STAT 443 Lab 4 : SARIMA

Caden Hewlett

2024-01-26

Question 1

Without using any mathematical notation, describe in words what it means for a time series to be stationary.

Solution: In short, a stationary time series is one that can be well-described by some stationary stochastic process (i.e. AR, MA, ARMA, etc.)

There are three primary components to a (weakly) stationary time series. Let's describe them without any mathematical notation. Firstly, the expected value of the series must be some real constant - in other words, independent of time for all time steps. This means that the expected value of the series *right now* is exactly the same as it will be a whole month from now. We normally detect lack of stationarity with this first condition, since it is easy to detect the presence of a series or trend - both of which are dependent on time and hence disqualify a series from being stationary.

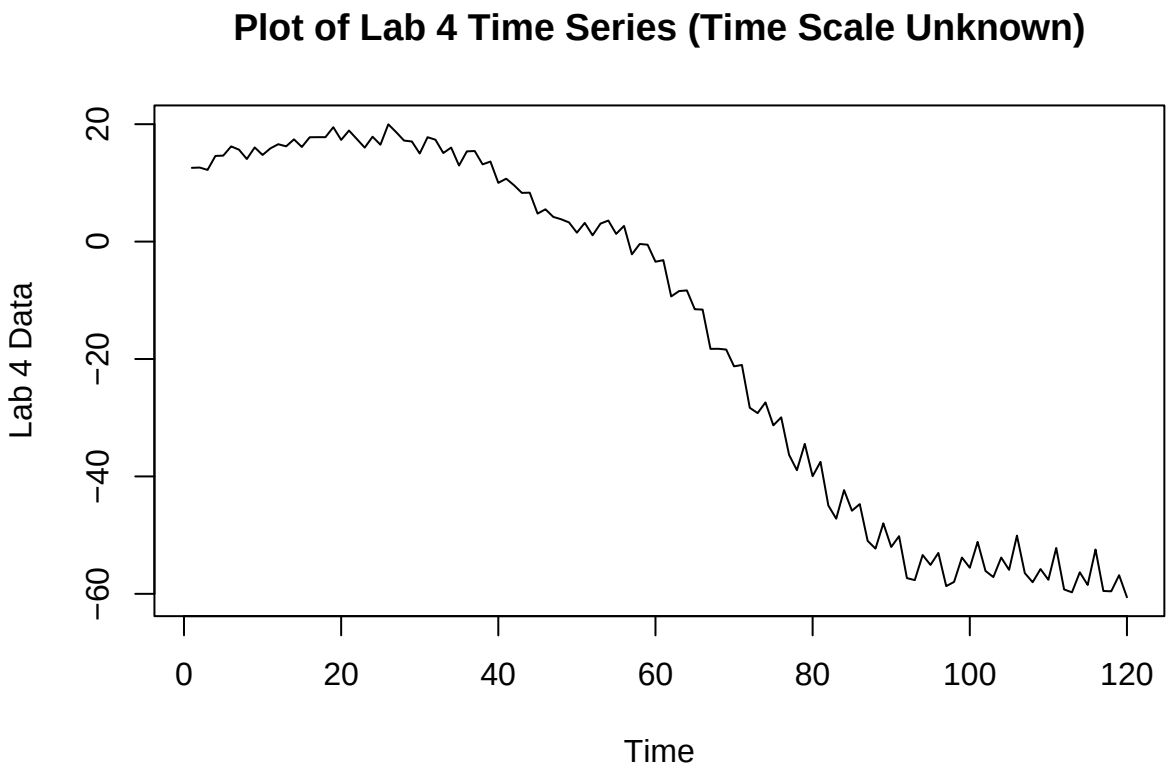
Secondly, the series must have finite variance. While not getting bogged down in the math, let's consider why this is necessary. If the series had infinite variance, it would be too volatile to "nicely" describe in terms of some stationary stochastic process. It would be bouncing around everywhere across time steps, in a way that's impossible to algebraically (or realistically) describe.

Finally, and most technically, the so-called "autocovariance function" or *acf* must only be a function of what is called the "lag." For example, this means that the correlation between the value of the series at the *current time* and, let's say two time steps from now, is only dependent on the difference in the time step of two. There's no additional function of time playing a role in the correlation between these points. A nice example of this would be to consider the correlation between a time of one and a time of three. This is called a "lag" of two. We'd expect the correlation between these to be the exact same as if we had decided to inspect the correlation between a time of three and a time of five (also a "lag" of two.) If the actual times themselves, e.g. one, three or five caused changes in the acf, the values here wouldn't be the same. This is arguably the most difficult to explain non-mathematically, but can be considered a more technical variation to the first claues (but not exactly the same.)

Question 2

Consider a realization of a process with values given in file `lab4data.csv`. Read in the data, coerce it into a ts object, plot and comment on whether the series appears to satisfy the requirements of stationarity.

```
df = read.csv("lab4data.csv")
lab4ts = ts(df$x)
plot(lab4ts, ylab  = 'Lab 4 Data', main = "Plot of Lab 4 Time Series (Time Scale Unknown)")
```



We note that no time scale was provided, hence we left the horizontal axis as indices.

Visually, these data do *not* appear to meet the prerequisites of stationarity. There is an obvious downward trend to the data, implying that the value of $\{X_t\}$ is very likely dependent on some trend component m_t, a function of time. Hence, it's not the case that $\mathbb{E}(X_t) = \mu$ for some time-independent constant $\mu \in \mathbb{R}$. Likely we'd have to de-seasonalize and/or de-trend the data.

Question 3

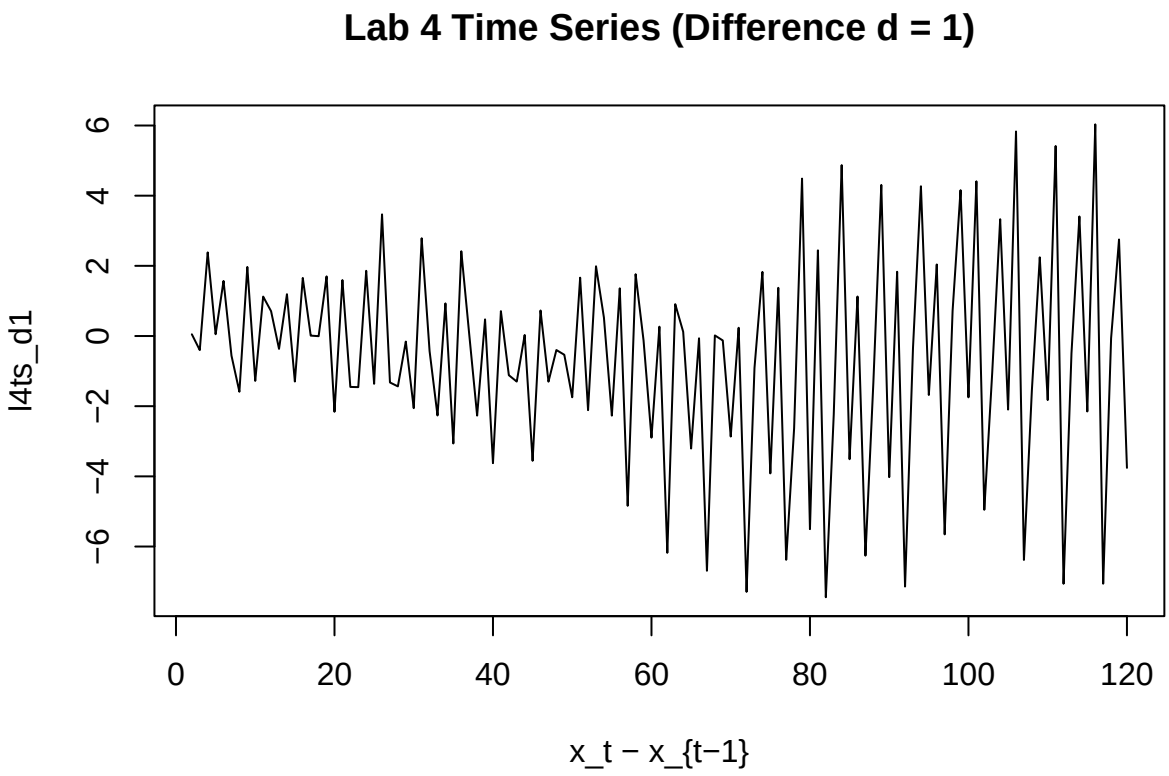
Let's take a difference $d = 1$ for these data, letting y_t be the differenced series.

In other words, we let ∇ be the difference operator, and we consider:

$$y_t = \nabla_1 x_t = x_t - x_{t-1}$$

Let's do this now.

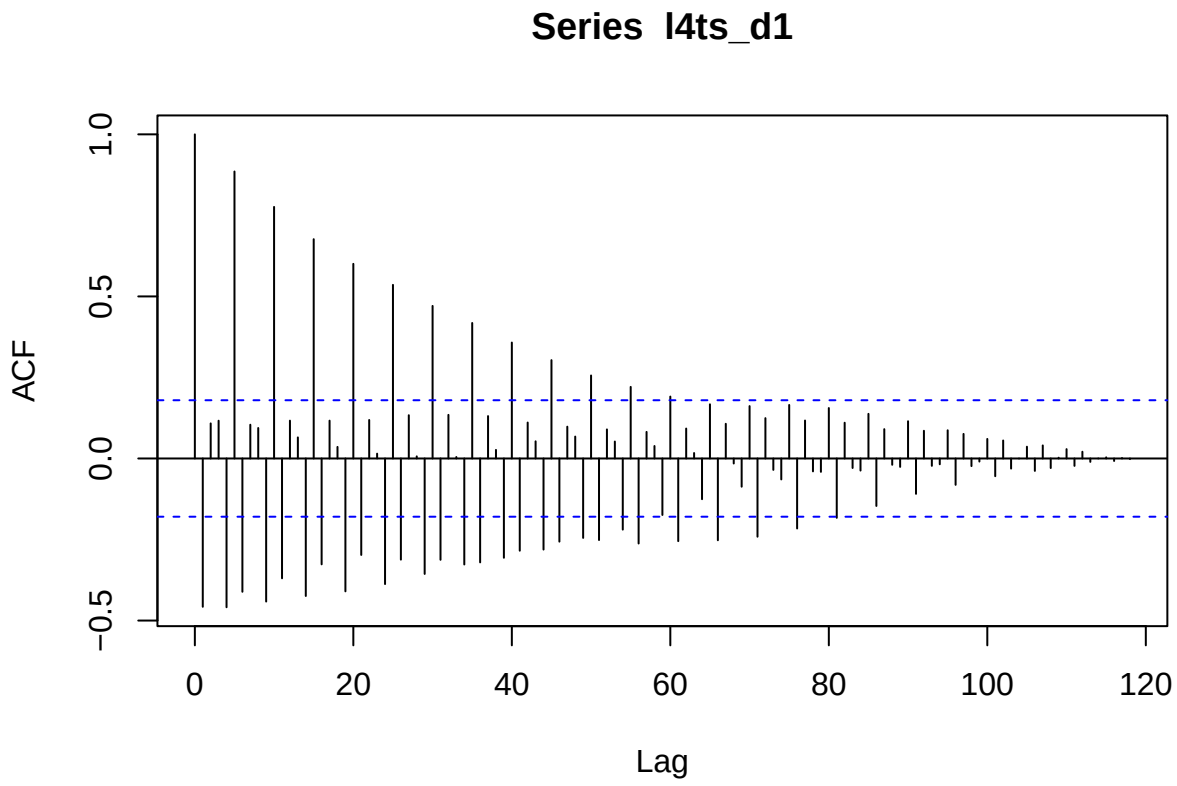
```
# lag   :an integer indicating which lag to use.
# differences   : an integer indicating the order of the difference.
l4ts_d1 = diff(lab4ts, lag = 1, differences = 1)
plot(l4ts_d1, xlab = "x_t - x_{t-1}", main = "Lab 4 Time Series (Difference d = 1)")
```



Now that we've de-trended the data, there appears to be some sort of multiplicative seasonality left in $\{y_t\}$, however, it is already looking much better in terms of a less notable trend.

Let's look at its acf. I prefer a larger `lag.max` in my acf plots.

```
acf(l4ts_d1, lag.max = length(l4ts_d1))
```



As we noted before, the cyclical pattern to the acf may be indicative of some residual seasonality in the model. This is reflected by the periodicity of the correlation values (above and below the zero line) as a function of lag h. We saw in lecture that acfs like this tend to point towards the presence of seasonality. Fortunately for us, however, the magnitude of autocorrelation is in fact decreasing as a funciton of lag as we'd hope.

Question 3

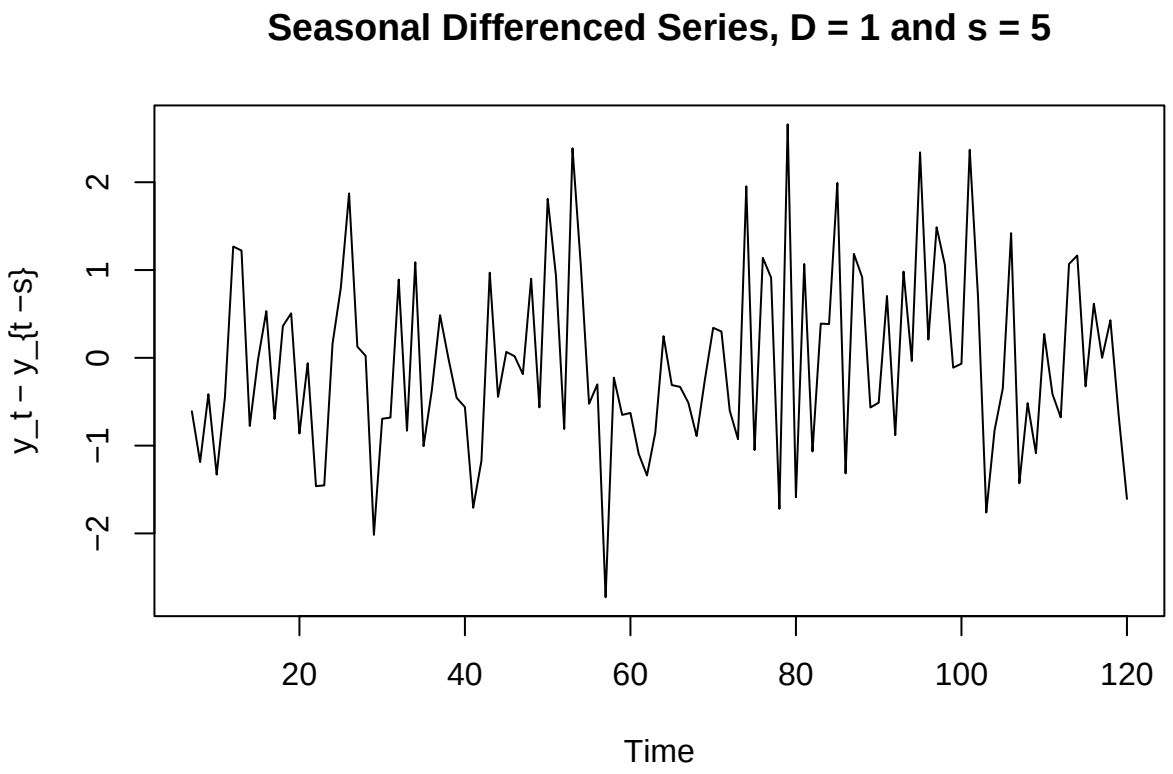
Let's take a seasonal difference of $s = 5$, $D = 1$ for these data. In other words, we will subtract $D = 1$ seasons of magnitude $s = 5$.

In other words, we let ∇ be the difference operator, and we consider:

$$\nabla_s y_t = y_t - \sum_{\delta=1}^{D} B^{\delta s} y_t = y_t - y_{t-s}$$

Let's experiment with some different values of s and inspect the resulting time series and acfs.

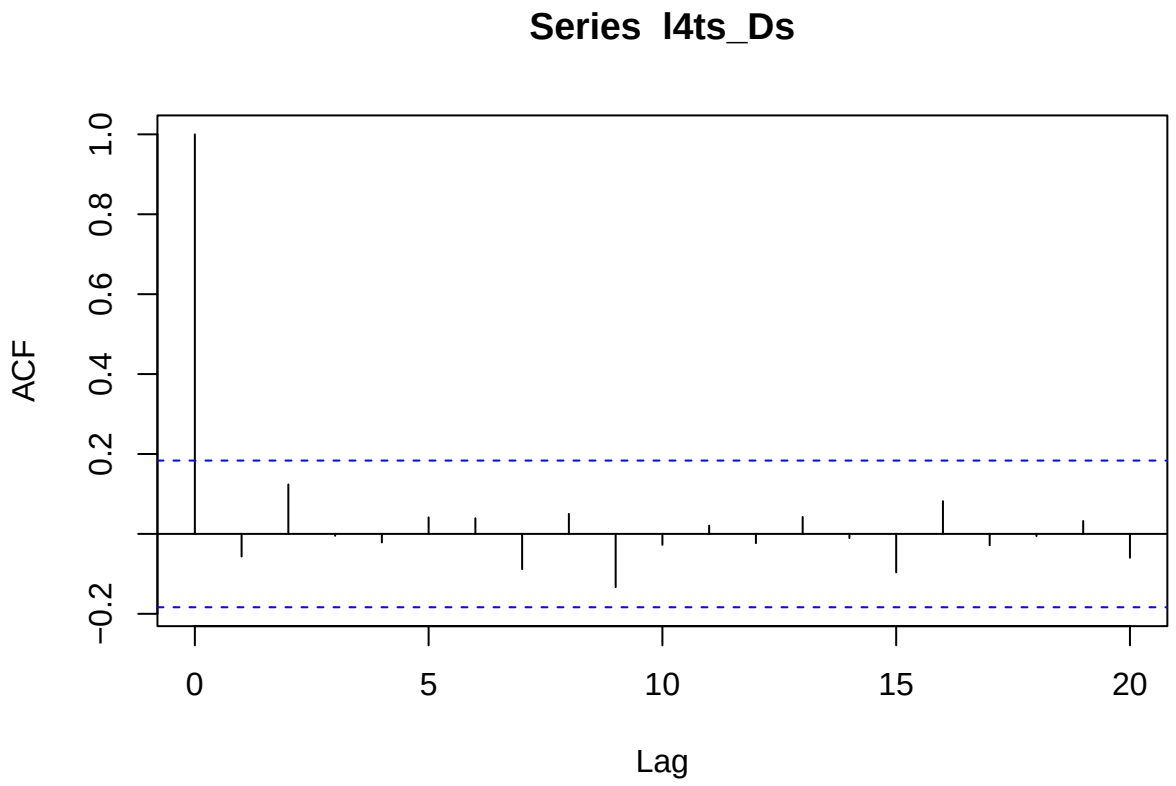
```
s = 5; D = 1
l4ts_Ds = diff(l4ts_d1, lag = s, differences = D)
plot(l4ts_Ds, ylab = "y_t - y_{t -s}",
     main = paste(
       "Seasonal Differenced Series, D =", D,
       "and s =", s))
```



We see this time series is looking more and more like a stationary (or close to stationary) process as we deseasonalize.

Let's inspect the acf.

```
acf(l4ts_Ds)
```



From this acf, we see that our new time series $\{\nabla_5 y_t\}$ resembles white noise. This is evident from the fact that there are no significant autocorrelation coefficients past $\rho(0) = 1$ according to the $\pm 2/\sqrt{n}$ bars (indicated with the dotted blue lines.)

Question 5

What model from the SARIMA family would you use?

Recall that a SARIMA model is of the form:

$$\text{SARIMA}(p, d, q) \times (P, D, Q)_s$$

We've been tracking throughout this work our values of d, D and s. Now, we need to consider from the last acf if there are serial AR(p) or MA(q) components, or, rather, seasonal AR(P^s) or MA(Q^s) components.

From our last question, we're pretty confident we've arrived at white noise. Let's stop for a moment and think about what a white-noise process actually is.

We would have the following form, allowing $\{Y_t'\}$ to be a stand-in for our transformed model $\{\nabla_1 \nabla_{s=5} Y_t\}$ for simplicity.

$$Y_t' = Z_t, \text{ where } Z_t \sim \text{WN}(0, \sigma^2)$$

In terms of an MA(q) process (letting $\beta_0 = 1$), this is...

$$\text{MA}(q) : Y_t' = Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$$

So, from the definition of MA processes, we know that both q and Q are zero. There's no serial MA nor seasonal MA, since we're in a pure white-noise situation.

Similarly, recalling the definition of an AR(p) process,

$$\text{AR}(p) : Z_t = Y'_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p}$$

Again, we see that since we have a model of the form $Z_t = Y'_t$, it must be the case that p and P are both zero.

Hence, we can describe our *original* series $\{X_t\}$ with the following member of the SARIMA family:

$$\text{SARIMA}\underset{p}{(0,}\underset{d}{1,}\underset{q}{0)} \times \underset{P}{(0,}\underset{D}{1,}\underset{Q}{0)}_{\underset{s}{5}}$$

Question 6

(Theoretical exercise) We have seen that removing trends and seasonality can be as simple as differencing successively at different lags. In such cases, suitable models are integrated ARMA (ARIMA) models, or SARIMA if we include seasonal differencing. One difficult aspect of these models is describing them in mathematical terms. Let us try re-using the notation we used with ARMA models to describe these processes.

Part 1

Firstly, we took our original time series X_t and differenced it to obtain $Y_t = X_t - X_{t-1}$ then differenced again at a longer lag to obtain $W_t = Y_t - Y_{t-s}$. Combine these two operations to express our final model Wt in terms of the original series X_t.

$$\begin{aligned}
W_t &= Y_t - Y_{t-s} \\
W_t &= (X_t - X_{t-1}) - \nabla_s (X_t - X_{t-1}) \\
W_t &= X_t - X_{t-1} - X_{t-s} + X_{t-s-1} \\
W_t &= X_t - X_{t-1} - X_{t-5} + X_{t-6} \text{ letting } s = 5
\end{aligned}$$

Part 2

Recall the differencing operator B which takes a series and returns the series at lag 1. For instance, $BX_t = X_{t-1}$. Express the differenced series Y_t in terms of B and X_t.

$$\begin{aligned}
Y_t &= X_t - X_{t-1} \\
Y_t &= B^0 X_t - B^1 X_t \\
Y_t &= (B^0 - B^1) X_t \\
Y_t &= (1 - B) X_t
\end{aligned}$$

Part 3

Taking our un-simplified answer from **Part (a)**

$$
\begin{aligned}
W_t &= X_t - X_{t-1} - X_{t-s} + X_{t-s-1} \\
W_t &= B^0 X_t - B^1 X_t - B^s X_t + B^s B^1 X_t \\
W_t &= (1 - B - B^s + B^{s+1}) X_t
\end{aligned}
$$