Math 545 — Data Assignment 2 — Spring 2023

Farhad de Sousa
Department of Mathematics
University of Southern California
fdesousa@usc.edu

1 Introduction & Task 1

In this project, we aim to predict electricity consumption in Arkansas using methods from time series analysis. In particular, we take monthly data from 1990 to 2012 as our initial sample, and use an ARMA process to predict consumption for the period from 2013 to 2015.

In our first data assignment, we performed a classical decomposition of our data (after taking the logarithm), to get a trend and seasonality free data set which we use to build our model. We also calculated local and global polynomial estimates for the season-less data, the latter of which we will use in the construction of our final predictions. The plot below, of the data we start our present assignment with, makes it evident that there is no trend or seasonality left.

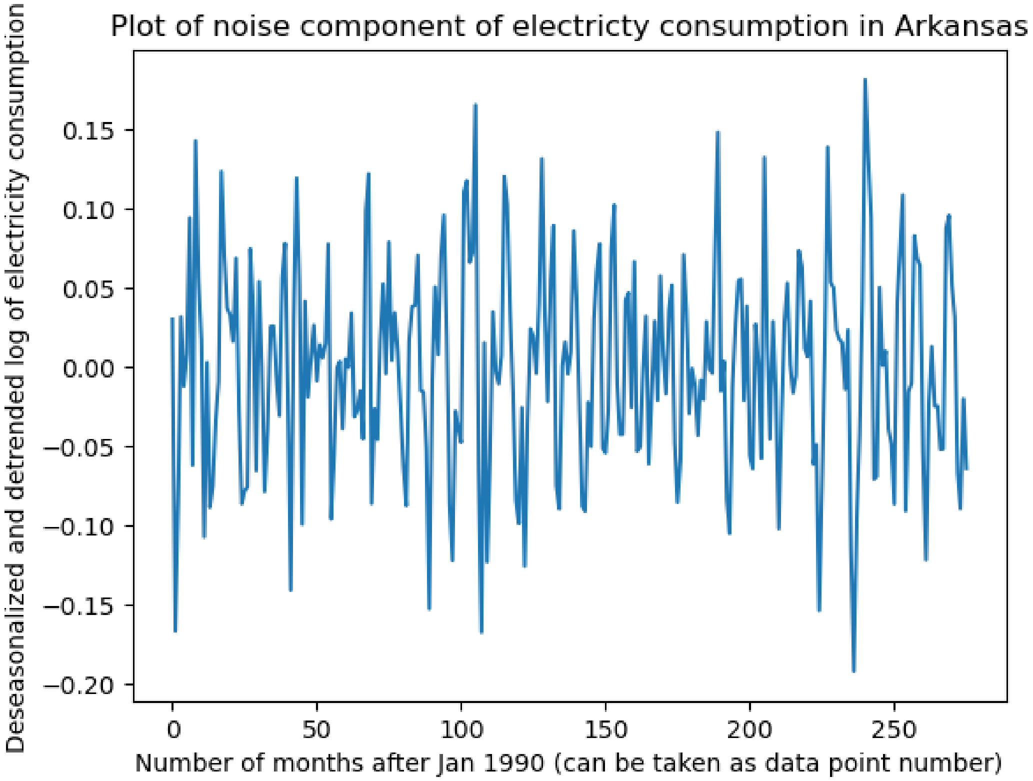

Figure 1: A plot of the logarithm of our original data, with both trend and seasonality removed.

2 Task 2

Our next goal is to find the empirical autocorrelation function (ACF) of the data (up to lag 30), and use it to determine the order of a Moving Average model.

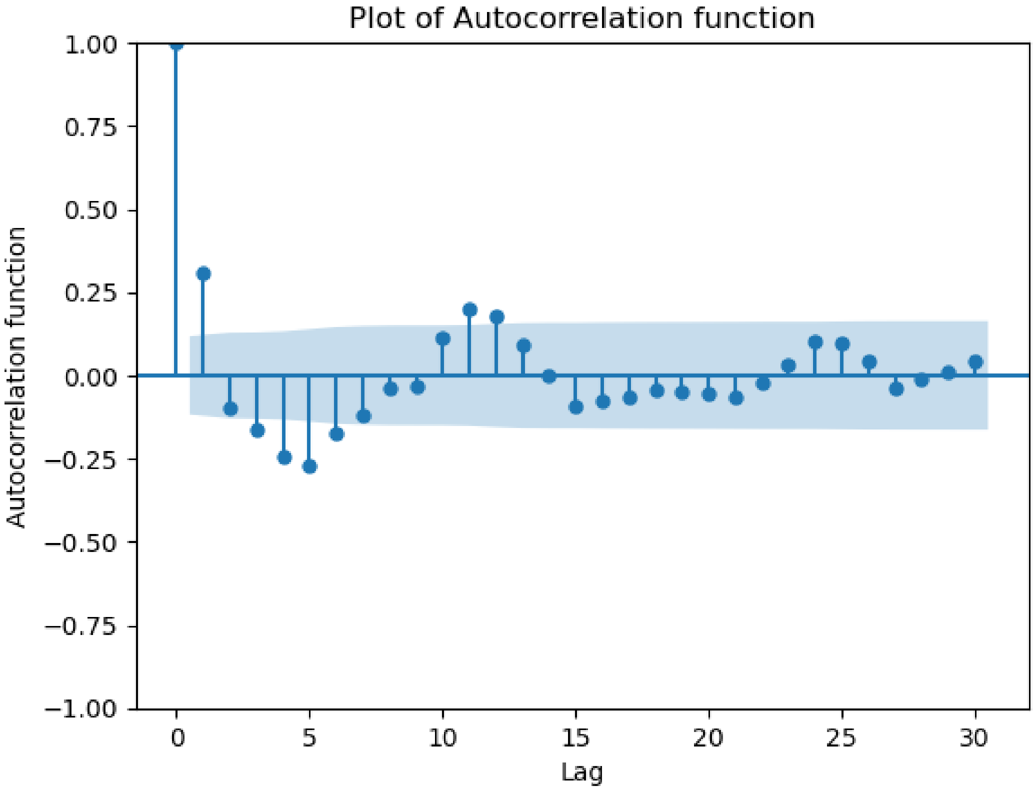

Figure 2: A plot of the empirical autocorrelation function of our data

We see that the values of the empirical ACF stay within the required band for lag 12 onwards. Therefore we choose a **MA(12) model**. Building this model using the ARIMA routine in Python, we see that its Akaike Information Criterion (AIC) is -828.12, and its Bayesian information criterion (BIC) is -777.44.

3 Task 3

We now look at the empirical partial autocorrelation function (PACF), and see that the best Autoregressive model to use is **AR(9)**, with AIC -809.27, and BIC -769.43. Given that the MA(12) has a lower AIC score, we would choose it as our model.

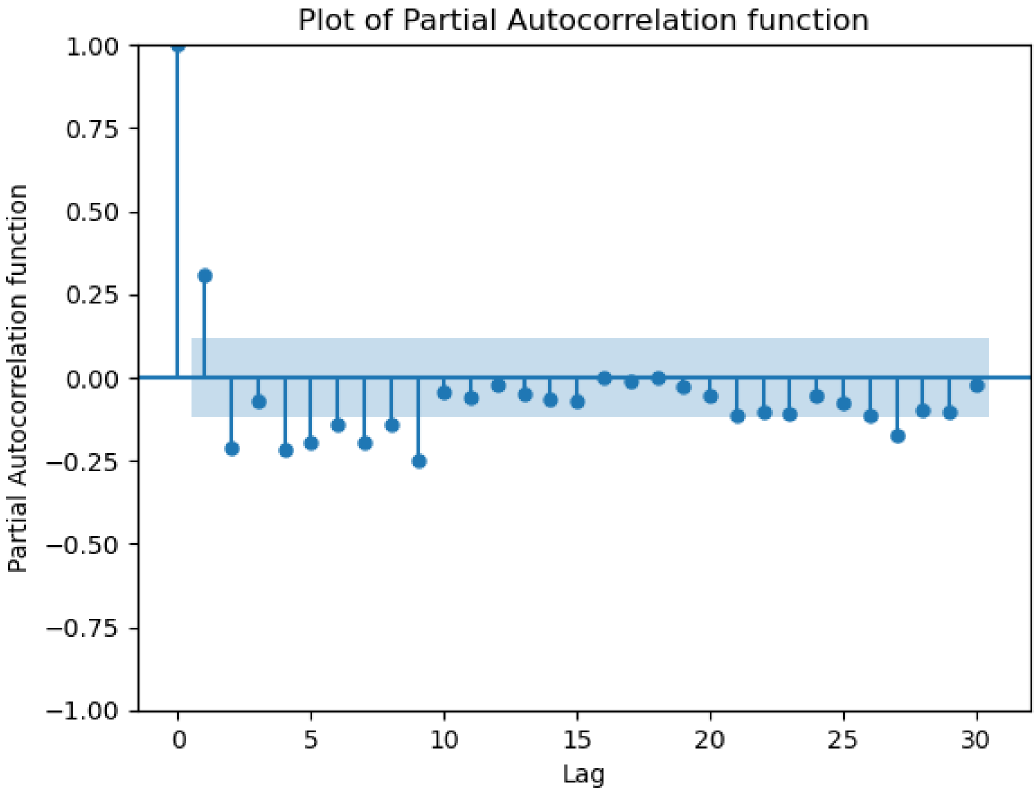

Figure 3: A plot of the empirical partial autocorrelation function of our data

Note that the PACF slightly escapes the bounds at lag = 26, but we ignore this in order to have a more practical and interpretable model.

4 Task 4

We next aim to build our final ARMA(p,q) model. If we followed our graphs, we would choose an upperbound for p + q as 9, as it is the lower of the two values chosen earlier for our pure MA and AR processes. However, following the instructions for the assignment, we use an upper bound of 5. The best values for p and q, when $p + q \leq 5$, are $p^* = 2$ and $q^* = 1$ with an AIC value of -825.13.

However, the code below follows the prompt of the **BONUS** question, and we find a better model is $p^* = 3$ and $q^* = 4$ with an AIC value of -832.96.

```
P_max = 9
AIC_small = ARIMA(train, order = (1,0,1)).fit().aic
param = []
for i in range (1, P_max):
    for j in range(1, P_max - i + 1):
        ARMA_model = ARIMA(train, order = (i,0,j))
        ARMA = ARMA_model.fit()
        AIC = ARMA.aic
        if (AIC < AIC_small):
            AIC_small = AIC
            best_param = [i,j]
            bestARMA = ARMA
```

5 Task 5

We now proceed with a diagnostic evaluation of our ARMA (2,1) model by plotting its normalized residues, and checking if the graph looks like white noise:

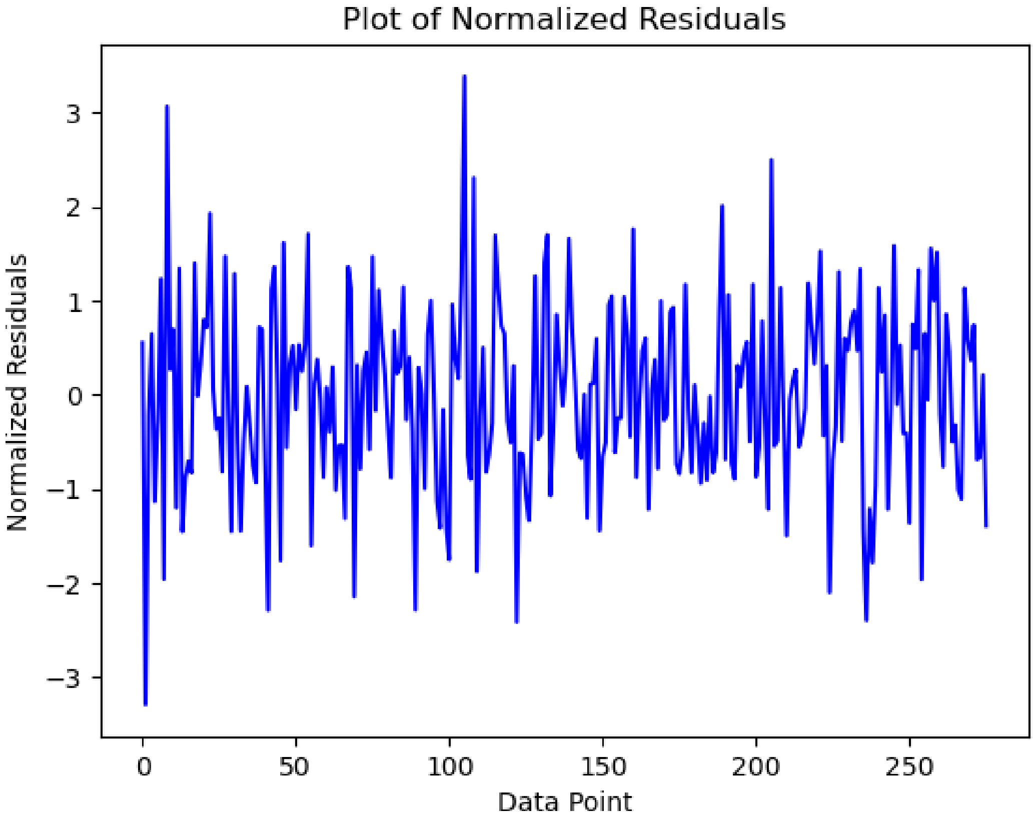

Figure 4: A plot of the normalized residuals for our ARMA(2,1) model

The data seems to be reasonably whitenoise like, apart from a few unusual spikes around months 1, 10, and 110. We plot the autocorrelation function of the normalized residues next to verify that these spikes don't amount to any serious concern:

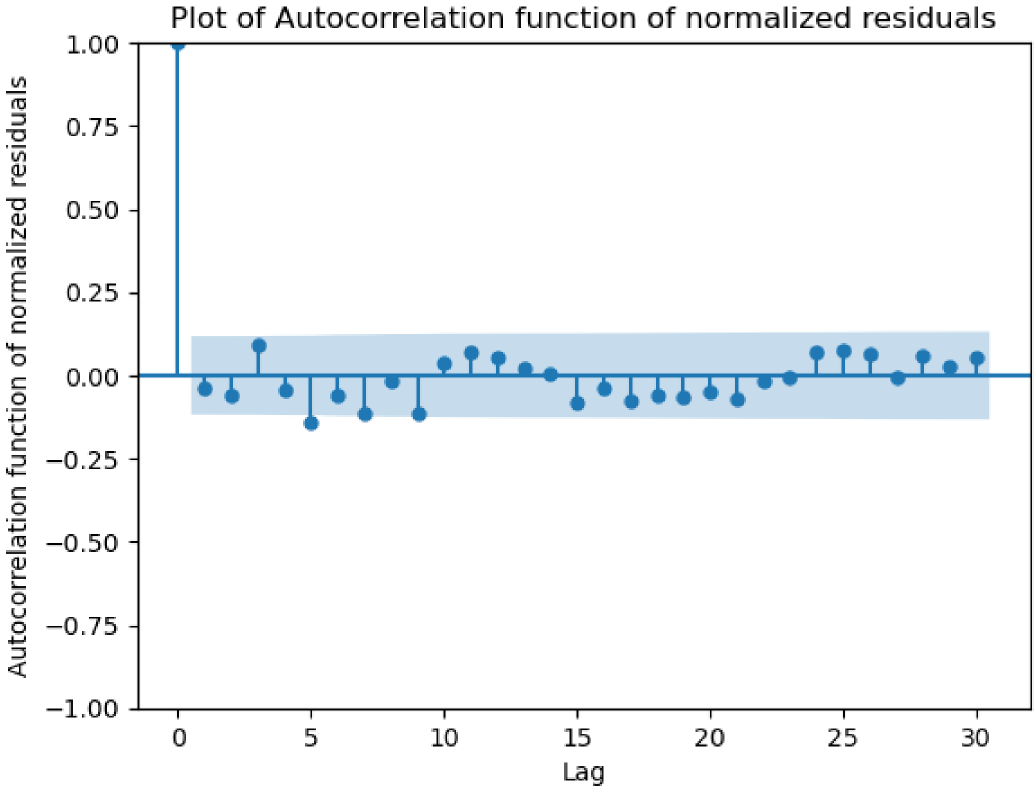

Figure 5: A plot of the empirical autocorrelation function of the normalized residuals for our ARMA(2,1) model

This confirms that the residues are indeed like white noise, as the estimated ACF has all its values within the band for a lag ≥ 1.

6 Task 6

Finally, we use our ARMA (2,1) model to predict values in our stationary seasonless time series for the next 3 years, and then reverse all the transformations we have done so far, to get our last plot:

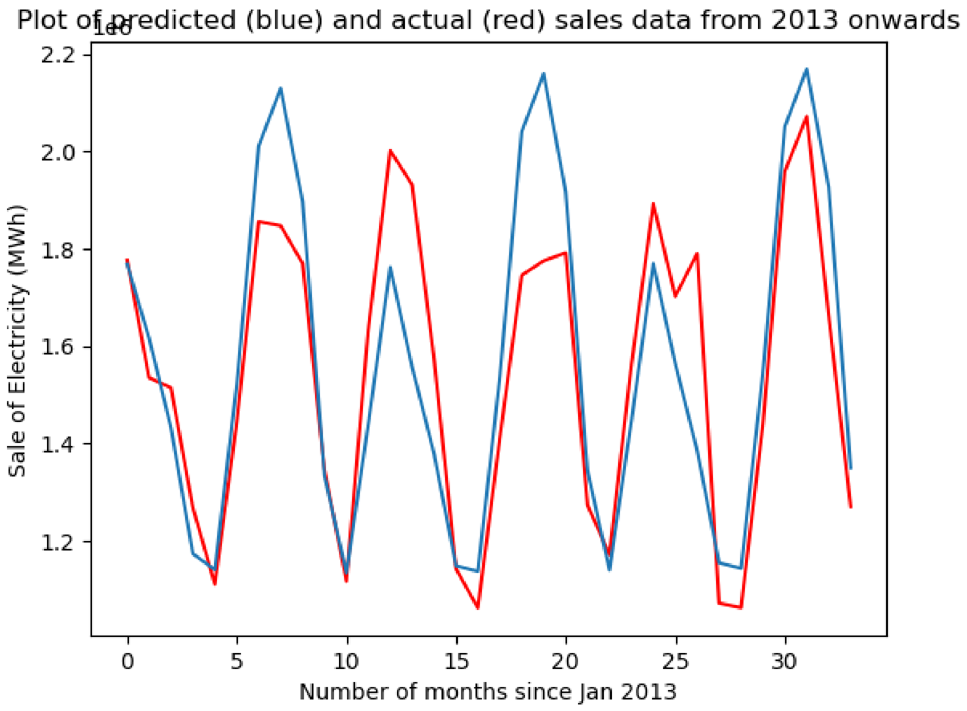

Figure 6: A plot of the actual electricity sales compared to our predictions

The red line is the actual electricity consumption in Arkansas from Jan 2013 to October 2015 (data that our model has not seen before). Our predictions, in the blue, seem to fit the new data reasonably well, apart from overestimating the two peaks at 7 and 18 months, and underestimating the peak at 13 months. The fact that the predictions follow the actual data so closely elsewhere gives us the impression that there was no series error in the building of the model, but perhaps these few months were just especially hard to predict given the data we had. To investigate further, we look at the plot of the entire data set, along with the global polynomial used in our model, as well as a simple 12-term MA:

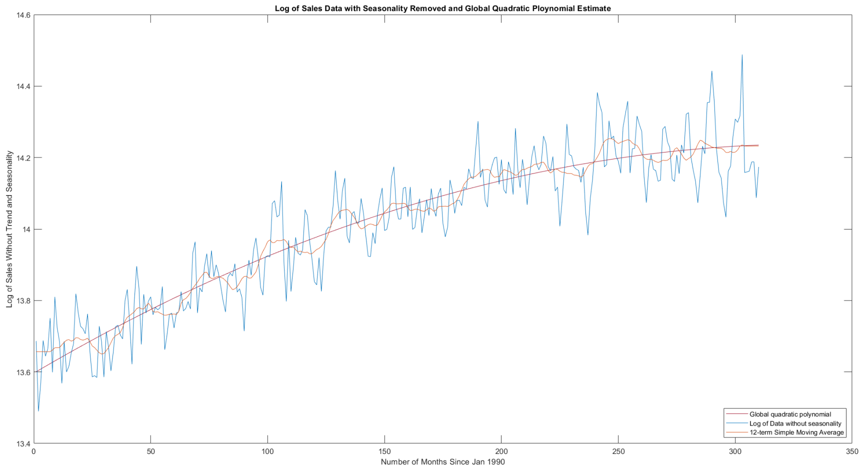

Figure 7: A plot of the log of the actual data, our global polynomial, and a simple MA(12)

We see that, even though we took the log of the data, the variance towards the end is still significantly higher than all the data before. This could mean that our original data is considerably far from the satisfying our stationarity assumption, which could be the reason why we weren't properly able to predict those three peaks.

7 Task 7

The most basic quantitative evaluation of the performance of our model is the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x_t})^2,$$

where the sum is over all the unseen test points. When we evaluate this for our model, we get the enormous number 30,322,101,796 (30 billion), which is understandable, given that our values are of the order 10^6, and squaring them will give values in the billions.

In this situation, a better measure of the performance of our model is the R^2 statistic, which gives us the fraction of variance of our target data that is explained by our model.

$$R^2 = 1 - \frac{\sum(x_t - \hat{x_t})^2}{\sum(x_t - \bar{x_t})^2},$$

In our case, we get $R^2 = 0.673$, which tells us about 67% of the variance in our data is explained by our model. This result, along with the MSE, indicates that there is a lot of room for improvement in our model.

Perhaps a possible path forward is to fit another ARMA(p,q) process to the residues of our current model, and then combine the results of the two to get a better model.