ATS655 HW3

Tyler Barbero

February 2022

0. Time Management

Code here: Link
Estimate of Time to Completion: 10 hrs
Maximum Allotted Time to Completion: 20 hrs
Actual Time to Completion: 13 hrs
Collaborators: James Larson, Andrey Marsavin, Anindita Chakraborty, En Li, Amanda Bowen, Spencer Jones
Latex used: Yes

Problem 1: Influence of autocorrelation on basic statistics

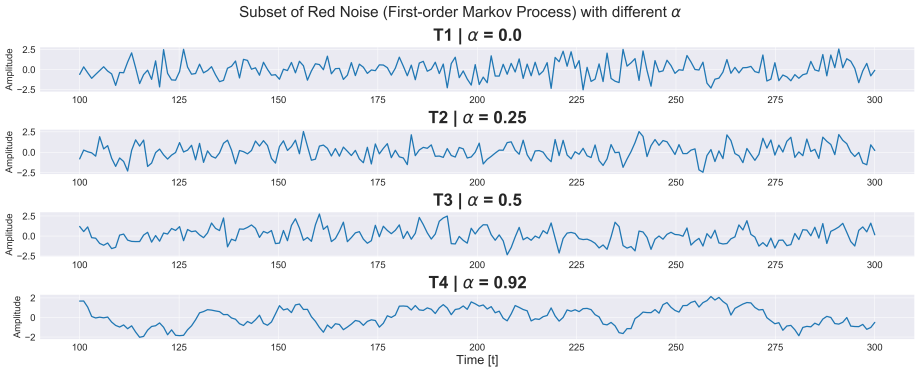

Figure 1: Subset of Red-Noise Time Series

a. Do bootstrapping then plot histograms of the sample means and standard deviations for each of the time series.
b. Discuss your results (1-2 paragraphs).
Looking at the (left) plot in 1a, we see the histogram of bootstrapped sample

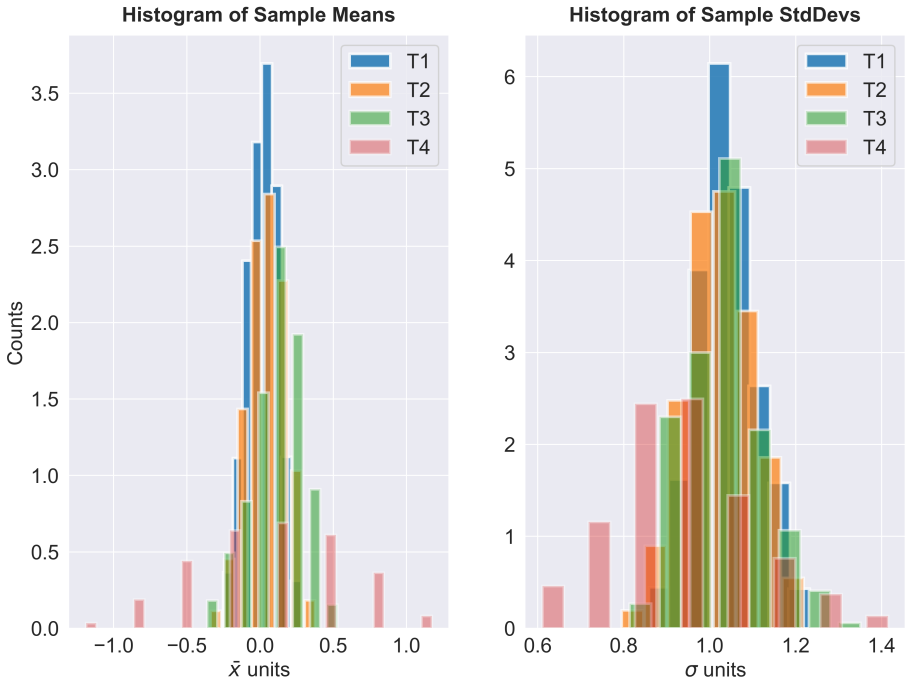

Figure 2: Plot of histograms of sample means and standard deviations for each time series.

means looks to become less Gaussian with larger autocorrelation value (with more memory). The blue histogram is essentially what we've been doing before, sampling from a random distribution assuming all values are independent from each other. However the other histograms (i.e., pink) incorporate autocorrelation (α), that is some of the data is dependent on each other. We see that for the same sample size N=100, the pink histogram with α=0.92 is far less Gaussian than blue with α=0. This is because with autocorrelation, there is more persistence (memory) of the data values with each other. Therefore when we take a consecutive sample of size N=100 and some nonzero α, we are sampling a more specific area (some of these N=100 may be repeats) of the underlying distribution than if we were truly sampling the entire distribution randomly (blue histogram). The result is a lower effective sample size, leading to slower convergence to a Gaussian via the Central Limit Theorem, even when $N = 100 > 30$.

The net result is similar for the standard deviation plots. With the blue histogram, it is the most chi-squared distributed while the others with persistence ($\alpha > 0$) are slower to convergence to a chi-squared. The pink histogram ($\alpha = 0.92$) seems to be shifted to the left a little and a little Gaussian as well. I think this is because

c. Return to Problem 1c of Homework 2. If you got anything incorrect on this problem, discuss in detail what was incorrect and why. If you got everything correct, just write "I did the problem correctly."
I did the problem correctly.
d. Returning to Problem 1c of Homework 2, re-do the calculations with your results from (a) and (b) of this homework in mind. Do your conclusions change?
When doing the bootstrapping I incorporate memory into the process. To do this I count all the times where we have consecutive instances of rain. Then I average these counts, and find the average length of a run to be 3 hours (hourly data). Then I take the mean of a 3-hour consecutive run (randomly grabbed), and do this N=384 times. Then I average these 384 values, where this is one sample. I repeat this experiment 5000 times and plot the distribution of the sample means. I find that the average pressure when it rains is actually within the 95 % confidence bounds of the sample means with memory included. Ultimately, I find that incorporating memory has an impact on our solution and we cannot reject the null hypothesis.

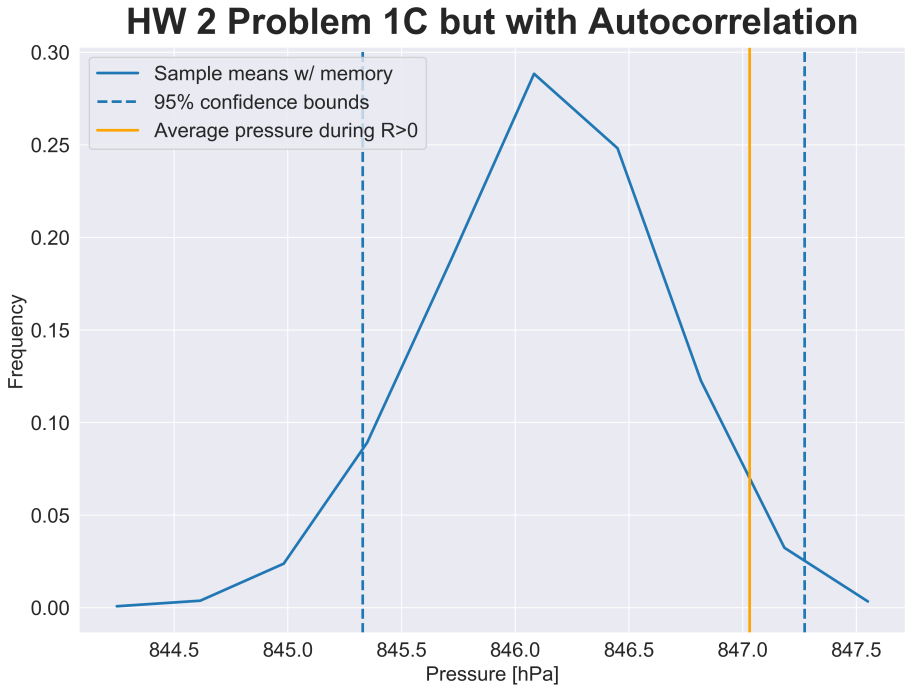

Figure 3: Plot of redo of HW2 1c but incorporating the concept of memory into the bootstrapping.

Problem 2: Correlations vs Composites

a. Plot a scatter plot of values of Y (y-axis) vs. X (x-axis).

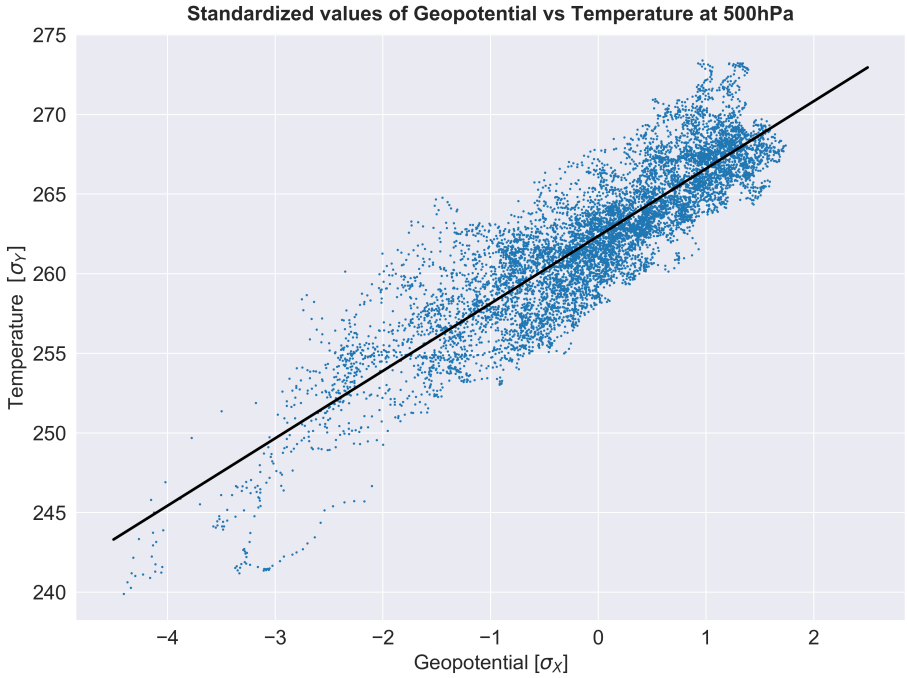

Figure 4: Scatter plot of standardized data.

b. Calculate the correlation and regression coefficients between Y and standardized values of X (the regression should be in units of Y per unit standard deviation of X). Calculate the fraction of the variance in Y linearly explained by X.

$r = 0.8843, a = 4.2348, R^2 = 0.782$

c. Estimate the statistical significance of the correlation you compute in (b) - be sure to consider auto- correlation/memory when determining your degrees of freedom.

I calculated my effective sample size $N^* = 4$ by plotting the autocorrelation of both standardized variables and choosing the larger value where ρ approximately hit zero, which was a value of 2000 hrs time lag. I divided the length of the time series by this and effective N equals 4. To calculate the t-statistic, I used

$$t = r * \sqrt{N^* - 2}/\sqrt{1 - r^2}$$

and $\nu = N^* - 2 = 2$ so the $t_{crit} = 2.92$. I got that t-statistic = 2.93, which is statistically significant albeit very close to the t_{crit} value.

d. Using composite analysis, calculate the mean (or median) Y value

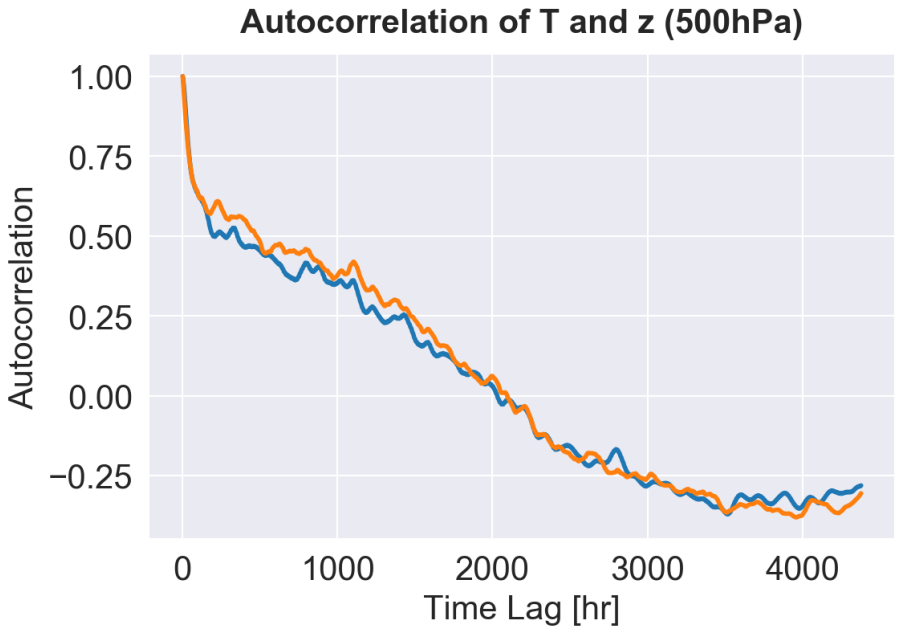

Figure 5: Autocorrelation of temperature and geopotential.

for samples when X exceeds +1 standard deviation. Repeat for when X is smaller than -1 standard deviation. Can you estimate the statistical significance of both values? If yes, do it. If not, why not?
I got that the t-statistic is 1.104 for Y values greater than +1 standard deviation and -0.220 for Y values less than -1 standard deviation. Neither of these surpass the t_{crit} values and are not statistically significant.
d. Provide a brief (1-2 paragraphs) discussion about the similarities and differences between the results from the composite analysis and the regression/correlation analysis. Include in your discussion specific reasons why they don't provide identical results and what these two analyses tell you about physical relationships between X and Y.
My null is that there is no memory in the data. I find that when I incorporate memory through an effective N, the t-statistic is statistically significant therefore we can reject the Null. When I don't incorporate memory (problem 2d), I find that there is t-statistic is not statistically significant therefore we cannot reject the Null. These do not provide identical results because my data are pressure and temperature time series, and we know that these geophysical system do have memory. Typically these last on the order of a few days to a week and thus influence the variables I am using. Thus the assumption (null) that there is no memory is a bad assumption, therefore the t-statistic values in the two analyses are bound to be different. These analyses also tell me temperature and pressure are highly positively correlated and as one go up, the other goes up.

Problem 3: Cost Functions and Regression

a. Calculate the RMSE and MAE for this record of forecasts. How do they compare?
I found the RSME = 18.9 kts (assuming intensity measured in kts, data looks to be reasonable) and MAE = 13.9 kts. The RSME has a higher error, probably due to the square emphasizing outliers in the data.
b. Use Ordinary Least Squares (OLS) linear regression to determine your best-guess hedged forecast v based on the forecast-verification pairs in the NHC dataset. That is

$$\nu = \alpha_o + \alpha_1 \nu*$$

- Calculated using scipy.stats.linregress and I got $a_0 = 0.60$ and $a_1 = 22.7$
(i) How do you interpret the regression coefficients and what do the values say about NHC forecasts?
- α_1 is the slope that best fits the given data and α_0 is the y-intercept of the data.
(ii) What is the RMSE of the predictions made from this OLS best-fit line?
- I take the difference of RHS (truth) and LHS (prediction of best fit line).
- $RSME_{OLS} = 18.6$ kts
(iii) What is the MAE of the predictions made from this OLS best-fit line?
- $MAE_{OLS} = 14.0$ kts
c. Least Absolute Deviations (LAD) regression
(i) What is the RMSE of the predictions made from the LAD best-fit line?
- $RSME_{LAD} = 18.87$ kts
(ii) What is the MAE of the predictions made from the LAD best-fit line?
- $MAE_{LAD} = 13.90$ kts
(iii) Make a scatterplot of the forecasts and observations from the csv file and plot the OLS and LAD regression lines on the same plot. How do they differ?
- We see that the LAD best-fit line has a high slope than the OLS best-fit line and an intercept closer to 0.

d. Add these pairs of verifications/forecasts to your dataset and re-calculate the best fit OLS line. The best-fit LAD line with this "new" hurricane included is a_1 = 1.0000 and a_0 = 5.484e(6). What can you conclude about the sensitivity of each approach to outliers?
Just by looking at the plots it looks like OLS has a higher sensitivity to outliers. This makes sense because we're taking the square of the errors.

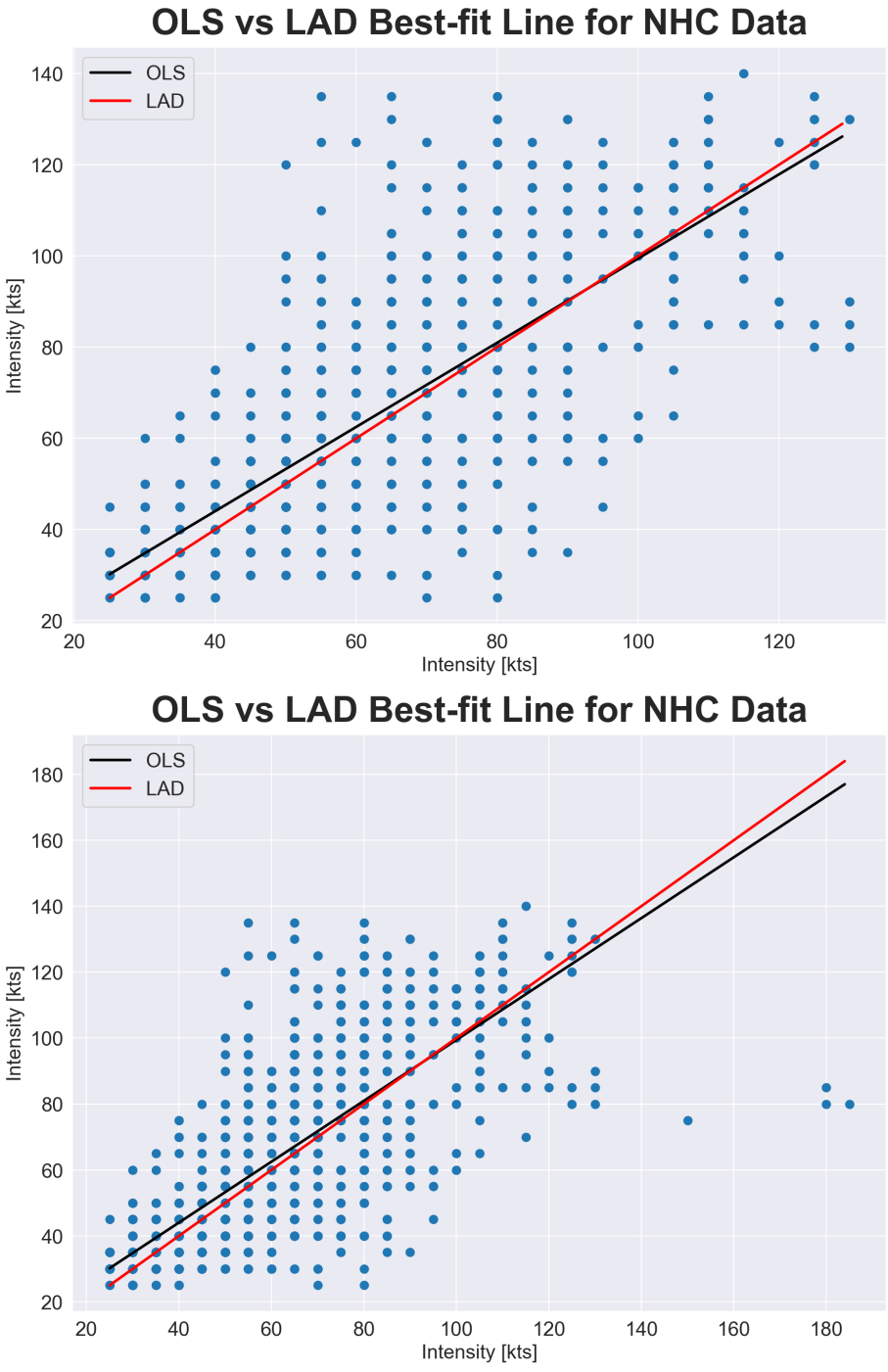

Figure 6: Scatter plot of NHC Hurricane Intensity Data overlaid with OLS and LAD best-fit lines (upper) and same but incorporating anomaly data (lower).