STAT 153 & 248 - Time Series
Homework Five

Spring 2025, UC Berkeley

Due by 11:59 pm on 28 April 2025
Total Points = 92 (STAT 153) and 114 (STAT 248)

1. Let $\{y_t\}$ be a doubly infinite sequence of random variables that is stationary with AutoCovariance function $\gamma_y(h)$. Let

$$x_t = (a + bt)\alpha_t + y_t$$

where a and b are fixed real numbers, and α_t is a deterministic seasonal function with period s i.e., $\alpha_t = \alpha_{t+s}$ for all t.

   a) Is $\{x_t\}$ stationary? Why or why not? (**2 points**)

   b) Let $u_t = \nabla_s^2 x_t$ where $\nabla_s^2 = (I - B^s)^2$ (B being the Backshift operator). Show that u_t is stationary. (**3 points**)

   c) Write the AutoCovariance function of $\{u_t\}$ in terms of the AutoCovariance function γ_y of $\{y_t\}$. (**3 points**).

2. Consider the AR(2) equation:

$$y_t = 0.75y_{t-1} - 0.125y_{t-2} + \epsilon_t$$

where $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$.

   a) Argue that this AR(2) equation admits a unique causal stationary solution. (**3 points**).

   b) Explicitly write down the formula for the causal stationary solution in terms of $\{\epsilon_t\}$. (**4 points**).

   c) Calculate the ACF of the causal stationary solution. (**4 points**).

3. We have seen in class that $\sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$ is the unique stationary solution to the AR(1) difference equation: $y_t - \phi_1 y_{t-1} = \epsilon_t$ for $|\phi_1| < 1$ and that there can be many *non-stationary* solutions.

   a) Show that $y_t = c\phi_1^t + \sum_{j=0}^{\infty} \phi_1^j \epsilon_{t-j}$ is a solution to the difference equation for every real number c. (**2 points**)

   b) Show that this y_t is non-stationary for $c \neq 0$. (**2 points**).

4. Let $\{y_t\}$ be a doubly infinite sequence of random variables that is stationary. Let

$$x_t = \beta_0 + \beta_1 t + \cdots + \beta_q t^q + y_t$$

where $\beta_0, \ldots, \beta_q$ are real numbers with $\beta_q \neq 0$.

a) Show that $(I - B)^k y_t$ is stationary for every $k \geq 1$. (**3 points**)

b) Show that $(I - B)^k x_t$ is not stationary for $k < q$ and that it is stationary for $k \geq q$. (**3 points**).

5. A time series data set is plotted in the top panel of Figure 1. Its sample autocorrelation and partial autocorrelation functions are plotted in the middle and bottom panels of Figure 1 respectively.

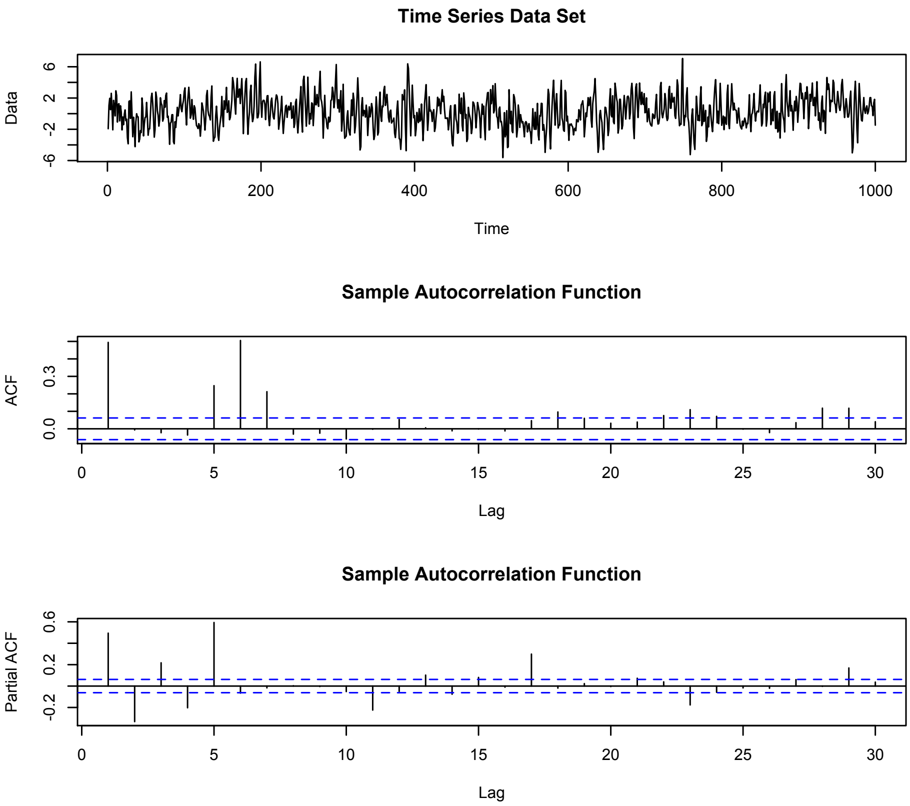

Figure 1: A Time Series Data set along with its sample autocorrelation and partial autocorrelation functions

For each of the following models given by their ARIMA statsmodels functions, provide reasons for using or not using that model for this dataset: (3 × 2 = **6 points**)

a) `ARIMA(dataset, order = (0, 0, 7))`

b) `ARIMA(dataset, order = c(0, 0, 1), seasonal_order = (0, 0, 1, 6))`

c) `ARIMA(dataset, order = c(0, 1, 1), seasonal_order = (0, 1, 1, 6))`

6. Consider the sunspots data that we looked at in class. Split this dataset by removing the last 50 datapoints and keeping them aside as a test dataset. The remaining observations will form the training dataset.

a) Plot the sample acf and pacf for this dataset. Based on these plots, propose an appropriate AR(p) or MA(q) model for the data. (**3 points**)

b) Fit your model from the above part to the training dataset and use the fitted model to obtain predictions for the test time points. Compare the predictions to the actual test observations and report the mean squared error of prediction. (**3 points**). (**3 points**)

c) Fit the ARMA(p, q) to this dataset for all $0 \leq p, q \leq 12$, and use the AIC and BIC to select the best of these models (you should report one best AIC model, and another best BIC model). (**6 points**)

d) Obtain predictions for the test time points, and compare them to the actual observations in the test dataset. Which of the three models (best AIC, best BIC and your model from part (a), if they are different) performs best in terms of mean squared error of prediction? (**3 points**)

7. Download the FRED dataset $\{y_t\}$ on "Long-Term Government Bond Yields: 10-year Main (including benchmark) for the United States" from `https://fred.stlouisfed.org/series/IRLTLT01USM156N`.

   a) Examine the sample ACF and sample PACF plots of the differenced data $\nabla y_t := y_t - y_{t-1}$. Based on these, propose an appropriate AR(p) and an appropriate MA(q) model for ∇y_t. (**4 points**)

   b) Fit your AR(p) and MA(q) models to the differenced data ∇y_t to get parameter estimates. Obtain an MA(q) approximation to your fitted AR(p) model using the `arma2ma` function. Compare this approximation to your fitted MA(q) model. Are they similar? (**4 points**).

   c) Using your AR and MA models for the differenced data, obtain predictions of y_t for the next 100 future points. You can refit your models to the original data y_t directly using `ARIMA`. Plot the predictions and comment on whether they are similar. (**4 points**)

   d) Run a model selection procedure for $p \in \{0, 1, \ldots, 10\}$ and $q \in \{0, 1, \ldots, 10\}$ and report the best AR model, best MA model and the best ARMA model. (**6 points**)

   e) Compare (visually) your predictions from (c) with the predictions given by the best models from part (d). Comment on the closeness of the predictions. (**4 points**).

8. Download the Google trends dataset (2004-present and for the United States) for the query *golf*. Remove the data for the last 36 months and set it aside as the test dataset. Use the Box-Jenkins modeling strategy to fit a time series model of the form ARIMA$(p, d, q) \times (P, D, Q)_s$ for appropriate p, d, q, P, D, Q to the training dataset. Clearly describe the main steps in your analysis and include relevant plots. Use your fitted model to predict the observations in the test dataset and report and comment on the accuracy of the predictions. (**10 points**).

9. Download the Google trends dataset (2004-present and for the United States) for the query *aquarium*. Remove the data for the last 36 months and set it aside as the test dataset. Use the Box-Jenkins modeling strategy to fit a time series model of the form ARIMA$(p, d, q) \times (P, D, Q)_s$ for appropriate p, d, q, P, D, Q to the training dataset. Clearly describe the main steps in your analysis and include relevant plots. Use your fitted model to predict the observations in the test dataset and report and comment on the accuracy of the predictions. (**10 points**).

10. [**This question is only for students taking STAT 248**] The goal of this question is parameter estimation for MA(q) models for a fixed q: $y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$

with $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. It is difficult to work with the full likelihood so we shall work with the conditional likelihood of $y_1, \dots, y_n$ conditioned on $\epsilon_t = 0$ for all $t \leq 0$.

a) Write down the conditional likelihood of $y_1, \dots, y_n$ conditioned on $\epsilon_t = 0$ for all $t \leq 0$. As a hint, you can use the fact that under this conditioning (and for fixed parameters $\mu, \theta_1, \dots, \theta_q$), one can explicitly write $\epsilon_1, \dots, \epsilon_n$ in terms of $y_1, \dots, y_n$. Express the conditional likelihood as

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{S(\mu, \theta_1, \dots, \theta_q)}{2\sigma^2}\right),$$

for some $S(\mu, \theta_1, \dots, \theta_q)$ which plays the role of Conditional Sum of Squares. (**6 points**).

b) Under the prior $\mu, \theta_1, \dots, \theta_q, \log \sigma \overset{\text{i.i.d}}{\sim} \text{uniform}(-C, C)$ for a large $C \to \infty$, calculate the posterior of $\mu, \theta_1, \dots, \theta_q$. (**4 points**).

c) Let $\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q$ denote the point estimates defined as the minimizers of $S(\mu, \theta_1, \dots, \theta_q)$. Using a second order Taylor approximation of $S(\mu, \theta_1, \dots, \theta_q)$ around $(\hat{\mu}, \hat{\theta}_1, \dots, \hat{\theta}_q)$, approximate the posterior from part (b) by an explicit multivariate t-density. (**5 points**).

d) Consider the `varve` dataset that we used in Lecture 20. Work with $y_t = \nabla \log(\text{varve}_t)$ (i.e., differenced log data). Use the above scheme to obtain parameter estimates and standard errors for fitting the MA(1) and MA(2) models to y_t (you can use optimization libraries such as `scipy.optimize.minimize` for minimizing $S(\mu, \theta_1, \dots, \theta_q)$ as well as some numerical differentiation libraries such as `numdifftools` for computing the Hessian term appearing the t-density in (d)). Compare your results with those obtained by using the ARIMA function. (**7 points**).