CHALMERS | UNIVERSITY OF GOTHENBURG

Financial Time Series TMS088/MSA410 – LP4 2020/21

Project 1: Missing trading volume data

Theoretical part

1. (13 points) Let $\mu \in \mathbb{R}, \sigma^2 > 0$ and let $Z \sim \mathrm{WN}(\mu, \sigma^2)$. Let then Y be the process defined by

$$Y_t = \sum_{j=0}^{q} \theta_j Z_{t-j}, \tag{1}$$

for some coefficients $\theta_1, \dots, \theta_q \in \mathbb{R}$, $q \in \mathbb{N}$, with $\theta_0 = 1$. As we will see in the lectures, Y is called a moving average process of order q.

a) (2 points) Show that for any $(t,h) \in \mathbb{Z}^2$, $\mathbb{E}(Y_t) = \mathbb{E}(Y_{t+h})$.

b) (4 points) Show that $(t,s,h) \in \mathbb{Z}^3$, $\mathsf{Cov}(Y_t, Y_{t+h}) = \mathsf{Cov}(Y_s, Y_{s+h})$.

c) (2 points) Show that Y is stationary and give its autocovariance function.

d) (5 points) Prove that if Z is a Gaussian process, then Y_t is independent of Y_{t+h} for any $t \in \mathbb{Z}$ and $|h| > q$.
Hint: Start by showing that Y is also a Gaussian process, then write the joint probability function of (Y_t, Y_{t+h}).

Remember to carefully motivate each step in your calculations.

Practical part

The file *intel.csv* contains daily trading volume data for the Intel Corporation stock at Nasdaq[1] between March 19th, 2001, and November 30th, 2020, with $N = 4958$ data points. The variable *Volume* contains the original volume trading. The variable *VolumeMissing* is a copy of *Volume* but has 100 data points missing, replaced by NaN entries. Your task is to work with the differenced log time series $Y = (Y_t, t = 1, \dots, N-1)$, given by

$$Y_t := \log X_{t+1} - \log X_t,$$

[1] Source: https://finance.yahoo.com/

Please turn!

where $(X_t, t = 1, \ldots, N)$ are the data points in *VolumeMissing*, and reconstruct the missing points using the theory of linear time series. Note that Y will have more missing values than X since, for a given $t \in \{1, \ldots, N-1\}$, Y_t is missing if either X_{t+1} or X_t is missing. You are to work with this data set using either MATLAB or R. To import the contents of the csv-file, make sure your working directory contains the file and use the `readtable` function in MATLAB, or the `read.csv` function in R.

2. (2 points) Compute the time series Y and plot the sample autocorrelation function (ACF) $\hat{\rho}_Y(h)$ for $h = 0, \ldots, 20$. As we will see in the lectures, the ACF is consistent with Y being a moving average process of order q, as defined in (1) (let us assume here and below that $\mu = 0$). Such a process has the feature that the sample ACF values $\hat{\rho}_Y(h)$ are approximately IID $\mathcal{N}(0, N^{-1})$ for $h > q$. Based on this fact, choose a reasonable value of q based on your computed ACF $\hat{\rho}_Y$, assuming that the data is a realization of (1).

3. (8 points) Let $\mathbb{M}$ be the set of indices of the missing values of Y. Use Corollary 2.4.6 in the lecture notes to write a program that, for each $t \in \mathbb{M}$, calculates $b_t^\ell(Y^q)$ (but use your computed sample ACF $\hat{\rho}$ instead of the theoretical ACF ρ). Here $Y^q := (Y_s, \max(1, t-q) \leq s \leq \min(N, t+q)$ and $s \notin \mathbb{M})$.

4. (2 points) Now compute the differenced series using all available data, i.e., compute $\mathcal{Y} = (\mathcal{Y}_t, t = 1, \ldots, N-1)$, given by

$$\mathcal{Y}_t := \log \mathcal{X}_{t+1} - \log \mathcal{X}_t,$$

where $(\mathcal{X}_t, t = 1, \ldots, N)$ are the data points in *Volume*. Let $\hat{Y}$ denote the modified process consisting of Y where each missing data point Y_t has been replaced by $b_t^\ell(Y^q)$, computed as in the previous task. Let $\check{Y}$ denote the modified process consisting of Y where each missing data point Y_t has been replaced by a value calculated by simple linear interpolation. Calculate and report the root mean squared errors for the two series: $\sqrt{M^{-1}\sum_{t\in\mathbb{M}}(\mathcal{Y}_t - \hat{Y}_t)^2}$ and $\sqrt{M^{-1}\sum_{t\in\mathbb{M}}(\mathcal{Y}_t - \check{Y}_t)^2}$, where $M = |\mathbb{M}|$, the number of indices in $\mathbb{M}$.

See next page!

Some useful MATLAB functions in no particular order:

- *autocorr* - Computes the sample ACF.
- *find* - Finds indices of nonzero elements in a vector.
- *isnan* - Returns 1 if the input is NaN, otherwise 0.
- *fillmissing* - Fills in missing values in a vector with several interpolation options available.

Deadline: April 23, 2021.
Requirement: You must do this project in MATLAB or R. For this project there are 25 points available. To qualify for bonus points you need to score at least 10 points. After that, every 2.5 points you score on this project will translate into 0.5 bonus points on the exam.
Formalities: You may work alone but you are strongly encouraged to work in pairs. Write your project report as a single pdf document, preferably in using LaTeX, MATLAB's LiveEditor or R's Rmarkdown. It should include all plots, explanations, and answers to the questions as well as your implemented (and *commented*) code. If you do not write your report in LaTeX, it is acceptable to scan your handwritten solutions to the theoretical parts of the project and include them in the pdf file. Upload this report in Canvas. Reports without code will not be graded and the code should be structured and include comments that make it readable. Your report will be subject to a plagiarism review with Urkund.