Sullivan Change Point Project

IMSE 641: Quality Engineering

Kansas State University

Dr. Chang

By: Kiel Johnson

3/6/2020

**I.** **Sullivan Data Validation**

The Sullivan change point method paper provided had an example data set and analysis with 60 values. These values were created and imported into an excel and eventually a text file in order to utilize them within python. The *Ruptures* algorithm was utilized and modified in order to accommodate for the Sullivan data and analysis. See the code utilized below:

#modified ruptures by KNJ

import matplotlib.pyplot as plt

import ruptures as rpt

## read data from Sullivan\_xi excel file that has the data from the paper

n\_samples, dim, sigma = 60, 2, 1.23 #dim was 3, sigma assumed as 1.23 per the paper

with open('Sullivan\_xi.txt') as newfile:

signal = newfile.read()

n\_bkps = 4 # number of breakpoints to find before stopping

signal, bkps = rpt.pw\_constant(n\_samples, dim, n\_bkps, noise\_std=sigma)

# detection

algo = rpt.Pelt(model="rbf",min\_size=2, jump=1).fit(signal) #added 'min\_size=2, jump=1'

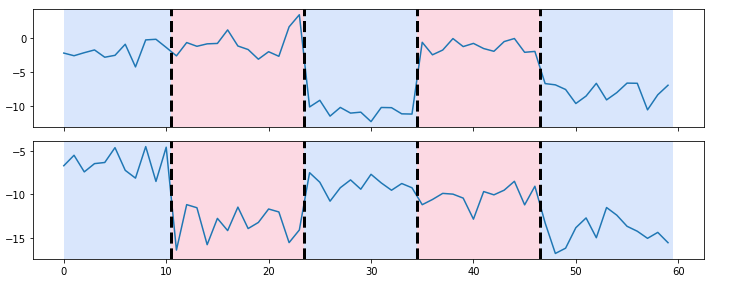
result = algo.predict(pen=1) #pen was = 1, changed due to smaller data values

# display

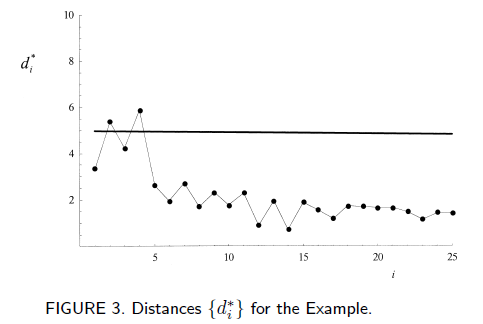
rpt.display(signal, bkps, result)

plt.show()

The code was modified to import the Sullivan data from a text file. The number of samples, dimensions and sigma were all modified to correlate to the Sullivan paper. The sigma was calculated in the paper and utilized in the code above. Two dimensions were utilized and the graphs that were output are below.



Four change points were identified just as the example in the Sullivan paper. I feel that my data output is still not correlating with the Sullivan paper with regard to results though. See the Sullivan graph below:



Four change points were identified with the modified ruptures code and with the Sullivan paper. The outputs are different and a lot of effort was put into trying to understand and modify the ruptures algorithm. The model, min\_size, jump and penalty values were all modified multiple times in order to try and create the same output as the Sullivan paper. The model is very sensitive to alterations in these values and an iterative process was utilized in order to establish a similar result. Ruptures utilizes a trimming method to determine the A further understanding of python and more time and experience would hopefully enable me to be able to tailor the code to provide exact results when compared to Sullivan.

**II.** **Utilize A5 data set**

A similar modification to the Ruptures algorithm was utilized for validating the A5 data set provided. The code is listed below:

#modified ruptures by KNJ

import matplotlib.pyplot as plt

import ruptures as rpt

## read data from A5 exercise

n\_samples, dim, sigma = 40, 2, 1 #dim was 3, sigma assumed as 1

with open('A5 Exercise\_test641data.txt') as newfile:

signal = newfile.read()

n\_bkps = 1 # number of breakpoints to find before stopping

signal, bkps = rpt.pw\_constant(n\_samples, dim, n\_bkps, noise\_std=sigma)

# detection

algo = rpt.Pelt(model="rbf",min\_size=1, jump=1).fit(signal) #added 'min\_size=1, jump=1'

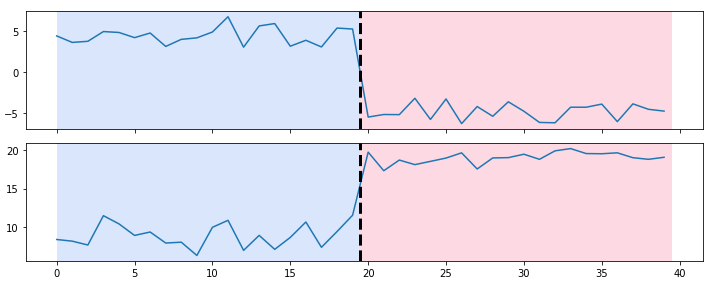
result = algo.predict(pen=1) #pen was = 1, changed due to smaller data values

# display

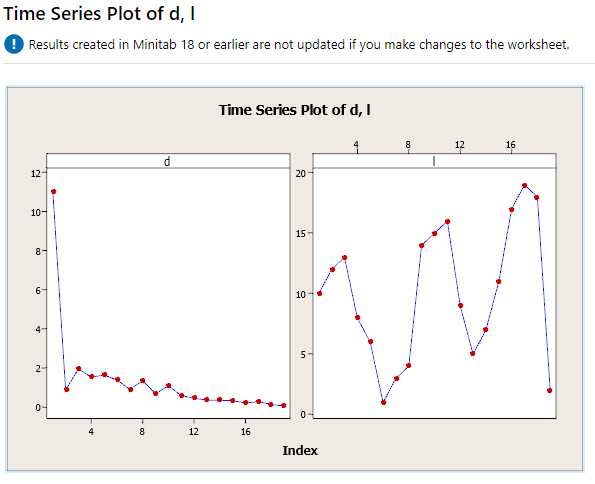
rpt.display(signal, bkps, result)

plt.show()

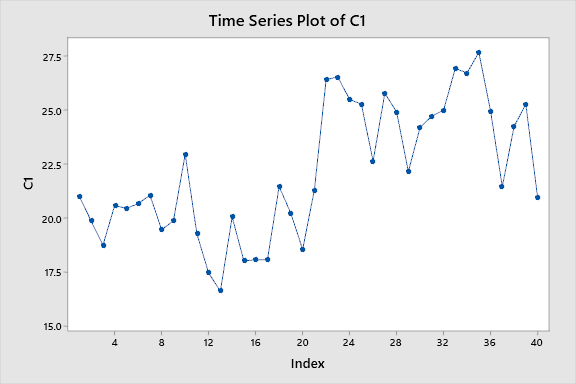
As with the Sullivan example, modifications were made to the code. The sample size was changed to 40, dimension to 2 and the sigma was modified to 1 for a starting point. The input from the A5 text file was utilized for the signal in the form of a text file. The number of breakpoints identified was 1, similar to the Minitab information provided. The min\_size was modified to 1 and the jump value was set at 1 to provided a more accurate answer. If the data set was much larger, a larger jump value would have been utilized. Finally, the penalty value was modified to 1 after an iterative process to determine the best case. The final resulting graphs are listed below:



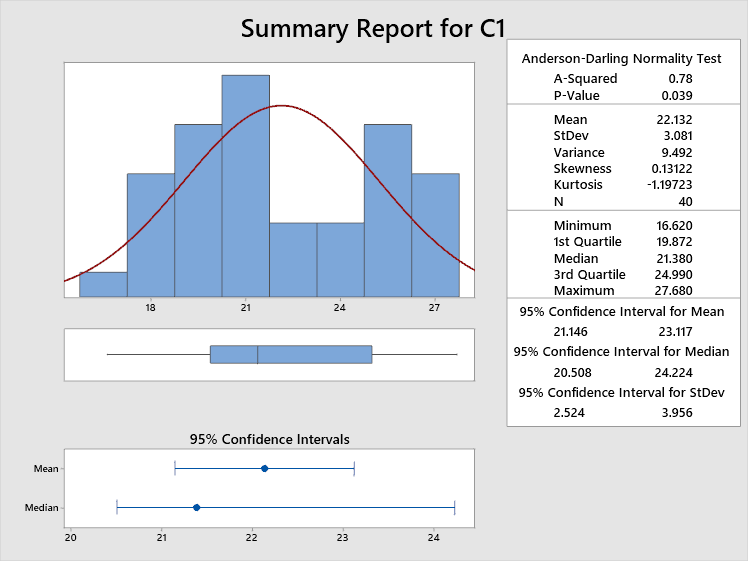
The results from Minitab are below:



I feel that the results from the Ruptures modification and the Minitab example were better correlated than the Sullivan comparison. A single change point was identified in both sets and there is a clear distinction in the graphs. The mean of the data set was 22.13 and the max deviation from the mean was 3.08. The data set appears to be in a steady state, aka In-Control (IC) due to the fact that the values are relatively grouped and stable. See the time series plot below.



The data appears to be independent and identically distributed (iid). See the chart below that shows the basic statistical information:



Although its not an ideal normal curve, the data still displays iid characteristics. The data points appear to be independent of each other with the same statistical probability regardless of any other variables value.