For this assignment, you may work in groups of two or three students. Part 1 is theoretical and part 2 is an application of your results from part 1. For each sub-question, you can earn three points. It is crucial to motivate your answers: No points will be awarded for answers without motivation. Furthermore, reports that are not typeset as a single PDF file in LaTeX will not be considered. Please also make sure to submit your R (and/or Python) codes for question 2, separately, together with a CSV file of the data. For question 2, your code should reproduce all reported results without returning errors. Results that are not returned by your code cannot be graded.

The deadline for this assignment is Monday, February 12, 15:00.

Part 1. For the next questions, consider an AR(p) model:

$$y_t = \phi_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t, \qquad t = 1, \dots T,$$

for some given initial values $y_0, \dots y_{-p+1}$ and $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$.

(a) Write the model in terms of $\mathbf{x}_t = (1, y_{t-1}, \dots, y_{t-p})'$ and a vector of coefficients $\boldsymbol{\phi} = (\phi_0, \dots, \phi_p)'$ and show that the OLS estimator for $\boldsymbol{\phi}$ can be written as:

$$\hat{\boldsymbol{\phi}}_{\text{OLS}} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t y_t$$

(b) Write the AR(p) model for $\mathbf{y} = (y_T, y_{T-1}, \dots, y_1)'$ in terms of the matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & y_{T-1} & \cdots & y_{T-p} \\ 1 & y_{T-2} & \cdots & y_{T-p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_0 & \cdots & y_{-p+1} \end{pmatrix}$$

and derive the OLS estimator in terms of $\mathbf{X}$ and $\mathbf{y}$. Is the OLS estimator unbiased in the current setting? Explain.

(c) Prove that

$$f(\mathbf{y}) = f(y_T | y_{T-1}, \dots, y_1) f(y_{T-1}, \dots, y_1) = f(y_T | y_{T-1}, \dots, y_1) \cdots f(y_2 | y_1) f(y_1).$$

(d) Show that

$$f(\mathbf{y}|\mathcal{F}_0) = \prod_{t=1}^{T} f(y_t|\mathcal{F}_{t-1}), \tag{1}$$

where $\mathcal{F}_t = \{y_s | s \leq t\}$ is the available information at time t, e.g. $\mathcal{F}_0 = \{y_{-p+1}, \ldots, y_0\}$. Use Eq. (1) to derive the Maximum Likelihood Estimator for $\boldsymbol{\phi}$.

(e) Now suppose that y_t itself plays the role of a noise term in the complete model for some asset return, that is,

$$r_t = \mathbf{z}_t'\boldsymbol{\beta} + y_t,$$

where $y_t \sim \text{AR}(p)$. Derive the FGLS estimator for $\boldsymbol{\beta} = \boldsymbol{\beta}(\hat{\boldsymbol{\phi}}_{\text{OLS}})$. Do you prefer this estimator for $\boldsymbol{\beta}$ over OLS in this case? Explain why.

Now, some questions about stationarity will follow.

(f) Consider the AR(4) model: $y_t = \phi_0 - 1.3y_{t-2} - 0.4y_{t-4} + \varepsilon_t$. Is this AR model stationary? Show why.

(g) The stationarity conditions of an AR(1) process follow naturally from its ACF. Explain.

(h) Using the Yule-Walker equations, find an analytical expression of the ACF ρ_k of the AR(2) process: $y_t = 0.6y_{t-1} - 0.25y_{t-2} + \varepsilon_t$. Fully work out the expression for the (real-valued) ACF. In particular, don't leave any complex numbers in your answer.

The h-step-ahead forecast is defined by: $\hat{y}_{n+h} = \mathbb{E}[y_{n+h}|Y_n]$.

(i) Give a general expression for the h-step-ahead forecast for a stationary and invertible ARMA$(1,1)$ model, in terms of known values at time n. Hint: Since ε_n can be written as a function of $y_n, y_{n-1}, \ldots$, it is contained in the information set Y_n.

Part 2. In this part, you are asked to download data yourself and make estimates and forecasts, assuming an AR(2) model for the log-returns. Unless mentioned otherwise, it is not allowed to use special packages for estimating and forecasting AR models, so you should write your own functions and code. It is useful to implement your results from part 1 to do this.

(a) Download daily data from the website of Yahoo Finance for the 5-Year U.S. Dollar Treasury Yield (ticker: ^FVX) from February 1, 2000, up to and including January 31, 2024. Save it as a CSV file and enable your script to read the data. Use the adjusted closing prices as our time series. Note that the data contains days without a value (non-trading days). Make sure you remove these empty entries.

(b) Calculate the daily log-returns and plot both the rates and the log-returns over time. Also make an autocorrelation plot of the log-returns. What can you say about the series' characteristics based on these plots (e.g. ACF, stationarity)?

The last 20 data points are not used for estimating, but for comparing to our forecasts later. Split your data points into a set containing these last 20 values and an estimation set, containing all other data points.

(c) Estimate the model using both the OLS and MLE methods. For MLE, you can use an optimization function defined in R. Interpret the results and compare the two estimation methods. You can use the first return as the initial value for the log-return and the sample variance over the whole estimation sample as the variance for the MLE. Furthermore, you may assume that the initial error term is equal to zero.

(d) Now, estimate the model using the "tseries" package in R (so now you are allowed to use the function defined in R to estimate the AR model). You can also use the"arima" function. Compare these results to the results from OLS and MLE. Also, plot the fitted returns and the actual returns in one figure. Comment on the quality of the fit.

(e) Check for normality of the fitted residuals.

(f) Use the estimated parameters from (d) to make forecasts for the log-returns up to 20 days forward. Note: for constructing the forecasts, you must write your own code again! Plot the forecasts and use them in your report. Also plot the actual returns and compare them to your forecasts.

