Stat4DS / Homework 2

Maria Luisa Croci, Valerio Antonini

Stock, Dependency and Graphs

First of all we import all the libraries that we will need:

```
library(zoo)
library(tseries)
library(energy)
library(ggplot2)
library(GGally)
library(magrittr)
library(dplyr)
```

[1] *Look at the end of the page*

```
#library(shiny)
```

The informations about each companies of S&P have been downloaded from https://en.wikipedia.org/wiki/List_of_S%26P_500_companies (https://en.wikipedia.org/wiki/List_of_S%26P_500_companies).

```
companies <- read.csv("~/Desktop/Universita/SM - Brutti/HW2/table-1.csv", encoding="UTF-8")
head(companies[,1:4])
```

```
##   Symbol            Security SEC.filings            GICS.Sector
## 1    MMM          3M Company     reports            Industrials
## 2    ABT Abbott Laboratories     reports            Health Care
## 3   ABBV         AbbVie Inc.     reports            Health Care
## 4   ABMD         ABIOMED Inc     reports            Health Care
## 5    ACN       Accenture plc     reports Information Technology
## 6   ATVI Activision Blizzard     reports Communication Services
```

We can extract a sample of 5 stocks for each sector, as follow:

```
#take a sample of 5 companies for each sector. We have 11 sectors, so the df returns 55 observations
companies_sample <- companies %>% group_by(GICS.Sector) %>% sample_n(5)
symbol = companies_sample["Symbol"]
```

So we are condisering the following companies:

1. Communication Services: Netflix (NFLX), Twenty-First Century Fox Class A (FOXA), Nike (NKE), The Walt Disney Company (DIS), Twenty-First Century Fox Class B (FOX)

2. Consumer Discretionary: Harley Davidson (HOG, Amazon (AMZN), eBay Inc. (EBAY), Hasbro (HAS), McDonald's Corp. (MCD)

3. Consumer Staples: Coca Cola Company (KO), Este-laudeer (EL), Colgate (CL), Kellogs (K), Campbell Soup (CPB)

4. Energy Anadarko Petroleum Corp (APC), Chevron Corp (CVX), EOG Resources (EOG), HollyFrontier Corp (HFC), Noble Energy Inc (NBL)

5. Financials American Express Co (AXP), Bank of America Corp (BAC), Capital One Financial (COF), Lincoln National (LNC), Nasdaq Inc. (NDAQ)

6. Health Care Biogen Inc. (BIIB), Johnson & Johnson (JNJ), Universal Health Services Inc. (UHS), Waters Corporation (WAT), WellCare (WCG)

7. Industrials FedEx Corporation (FDX), Eaton Corporation (ETN), Caterpillar (CAT), General Dynamics (GD), Masco Corp. (MAS)

8. Information Technology Adobe Systems Inc (ADBE), Apple (AAPL), Autodesk Inc. (ADSK), Intel Corp. (ANTC), Microsoft Corp. (MSFT)

9. Materials Air Products & Chemicals Inc (APD), Albemarle Corp (ALB), Ball Corp (BLL), Nucor Corp. (NUE), The Mosaic Company (MOS)

10. Real Estate Boston Properties (BXP), Equinix (EQIX), Extra Space Storage (EXR), Kimco Realty (KIM), Public Storage (PSA)

11. Utilities Consolidated Edison (ED), Duke Energy (DUK), Exelon Corp. (EXC), FirstEnergy Corp (FE), SCANA Corp (SCG)

In order to give the opportunity of running the code with the same companies that we have sampled, the vector "symbol" has been overwritten.

Morover 8 functions have been computed in order to replicate easly the same analysis on other time series and to avoid the repetition of the code:

- "clean" use as input a zoo serie that is converted into a data frame, remove the columns (companies) with missing values (NA) and reconvert the data frame into a zoo serie;
- "log_fun" use as input a zoo series, calculate the X matrix of the logarithms defined below and converts it as a data frame;

- “bootstrap_fun” use as input the matrix X of the logarithms and the matrix of the correlations of X. The function calculate the bootstrap procedure in order to build the marginal correlation graphs;
- “graph” use as input the matrix X, the matrix of the correlations of X, the matrix of the correlations estimated with the bootstrap function and a fixed value ϵ. It return the marginal correlation graph based on Pearson coefficient;
- “name_sectors” use as input two data frames, and return the first data with two columns that indicates the company’s name and the sector it belongs to. This function is use inside the functions “graph”, “graph_test” and “graph_test_bonf” in order to display different colors for different sectors;
- “test_dcov” use as input the matrix X and return the matrix with the p-values of the multiple hypothesis testing between each pair of column;
- “graph_test” use as input the matrix X and the p-values from “test_dcov”. It return the marginal correlation graph based on the distance covariance;
- “graph_test_bonf” use as input the same of “graph_test” plus “m” that is the number of multiple hypothesis test. It return the marginal correlation graph based on the distance covariance with the Bonferroni correction.

```
symbol = c("NFLX", "FOXA", "NKE", "DIS", "FOX", "HOG", "AMZN", "EBAY", "HAS", "MCD", "KO", "EL", "CL", "K", "CPB",
"APC", "CVX", "EOG", "HFC", "NBL", "AXP", "BAC", "COF", "LNC", "NDAQ", "BIIB", "JNJ", "UHS", "WAT", "WCG", "FDX",
"ETN", "CAT", "GD", "MAS", "ADBE", "AAPL", "ADSK", "INTC", "MSFT", "APD", "ALB", "BLL", "NUE", "MOS", "BXP", "EQIX
", "EXR", "KIM", "PSA", "ED", "DUK", "EXC", "FE", "SCG")

invisible(capture.output(for (i in symbol) {
  #invisible avoid of printing all the series in output
  assign(i,
    get.hist.quote(instrument = i, start="2003-01-01", end="2008-01-01",
                   quote= c("Close"), provider="yahoo", drop=TRUE)
)

}))

prova <- merge(NFLX, FOXA, NKE, DIS, FOX, HOG, AMZN, EBAY, HAS, MCD, KO, EL, CL, K, CPB, APC, CVX, EOG, HFC, NBL,
AXP, BAC, COF, LNC, NDAQ, BIIB, JNJ, UHS, WAT, WCG, FDX, ETN, CAT, GD, MAS, ADBE, AAPL, ADSK, INTC, MSFT, APD, ALB
, BLL, NUE, MOS, BXP, EQIX, EXR, KIM, PSA, ED, DUK, EXC, FE, SCG)


#convert to df
clean <- function(a) {
  a <- as.data.frame(a)

  #remove columns with na values from the df (only two companies), at the end we will have 53 companies
  a <- a[ , colSums(is.na(a)) == 0]

  #reconvert to zoo series
  a <- as.zoo(a)
  return(a)
}

prova <- clean(prova)
```

“prova” contains 1258 observations (time series of closing prices) of 53 variables (companies)

Building of the matrix X

Since $x_{t,j} = log(\frac{c_{t,j}}{c_{t-1,j}}) = log(c_{t,j}) - log(c_{t-1,j})$ we define the logarithm as follow in order to calculate the matrix $X = [x_{t,j}]_{t,j}$:

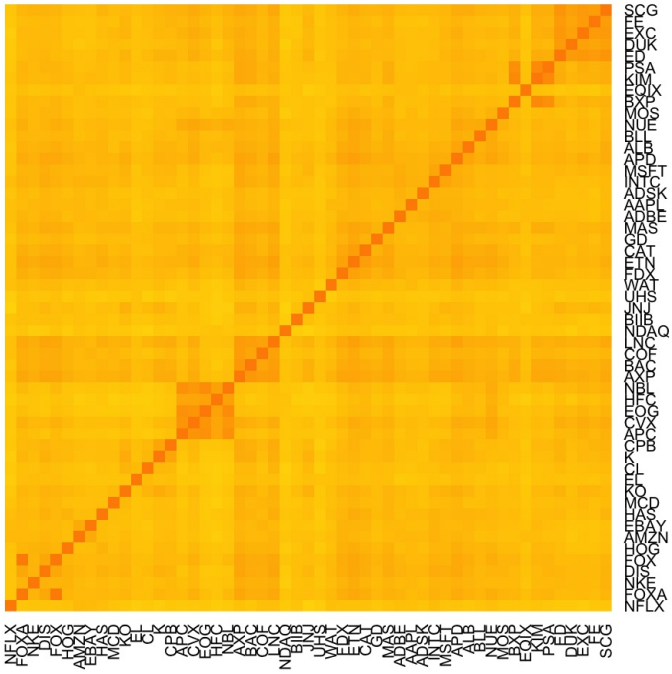
```
log_fun <- function(a) {

  a = diff(log(a))
  #convert to data frame
  a <- as.data.frame(a)

  return(a)
}

log <- log_fun(prova)
```

Calculating the Pearson correlation on X

```
#correlation pearson - R
corr <- cor(log, method = "pearson")
```

The heatmap shows the correlation between each pair of company: darker color means high correlation and viceversa. As we expected, the main diagonal confirm the maximun correlation (equal to 1) because the pair is between the same company. Morover we have all the other pairs with a low correlation.

```
#heatmap of correlations
col_heatmap <- colorRampPalette(c("gold", "dark orange"))
heatmap(corr, col = col_heatmap(100), scale="none", Rowv = NA, Colv = NA)
```



Implementing bootstrap procedure

Each element of the “delta” vector is calculated according to the following formula:

$$\Delta_b = \sqrt{n} * max_{j,k} | \hat{R}_b^*[j,k] - \hat{R}[j,k] |$$

In our case “corrB” represents $\hat{R}_b^*[j,k]$ and “corr” represents $\hat{R}[j,k]$.

n is the number of observations equal to 1258.

```
bootstrap_fun <- function(data, stat) {
  n <- length(data[ ,1])

  #num repetitions
  b <- 1000

  # vector of statistics
  d <- rep(NA, b)

  for (i in 1:b) {
    #unit for each boostrap sample
    u <- sample(1:n, replace = TRUE)
    # bootstrap sample
    x <- data[u, ]
    #matrix corr R_hat for bootstrap sample
    corrB <- cor(x, method = "pearson")
    #  delta statistics
    d[i] <- max(abs(corrB - stat))*sqrt(n)

    return(list(corrB, d))

  }
}

boot1 <- bootstrap_fun(log, corr)
corrB <- boot1[[1]]
delta <- boot1[[2]]
```

Create marginal correlation graph for Bootstrap procedure

A graph is a representation of elements called nodes or vertex connected through a link or edge. A graph could be connected and/or completed. The first case each vertex can be reached by any path, the second case each vertex is linked with the others. To build the marginal correlation graph we start from the following requirements:

– $|\rho(j,k)| \geq \epsilon$

– $\rho(j,k) \in [L,U]$

– $[-\epsilon,\epsilon] \cap (C_{n,\alpha} = [L,U]) = \varnothing$

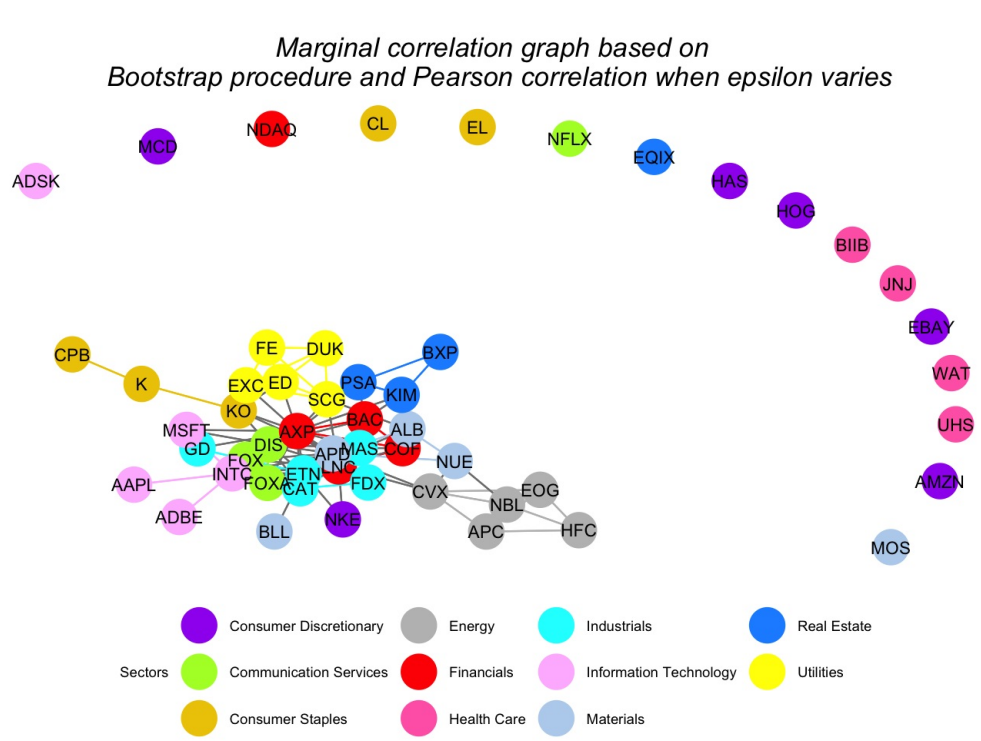
So, replacing ρ with the set of the confidence interval:

- $\epsilon \leq U$ and $\epsilon \leq L$
- $-\epsilon \geq U$ and $-\epsilon \geq L$

```
graph <- function(data, statistic, corr_est, epsilon) {

  n <- length(data[ ,1])
  # delta's ecdf
  ecdf_delta <- ecdf(statistic)

  #fixed alpha
  alpha = 0.05

  #intervals confidence
  t <- quantile(ecdf_delta, 1-alpha/2)

  L <- corr_est - t/sqrt(n) #lower bound
  U <- corr_est + t/sqrt(n) #upper bound

  len_ad <- length(data)

  adjacency <- matrix(NA, len_ad, len_ad)

  for (i in 1:len_ad) {
    for (j in 1:len_ad) {
      if (( epsilon <= L[i, j] && epsilon <= U[i, j]) || ( -epsilon >= L[i, j] && -epsilon >= U[i, j] ))  {
        # if epsilon in the confidence interval, an edge is put between i and j
        adjacency[i, j] = 1
      } else {
        adjacency[i, j] = 0
      } }
  }

  adjacency <- name_sectors(adjacency, log)

  plot(ggnet2(adjacency[, 1:53], label = T, color = adjacency$sector, palet = col, alpha.palette = 10, size = 8, s
hape = 16, edge.color = c("color", "grey50"), mode = "kamadakawai", edge.size = 0.4, color.legend = "Sectors", leg
end.position = "bottom", label.size = 3, legend.size = 7)+ggtitle("Marginal correlation graph based on \n Bootstra
p procedure and Pearson correlation when epsilon varies") +theme(plot.title=element_text( hjust=0.5, vjust=1, face
='italic')))

}

name_sectors <- function(adj, data) {

  rownames(adj) <- colnames(data)
  colnames(adj) <- colnames(data)

  adj <- as.data.frame(adj)
  adj$symbol <- colnames(data)

  pos <- match(adj$symbol, companies$Symbol)
  adj$sector <- companies$GICS.Sector[pos]

  return(adj)

}

col <- c(`Consumer Discretionary` = "purple",`Communication Services` = "greenyellow", `Consumer Staples` = "gold2
", Energy = "grey", Financials = "red", `Health Care` = "hotpink", Industrials = "cyan", `Information Technology`
= "plum1", Materials = "slategray2", `Real Estate` = "dodgerblue1", Utilities = "yellow")
```

When ϵ decrease we obtain a graph connected and viceversa, with an higher value the graph doesn't have any link. From our analysis if ϵ goes from 0.25 to 0.4 we get acceptable results. The following rapresentation is for $\epsilon = 0.3$.

```
graph(log, delta, corrB, 0.3)
```



(2)

```
#sliderInput("epsilon", "Fix an epsilon", min = 0, max = 0.5, value = 0.25, step = 0.05)
#renderPlot({graph(log, delta, corrB, epsilon=input$epsilon)})
```

In this case companies of the same sector approache to cluster together as “Financial”, “Utilities”, “Energy” and “Industrials”, but at the same time other stocks are isolated.

Multiple hypothesis testing of distance covariance

Define the following null hypothesis:

- $H_0, \dots, H_m : \gamma^2_{i,j} = 0$

We want to replicate the bootstrap permutation $R = 1000$ times. The index represents the power of the Euclidean distance $||x_i - x_j||^s$ with $s \in (0, 2]$, we fixed it equal to 0.001.

If $R \uparrow$ and $s \to 0$, then the p-values are different and accurate.

Since the following tests require around an hour and a half, we decide to save locally the results of the p-values and not running in mardowndown this code.

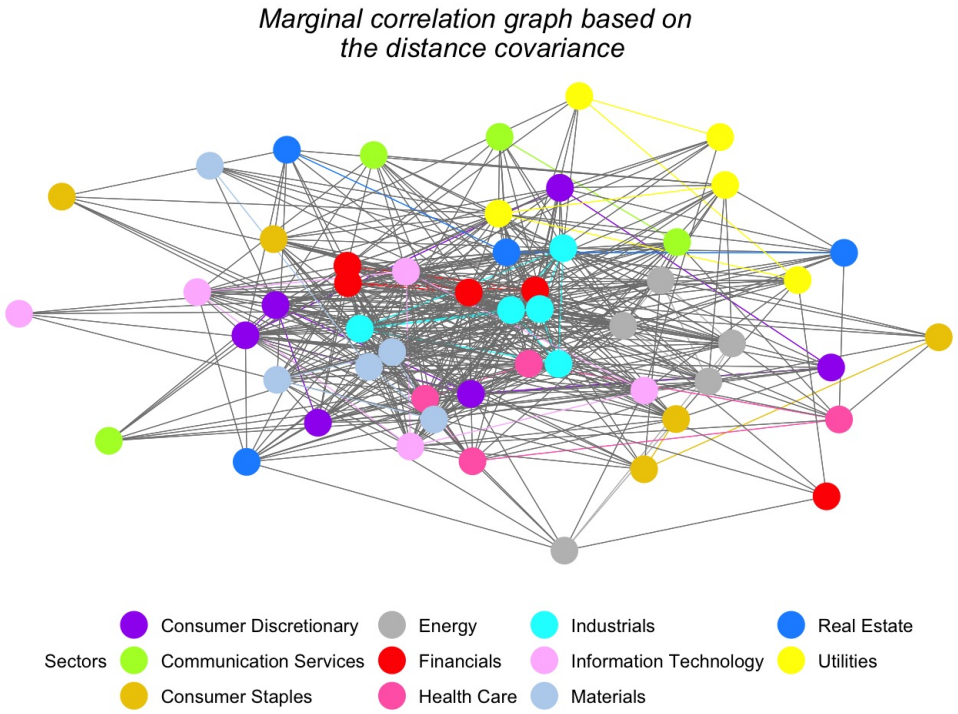
```
test_dcov <- function(data) {
  len_ad <- length(data)
  pMatrix <- matrix(NA, len_ad, len_ad)

  for (i in 1:len_ad) {
    for (j in 1:len_ad) {
      if (j >= i) {
        hp <- dcov.test(data[ ,i], data[ ,j], index=0.001, R=1000)
        pMatrix[i,j] <- hp$p.value
      }
      else {
        pMatrix[i,j] <- pMatrix[j,i]
      }
    }
  }

  return(pMatrix)

}

pMatrix <- test_dcov(log)
```

Load the p-values (this file contains also the pvalues for the time series of the next analysis).

```
load("pvalues.RData")
```

Given the matrix of p-values, we want to show the marginal correlation graph. We put an edge between i and j if and only if we reject the null hypothesis, that is when the p-values are less or equal than alpha.

```
graph_test <- function(data, pvalues) {

  len_ad <- length(data)
  padj <- matrix(NA, len_ad, len_ad)
  alpha <- 0.05

  for (i in 1:len_ad) {
    for (j in 1:len_ad) {
      if (pvalues[i,j] <= alpha)  {
        # if p value is less than alpha we reject the null hypotesis and we put an edge between i and j
        padj[i, j] = 1
      } else {

        padj[i, j] = 0
      } }
  }


  padj <- name_sectors(padj, log)

  plot(ggnet2(padj[, 1:53], color = padj$sector, palet = col, alpha.palette = 10, size = 6, shape = 16, edge.color
= c("color", "grey50"), mode = "fruchtermanreingold", edge.size = 0.2, color.legend = "Sectors", legend.position =
"bottom")+ggtitle("Marginal correlation graph based on \n the distance covariance") +theme(plot.title=element_text
( hjust=0.5, vjust=1, face='italic')))

}

graph_test(log, pMatrix)
```



As we can see the graph is connected and has an high number of links.

What's happened if we use the **Bonferroni correction**?

We expect to obtain a graph that isn't connected.

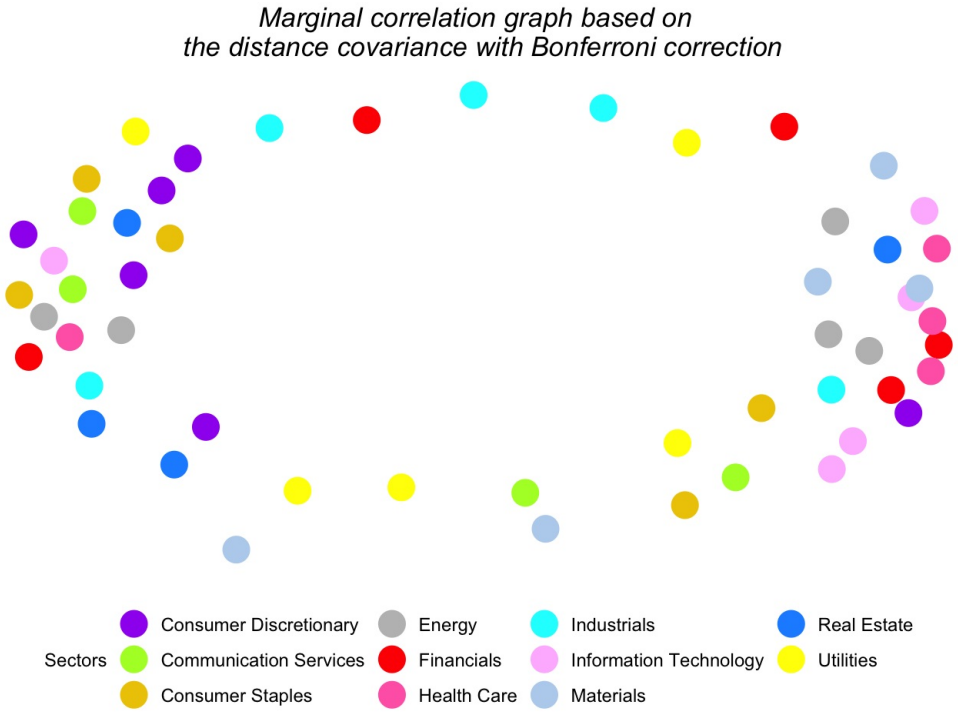
In general an hypotesys testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low. If multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis increases.

So Bonferroni correction rejects the null hypothesis for each $p $ where m is the number of tests. Since compare the test between stocks i and j is the same of stocks j and i, we calculate m as the combinations with replacing of k elements from a set of n:

-$C'_{n,k} = \binom{n+k-1}{k}$ where $n = 53$ is the number of total stocks, $k = 2$ the number of stocks in the test

- $m = \frac{54!}{2*52!} = 1431$

```
graph_test_bonf <- function(data, pvalues, m) {

  len_ad <- length(data)
  bonf_adj <- matrix(NA, len_ad, len_ad)
  alpha <- 0.05

  for (i in 1:len_ad) {
    for (j in 1:len_ad) {
      if (pvalues[i,j] <= alpha/m)  {
        # if p value is less or equal than alpha we reject the null hypotesis and we put an edge between i and j
        bonf_adj[i, j] = 1
      } else {
        bonf_adj[i, j] = 0
      } }
  }

  bonf_adj <- name_sectors(bonf_adj, log)

  plot(ggnet2(bonf_adj[, 1:53], color = bonf_adj$sector, palet = col, alpha.palette = 10, size = 6, shape = 16, ed
ge.color = "black", mode = "fruchtermanreingold", edge.size = 0.2, color.legend = "Sectors", legend.position = "bo
ttom")+ggtitle("Marginal correlation graph based on \n the distance covariance with Bonferroni correction") +theme
(plot.title=element_text( hjust=0.5, vjust=1, face='italic')))

}


graph_test_bonf(log, pMatrix, 1431)
```



Now the graph doesn't have any edge and the hypotesys testing has been corrected according to our expectation.

Repeat the analysis for the time series from 1st January 2013 to 1st January 2018 on the same portfolio of stocks

Collect the data

```
invisible(capture.output(for (i in symbol) {
  assign(i,
    get.hist.quote(instrument = i, start="2013-01-01", end="2018-01-01",
                   quote= c("Close"), provider="yahoo", drop=TRUE)
)

}))

prova2 <- merge(NFLX, FOXA, NKE, DIS, FOX, HOG, AMZN, EBAY, HAS, MCD, KO, EL, CL, K, CPB, APC, CVX, EOG, HFC, NBL,
AXP, BAC, COF, LNC, NDAQ, BIIB, JNJ, UHS, WAT, FDX, ETN, CAT, GD, MAS, ADBE, AAPL, ADSK, INTC, MSFT, APD, ALB, BLL
, NUE, MOS, BXP, EQIX, KIM, PSA, ED, DUK, EXC, FE, SCG)

prova2 <- clean(prova2)

log2 <- log_fun(prova2)

corr2 <- cor(log2, method = "pearson")
```

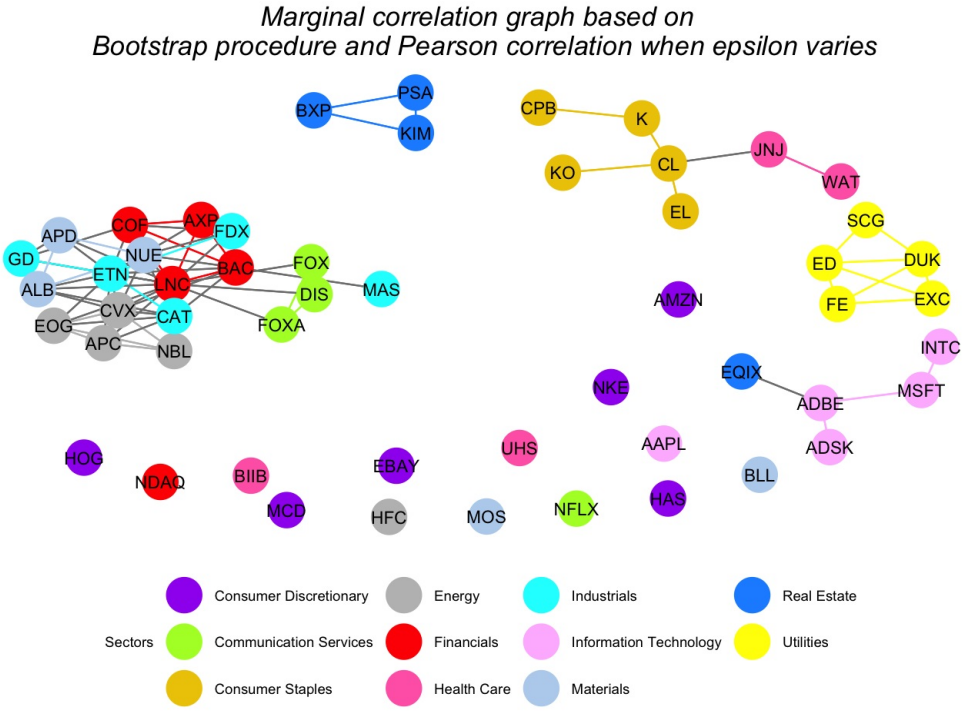
Bootstrap procedure

```
boot2 <- bootstrap_fun(log2, corr2)
corrB2 <- boot2[[1]]
delta2 <- boot2[[2]]
```

Marginal Correlation graph

Also in this case a reasonable value is for $\epsilon = 0.3$.

```
graph(log2, delta2, corrB2, 0.3)
```



(3)

```
#sliderInput("epsilon", "Fix an epsilon", min = 0, max = 0.5, value = 0.25, step = 0.05)
#renderPlot({graph(log2, delta2, corrB2, epsilon=input$epsilon)})
```

In this time series (2013 - 2018) the graph has more companies of the same area clustered together.

Let's see the main differences from the previous graph (2003 - 2008):

– We can observe that the clusters of the first analysis are still in the second one after the financial crisis;

– In addition we have other 3 sectors clustered:

- "Real Estate"
- "Consumer Staple"
- "Materials"

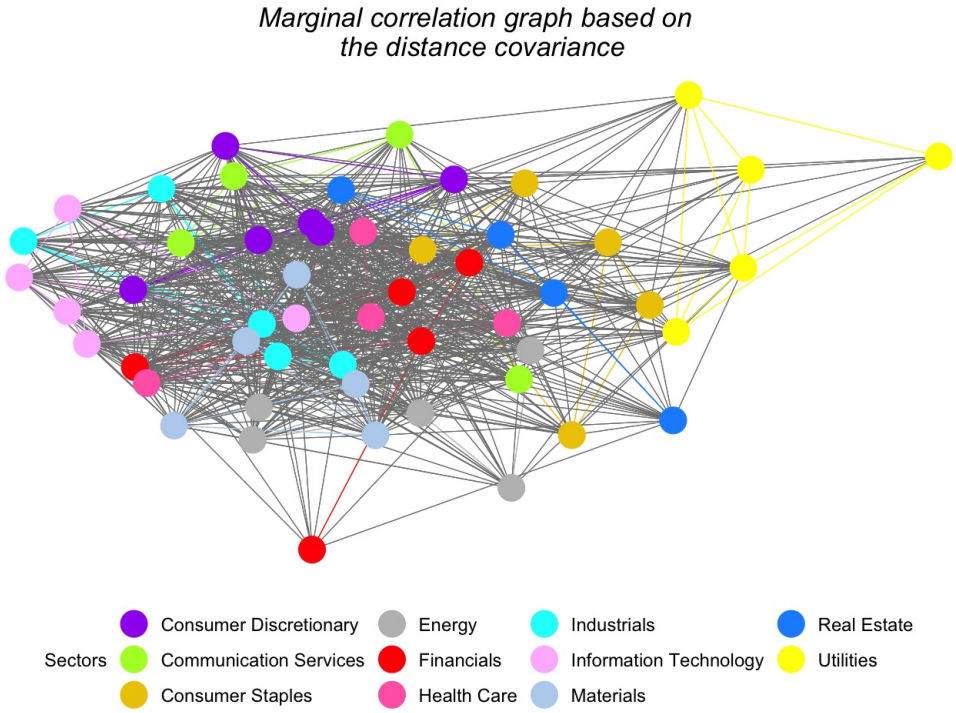
– However the "Energy" sector after the crisis has the "HFC" (*HollyFrontier Corp*) company separated from its cluster.

Multiple hypothesis testing of distance covariance and graph

```
pMatrix2 <-  test_dcov(log2)
```

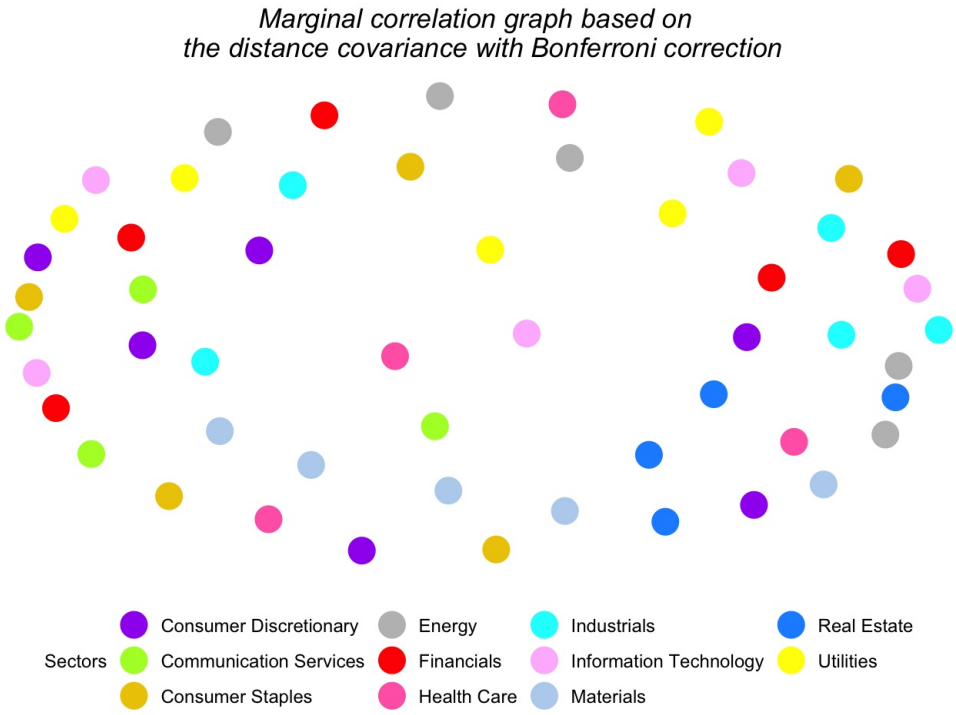
Also in this case we still expect similar results for the marginal correlation graph (connected graph)…

```
graph_test(log2, pMatrix2)
```



…and for the **Bonferroni Correction** (no links).

```
graph_test_bonf(log2, pMatrix2, 1431)
```

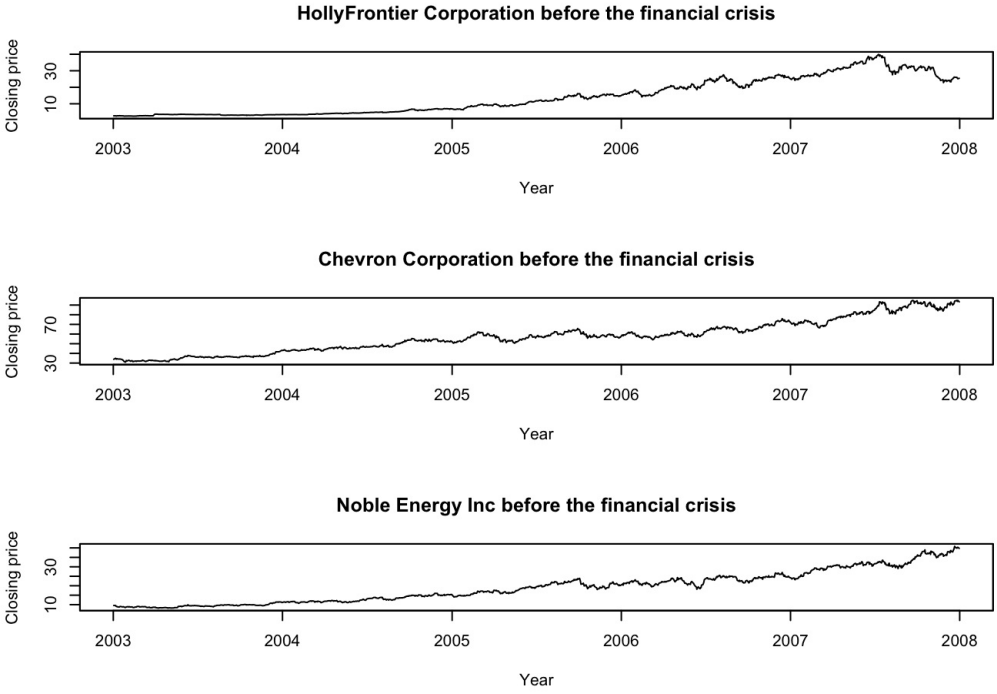


Conclusion

As we has seen above, the main change after the crisis is given by the increase of the clustering among different fields, infact companies of the same area seems to be more linked.

One of the exeptions is the company HollyFrontier Corp. that is separated from his sector, we can look into the trend before and after the financial crisis comparing it with another companies of the same sector which is linked with in 2003 - 2008:

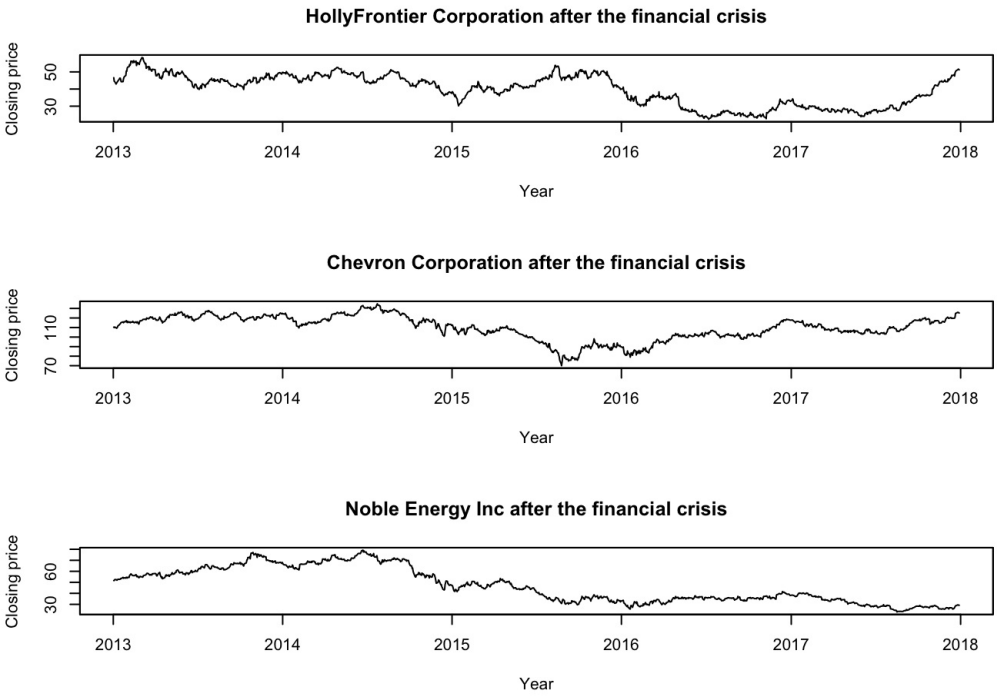
```
par(mfrow=c(3,1))
plot(HFC_before, main = "HollyFrontier Corporation before the financial crisis", ylab = "Closing price", xlab = "Y
ear")
plot(CVX_before, main = "Chevron Corporation before the financial crisis", ylab = "Closing price", xlab = "Year")
plot(NBL_before, main = "Noble Energy Inc before the financial crisis", ylab = "Closing price", xlab = "Year")
```



In the period 2003 - 2008 the three companies have the same trend and for this reason they are clustered together.

Now let's see what happens in the period 2013 - 2018:

```
par(mfrow=c(3,1))
plot(HFC_after, main = "HollyFrontier Corporation after the financial crisis", ylab = "Closing price", xlab = "Yea
r")
plot(CVX_after, main = "Chevron Corporation after the financial crisis", ylab = "Closing price", xlab = "Year")
plot(NBL_after, main = "Noble Energy Inc after the financial crisis", ylab = "Closing price", xlab = "Year")
```



As we can see the trend for HollyFrontier Corporation has a decrease around the end of 2016 and the beginning of 2017, instead for the other two companies have already a decrease from the end of 2014 (reason why they are still clusterd).

In conclusion the analysis can obviosly be quite different according to the companies sampled.

Collaborations: *Alice Schirinà & Eleonora Barocco, Daniele Sanna & Giorgio Zannini Quirini*

N.B. *The following HTML can be also run in order to have a dynamic output on the "graph" function defined above. If you want to see how the graph change as ϵ varies using a slider on the plot, you just need to remove the comment hashtag from* `(1), (2), (3)` *and from "runtime = shiny" at the beggining or the Rmd file.*