Time Series Forecast Using a Bayesian Approach

Topic:

Monthly passenger totals of a US airline from 1949 to 1960

Ahmed Jamoussi

April 26, 2021

Course: Baysesian Statistics

1 Introduction

Time series data often arise when studying environmental factors and physiological functions, monitoring industrial process and tracking corporate business metrics. Time series data requires adequate analytical and statistical models that account for the fact that the data taken over time has internal structures such as auto-correlation, trend and seasonal variation. Given a time series of data, the Auto-Regressive–Moving-Average (ARMA) model is a powerful tool for understanding and forecasting future values in this series. Despite the advantages that the ARMA model offers, we can still note some limitations. The ARMA time series models are non-standardized and must rely on asymptotic behaviors. However, a Bayesian approach is statistically almost unaffected by these factors. Thus, the objective of this analysis is to study the performance of Bayesian methods in time series data and how it compares to classical time series analysis.
For this project, I am using the *Monthly passenger totals of a US airline from 1949 to 1960* which accounts for 144 data points.

2 ARMA Model

In statistical time series analysis the autoregressive–moving-average ($ARMA(p,q)$) model provides an accurate analysis of (weakly) stationary stochastic processes in terms of two polynomials:

- Autoregression (AR) of order $p \in N$
- Moving Average (MA) of order $q \in N$

Given a time series data X_t the ($ARMA(p,q)$) model is generally introduced as the following:

$$X_t = c + a_t + \sum_{i=1}^{p} \alpha_i X_{t-ik} + \sum_{i=1}^{q} \beta_i a_{t-ik}$$

such that $\forall i \in [p], \alpha_i \in R$ and $\forall i \in [q], \beta_i \in R$. Furthermore, it is important to note the following:

- X_i's stand for the p Auto-regressive terms of the model.
- a_i's stand for the q Moving Average terms of the model.
- k describes the interval of seasonality.
- c is a constant.

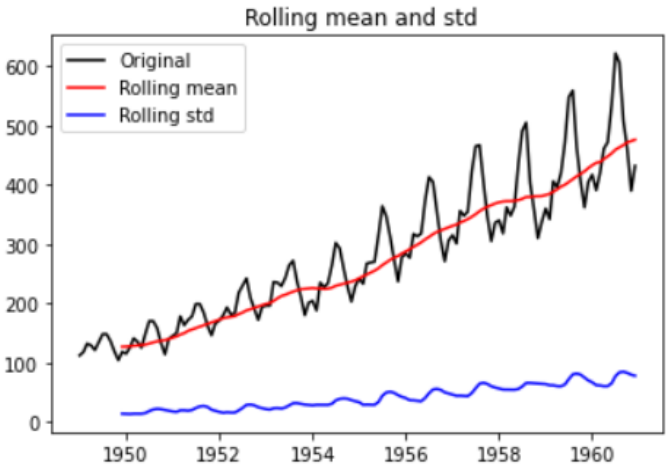
By analysing and investigating the structure of our time series data, we can infer the values of the model parameters p, q and k that best fit the chosen dataset (*Monthly passenger totals of a US airline from 1949 to 1960*). We describe the different steps of the analysis in the following section.

3 Data Analysis

3.1 Stationarity

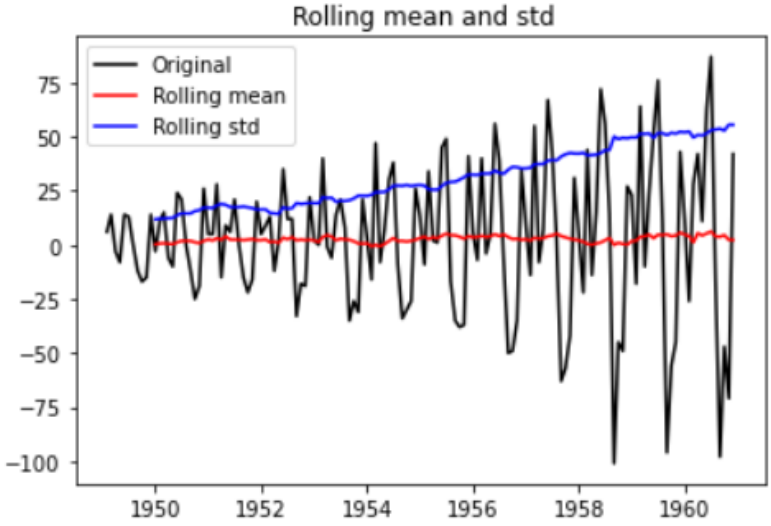
To test the sationarity of the time series we use the following two methods:

- Plotting the rolling mean and standard deviation.
- Performing Dickey-Fuller test and checking if the p-value is below 0.05.




```
Dickey-Fuller test results:
Test Statistic                  0.815369
p-value                         0.991880
Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value 1%               -3.481682
Critical Value 5%               -2.884042
Critical Value 10%              -2.578770
dtype: float64
```

Based on the plotting of the rolling mean and standard deviation seen above, we deduce that the rolling mean and standard deviation are clearly not stationary (upward trend). In addition, by performing the Dickey-Fuller test we obtain a p-value that is significantly greater than 0.05. Thus, we conclude that the dataset is non-stationary. A common approach to deal with non-stationarity of the data is to analyse the difference between two consecutive data points. In other words, we study the differencing of the data instead of the data itself. The next step would be the transformed data set:

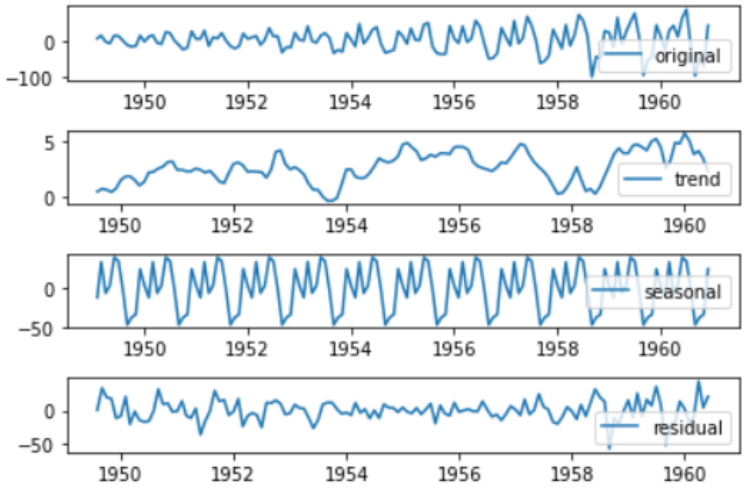



```
Dickey-Fuller test results:
Test Statistic                  -2.829267
p-value                          0.054213
Lags Used                       12.000000
Number of Observations Used    130.000000
Critical Value 1%               -3.481682
Critical Value 5%               -2.884042
Critical Value 10%              -2.578770
dtype: float64
```

Based on the plotting of the rolling mean and standard deviation seen above, we deduce that the rolling mean is relatively constant over time. However, the standard deviation still shows a non stationary behavior by having a slight upward trend. Furthermore, by performing the Dickey-Fuller test we obtain a p-value that is almost equal to 0.05. Despite having a variance that increases over time we can still consider that the data is stationary based on the rolling mean plot and the Dickey-Fuller test results (seen above).

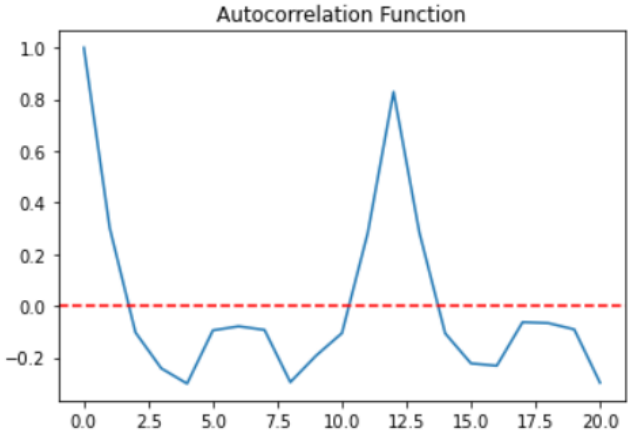
3.2 Seasonality

We now study the seasonality of the time series data by decomposing it into four components by using `statsmodels.tsa.seasonal.seasonal_decompose` tool in python:



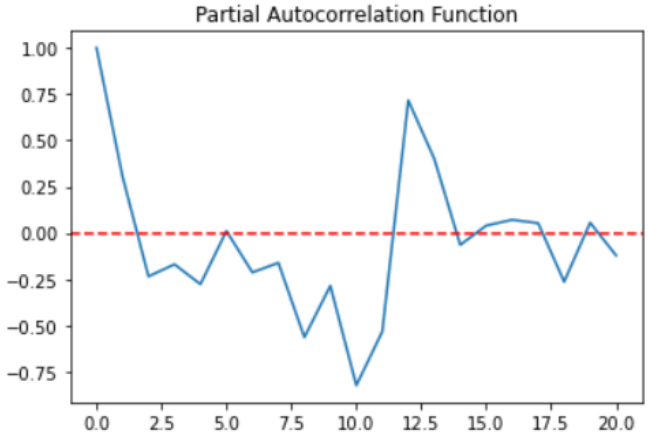
Based on the seasonal decomposition plot seen above we observe a significant seasonality in the data. The data follows a similar pattern every six time steps (6 months). Thus, we conclude that the seasonality interval k = 6.

3.3 Moving Average Parameter



The ACF function equals 0 for the first time around the value of $x = 2$. Thus, we pick $q = 2$ for our model.

3.4 Autoregressive Parameter



The PACF function equals 0 for the first time around the value of $x = 2$. Thus, we pick $p = 2$ for our model.

Following our data investigation we choose the following $ARMA(2, 2)$ model to fit our time series data:

$$X_t = c + a_t + \alpha_1 X_{t-6} + \alpha_2 X_{t-12} + \beta_1 a_{t-6} + \beta_1 a_{t-12}$$

4 Model Fitting & Results

4.1 Classic Approach (Python)

We first implemented and fitted the $ARMA(2, 2)$ model using the training set *(of size 137)* following the classical approach on python by using the function `statsmodels.tsa.arima_model.ARMA`. Next, we used the same fitted model to forecast 6 steps in the future (6 months) and then compare them to the testing set *(of size 6)* by computing the Residual Square Error (RSS):

$$RSS = \sum_{i=1}^{6} (y_i + \hat{y}_i)^2$$

where y_i's are the expected values from the testing set and $\hat{y}_i$'s are the predicted values obtained by our fitted model. Here are the obtained results:

```
Here's the forcasted values:

            #Passengers
Month
1960-07-01    13.677847
1960-08-01    -8.927432
1960-09-01   -26.661613
1960-10-01   -35.305122
1960-11-01   -33.366620
1960-12-01   -22.242712

RSS:  16195.492323785733
```

4.2 Bayesian Approach (WinBUGS)

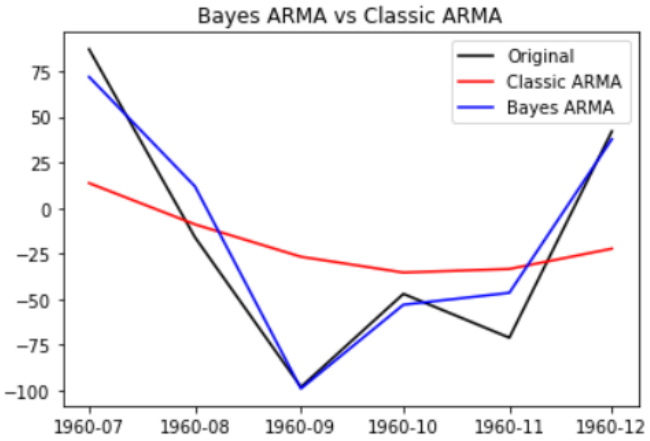
We then implemented and fitted the $ARMA(2,2)$ model using the training set *(of size 137)* following the Bayesian approach on WinBUGS. Next, using the fitted Bayesian model we forecasted 6 steps in future (6 months). We calculated the RSS as described above. Here are the obtained results:

```
Here's the forcasted values:

            #Passengers
Month
1960-07-01        71.88
1960-08-01        11.88
1960-09-01       -98.91
1960-10-01       -52.93
1960-11-01       -46.43
1960-12-01        37.74

RSS:  1663.7343000000003
```

4.3 Comparison: Bayes `ARMA(2,2)` vs Classic `ARMA(2,2)`




```
CLASSIC ARMA MODEL
RSS:  16195.492323785733

BAYESIAN ARMA MODEL
RSS:  1663.7343000000003
```

We clearly note that the Bayesian model generates a very accurate forecasting and has a significantly low RSS at 1663.73 compared to the Classic $ARMA(2,2)$ model that has an RSS of 16195.49. Therefore, we conclude that the Bayesian $ARMA(2,2)$ performs better on the chosen time series.

5 Conclusion

In this project , we have been able to demonstrate that the Bayesian ARMA outperforms the ARMA model that follows a classical approach on our chosen data set. (*Monthly passenger totals of a US airline from 1949 to 1960*). Despite all the advantages that a Bayesian approach provides, we have noted some limitations in our analysis.
First, our conclusion cannot be generalized since it has been tested on a single dataset. Further analysis on different data sets that have different internal structures is required to generalize our statement. Moreover, the performance of the classical ARMA model may have been affected by the small size of the data sets. Additional data is required for the classical ARMA model to perform better.

References

[1] : *https://www.kaggle.com/chirag19/air-passengers*