Factor Models & PCA

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

Preliminary Setup

Load Data

```
load("HW08_RT.Rdata")
source("FM_functions.R")
# Get data from Fama/French website
FF_data = read.table("F-F_Research_Data_Factors_weekly.txt", header=T)   # read in downloaded da
ta
ind.0   = which(rownames(FF_data) == 20060106)                           # remove earlier histor
ical data
FF_data = FF_data[-(1:(ind.0-1)),]
cat("dimension of FF_data:", dim(FF_data))
```

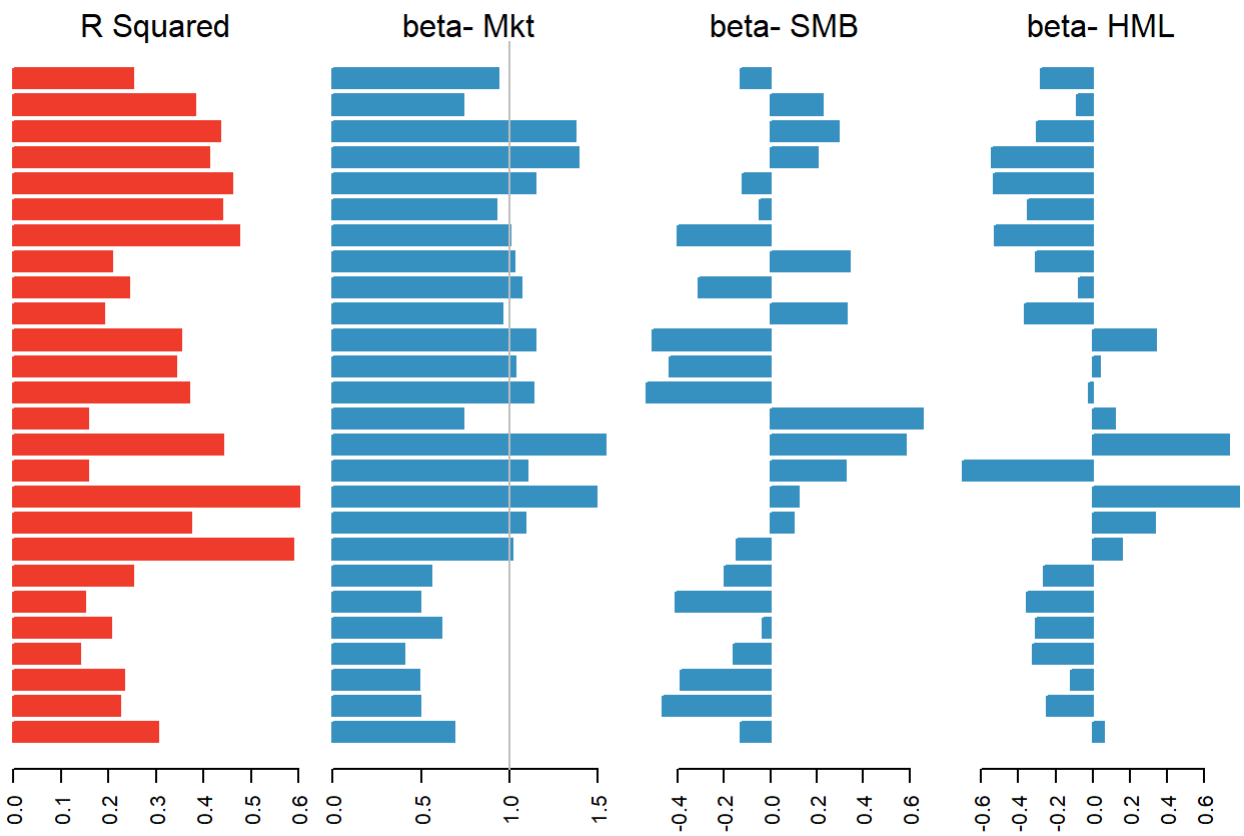
```
## dimension of FF_data: 791 4
```

```
attach(FF_data)
```

Question 1: Fama-French 3 Factor Model

1a) Plot a summary of the Rsq. and Betas.

```
Yt = apply(Rt,2, function(x) x-RF)    # subtract rf from weekly returns to get excess returns
dimnames(Yt)[[2]] = syb               # add symbol names
n = dim(Yt)[1]                        # define n = # returns for each stock
N = dim(Yt)[2]                        # define N = # of stocks
p = 3                                 # define p = # of factors
fit = lm(Yt~Mkt.RF+SMB+HML)           # fit the regression model on the factors
sfit = summary(fit)
# define Betas Vector
betas = coef(fit)[-1,]
# define R-squared vector
R.Squared = c()
for(i in 1:N) R.Squared[i] = sfit[[i]]$r.squared
# plot summary
coef.plot(R.Squared, betas)
```




```
# Check Aggressiveness by Industry
industry = c("Food", "Media", "Healthcare", "Tech")  # vector of industry names
Ns = c(7,6,6,7)                                      # count of securities in each industry
by_industry = c()
for(i in 1:4) by_industry = c(by_industry, rep(industry[i],Ns[i]))
table(Aggressive = coef(fit)[2,] > 1, by_industry)
```

```
##           by_industry
## Aggressive Food Healthcare Media Tech
##      FALSE    7          1     1    3
##      TRUE     0          5     5    4
```

```
# Summarize R2 by industry
R.Squared = c()
for(i in 1:N) R.Squared[i] = sfit[[i]]$r.squared
table(Hi_R.Sq = R.Squared > 0.5, by_industry)
```

```
##        by_industry
## Hi_R.Sq Food Healthcare Media Tech
##   FALSE    7          6     4    7
##   TRUE     0          0     2    0
```

Media has the two highest R-squared values (Disney at .5875 and Viacom at .60). The rest of the stocks all have R-squared values below .5, meaninig that less than 50% of the asset return variance is explained by the 3 factor model for these stocks.

Healthcare and Media are generally aggressive industries, each with 5 out of 6 stocks having an estimated coefficient on the excess return factor that is greater than 1. Food is not aggressive, with all 7 stocks below 1. Tech is approximately equally weighted with 4 stocks labeled aggressive and 3 not. In total, 14 of the 26 stocks are aggressive.

1b) Identify those individual assets that do not follow the FF-3-factor model and their sectors.

```
alpha = sfit[[1]]$coef[1,]
for(i in 2:N){
  alpha= rbind(alpha, sfit[[i]]$coef[1,])
}
dimnames(alpha)[[1]] = syb
cat("alpha significant at 5%: \n", syb[(alpha[,4] < 0.05)])  # output the stocks that don't work for model
```

```
## alpha significant at 5%:
##
```

```
alpha[(alpha[,4]< 0.05),]
```

```
##      Estimate Std. Error t value Pr(>|t|)
```

All of the individual stocks follow the FF-3-factor model using an alpha = 5%.

1c) Test if the FF-3-factor model holds for ALL the 26 assets. Use both the Wald and the LRT.

```
# Wald Test for all alpha's are zero
ahat = alpha[,1]                          # intercepts from each LR model combined in one vector
res = resid(fit)                          # residual from each LR model combined for all models
Sig.e = 1/n*t(res)%*%res                   # Sigma_epsilon hat
Ft = FF_data[,1:3]                         # the three factors
F.bar = apply(Ft,2,mean); F.S2 = cov(Ft)   # sample mean and variance of Ft
wald = (n-N-p)/N*1/(1+t(F.bar)%*%solve(F.S2)%*%F.bar)*(t(ahat)%*%solve(Sig.e)%*%ahat)   # Wald statistic
cat("Wald Test: \n", c(statistic = wald, p.value = 1-pf(wald, N, n-N-p)))
```

```
## Wald Test:
##  0.5181021 0.9783284
```

```
# Likelihood Ratio Test for all alpha's are zero
res.0 = resid(lm(Yt~Mkt.RF+SMB+HML-1))     # residuals from the restricted model
Sig.e0 = 1/n*t(res.0)%*%res.0              # Sigma_epsilon hat under H0
lr = (n-N/2-p-1)*(log(det(Sig.e0))-log(det(Sig.e)))  # LRT statistic
cat("\n \n LRT Test: \n", c(statistic = lr, p.value =1-pchisq(lr, N)))
```

```
##
##
##  LRT Test:
##  13.56307 0.9783199
```

Our null hypothesis is that the 3 factor model is valid, meaning that the vector of intercepts is 0 (H_0: a = 0). The p-values from both tests are equivalent and high. We do not have sufficient evidence to reject the null hypothesis that the 3 factor model is valid.

1d) Test if FF-3-factor model holds for each industry by testing the assets from the same industry.

```
cat(" Testing alpha's in Food industry are zero \n")
```

```
##  Testing alpha's in Food industry are zero
```

```
ind = 1:7
Ni = length(ind)
wald = (n-Ni-p)/Ni*1/(1+t(F.bar)%*%solve(F.S2)%*%F.bar)*
(t(ahat[ind])%*%solve(Sig.e[ind,ind])%*%ahat[ind])                    #Wald statistic
lr = (n-Ni/2-p-1)*(log(det(Sig.e0[ind,ind]))-log(det(Sig.e[ind,ind])))  #LRT
cat("Wald test:\n")
```

```
## Wald test:
```

```
c(statistic = wald, p.value = 1-pf(wald, Ni, n-Ni-p))
```

```
## statistic   p.value
## 0.8838214 0.5186585
```

```
cat("LR test:\n")
```

```
## LR test:
```

```
c(statistic = lr, p.vallue = 1-pchisq(lr, Ni))
```

```
## statistic  p.vallue
## 6.1820174 0.5186644
```

```
cat("\n Testing alpha's in Media industry are zero \n")
```

```
##
##  Testing alpha's in Media industry are zero
```

```
ind = 8:13
Ni = length(ind)
wald = (n-Ni-p)/Ni*1/(1+t(F.bar)%*%solve(F.S2)%*%F.bar)*
(t(ahat[ind])%*%solve(Sig.e[ind,ind])%*%ahat[ind])                    #Wald statistic
lr = (n-Ni/2-p-1)*(log(det(Sig.e0[ind,ind]))-log(det(Sig.e[ind,ind])))  #LRT
cat("Wald test:\n")
```

```
## Wald test:
```

```
c(statistic = wald, p.value = 1-pf(wald, Ni, n-Ni-p))
```

```
## statistic   p.value
## 0.8592406 0.5245157
```

```
cat("LR test:\n")
```

```
## LR test:
```

```
c(statistic = lr, p.vallue = 1-pchisq(lr, Ni))
```

```
## statistic  p.vallue
## 5.1515968 0.5245221
```

```
cat("\n Testing alpha's in Healthcare industry are zero \n")
```

```
##
##  Testing alpha's in Healthcare industry are zero
```

```
ind = 14:19
Ni = length(ind)
wald = (n-Ni-p)/Ni*1/(1+t(F.bar)%*%solve(F.S2)%*%F.bar)*
(t(ahat[ind])%*%solve(Sig.e[ind,ind])%*%ahat[ind])                    #Wald statistic
lr = (n-Ni/2-p-1)*(log(det(Sig.e0[ind,ind]))-log(det(Sig.e[ind,ind])))  #LRT
cat("Wald test:\n")
```

```
## Wald test:
```

```
c(statistic = wald, p.value = 1-pf(wald, Ni, n-Ni-p))
```

```
##  statistic    p.value
## 0.04981325 0.99949914
```

```
cat("LR test:\n")
```

```
## LR test:
```

```
c(statistic = lr, p.vallue = 1-pchisq(lr, Ni))
```

```
## statistic  p.vallue
## 0.2995827 0.9994992
```

```
cat("\n Testing alpha's in Tech industry are zero \n")
```

```
##
##  Testing alpha's in Tech industry are zero
```

```
ind = 20:26
Ni = length(ind)
wald = (n-Ni-p)/Ni*1/(1+t(F.bar)%*%solve(F.S2)%*%F.bar)*
(t(ahat[ind])%*%solve(Sig.e[ind,ind])%*%ahat[ind])                    #Wald statistic
lr = (n-Ni/2-p-1)*(log(det(Sig.e0[ind,ind]))-log(det(Sig.e[ind,ind])))  #LRT
cat("Wald test:\n")
```

```
## Wald test:
```

```
c(statistic = wald, p.value = 1-pf(wald, Ni, n-Ni-p))
```

```
## statistic   p.value
## 0.3083866 0.9502939
```

```
cat("LR test:\n")
```

```
## LR test:
```

```
c(statistic = lr, p.vallue = 1-pchisq(lr, Ni))
```

```
## statistic  p.vallue
## 2.1625997 0.9502952
```

Both the Wald and Likelihood Ratio tests are not significant for any of the 4 industries. We cannot reject the 4 null hypotheses that each industry's weekly returns follow the 3 factor Fama-French model.

1e) give the numbers of parameters of the sample covariance and the model based. The model based estimation relies on the assumption that the covariance is diagonal. Check the assumption.

```
covparam = (N*(N+1)/2)
modparam = (p+1)*(N+p/2)
cat("The number of sample covariance parameters to estimate is:", covparam, "\n")
```

```
## The number of sample covariance parameters to estimate is: 351
```

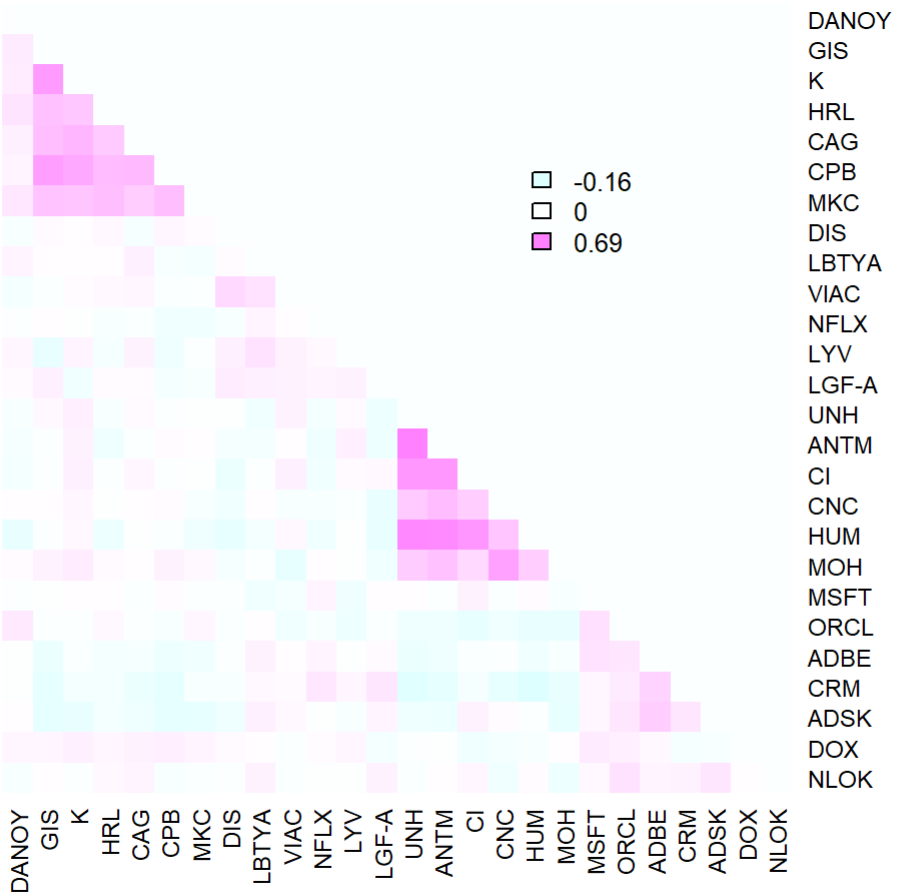
```
cat("The number of model based parameters to estimate is:", modparam, "\n")
```

```
## The number of model based parameters to estimate is: 110
```

Now check the assumption of a diagonal covariance using the heat map.

```
resid.summary(res)
```

```
##
## Significant pairs at 1% level:  75 of  325 pairs
## Significant pairs at 5% level:  113 of  325 pairs
```



The plot shows that there is a block structure in the variance-covariance matrix. There are moderately strong positive correlations in the diagonal blocks indicating that industry type should be a factor, especially for food and healthcare. The correlation between two returns from the same industry is higher than those from differing

industries.
There is also some weak positive correlation between Media and Tech companies.

Question 2: Factor Analysis

Setup

```
p = 3
Zt = apply(Yt, 2, function(u) (u-mean(u))/sd(u))                  # using standardized excess re
turns
fa = factanal(Zt, p, scores = "Bartlett", rotation = "none")      # apply Bartlett and no rotati
on
B = t(fa$loading)                                                 # extract loadings
Ft.fa = fa$scores                                                 # proxy Ft - extract our estim
ated factor realizations
R.Sq.fa = diag(t(B)%*%var(Ft.fa)%*%B)                             # R2 calc'd in a similar way t
o PCA
Ehat =   Zt - Ft.fa %*% B                                         # calculate residuals
```

2a) Examine the loading estimate. Can you find interpretation about these coefficients?

```
print(B)
```

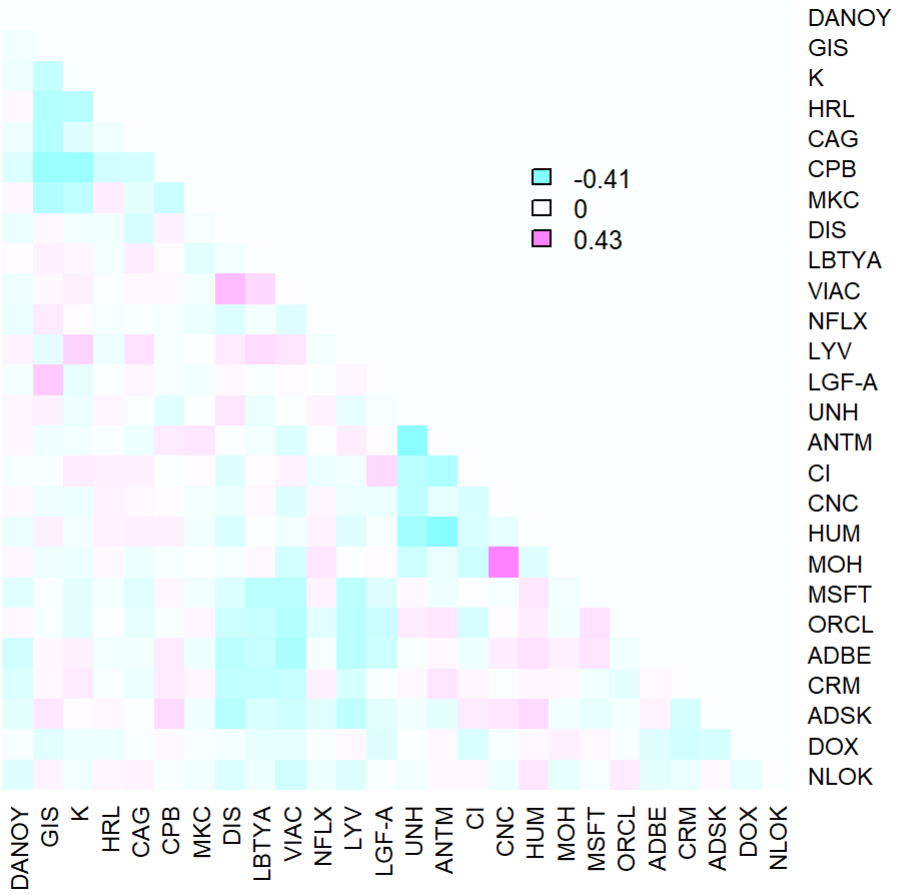
```
##
## Loadings:
##         DANOY  GIS    K      HRL    CAG    CPB    MKC    DIS    LBTYA  VIAC
## Factor1  0.503  0.485  0.551  0.361  0.478  0.399  0.498  0.667  0.535  0.670
## Factor2  0.246  0.437  0.373  0.408  0.361  0.428  0.374  0.211  0.183  0.122
## Factor3  0.101 -0.458 -0.402 -0.260 -0.236 -0.488 -0.219  0.226  0.271  0.263
##         NFLX   LYV    LGF-A  UNH    ANTM   CI     CNC    HUM    MOH    MSFT
## Factor1  0.306  0.574  0.308  0.807  0.795  0.751  0.521  0.710  0.529  0.607
## Factor2  0.170         0.176 -0.360 -0.402 -0.305 -0.169 -0.449 -0.112  0.215
## Factor3  0.187  0.269  0.246 -0.111 -0.103                -0.113         0.163
##         ORCL   ADBE   CRM    ADSK   DOX    NLOK
## Factor1  0.570  0.587  0.518  0.571  0.553  0.469
## Factor2  0.283  0.255  0.267  0.183  0.223  0.167
## Factor3  0.260  0.336  0.388  0.390  0.162  0.190
##
##                DANOY   GIS     K   HRL   CAG   CPB   MKC   DIS LBTYA  VIAC
## SS loadings    0.324 0.636 0.604 0.364 0.414 0.581 0.436 0.541 0.393 0.533
## Proportion Var 0.108 0.212 0.201 0.121 0.138 0.194 0.145 0.180 0.131 0.178
## Cumulative Var 0.108 0.320 0.521 0.643 0.781 0.974 1.120 1.300 1.431 1.609
##                 NFLX   LYV LGF-A   UNH  ANTM    CI   CNC   HUM   MOH  MSFT
## SS loadings    0.158 0.410 0.186 0.793 0.805 0.658 0.301 0.719 0.299 0.441
## Proportion Var 0.053 0.137 0.062 0.264 0.268 0.219 0.100 0.240 0.100 0.147
## Cumulative Var 1.661 1.798 1.860 2.124 2.393 2.612 2.712 2.952 3.051 3.199
##                 ORCL  ADBE   CRM  ADSK   DOX  NLOK
## SS loadings    0.473 0.523 0.490 0.511 0.382 0.284
## Proportion Var 0.158 0.174 0.163 0.170 0.127 0.095
## Cumulative Var 3.356 3.531 3.694 3.864 3.992 4.086
```

Factor 1 is positive for all stocks and ranges from 0.306 to 0.807. This is likely a market component as it represents general movement.
Factor 2 is positive for all stocks except those in the healthcare industry. This is likely an industry component.
Factor 3 is mainly negative for food and healthcare stocks but positive for media and tech. Similar to factor 2, this is likely an industry component.

2b) Check the covariance structure of the errors with the resid.summary() function. Describe your findings.

```
resid.summary(Ehat)
```

```
##
## Significant pairs at 1% level:  84 of  325 pairs
## Significant pairs at 5% level:  120 of  325 pairs
```



Unlike part 1e, there isn't strong evidence of block correlation by industry here. There is a some minor, negative block correlation in the food industry but it's very mild. This covariance structure is pretty good as there is little correlation along the diagonals (with the exception of a few pairs of stocks) as well as between industries. The color scale for positive correlation is lower here than in 1e but the negative correlation is higher.

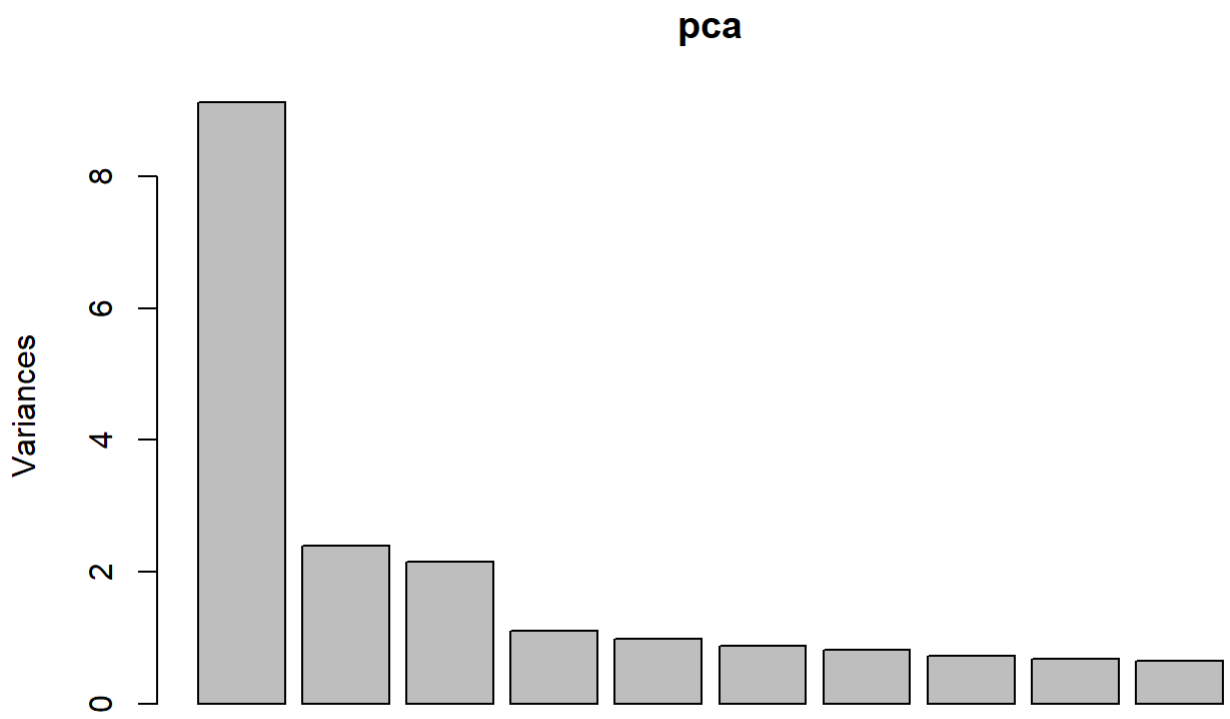
Question 3: Principal Component Analysis

Setup

```
p = 3
pca = prcomp(Zt)
```

3a) Plot the scree plot. How many principal components would you choose?

```
plot(pca)
```



It appears that 3 principal components would be a good choice.

3b) Approximate the factor model with p = 3. Compute the estimated factors and R Squared.

```
B.3 = t(pca$rotation[,1:p])                          # t(Op)
Ft.pc = pca$x[,1:p]                                  # proxy Ft  -> these are the approximate fa
ctors
R.Sq.pc = diag(t(B.3)%*%diag(pca$sd[1:p]^2)%*%B.3)
```

3c) Examine the weights of principal components - can you find interpretations form them?

```
B.3                  # calculated in part 3b as:     B.3 = t(pca$rotation[,1:p])
```

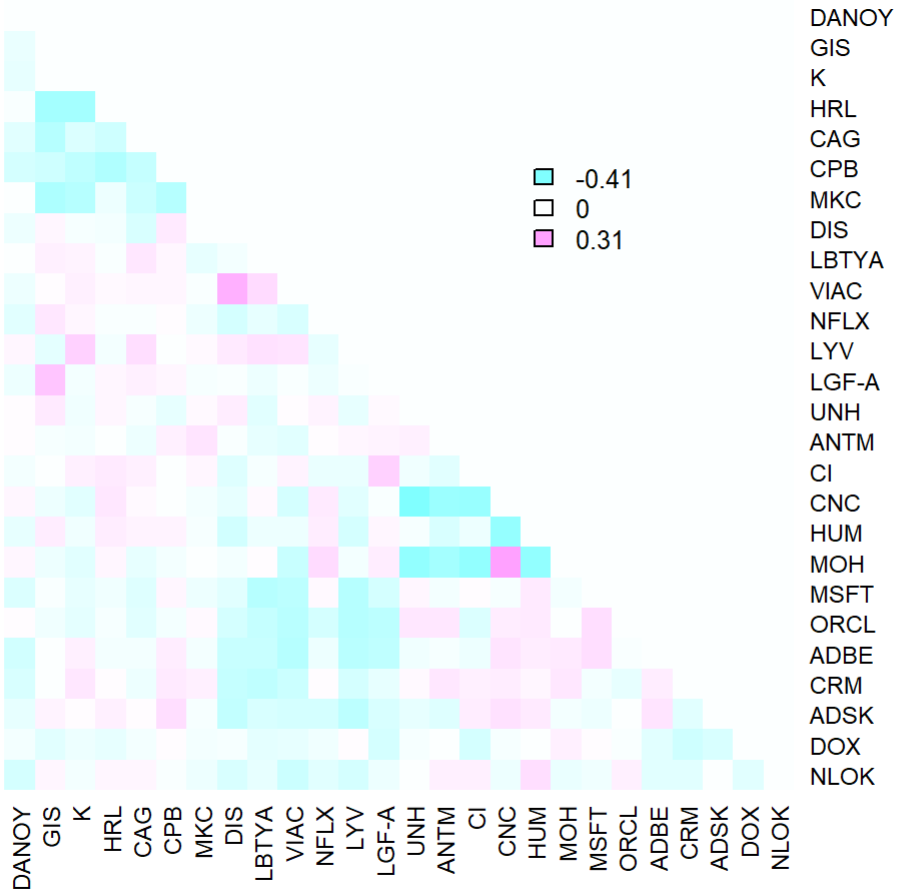
```
##           DANOY        GIS          K         HRL         CAG        CPB
## PC1 -0.19141706 -0.1743467 -0.1935079 -0.14345367 -0.17937374 -0.1457178
## PC2 -0.07226542 -0.3691323 -0.3096675 -0.34856002 -0.27962125 -0.3982038
## PC3  0.08969875 -0.1723853 -0.1867894 -0.07637511 -0.08748834 -0.2038614
##             MKC         DIS      LBTYA        VIAC         NFLX        LYV
## PC1 -0.18582198 -0.24105916 -0.2019055 -0.23695819 -0.124357306 -0.2074512
## PC2 -0.28959366  0.02534098  0.0576096  0.09058507  0.005760991  0.1083850
## PC3 -0.07563685  0.12998074  0.1718240  0.12122419  0.176666545  0.1338515
##           LGF-A        UNH       ANTM         CI        CNC        HUM
## PC1 -0.12600028 -0.2317837 -0.2263353 -0.2237193 -0.1680730 -0.1970159
## PC2  0.02523904  0.2200359  0.2482334  0.2303018  0.1749862  0.2725365
## PC3  0.23011316 -0.2855494 -0.2987938 -0.2216748 -0.2328848 -0.3213528
##            MOH        MSFT       ORCL        ADBE          CRM        ADSK
## PC1 -0.1739879 -0.22152795 -0.2166029 -0.22153886 -0.20176202 -0.21352280
## PC2  0.1221613 -0.01536763 -0.0292956  0.03217948  0.03480693  0.09520833
## PC3 -0.2343440  0.10625239  0.2007703  0.22077996  0.28754218  0.23153333
##             DOX         NLOK
## PC1 -0.20719952 -0.175951587
## PC2 -0.02340197  0.008950645
## PC3  0.11193220  0.147821917
```

The weights of Principal Component 1 are negative for all stocks, indicating that this is likely a market related component (negative of the market factor).
The weights of PC2 are negative for all of the food stocks and a few of the tech stocks. This is likely an industry component.
The weights of PC3 are all negative for food and healthcare and all positive for media and tech, again indicating that this is likely an industry component.

3d) Check the covariance structure of the errors. Describe your findings.

```
Ehat3 =   Zt - Ft.pc %*% B.3                                    # calculate residuals
resid.summary(Ehat3)
```

```
##
## Significant pairs at 1% level:  91 of  325 pairs
## Significant pairs at 5% level:  125 of  325 pairs
```

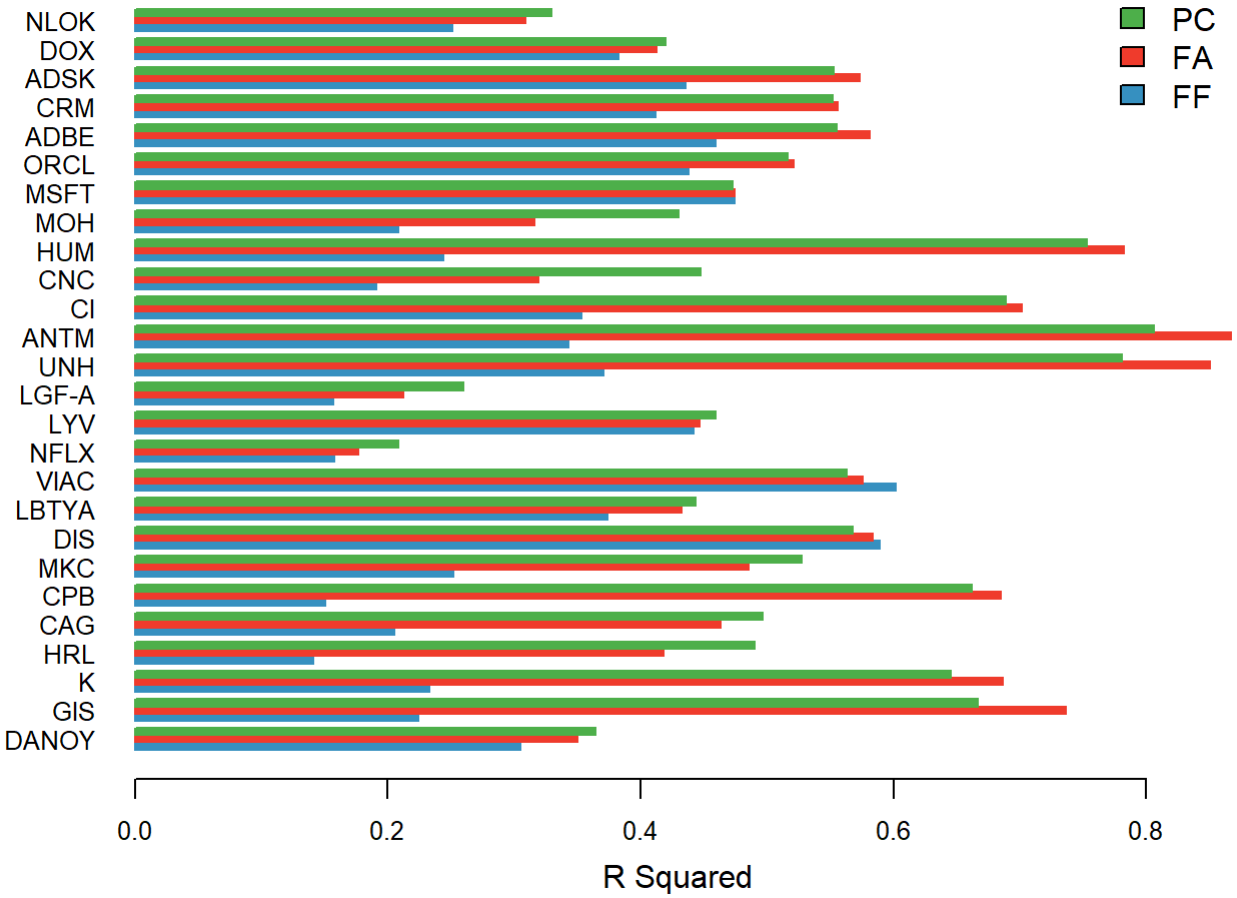


Similar to 2b, there isn't strong evidence of block correlation by industry here. There is a some minor, negative block correlation in the food industry and some positive correlation in the tech industry but both are very mild. This covariance structure is again pretty good as there is little correlation along the diagonals (with the exception of a few pairs of stocks) as well as between industries. Note that the color scale for positive correlation is also .1 lower here than in part 2b.

Question 4: Comparison of Methods

4a) Plot the R Squared values of the 3 models.

```
RSq.all = t(cbind(R.Squared, R.Sq.fa, R.Sq.pc))
RSq.plot(RSq.all)
```



4b) Compare and comment the three approaches in pricing modeling from all the analysis you have done.
For food and Media companies, the PCA price modeling approach has the highest R-squared values, indicating that this model explains more variation in excess weekly returns than the other two models.

For healthcare and tech, the factor analysis method has the higher R-squared values.

The 3 factor Fama-French model generally performs the worst, except for Disney and Viacom. Both the PCA and FA methods do best for 12 stocks each. Either of these two methods seems more appropriate than the 3 factor Fama French model. They also have better correlation plots with very few strongly correlated pairs or industry correlations (positive or negative). This makes sense as the Fama-French model has rigid factors, whereas PCA and FA can blend factor compositions as they are linear combinations.