CHALMERS | UNIVERSITY OF GOTHENBURG

Financial Time Series TMS088/MSA410 – LP4 2021/22

Project 1

Theoretical part: differencing time series

1. (8 points) Assume that the stochastic process $X = (X_t, t \in \mathbb{Z})$ is given by

$$X_t := m_t + Y_t,$$

where Y is stationary with $\mathbb{E}[Y_t] = 0$ for all $t \in \mathbb{Z}$ and the trend m is given by the polynomial

$$m_t := \sum_{j=0}^{q} a_j t^j$$

for some coefficients $a_j \in \mathbb{R}$, $j = 0, \ldots, q$.

a) (3 points) Let the differencing operator be given by $\nabla f_t = f_t - f_{t-1} = (1-B)f_t$, where B is the backward shift operator. Show that $\nabla^q m_t = q!\, a_q$. (*Hint: Start with $q = 1$ and ∇m_t, proceed by induction.*)

b) (1 point) Show that $\mathbb{E}[\nabla^q Y_t]$ is constant in t.

c) (1 point) Show that $\mathrm{Var}[\nabla^q Y_t] < +\infty$ for all $t \in \mathbb{Z}$.

d) (2 points) Show that $\gamma_{\nabla^q Y}(r, s) := \mathsf{Cov}(\nabla^q Y_r, \nabla^q Y_s) = \gamma_{\nabla^q Y}(r + h, s + h)$ for every $r, s, h \in \mathbb{Z}$.

e) (1 point) Conclude that $\nabla^q X$ is a stationary process with mean $q!\, a_q$.

Remember to carefully motivate each step in your calculations.

Please turn!

Practical part: Interest rates and autoregressive processes

Background

A statistical model for interest rates is the Vasicek model, which is given by the *stochastic differential equation* (SDE)

$$\mathrm{d}r_t = a(b - r_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{1}$$

where r_t is the interest rate, with initial value r_0, $a, \sigma > 0$, $b \in \mathbb{R}$ and W_t is the random market risk, modeled by a *Brownian motion*. The *drift term* $a(b - r_t)$ is the expected change in the interest rate, and σ is the *volatility*, which influences how randomly r behaves. The key feature of the model is that it exhibits *mean reversion*, meaning that the interest rate will move back to the *long-term mean* b, with the reversion speed governed by a.

The solution of an SDE can be seen as a time-continuous time series model, and in fact, a discrete version of Equation (1) is given by

$$R_t = ab + (1 - a)R_{t-1} + Z_t = ab + \varphi R_{t-1} + Z_t, \tag{2}$$

where Z is normally distributed IID noise with mean 0 and variance σ^2. Note that this is an autoregressive process of order 1 with mean b and parameter $\varphi := (1 - a)$. The goal of this project is to use Equation (2) to model and predict interest rates.

The file *interest.csv* contains monthly interest rate data from the US, between July 1983 and August 2008[1], with $N = 302$ data points. 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.

Exercises

2. (2 points) Load the data and calculate its sample mean $\hat{\mu}$. Calculate the mean-corrected time series

$$S_t := R_t - \hat{\mu},$$

[1]OECD (2022), Short-term interest rates (indicator). DOI: `http://dx.doi.org/10.1787/2cc37d77-en` (Accessed on 11 March 2022)

See next page!

where $(R_t, t = 1, \ldots, N)$ are the data points in *interest.csv*. Plot the sample autocorrelation function (ACF) $\hat{\rho}_S(h)$ for $h = 0, \ldots, 20$ as well as the sample partial autocorrelation function (PACF) $\hat{\alpha}_S(h)$ for the same values of h. As seen in the lectures, an autoregressive process of order p exhibits a slowly decreasing ACF whereas the sample PACF should be significantly different from 0 for all lags $0 \leq h \leq p$ and negligible if $h > p$. We have already assumed that $p = 1$, but is this reasonable given the plots? If not, what would you suggest instead?

3. (5 points) Assuming that the time series model is given by Equation (2), we note that after mean-correction, the relation

$$S_t = \varphi S_{t-1} + Z_t \tag{3}$$

should hold for some parameters φ and σ. This is a linear model, where S_t is called the *dependent variable* and S_{t-1} the *independent variable* in this context. The goal of this task is to estimate φ and σ with linear regression.

a) (1 point) Create a data set containing the pairs $((S_{i-1}, S_i))_{i=2}^N$. Plot the pairs in a scatter plot Does a linear relationship look reasonable?

b) (2 points) Using the data set, estimate φ and σ using linear regression without intercept. Report the values. Simulate and plot a few sample paths of the estimated model together with the mean-corrected data. (*Hint: Read up on linear regression without intercept at for instance Wikipedia.*[2])

c) (2 points) Calculate and plot the residuals $E_i := S_{i+1} - \hat{\varphi} S_i$, where $\hat{\varphi}$ is the estimated value of φ. As $S_{i+1} - \varphi S_i = Z_i$ is assumed to be an IID noise, E_i should also behave as IID noise. Does it seem reasonable given the plot of the residuals? Plot the ACF and perform a Ljung–Box test on the residuals to test whether or not this seems to be the case.

4. (7 points) In this exercise we compute predictions for one month in the future given the values of the previous 20 months. First, divide the data set into a *training set*, containing the first 201 data points and a *test set* consisting of the remaining 101 data points. Next, compute the sample ACVF $\hat{\gamma}$ ***using the data from the training set only***. Then compute all linear forecasts

$$b_n^l(z_{n-1}, \ldots, z_{n-20}) = a_1 z_{n-1} + a_2 z_{n-2} + \ldots + a_{20} z_{n-20}$$

for $n = 202, 203, \ldots, 302$ by solving the equations in Proposition 2.4.5 in the lecture notes, replacing the exact ACVF γ with the sample ACVF $\hat{\gamma}$ you computed from the

[2] `https://en.wikipedia.org/wiki/Simple_linear_regression`

Please turn!

training set. Evaluate the performance of your predictions by plotting them in the same figure as the test data and by computing the error

$$\frac{1}{101}\sum_{n=202}^{302}\left(b_n^l(z_{n-1},\ldots,z_{n-20})-z_n\right)^2.$$

Compare this to the error you get from the naive 'prediction' of just using the mean (which is zero since you have a mean corrected series), i.e., compute the error

$$\frac{1}{101}\sum_{n=202}^{302}\left(\hat{\mu}-z_n\right)^2=\frac{1}{101}\sum_{n=202}^{302}z_n^2.$$

Can we conclude that the linear forecast is better than the naive prediction? Please note that there is no parametric assumption in this exercise.

5. (3 points) We will now make the same predictions as in the previous exercise, but we use the parametric assumption that S follows an autoregressive process of order 1 with the parameters $\hat{\varphi}$ and $\hat{\sigma}$ that were estimated in the third exercise. Calculate the best linear predictor in this case, i.e., instead of using the sample ACVF $\hat{\gamma}$ take the ACVF of Equation (3) with the estimated parameters $\hat{\varphi}$ and $\hat{\sigma}$. Plot the predictions in the same figure as the test data and calculate the same errors as in the previous exercise. Have we improved the quality of the predictions? Try to explain why or why not. Is there a point in making the parametric assumption? (*Hint: Example 2.5.1 in the course book may be helpful!*)

See next page!

Some useful MATLAB functions in no particular order:

- *autocorr* - Computes the sample ACF.
- *parcorr* - Computes the sample PACF.
- *fitlm* - Fits a linear model.
- *randn* - Generates one normal random number with mean 0 or 1.
- *lbqtest* -Performs a Ljung–Box test.
- *toeplitz* -Calculates a Toeplitz matrix given a certain vector.

Some useful R functions in no particular order:

- *acf* - Computes the sample ACF.
- *pacf* - Computes the sample PACF.
- *lm* - Fits a linear model.
- *rnorm* - Generates one normal random number with mean 0 or 1.
- *box.test* -Performs a Ljung–Box test.
- *matrix* - Can be used to preallocate a matrix.
- *toeplitz* -Calculates a Toeplitz matrix given a certain vector.

Deadline: April 29 2022 at 23.59.
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 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 *readable* 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 *and understandable*. Your report will be subject to a plagiarism check. Please note that no form of plagiarism will be tolerated and that all work, including code, must be your own.