Time Series Analysis
Coursework

Handout: Monday 21 November 2022.
Deadline: **Submit electronic copy to the turnitin assignment dropbox on Blackboard before Friday 16 December 2022, 1pm.**

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

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

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

Comment your code. A small number of 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**.

Your report must be no more than 12 pages (not including a cover page). This does not mean you have to use all 12 pages - you should be able to do this in fewer than 12 pages.

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

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 IS A VERY SERIOUS OFFENCE.
YOU MUST SUBMIT YOUR OWN PIECE OF WORK.
CASES OF PLAGIARISM WILL BE TREATED AS AN EXAMINATION OFFENCE.

Question 1

This question relates to material learnt in weeks 2 and 8.
In this question you will code three functions that will be of use in your report.

(a) Simulating time series is an essential component of research, allowing methodology to be tested and benchmarked. In this question you will write your own code to simulate a sequence $X_1, \ldots, X_N$ from a stationary Gaussian/normal ARMA(1,1) process.

The tricky thing here is that in order to simulate X_1 you need to know X_0 and ϵ_0. A stochastic process goes from times $-\infty$ to ∞ so that these three variables must be generated to satisfy the stationarity conditions. Let's see how we can do this. Consider the covariance matrix for $[X_0, \epsilon_0]^T$. It takes the form

$$D = \begin{bmatrix} \text{Var}\{X_0\} & \text{Cov}\{X_0, \epsilon_0\} \\ \text{Cov}\{\epsilon_0, X_0\} & \text{Var}\{\epsilon_0\} \end{bmatrix}.$$

We can carry out a Cholesky decomposition of D by finding C such that $D = CC^T$ where C is a lower triangular matrix.

Now consider a vector $Y = [Y_1, Y_2]^T$ where Y_1 and Y_2 are each standard Gaussian/normal and independent. Then the covariance of the vector CY is $CIC^T = CC^T = D$, the required covariance structure.

Using the above scheme, *write your own* function `ARMA11(phi,theta,sigma2,N)` to simulate N values $\mathcal{X} = [X_1, \ldots, X_N]$ from a zero mean Gaussian ARMA(1,1) process $X_t = \phi X_{t-1} + \epsilon_t - \theta\epsilon_{t-1}$. The input parameters are ϕ, θ, σ_ϵ^2 and N. You may use an inbuilt function to compute the Cholesky decomposition.

HINT: You will need to consider the general linear process form of the ARMA(1,1) process to compute the elements of D.

Your answer to this question is your code together with any mathematical derivations required to compute D.

[4 marks]

(b) Write the function `acvs(X,tau)` that will compute the estimator $\hat{s}_\tau^{(p)}$ for a time series stored in a vector `X` at the lags in vector `tau`.

Your answer to this question is your code.

[2 marks]

(c) Write the function `periodogram(X)` that will use the `fft`[1] function to compute the periodogram of a time series stored in a vector `X`. You may also find it useful for the function to apply `fftshift` and output the relevant Fourier frequencies as a vector.

Your answer to this question is your code.

[2 marks]

[1] This is the syntax for MATLAB and R. If using Python, you need to make use of `numpy.fft`.

Question 2

This question relates to material learnt in week 8.
For a portion $X_1, ..., X_N$ of a zero mean stationary random process $\{X_t\}$, the periodogram is defined as

$$\hat{S}^{(p)}(f) = \frac{1}{N}\left|\sum_{t=1}^{N} X_t \mathrm{e}^{-\mathrm{i}2\pi f t}\right|^2 .$$

Asymptotically as $N \to \infty$

$$\begin{aligned} E\{\hat{S}^{(p)}(f)\} &= S(f) \\ \mathrm{Var}\,\{\hat{S}^{(p)}(f)\} &= S^2(f), \qquad 0 < f < \frac{1}{2} \\ \hat{S}^{(p)}(f) &\overset{\mathrm{d}}{=} \frac{S(f)}{2}\chi^2_2, \qquad 0 < f < \frac{1}{2}, \end{aligned}$$

where $S(f)$ is the spectral density function of $\{X_t\}$.

If we restrict ourselves just to the Fourier frequencies $f_k = k/N$, we also find that the $(N/2)+1$ random variables (N even), $\hat{S}^{(p)}(f_0), \hat{S}^{(p)}(f_1), \ldots, \hat{S}^{(p)}(f_{N/2})$, are all approximately pairwise uncorrelated for N large enough; i.e.,

$$\mathrm{corr}\,\{\hat{S}^{(p)}(f_j), \hat{S}^{(p)}(f_k)\} \approx 0, \;\; j \neq k \text{ and } 0 \leq j, k \leq N/2.$$

Consider the following sequence of tasks (A)-(C) for some N (divisible by 4):

(A) Use `ARMA11(phi,theta,sigma2,N)` developed in Question 1 to simulate a time series of length N. You must use the parameters given to you in `coursework_numbers.pdf` on blackboard.

(B) Compute the periodogram from $X_1, ..., X_N$ at the Fourier frequencies.

(C) Repeat steps (A)-(B) a total of $N_r = 10000$ times, storing the sequences $\{S_j^{(p)}(f_{\frac{N}{4}}) : j = 1, \ldots, N_r\}$ and $\{S_j^{(p)}(f_{\frac{N}{4}+1}) : j = 1, \ldots, N_r\}$.

Perform steps (A)-(C) for $N = 4, 8, 16, 32, 64, 128, 256$ and 512. Produce the following 6 plots on separate axes and give a brief comment on them.

(a) The sample mean of $\hat{S}^{(p)}(f_{\frac{N}{4}})$ (y-axis) against N (x-axis). Mark on it with a horizontal line the large sample result for the mean given above.

(b) The sample variance of $\hat{S}^{(p)}(f_{\frac{N}{4}})$ (y-axis) against N (x-axis). Mark on it with a horizontal line the large sample result for the variance given above.

(c) The sample correlation coefficient $\hat{\rho}$ (the Pearson product moment correlation coefficient) between $\{S_j^{(p)}(f_{\frac{N}{4}})\}$ and $\{S_j^{(p)}(f_{\frac{N}{4}+1})\}$ (y-axis), against N (x-axis). Mark on it with a horizontal line the large sample results for the correlation given above.

(d) A histogram for the sampled values of $\{S_j^{(p)}(f_{\frac{N}{4}})\}$ for $N = 4$. Plot on top the probability density function of the asymptotic distribution given above.

(e) A histogram for the sampled values of $\{S_j^{(p)}(f_{\frac{N}{4}})\}$ for $N = 32$. Plot on top the probability density function of the asymptotic distribution given above.

(f) A histogram for the sampled values of $\{S_j^{(p)}(f_{\frac{N}{4}})\}$ for $N = 256$. Plot on top the probability density function of the asymptotic distribution given above.

[12 marks]

Question 3

This question relates to material learnt in weeks 8, 9 and 10.

You each have your own individual time series $x_1, ..., x_{730}$ which you can retrieve from

`wwwf.imperial.ac.uk/~eakc07/time_series/2022/time_series_number.csv`

where instead of "number" you insert the number designated to you `coursework_numbers.pdf` on blackboard, e.g. `wwwf.imperial.ac.uk/~eakc07/time_series/2022/time_series_72.csv`

This is a real time series for daily energy consumption in Germany, measured in GWh and provided by the OPSD[2].

(a) Using your `periodogram` function, compute a direct spectral estimator using a 50% cosine taper. Plot it, clearly labelling the frequency axis and its units. Comment on any peaks you find in your spectrum, matching them to physical phenomena.

Note, you will have to *centre* (remove the mean) of the time series before applying the direct spectral estimator. What happens if you don't do this, and why?

[4 marks]

Imagine you are a statistician tasked with forecasting the next 30 days of energy consumption.

(b) **Code your own functions** to fit an AR(p) model using:

- approximate maximum likelihood
- Yule-Walker (tapered with 50% cosine taper).

[6 marks]

(c) You want to determine which order p of the AR model provides a good fit for you data. A common approach is to do a residual analysis.

Recall an AR(p) model 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 $\{e_{p+1}, ..., e_N\}$, where

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

Should the AR(p) process be a good fit, we would expected $\{e_{p+1}, ..., e_N\}$ to be a sequence of uncorrelated random variables. Conversely, if p is not large enough, we would expect $\{e_{p+1}, ..., e_N\}$ to contain autocorrelation.

Starting at $p = 1$, apply the following procedure.

- Using the centred process, fit an AR(p) process.
- Compute all the residuals.
- The Ljung-Box test is constructed as

$$\begin{aligned} H_0 &: \quad e_{p+1}, ..., e_N \text{ is a sequence of uncorrelated random variables} \\ H_A &: \quad H_0 \text{ is not true.} \end{aligned}$$

[2] https://open-power-system-data.org

The test statistic is

$$L = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k},$$

where n is the number of residuals, h is the number of lags being considered, and $\hat{\rho}_k$ is the estimate of the autocorrelation at lag-k for the residual sequence. Take $\hat{\rho}_k = \hat{s}_k^{(p)}/\hat{s}_0^{(p)}$. Hypothesis H_0 is rejected if $L > \chi^2_{1-\alpha,h}$, where $\chi^2_{1-\alpha,h}$ is the $(1-\alpha)$-quantile of the χ^2_h distribution.

- Use the smallest p for which we fail to reject H_0.

Apply the Ljunge-Box test for $h = 14$ (typical for daily time series) at the $\alpha = 0.05$ level.

For both maximum likelihood and tapered Yule-Walker, report the smallest p and the corresponding estimated parameter values for which you do not reject H_0

[6 marks]

(d) Using the maximum likelihood model fitted in part (c), forecast $X_{731}, ..., X_{760}$. Remember to include the mean back in. On a single axes, from $t = 710$ to $t = 760$, plot the time series followed by the forecasted values.

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. 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 $\hat{\sigma}_e\sqrt{l}$ where $\hat{\sigma}_e$ is the sample standard deviation of the residuals. On your plot, also show the upper and lower bounds of your 95% prediction interval using the method described here.

How would you report your findings back to those who have tasked you?

[4 marks]