Stat4DS / Homework 02

Pierpaolo Brutti

Due Friday, December 07, 2018, 23:00 PM on Moodle

General Instructions

I expect you to upload your solutions on Moodle as a **single running** `R Markdown` file (`.rmd`) + its `html` output, named with your surnames.

You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Your responses must be supported by both textual explanations and the code you generate to produce your results. *Just examining your various objects in the "Environment" section of RStudio is insufficient – you must use scripted commands and functions.*

R Markdown Test

To be sure that everything is working fine, start `RStudio` and create an empty project called `HW1`. Now open a new `R Markdown` file (`File > New File > R Markdown...`); set the output to `HTML mode`, press `OK` and then click on `Knit HTML`. This should produce a web page with the knitting procedure executing the default code blocks. You can now start editing this file to produce your homework submission.

Please Notice

- For more info on `R Markdown`, check the support webpage that explains the main steps and ingredients: R Markdown from RStudio.
- For more info on how to write math formulas in LaTex: Wikibooks.
- Remember our **policy on collaboration**: *Collaboration on homework assignments with fellow students is **encouraged**. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had discussions concerning your solution. You may **not**, however, share written work or code after discussing a problem with others. The solutions should be written by **you**.*

Stock, Dependency and Graphs

Our question: study the dependency among stocks via marginal correlation graphs

We want to study the dependency among some standard measure of stock *relative performance* – see **`Appendix (B)`** for more info. To this end, we may collect the *daily closing prices* for D stocks[1], selected within those consistently in the S&P500 index from January 1, 2003 through today.

The stocks are categorized into 10 *Global Industry Classification Standard* (GICS) sectors, including `Consumer Discretionary`, `Energy`, `Financials`, `Consumer Staples`, `Telecommunications Services`, `Health Care`, `Industrials`, `Information Technology`, `Materials`, and `Utilities`. It is expected that stocks from the same GICS sectors should tend to be clustered together, since stocks from the same GICS sector tend to interact more with each other. This is the hypothesis we'd like to verify. So, ideally, we want to collect something like D/10 stocks for each GICS (or a relevant subset of GICS).

Each data point will correspond to the vector of closing prices on a trading day. More specifically, with $c_{t,j}$ denoting the *closing price* of stock j on day t, we consider the variables $x_{t,j} = \log(c_{t,j}/c_{t-1,j})$ and we want to build correlation graphs over the stock indices j (i.e. each node is a stock). In other words, we simply treat the instances $\{x_{t,j}\}_t$ as independent replicates, even though they form a time series.

Your job:

1. Take a look at basic tools to deal with graphs in `R` such as the `igraph` and `ggnet` packages.
2. Select a sensible portfolio of stocks and take data from January 1, 2003 through January 1, 2008, before the onset of the "financial crisis". Build the data matrix $\mathbb{X} = \left[x_{t,j}\right]_{t,j}$.

[1] The number of stocks D will affect the computational time and the statistical performance, hence, choose it wisely!

3. With this data, consider the usual Pearson correlation coefficient between stocks, and implement the **bootstrap procedure** described at page 3 of our notes to build *marginal correlation graphs*. In particular, visualize the dynamic of the graph as ϵ varies, highlighting the GICS sectors with annotation/node color. Draw some conclusion: is there any statistical evidence to support the claim that stocks from the same sector cluster together? Explain.
4. Again with the data in $\mathbb{X}$, we now want to build a *marginal correlation graph* based on γ^2, the *distance covariance* (see the Appendix in our notes). This time we don't have a confidence interval available, hence we will simply go for a multiple hypothesis testing (with and without Bonferroni correction) placing an edge $\{j, k\}$ between stock j and stock k if and only if we reject the null hypothesis that $\gamma^2_{i,j} = 0$. Use the functions in the package `energy` to perform these tests, then build and visualize the graph commenting on the results.
5. Finally, if possible using the same portfolio of stocks, grab data from January 1, 2013 through January 1, 2018. Build the new matrix $\mathbb{Y} = \left[y_{t,j}\right]_{t,j}$ and repeat the previous analysis commenting on observed differences between the two time-frames.

Appendix (A) - Financial data & R

As you can imagine, we can pull financial data into R in many different ways. For a general overview of the available tools, there are always the two `Task View` on Empirical Finance and Web Technologies and Services: on the latter, scroll down till you find a section entitled **Finance**.

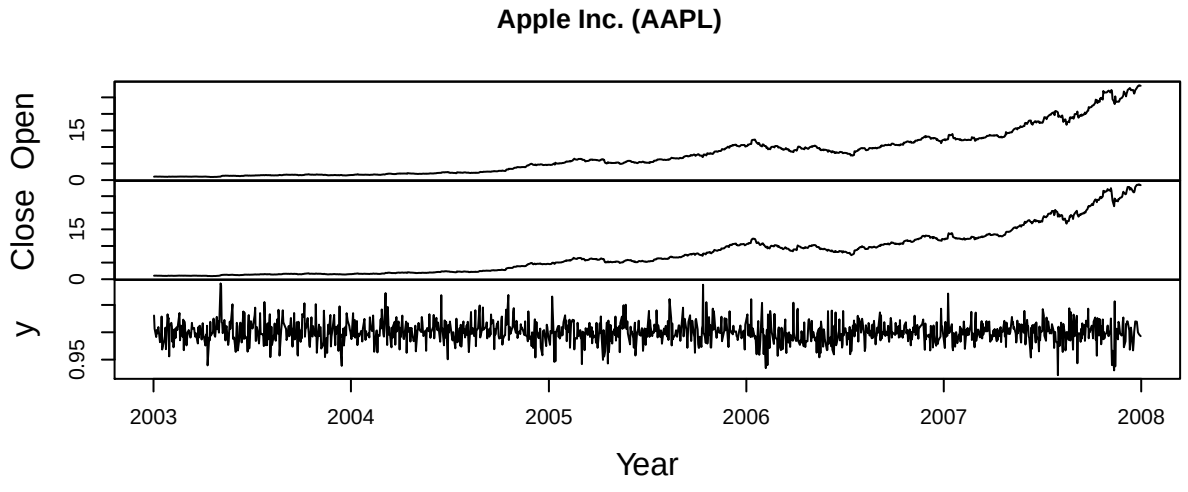
For specific suggestions I tried in the past, I would say:

- the `get.hist.quote()` function in the tseries package, or
- the `getSymbols()` function in the quantmod package – more info – or
- the `yahooImport()` function in the fImport package.

For general time series handling, I would suggest to explore a little bit the zoo package.

To get the stocks info you need, you have to know (in advance) the symbol associated to that stock (e.g. `Apple Inc.`, `IBM`, etc.) in a particular market (e.g. `NYSE`, `NASDAQ`). You can easily obtain this from portals like Yahoo! Finance. For example:

```
# Load the package
require(tseries, quietly = TRUE)
?get.hist.quote
# Get Apple Inc. from NYSE
aapl <- suppressWarnings(
  get.hist.quote(instrument="AAPL", start="2003-01-01", end="2008-01-01",
                 quote= c("Open","Close"), provider="yahoo", drop=TRUE) )
# Take a look
class(aapl); names(aapl); head(aapl)
# Build some relative performance measure like: Close/Open
aapl$y <- aapl$Close/aapl$Open
# Plot
plot(aapl, main = "Apple Inc. (AAPL)", xlab = "Year")
```



```
## time series starts 2003-01-02
## time series ends   2007-12-31
## [1] "zoo"
## [1] "Open"  "Close"
##                Open    Close
## 2003-01-02 1.025714 1.057143
## 2003-01-03 1.057143 1.064286
## 2003-01-06 1.073571 1.064286
## 2003-01-07 1.056429 1.060714
## 2003-01-08 1.041429 1.039286
## 2003-01-09 1.044286 1.048571
```

Appendix (B) - Returns & Price relatives of a stock.

Let's start from the basics: what is a **return**? The goal of investing is, of course, to make a profit. The revenue from investing, or the loss in the case of a negative revenue, depends upon both the change in prices and the amounts of the assets/stocks being held. Investors are interested in revenues that are high relative to the size of the initial investments. Returns measure this, because returns on an asset, e.g., a stock, a bond, a portfolio of stocks and bonds, are changes in price expressed as a fraction of the initial price.

There are many variants of the definition of returns, according to whether we allow some extra parameters to be considered in their calculation, like dividends or costs of transactions. Here we just look at the most simple definition of returns where only the price is considered.

- **Simple return**: Let P_t be the price of an asset at time t. Given a time scale τ, the τ–period simple return at time t, $R_t(\tau)$ is the rate of change in the price obtained from holding the asset from time $t-\tau$ to time t:

$$R_t(\tau) = \frac{P_t - P_{t-\tau}}{P_{t-\tau}} = \frac{P_t}{P_{t-\tau}} - 1.$$

  The τ–period *simple gross return* at time t is $R_t(\tau)+1$. If $\tau = 1$ we have a 1–period simple return (respectively, a *simple gross return*), and denote it R_t (resp., $R_t + 1$). Notice that if $\{P_t\}_t$ represents the series of *closing price* of a stock, the 1–period simple gross return gives essentially the price relatives used in `Borodin et al.(2004)` and mentioned in the very first section of this appendix.

There is a practical reason for defining returns backwards, and it is that more often than not we want to know the return obtained today for an asset bought some time in the past. Note that return values range from -1 to ∞; so, in principle, you can not lose more than what you have invested, but you can have unlimited profits.

Notice also that the τ–period *simple gross return* at time t equals the product of τ one–period simple gross returns at times $t-\tau+1$ to t; that is,

$$R_t(\tau)+1 = \frac{P_t}{P_{t-\tau}} = \frac{P_t}{P_{t-1}} \cdot \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-\tau+1}}{P_{t-\tau}} = (R_t+1)\cdot(R_{t-1}+1)\cdots(R_{t-\tau+1}+1).$$

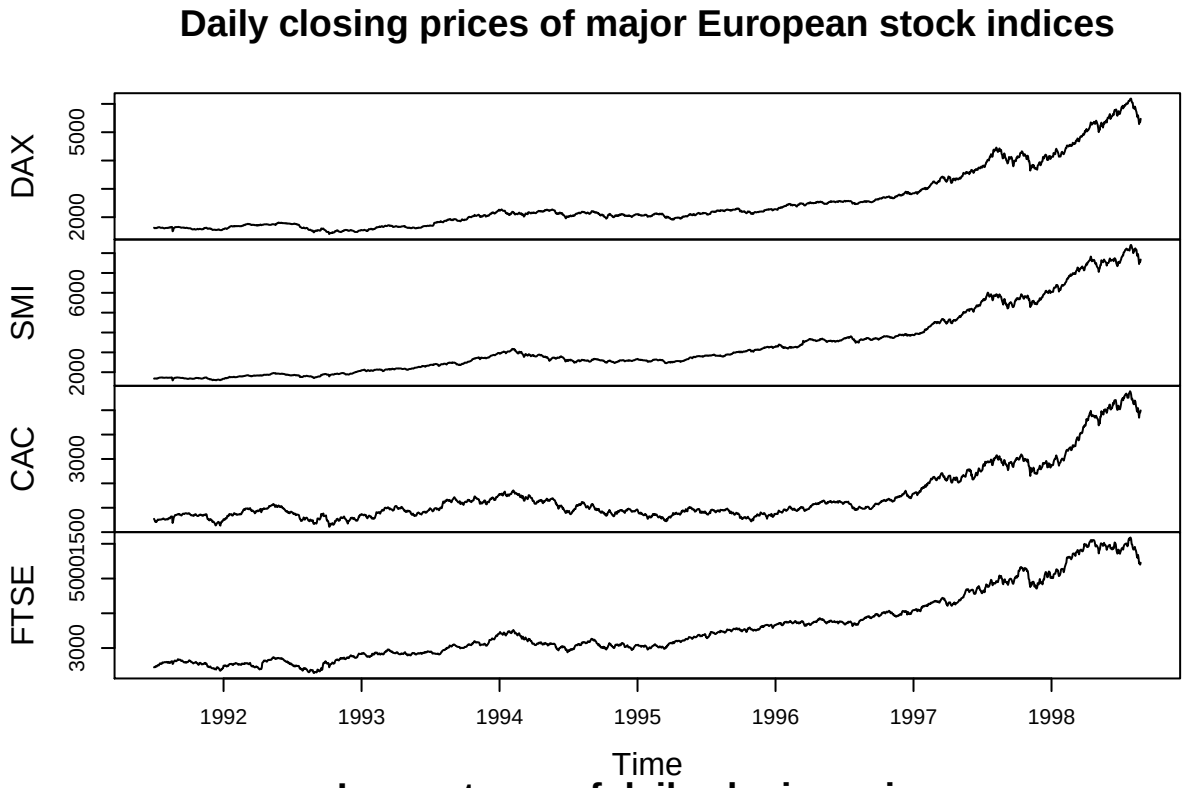
For this reason these *multiperiod* returns are known also as **compounded returns**. Returns are independent from the magnitude of the price, but they depend on the time period τ. For this reason one must always add to the numerical information the time span considered, if daily, weekly, monthly and so on. An example:

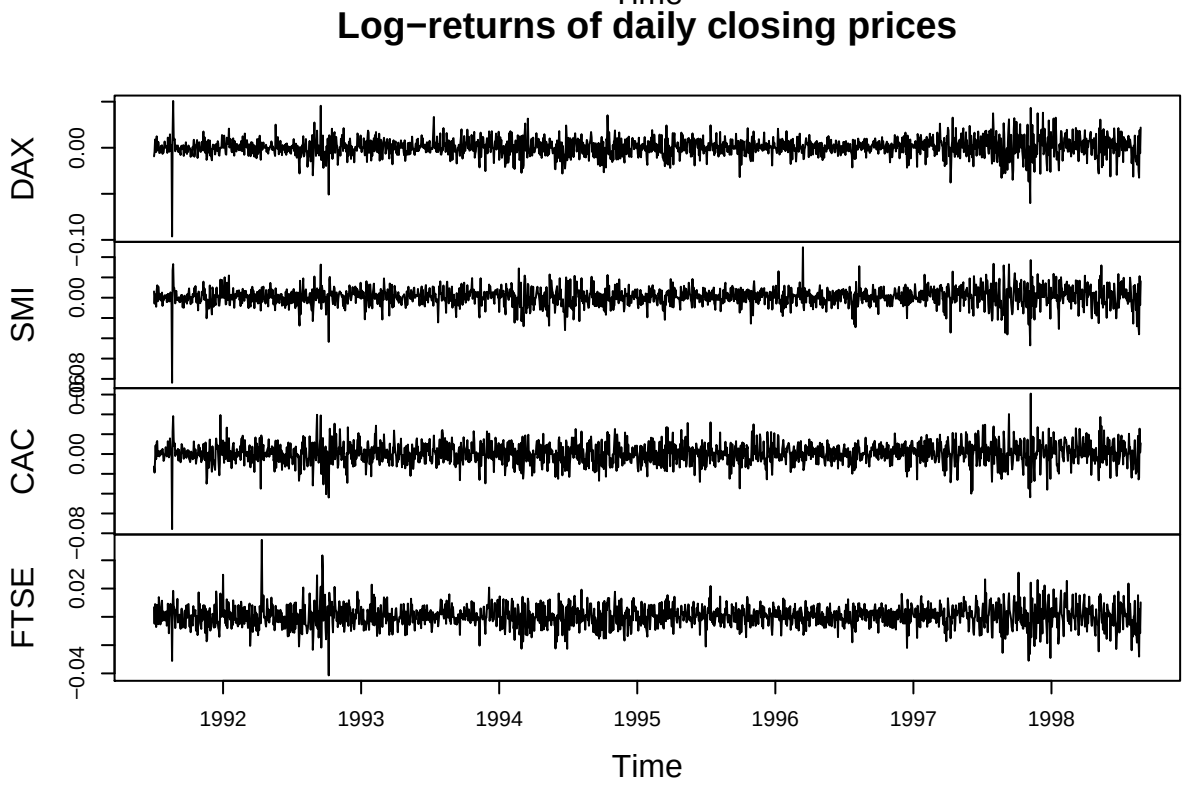
```
# Daily closing prices of major European stock indices
data(EuStockMarkets)
?EuStockMarkets

# mode(EuStockMarkets)
# class(EuStockMarkets)

# Daily closing prices
plot(EuStockMarkets, main = "Daily closing prices of major European stock indices")

# Log-returns
logR = diff(log(EuStockMarkets))
plot(logR, main = "Log-returns of daily closing prices")
```





Finally, one of the most important features of financial assets, and possibly the most relevant for professional investors, is the asset **volatility**. *In practice volatility refers to a degree of fluctuation of the asset returns.* However it is **not** something that can be directly observed. One can observe the return of a stock every day, by comparing the change of price from the previous to the current day, but one can not observe how the return fluctuates in a specific day. We need to make further observations of returns (and of the price) at different times on the same day to make an estimate of the way returns vary daily (so that we can talk about daily volatility), but these might not be sufficient to know precisely how returns will fluctuate. Therefore volatility can not be observed but estimated from some model of the asset returns. A general perspective, useful as a framework for volatility models, is to consider volatility as the **conditional standard deviation** of the asset returns. But better if we stop here!

Price relatives

Now, consider a portfolio containing D stocks quoted on some market, NYSE say. Each trading day $t \in \{1, \dots, T\}$ we observe the opening price $o_{j,t}$ and the closing[2] price $c_{j,t}$ of each stock $j \in \{1, \dots, \mathrm{D}\}$.

To evaluate the performance of a stock, it is more convenient to work with some sort of **relative price** that, broadly speaking, should represent the *factor* by which the wealth/money invested in the $j^{\texttt{th}}$ stock increases during the $t^{\texttt{th}}$ period. In the literature we see different options. Here's some examples:

1. In `Borodin et al.(2004)`, the Authors consider (possibly on a log-scale)

$$x_{t,j} = \frac{c_{t,j}}{c_{t-1,j}},$$

   so that an investment of $d\$$ in the $j^{\texttt{th}}$ stock just **before** the $t^{\texttt{th}}$ day yields $(d \cdot x_{j,t})$ dollars. But notice that there is nothing special about "daily closing prices" and the problem can be defined with respect to **any** (sub)sequence of the (intra-day) sequence of all price offers which appear in the stock market. In fact...

2. ...in `Cover (1991)`, they consider the ratio between opening and closing prices

$$x_{t,j} = \frac{c_{t,j}}{o_{t,j}},$$

3. ...but *today* closing price does not necessarily match *tomorrow* opening price. For this reason in `Helmbold et al.(1998)`, they use

$$x_{t,j} = \frac{o_{t+1,j}}{o_{t,j}},$$

   so that moving from one morning to the next, the value of a stock increases or falls to $x_{t,j}$ times its previous value.

[2]These are usually **adjusted closing prices**. The price that is quoted at the end of the trading day is the price of the last lot of stock that was traded for the day. This is called a stock's **closing price** and can be used by investors to compare a stock's performance over a period of time (usually from one trading day to another). During the course of a trading day, many things can happen to affect a stock's price: good and bad news relating to the operations of a company, any sort of *distribution* that is made to investors such as *cash dividends*, *stock dividends* and *stock splits*. When any these things happen, the closing prince has to be appropriately **adjusted**. Fortunately, historical price services provided by financial sites such as Yahoo! Finance eliminate the confusion by directly calculating **adjusted closing prices** for investors.