hw1

Qiuying Li UNI ql2280

2/6/2017

1. Apply the three subset selection methods mentioned above to Credit data set. Plot the RSS as a function of the number of variables for these three methods in the same figure.

```
knitr::opts_chunk$set(echo = TRUE)
setwd("~/Desktop/2017 spring/GR 5241/HW/HW1")
credit = read.csv("credit.csv", head = TRUE )
balance = credit$Balance
library(MASS)
library(leaps)
lm3 = regsubsets(balance ~., data = credit)
summary(lm3)$rss
```

```
## [1] 4.004304e-23 1.291686e-23 9.089095e-24 4.206989e-24 2.876464e-24
## [6] 2.670042e-24 2.634288e-24 0.000000e+00
```

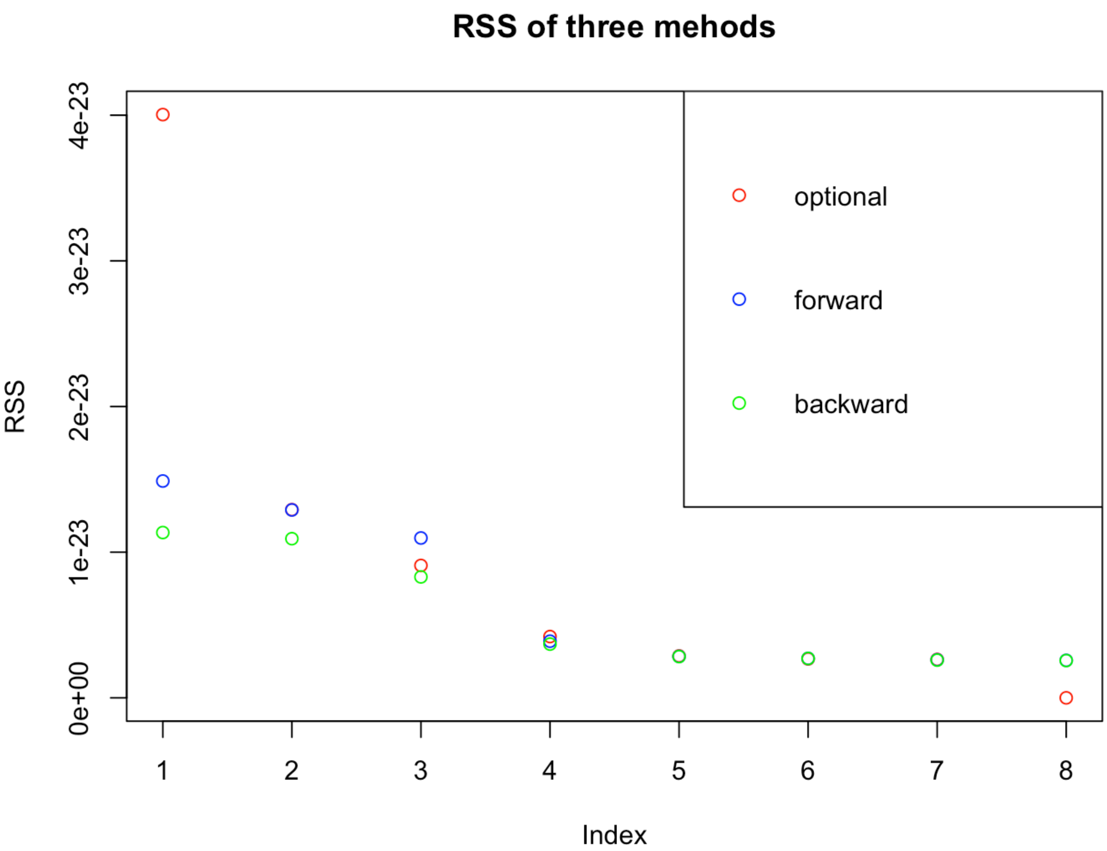
```
lm4 = regsubsets(balance ~., data = credit,method = "forward")
summary(lm4)$rss
```

```
## [1] 1.488617e-23 1.290117e-23 1.097386e-23 3.887676e-24 2.841082e-24
## [6] 2.708195e-24 2.610526e-24 2.573541e-24
```

```
lm5 = regsubsets(balance ~., data = credit,method = "backward")
summary(lm5)$rss
```

```
## [1] 1.134851e-23 1.092685e-23 8.301538e-24 3.689724e-24 2.833115e-24
## [6] 2.704195e-24 2.591497e-24 2.560741e-24
```

```
par(mfrow = c(1,1))
plot(summary(lm3)$rss,col = "red")
```

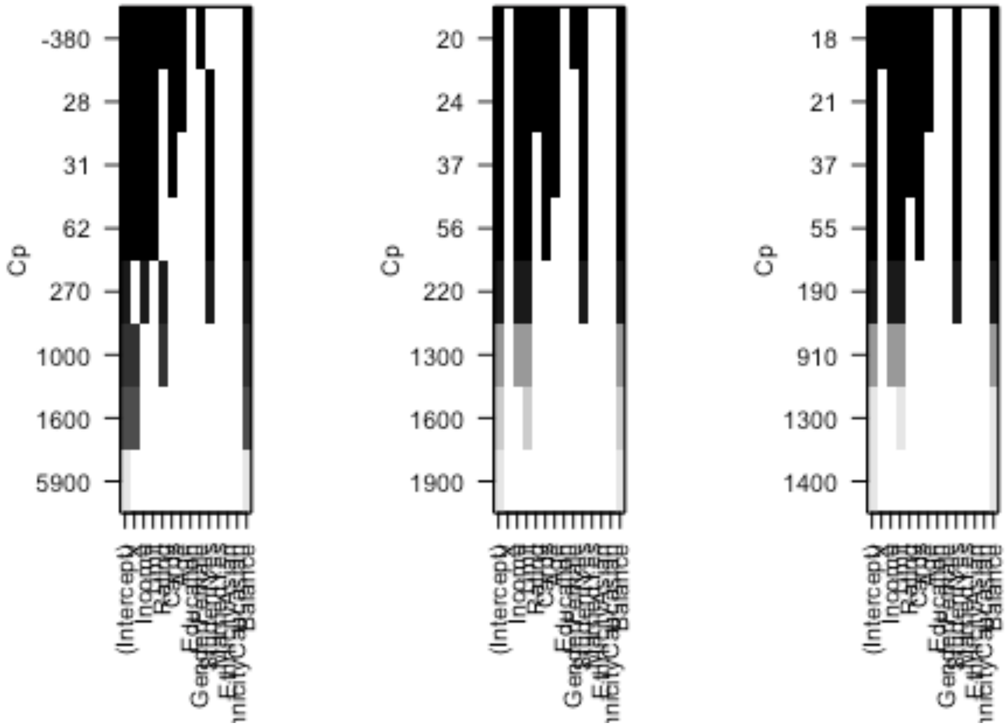


```
#points(summary(lm4)$rss,col = "blue")
#points(summary(lm5)$rss,col = "green")
#legend("topright",leg.txt,pch = 1,col= c("red","blue","green"))
```

From the above plot, we could tell that the Rss of three methods are very closed to each other.

2. Each subset selection method results in a set of models. For each approach, choose a single optimal model by using Cp and BIC statistics respectively. Report the optimal models for each approach (i.e. specify the predictors in the optimal model).

```
par(mfrow = c(1,3))
plot(lm3,scale = "Cp")
plot(lm4,scale = "Cp")
plot(lm5,scale = "Cp")
```



```
summary(lm3,scale = "Cp")

## Subset selection object
## Call: regsubsets.formula(balance ~ ., data = credit)
## 13 Variables  (and intercept)
##                     Forced in Forced out
## X                       FALSE      FALSE
## Income                  FALSE      FALSE
## Limit                   FALSE      FALSE
## Rating                  FALSE      FALSE
## Cards                   FALSE      FALSE
## Age                     FALSE      FALSE
## Education               FALSE      FALSE
## GenderFemale            FALSE      FALSE
## StudentYes              FALSE      FALSE
## MarriedYes              FALSE      FALSE
## EthnicityAsian          FALSE      FALSE
## EthnicityCaucasian      FALSE      FALSE
## Balance                 FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          X   Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   " "    " "   " " " "        " "
## 2  ( 1 ) "*" " "    " "   " "    " "   " " " "       " "         
## 3  ( 1 ) "*" " "    " "   "*"    " "   " " " "       " "         
## 4  ( 1 ) " " "*"    " "   "*"    " "   " " " "       " "         
## 5  ( 1 ) "*" "*"    "*"   " "    " "   " " " "       " "         
## 6  ( 1 ) "*" "*"    "*"   " "    "*"   " " " "       " "         
## 7  ( 1 ) "*" "*"    "*"   " "    "*"   "*" " "       " "         
## 8  ( 1 ) "*" "*"    "*"   "*"    "*"   "*" " "       "*"         
##           StudentYes MarriedYes EthnicityAsian EthnicityCaucasian Bal
ance
## 1  ( 1 ) " "        " "        " "            " "                "*"
    
## 2  ( 1 ) " "        " "        " "            " "                "*"
    
## 3  ( 1 ) " "        " "        " "            " "                "*"
    
## 4  ( 1 ) "*"        " "        " "            " "                "*"
    
## 5  ( 1 ) "*"        " "        " "            " "                "*"
    
## 6  ( 1 ) "*"        " "        " "            " "                "*"
    
## 7  ( 1 ) "*"        " "        " "            " "                "*"
    
## 8  ( 1 ) " "        " "        " "            " "                "*"
```

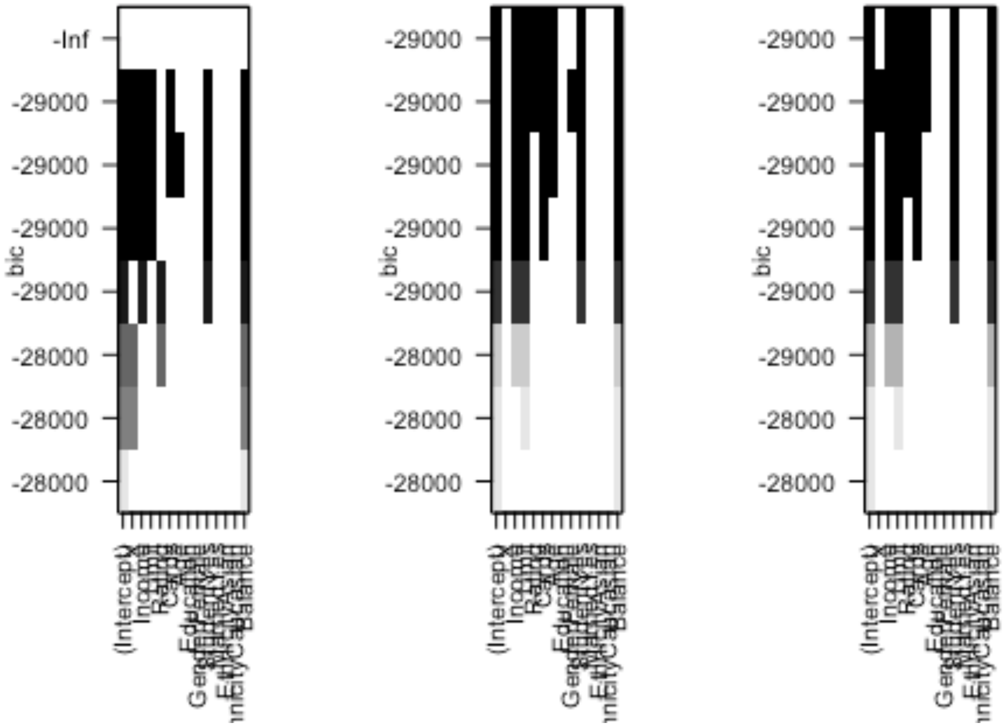
```
summary(lm4,scale = "Cp")
```

```
## Subset selection object
## Call: regsubsets.formula(balance ~ ., data = credit, method = "forwa
rd")
## 13 Variables  (and intercept)
##                    Forced in Forced out
## X                      FALSE      FALSE
## Income                 FALSE      FALSE
## Limit                  FALSE      FALSE
## Rating                 FALSE      FALSE
## Cards                  FALSE      FALSE
## Age                    FALSE      FALSE
## Education              FALSE      FALSE
## GenderFemale           FALSE      FALSE
## StudentYes             FALSE      FALSE
## MarriedYes             FALSE      FALSE
## EthnicityAsian         FALSE      FALSE
## EthnicityCaucasian     FALSE      FALSE
## Balance                FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##           X   Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   " "    " "   " " " "       " "         
## 2  ( 1 ) " " " "    "*"   " "    " "   " " " "       " "         
## 3  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 4  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 5  ( 1 ) " " "*"    "*"   " "    "*"   " " " "       " "
## 6  ( 1 ) " " "*"    "*"   " "    "*"   "*" " "       " "
## 7  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       " "
## 8  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       "*"
##          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian Bal
ance
## 1  ( 1 ) " "        " "        " "            " "                "*"

## 2  ( 1 ) " "        " "        " "            " "                "*"

## 3  ( 1 ) " "        " "        " "            " "                "*"

## 4  ( 1 ) "*"        " "        " "            " "                "*"

## 5  ( 1 ) "*"        " "        " "            " "                "*"

## 6  ( 1 ) "*"        " "        " "            " "                "*"

## 7  ( 1 ) "*"        " "        " "            " "                "*"

## 8  ( 1 ) "*"        " "        " "            " "                "*"
```

```
summary(lm5,scale = "Cp")
```

```
## Subset selection object
## Call: regsubsets.formula(balance ~ ., data = credit, method = "backw
ard")
## 13 Variables  (and intercept)
##                    Forced in Forced out
## X                      FALSE      FALSE
## Income                 FALSE      FALSE
## Limit                  FALSE      FALSE
## Rating                 FALSE      FALSE
## Cards                  FALSE      FALSE
## Age                    FALSE      FALSE
## Education              FALSE      FALSE
## GenderFemale           FALSE      FALSE
## StudentYes             FALSE      FALSE
## MarriedYes             FALSE      FALSE
## EthnicityAsian         FALSE      FALSE
## EthnicityCaucasian     FALSE      FALSE
## Balance                FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          X   Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   " "    " "   " " " "       " "
## 2  ( 1 ) " " " "    "*"   " "    " "   " " " "       " "
## 3  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 4  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 5  ( 1 ) " " "*"    "*"   " "    "*"   " " " "       " "
## 6  ( 1 ) " " "*"    "*"   "*"    "*"   " " " "       " "
## 7  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       " "
## 8  ( 1 ) "*" "*"    "*"   "*"    "*"   "*" " "       " "
##          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian Bal
ance
## 1  ( 1 ) " "        " "        " "            " "                "*"

## 2  ( 1 ) " "        " "        " "            " "                "*"

## 3  ( 1 ) " "        " "        " "            " "                "*"

## 4  ( 1 ) "*"        " "        " "            " "                "*"

## 5  ( 1 ) "*"        " "        " "            " "                "*"

## 6  ( 1 ) "*"        " "        " "            " "                "*"

## 7  ( 1 ) "*"        " "        " "            " "                "*"

## 8  ( 1 ) "*"        " "        " "            " "                "*"
```

```r
plot(lm3,scale = "bic")
plot(lm4,scale = "bic")
plot(lm5,scale = "bic")
```



```
summary(lm3,scale = "bic")
```

```
## Subset selection object
## Call: regsubsets.formula(balance ~ ., data = credit)
## 13 Variables  (and intercept)
##                       Forced in Forced out
## X                         FALSE      FALSE
## Income                    FALSE      FALSE
## Limit                     FALSE      FALSE
## Rating                    FALSE      FALSE
## Cards                     FALSE      FALSE
## Age                       FALSE      FALSE
## Education                 FALSE      FALSE
## GenderFemale              FALSE      FALSE
## StudentYes                FALSE      FALSE
## MarriedYes                FALSE      FALSE
## EthnicityAsian            FALSE      FALSE
## EthnicityCaucasian        FALSE      FALSE
## Balance                   FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          X   Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   " "    " "   " " " "       " "
## 2  ( 1 ) "*" " "    " "   " "    " "   " " " "       " "
## 3  ( 1 ) "*" " "    " "   "*"    " "   " " " "       " "
## 4  ( 1 ) " " "*"    " "   "*"    " "   " " " "       " "
## 5  ( 1 ) "*" "*"    "*"   " "    " "   " " " "       " "
## 6  ( 1 ) "*" "*"    "*"   " "    "*"   " " " "       " "
## 7  ( 1 ) "*" "*"    "*"   " "    "*"   "*" " "       " "
## 8  ( 1 ) "*" "*"    "*"   "*"    "*"   "*" " "       "*"
##          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian Bal
ance
## 1  ( 1 ) " "        " "        " "            " "                "*"

## 2  ( 1 ) " "        " "        " "            " "                "*"

## 3  ( 1 ) " "        " "        " "            " "                "*"

## 4  ( 1 ) "*"        " "        " "            " "                "*"

## 5  ( 1 ) "*"        " "        " "            " "                "*"

## 6  ( 1 ) "*"        " "        " "            " "                "*"

## 7  ( 1 ) "*"        " "        " "            " "                "*"

## 8  ( 1 ) " "        " "        " "            " "                "*"
```

```
summary(lm4,scale = "bic")
```

```
## Subset selection object
## Call: regsubsets.formula(balance ~ ., data = credit, method = "forwa
rd")
## 13 Variables  (and intercept)
##                    Forced in Forced out
## X                      FALSE      FALSE
## Income                 FALSE      FALSE
## Limit                  FALSE      FALSE
## Rating                 FALSE      FALSE
## Cards                  FALSE      FALSE
## Age                    FALSE      FALSE
## Education              FALSE      FALSE
## GenderFemale           FALSE      FALSE
## StudentYes             FALSE      FALSE
## MarriedYes             FALSE      FALSE
## EthnicityAsian         FALSE      FALSE
## EthnicityCaucasian     FALSE      FALSE
## Balance                FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          X   Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   " "    " "   " " " "       " "
## 2  ( 1 ) " " " "    "*"   " "    " "   " " " "       " "
## 3  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 4  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 5  ( 1 ) " " "*"    "*"   " "    "*"   " " " "       " "
## 6  ( 1 ) " " "*"    "*"   " "    "*"   "*" " "       " "
## 7  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       " "
## 8  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       "*"
##           StudentYes MarriedYes EthnicityAsian EthnicityCaucasian Bal
ance
## 1  ( 1 ) " "        " "        " "            " "                "*"

## 2  ( 1 ) " "        " "        " "            " "                "*"

## 3  ( 1 ) " "        " "        " "            " "                "*"

## 4  ( 1 ) "*"        " "        " "            " "                "*"

## 5  ( 1 ) "*"        " "        " "            " "                "*"

## 6  ( 1 ) "*"        " "        " "            " "                "*"

## 7  ( 1 ) "*"        " "        " "            " "                "*"

## 8  ( 1 ) "*"        " "        " "            " "                "*"
```

```
summary(lm5,scale = "bic")
```

```
## Subset selection object
## Call: regsubsets.formula(balance ~ ., data = credit, method = "backw
ard")
## 13 Variables  (and intercept)
##                       Forced in Forced out
## X                         FALSE      FALSE
## Income                    FALSE      FALSE
## Limit                     FALSE      FALSE
## Rating                    FALSE      FALSE
## Cards                     FALSE      FALSE
## Age                       FALSE      FALSE
## Education                 FALSE      FALSE
## GenderFemale              FALSE      FALSE
## StudentYes                FALSE      FALSE
## MarriedYes                FALSE      FALSE
## EthnicityAsian            FALSE      FALSE
## EthnicityCaucasian        FALSE      FALSE
## Balance                   FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##           X   Income Limit Rating Cards Age Education GenderFemale
## 1  ( 1 ) " " " "    " "   " "    " "   " " " "       " "
## 2  ( 1 ) " " " "    "*"   " "    " "   " " " "       " "
## 3  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 4  ( 1 ) " " "*"    "*"   " "    " "   " " " "       " "
## 5  ( 1 ) " " "*"    "*"   " "    "*"   " " " "       " "
## 6  ( 1 ) " " "*"    "*"   "*"    "*"   " " " "       " "
## 7  ( 1 ) " " "*"    "*"   "*"    "*"   "*" " "       " "
## 8  ( 1 ) "*" "*"    "*"   "*"    "*"   "*" " "       " "
##          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian Bal
ance
## 1  ( 1 ) " "        " "        " "            " "                "*"

## 2  ( 1 ) " "        " "        " "            " "                "*"

## 3  ( 1 ) " "        " "        " "            " "                "*"

## 4  ( 1 ) "*"        " "        " "            " "                "*"

## 5  ( 1 ) "*"        " "        " "            " "                "*"

## 6  ( 1 ) "*"        " "        " "            " "                "*"

## 7  ( 1 ) "*"        " "        " "            " "                "*"

## 8  ( 1 ) "*"        " "        " "            " "                "*"
```

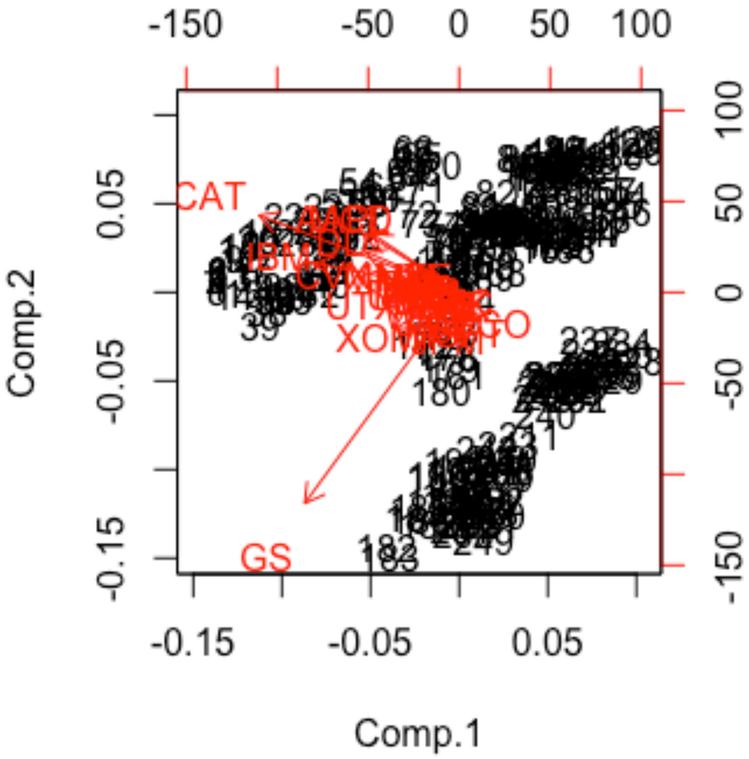
Problem 3 (PCA, 15 points) 1. For each of the 30 stocks in the Dow Jones Industrial Average, download the closing prices for every trading day from January 1, 2010 to January 1, 2011. Y

```
sym1 = c("MMM","AXP","AAPL","BA","CAT","CVX","CSCO","KO","DD","XOM","GE
",
         "GS","HD","IBM","INTC","JNJ","JPM","MCD","MRK","MSFT","NKE","P
FE",
         "PG","TRV","UNH","UTX","VZ","V","WMT","DIS")
web = NULL
title = NULL
stock = NULL
for (i in 1:length(sym1)){
  web[i] = paste("http://chart.finance.yahoo.com/table.csv?s=", sym1[i],
"&a=0&b=1&c=2010&d=0&e=1&f=2011&g=d&ignore=.csv",sep = "")
  title[i] = paste(sym1[i],".csv",sep = "")
  download.file (web[i],title[i],quiet = FALSE)
}
d = c(1:252)
 pr= matrix(d,nrow = 252,ncol = 30)
for( i in 1:length(sym1))
{
  stock = read.csv(title[i],header = T)
  m = stock$Adj.Close
  pr[,i] = m
}
colnames(pr) = sym1
str(pr)
```

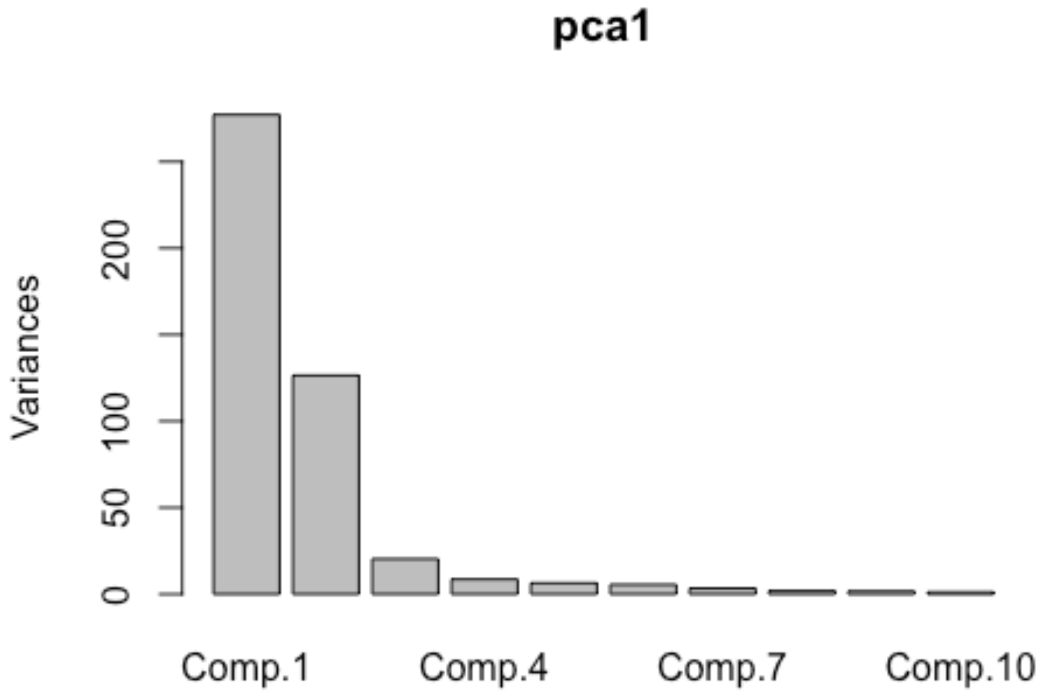
```
##  num [1:252, 1:30] 74.2 74.4 74.6 74.6 74.8 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:30] "MMM" "AXP" "AAPL" "BA" ...
```

2. Perform a PCA on the prices and create the biplot

```
par(mfrow = c(1,1))
pca1 = princomp(pr, cor = F,center = TRUE,scale. = TRUE)
biplot(pca1)
```

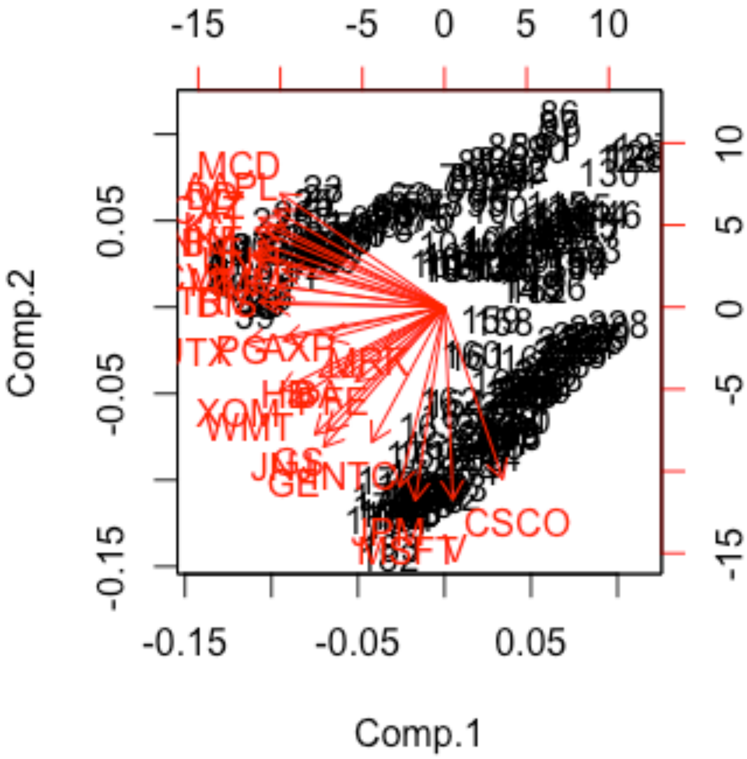


```r
screeplot(pca1)
```

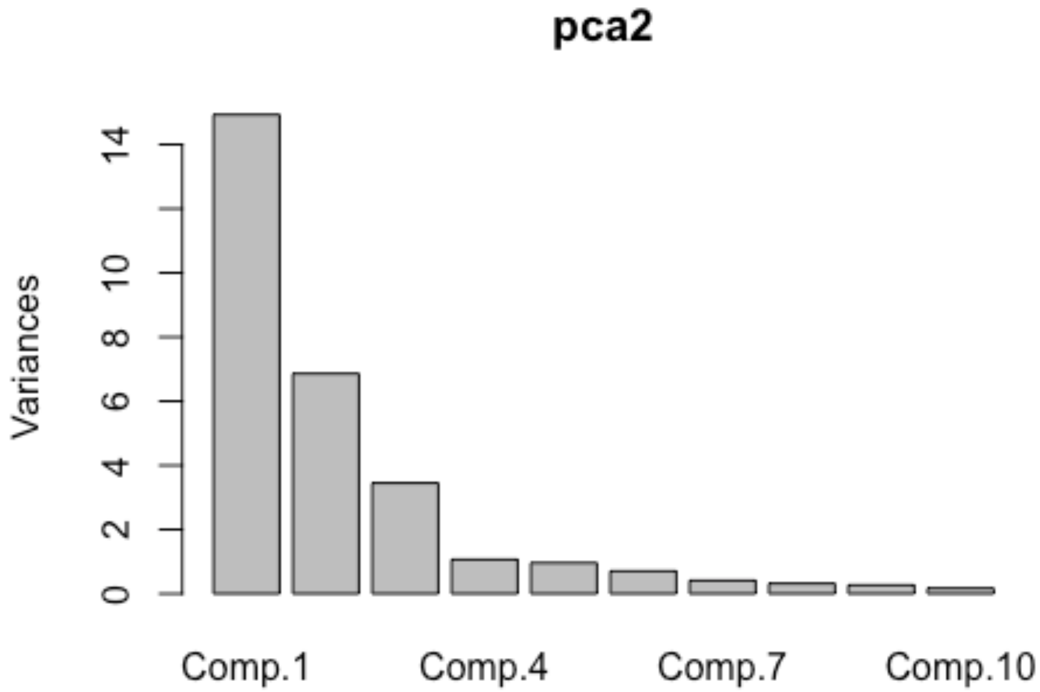


3.Repeat part 2 with cor=TRUE. This is equivalent to scale each column of the data matrix

```
pca2 = princomp(pr, cor = T,center = TRUE,scale. = TRUE)
biplot(pca2)
```



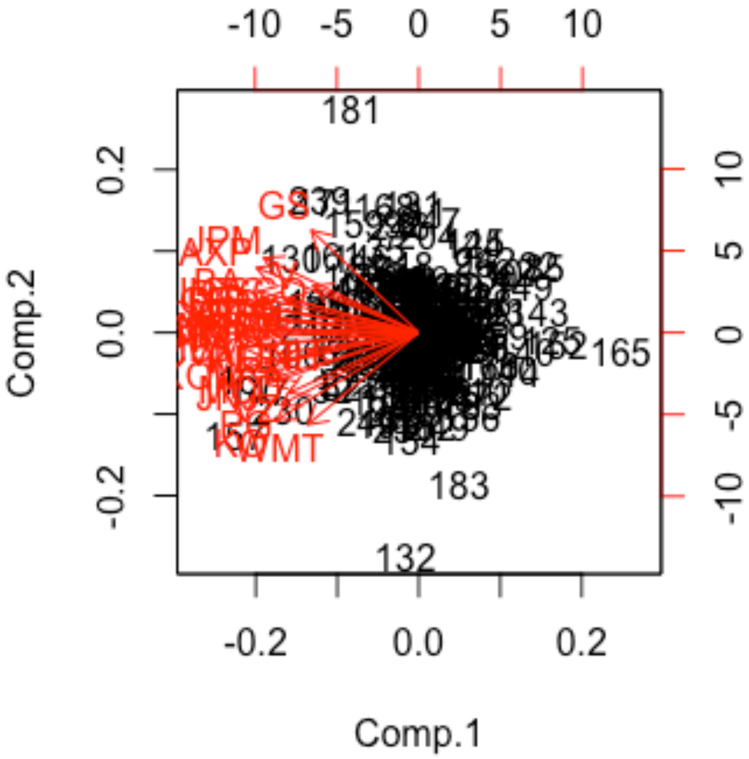
```
screeplot(pca2)
```



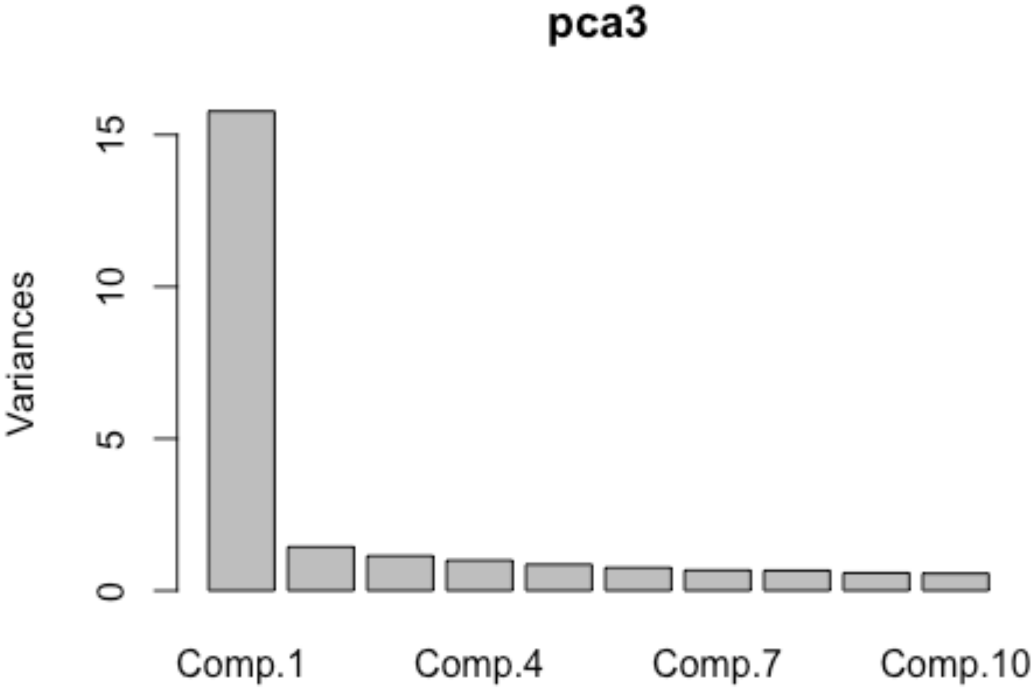
4.

Calculate the return for each stock, and repeat part 3 on the return data. I

```
pr1 = pr[-1,]
pr2 = pr[-252,]
ret1 = pr1 - pr2
pca3= princomp(ret1, cor = T,center = TRUE,scale. = TRUE)

## Warning: In princomp.default(ret1, cor = T, center = TRUE, scale. =
TRUE) :
##  extra arguments 'center', 'scale.' will be disregarded

biplot(pca3)
```



```r
screeplot(pca3)
```


pca3
Variances
0
5
10
15
Comp.1
Comp.4
Comp.7
Comp.10