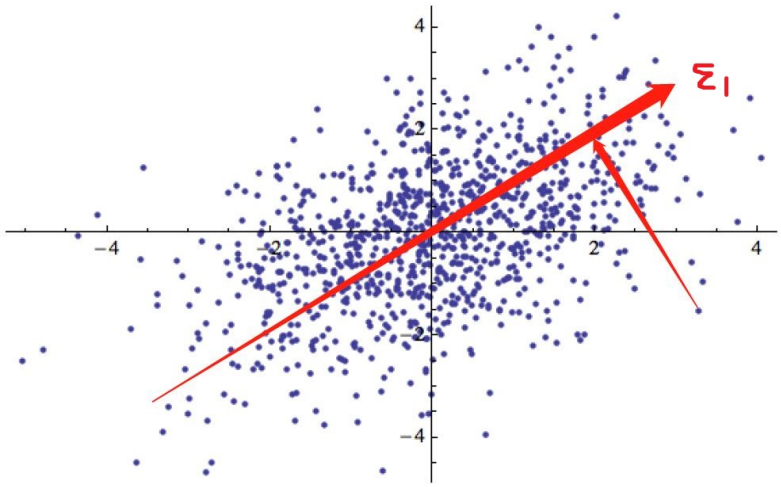
GR5241 HW2

JIE LI

February 12, 2019

Problem 1 PCA

1. Please graph approximately into the following picture:
   - The first principle component (the 1-dimensional subspace onto which PCA with one component would project). Mark this ξ_1.
   - Pick an arbitrary data point (reasonably far away from the principal component) and graph the direction in which it will be projected.



Caption for the picture.

2

a)

There are 10304 principle components in total

b)

$$\hat{X} = UDV^T$$

U is a 100x10304 left sigular vector.

D is a 10304x10304 diagonal matrix contains eigenvalues for each principle component such that $d_1 \geq d_2 \geq \ldots \geq d_{10304}$

V^T is a 10304x10304 right sigular orthogonal matrix, each row stands for each principle component spaces, and column stands for feature spaces.

Therefore, the first 48 principle componets to represent:

$$\hat{X}_i = \bar{x_i} + x_{i,1}v_{i,1} + x_{i,2}v_{i,2} + \ldots + x_{i,48}v_{i,48}$$

Problem 2

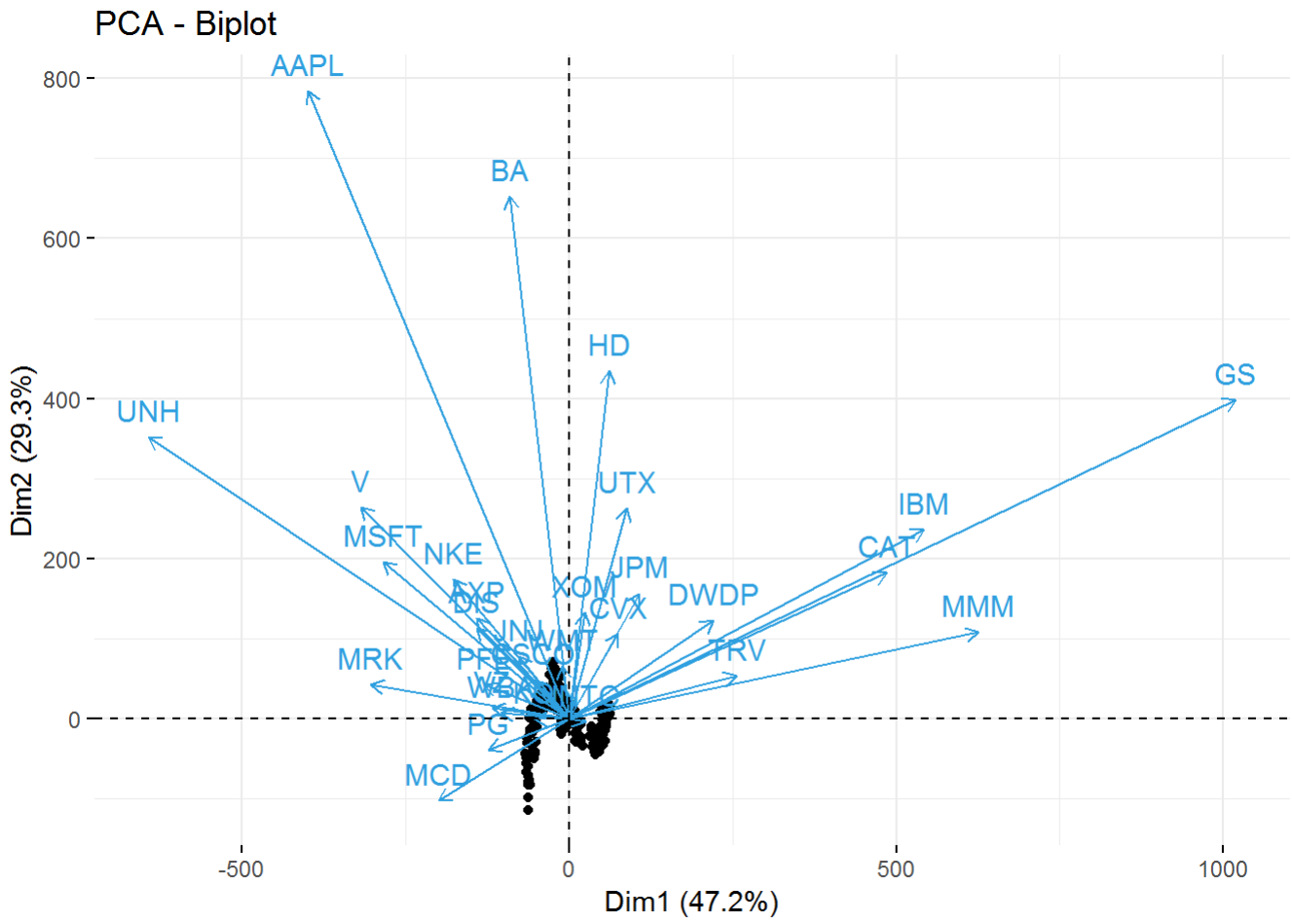
1

```
library(quantmod)
library(factoextra)
library(ggplot2)

company = c("MMM","AXP","AAPL","BA","CAT", "CVX", "CSCO","KO","DWDP","XOM","GS","HD","IBM","INTC","JNJ","JPM","MCD","MRK","MSFT","NKE","PFE","PG","TRV","UNH","UTX","VZ","V","WMT","WBA","DIS")

data = c()
for(i in company)
{
  data = cbind(data, getSymbols(i, auto.assign = F, from = "2018-01-01", to = "2019-01-01")[,4])
}
colnames(data) = gsub(".Close", "", colnames(data))
```

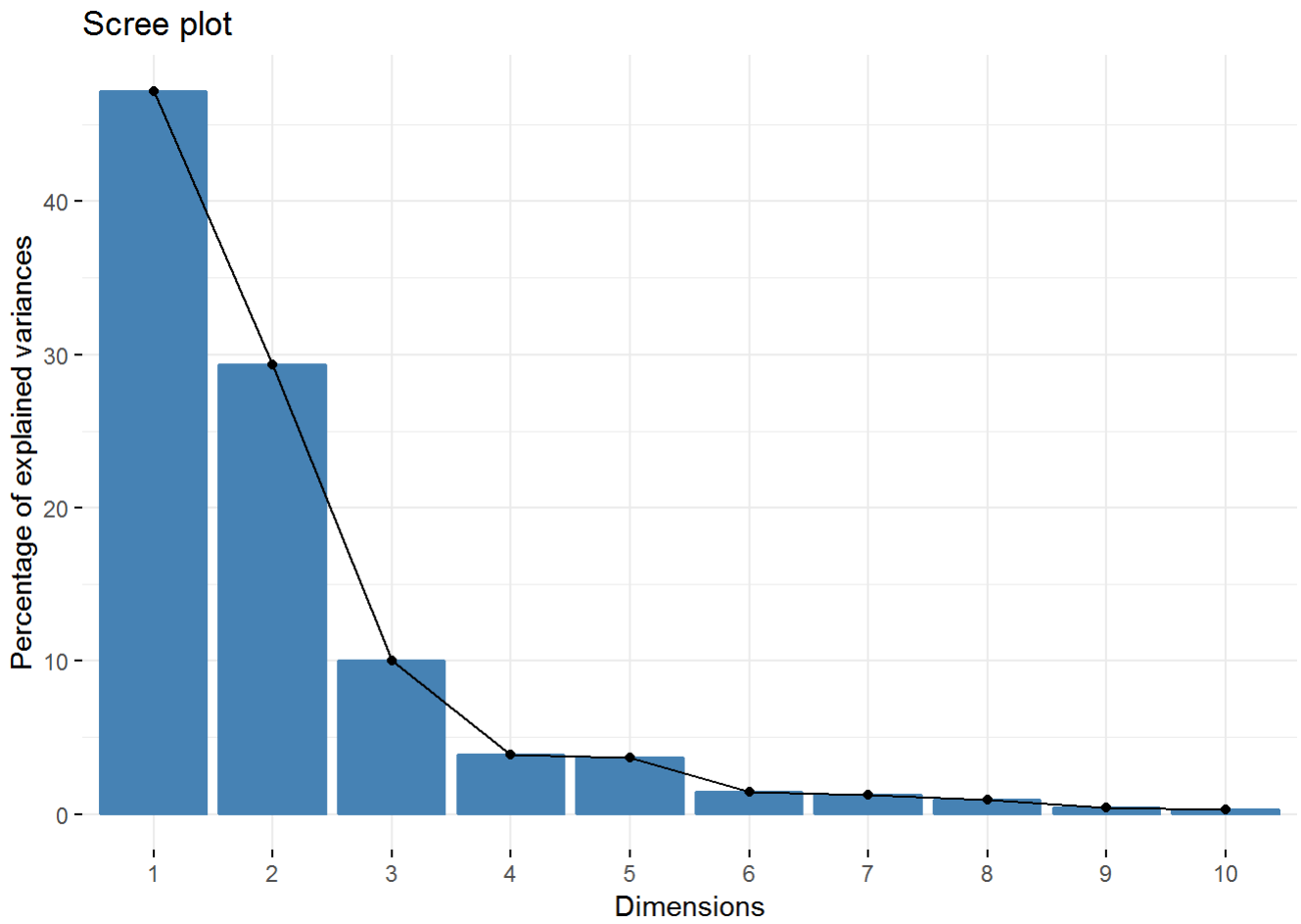
2

```
pc.pipline = function(data, scale, summary)
{
  pc = princomp(data, cor= scale)
  a = fviz_pca_biplot(pc, geom = "point", repel = F, col.var = "#2E9FDF")
  b = fviz_eig(pc)
  if(summary == T)
    c = summary(pc)
  else
    c = NULL
  return(list(a, b, c))
}

pc.pipline(data, F, T)
```

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


PCA - Biplot
AAPL
BA
HD
GS
UNH
V
UTX
IBM
MSFT
NKE
CAT
JPM
XOM
DWDP
CVX
MMM
DIS
MRK
TRV
PG
MCD
800
600
400
200
0
-500
0
500
1000
Dim2 (29.3%)
Dim1 (47.2%)

```
##
## [[2]]
```


Scree plot
Percentage of explained variances
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Dimensions

```
##
## [[3]]
## Importance of components:
##                            Comp.1     Comp.2     Comp.3      Comp.4
## Standard deviation     38.9464635 30.7083169 17.9142695 11.18089027
## Proportion of Variance  0.4720066  0.2934432  0.0998643  0.03890136
## Cumulative Proportion   0.4720066  0.7654498  0.8653141  0.90421549
##                            Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     10.89296635 6.85795468 6.41178665 5.54224433
## Proportion of Variance  0.03692363 0.01463529 0.01279293 0.00955836
## Cumulative Proportion   0.94113912 0.95577441 0.96856734 0.97812570
##                            Comp.9     Comp.10     Comp.11     Comp.12
## Standard deviation     3.715171986 3.158836458 3.028046131 2.609657054
## Proportion of Variance 0.004295066 0.003105034 0.002853232 0.002119234
## Cumulative Proportion  0.982420767 0.985525801 0.988379032 0.990498266
##                           Comp.13     Comp.14     Comp.15     Comp.16
## Standard deviation     2.271094013 2.171198293 1.942200991 1.817296039
## Proportion of Variance 0.001605027 0.001466935 0.001173817 0.001027693
## Cumulative Proportion  0.992103293 0.993570229 0.994744046 0.995771739
##                           Comp.17      Comp.18      Comp.19      Comp.20
## Standard deviation     1.6735621994 1.4889989187 1.2942827797 1.1944458737
## Proportion of Variance 0.0008715569 0.0006899233 0.0005212792 0.0004439612
## Cumulative Proportion  0.9966432955 0.9973332188 0.9978544980 0.9982984591
##                           Comp.21      Comp.22     Comp.23      Comp.24
## Standard deviation     1.0519123494 0.9707014056 0.922032483 0.8474031657
## Proportion of Variance 0.0003443271 0.0002932131 0.000264548 0.0002234561
## Cumulative Proportion  0.9986427862 0.9989359993 0.999200547 0.9994240034
##                           Comp.25      Comp.26      Comp.27      Comp.28
## Standard deviation     0.7620331386 0.6860093069 5.298655e-01 5.189041e-01
## Proportion of Variance 0.0001807007 0.0001464442 8.736619e-05 8.378885e-05
## Cumulative Proportion  0.9996047040 0.9997511482 9.998385e-01 9.999223e-01
##                           Comp.29      Comp.30
## Standard deviation     3.802759e-01 3.241517e-01
## Proportion of Variance 4.499971e-05 3.269705e-05
## Cumulative Proportion  9.999673e-01 1.000000e+00
```

There are some structures in the biplot

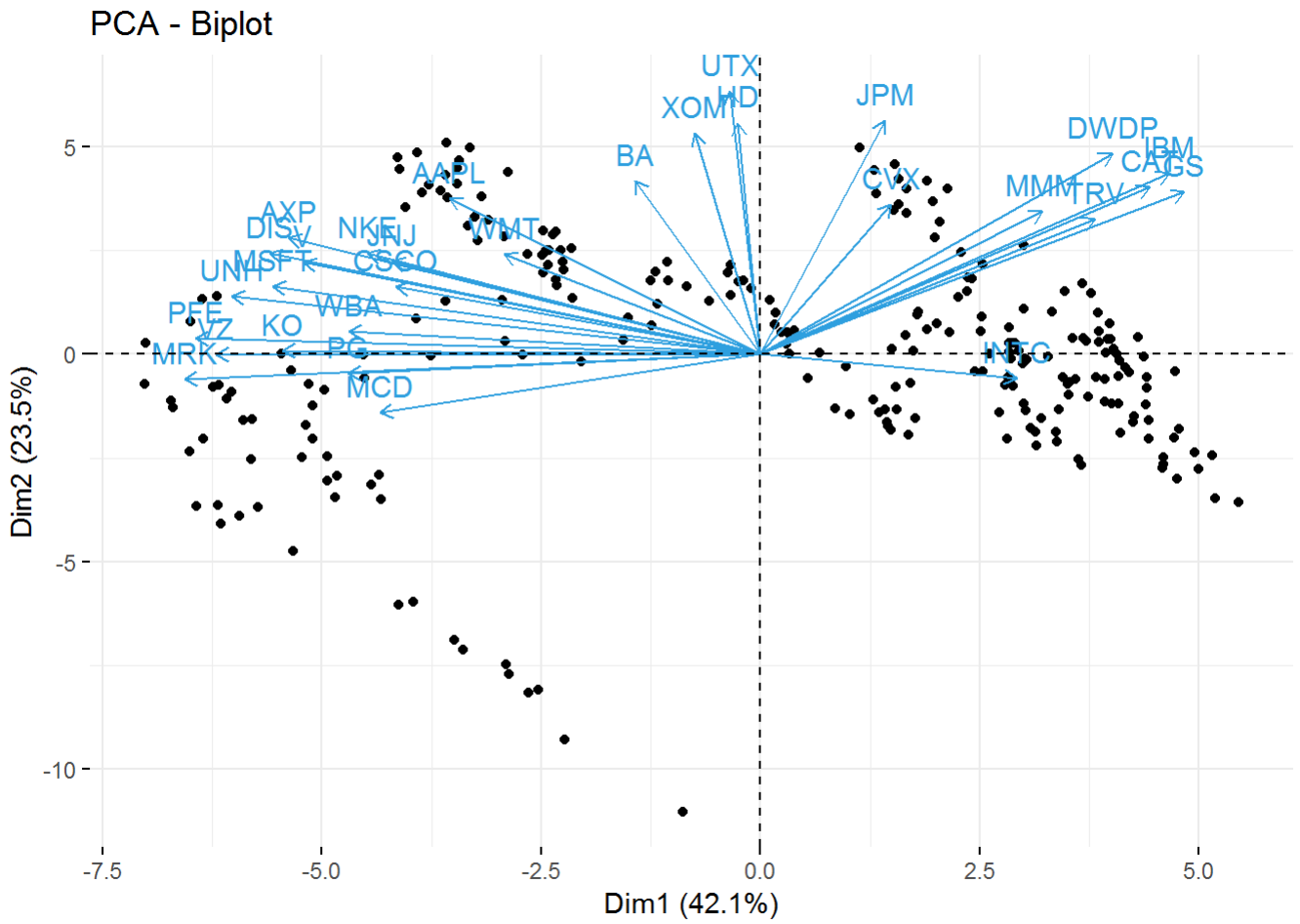
- left direction is most likely related to food, healthcare and pharamcy
- up direction is about chemical, energy and retail industry
- right direction is related to information technologies

Based on summary table, when choose 6 principle components, the cumulative Proportion is reaching 95%

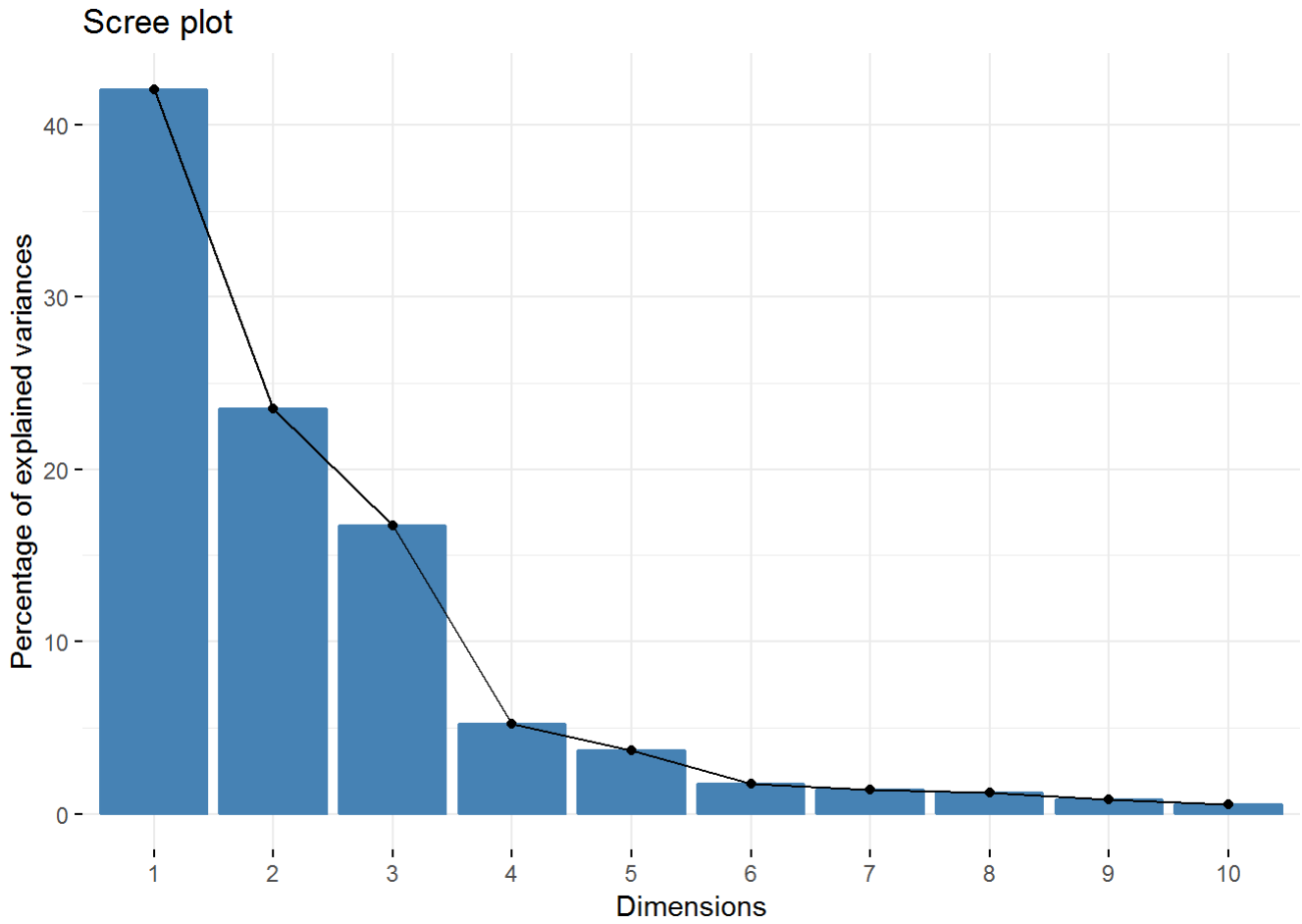
3

```
pc.pipline(data, T, T)
```

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


PCA - Biplot
UTX
HD
XOM
JPM
DWDP
IBM
CAT
GS
BA
AAPL
CVX
MMM
TRV
AXP
DIS
NKE
JNJ
WMT
MSFT
CSCO
UNH
WBA
PFE
VZ
KO
PG
MRK
MCD
INTC
5
0
-5
-10
-7.5
-5.0
-2.5
0.0
2.5
5.0
Dim2 (23.5%)
Dim1 (42.1%)

```
##
## [[2]]
```


Scree plot
Percentage of explained variances
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Dimensions

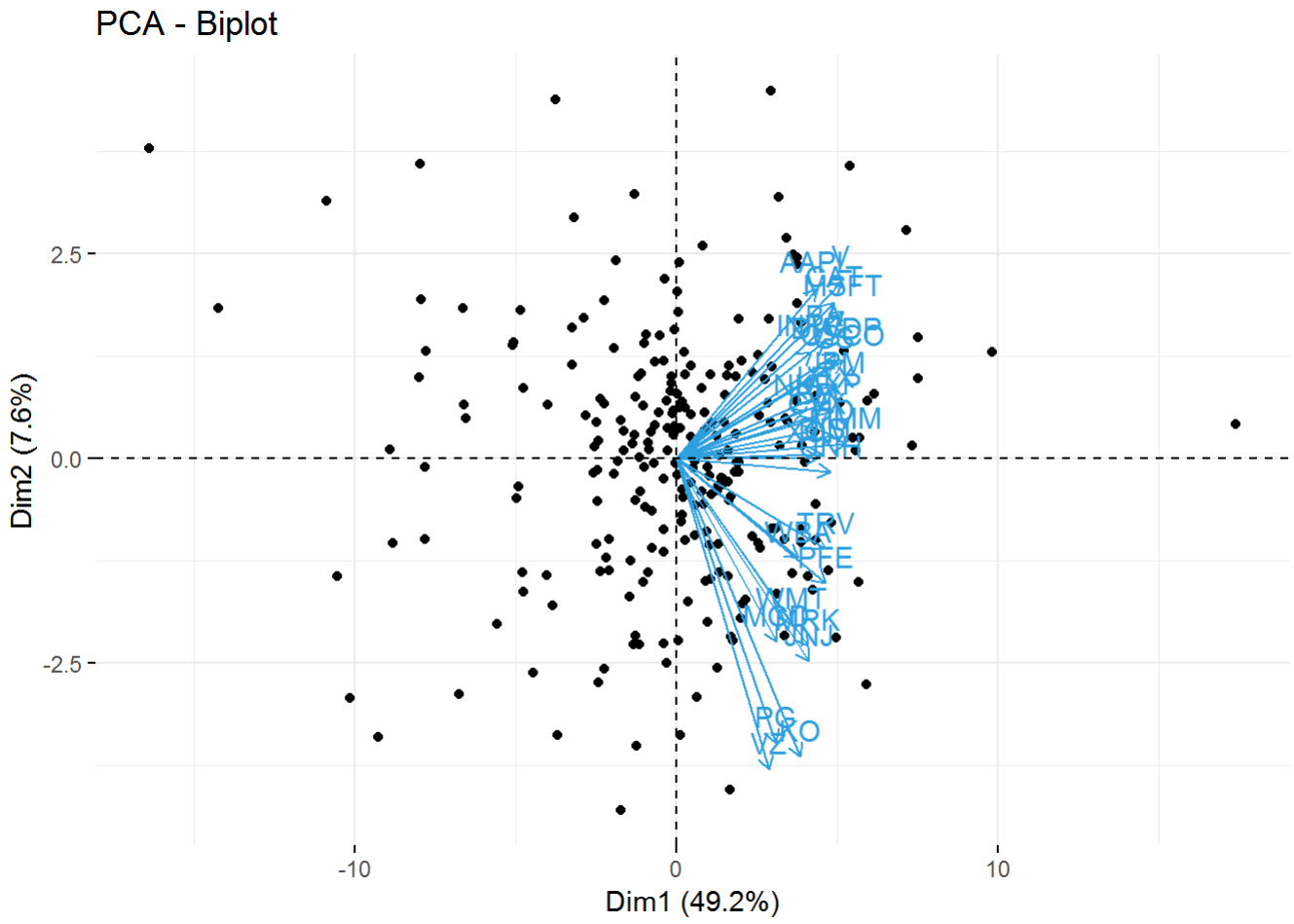
```
##
## [[3]]
## Importance of components:
##                            Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
## Standard deviation     3.5523912 2.6555330 2.2395918 1.25389698 1.04804610
## Proportion of Variance 0.4206495 0.2350618 0.1671924 0.05240859 0.03661335
## Cumulative Proportion  0.4206495 0.6557113 0.8229037 0.87531227 0.91192562
##                            Comp.6      Comp.7      Comp.8       Comp.9
## Standard deviation     0.72565783 0.64167393 0.60387369 0.505965714
## Proportion of Variance 0.01755264 0.01372485 0.01215545 0.008533377
## Cumulative Proportion  0.92947827 0.94320311 0.95535856 0.963891939
##                           Comp.10      Comp.11     Comp.12      Comp.13
## Standard deviation     0.410284132 0.396849631 0.36373231 0.328427242
## Proportion of Variance 0.005611102 0.005249654 0.00441004 0.003595482
## Cumulative Proportion  0.969503041 0.974752696 0.97916274 0.982758217
##                           Comp.14      Comp.15      Comp.16      Comp.17
## Standard deviation     0.272696214 0.260903634 0.230339465 0.224142234
## Proportion of Variance 0.002478774 0.002269024 0.001768542 0.001674658
## Cumulative Proportion  0.985236991 0.987506015 0.989274557 0.990949215
##                           Comp.18     Comp.19      Comp.20       Comp.21
## Standard deviation     0.21427928 0.20014298 0.191223574 0.1670968650
## Proportion of Variance 0.00153052 0.00133524 0.001218882 0.0009307121
## Cumulative Proportion  0.99247974 0.99381498 0.995033858 0.9959645697
##                             Comp.22       Comp.23       Comp.24       Comp.25
## Standard deviation     0.1514865235 0.1452216399 0.1344319571 0.1244578452
## Proportion of Variance 0.0007649389 0.0007029775 0.0006023984 0.0005163252
## Cumulative Proportion  0.9967295086 0.9974324861 0.9980348844 0.9985512096
##                             Comp.26       Comp.27       Comp.28       Comp.29
## Standard deviation     0.1081772731 0.1059578703 0.0973126302 0.0763609557
## Proportion of Variance 0.0003900774 0.0003742357 0.0003156583 0.0001943665
## Cumulative Proportion  0.9989412870 0.9993155227 0.9996311810 0.9998255475
##                             Comp.30
## Standard deviation     0.0723434564
## Proportion of Variance 0.0001744525
## Cumulative Proportion  1.0000000000
```

Based on summary table, when choose 8 principle components, the cumulative Proportion is reaching 95%

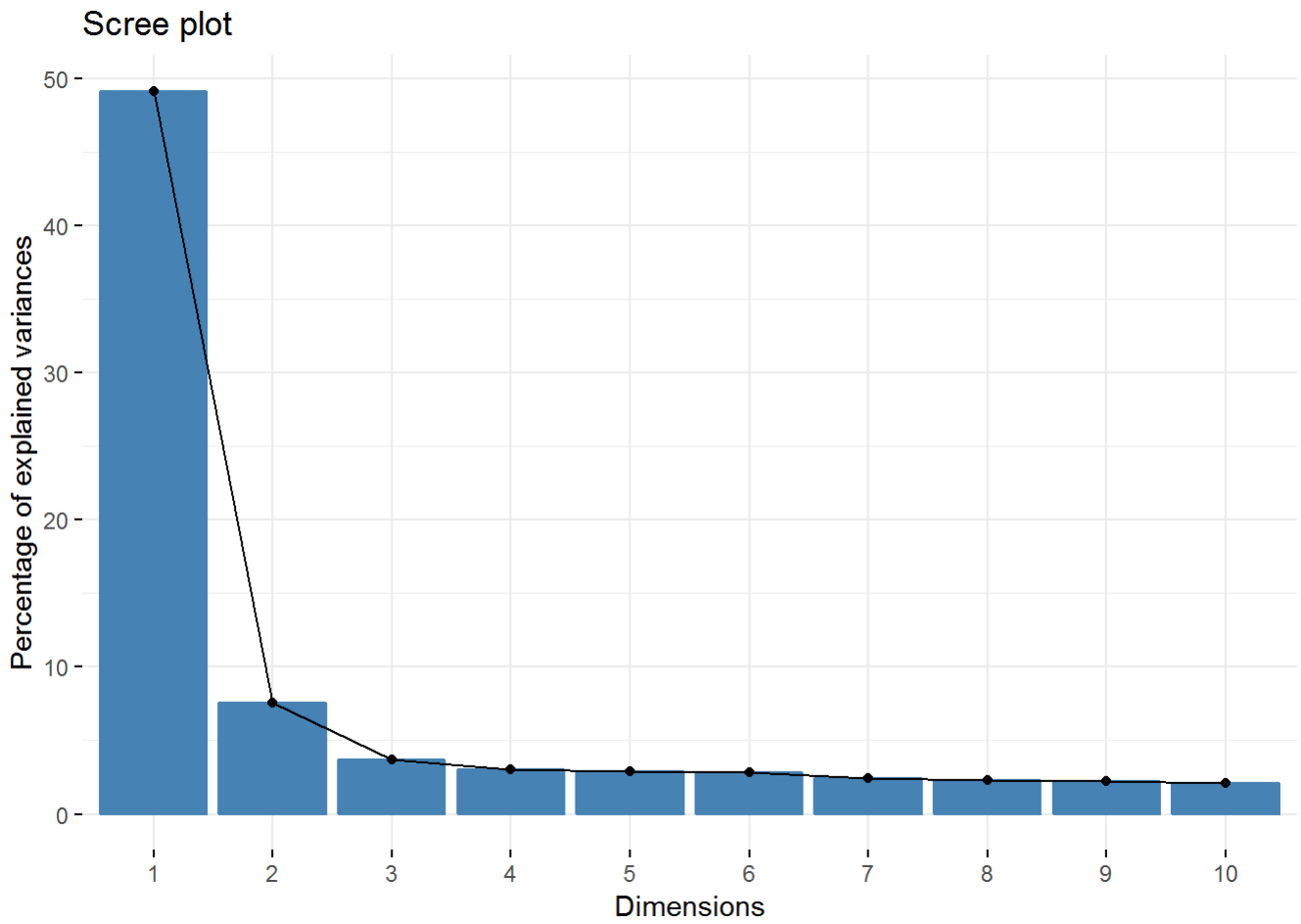
4

```
data = as.matrix(data)
data.today = as.matrix(data[1:250,])
data.tmr = as.matrix(data[2:251,])
return = (data.tmr - data.today)/data.today

pc.pipline(return, T, T)
```

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




```
##
## [[2]]
```


Scree plot
Percentage of explained variances
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Dimensions

```
##
## [[3]]
## Importance of components:
##                            Comp.1     Comp.2     Comp.3     Comp.4
## Standard deviation     3.8404888 1.50885324 1.05290551 0.95275091
## Proportion of Variance 0.4916451 0.07588794 0.03695367 0.03025781
## Cumulative Proportion  0.4916451 0.56753308 0.60448675 0.63474456
##                            Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     0.93304126 0.91670678 0.85279863 0.83329930
## Proportion of Variance 0.02901887 0.02801171 0.02424218 0.02314626
## Cumulative Proportion  0.66376343 0.69177514 0.71601732 0.73916358
##                            Comp.9    Comp.10    Comp.11    Comp.12
## Standard deviation     0.81591802 0.79273677 0.7739225 0.76333486
## Proportion of Variance 0.02219074 0.02094772 0.0199652 0.01942267
## Cumulative Proportion  0.76135432 0.78230204 0.8022672 0.82168991
##                           Comp.13    Comp.14    Comp.15    Comp.16
## Standard deviation     0.74118206 0.68191948 0.66388756 0.63920272
## Proportion of Variance 0.01831169 0.01550047 0.01469156 0.01361934
## Cumulative Proportion  0.84000160 0.85550208 0.87019363 0.88381297
##                           Comp.17    Comp.18    Comp.19    Comp.20
## Standard deviation     0.62841102 0.61011078 0.58701922 0.55889731
## Proportion of Variance 0.01316335 0.01240784 0.01148639 0.01041221
## Cumulative Proportion  0.89697632 0.90938416 0.92087054 0.93128275
##                           Comp.21    Comp.22    Comp.23    Comp.24
## Standard deviation     0.545656307 0.517781493 0.500702736 0.489181836
## Proportion of Variance 0.009924694 0.008936589 0.008356774 0.007976629
## Cumulative Proportion  0.941207441 0.950144030 0.958500804 0.966477433
##                           Comp.25    Comp.26    Comp.27    Comp.28
## Standard deviation     0.46764570 0.457587555 0.418239405 0.406139487
## Proportion of Variance 0.00728975 0.006979546 0.005830807 0.005498309
## Cumulative Proportion  0.97376718 0.980746729 0.986577536 0.992075845
##                           Comp.29    Comp.30
## Standard deviation     0.36465450 0.323653743
## Proportion of Variance 0.00443243 0.003491725
## Cumulative Proportion  0.99650828 1.000000000
```

The biplot is telling that all the return is focusing on chemical, energy, retail and information technologies industry. Based on screeplot, first three principle components have already catch most information from data.

If each stock were fluctuating up and down randomly and independent of all the stocks, then expect each principle component has similar weight because each direction catchs different information from data.

Problem 3

Let T be the centered inner product matrix with elements $< Z_i - \bar{Z}_i, Z_i - \bar{Z} >$, then $\sum_{i,i'}^{n} (s_{i,i'} - < z_i - \bar{z}, z_{i'} - \bar{z} >)$

Assume that the eign decomposition of S and T are UD^2U^T and $\tilde{U}\tilde{D}^2\tilde{U}^T$

$$S_c(z_1, z_2, \ldots, z_n) = tr[(UD^2U^T - \tilde{U}\tilde{D}^2\tilde{U}^T)^2] = tr[D^4U^2U^{2T} + \tilde{D}^4\tilde{U}^2\tilde{U}^{2T} - 2D^2U^T\tilde{U}\tilde{D}^2\tilde{U}^TU]$$

$$= tr[D^4 + \tilde{D}^4 - 2D^2U^T\tilde{U}\tilde{D}^2\tilde{U}^TU]$$

Let $A = U^T\tilde{U}$, where a_{ij} is the $(i,j)th$ element of A

$$\frac{\partial S_c}{\partial \tilde{d}^2} = d^2 + \tilde{d}^2 - 2\sum_i^n d_i^2 a_{ij}^2$$

$$= 2\tilde{d}^2 - 2\sum_i^n d_i^2 a_{ij}^2$$

Minimizing $S_c(z_1, z_2, \ldots, z_n) \rightarrow \frac{\partial S_c}{\partial \tilde{d}^2} = 0$

$\rightarrow \tilde{d}^2 = \sum_i^n d_i^2 a_{ij}^2$

$\rightarrow max \sum_i^n d_i^2 a_{ij}^2 = \sum_i^n (\tilde{u}_i^T U D^2 U^T \tilde{u}_i)^2$

$\rightarrow S_c(z_1, z_2, \ldots, z_n) = X^T\tilde{U}$, where $\tilde{U}$ is $k \times k$ orthogonal matrix, and ith column stands for each principle component. Therefore, $X^T\tilde{U}$ is mapping to the first k principle components with the largest variance d^2