Time Series Analysis
Coursework

Handout: Monday 20 November 2023.
Deadline: **Submit electronic copy to the turnitin assignment on Blackboard before Friday 15 December 2023, 1pm.**

This coursework is worth 10% of your mark for Time Series Analysis.

There are 3 questions. **Together they are worth 40 marks**

Put your CID number clearly at the top of your report.

Plots and tables should be clear, well labelled and captioned. **Marks will be deducted for a poorly presented report.**

Comment your code. Marks will be deducted for uncommented code and *very* inefficient coding.

For the computational elements, you must use `R`, `MATLAB` or `Python`. It is up to you which you choose. All three are equally valid.

You may use any results from the notes you wish, but you must properly cite them.

You must type up your report (MS Word, LaTeXor a notebook (e.g. Jupyter) is fine), including all your code within the main text (not as appendices or screen shots).
YOU MUST SUBMIT A PDF ON TURNITIN.

Your report must be no more than 15 pages (not including a cover page). This does not mean you have to use all 15 pages - you should be able to do this in fewer than 15 pages.

The use of large language models is prohibited.

NB: Stationarity by itself always means 'second-order' stationarity. $\{\epsilon_t\}$ denotes white noise with mean zero and variance σ_ϵ^2. Assume $\{\epsilon_t\}$ is Gaussian/normal here. You can assume a sampling interval of $\Delta t = 1$ throughout.

Plagiarism, including the use of large language models, is a very serious offence that will be reported to registry. You must submit your own piece of work. You may be invited to an oral examination to demonstrate the authenticity of your work.

1 Question 1

This question relates to material learnt in week commencing 20th November.

(a) The pseudo-cyclical behaviour of an AR(2) process with complex conjugate roots, as discussed in lectures, extends to higher order AR processes. Write a function that simulates an AR(4) process, that oscillates with 2 dominant frequencies, f_1 and f_2. These frequencies should be inputs, along with r_1 and r_2, each in $(0, 1)$, that control the strength of the cyclical behaviours. The length N of the output time series, and σ_ϵ^2, the variance of the white noise process.

To simulate the process, use a burn in method that sets $X_{-3} = X_{-2} = X_{-1} = X_0 = 0$, and then discards the first 1000 values, to ensure an (approximately) stationary process.

In your report, you should include a short mathematical description of how you computed the parameters of your AR(4) process. **4 marks**

(b) Write a function `S_AR(f,phis,sigma2)` that computes the (theoretical) spectral density function for an AR(p) process.

The inputs should be:

`f`: the vector of frequencies at which it should be evaluated.

`phis`: the vector $[\phi_{1,p}, ..., \phi_{p,p}]$.

`sigma2`: a scalar for the variance σ_ϵ^2 of the white noise $\{\epsilon_t\}$.

Order p must not be an input, instead it needs to be computed from the length of `phis`. **2 marks**

(c) Using `fft`[1], write two functions.

- `periodogram(X)` that computes the periodogram at the Fourier frequencies for a time series `X`. You may also find it useful for the function to apply `fftshift` and output the relevant Fourier frequencies as a vector
- `direct(X,p)` that computes the direct spectral estimate at the Fourier frequencies using the $p \times 100\%$ cosine taper for a time series `X`.

3 marks

(d) In this question, we will explore how dynamic range effects bias in the spectral estimator.

You need to write a script that calls the functions you made in (a), (b) and (c). It should perform the following tasks:

A. Simulate 5,000 realizations, each of length $N = 64$, of an AR(4) process ($\sigma_\epsilon^2 = 1$) that shows pseudo-cyclical behaviour at $f_1 = 6/64$ and $f_2 = 26/64$ where $r_1 = r_2 = 0.8$. For each realization, compute the periodogram and four direct spectral estimates using a cosine taper with $p = 0.05$, $p = 0.1$, $p = 0.25$ and $p = 0.5$. Store the values for each at frequencies 6/64, 8/64, 16/64 and 26/64 (i.e. on the oscillating frequencies, at a frequency near one of the oscillating frequencies, and at a frequency far away from the oscillating frequencies).

B. Compute the *sample* percentage bias (using `S_AR`) of the periodogram and four direct spectral estimators at frequencies 6/64, 8/64, 16/64 and 26/64 from the 5,000 realizations. The percentage bias of an estimator $\hat{\theta}$ of θ is

$$100 \times \frac{\text{bias}(\hat{\theta})}{\theta}.$$

[1] The fft algorithm computes the Fourier transform at the Fourier frequencies $f_k = k/N$, $k = 0, ..., N-1$. This is the syntax for MATLAB and R. If using Python, you need to make use of `numpy.fft`

C. Repeat steps A and B for r_1 and r_2 equal to $0.81, 0.82, 0.83, ..., 0.99$.

D. Present four plots, one for each frequency, that compare the periodogram and the direct spectral estimators for different values of p, as a function r.

6 marks

(e) In no more than 100 words, comment on the results. **3 marks**

In Questions 2 and 3, you will be analysing a real time series. Your time series must be downloaded from this OneDrive folder. They are all taken from the National Oceanographic Centre's repository `https://psmsl.org/data/obtaining/`. Your time series is 10 years of monthly sea level gauge data from a particular buoy. Your number aligns with the ID on the NOC's repository. The first column is the time stamp (in years), and the second column is the gauge reading. I have selected time series segments that have no missing values.

You can assume the process is stationary.

Question 2

This question relates to material learnt in week commencing 20th November.

Estimate the spectral density function on your time series using a direct spectral estimator with a cosine taper ($p = 0.25$). Plot your spectral estimator, with the frequency axis units of month^{-1}. Give a caption that includes an explanation of your data, e.g. where it is from and between what dates.

You will have to *centre* your time series (remove the mean) before your analysis. What happens if you do not? Why is this?

Comment on your results. What are the dominant frequencies?

6 marks

Question 3

This question relates to material learnt in weeks commencing 27th November and 4th December.

Hint: model fitting will need to be carried out of the centred data, but make sure you add the mean back in when it is needed for forecasting.

(a) Your task is to forecast the next sea level values of your time series. To do this, you are going to fit an AR(p) model using both an (untapered) Yule-Walker and (approximate) maximum likelihood estimation scheme (write your own code for doing these).

You will have to choose p, the order of the AR process. You will do this by assessing which order of p gives the best 1-step ahead *out of sample* predicative performance with a rolling origin. The method proceeds as follows. Start with $p = 1$ and fit an AR(1) process using the first 60 months of data, $x_1, ..., x_{60}$. Forecast the value of x_{61}, which we will denote $\hat{x}_{61}$. Compute the forecast error $e_{61} = \hat{x}_{61} - x_{61}$. Now fit an AR(1) process to $x_1, ..., x_{61}$, forecast x_{62}, and compute the forecast error $e_{62} = \hat{x}_{62} - x_{62}$. Keep repeating all the way until you're forecasting x_{120} by fitting an AR(1) process to $x_1, ..., x_{119}$. Compute the root mean square error (RMSE) $(\text{mean}(e_t^2))^{1/2}$.

Now repeat this procedure for $p = 2, ..., 10$. Tabulate the RMSE for the Yule-Walker and (approximate) maximum likelihood methods for each p.

Choose a value of p based on these results. **7 marks**

(b) For the chosen value of p, plot the corresponding AR(p) spectral density function, using the parameter values as estimated using (approximate) maximum likelihood on the entire time series $x_1, ..., x_{120}$. **2 marks**

(c) For these same parameter values, forecast the next 12 months, $X_{121}, ..., X_{132}$. On a single axes, from $t = 80$ to $t = 132$, plot the time series followed by the forecasted values. **2 marks**

(d) Point forecasts, like this, on their own only have limited use. What is much more useful is supplying an accompanying prediction interval. This is an interval which has some designated probability of containing the realised trajectory.

Recall a zero mean AR(p) model is of the form $X_t - \phi_{1,p}X_{t-1} - ... - \phi_{p,p}X_{t-p} = \epsilon_t$, where $\{\epsilon_t\}$ is a white noise process. Given a time series $x_1, ..., x_N$, for a chosen order p and associated set of estimated parameters $\{\hat{\phi}_{1,p}, ..., \hat{\phi}_{p,p}\}$, the residuals are $\{\xi_{p+1}, ..., \xi_N\}$, where

$$\xi_t = x_t - \hat{\phi}_{1,p}x_{t-1} - ... - \hat{\phi}_{p,p}x_{t-p}.$$

The 95% prediction interval is given as $X_N(l) \pm 1.96\sigma_l$, where σ_l is the standard deviation of the l-step forecast distribution. A naive estimator of σ_l can be shown to be $\hat{\sigma}_\xi\sqrt{l}$ where $\hat{\sigma}_\xi$ is the sample standard deviation of the residuals. On your plot from (c), also show the upper and lower bounds of your 95% prediction interval using the method described here. **3 marks**

(e) Feel free to play around with larger values of p, and perhaps computing the RMSE on further step ahead predictions. Briefly report on your findings. **2 marks**