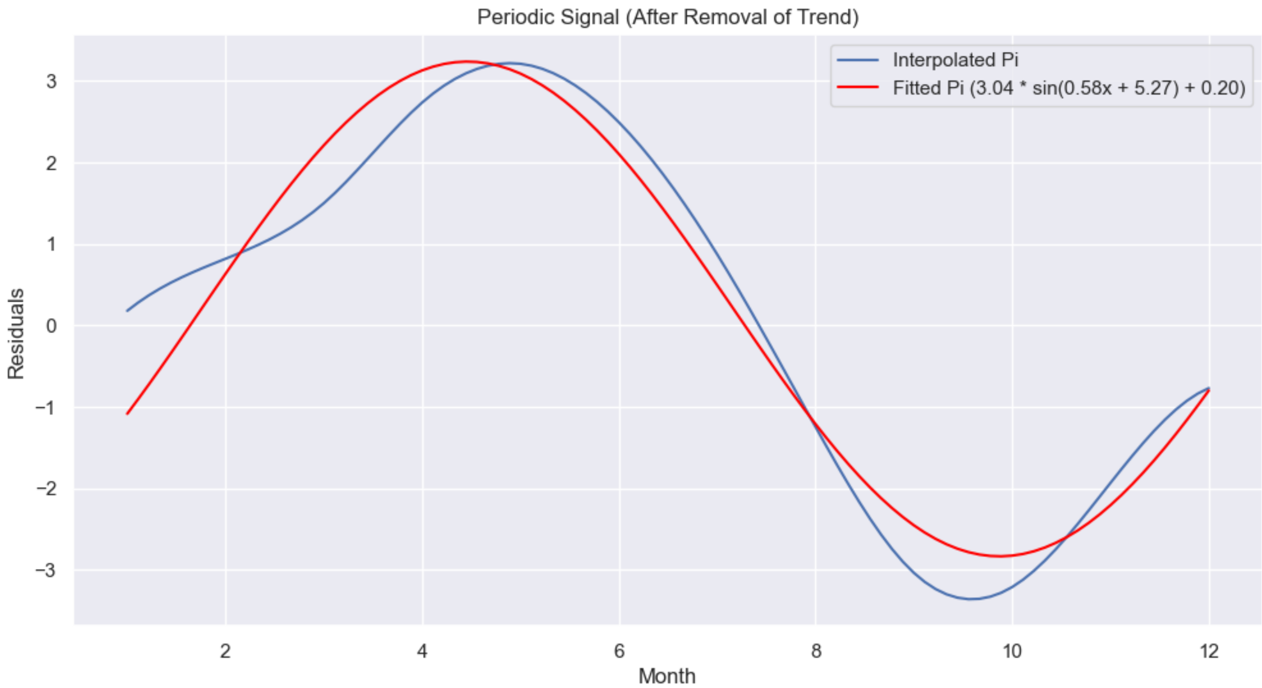
Module 4 - Time Series: Written Analysis, Peer Review and Discussion

Name: eddysanoli

Problem 1: The Mauna Loa CO2 Concentration

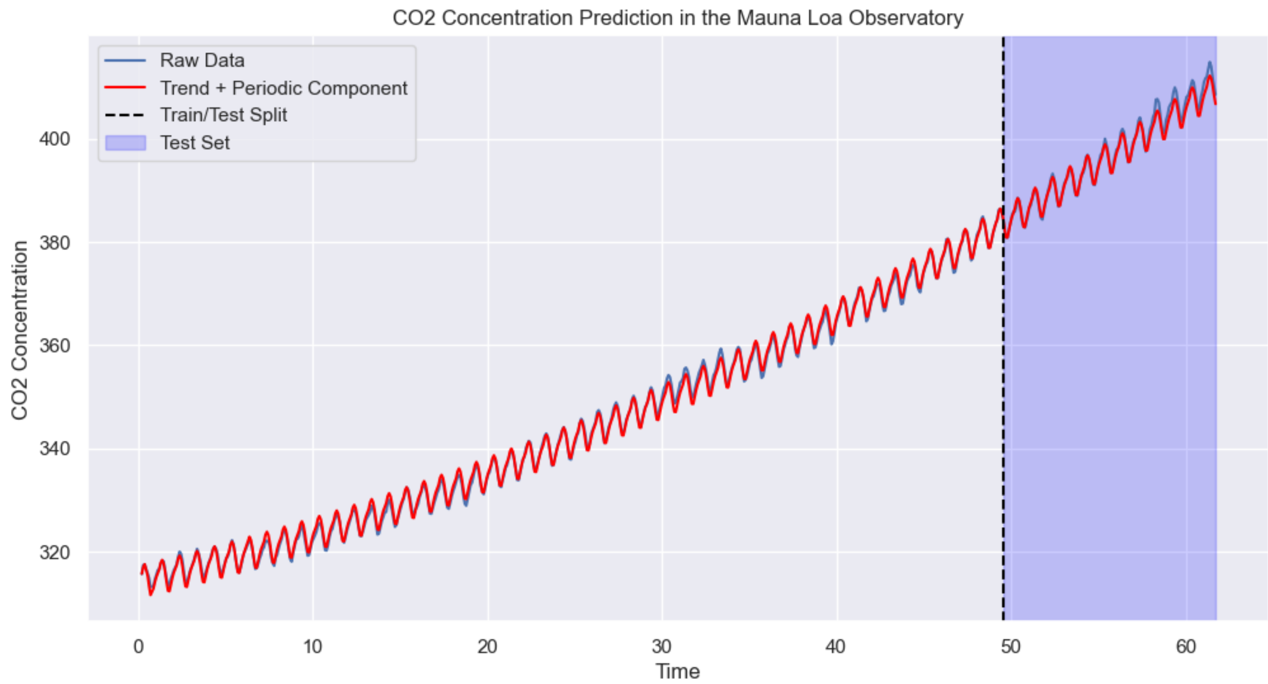
The final model

1. (3 points) Plot the periodic signal Pi. (Your plot should have 1 data point for each month, so 12 in total.) Clearly state the definition the Pi, and make sure your plot is clearly labeled.



   - After interpolating the periodic signal Pi I realized that the periodic signal was really close to a sinusoid, so I did an additional fit of a sinusoid. The resulting signal is shown in red in the plot above.

2. (2 points) Plot the final fit Fn(ti) + Pi. Your plot should clearly show the final model on top of the entire time series, while indicating the split between the training and testing data.



3. (4 points) Report the root mean squared prediction error RMSE and the mean absolute percentage error MAPE with respect to the test set for this final model. Is this an improvement over the previous model Fn(ti) without the periodic signal? (Maximum 200 words.)

   - After calculating the RMSE and MAPE for the model that includes the periodic signal, its evident that there was an improvement on the test set when compared to the model that only includes the trend component. Both the RMSE and MAPE were reduced by approximately 58 and 64 percent respectively. A statistical test should be performed to confirm if the improvement is significant, but the results seem to indicate that the addition of the periodic signal is a good idea.

```
Trend + Periodic Component
RMSE:  1.054729182725852
MAPE:  0.1911171270127314 %

Trend
RMSE:  2.501332219489783
MAPE:  0.5320319129740851 %

Percentage Difference
RMSE:  57.83330280929281 %
MAPE:  64.07788285774494 %
```

4. (3 points) What is the ratio of the range of values of Fn to the amplitude of Pi and the ratio of the amplitude of Pi to the range of the residual Ri (from removing both the trend and the periodic signal)? Is this decomposition of the variation of the CO2 concentration meaningful? (Maximum 200 words.)

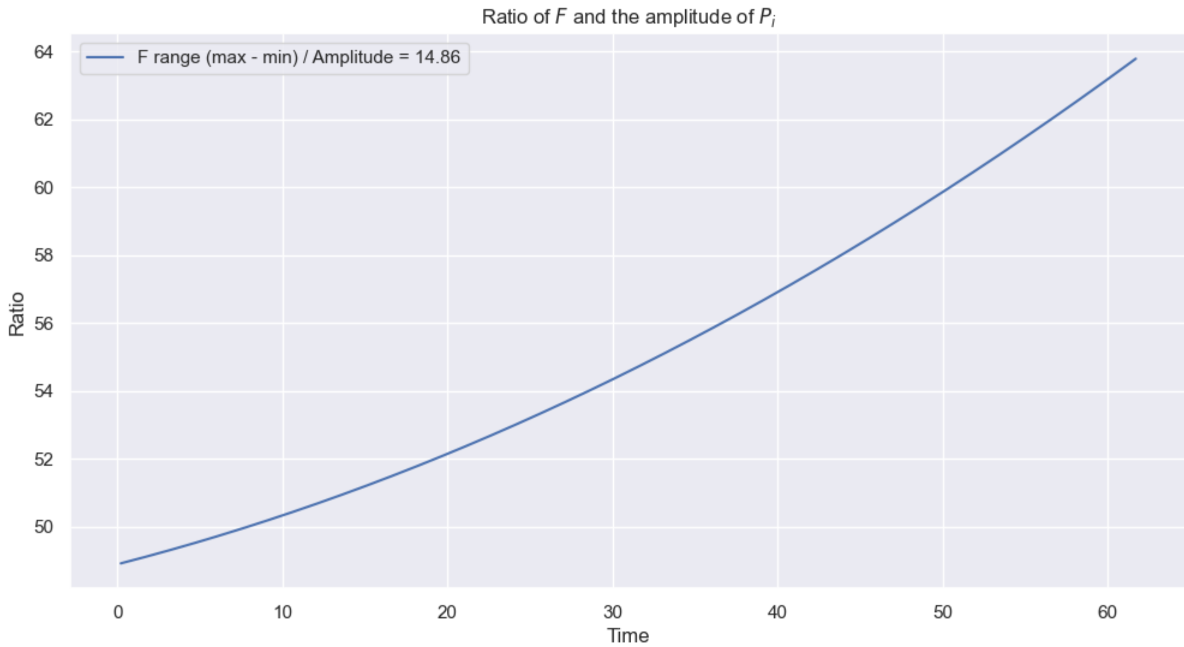
   - According to the problem setup, the decomposition into a trend (F) and a periodic component (Pi) is "meaningful only if range of F is much larger than the amplitude of Pi and this amplitude in turn is substantially larger than that of Ri". When calculating the ratio of the range of F and the amplitude of Pi for the entire dataset, we get a value of ~14.86, which could help confirm that F is much larger than the amplitude of Pi.

```
F range / Amplitude ratio:  14.862612733264376
Amplitude / R range ratio:  1.3807759409578109
```

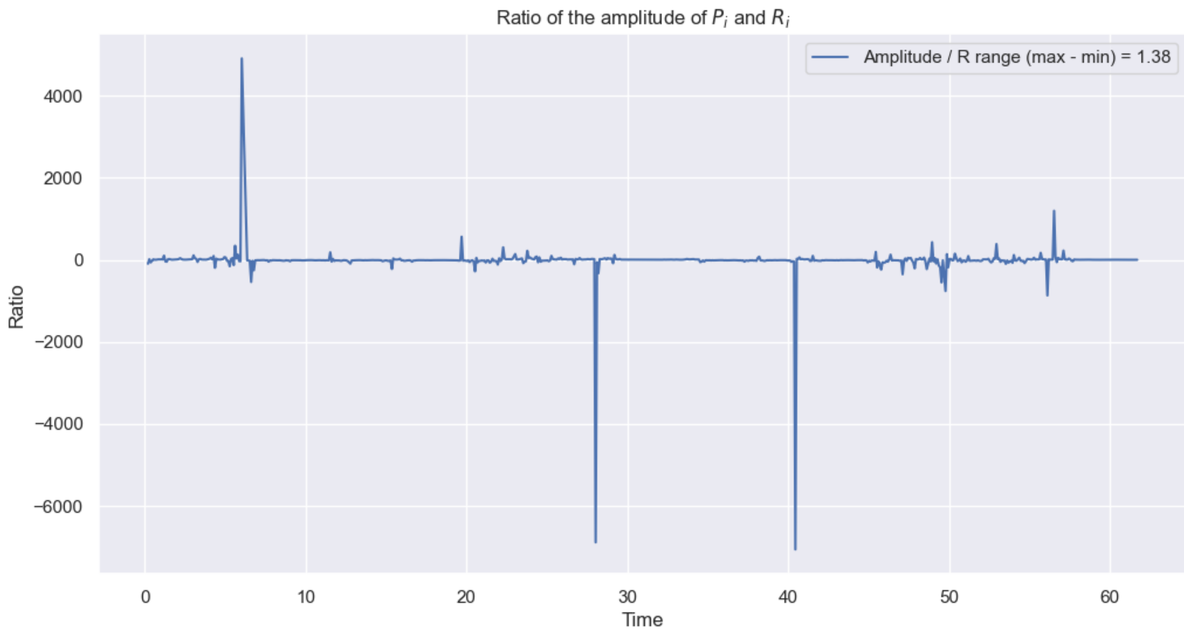
However, this changes depending on the dataset we are considering, as the trend increases in a quadratic fashion. We can actually get an equation for range by subtracting two "y" values from the trend equation (F):

- *range = (A*x1^2 + B*x1 + C) - (A*x2^2 + B*x2 + C)*
- *range = A*(x1^2 - x2^2) + B*(x1 - x2)*
- *range = A*(x1 - x2)(x1 + x2) + B*(x1-x2).*
- *range = (x1 - x2)(A*(x1 + x2) + B)*

Considering that in our case A = 0.012, B = 0.802 and C = 314, we can calculate the range for both the training and test sets. The training set starts at x1 = 0 and ends at x2 = 49.46, making the range ~69. The test set starts at x1 = 49.46 and ends at x2 = 61.71, making the range ~12. We can probably just consider the training dataset since that was the one used to create the model itself. Since the amplitude of Pi is 6.42, the ratio of the range of F and the amplitude of Pi now becomes ~10.8, which could still be considered "much larger".



This is great, but what about the ratio of the amplitude of Pi to the range of the residual Ri? The amplitude of Pi is 6.42, and the range of the residual Ri is 4.65, which gives us a ratio of ~1.4. This is not substantially larger, so it could be argued that the decomposition is not meaningful, or that the residuals still hold some information that could be useful. An additional ARIMA model could be fit to the residuals to see if there is any information left there.



Problem 2: Autovariance Functions

1. (4 points) Find the autocovariance function of the MA(1) model,

   Xt = Wt + theta * W{t-1}

$$X_t = Wt + \theta W_{t-1}$$

Additional considerations:

$$Cov(W_t, W_t) = Var(W_t) = \sigma^2$$

$$Cov(W_t, W_{t-s}) = 0$$

$$Cov(AW_t, AW_t) = A^2 Var(W_t)$$

$$|t-s| = 0$$

$$Cov(X_t, X_t) = Cov(W_t, W_t) + \theta Cov(W_t, W_{t-1}) + \theta Cov(W_{t-1}, W_t) + \theta^2 Cov(W_{t-1}, W_{t-1})$$

$$Cov(X_t, X_t) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1+\theta^2)$$

$$\gamma(0) = \sigma^2(1+\theta^2)$$

$$|t-s| = 1$$

$$Cov(X_t, X_{t-1}) = Cov(W_t, W_{t-1}) + \theta Cov(W_t, W_{t-2}) + \theta Cov(W_{t-1}, W_{t-1}) + \theta^2 Cov(W_{t-1}, W_{t-2})$$

$$Cov(X_t, X_{t-1}) = \theta\sigma^2$$

$$\gamma(1) = \theta\sigma^2$$

$$|t-s| > 1$$

$$\gamma(t) = 0$$

2. (4 points) Find the autocovariance function of the AR(1) model,

   Xt = phi * X{t-1} + Wt,

$$\gamma(0) = Cov(X_t, X_t)$$
$$= Var(X_t)$$
$$\gamma(1) = Cov(X_t, X_{t-1})$$
$$= Cov(\phi X_{t-1} + W_t, X_{t-1})$$
$$= Cov(\phi X_{t-1}, X_{t-1}) + Cov(W_t, X_{t-1})$$
$$= \phi Cov(X_{t-1}, X_{t-1})$$
$$= \phi\gamma(0)$$
$$\gamma(2) = Cov(X_t, X_{t-2})$$
$$= Cov(\phi X_{t-1} + W_t, X_{t-2})$$
$$= Cov(\phi(\phi X_{t-2} + W_t) + W_t, X_{t-2})$$
$$= Cov(\phi^2 X_{t-2} + \phi W_t + W_t, X_{t-2})$$
$$= \phi^2 Cov(X_{t-2}, X_{t-2})$$
$$= \phi^2\gamma(0)$$

Here we can see a pattern emerge. We have an initial value $\gamma(0)$ that consists of the variance of the auto-regressive process plus the variance of the white noise. Then, for each subsequent lag, we have a value that is the auto-regressive coefficient squared times the previous value. So we can write the general formula for the autocovariance function of an AR(1) process in the following way:

$$\gamma(h) = \phi^h\gamma(0) \qquad \text{for h} = 1, 2, 3, \ldots$$

Since we know that $|\phi| < 1$, we can say that the AR(1) model will exponentially decay to zero, as a fraction (ϕ) elevated to an ever increasing exponential will eventually tend to zero.

To get $\gamma(0)$, we need to find the variance $Var(X_0)$. For this, we can use the formula for the variance based on the expectation:

$$Var(X_0) = E[X_0^2] - E[X_0]^2$$

Since we know that X is a stationary process, its mean is assumed to be 0 (also assuming that X_0 = 0 and no drift exists). So we can simplify the formula to:

$$Var(X_0) = E[X_0^2]$$

Now we can just square the AR(1) formula and take the expectation. We can also use the fact that X_0 is independent of W_t to remove the $E[X_0 W_t]$ term from the formula, as well as the fact that the expectation of the square of the white noise is equal to its variance:

$$(X_t)^2 = (\phi X_{t-1} + W_t)^2$$
$$= \phi^2 X_{t-1}^2 + 2\phi X_{t-1}W_t + W_t^2$$
$$E[(X_t)^2] = E[\phi^2 X_{t-1}^2 + 2\phi X_{t-1}W_t + W_t^2]$$
$$= \phi^2 E[X_{t-1}^2] + 2\phi E[X_{t-1}W_t] + E[W_t^2]$$
$$= \phi^2 E[X_{t-1}^2] + E[W_t^2]$$
$$= \phi^2 E[X_{t-1}^2] + \sigma^2$$
$$Var(X_t) = \phi^2 Var(X_t) + \sigma^2$$

Now we just need to solve for $Var(X_t)$:

$$Var(X_t) = \gamma(0) = \frac{\sigma^2}{1-\phi^2}$$

And we can finally get the full autocovariance function:

$$\gamma(h) = \phi^h\frac{\sigma^2}{1-\phi^2} \qquad \text{for h} = 1, 2, 3, \ldots$$

Problem 3: CPI and BER Data Analysis

Converting to Inflation Rates

1. Repeat the model fitting and evaluation procedure from the previous page for the monthly inflation rate computed from CPI.

   Your response should include:

   - (1 point) Description of how you compute the monthly inflation rate from CPI and a plot of the monthly inflation rate. (You may choose to work with log of the CPI.)
   - (2 points) Description of how the data has been detrended and a plot of the detrended data.
   - (3 points) Statement of and justification for the chosen $AR(p)$ model. Include plots and reasoning.
   - (3 points) Description of the final model; computation and plots of the 1 month-ahead forecasts for the validation data. In your plot, overlay predictions on top of the data.

2. (3 points) Which $AR(p)$ model gives the best predictions? Include a plot of the $RMSE$ against different lags p for the model.

3. (3 points) Overlay your estimates of monthly inflation rates and plot them on the same graph to compare. (There should be 3 lines, one for each datasets, plus the prediction, over time from September 2013 onward.)