Stat4DS Homework 03 – Due Friday, December 23, 2016, 06:00 PM on Moodle

General Instructions

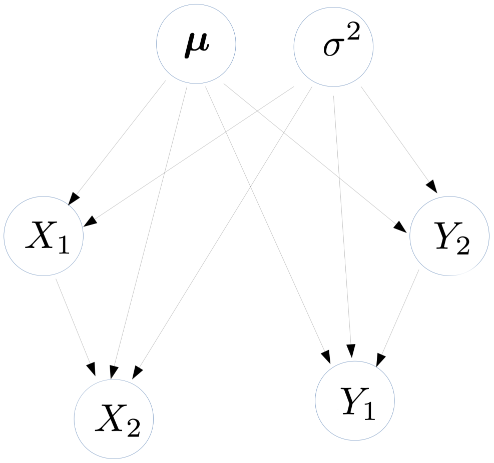
For the exercises involving `R`, I expect you to upload a **running** `R Markdown` file (`.rmd`), named with your group ID, to Moodle. You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Your responses must be supported by both textual explanations and the code you generate to produce your results.

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.

Exercise I: Generate samples from a DAG

Consider the following (Bayesian) model in DAG format



where, conditionally on $\boldsymbol{\mu} = [\mu_1, \mu_2]$ and σ^2, we assume that

- $\boldsymbol{X} = [X_1, X_2]^\intercal \sim \mathrm{N}_2(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_2)$
- $\boldsymbol{Y} = [Y_1, Y_2]^\intercal \sim \mathrm{N}_2(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_2)$

1. Use what we studied about conditional distributions of multivariate Normal vectors to write down in formula the joint distribution corresponding to this DAG.
2. Choose two suitable distributions for $\boldsymbol{\mu}$ and σ^2 and write the corresponding R code to simulate a sample of $n = 10000$ random vectors from the joint distribution.
3. Show a suitable plot with the empirical distribution of each component of the $\mathbf{X}$ vector. What theoretical distribution are you approximately describing?

Exercise II: How to estimate a population mean?

Given $\{X_1, \ldots, X_n\}$ IID from some distribution $F_X(\cdot)$, in this exercise we consider a seemingly trivial goal: estimate the population mean $\mu = \mathbb{E}(X)$.

An obvious choice would be the plug-in estimator, the *empirical mean*

$$\widehat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This estimator is computationally attractive, requires no prior knowledge and automatically scale with the population variance σ. In addition, tweaking a bit the *Central Limit Theorem*, we also know that

$$\lim_{n\to+\infty}\mathbb{P}\left(\frac{\sqrt{n}\,\left|\bar{X}_n-\mu\right|}{\sigma}\leqslant\sqrt{2\,\log\left(\tfrac{2}{\alpha}\right)}\right)=\lim_{n\to+\infty}\mathbb{P}\left(\left|\bar{X}_n-\mu\right|\leqslant\sigma\sqrt{\tfrac{2}{n}\,\log\left(\tfrac{2}{\alpha}\right)}\right)\geqslant 1-\alpha,$$

result that also holds *non-asymptotically* under some suitable technical conditions. If these conditions are not met, we still have Chebyshev's inequality, which says that with probability at least $1-\alpha$

$$\left|\bar{X}_n-\mu\right|\leqslant\sigma\sqrt{\frac{2}{n\,\alpha}},$$

an exponentially weaker bound that will especially hurt in modern applications where many means have to be estimated simultaneusly (e.g. empirical risk minimization methods).

Our question: can we do better?

An interesting alternative estimator is the **median-of-means** estimator. So assume that we chop the original n observations in k independent blocks of size N (approximately), then define

$$\widehat{\mu}_{\texttt{MM}}(N)=\{\text{median of the } k \text{ block-means}\}=\text{median}\left\{\frac{1}{N}\sum_{i=1}^{N}X_i,\ldots,\frac{1}{N}\sum_{i=(k-1)N}^{kN}X_i\right\}.$$

This new estimator is in general biased but, if we carefully choose the block size N, then for any distribution with finite variance (and also in some infinite variance case) with probability at least $1-\alpha$ we have

$$\left|\widehat{\mu}_{\texttt{MM}}-\mu\right|\leqslant 8\,\sigma\sqrt{\tfrac{1}{n}\log\left(\tfrac{2}{\alpha}\right)},$$

an inequality exactly of the form we like. The theoretical optimal block size is then $N^{\star}=\frac{n}{8\log(\alpha^{-1})}$.

Your turn: explore the performance of $\widehat{\mu}_{\texttt{MM}}(N)$

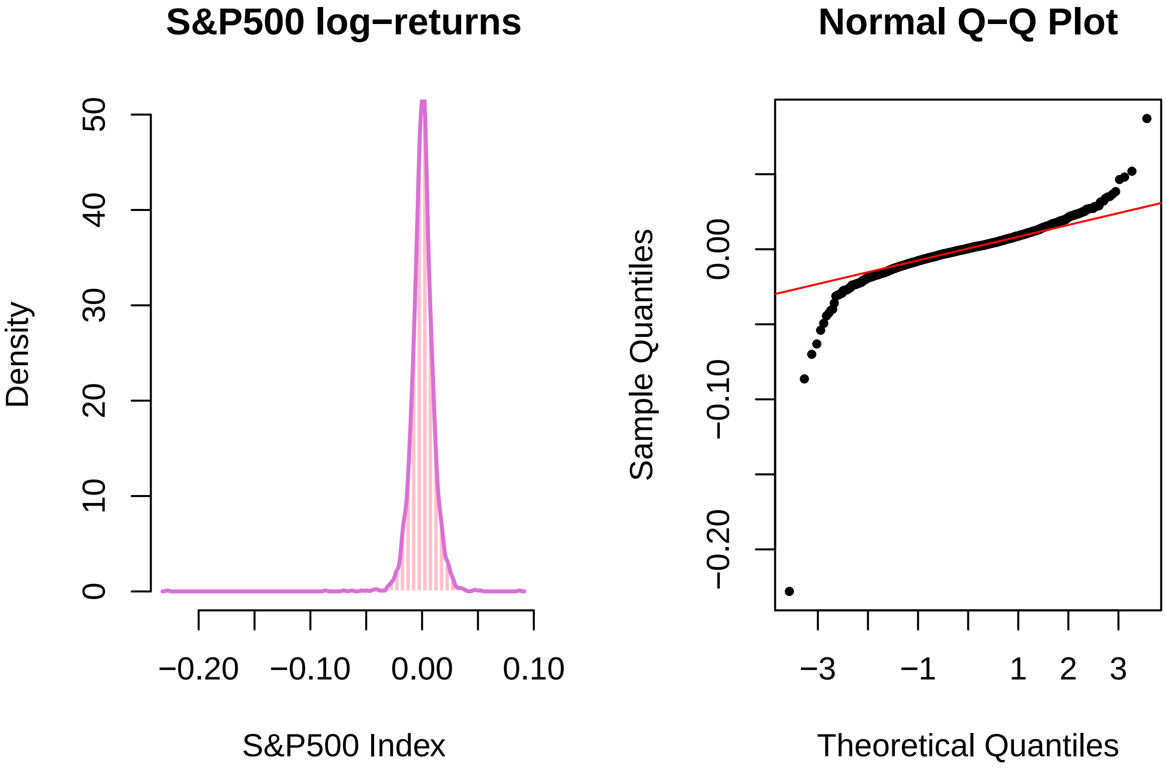
1. Setup a sensible simulation study to compare $\bar{X}_n$ and $\widehat{\mu}_{\texttt{MM}}(N)$ in terms of their bias, variance, MSE and tail probabilities; that is, the probability that $\left(\,|\,\bar{X}_n-\mu\,|>t\,\right)$ versus $\left(\,|\,\widehat{\mu}_{\texttt{MM}}(N)-\mu\,|\,>t\,\right)$ where μ is the true population mean and t is some prefixed threshold. You should carry out the simulation under at least two population models, one with light tails like the Gaussian and the other one with fatter tails like the Laplace or the Student–t with a few degrees of freedom, but you may also wanna try something more extreme, or discrete or skewed. For each scenario you should vary the sample size n (starting quite small), the variance σ and, for the *median-of-means*, the block size N.

2. Consider now the daily log-returns on the Standard & Poor's 500 index from January 1981 to April 1991 contained in the file `sp500.RData`. First of all take a look

```
# Load the data
load("sp500.RData")

# Now some diagnostic plot
```

```
par(mfrow = c(1,2))
hist(sp.ind, prob = T, col = "pink", border = "white", breaks = 50,
     xlab = "S&P500 Index", main = "S&P500 log-returns")
lines(density(sp.ind), col = "orchid", lwd = 2)
# Compare the empirical quantiles with the normal quantiles at the
# same level: we see heavy tails and a bit of skewness
qqnorm(sp.ind, pch = 19, cex = .5)
qqline(sp.ind, col = "red")
```



Now we want to estimate the mean using the median-of-means but we need to choose the block size N. To this end, consider a sequence of candidate values $N_{\texttt{seq}} = \{N_1, N_2, \ldots, N_\ell\}$. For each one of them use the *nonparametric bootstrap* to estimate the MSE of $\widehat{\mu}_{\text{MM}}(N)$. Pick the block size with associated the smallest (estimated) MSE.

Exercise III: 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], possibly selected within those consistently in the S&P500 index between January 1, 2003 through January 1, 2008, before the onset of the "financial crisis".

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: Select a sensible portfolio of stocks, build the data matrix $\mathbb{X} = \left[x_{t,j}\right]_{t,j}$ and visualize what's going on in there. With this data, choose a suitable association measure and implement the **bootstrap procedure** described at page 3 and 4 of our notes to build a **marginal correlation graph**. Visualize the graph using the GICS sectors to annotate/color the nodes, and draw some conclusion: is there any statistical evidence to support the claim that stocks from the same sector cluster together?

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