PSTAT274 Time Series

Lab Worksheet - Week 7

John Inston

This lab is due at 11:59pm on Wednesday, March 6th, 2024 and should be submitted as a pdf document via Gradescope.

For this section we will be extending our application of time series modelling procedures to seasonal ARIMA (SARIMA) models. We will need to load the `astsa` package using the following chunk.

```
# libraries
library(astsa)
library(forecast)
library(ggplot2)
library(ggfortify)
```

Brief Theoretical Introduction

The seasonal ARIMA (SARIMA) model incorporates the following 4 new parameters to our modelling frame work:

1. P the seasonal autoregressive parameter,
2. D the seasonal differencing,
3. Q the seasonal moving average parameter, and
4. s the seasonal period.

Intuitively, the SARIMA modelling framework incorporates an additional ARIMA model to our existing framework which exclusively deals with seasonal lags determined by s.

For example, a SARIMA(0,0,0)(1,0,0)12 is equivalent to a SMA(1)12 model with the form

$$X_t = Z_t + \Theta_1 Z_{t-12}, \quad Z_t \sim WN(0, \sigma^2).$$

If we consider the form of a SARIMA(0,0,0)(1,1,1)4 model we see that we have a seasonal period of 4 (e.g. quarterly), we have seasonally differenced once, we have one seasonal MA coefficient and one seasonal AR coefficient and so our model written in terms of the backshift operator B takes the form

$$(1 - \Phi_1 B^4)\nabla_4 X_t = (1 + \Theta_1 B^4) Z_t, \quad Z_t \sim WN(0, \sigma^2).$$

The idea is that with seasonal data we expect there to be high correlation between time series values lagged at a certain period. For example sales data with annual seasonality, we would expect last years sales data in November to be most useful for predicting sales next November.

New Parameter Selection

We now briefly discuss how we select these new parameters when fitting SARIMA models to data.

The easiest is the seasonal period s which we determine by looking at the period of the seasonality of the data (i.e. how many observations until the data repeats the pattern). For monthly data with annual seasonality we would choose $s = 12$. For daily data with weekly seasonality we would choose $s = 7$ (data with multiple seasonal trends is beyond the scope of this course).

The next easiest is the seasonal differencing D which much like its counterpart d represents the number of times we had to seasonally difference the data to obtain stationarity.

For the seasonal ARMA orders P and Q we look at the sample ACF and PACF plots and perform a similar analysis as we do for p and q but this time looking for spikes of correlation at lags that are multiples of the seasonality. For example autocorrelation and partial autocorrelation spikes each at lag 12 would suggest the use of $P = Q = 1$.

Problem 1

In this question we shall be analyzing the `AirPassengers` data set from the `astsa` package. Our aim is to determine an appropriate model from the SARIMA family by applying a broadly similar methodology as we have seen thus far in section.

a. Begin by producing a time series plot of the data using the `tsplot()` function. Note your observations about any trends, seasonality and stationarity (is the variance constant?).

b. We see evidence of both a linear trend and a seasonal trend with an annual period (i.e. 12 months) which we decide to investigate further. Use the function `decompose()` to compute the decomposed data and save this to the variable `decomp.data`. Use the `autoplot()` function to produce the decomposition plot.

   Next, use the `ggseasonplot()` from the `forecast` package to produce a seasonality plot which plots the time series over each 12 months and overlays them, allowing us to closely examine the similarities in the behavior of the time series each year.

c. We investigate possible transformations to obtain a stationary time series on which we can consider potential SARIMA models. Compute 3 transformed time series: (1) `log.data` taking the natural log of the data; (2) `dlog.data` taking the difference (lag 1) of the log data; and (3) `ddlog.data` taking the difference (lag 12) of the differenced log data. Create a matrix of all four series called `plot.data` using the `cbind()` function and use `plot.ts()` to produce our combined plot.

d. We continue our data exploration by producing both an ACF and PACF plot for our twice differenced log data using the `acf2()` function. Note your observations about the significant spikes in autocorrelation and partial autocorrelation.

e. Let us first consider a non-seasonal ARMA(1,1) model which we fit using the `sarima()` function (this time fitting the model to the twice differenced log data). What do we observe in the ACF of the residuals?

f. Use your ACF and PACF plot above to suggest possible parameters for the parameters SARIMA(p,d,q)(P,D,Q)s. Consider two potential models and fit each in turn using the `sarima()` function. For each evaluate model performance using the following criteria:

   i. Parameter significance (i.e. does our coefficient lie within one standard error of 0?, if so then it is not significant and you should consider removing it from the model.)

   ii. Residual stationarity (does the plot of the standardized residuals resemble white noise?).

   iii. Residual ACF plot (is there evidence to suggest that our model has failed to capture any autocorrelation?

   iv. Do the standardized residuals appear to follow a Gaussian distribution?

   v. Does the Ljung-Box statistic conclude that there is evidence that any group of autocorrelations of the residuals is different from zero? (i.e. if any p-values are within the significance region)

   vi. Which model achieved the lowest information criterion?

g. Produce a 24 month ahead forecast for our data using the `sarima.for` function. Does our forecast appear reasonable?

h. Below write out the form of our fitted model in terms of the time series X_t and white noise Z_t.