

Change Point Detection with Applications

Master Project

Han Ambrose

Supervised by: Lajos Horváth (Committee Chair), Fernando Guevara Vasquez, Jingyi Zhu

Submitted to the faculty of The University of Utah in partial fulfillment of the requirements for the degree of Master
of Statistics Department of Mathematics

February 19, 2021

Overview

1. The CUSUM Method
2. Weighted CUSUM Method with independent errors
3. Weighted CUSUM Method with dependent errors
4. Applications

CUSUM Method

Let X_1, X_2, \dots, X_N be independent normal random defined as

$$X_i = \left\{ \begin{array}{ll} 2 + e_i, & \text{if } 1 \leq i \leq N/3, \\ 1 + e_i, & \text{if } N/3 + 1 \leq i \leq 2N/3 \\ e_i, & \text{if } 2N/3 + 1 \leq i \leq N, \end{array} \right\} \quad (1)$$

where x denotes the integer part of x . We assume that e_1, e_2, \dots, e_N are independent and identically distributed normal random variables with mean 0 and variance σ^2 . According to our model we start with mean 2, this changes to 1 at $N/3 + 1$ and to 0 at $2N/3 + 1$. So we have exactly two changes at $N/3 + 1$ and $2N/3 + 1$. We wish to estimate the number of changes in the sequence. The testing and estimation method based on the CUSUM sequence

CUSUM Method

CUSUM Sequence

$$T(k) = \sum_{i=1}^k X_i - \frac{k}{N} \sum_{i=1}^N X_i, \quad 1 \leq k \leq N.$$

$$T_N = \frac{1}{\sigma} N^{-1/2} \max_{1 \leq k \leq N} |T(k)|,$$

where σ is a scaling parameter.

Null Hypothesis: No change in the mean

CUSUM Method

Critical Values

If the no change null hypothesis is true

$$T_N \xrightarrow{\mathcal{D}} \sup_{0 \leq t \leq 1} |B(t)|,$$

where $\{B(t), 0 \leq t \leq 1\}$ is a Brownian bridge. Let $c(\alpha)$ the critical value for the supremum of the absolute value of a Brownian bridge, i.e.

$$P \left\{ \sup_{0 \leq t \leq 1} |B(t)| \geq c(\alpha) \right\} = \alpha.$$

These critical values $c(\alpha)$ are the asymptotic values for the classical Kolmogorov–Smirnov statistic.

Figure: Kolmogorov-Smirnov Table

$n \backslash \alpha$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	1.94947	1.62762	1.51743	1.35810	1.22385	1.13795	1.07275
	\sqrt{n}	\sqrt{n}	\sqrt{n}	\sqrt{n}	\sqrt{n}	\sqrt{n}	\sqrt{n}

CUSUM Method

Simulation

Sample size is $N = 100$ and the variance is chosen as $\sigma^2 = 0.1$.

First segment is from $X_1 \dots X_{33} \sim N(2, 0.1)$

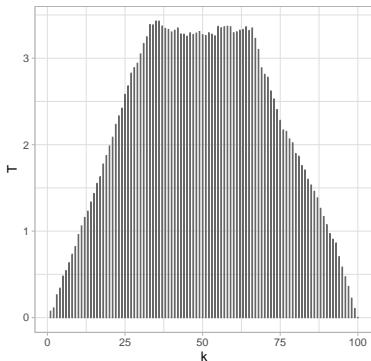
Second segment is from $X_{34} \dots X_{66} \sim N(1, 0.1)$

Last segment is from $X_{67} \dots X_{100} \sim N(0, 0.1)$

CUSUM Method

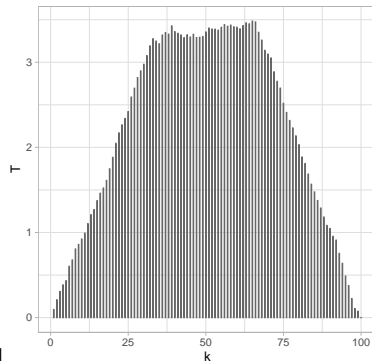
Close to the true values of the times of the changes in the mean.

Function $T(k)$
Sample Size = 100 , Variance = 0.1
Max_k = 35 , Max_T = 3.43



[k = 35]

Function $T(k)$
Sample Size = 100 , Variance = 0.1
Max_k = 65 , Max_T = 3.48



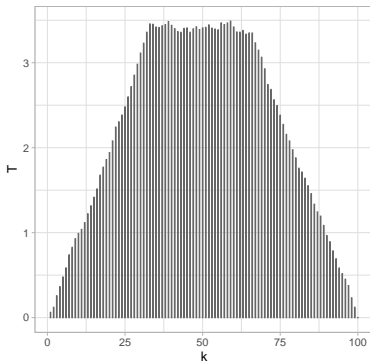
[k = 65]

Figure: A realization of the absolute value of the CUSUM process in the model (1) with $N = 100$ and $\sigma^2 = 0.1$.

CUSUM Method

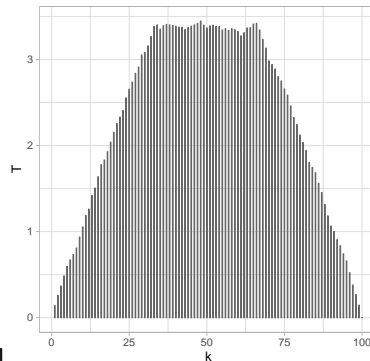
Not close to any of the real times of changes.

Function $T(k)$
Sample Size = 100 , Variance = 0.1
Max_k = 59 , Max_T = 3.49



[k = 59]

Function $T(k)$
Sample Size = 100 , Variance = 0.1
Max_k = 48 , Max_T = 3.45



[k = 48]

Figure: A realization of the absolute value of the CUSUM process in the model (1) with $N = 100$ and $\sigma^2 = 0.1$.

CUSUM Method

Repeat CUSUM process to find all changes

Cut the data into two subsets $X_1, X_2, \dots, X_{\bar{k}_1}$ and $X_{\bar{k}_1+1}, X_{\bar{k}_1+2}, \dots, X_N$ and again we compute the CUSUM sequence from each subset

$$T^{(1,1)}(k, X_1, X_2, \dots, X_{\bar{k}_1}) = \frac{1}{\bar{k}_1^{1/2}} \left| \sum_{i=1}^k X_i - \frac{k}{\bar{k}_1} \sum_{i=1}^{\bar{k}_1} X_i \right|.$$

From $X_{\bar{k}_1+1}, X_{\bar{k}_1+2}, \dots, X_N$ we compute

$$T^{(1,2)}(k, X_{\bar{k}_1+1}, X_{\bar{k}_1+2}, \dots, X_N) = \frac{1}{(N - \bar{k}_1)^{1/2}} \left| \sum_{i=\bar{k}_1+1}^k X_i - \frac{k - \bar{k}_1 + 1}{N - \bar{k}_1} \sum_{i=\bar{k}_1+1}^N X_i \right|.$$

Repeated this experiment M times and we got M vectors of the times of changes.

CUSUM Method

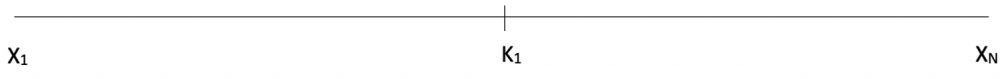


Diagram illustrating the CUSUM method. A horizontal line represents the data sequence, with a vertical marker at K_1 . The sequence starts at X_1 and ends at X_N .

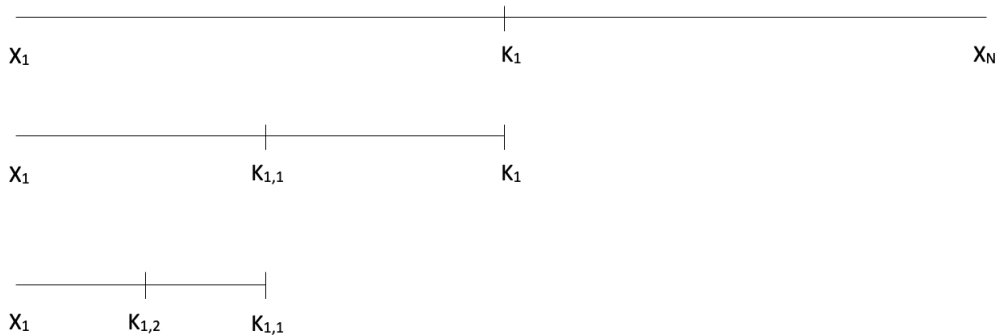
The CUSUM statistic for the first segment is defined as:

$$T^{(1,1)}(k, X_1, X_2, \dots, X_{\bar{k}_1}) = \frac{1}{\bar{k}_1^{1/2}} \left| \sum_{i=1}^k X_i - \frac{k}{\bar{k}_1} \sum_{i=1}^{\bar{k}_1} X_i \right|.$$

The CUSUM statistic for the second segment is defined as:

$$T^{(1,2)}(k, X_{\bar{k}_1+1}, X_{\bar{k}_1+2}, \dots, X_N) = \frac{1}{(N - \bar{k}_1)^{1/2}} \left| \sum_{i=\bar{k}_1+1}^k X_i - \frac{k - \bar{k}_1 + 1}{N - \bar{k}_1} \sum_{i=\bar{k}_1+1}^N X_i \right|.$$

CUSUM Method



CUSUM Method

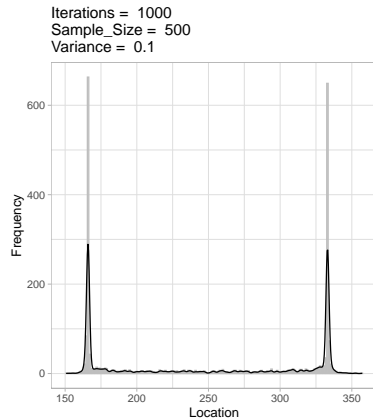
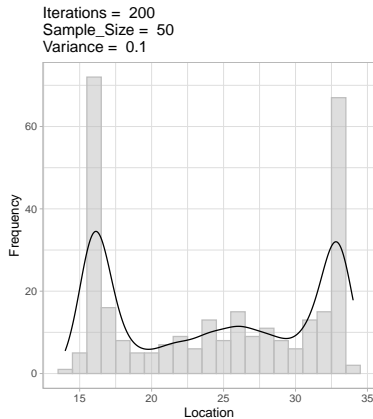


Figure: The distribution of the estimated times of changes with $\sigma^2 = 0.1$.

CUSUM Method

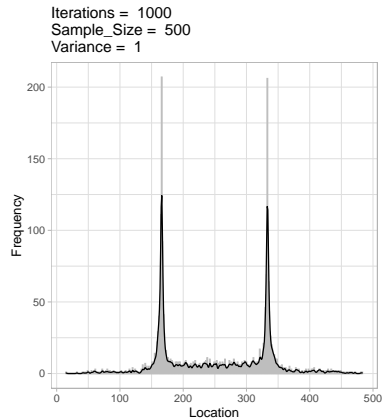
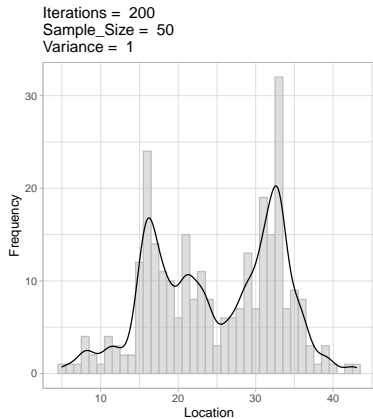


Figure: The distribution of the estimated times of changes with $\sigma^2 = 1$.

CUSUM Method

The higher the variance, the more likely we are able to find changes that is not close to any of the real times of changes.

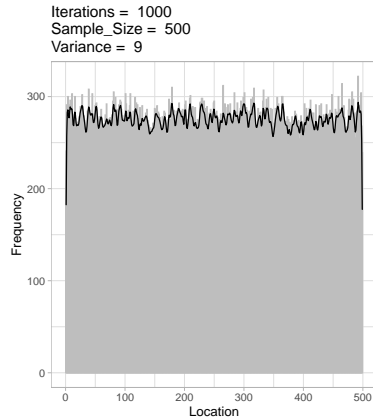
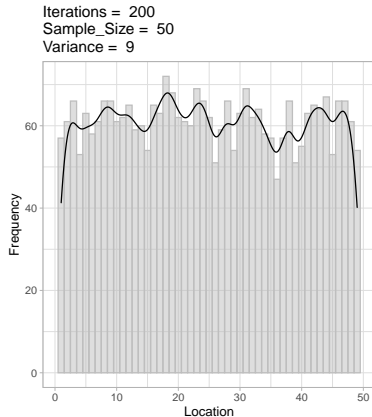


Figure: The distribution of the estimated times of changes with $\sigma^2 = 9$.

CUSUM Method

The probability that the change occurs between $N/3$ and $2N/3$ looks uniform. "Change point" which are not really of the change points. Hence the CUSUM finds more change points than we have in the data.

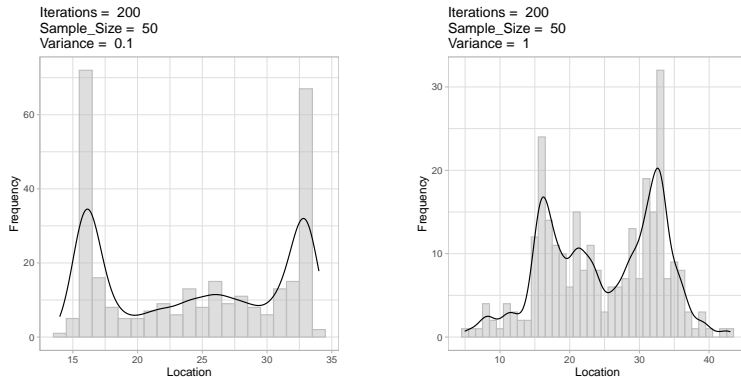
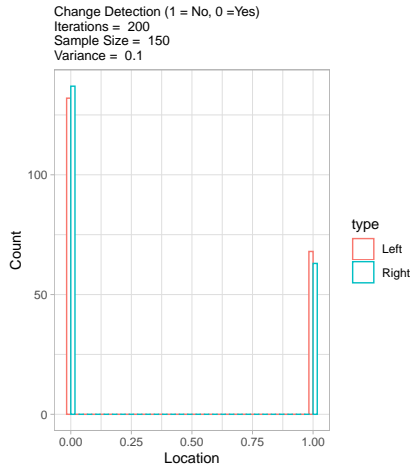


Figure: The distribution of the estimated times of changes with $\sigma^2 = 9$.

CUSUM Method

Figure: The distribution of the first and second estimated times of changes.



Weighted CUSUM

Weighted CUSUM Sequence

$$T_N(\kappa) = \frac{1}{\sigma} \max_{1 \leq k < N} \frac{N^{-1/2}}{((k/N)(1 - (k/N)))^\kappa} \left| \sum_{i=1}^k X_i - \frac{k}{N} \sum_{i=1}^N X_i \right|,$$

$0 \leq \kappa \leq 1/2$, where σ is scaling as before.

Weighted CUSUM

Critical Values

There is no formula for $T_N(\kappa)$ nor for its limit

$$T(\kappa) = \sup_{0 < t < 1} \frac{1}{(t(1-t))^\kappa} |B(t)|,$$

where $B(t)$ is a Brownian bridge. To do the testing step, we need to find $c_\kappa(\alpha)$ such that

$$P \left\{ \sup_{0 < t < 1} \frac{|B(t)|}{(t(1-t))^\kappa} \leq c_\kappa(\alpha) \right\} = 1 - \alpha,$$

Weighted CUSUM

Finding critical values through simulation

We simulated $M = 500$ independent copies of $T_N(\kappa)$, and computed the empirical distribution of the generated copies of $T_N(\kappa)$ when $N = 500$. Thus we got $c_\kappa(\alpha)$ for $\alpha = 0.0, 0.1, 0.05, 0.001$ and $\kappa = 0, 0.1, 0.2, \dots, 0.49$. The results are given in table below The simulation result for $\kappa = 0$ is compared to the known values of $\kappa_0(\alpha)$.

Weighted CUSUM

Table: Selected critical values for $T(\kappa)$.

α	Empirical Critical Values	Theoretical Critical Values
$\kappa = 0$		
0.01	1.604	1.628
0.05	1.329	1.358
0.1	1.190	1.224
$\kappa = 0.1$		
0.01	1.843	
0.05	1.552	
0.1	1.410	
$\kappa = 0.3$		
0.01	2.617	
0.05	2.146	
0.1	1.947	
$\kappa = 0.45$		
0.01	3.260	
0.05	2.755	
0.1	2.596	

Weighted CUSUM

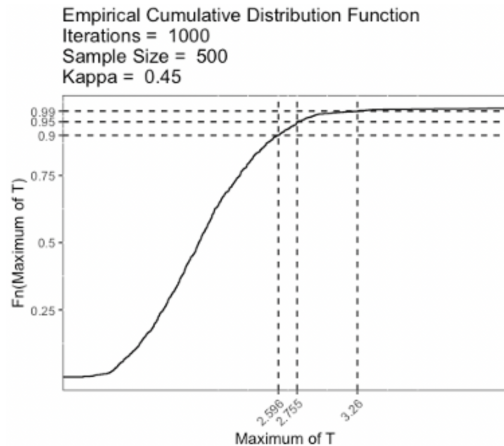
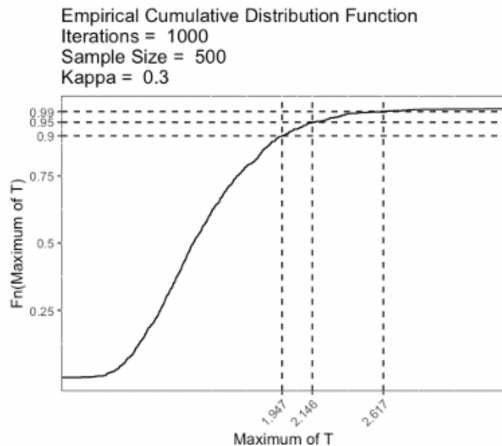


Figure: The empirical distribution of $T(0.3)$ and $T(0.45)$ based on Monte Carlo simulations.

Weighted CUSUM

The shapes of the realizations confirms the theoretical result that the “flat” of the unweighted CUSUM disappears. Using larger weight, the largest value of the weighted CUSUM is closer to and sharper at the change point.

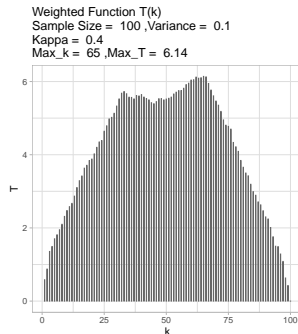
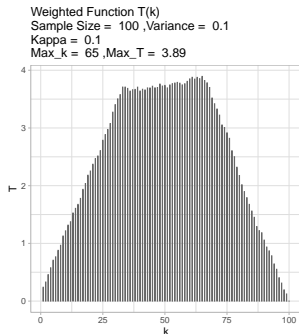
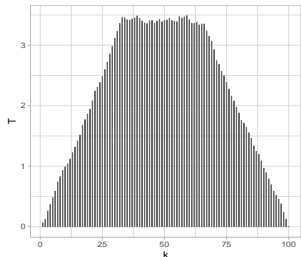
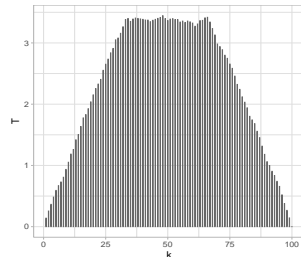


Figure: A realization of the absolute value of the weighted CUSUM $\kappa = .1$ and $\kappa = .4$ process with $N = 100$ and $\sigma^2 = 0.1$

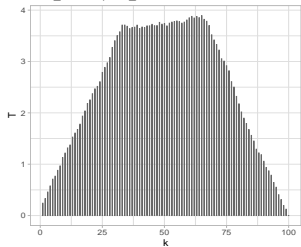
Function $T(k)$
 Sample Size = 100 , Variance = 0.1
 Max_k = 59 , Max_T = 3.49



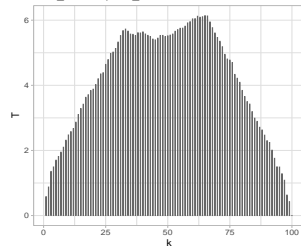
Function $T(k)$
 Sample Size = 100 , Variance = 0.1
 Max_k = 48 , Max_T = 3.45



Weighted Function $T(k)$
 Sample Size = 100 , Variance = 0.1
 Kappa = 0.1
 Max_k = 65 , Max_T = 3.89



Weighted Function $T(k)$
 Sample Size = 100 , Variance = 0.1
 Kappa = 0.4
 Max_k = 65 , Max_T = 6.14



Top two figures: non_weighted CUSUM

Bottom two figures: weighted CUSUM

Weighted CUSUM

We repeated our experiments but now we used exponential (λ) observations. For the sake of comparison, $\lambda = \sigma^2$, where σ^2 is the variance in the normal case.

Table: Selected critical values for $T(\kappa)$ in case of exponential observations compared to values from the limit distribution.

$\kappa = 0$		
0.01	1.604	1.698
0.05	1.329	1.332
0.1	1.190	1.212
$\kappa = 0.1$		
0.01	1.843	1.895
0.05	1.552	1.602
0.1	1.410	1.421
$\kappa = 0.3$		
0.01	2.617	2.548
0.05	2.146	2.132
0.1	1.947	1.933
$\kappa = 0.45$		
0.01	3.260	3.877
0.05	2.755	3.067
0.1	2.596	2.733

Weighted CUSUM

Weighted Function $T(k)$
Sample Size = 100 , Variance = 0.1
Kappa = 0
Max_k = 34 , Max_T = 7.86

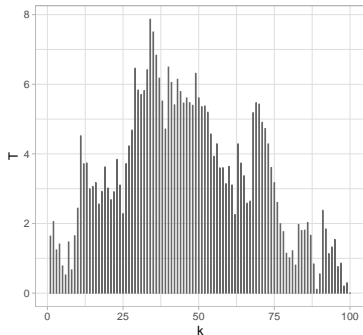


Figure: $\kappa = 0$, i.e. no weight is used, our estimates are not change points with relatively large probabilities

Weighted Function $T(k)$
Sample Size = 100 , Variance = 0.1
Kappa = 0.1
Max_k = 20 , Max_T = 6.68

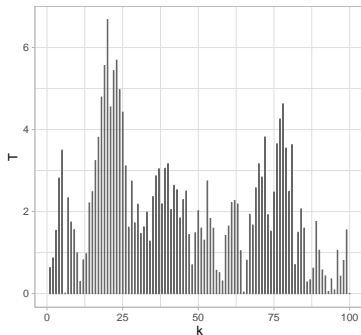


Figure: $\kappa = 0.1$, changes are clear

Weighted Function $T(k)$
Sample Size = 100 , Variance = 0.1
Kappa = 0.4
Max_k = 23 , Max_T = 12.85

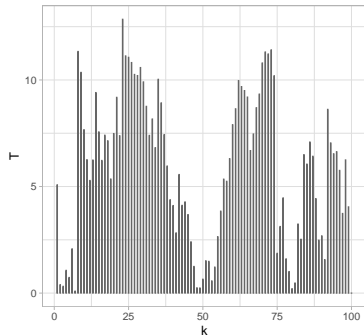


Figure: $\kappa = 0.4$, there are artificial changes found at the beginning and end of the data.

Error Terms Assumption

Are the error terms correlated or uncorrelated?

1. Correlated: ARMA(p,q) with long run variance estimators
2. Uncorrelated: GARCH with sample variance estimators

ARMA(p,q) with long run variance estimators

In the dependent case σ^2 is not the variance of the individual observations, but the long run variance of the stationary sequence. Using the sample variance defined before, we might under or overestimate the long run variance. The standard estimator for the long run variance is the kernel estimator.

The long run variance estimator is

$$\hat{\sigma}_N^2 = \sum_{\ell=-(N-1)}^{N-1} K\left(\frac{\ell}{h}\right) \hat{\gamma}_\ell,$$

where $K(t)$ is the kernel and h is the smoothing parameter (window).

ARMA(p,q) with long run variance estimators

Under the null hypothesis

$$\frac{1}{\hat{\sigma}_N N^{1/2}} \max_{1 \leq k \leq N} \left(\frac{N^2}{k(N-k)} \right)^\kappa \left| \sum_{i=1}^k e_i - \frac{k}{N} \sum_{i=1}^N e_i \right| \xrightarrow{\mathcal{D}} \sup_{0 < t < 1} \frac{1}{(t(1-t))^\kappa} |B(t)|.$$

However, the behaviour of σ_N^2 is different from the independent case since under the alternative

$$\frac{\hat{\sigma}_N^2}{h} \xrightarrow{P} \tau^2 > 0.$$

h is the smoothing parameter (window)

This means that the power of the test will be seriously reduced if h is too large. It is known that if the errors are from an ARMA(p, q), then the errors are correlated.

GARCH with sample variance estimators

Using GARCH, e_i is an uncorrelated sequence and there is no need to use the long run variance estimator since in case of uncorrelated the sample variance works, i.e.

$$S_N^2 \xrightarrow{P} \sigma^2$$

and

$$\sigma^2 = E_i^2.$$

GARCH with sample variance estimators

So in case of a GARCH(1,1),

$$\frac{1}{S_N^{1/2}} \max_{1 \leq k \leq N} \left(\frac{N^2}{k(N-k)} \right)^\kappa \left| \sum_{i=1}^k e_i - \frac{k}{N} \sum_{i=1}^N e_i \right| \xrightarrow{\mathcal{D}} \sup_{0 < t < 1} \frac{1}{(t(1-t))^\kappa} |B(t)|$$

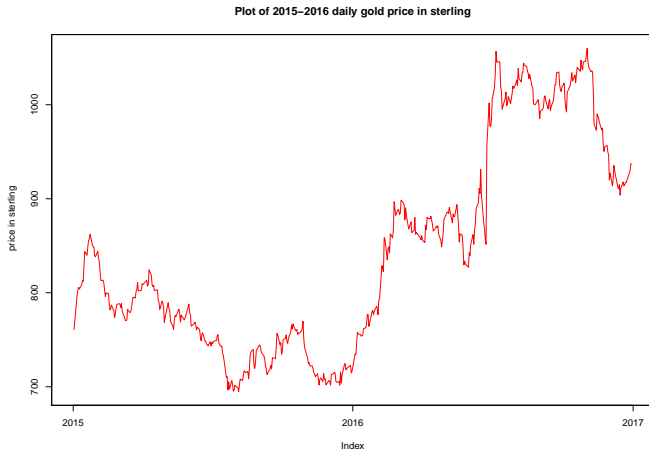
under the null hypothesis and

$$\frac{1}{S_N^{1/2}} \max_{1 \leq k \leq N} \left(\frac{N^2}{k(N-k)} \right)^\kappa \left| \sum_{i=1}^k e_i - \frac{k}{N} \sum_{i=1}^N e_i \right| \xrightarrow{P} \infty,$$

where S_N^2 is the sample variance. Also, the power is not reduced as in case when the long run variance estimator is used.

Application - Gold Data

The mean value is not zero and the variance is very high. This indicates that the time series is non-stationary with varying mean and variance. Thus, to stationarize the process, we study the log return of the price.



Application - Gold Data

Log Return Calculation

Taking the ratio of the daily price at time $i + 1$ and the daily return at time i in order to calculate the ratio.

Let X_1, X_2, \dots, X_n be the gold prices. The return value is

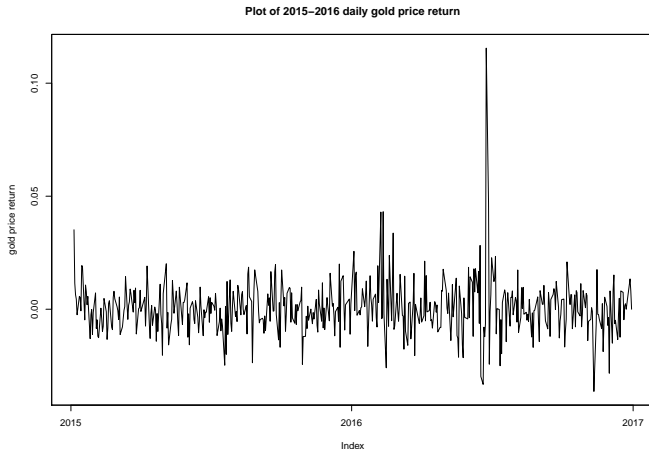
$$\frac{X_{i+1}}{X_i}$$

Then, we take the log of this ratio to get the return

$$Y_i = \log\left(\frac{X_{i+1}}{X_i}\right) = \log(X_{i+1}) - \log(X_i)$$

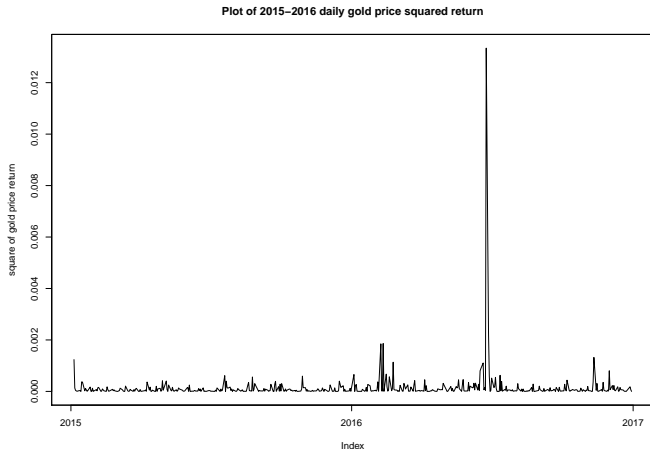
Application - Gold Data

The returns vary along the zero line with the largest log return of gold prices observed around beginning, mid and end of 2016, which shows signs of volatility.



Application - Gold Data

To better identify its volatility, we also use the square of log return. Peaks are clearly shown figure below.



Application - Gold Data

CUSUM using sample variance

The weighted CUSUM is

$$\frac{1}{S_{N-1}} \frac{N^{-1/2}}{((k/N)(1 - (k/N)))^\kappa} \left| \sum_{i=1}^k Y_i - \frac{k}{N-1} \sum_{i=1}^{N-1} Y_i \right|$$

with the sample mean and sample variance as follow

$$\bar{Y}_{N-1} = \frac{1}{N-1} \sum_{i=1}^{N-1} Y_i$$

$$S_{N-1}^2 = \frac{1}{N-2} \sum_{i=1}^{N-1} (Y_i - \bar{Y}_{N-1})^2$$

Application - Gold Data

CUSUM using sample variance

We reject if

$$\max_{1 \leq k < N-1} \frac{1}{S_{N-1}} \frac{N^{-1/2}}{((k/N)(1 - (k/N)))^\kappa} \left| \sum_{i=1}^k Y_i - \frac{k}{N-1} \sum_{i=1}^{N-1} Y_i \right| \geq c(\kappa, \alpha)$$

Application - Gold Data

Weighted CUSUM process with kappa =0 using log return of gold data

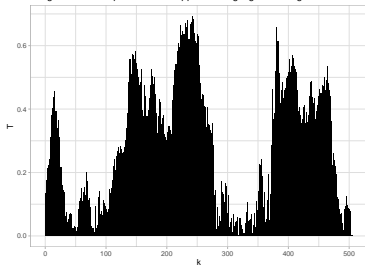


Figure: $\kappa = 0$, Fail to Reject

Weighted CUSUM process with kappa =0.1 using log return of gold data

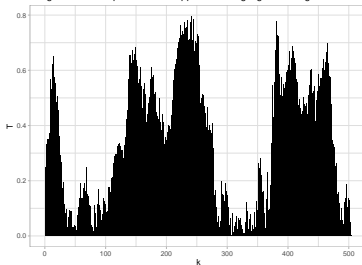


Figure: $\kappa = 0.1$, Fail to Reject

Weighted CUSUM process with kappa =0.45 using log return of gold data

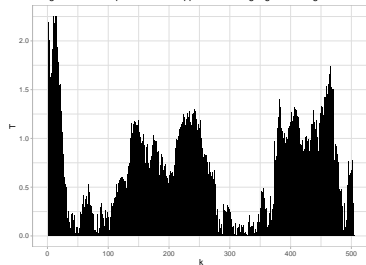
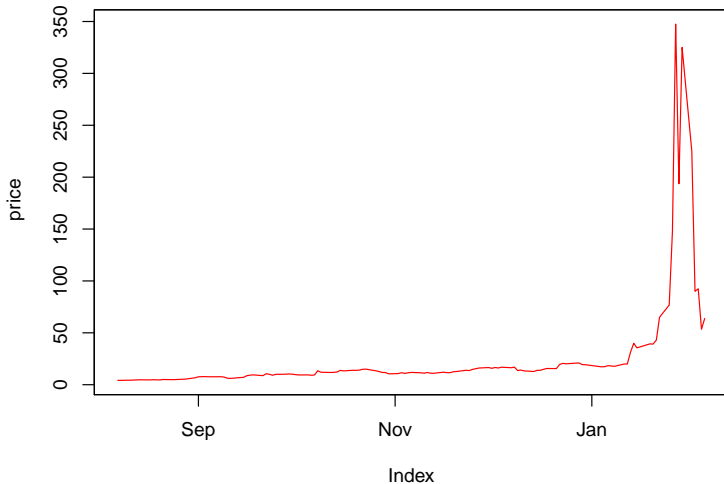


Figure: $\kappa = 0.4$, Fail to Reject

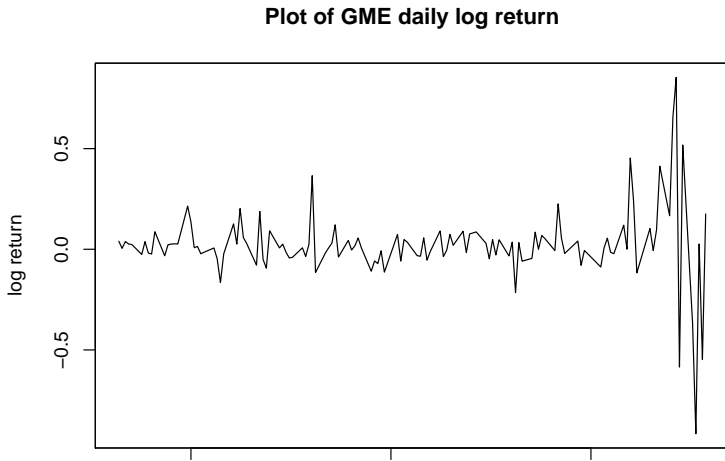
Application - GME Data

Plot of GME stock prices from 08-2020 to 02-2021



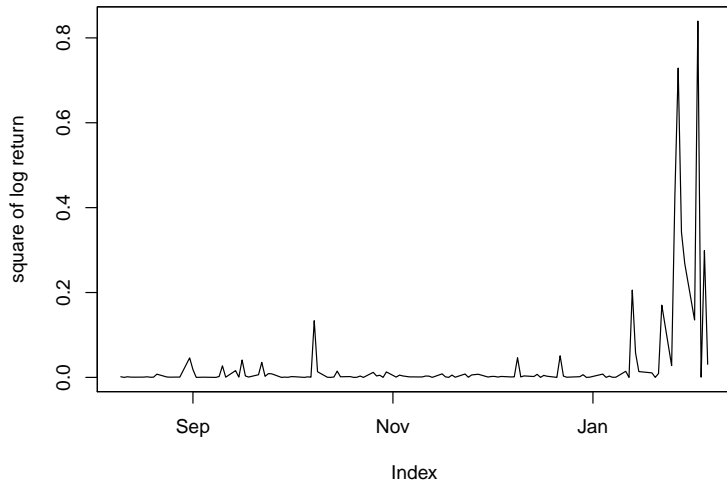
Application - GME Data

As expected, we see some big swings at the end, but also some small swings in the beginning of the data.



Application - GME Data

Plot of GME daily squared log return



Application - GME Data

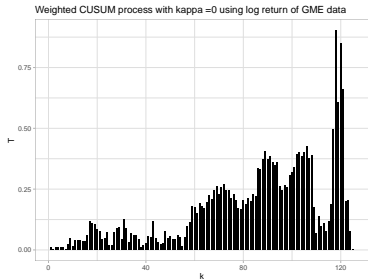


Figure: $\kappa = 0$, Fail to Reject

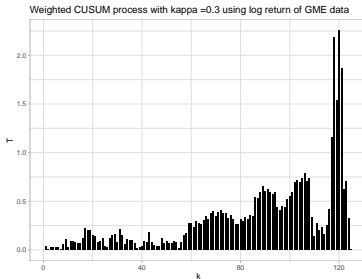


Figure: $\kappa = 0.3$, Reject

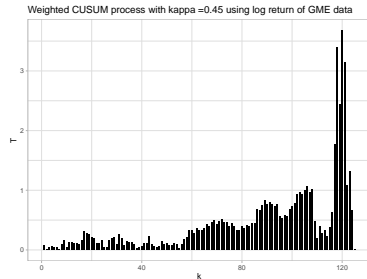








Figure: $\kappa = 0.4$, Reject

Conclusion

- Weighted CUMSUM method should be used for better accuracy. It is more likely it is to detect change when the weight is adjusted.
- This research is limited to the assumption that the error term is uncorrelated and follows the GARCH process
- The research could also be extended towards monthly, quarterly and yearly data set to see the effectiveness of certain data types
- Overall, it is a powerful tool to detects anomalous patterns reliably and efficiently for continuous financial surveillance.

References

-  Aue, A. and Horváth, L.: Structural breaks in time series. *Journal of Time Series Analysis* **23**(2013), 1–16.
-  Csörgő, M. and Horváth, L.: *Limit Theorems in Change–Point Analysis*. Wiley, New York, 1997.
-  Horváth, L. and Rice, G.: Extensions of some classical methods in change point analysis (with discussions) *TEST* **23**(2014), 219–290.
-  Horváth, L., Rice, G. and Zhao, Y.: Detecting multiple changes in linear models. Preprint.
-  DasGupta, A.: *Asymptotic Theory of Statistics and Probability*, 2008.
-  One-Sample Kolmogorov-Smirnov table. (n.d.). Retrieved February 10, 2021, from <https://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/>

The End