Midterm 1, STATS 531/631 W26

In class on 2/16

Instructions. The test is closed book, and you are not allowed access to any notes. Any electronic devices in your possession must be turned off and remain in a bag on the floor.

For each question, circle one letter answer and provide some supporting reasoning.

Q1. Stationarity and unit roots.

Suppose that a dataset $y^*_{1:N}$ is well described by the statistical model

$$Y_n = a + bn + \epsilon_n,$$

where ϵ_n is a Gaussian ARMA process and $b \neq 0$. Which of the following is the best approach to time series modeling of $y^*_{1:N}$?

A. The data are best modeled as non-stationary, so we should take differences. The differenced data are well described by a stationary ARMA model.

B. The data are best modeled as non-stationary, and we should use a trend plus ARMA noise model.

C. The data are best modeled as non-stationary. It does not matter if we difference or model as trend plus ARMA noise since these are both linear time series models which become equivalent when we estimate their parameters from the data.

D. We should be cautious about doing any of A, B or C because the data may have nonstationary sample variance in which case it may require a transformation before it is appropriate to fit any ARMA model.

Q2. Calculations for ARMA models

Let Y_n be an ARMA model solving the difference equation

$$Y_n = (1/4)Y_{n-2} + \epsilon_n + (1/2)\epsilon_{n-1}.$$

This is equivalent to which of the following:

A. $Y_n = (1/2)Y_{n-1} + \epsilon_n$

B. $Y_n = -(1/2)Y_{n-1} + \epsilon_n$

C. $Y_n = (1/2)Y_{n-2} - (1/16)Y_{n-4} + \epsilon_n + \epsilon_{n-1} + (1/4)\epsilon_{n-2}$

D. $Y_n = -(1/2)Y_{n-2} - (1/16)Y_{n-4} + \epsilon_n + \epsilon_{n-1} + (1/4)\epsilon_{n-2}$

E. None of the above

Q3. Likelihood-based inference for ARMA models

```
##
## Call:
## arima(x = huron_level, order = c(2, 0, 1))
##
```

```
## Coefficients:
##          ar1     ar2     ma1  intercept
##       0.3388  0.4092  0.6320   176.4821
## s.e.  0.4646  0.4132  0.4262     0.1039
##
## sigma^2 estimated as 0.04479:  log likelihood = 21.42,  aic = -32.84
```

```
##
## Call:
## arima(x = huron_level, order = c(2, 0, 2))
##
## Coefficients:
##           ar1     ar2     ma1     ma2  intercept
##       -0.1223  0.7646  1.1310  0.1310   176.4815
## s.e.   0.0682  0.0550  0.1084  0.1004     0.1004
##
## sigma^2 estimated as 0.04364:  log likelihood = 22.64,  aic = -33.28
```

The R output above uses `stats::arima` to fit ARMA(2,1) and ARMA(2,2) models to the January level (in meters above sea level) of Lake Huron from 1860 to 2024. Residual diagnostics (not shown) show no major violation of model assumptions. We aim to choose one of these as a null hypothesis of no trend for later comparison with models including a trend.

Which is the best conclusion from the available evidence:

A: The ARMA(2,2) model has a lower AIC so it should be preferred.

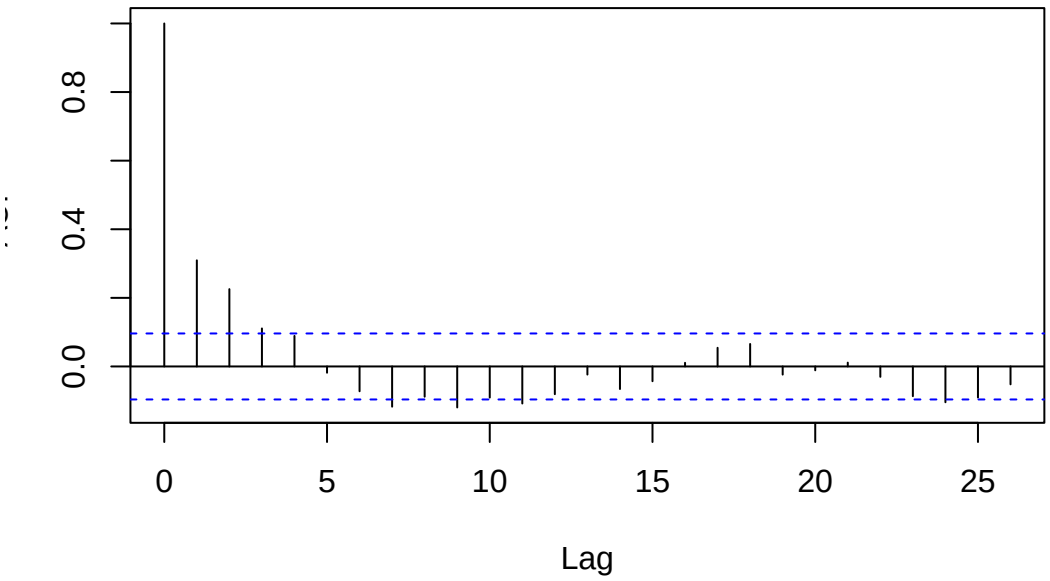
B: We cannot reject the null hypothesis of ARMA(2,1) since the ARMA(2,2) model has a likelihood less than 1.92 log units higher than ARMA(2,1). Since there is not sufficient evidence to the contrary, it is better to select the simpler ARMA(2,1) model.

C: Since the comparison of AIC values and the likelihood ratio test come to different conclusions in this case, it is more-or-less equally reasonable to use either model.

D: When the results are borderline, numerical errors in the `stats::arima` optimization may become relevant. We should check using optimization searches from multiple starting points in parameter space, for example, using `arima2::arima`.

Q4. Interpreting diagnostics

We consider data $y_{1:415}$ where y_n is the time, in miliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. Let $z_n = \log(y_n)$, with log being the natural logarithm. The sample autocorrelation function of $z_{1:415}$ is shown below.



We are interested about whether it is appropriate to model the time series as a stationary causal ARMA process. Which of the following is the best interpretation of the evidence from these plots:

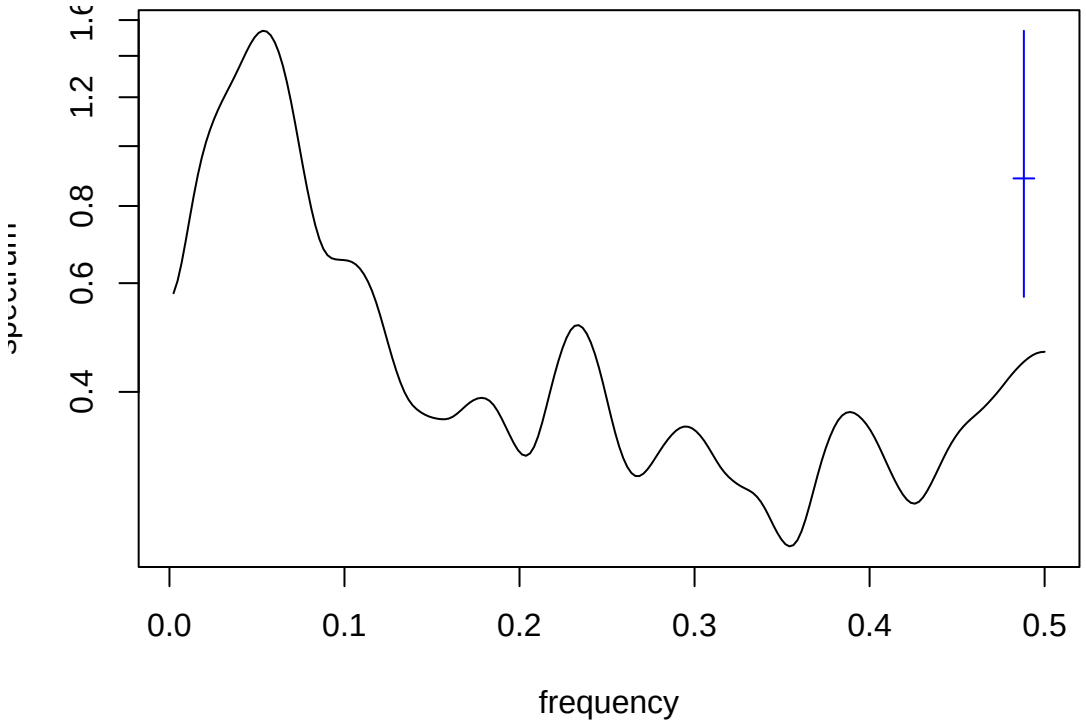
A. There is clear evidence of a violation of stationarity. We should consider fitting a time series model, such as ARMA, and see if the residuals become stationary.

B. This plot suggests there would be no benefit from detrending or differencing the time series before fitting a stationary ARMA model. It does not rule out a sample covariance that varies with time, which is incompatible with ARMA.

C. This plot is enough evidence to demonstrate that a stationary model is reasonable. We should proceed to check for normality, and if the data are also not far from normally distributed then it is reasonable to fit an ARMA model by Gaussian maximum likelihood.

Q5. The frequency domain

We consider data $y_{1:415}$ where y_n is the time interval, in milliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. Let $z_n = \log(y_n)$, with log being the natural logarithm. A smoothed periodogram of $z_{1:415}$ is shown below. Units of frequency are the default value in R, i.e., cycles per unit observation. We see a peak at a frequency of approximately 0.07.



Which if the following is the best inference from this figure

A. Transitions between rapid neuron firing (short intervals between firing) and slow neuron firing (long intervals between firing) occur every $1/0.07 \approx 14$ firing events.

B. The neuron has a characteristic duration between firing events of $1/0.07 \approx 14$ milliseconds.

C. The neuron has a characteristic duration between firing events of $1/\exp(0.07) \approx 0.9$ milliseconds.

Q6. Scholarship for time series projects

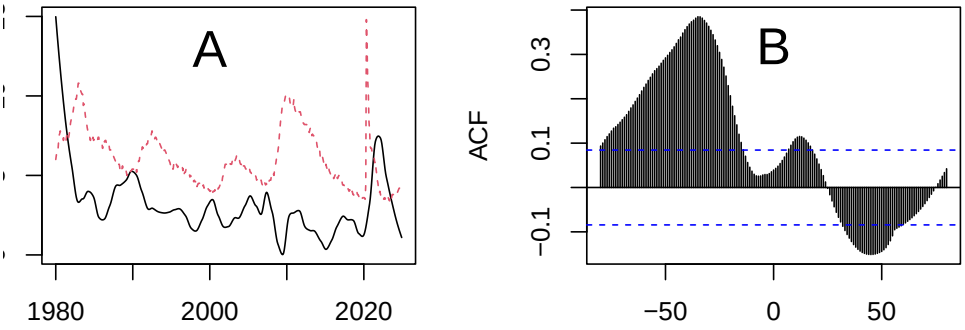
You discover that your team-mate is using Google Translate to carry out their share of the writing. The translation looks poorly done, similar in quality to ChatGPT, and does not use technical time series terminology correctly. What is the best course of action among the options below

A. Alert the instructor that you have a team mate adopting questionable scholarship strategies, in order to make sure you are not personally held responsible.

B. Ask ChatGPT to rewrite this problematic section to improve its quality

C. Help your team mate to rewrite the section in their own voice (shared with your voice).

Q7. Data analysis



(A) Inflation (black) and unemployment (red) for the USA, 1980-2024. (B) Cross-correlation function, `ccf(inflation,unemployment)`. What is the best interpretation of this plot?

A: High inflation generally led high unemployment, with a lag of about 4 yr.
B: High inflation generally followed high unemployment, with a lag of about 4 yr.
C: Association is not causation, so we should not interpret a cross-correlation plot in terms of lead and lag relationships.
