Stat 154 Problem Set One

Jinze Gu SID:24968967

February 20, 2014

```
# Problem One Data Cleaning:
setwd("/Volumes/æIJĽèČ¡åĞžæšą/Stat 154/Stat 154")
meta <- read.csv("stock.csv")

## Warning:  æŮăæşŢæĽŞåijĂæŰĞäżű'stock.csv':  No such file or directory
## Error:  æŮăæşŢæĽŞåijĂéŞ¿çżŞ

# To calculate daily returns, I only need price, so I extract the data
# with price, date and company name.
stock <- data.frame(date = meta$date, COMP = meta$COMNAM, PRC = meta$PRC)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'meta'

# Just remove the raw data since it takes too much memory.
rm(meta)

## Warning:  æĽ¿äÿŊåĹřåŕźèśą'meta'

gc()

##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 262592 14.1     467875 25.0   407500 21.8
## Vcells 629202  4.9    1031040  7.9   786432  6.0

compname <- levels(factor(stock$COMP))

## Error:  æĽ¿äÿŊåĹřåŕźèśą'stock'

# f is a function that extract the price of the same company with 1342
# price records. It turns out that there are 422 companies with 1342
# sotck price records.
f <- function(x) {
    if (length(stock$PRC[stock$COMP == compname[x]]) == 1342) {
        return(stock$PRC[stock$COMP == compname[x]])
    }
}
stock_l <- sapply(c(1:length(compname)), f)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'compname'

# g is a function that extract the name of the company with 1342 price
# records.
g <- function(x) {
    if (length(stock$PRC[stock$COMP == compname[x]]) == 1342) {
        return(compname[x])
    }
}
names <- sapply(c(1:length(compname)), g)
```

```
## Error:  æĽ¿äÿŊåĹřåŕźèśą'compname'

# Now I construct the price_matrix with price record in different date as
# row and with different company as column.
price_ma <- matrix(unlist(stock_l), nrow = 1342)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'stock_l'

colnames(price_ma) <- unlist(names)

## Error:  åŔĆæŢřäÿŊæŸŕäÿšåĹŮ

# Problem One 1).  I calculated the daily return matrix based on
# price_ma, and I transformed daily return matrix into percentage
# representation so that I can avoid numerical issue.
daily_return <- (price_ma[-1, ] - price_ma[1:1341, ])/price_ma[1:1341, ] *
    100

## Error:  æĽ¿äÿŊåĹřåŕźèśą'price_ma'

# The I do PCA to the daily return matrix with scaled value.
prin_stock <- prcomp(daily_return, rtex = TRUE, scale = TRUE)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'daily_return'

screeplot(prin_stock)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'

explain_var_stock <- sapply(1:422, function(i) sum(prin_stock$sdev[1:i]^2)/sum(prin_stock$sdev^2))

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'

# I can check the variance contribution of the first several principal
# components. Besides, it is reasonable to test how many principal
# components we want to leave in order to keep sat 80% of information
head(explain_var_stock)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'explain_var_stock'

sum(explain_var_stock < 0.8)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'explain_var_stock'

# Interpretation of first few principal component loadings:
hist(prin_stock$rotation[, 1], main = "First Principal Direction")

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'

max(prin_stock$rotation[, 1])

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'

min(prin_stock$rotation[, 1])

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'

hist(prin_stock$rotation[, 2], main = "Second Principal Direction")

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'

max(prin_stock$rotation[, 2])

## Error:  æĽ¿äÿŊåĹřåŕźèśą'prin_stock'
```

```
min(prin_stock$rotation[, 2])

## Error:  æĽ¿äÿŊåĹřåŕźèśą’prin_stock’

hist(prin_stock$rotation[, 3], main = "Third Principal Direction")

## Error:  æĽ¿äÿŊåĹřåŕźèśą’prin_stock’

max(prin_stock$rotation[, 3])

## Error:  æĽ¿äÿŊåĹřåŕźèśą’prin_stock’

min(prin_stock$rotation[, 3])

## Error:  æĽ¿äÿŊåĹřåŕźèśą’prin_stock’

# Comment: Based on the principal loadings, in the first three principal
# component loadings, most of the variables(companies in this case) have
# similar variance. It means that each variable are equally important in
# accounting for the variability in the PC.
# Besides, I plotted the projected data on the first and second principal
# component direction. I don't think there is an obvious clustering.
plot(prin_stock$x[, 1], prin_stock$x[, 2], xlab = "First Principal Component",
    ylab = "Second Principal Component")

## Error:  æĽ¿äÿŊåĹřåŕźèśą’prin_stock’

# Compar the plot that I only use the projected data in the first and
# second PC direction rather than using all of the eigenvectors to
# project the data.
prin_stock <- prcomp(t(daily_return), rtex = TRUE, scale = TRUE)

## Error:  æĽ¿äÿŊåĹřåŕźèśą’daily_return’

stock_proj <- t(prin_stock$rotation[, 1:2]) %*% scale(daily_return)

## Error:  æĽ¿äÿŊåĹřåŕźèśą’prin_stock’

plot(t(stock_proj), main = "Projected data in the first and second PCs")

## Error:  æĽ¿äÿŊåĹřåŕźèśą’stock_proj’

# Comment: I did not see any obviously great clustering.
# Problem One 2).  If I use all 422 companies to do hieararchical
# clustering, then I have the following graph.
dist_stock <- dist(t(daily_return), method = "manhattan")

## Error:  æĽ¿äÿŊåĹřåŕźèśą’daily_return’

hc_stock <- hclust(dist_stock, method = "complete")

## Error:  æĽ¿äÿŊåĹřåŕźèśą’dist_stock’

plclust(hc_stock, labels = FALSE)

## Error:  æĽ¿äÿŊåĹřåŕźèśą’hc_stock’

# Comment: I chose manhattan metric since I believe daily return should
# be equally weighted for everyday and it is reasonable to calculate the
# absolute difference between daily return rather than the euclidean
# distance. From the plot, I can see that those companies can be
# classified as different groups based on their variance of daily return.
# I used only 10 companies which would generate a better plot.
sample <- c(1, 40, 60, 114, 210, 275, 89, 320, 170, 413)
dist_stock_sub <- dist(t(daily_return[, sample]), method = "manhattan")
```

```
## Error:  æĽ¿äÿŊåĹřåŕźèśą'daily_return'

hc_stock_sub <- hclust(dist_stock_sub, method = "complete")

## Error:  æĽ¿äÿŊåĹřåŕźèśą'dist_stock_sub'

plclust(hc_stock_sub, labels = NULL)

## Error:  æĽ¿äÿŊåĹřåŕźèśą'hc_stock_sub'
```

Problem One 1): Interpretation of first few vectors of PC loadings: Comment: Based on the principal loadings, in the first three principal component loadings, most of the variables(companies in this case) have similar variance. It means that each variable are equally important in accounting for the variability in the PC.

Problem One 2): Comment: I chose manhattan metric to do hclustering since I believe daily return should be equally weighted for everyday and it is reasonable to calculate the absolute difference between daily return rather than the euclidean distance. From the plot, I can see that those companies can be classified as different groups based on their variance of daily return.