PCA example I

Stock data

STAT 32950-24620

Spring 2023 (3/23, wk1)

Stock price data

Data: Weekly rates of return for five stocks

```
stock = read.table("T8-4.DAT")
colnames(stock) =
  c("JPMorgan","Citibank","WellsFargo","Shell", "Exxon")
attach(stock)
```

```
str(stock)
```

```
## 'data.frame':    103 obs. of  5 variables:
##  $ JPMorgan  : num  0.01303 0.00849 -0.01792 0.02156 0.(
##  $ Citibank  : num  -0.00784 0.01669 -0.00864 -0.00349 (
##  $ WellsFargo: num  -0.00319 -0.00621 0.01004 0.01744 -(
##  $ Shell     : num  -0.0448 0.012 0 -0.0286 0.0292 ...
##  $ Exxon     : num  0.00522 0.01349 -0.00614 -0.00695 0.
```



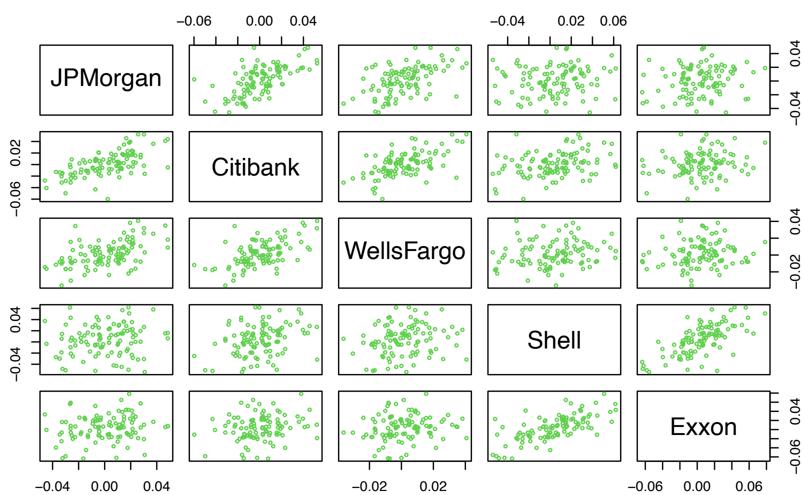
```
summary(stock[,1:3]);summary(stock[,4:5])
```

```
##     JPMorgan           Citibank           WellsFargo
##  Min.   :-0.04587   Min.   :-0.059792   Min.   :-0.0362
##  1st Qu.:-0.01356   1st Qu.:-0.013241   1st Qu.:-0.0080
##  Median : 0.00336   Median : 0.001734   Median : 0.0003
##  Mean   : 0.00106   Mean   : 0.000655   Mean   : 0.0016
##  3rd Qu.: 0.01680   3rd Qu.: 0.014029   3rd Qu.: 0.0100
##  Max.   : 0.04848   Max.   : 0.052527   Max.   : 0.0406

##      Shell              Exxon
##  Min.   :-0.05395   Min.   :-0.06360
##  1st Qu.:-0.01447   1st Qu.:-0.01254
##  Median : 0.00634   Median : 0.00522
##  Mean   : 0.00405   Mean   : 0.00404
##  3rd Qu.: 0.02224   3rd Qu.: 0.02162
##  Max.   : 0.06199   Max.   : 0.07842
```


Plot raw data

```
pairs(stock,cex=.5,col=3)
```



Sample variance-covariance matrix S

```r
cov(stock)   # same as var(stock)
```

```
##                JPMorgan   Citibank WellsFargo      Shell 
## JPMorgan   4.3327e-04 0.00027567 1.5903e-04 6.4119e-05 8
## Citibank   2.7567e-04 0.00043872 1.7997e-04 1.8145e-04 1
## WellsFargo 1.5903e-04 0.00017997 2.2397e-04 7.3413e-05 6
## Shell      6.4119e-05 0.00018145 7.3413e-05 7.2250e-04 5
## Exxon      8.8966e-05 0.00012326 6.0546e-05 5.0828e-04 7
```


To have a sense of relative magnitudes: $S \times 10^5$

```r
round(cov(stock)*10^5);
```

```
##            JPMorgan Citibank WellsFargo Shell Exxon
## JPMorgan         43       28         16     6     9
## Citibank         28       44         18    18    12
## WellsFargo       16       18         22     7     6
## Shell             6       18          7    72    51
## Exxon             9       12          6    51    77
```


Sample correlation matrix R

```r
cor(stock)
```

```
##            JPMorgan Citibank WellsFargo   Shell   Exxon
## JPMorgan    1.00000  0.63229    0.51050 0.11460 0.15446
## Citibank    0.63229  1.00000    0.57414 0.32229 0.21267
## WellsFargo  0.51050  0.57414    1.00000 0.18250 0.14621
## Shell       0.11460  0.32229    0.18250 1.00000 0.68338
## Exxon       0.15446  0.21267    0.14621 0.68338 1.00000
```


Better view:

```r
round(cor(stock),2)
```

```
##            JPMorgan Citibank WellsFargo Shell Exxon
## JPMorgan       1.00     0.63       0.51  0.11  0.15
## Citibank       0.63     1.00       0.57  0.32  0.21
## WellsFargo     0.51     0.57       1.00  0.18  0.15
## Shell          0.11     0.32       0.18  1.00  0.68
## Exxon          0.15     0.21       0.15  0.68  1.00
```

Eigenvalues and eigenvectors of S

```
eigen(cov(stock))
```

```
## eigen() decomposition
## $values
## [1] 0.00136768 0.00070116 0.00025380 0.00014260 0.000118
##
## $vectors
##          [,1]     [,2]      [,3]     [,4]      [,5]
## [1,] 0.22282  0.62523  0.326112  0.66276  0.117660
## [2,] 0.30729  0.57039 -0.249590 -0.41409 -0.588608
## [3,] 0.15481  0.34450 -0.037639 -0.49705  0.780304
## [4,] 0.63897 -0.24795 -0.642497  0.30887  0.148455
## [5,] 0.65090 -0.32185  0.645861 -0.21638 -0.093718
```


Principal Component Analysis on original data (using S)

```
summary(princomp(stock),loading=T)
```

```
## Importance of components:
##                           Comp.1   Comp.2   Comp.3   Comp
## Standard deviation     0.036802 0.026351 0.015854 0.011
## Proportion of Variance 0.529261 0.271333 0.098216 0.055
## Cumulative Proportion  0.529261 0.800594 0.898809 0.953
##
## Loadings:
##            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## JPMorgan    0.223  0.625  0.326  0.663  0.118
## Citibank    0.307  0.570 -0.250 -0.414 -0.589
## WellsFargo  0.155  0.345        -0.497  0.780
## Shell       0.639 -0.248 -0.642  0.309  0.148
## Exxon       0.651 -0.322  0.646 -0.216
```


Comparison: PCA and Eigen-Analysis

Eigenvalues of covariance matrix S:

```
round(eigen(cov(stock))$values,4)
```

```
## [1] 0.0014 0.0007 0.0003 0.0001 0.0001
```

Standard deviations of Principal Component variables:

```
round(princomp(stock)$sdev,4)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## 0.0368 0.0264 0.0159 0.0119 0.0109
```

Variance of PC variables:

```
round(princomp(stock)$sdev^2,4)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## 0.0014 0.0007 0.0003 0.0001 0.0001
```


Eigenvalue of S and PC variance

$$\lambda_i = Var(Y_i), \qquad i = 1, \cdots, p$$

Proportions of variation:

$$\frac{\lambda_i}{\sum_j \lambda_j} = \frac{Var(Y_i)}{\sum_j Var(Y_j)}, \qquad i = 1, \cdots, p$$

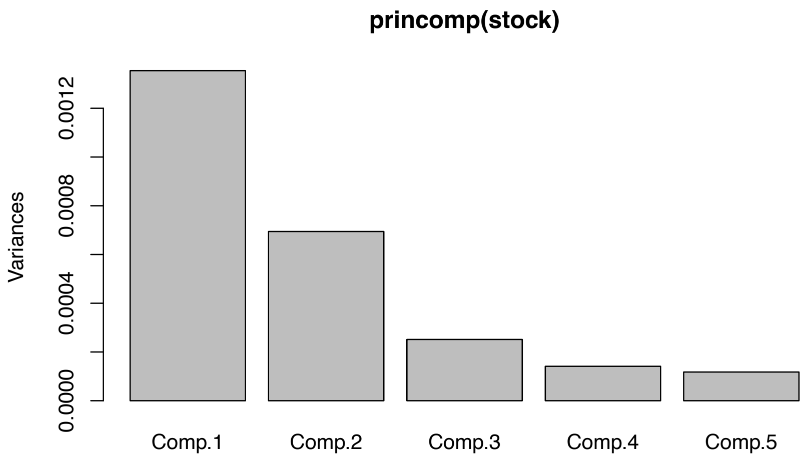
```
round(princomp(stock)$sdev^2/sum(princomp(stock)$sdev^2),3)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
##  0.529  0.271  0.098  0.055  0.046
```

Scree plot

PCA on raw data (no scaling, using the covariance matrix)

```
par(mfrow=c(1,1)); screeplot(princomp(stock))
```



Relation between PCs and original variables

PC of the raw data (using covariance matrix)

The first principal component: $Y_1 = a_1'X$

$\lambda_1 = 0.00136 = V(Y_1)$

$Y_1 = .22(JPM) + .31(Citibk) + .15(WellsF) + .64(Shell) + .65(Exxon)$

Interpretations?

Y_1 could be viewed as " **Market component** "

Relation between PCs and original variables (cont.)

The Second principal component: $Y_{=}a_2'X$

$\lambda_2 = 0.000701 = V(Y_2)$

$Y_2 = .63(JPM) + .57(Citibk) + .34(WellsF) - .25(Shell) - .32(Exxon)$

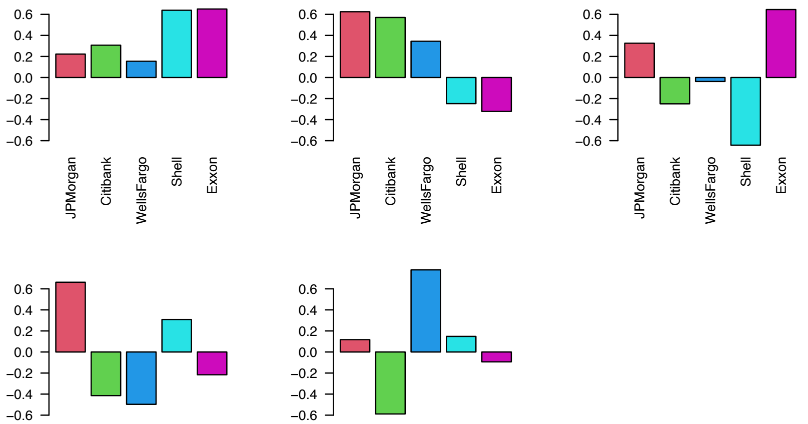
Interpretations:

Y_2 could be viewed as " **Industry component** "

PC loadings by original variables -Code

```
par(mfrow=c(2,3));
barplot(princomp(stock)$loadings[,1],
        las=2,ylim=c(-.6,.6),col=2:6)
barplot(princomp(stock)$loadings[,2],
        las=2,ylim=c(-.6,.6),col=2:6)
barplot(princomp(stock)$loadings[,3],
        las=2,ylim=c(-.6,.6),col=2:6)
barplot(princomp(stock)$loadings[,4],
        las=2,ylim=c(-.6,.6),axisnames = F,col=2:6)
barplot(princomp(stock)$loadings[,5],
        las=2,ylim=c(-.6,.6),axisnames = F,col=2:6)
```

PC loadings by original variables - Plots



PC variance proportions

Other components:

$$\hat{\lambda}_3 = 0.000253 = \hat{V}(Y_3) \qquad \approx 9.8\% \text{ of } \sum_{i=1}^{5} V(Y_i)$$

$$\hat{\lambda}_4 = 0.000143 = \hat{V}(Y_4) \qquad \approx \text{ of } 5.5\% \sum_{i=1}^{5} V(Y_i)$$

$$\hat{\lambda}_5 = 0.000119 = \hat{V}(Y_5) \qquad \approx 4.6\% \text{ of } \sum_{i=1}^{5} V(Y_i)$$

Compared with

$$\hat{V}(Y_2) \approx 27.1\% \text{ of } \sum_{i=1}^{5} V(Y_i), \qquad \hat{V}(Y_1) \approx 52.9\% \text{ of } \sum_{i=1}^{5} V(Y_i)$$

Eigenvalues and eigenvectors of correlation matrix

```
eigen(cor(stock))
```

```
## eigen() decomposition
## $values
## [1] 2.43727 1.40701 0.50051 0.40003 0.25517
##
## $vectors
##          [,1]     [,2]      [,3]     [,4]      [,5]
## [1,] -0.46908  0.36801 -0.604315  0.36302  0.384122
## [2,] -0.53241  0.23646 -0.136106 -0.62921 -0.496188
## [3,] -0.46516  0.31518  0.771828  0.28897  0.071169
## [4,] -0.38735 -0.58504  0.093362 -0.38125  0.594664
## [5,] -0.36068 -0.60585 -0.108826  0.49341 -0.497552
```


PCA using scaled data of variable variance = 1

That is, **PCA using correlation matrix**.

```
summary(princomp(stock,cor=T),loading=T)
```

```
## Importance of components:
##                           Comp.1  Comp.2  Comp.3   Comp.4
## Standard deviation     1.56118 1.18618 0.70747 0.632481
## Proportion of Variance 0.48745 0.28140 0.10010 0.080006
## Cumulative Proportion  0.48745 0.76886 0.86896 0.948966
##
## Loadings:
##            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## JPMorgan    0.469  0.368  0.604  0.363  0.384
## Citibank    0.532  0.236  0.136 -0.629 -0.496
## WellsFargo  0.465  0.315 -0.772  0.289
## Shell       0.387 -0.585        -0.381  0.595
## Exxon       0.361 -0.606  0.109  0.493 -0.498
```

Correlation of PC and original variables

For PCA on scaled data (using correlation matrix):

The correlation between

the ith PC variable Y_i and the (scaled) kth variable X_k is

$$\rho_{Y_i,X_k} = a_{ik}\sqrt{\lambda_i}$$

For PCA on raw data (using covariance matrix):

$$\rho_{Y_i,X_k} = a_{ik}\sqrt{\lambda_i/\sigma_{kk}}$$

(both ignoring the presence of other X_j variables)

PCA vs eigen-alanysis on correlation matrix R

Eigenvalues of correlation matrix R:

```
round(eigen(cor(stock))$values,4)
```

```
## [1] 2.4373 1.4070 0.5005 0.4000 0.2552
```

Standard deviations of PC variables (scaled data):

```
round(princomp(stock,cor=T)$sdev,4)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## 1.5612 1.1862 0.7075 0.6325 0.5051
```

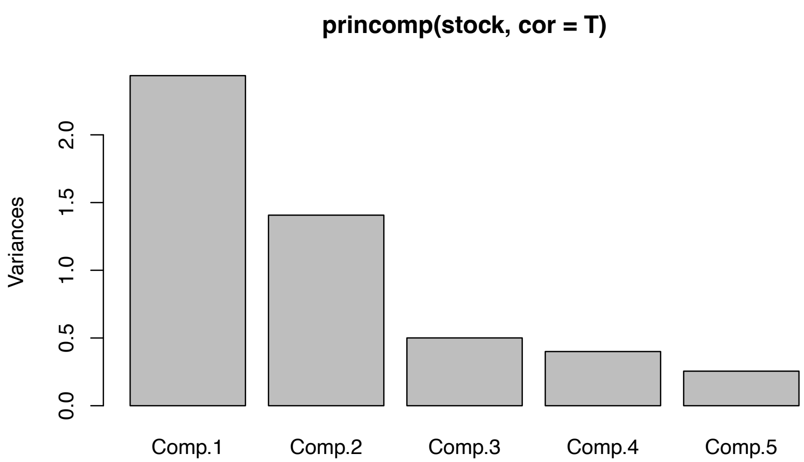
Variance of PC variables (scaled data):

```
round(princomp(stock,cor=T)$sdev^2,4)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## 2.4373 1.4070 0.5005 0.4000 0.2552
```

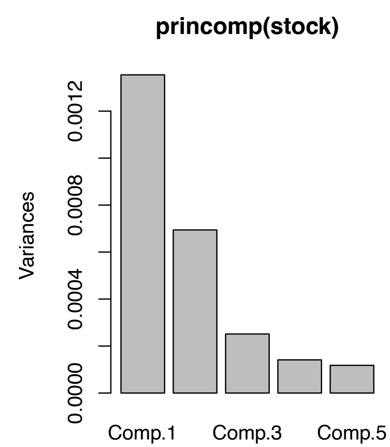

Scree plot (scaled, using correlation matrix)

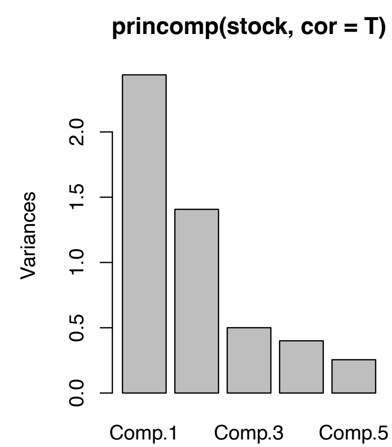
```
par(mfrow=c(1,1))
screeplot(princomp(stock,cor=T))
```



Scree plots (scaled vs not scaled)

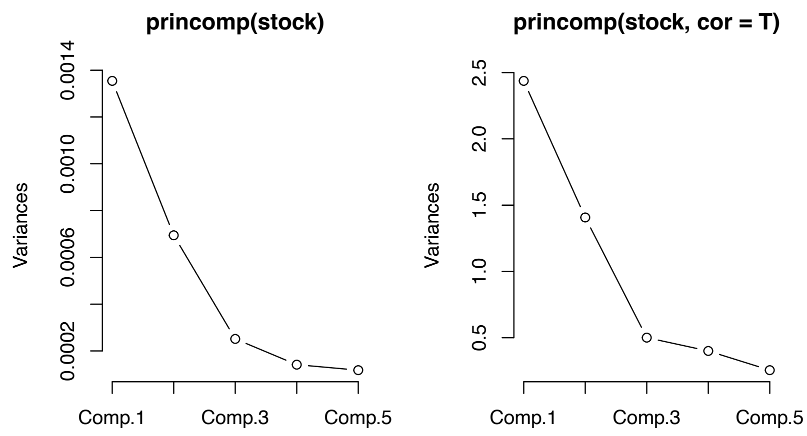
```
par(mfrow=c(1,2))
screeplot(princomp(stock));screeplot(princomp(stock,cor=T))
```





To scale or not to scale?

```
par(mfrow=c(1,2))
screeplot(princomp(stock),type="l")
screeplot(princomp(stock, cor=T),type="l")
```



Proportion of total variation explained

- Of sample covariance matrix S

```
cumsum((princomp(stock)$sdev)^2)/sum((princomp(stock)$sdev)
```

```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5
## 0.52926 0.80059 0.89881 0.95399 1.00000
```

- Of sample correlation matrix R

```
cumsum((princomp(stock,cor=T)$sdev)^2)/5
```

```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5
## 0.48745 0.76886 0.86896 0.94897 1.00000
```


Sum of eigenvalues

- Of covariance matrix S

```
sum(eigen(cov(stock))$values)
```

```
## [1] 0.0025841
```

- Of correlation matrix R

```
sum(eigen(cor(stock))$values)
```

```
## [1] 5
```

- Comparison: What is the dimension of variables?

Scale sizes

Scale = Using sample correlation matrix R

```
princomp(stock,cor=T)$scale
```

```
##    JPMorgan    Citibank WellsFargo       Shell       Exxon
##    0.020714    0.020844   0.014893    0.026748    0.027536
```

- What should the values of the "scale" be?
- Should scaling make a significant difference for this dataset?

No re-scale = Using original sample covariance matrix S

```
princomp(stock)$scale
```

```
##    JPMorgan    Citibank WellsFargo       Shell       Exxon
##           1           1          1           1           1
```

Using correlation matrix

Relation between principal components and original variables

$$\hat{\lambda}_1 = 2.437$$

$$Y_1 = .469(JPMorgan) + .532(Citibank) + .465(WellsFargo)$$
$$+ .387(Shell) + .361(Exxon)$$

Y_1: " **Market component** " (or similar interpretations)

(Using correlation matrix)

$$\hat{\lambda}_2 = 1.407$$

$$Y_2 = .368(JPMorgan) + .236(Citibank) + .315(WellsFargo)$$
$$-.585(Shell) - .606(Exxon)$$

Y_2 : " **Industry component** "

$$\hat{\lambda}_3 = 0.501 \quad > \quad \hat{\lambda}_4 = 0.400 \quad > \quad \hat{\lambda}_5 = 0.255$$

Observed values vs PC values

Data:

n measurements of the original random vector $(X_1, \cdots, X_p)$:

$$\begin{matrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{matrix}$$

PC scores

n “measurements” of principal components $(Y_1, \cdots, Y_p)$:

$$\begin{matrix} y_{11} & \cdots & y_{1p} \\ \vdots & \cdots & \vdots \\ y_{n1} & \cdots & y_{np} \end{matrix}$$

Weekly rate of returns of the original five stocks

```
head(stock,3)
```

```
##      JPMorgan   Citibank WellsFargo     Shell      Exxon
## 1  0.0130338 -0.0078431 -0.0031889 -0.044769  0.0052151
## 2  0.0084862  0.0166886 -0.0062100  0.011956  0.0134890
## 3 -0.0179153 -0.0086393  0.0100360  0.000000 -0.0061428
```

```
tail(stock)
```

```
##        JPMorgan   Citibank WellsFargo      Shell      Ex:
## 98   0.0217449  0.0229645  0.0291983  0.0084395  0.0319:
## 99   0.0033740 -0.0153061 -0.0238245 -0.0016738 -0.0172:
## 100  0.0033626  0.0029016 -0.0030507 -0.0012193 -0.0097(
## 101  0.0170147  0.0095061  0.0181994 -0.0161758 -0.0075(
## 102  0.0103929 -0.0026612  0.0044290 -0.0024818 -0.0164!
## 103 -0.0127948 -0.0143678 -0.0187402 -0.0049759 -0.0163'
```

Weekly data of five "PC stocks' ' (using correlation matrix)

```
head(princomp(stock,cor=T)$scores,3)
```

```
##         Comp.1    Comp.2   Comp.3    Comp.4     Comp.5
## [1,] -0.78790  1.056230  0.71834  1.089819 -0.7052807
## [2,]  0.57118 -0.232923  0.73713 -0.449297 -0.2764345
## [3,] -0.59651  0.047938 -1.07632 -0.013571  0.0034684
```

```
tail(princomp(stock,cor=T)$scores)
```

```
##           Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
##   [98,]  2.32830  0.494367 -0.584970  0.661200 -0.422135
##   [99,] -1.51171 -0.085575  1.218117 -0.270965  0.558224
## [100,] -0.29287  0.384882  0.288229 -0.289340  0.097955
## [101,]  0.66015  1.432133 -0.310988  0.414374 -0.075710
## [102,] -0.14883  0.781089  0.047103  0.043971  0.490387
## [103,] -1.73173 -0.201152  0.503933 -0.421642  0.171508
```


$Y_k = a_k' X$ relates PC scores and data

```
stock[33,]
```

```
##    JPMorgan Citibank WellsFargo     Shell   Exxon
## 33 0.027618 0.016832   0.010498 0.0004153 0.00433
```

```
princomp(stock,cor=T)$scores[33,]
```

```
##   Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
## 1.242857 0.916126 0.434434 0.206232 0.063711
```

```
princomp(stock)$scores[33,]
```

```
##      Comp.1      Comp.2      Comp.3      Comp.4      Cor
##  1.0129e-02  2.9694e-02  6.8116e-03  5.3063e-03 -4.1301e
```


PC score correlation property

```
dim(stock); dim(princomp(stock,cor=T)$scores) # 103 5
```

```
## [1] 103   5
```

```
## [1] 103   5
```

```
round(cor(princomp(stock,cor=T)$scores),3)
```

```
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Comp.1      1      0      0      0      0
## Comp.2      0      1      0      0      0
## Comp.3      0      0      1      0      0
## Comp.4      0      0      0      1      0
## Comp.5      0      0      0      0      1
```


Un-correlatedness of PCs is true regardless of scaling

```
round(cor(princomp(stock)$scores),3)
```

```
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Comp.1      1      0      0      0      0
## Comp.2      0      1      0      0      0
## Comp.3      0      0      1      0      0
## Comp.4      0      0      0      1      0
## Comp.5      0      0      0      0      1
```

PC score covariance property

```
round(cov(princomp(stock,cor=T)$scores),2)
```

```
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Comp.1   2.46   0.00   0.00    0.0   0.00
## Comp.2   0.00   1.42   0.00    0.0   0.00
## Comp.3   0.00   0.00   0.51    0.0   0.00
## Comp.4   0.00   0.00   0.00    0.4   0.00
## Comp.5   0.00   0.00   0.00    0.0   0.26
```

Note the diagonal patterns.

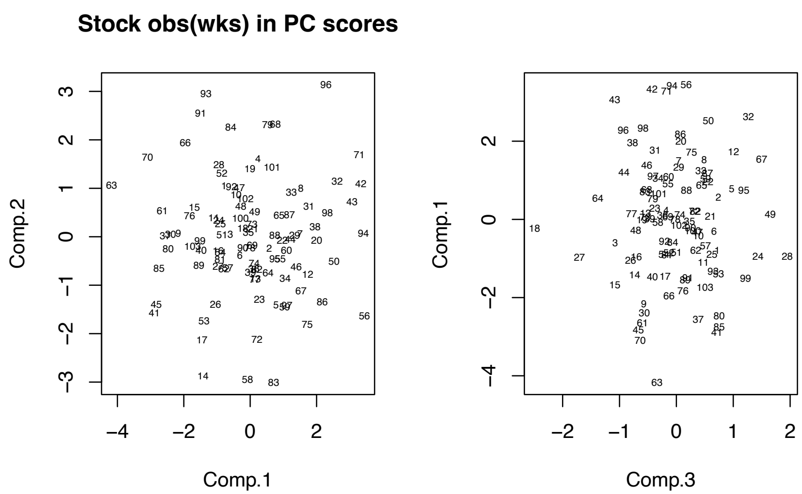
Observations using PC scores as coordinates - Code

PC scores can be uses as coordinates of data observation points

```
par(mfrow=c(1,2));
plot(princomp(stock,cor=T)$score[,1:2],
     type="n",main="Stock obs(wks) in PC scores")
text(princomp(stock,cor=T)$score[,1:2],
     labels=row.names(stock),cex=.5)
plot(princomp(stock,cor=T)$score[,c(3,1)],type="n")
text(princomp(stock,cor=T)$score[,c(3,1)],
     labels=row.names(stock),cex=.5)
```


Observations using PC scores as coordinates - Plots

The plots may reveal interesting data patterns sometimes (not here).



Variables using PC loading as coordinates - Code

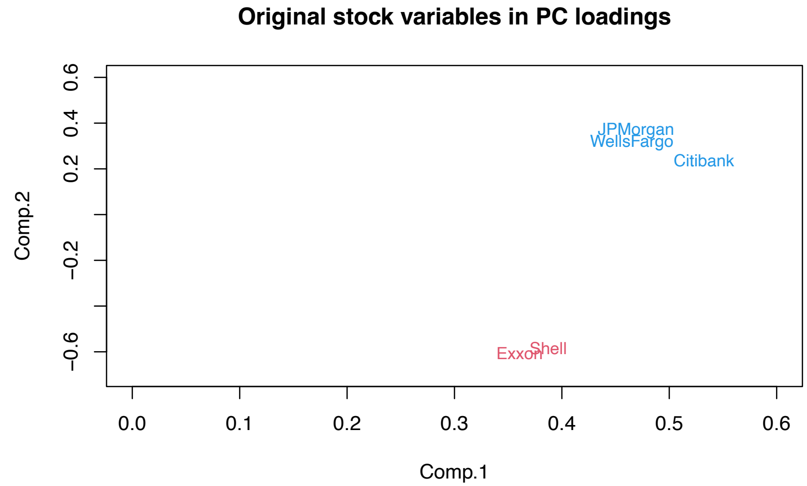
PC loadings (a_{ij}) can be used as coordinates of original variables.

```
par(mfrow=c(1,1))
plot(princomp(stock,cor=T)$loading[,1:2],
     xlim=c(0,.6),ylim=c(-.7,.6),type="n")
title(main="Original stock variables in PC loadings")
text(princomp(stock,cor=T)$loading[,1:2],
     labels=(colnames(stock)),cex=.8,col=c(4,4,4,2,2))
```

```
par(mfrow=c(1,2))
barplot(princomp(stock,cor=T)$loadings[,1],
        las=2,ylim=c(-.6,.6),col=2:6,cex.names =.7)
barplot(princomp(stock,cor=T)$loadings[,2],
        las=2,ylim=c(-.6,.6),col=2:6,cex.names =.7)
```

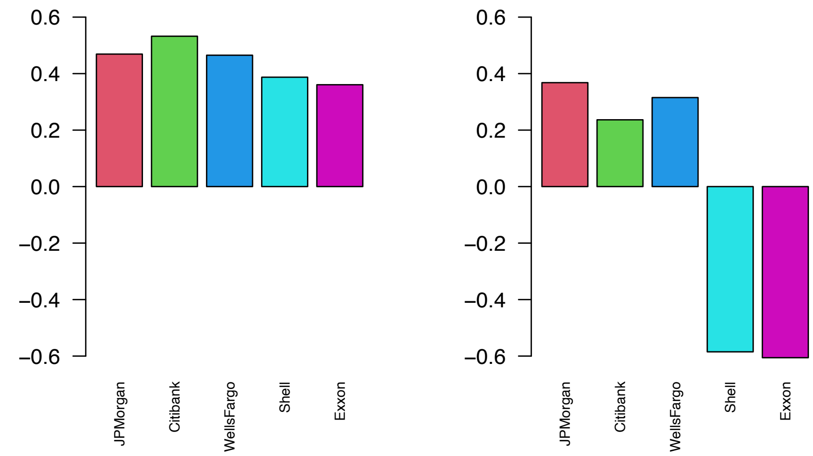
Variables using PC loading as coordinates - Plot

The plot may reveal interesting variable patterns.



Variable loading a_{ij}'s of top PCs

$$Y_1 = a_1'X, \qquad Y_2 = a_2'X$$



Norm-one property of PC loadings

In the derivation of PC variables,

$$Y_k = a_k'X, \quad \|a_k\| = 1, \quad a_i'a_k = 0, \quad i \neq k, \quad i, k = 1, \cdots, p.$$

- The p-vector a_i's are restricted to be of length 1.
- The loading vectors of different PCs are mutually orthogonal.
- The loading coefficients should form an orthogonal matrix.

Loading coefficients are norm-1: $Y_k = a_k'X, \ \|a_k\| = 1$

```
round(princomp(stock)$loading[,1:5],3)
```

```
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## JPMorgan   0.223  0.625  0.326  0.663  0.118
## Citibank   0.307  0.570 -0.250 -0.414 -0.589
## WellsFargo 0.155  0.345 -0.038 -0.497  0.780
## Shell      0.639 -0.248 -0.642  0.309  0.148
## Exxon      0.651 -0.322  0.646 -0.216 -0.094
```

```
a1 = princomp(stock,cor=T)$loading[,1]; round(a1,3)
```

```
##   JPMorgan   Citibank WellsFargo      Shell      Exxon
##      0.469      0.532      0.465      0.387      0.361
```

```
sum(a1^2)
```

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

Orthogonality of PC loadings $Y_k = a_k' X, \; a_i' a_k = 0, i \neq k$

```
a2 = princomp(stock,cor=T)$loading[,2]; round(a2,3)
```

```
##    JPMorgan    Citibank WellsFargo       Shell       Exxon
##       0.368       0.236       0.315      -0.585      -0.606
```

```
a1%*%a2;
```

```
##              [,1]
## [1,] 1.9429e-16
```

```
round(a1%*%a2,3)
```

```
##      [,1]
## [1,]    0
```


Loading coefficients form an orthogonal matrix

```
L=as.matrix(princomp(stock,cor=T)$loading,5,5);
round(L%*%t(L),3)
```

```
##            JPMorgan Citibank WellsFargo Shell Exxon
## JPMorgan          1        0          0     0     0
## Citibank          0        1          0     0     0
## WellsFargo        0        0          1     0     0
## Shell             0        0          0     1     0
## Exxon             0        0          0     0     1
```

$$L^{-1} = L^T, \qquad L^T L = L L^T = I_p$$

PC loadings by original variables (scaled data) - Code

```
par(mfrow=c(2,3))
barplot(princomp(stock,cor=T)$loadings[,1],
        las=2,ylim=c(-.6,.6),col=2:6)
barplot(princomp(stock,cor=T)$loadings[,2],
        las=2,ylim=c(-.6,.6),col=2:6)
barplot(princomp(stock,cor=T)$loadings[,3],
        las=2,ylim=c(-.6,.6),col=2:6)
barplot(princomp(stock,cor=T)$loadings[,4],
        las=2,ylim=c(-.6,.6),axisnames = F,col=2:6)
barplot(princomp(stock,cor=T)$loadings[,5],
        las=2,ylim=c(-.6,.6),axisnames = F,col=2:6)
```


PC loadings by original variables (scaled data) - Plots