ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 6 - Due date 03/06/23

Tony Jiang

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A06_Sp23.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Submit this pdf using Sakai.

R packages needed for this assignment: "xlsx" or "readxl", "ggplot2", "forecast","tseries", and "Kendall". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.

Questions

This assignment has general questions about ARIMA Models.

Packages needed for this assignment: "forecast","tseries". Do not forget to load them before running your script, since they are NOT default packages.\

```
#Load/install required package here
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(tseries)
library(stats)
library(sarima)
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'sarima'
```

```
## The following object is masked from 'package:stats':
##
##     spectrum
```

Q1

Describe the important characteristics of the sample autocorrelation function (ACF) plot and the partial sample autocorrelation function (PACF) plot for the following models:

- AR(2)

  Answer: ACF - ACF of AR(2) should have exponential and gradual decays over lags, which means ACF should gradually become statistically insiginificant over lags. PACF - PACF of AR(2) should have a cutoff at lag 2, which means PACF at lag 1 and lag 2 should be statistically significant (above the blue dashed line in pacf polt). And PACF at lag 3 should be statistically insignificant. Some spikes after lag 3 may be statistically significant, but most spikes should remain statistically insignificant.

- MA(1)

  Answer: ACF - ACF of MA(1) should have a cutoff at lag 1, which means ACF at lag 1 should be statistically significant (above the blue dashed line in pacf plot). And PACF at lag 2 and beyond should be statistically insignificant. Some spikes after lag 2 may be statistically significant, but most spikes should remain statistically insignificant. PACF - PACF of MA(1) should display have exponential and gradual decays over lags, which means PACF of MA(1) should gradually become statistically insiginificant over lags.

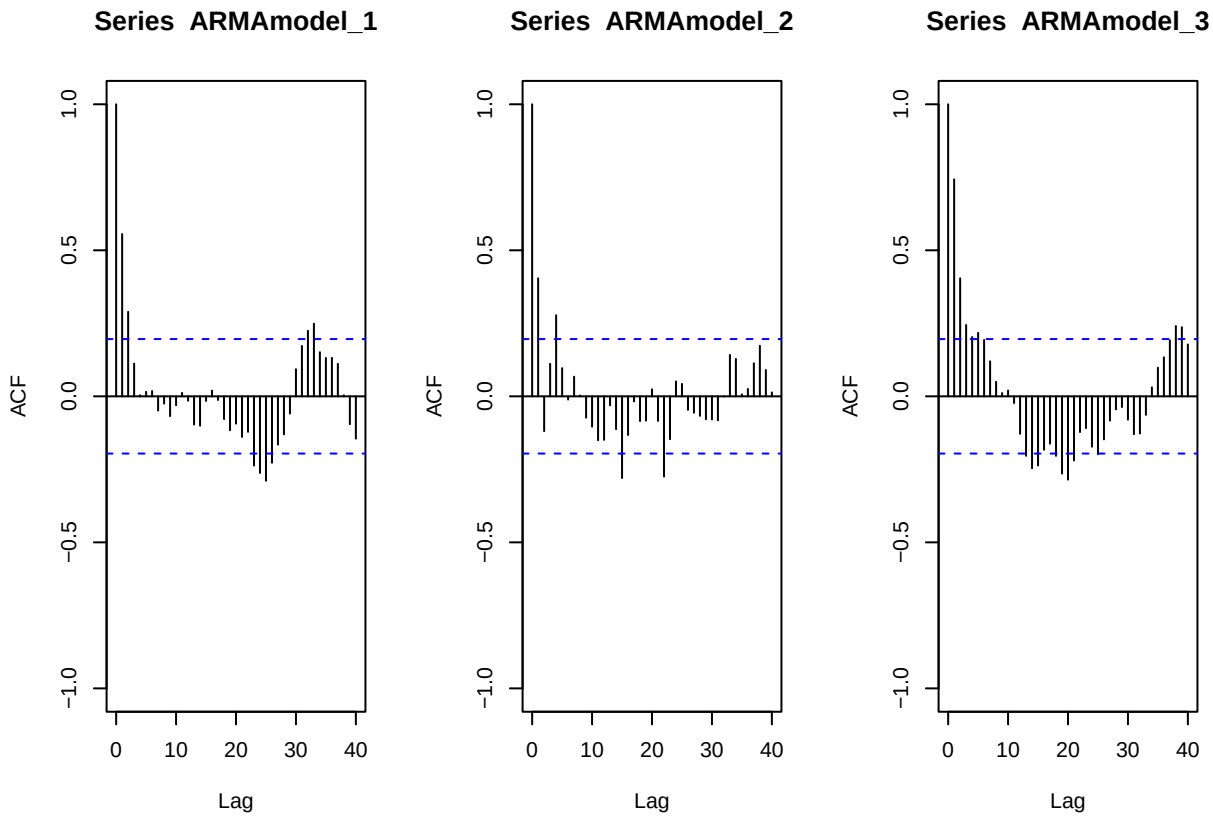
Q2

Recall that the non-seasonal ARIMA is described by three parameters ARIMA(p, d, q) where p is the order of the autoregressive component, d is the number of times the series need to be differenced to obtain stationarity and q is the order of the moving average component. If we don't need to difference the series, we don't need to specify the "I" part and we can use the short version, i.e., the ARMA(p, q). Consider three models: ARMA(1,0), ARMA(0,1) and ARMA(1,1) with parameters $\phi = 0.6$ and $\theta = 0.9$. The ϕ refers to the AR coefficient and the θ refers to the MA coefficient. Use R to generate $n = 100$ observations from each of these three models

```
#ARMA(1, 0)

ARMAmodel_1 <- arima.sim(model=list(ar=0.6), n=100) #the AR coefficient is 0.6

#ARMA(0, 1)

ARMAmodel_2 <- arima.sim(model=list(ma=0.9), n=100) #the MA coefficient is 0.9

#ARMA(1, 1)
ARMAmodel_3 <- arima.sim(model=list(ar = 0.6, ma = 0.9), n = 100)
#the AR coefficient is 0.6, the MA coefficient is 0.9
```

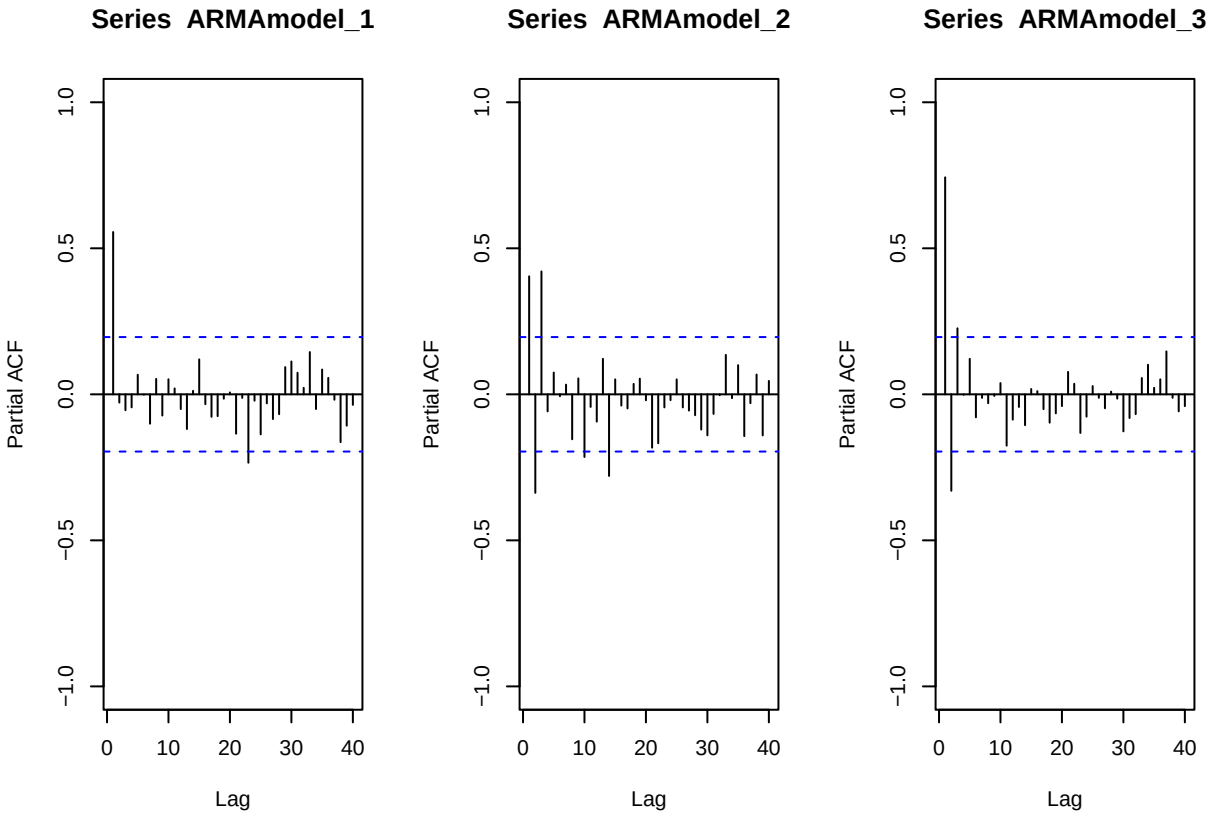
(a) Plot the sample ACF for each of these models in one window to facilitate comparison (Hint: use command $par(mfrow = c(1, 3))$ that divides the plotting window in three columns).

```
# set three plots to display in one graph
par(mfrow = c(1, 3))

#graph acf plots of three series
acf(ARMAmodel_1, lag.max = 40, ylim = c(-1, 1))
acf(ARMAmodel_2, lag.max = 40, ylim = c(-1, 1))
acf(ARMAmodel_3, lag.max = 40, ylim = c(-1, 1))
```



(b) Plot the sample PACF for each of these models in one window to facilitate comparison.

```
# set three plots to display in one graph
par(mfrow = c(1, 3))

#graph pacf plots of three series
pacf(ARMAmodel_1, lag.max = 40, ylim = c(-1, 1))
pacf(ARMAmodel_2, lag.max = 40, ylim = c(-1, 1))
pacf(ARMAmodel_3, lag.max = 40, ylim = c(-1, 1))
```



(c) Look at the ACFs and PACFs. Imagine you had these plots for a data set and you were asked to identify the model, i.e., is it AR, MA or ARMA and the order of each component. Would you be identify them correctly? Explain your answer.

Answer: ***I will identify series 1 as an AR(1) process.*** Because its acf has exponential decays and we can see a cutoff at lag 1 in its pacf. These two are identifiers of AR(1) processes. ***I will identify series 2 as an MA(1) process.*** Because its pacf has exponential decays and we can see a cutoff at lag 1 in its acf. These two are identifiers of MA(1) processes. ***For series 3***, I may identify it as an ARMA process, but I am not able to identify it as a ARMA(1, 1) process correctly. First, we can see exponential decays in its acf and pacf plots, which indicates the possibility of an ARMA model. But we can't observe any cutoffs in acf of pacf plots clearly, which makes it difficult for us to identify p and q of this ARMA process.

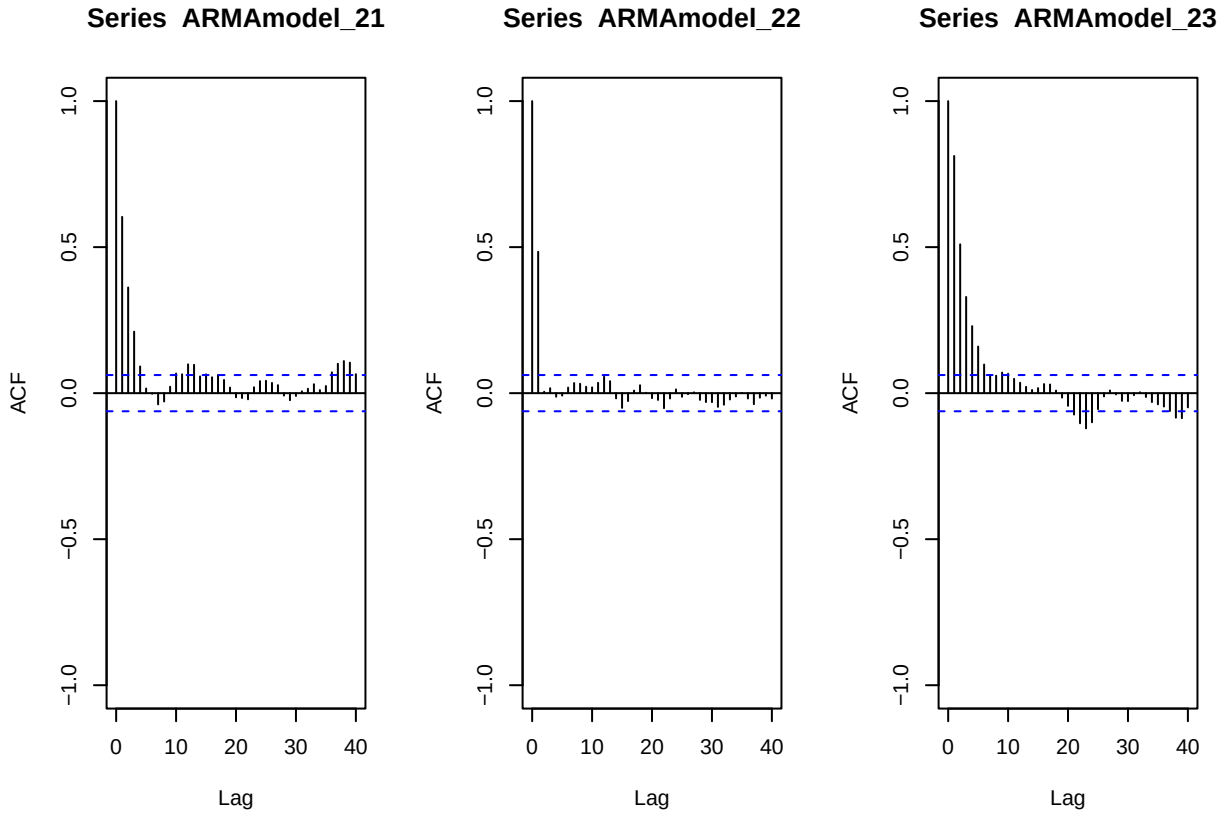
(d) Compare the ACF and PACF values R computed with the theoretical values you provided for the coefficients. Do they match? Explain your answer.

Answer: ***For series 1***, from the PACF, we can get the computed coefficient close to 0.6, which approximately matches the coeeficients provided by me. This is reasonable because we should be able to tell the coefficient of AR simply from its pacf plot, when no MA component contaminates the series. The slight difference is caused by the small number of observations in the series. PACF value at lag 1 computed by R is only an unbiased estimator of the coefficient. The difference between the estimate and the true value (residue) is likely to be larger when less observations are available (low precision). ***For series 2***, the computed ACF and PACF values computed by R at lag 1 do not match the coefficients provided by me. This is because, for a given MA(1) process, $\mathrm{Var}(y_t, y_{t-1})$, which is its acf at lag 1, is $\theta * {\sigma^2}_e$. Because the value of ${\sigma^2}_e$ may not be 1, we can't find 0.9 in its acf plot at lag 1. ***For series 3***, the computed ACF and PACF values computed by R at lag 1 do not match the coefficients provided by me. Its pacf value at lag 1 is distorted by its MA component, which makes it change from 0.6 to a number close to 0.8. As for its ACF value at lag 1, it also got distorted by its AR component, which makes its value highr than before.

To be more specific, both acf and pacf values are higher at lag 1 than before. The combination of AR and MA introduces more correlation between various lags than a sole AR or MA process, which should raesonably increase this process's acf and pacf values.

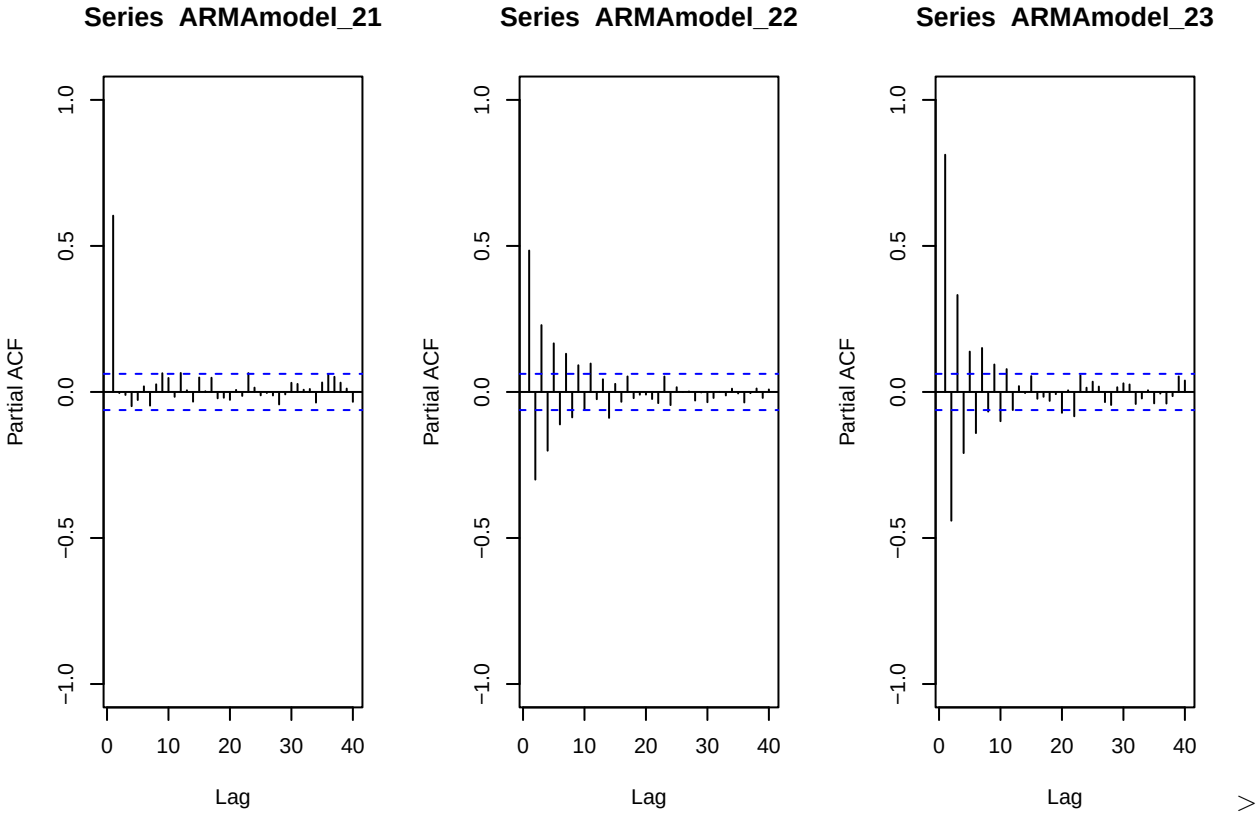
(e) Increase number of observations to $n = 1000$ and repeat parts (a)-(d).

```
#ARMA(1, 0)

ARMAmodel_21 <- arima.sim(model=list(ar=0.6), n=1000) #the AR coefficient is 0.6

#ARMA(0, 1)

ARMAmodel_22 <- arima.sim(model=list(ma=0.9), n=1000) #the MA coefficient is 0.9

#ARMA(1, 1)
ARMAmodel_23 <- arima.sim(model=list(ar = 0.6, ma = 0.9), n = 1000)
#the AR coefficient is 0.6, the MA coefficient is 0.9

# set three plots to display in one graph
par(mfrow = c(1, 3))

#graph acf plots of three series
acf(ARMAmodel_21, lag.max = 40, ylim = c(-1, 1))
acf(ARMAmodel_22, lag.max = 40, ylim = c(-1, 1))
acf(ARMAmodel_23, lag.max = 40, ylim = c(-1, 1))
```




```
# set three plots to display in one graph
par(mfrow = c(1, 3))
```

```r
#graph pacf plots of three series
pacf(ARMAmodel_21, lag.max = 40, ylim = c(-1, 1))
pacf(ARMAmodel_22, lag.max = 40, ylim = c(-1, 1))
pacf(ARMAmodel_23, lag.max = 40, ylim = c(-1, 1))
```



Answer: ***For series 1***, from the PACF, we can get the computed coefficient closer to 0.6 than the result computed with 100 observations. And its propoerty as a cutoff point is also more clear than before. This is reasonable because we should be able to tell the coefficient of AR simply from its pacf plot, when no MA component contaminates the series. Increasing number of observations enhances accuracy of the estimate, which leads to a PACF value at lag 1 closer to 0.6.

For series 2 and 3, I identify the same thing as I stated in my precious analysis. For MA(1) process, since the distribution of the error term is unknown, I cannot tell what the estimate should be. Therefore, I cannot tell what ACF value at lag 1 computed by R should be. So, I cannot say whether the precision of estimate gets improved or not. But I will assume, the accuracy is increased with more observations. The same reasoning is also applicable to series 3, ARMA(1, 1) process.

Q3

Consider the ARIMA model $y_t = 0.7 * y_{t-1} - 0.25 * y_{t-12} + a_t - 0.1 * a_{t-1}$

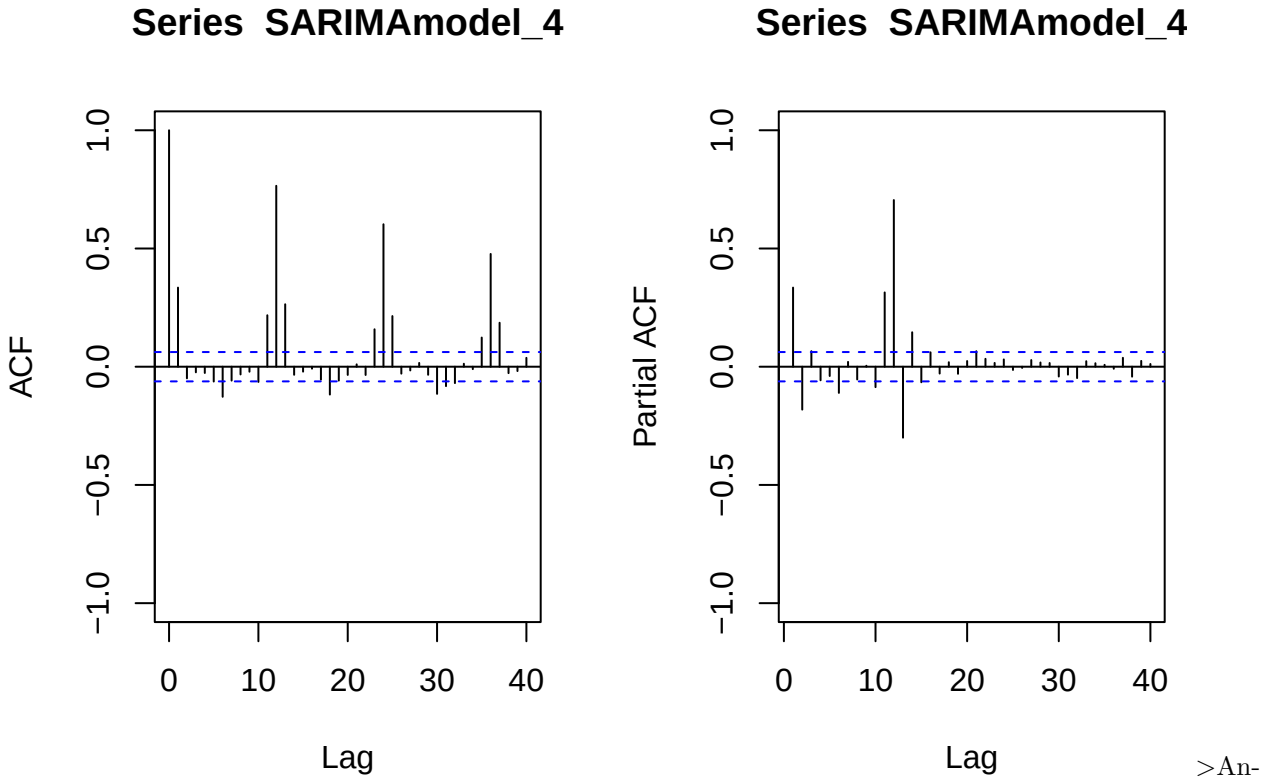
(a) Identify the model using the notation ARIMA$(p, d, q)(P, D, Q)_s$, i.e., identify the integers p, d, q, P, D, Q, s (if possible) from the equation. > ***Based on the equation, there is no differencing, so d = D = 0.*** y_{t-1} ***stands for AR component, so p = 1.*** a_{t-1} ***stands for MA component, so q = 1.*** y_{t-12} ***stands for seasonal AR component, so P = 1. So, this is an ARIMA(1,0,1)(1,0,0) model.***

(b) Also from the equation what are the values of the parameters, i.e., model coefficients. > ***Based on the model, 0.7 is the coefficient for nonseasonal AR component. -0.1 is the coefficient for nonseasonal MA component. -0.25 is the coefficient for seasonal AR component.***

Q4

Plot the ACF and PACF of a seasonal ARIMA$(0,1)\times(1,0)_{12}$ model with $\phi = 0.8$ and $\theta = 0.5$ using R. The 12 after the bracket tells you that $s = 12$, i.e., the seasonal lag is 12, suggesting monthly data whose behavior is repeated every 12 months. You can generate as many observations as you like. Note the Integrated part was omitted. It means the series do not need differencing, therefore $d = D = 0$. Plot ACF and PACF for the simulated data. Comment if the plots are well representing the model you simulated, i.e., would you be able to identify the order of both non-seasonal and seasonal components from the plots? Explain.

```
# A SARIMA model with 1000 observations, 0.5 for nonseasonal MA,
# and 0.8 for seasonal AR
SARIMAmodel_4<- sim_sarima(model=list(ma=0.5,sar=0.8, nseasons=12), n=1000)

par(mfrow = c(1, 2))
acf(SARIMAmodel_4, lag.max = 40, ylim = c(-1, 1))
pacf(SARIMAmodel_4, lag.max = 40, ylim = c(-1, 1))
```



>Answer: ***For seasonal component***: Order of 1 for seasonal AR component is well displayed in the plot. I can observe positive spikes in its ACF plot at lag 12, 24, 36, ... with exponential decays. I can observe only one positive spike at lag 12 in its PACF plot, as well. These two characteristics lead to P = 1 (order of 1 for seasonal AR component). ***For nonseasonal component***: I see a cutoff at lag 1 in its ACF plot, which may demonstrate order of 1 for its nonseasonal MA component. In contrast, I cannot observe a clear decay in its ACF plot, which dispells the possibility of the existence of nonseasonal AR component. ***Overall, based on the plot, I will conclude the process is an ARIMA$(0,0,1)(1,0,0)_{12}$ model, which indicates a good representation of our model.***