_Factors.CSV")
setnames(FF,c("V1","Mkt-RF"),c("date","MktRF"))
FF[,`:=`(year=date%/%100, month=date%%100)]
FF[,date:=as.yearmon(paste(year,month,sep="-"))]
FF[,`:=`(year=NULL,month=NULL)]
FF <- FF[year(date)>=1960]

# get industry returns
industry <- fread("industry.csv")
setnames(industry,"X","date")
industry[,`:=`(year=date%/%100, month=date%%100)]
industry[,date:=as.yearmon(paste(year,month,sep="-"))]
industry[,`:=`(year=NULL,month=NULL)]
industry <- industry[year(date) %in% 1960:2015]

# dropping  the industries that have missing values and are reported as -99.99
industry <- as.data.frame(industry)
i=2
while (i<=ncol(industry)){
  if(industry[1,i]==-99.99) industry[,i]=NULL else i=i+1
}

industry <- as.data.table(industry)
setkey(FF,date)
setkey(industry,date)
reg <- merge(FF,industry)
ind_exret <- reg[,lapply(.SD, function(x) x-RF),
```

```
                .SDcols=-c("MktRF","SMB","HML","RF"),by=date]


setkey(ind_exret,date)
ind_exret_nodate <- copy(ind_exret)
ind_exret_nodate[,date:=NULL]

# eigenvalue decomposition
Sigma <- cov(ind_exret_nodate)
eigs <- eigen(Sigma)

# fraction of variance explained by the first 3 PCs
# using eigenvalues
sum(eigs$values[1:3])/sum(eigs$values)

# another way: PCA
PCA <- prcomp(ind_exret_nodate)
PCA_ind_summ <- summary(PCA)
imp <- summary(PCA)$importance[2,]
sum(imp[1:3])/sum(imp)


names(imp)[1:9] <- c("PC01","PC02","PC03","PC04","PC05","PC06","PC07","PC08","PC09")
prop_var <- data.table(PC=names(imp),imp=imp)

# Fraction of variance explained by each PC
ggplot(prop_var, aes(x=PC, y=imp)) +
  geom_bar(stat="identity", fill="steelblue", position=position_dodge()) +
  theme_bw() + xlab("") + ylab("Faction of variance explained") +
  theme(axis.text.x = element_text(angle = 90, vjust = .5)) +
  theme(axis.text.x = element_text(size=14,face="plain"),
        axis.text.y = element_text(size=16,face="plain"),
        axis.title.x = element_text(size=16,face="plain"),
        axis.title.y = element_text(size=16,face="plain"),
        legend.text = element_text(size=9,face="plain"),legend.position = c(0.9,0.2)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(legend.margin=margin(t=-0.5, r=0, b=0, l=0, unit="cm"))

names_ind <- names(ind_exret_nodate)
ind_exret_nodate <- as.matrix(ind_exret_nodate)
```

```r
# calculate the first 3 PCs
PC1 <- ind_exret_nodate %*% PCA$rotation[,1]
PC2 <- ind_exret_nodate %*% PCA$rotation[,2]
PC3 <- ind_exret_nodate %*% PCA$rotation[,3]
PC1_3 <- cbind(PC1,PC2,PC3)

# summary stats for the first 3 PCs
# covaraince matrix of PC1-3
cov(PC1_3)

# sample mean of PC1-3
colMeans(PC1_3)

# std errors
std_err <- function(x) sd(x)/sqrt(length(x))
std_errs <- apply(PC1_3, 2, std_err)

# t_stats
(t_stat <- colMeans(PC1_3)/std_errs)

# Plot industry loadings on the first 3 PCs
PCs <- data.table(indx=1:43,PC1=PCA$rotation[,1],PC2=PCA$rotation[,2],
                  PC3=PCA$rotation[,3])

ggplot(PCs) + geom_line(aes(x=indx,y=PC1,color = "PC1"),size=1.2) +
  geom_line(aes(x=indx,y=PC2,color = "PC2"),size=1.2) +
  geom_line(aes(x=indx,y=PC3,color = "PC3"),size=1.2) +
  theme_bw() + xlab("Industry") + ylab("") + #ggtitle("Industry Loadings on PCs") +
  scale_colour_manual("",breaks = c("PC1", "PC2","PC3"),
                      values = c("PC1"= rgb(0,0.447,0.741), "PC2"=rgb(0.85,0.325,0.098),
                                 "PC3"=rgb(0.494,0.184,0.5560)),
                      labels = c("PC1","PC2","PC3")) +
  theme(title = element_text(size=16,face="bold"),
        axis.text.x = element_text(size=16,face="plain"),
        axis.text.y = element_text(size=16,face="plain"),
        axis.title.x = element_text(size=16,face="plain"),
        axis.title.y = element_text(size=16,face="plain"),
        legend.text = element_text(size=12,face="plain"),legend.position = c(0.9,0.2)) +
  theme(legend.key.size =  unit(0.4, "in"),legend.key.height =  unit(0.2, "in"),
        legend.key.width =  unit(0.4, "in")) +
```

```
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        legend.background = element_rect(fill=alpha(0.4)))

# Compute the predicted return from 3 PCs
PC1_3 <- t(PC1_3)
predict <- PCA$rotation[,1:3]%*%PC1_3
avgPredict <- apply(predict,1,mean)
avgActual <- apply(ind_exret_nodate,2,mean)
dt_ind <- data.table(predicted_ret=avgPredict, realized_ret=avgActual,
                     names=names_ind)

# plot predicted vs. realized returns
ggplot(dt_ind,aes(x=predicted_ret,y=realized_ret))+geom_point(size = 2) +
  xlab("Predicted") + ylab("Realized") +
  geom_abline(color="blue", size = .6) + theme_bw() +
  scale_x_continuous(limits = c(0,1), breaks = seq(0,1,by = .2)) +
  scale_y_continuous(limits = c(0,1), breaks = seq(0,1,by = .2)) +
  geom_text(aes(label=names),hjust=.5, vjust=-0.4, size = 4.5) +
  theme(axis.text.x = element_text(size=14,face="plain"),
        axis.text.y = element_text(size=16,face="plain"),
        axis.title.x = element_text(size=16,face="plain"),
        axis.title.y = element_text(size=16,face="plain"),
        legend.text = element_text(size=9,face="plain"),legend.position = c(0.9,0.2)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(legend.margin=margin(t=-0.5, r=0, b=0, l=0, unit="cm"))

# R^2
(Rsquared <- 1-var(avgActual-avgPredict)/var(avgActual))


##################
# Fama-French 25 #
##################
FF25 <- fread("FFPort.csv")
setnames(FF25,c("date","S1B1","S1B2","S1B3","S1B4","S1B5",
                "S2B1","S2B2","S2B3","S2B4","S2B5",
                "S3B1","S3B2","S3B3","S3B4","S3B5",
                "S4B1","S4B2","S4B3","S4B4","S4B5",
                "S5B1","S5B2","S5B3","S5B4","S5B5"))
FF25[,`:=`(year=date%/%100, month=date%%100)]
```

```
FF25[,date:=as.yearmon(paste(year,month,sep="-"))]
FF25[,`:=`(year=NULL,month=NULL)]
FF25 <- FF25[year(date) %in% 1960:2015]

setkey(FF,date)
setkey(FF25,date)
reg <- merge(FF,FF25)
FF25_exret <- reg[,lapply(.SD, function(x) x-RF),
                  .SDcols=-c("MktRF","SMB","HML","RF"),by=date]


setkey(FF25_exret,date)
FF25_exret_nodate <- copy(FF25_exret)
FF25_exret_nodate[,date:=NULL]

# eigenvalue decomposition
Sigma_FF25 <- cov(FF25_exret_nodate)
eigs_FF25 <- eigen(Sigma_FF25)

# fraction of variance explained by the first 3 PCs
# using eigenvalues
sum(eigs_FF25$values[1:3])/sum(eigs_FF25$values)

# PCA
PCA_FF25 <- prcomp(FF25_exret_nodate)
imp_FF25 <- summary(PCA_FF25)$importance[2,]
sum(imp_FF25[1:3])/sum(imp_FF25)

# Cumulative variance explained by each PC
imp_FF25 <- summary(PCA_FF25)$importance[2,]
imp_FF25[1]/sum(imp_FF25)
sum(imp_FF25[1:2])/sum(imp_FF25)
sum(imp_FF25[1:3])/sum(imp_FF25)
sum(imp_FF25[1:4])/sum(imp_FF25)


names(imp_FF25)[1:9] <- c("PC01","PC02","PC03","PC04","PC05","PC06","PC07","PC08","PC09")
prop_var_FF25 <- data.table(PC=names(imp_FF25),imp=imp_FF25)

# Fraction of variance explained by each PC
```

```
ggplot(prop_var_FF25, aes(x=PC, y=imp)) +
  geom_bar(stat="identity", fill="steelblue", position=position_dodge()) +
  theme_bw() + xlab("") + ylab("Faction of variance explained") +
  theme(axis.text.x = element_text(angle = 90, vjust = .5)) +
  theme(axis.text.x = element_text(size=14,face="plain"),
        axis.text.y = element_text(size=16,face="plain"),
        axis.title.x = element_text(size=16,face="plain"),
        axis.title.y = element_text(size=16,face="plain"),
        legend.text = element_text(size=9,face="plain"),legend.position = c(0.9,0.2)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(legend.margin=margin(t=-0.5, r=0, b=0, l=0, unit="cm"))


# Compute the predicted return from 4 PCs
# calculate the first 3 PCs
names_FF25 <- names(FF25_exret_nodate)
FF25_exret_nodate <- as.matrix(FF25_exret_nodate)

# calculate the first 3 PCs
PC1_FF25 <- FF25_exret_nodate %*% PCA_FF25$rotation[,1]
PC2_FF25 <- FF25_exret_nodate %*% PCA_FF25$rotation[,2]
PC3_FF25 <- FF25_exret_nodate %*% PCA_FF25$rotation[,3]
PC1_3_FF25 <- cbind(PC1_FF25,PC2_FF25,PC3_FF25)

PC1_3_FF25 <- t(PC1_3_FF25)
predict_FF25 <- PCA_FF25$rotation[,1:3]%*%PC1_3_FF25
avgPredict_FF25 <- apply(predict_FF25,1,mean)
avgActual_FF25 <- apply(FF25_exret_nodate,2,mean)

dt_FF25 <- data.table(predicted_ret=avgPredict_FF25, realized_ret=avgActual_FF25,
                      names=names_FF25)

# plot predicted vs. realized returns
ggplot(dt_FF25,aes(x=predicted_ret,y=realized_ret))+geom_point(size = 2) +
  xlab("Predicted") + ylab("Realized") +
  geom_abline(color="blue", size = .6) + theme_bw() +
  scale_x_continuous(limits = c(0,1.2), breaks = seq(0,1.2,by = .2)) +
  scale_y_continuous(limits = c(0,1.2), breaks = seq(0,1.2,by = .2)) +
  geom_text(aes(label=names),hjust=.5, vjust=-0.4, size = 4.5) +
  theme(axis.text.x = element_text(size=14,face="plain"),
```

```
        axis.text.y = element_text(size=16,face="plain"),
        axis.title.x = element_text(size=16,face="plain"),
        axis.title.y = element_text(size=16,face="plain"),
        legend.text = element_text(size=9,face="plain"),legend.position = c(0.9,0.2)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(legend.margin=margin(t=-0.5, r=0, b=0, l=0, unit="cm"))


# R^2 FF25
Rsquared_FF25 <- 1-var(avgActual_FF25-avgPredict_FF25)/var(avgActual_FF25)
```