

Using Monte Carlo Method to Compare CUSUM and EWMA Statistics

Xiaoyu Shen

Zhen Zhang

Abstract: Since ordinary datasets usually contain change points of variance, CUSUM and EWMA statistics can be used to detect these change points. In this project, we are going to use Monte Carlo Method to compare the efficiencies of these two statistics according to the average run length (ARL).

Key words: change point, CUSUM, EWMA, Monte Carlo Method, ARL

1. Introduction

Basic statistical theories told us that the variance of a data series is a very useful criterion indicating the deviation of the original data from the mean. However, a series of practical data usually may not have the constant variance, such that detecting the moment when the variance of data changes has its own significant value.

In the quality control theory, there exist two quality control charts based on two different statistics, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA). From these two charts, one can easily notice the moment when the properties of data become abnormal, which may help us to control the quality of data. Our topic in this project will focus on the CUSUM and EWMA statistics. Many statistical papers have revealed that after some simple transformations, both CUSUM and EWMA statistics can be used to detect the change points of variance.

2. Empirical Results

The expression of cumulative sum of square is

$$\text{CUSUM: } C_k = \sum_{t=1}^k \xi_t^2 .$$

We can get the D_k statistic after some simple transformation of CUSUM.

$$D_k = \frac{C_k}{C_n} - \frac{k}{n} = \frac{\sum_{t=1}^k \xi_t^2}{\sum_{t=1}^n \xi_t^2} - \frac{k}{n}, \quad k = 1, 2, \dots, n .$$

Also, we can get the expression of EWMA.

$$\text{EWMA: } W_t = (1-r)W_{t-1} + r \ln(\sigma_t^2), \quad \sigma_t^2 = \frac{1}{t-1} \sum_{i=1}^t (x_i - \bar{\mu})^2, \quad \bar{\mu} = \frac{1}{t} \sum_{i=1}^t x_i.$$

$$\text{EWMA}_n(r) = \max_{1 \leq t \leq n} \left| \frac{\sqrt{2-r}}{\sqrt{r(1-(1-r)^{2t})}} \sum_{i=0}^{t-1} r(1-r)^i \ln(\xi_{t-i}^2) \right|,$$

where $\frac{r(1-r)^i \sqrt{2-r}}{\sqrt{r(1-(1-r)^{2t})}}$ is the weight of the statistic. The method which may be

found in many books about quality control is not the topic of our project, so that it is not included in this paper.

After giving the expression of CUSUM and EWMA, we need to compare these two statistics. To finish this comparison, we introduce a new criterion, the average run length, which may help us to compare the detecting effect of these two statistics. The following is the definition of the average run length.

Definition: The average run length (ARL) of a sampling inspection scheme at a given level of quality is the average number of samples of n items taken in the period between the time when the process commences to run at the stated level and that at which the scheme indicates a change from acceptable to rejectable quality level is likely to have occurred.

In this project, we can define ARL as the mean of the first run length the change point take place after generating large numbers of samples. With the help of computers, we could do the simulation study of ARL to compare the detecting effect of CUSUM and EWMA.

3. Practical Results

To test the effect of CUSUM statistic, we first generate two series of random data, each of which has 1000 data. One of them has no change points of variance, another has two change points of variance. It is obvious to notice from the first pair of plot that 400 and 700 are the change points of variance in the second series of data.

We can see from the second pair of graph that the plot of CUSUM statistic without variance change points is close to a line, while that with variance change points has different slopes. After some simple transformation, we get the D_k plot (third pair of plots) which shows the change points more obviously. The D_k plot of data without change points is close to a line (bounded in $[-0.015, 0.015]$), while the D_k plot of data with change points greatly exceed that boundary and has the maximum at point 700. Even if D_k plot cannot convince us of the fact that point 400 is also a change point, CUSUM statistics and its derivative D_k really have good effects on detecting the change points of variance.

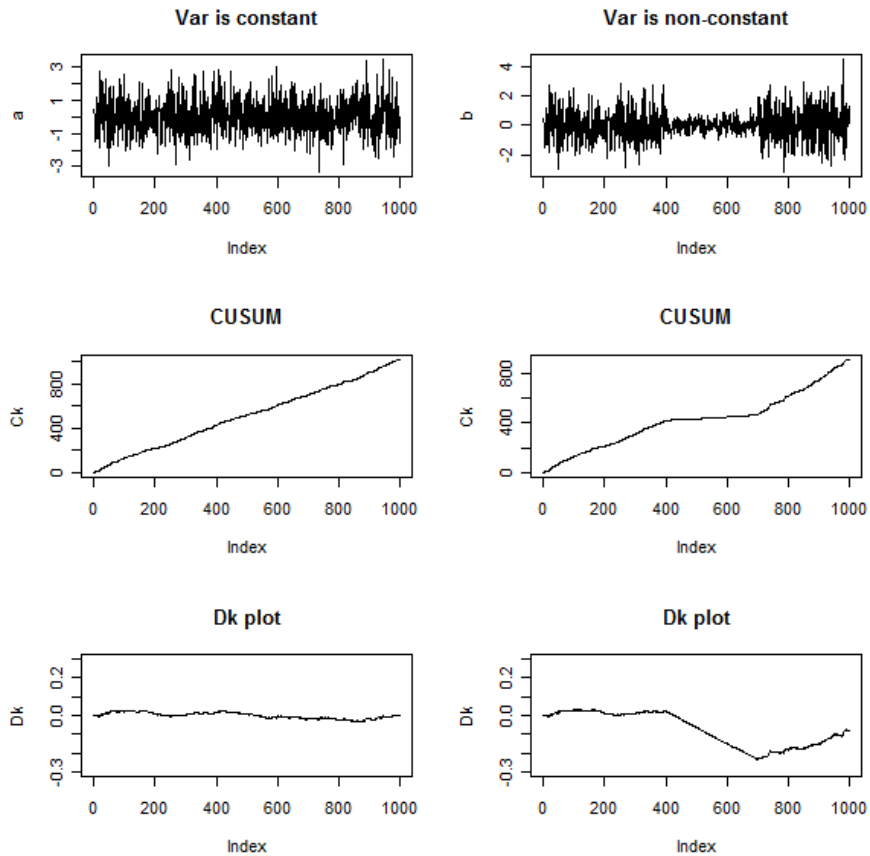


Figure 1: Effect of CUSUM

We use the same two series of data to test the effect of EWMA statistic. The first column in Figure 2 is the graphs of data without variance change points and the second column is those with change points. Although different weight r may have different effect of detecting, we can see the change points of variance obviously from all the plots in the second column. But also from these plots, the values of EWMA statistics seem to be abnormal when the length r is close to zero. So it is difficult for EWMA to detect the change point existing at the beginning of data.

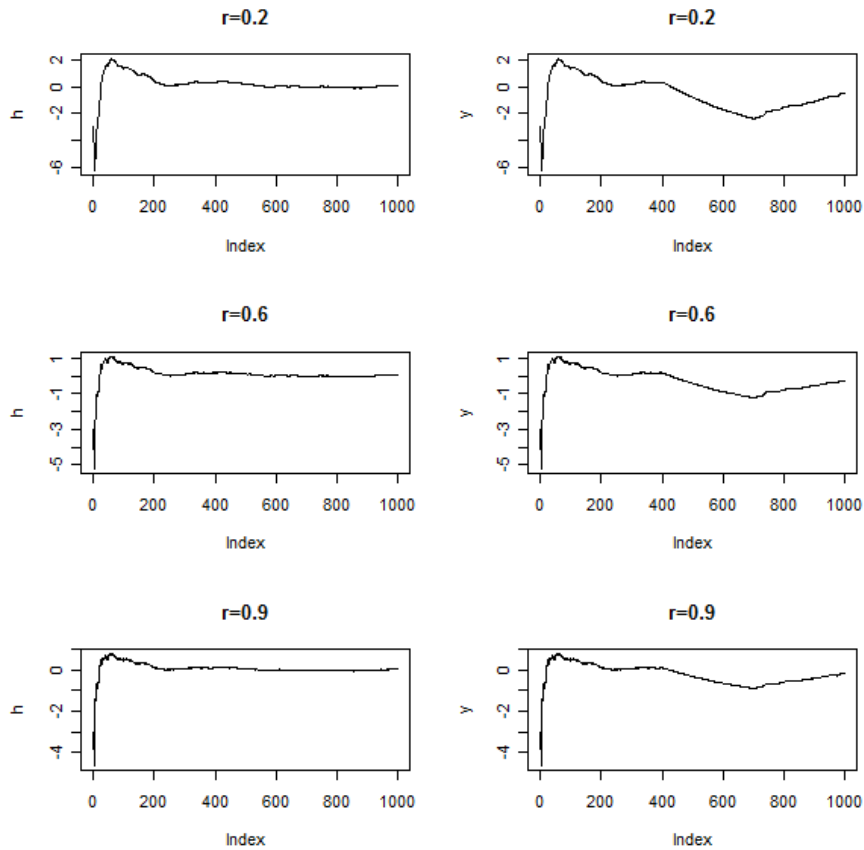


Figure 2: Effect of EWMA ($r=0.2, 0.6$ and 0.9)

We know from the previous discussion that CUSUM and EWMA both can help to detect the change point of variance, so it is time for us to apply Monte Carlo Method to the further research now. According to Monte Carlo simulation study of two statistics, we find the ARL of CUSUM is always smaller than that of EWMA, no matter how large the sample size n and weight r are. If the sample size n is small, the ARL of EWMA decreases when the weight r increases from 0.2 to 0.9. But if the sample size is large, the ARL of EWMA increases when the weight r increases.

	$n=50$	$n=100$	$n=200$
CUSUM	24.23	52.24	48.61
EWMA ($r=0.2$)	30.95	57.87	112.86
EWMA ($r=0.6$)	29.16	58.65	114.29
EWMA ($r=0.9$)	29.05	56.66	118.24

Table 1: $S=100$

	n=50	n=100	n=200
CUSUM	24.82	50.21	94.45
EWMA (r=0.2)	29.64	58.37	111.84
EWMA (r=0.6)	29.41	57.31	115.36
EWMA (r=0.9)	29.27	59.19	119.07

Table 2: S=300

	n=50	n=100	n=200
CUSUM	25.56	49.83	100.74
EWMA (r=0.2)	29.89	58.5	114.21
EWMA (r=0.6)	29.7	58.65	116.96
EWMA (r=0.9)	29.09	57.41	118.09

Table 3: S=500

4. Conclusion

CUSUM and EWMA can be both used to detect the change point of variance. However, after comparing the ARL of these two statistics with the help of Monte Carlo Method, we know that change points of variance might be more easily found by means of CUSUM. In the case of EWMA, when the sample size is small, EWMA with large weight has better effect of change points detecting. But when the sample size is large, EWMA with small weight works better than that with large weight.

In addition, we notice from the plot that EWMA statistic cannot detect the change point at the beginning of the dataset. The reasons may be the fact that when the length r is small, EWMA will be too large to reflect the effect of detecting. Also, information included in the beginning part of data may not be adequate for EWMA to detect the overall change points.

Appendix

R code for the project

```
# test the effect of CUSUM
a=rnorm(1000);b=a;
b[401:700]=rnorm(300,0,0.4);b[701:1000]=rnorm(300,0,1.2);
c=a;d=a;e=b;f=b;c[1]=a[1]^2;e[1]=b[1]^2;
for(i in 1:999)
{c[i+1]=c[i]+a[i+1]^2;e[i+1]=e[i]+b[i+1]^2;}
for(j in 1:1000)
{d[j]=c[j]/c[1000]-j/1000;f[j]=e[j]/1000-j/1000;}
```

```

plot(a,type="l",main="Var is constant")
plot(b,type="l",main="Var is non-constant")
plot(c,type="l",ylab="Ck",main="CUSUM")
plot(e,type="l",ylab="Ck",main="CUSUM")
plot(d,type="l",ylab="Dk",main="Dk plot",ylim=c(-0.3,0.3))
plot(f,type="l",ylab="Dk",main="Dk plot",ylim=c(-0.3,0.3))

```

```

# test the effect of EWMA (r=0.2, 0.6 and 0.9)
ewma=function(r){
c=a;d=a;e=b;f=b;c[1]=a[1]^2;e[1]=b[1]^2;
for(i in 1:999)
{c[i+1]=c[i]+a[i+1]^2;e[i+1]=e[i]+b[i+1]^2;}
for(j in 1:1000)
{d[j]=c[j]/j;f[j]=e[j]/j;}
g=a;h=a;h[1]=sqrt(2-r)/sqrt(r*(1-(1-r)^2))*r*log(d[1]^2);
for(i in 2:1000){g[1]=r*(1-r)^(i-1)*log(d[1]^2);
for(j in 1:(i-1)) g[j+1]=r*(1-r)^(i-j-1)*log(d[j+1]^2)+g[j];
h[i]=sqrt(2-r)/sqrt(r*(1-(1-r)^(2*i)))*g[i]}
x=b;y=b;y[1]=sqrt(2-r)/sqrt(r*(1-(1-r)^2))*r*log(f[1]^2);
for(i in 2:1000){x[1]=r*(1-r)^(i-1)*log(f[1]^2);
for(j in 1:(i-1)) x[j+1]=r*(1-r)^(i-j-1)*log(f[j+1]^2)+x[j];
y[i]=sqrt(2-r)/sqrt(r*(1-(1-r)^(2*i)))*x[i]}
plot(h,type="l",main="r=0.2")
plot(y,type="l",main="r=0.2")
}

```

```

# calculate the ARL of CUSUM (S=100,300,500, n=50,100,200)
arl.cusum = function(S,n){ # S=100, n=50
z=rep(0,n)
for(k in 1:S)
{a=rnorm(n);
c=a;d=a;c[1]=a[1]^2;
for(i in 1:(n-1))
{c[i+1]=c[i]+a[i+1]^2;}
for(j in 1:n)
{d[j]=c[j]/c[n]-j/n;}
z[k]=1;
for(i in 1:( n-1))
{ if (abs(d[i+1])>=abs(d[i]))
{z[k]=i+1;}
else
{d[i+1]=d[i]}
}}
list(ARL=mean(z))

```

```

}

# calculate the ARL of EWMA (S=100,300,500, n=50,100,200, r=0.2, 0.6, 0.9)
arl.ewma = function(S,n,r){
z=rep(0,S)
for(k in 1:S)
{a=rnorm(n);
c=a;d=a;c[1]=a[1]^2;
for(i in 1:(n-1))
{c[i+1]=c[i]+a[i+1]^2;}
for(j in 1:n)
{d[j]=c[j]/j;}
g=a;h=a;h[1]=sqrt(2-r)/sqrt(r*(1-(1-r)^2))*r*log(d[1]^2);
for(i in 2:n){g[1]=r*(1-r)^(i-1)*log(d[1]^2);
for(j in 1:(i-1)) g[j+1]=r*(1-r)^(i-j-1)*log(d[j+1]^2)+g[j];
h[i]=sqrt(2-r)/sqrt(r*(1-(1-r)^(2*i)))*g[i]}
z[k]=round(n/3);
for(i in round(n/3):(n-1))
{ if (abs(d[i+1])>=abs(d[i]))
{z[k]=i+1;}
else
{d[i+1]=d[i]}
}}
list(r=r,ARL=mean(z))
}

```