CUSUM procedures based on sequential ranks

C van Zyl

22231609

(Hons BSc Actuarial Science)

Dissertation submitted in partial fulfilment of the requirements for the degree *Magister Scientiae* in *Risk Analysis* at the Potchefstroom Campus of the North-West University

Supervisor: Prof F Lombard

October 2015

"Who rebels with mathematics?

...

She [Pari] said there was comfort to be found in the permanence of mathematical truths, in the lack of arbitrariness and the absence of ambiguity. In knowing that the answers may be elusive, but they could be found. They were there, waiting, chalk scribbles away."

Khaled Hosseini, And the Mountains Echoed.

Summary

The main objective of this dissertation is the development of CUSUM procedures based on signed and unsigned sequential ranks. These CUSUMs can be applied to detect changes in the location or dispersion of a process. The signed and unsigned sequential rank CUSUMs are distribution-free and robust against the effect of outliers in the data. The only assumption that these CUSUMs require is that the in-control distribution is symmetric around a known location parameter. These procedures specifically do not require the existence of any higher order moments. Another advantage of these CUSUMs is that Monte Carlo simulation can readily be applied to deliver valid estimates of control limits, irrespective of what the underlying distribution may be.

Other objectives of this dissertation include a brief discussion of the results and refinements of the CUSUM in the literature. We justify the use of a signed sequential rank statistic. Also, we evaluate the relative efficiency of the suggested procedure numerically and provide three real-world applications from the engineering and financial industries.

Keywords: CUSUM, distribution-free, sequential rank, symmetric distribution, location change, dispersion change.

Uittreksel

Die verhandeling het hoofsaaklik ten doel om CUSUM prosedures te ontwikkel gebaseer op betekende en onbetekende sekwensiële range. Hierdie CUSUMs kan aangewend word om ‘n verandering in die lokaliteit of spreiding van ‘n proses te identifiseer. Die betekende en onbetekende sekwensiële range is verdelingsvry en ook robuust teen die effek van uitskieters in data. Die enigste aanname wat hierdie CUSUMs vereis is dat die in-beheer verdeling simmetries gesentreer is om ‘n bekende lokaliteitsparameter. Hierdie prosedures maak spesifiek geen aannames aangaande enige hoër orde momente nie. Nog ‘n voordeel van hierdie CUSUMs is dat Monte Carlo simulasie met gemak toegepas kan word om kontrole grense te vind ongeag die aard van die onderliggende verdeling.

Ander doelstellings van hierdie verhandeling sluit ‘n bondige bespreking van die resultate en verfynings van die CUSUM in die literatuur in. Ons regverdig die gebruik van die betekende sekwensiële rangstatistiek. Verder evalueer ons die relatiewe doeltreffendheid van die voorgestelde prosedure numeries en verskaf drie wêreldsgetroue toepassings vanuit die ingenieurs- en finansiële industrië.

Sleutelwoorde: CUSUM, verdelingsvry, sekwensiële rang, simmetriese verdeling, lokaliteitsverandering, spreidingsverandering.

Acknowledgments

The adventure of completing this dissertation would never have been possible without generous support, encouragement and guidance. I thank the following people and institutions for their contributions.

Professor Freek Lombard, my supervisor, for his valuable guidance, insight, patience and enthusiasm. Thank you for guiding me with wisdom and investing ample time towards my studies. I sincerely appreciate the enjoyment and precise manner with which you teach and share some of your vast knowledge. I am privileged to be taught by a renowned scientist.

My fellow colleagues in the Department of Statistics at the North-West University for valuable discussions and computer help. A special word of gratitude to Elzabé Reynolds for proofreading the text.

The financial contributions of the following institutions are appreciated:

- The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.
- The DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS). Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the CoE-MaSS.
- The Centre for Business Mathematics and Informatics (BMI).

My non-statistician friends and extended family for support, encouragement, putting up with my frustrations and sharing in my excitements. To single anyone out, would be to leave someone out.

My sincere appreciation to my parents, Gerhard and Berna, and my sister, Esmari, for love, a keen interest in my studies and the investments you made towards my education, personal growth and development.

Henda, for her unconditional love and support throughout, sharing my passion in forever gaining new knowledge. Thank you for always believing in me, for lending me your imagination and for teaching me to follow my heart.

My grandmother in heaven, whose encouragement remains a fond memory. She would have been proud.

God, for granting me the opportunity, talent and privilege to study.

Contents

Frequently used notation

1. $N(\mu, \sigma^2)$ denotes a normal distribution (or random variable) with mean μ and variance σ^2.
2. t_{df} denotes a t-distribution (or random variable) with df degrees of freedom.
3. $U(a, b)$ denotes a uniform distribution (or random variable) over the interval $[a, b]$.
4. $\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$
5. $u := v$ means u is identically equal to v.
6. $\mathbb{I}(A) = \begin{cases} 1 & \text{if condition } A \text{ holds} \\ 0 & \text{if condition } A \text{ does not hold.} \end{cases}$
7. $X_{i:j}$ is the j^{th} order statistic of the random variables $X_1, X_2, \cdots, X_i$.
8. $X \overset{\mathcal{D}}{=} Y$ means that X and Y have the same distribution.
9. $X_i \to c$ means $\lim_{i \to \infty} X_i = c$.
10. $\xrightarrow{\mathcal{D}}$ means convergence in distribution.
11. i.i.d. abbreviates "independent and identically distributed".
12. "Cumulative distribution function" is abbreviated cdf.
13. "Average run length" is abbreviated ARL.
14. "Signed sequential rank" is abbreviated SSR.

15. “Unsigned sequential rank” is abbreviated USR.

16. “Locally most powerful” is abbreviated LMP.

Chapter 1

Introduction

1.1 Literature review

The CUSUM (cumulative sum) procedure is a fundamental tool of statistical process control. Its aim is to detect a persistent or substantial change in the output of a process. Such inspection schemes typically entail taking observations in a sequential manner on a measurable property of a particular process. The process is said to be in control as long as the location and scale parameters are at their desired levels. The fundamental theory and history of sequential analysis can be found in Ghosh and Sen (1991) and Siegmund (1985).

The first statistical process control procedure was developed by Shewhart (1931). If $n \geq 1$ observations are available at time t, the mean $\overline{X}_t$ is plotted against time t to obtain what is known as the Xbar chart. Assuming that the process is in control when the population mean μ is 0, the control limits are typically of the form $\mu \pm 1.96\sigma/\sqrt{n}$. As soon as a sample mean $\overline{X}_t$ falls outside of these control limits a change in distribution is signalled. Full details on this type of control chart can be found in Dudding and Jennett (1942), Duncan (1959) and Montgomery (1996). The main drawback of the Xbar chart is that it requires groups of observations at each time point. However, in many situations such multiple observations are not available. The CUSUM procedure suggested by Page (1954) as an extension of the sequential probability ratio test of Wald (1947), provided a solution to this problem.

There exists presently an extensive body of literature on CUSUM procedures. The CUSUM finds application in numerous scientific fields that include, amongst others, engineering (Hawkins and Olwell (1998), Timmer et al. (2001)), public health applications (Woodall, 2006), business (Kahya and Theodossiou, 1999) and finance (Yi et al. (2006), Lam and Yam (1997), Golosnoy and Schmid (2007), Mukherjee (2009) and Coleman et al. (2001)).

According to Hawkins and Olwell (1998, pp. 1-5) two types of variability exists in a process:

1. Common cause variability occurs when there exists random variation inherent to the nature of the process. The observations on the measurable property will have a statistical distribution characterized by a location and a scale parameter. Both the location and scale parameters of the distribution are important. If the location of the distribution differs from what is regarded as acceptable, the quality of the process is compromised. If the spread of the distribution tends to be too large then excessive variability will be present in the observations leading to an excessive number of observations in the tails of the distribution. The latter is also not desirable.
2. Special cause variability, on the other hand, occurs when some predictable or systematic failure results in a change in the location of the distribution or a change in the dispersion thereof. This type of variability may be due to manufacturing or laboratory errors and may be, amongst others, a source of quality problems or financial losses.

The fundamental objective of statistical process control is to identify the second source of variability by regular monitoring and evaluation of the process. However, once identified, common cause variability can only be rectified by making fundamental changes to the process design and operations.

The standard normal CUSUM relies on three assumptions in defining the in-control model. These assumptions are:

1. Statistically independent and identically distributed random variables.
2. An underlying normal distribution.
3. The mean and variance of the distribution are at their desired levels.

The majority of results in the literature on the CUSUM require specific parametric assumptions. The assumption most often made is that the observations are generated from a normal distribution. However, there are many instances in which this assumption is known to be false, see for instance Chapter 5.

To overcome the difficulties presented by parametric assumptions, nonparametric procedures have been developed. Sequential rank methods have been applied in CUSUM-like procedures by Reynolds (1975) and Bhattacharya and Frierson (1981). The sequential detection procedure developed by Reynolds (1975) relies on the assumption that the data come from an underlying symmetric distribution. His approach is based on a truncated test where an upper bound is placed on the number of observations and is therefore not a fully fledged CUSUM procedure. If the procedure reaches the upper bound on the number of observations without signalling a change in distribution then the procedure is restarted from scratch. Bhattacharya and Frierson (1981) developed a similar procedure based on unsigned sequential ranks. Lombard (1983) generalises the control chart of Bhattacharya and Frierson (1981).

Another CUSUM-like procedure assuming symmetry of the underlying distribution around a known value, but based on the ordinary ranks of groups of observations, was developed by Bakir and Reynolds (1979). Bakir (2006) extends this work to the case where the point of symmetry is unknown. Given groups of n observations the usual Wilcoxon signed rank statistic is calculated within each group and the CUSUM procedure is based on this sequence of signed rank statistics. The fixed group size together with the independence of the groups enables one to find control limits using the Markov chain approach of Brook and Evans (1972). A discussion can be found in Hawkins and Olwell (1998, Chapter 6). This procedure is reminiscent of the Xbar chart except that the sample mean in a group is replaced by the Wilcoxon signed rank statistic in that group. Bandyopadhyay and Mukherjee (2007) introduce a nonparametric sequential detection procedure analogous to that of Bakir and Reynolds (1979) in the sense that their method also applies to grouped observations rather than individual observations. The procedure of Bandyopadhyay and Mukherjee (2007), however, differs from that of Bakir and Reynolds (1979) in that no symmetry assumption of the underlying process is made.

A nonparametric CUSUM procedure based on individual observations is developed by McDonald (1990). He uses unsigned sequential ranks together with the fact that these are independently distributed and converge in distribution to independent uniformly distributed random variables. He then develops a CUSUM procedure for these uniform random variables, the expectation being that the results would also be applicable to CUSUM procedures based directly on the sequential ranks. However, the convergence to the uniform distribution occurs at a rather slow rate. This leads to the procedure requiring a large start-up time or initial sample. It is then possible that a change in distribution already occurs during this startup time and will then go unnoticed.

The preceding methodologies are concerned with signalling a change in the location parameter of a process, but neglects other types of structural change such as a change in dispersion. An advance in this regard is due to Ross and Adams (2012) who developed CUSUM procedures to detect arbitrary changes in a distribution, be it a change in dispersion or a change in skewness or some other characteristic of the distribution. This they accomplish by adapting the changepoint formulation of the CUSUM suggested by Hawkins et al. (2003), applying it to the Kolmogorov-Smirnov and Cramér-Von Mises statistics.

1.2 Overview of the dissertation

In the present work we propose CUSUM procedures based on signed and unsigned sequential ranks to detect a change in location or dispersion. These CUSUM procedures do not require any specific parametric assumptions. The only assumptions are that the underlying distribution is continuous and symmetric around its median. Moreover, the procedures do not require the existence of any moments and can therefore be applied to distributions with infinite variances such as the Cauchy distribution. In view of the latter fact we will use the term "dispersion" to describe variability.

Our approach differs from that of McDonald (1990) in that we do not use the uniformly distributed random variables, but we apply the CUSUM procedure directly to the signed and unsigned sequential ranks. We exploit the rapid convergence of the partial sums of these quantities to normality. We also take advantage

of the ease with which Monte Carlo simulation can be applied to our procedures in order to obtain control limits irrespective of the underlying distribution. Furthermore, our CUSUM procedure to detect a change in the median does not require an initial in-control sample of observations. The results in Chapters 3 and 4 of this dissertation are, to the best of our knowledge, new.

The symmetry assumption that we make is by no means merely a matter of convenience. The application that gave rise to the work in this dissertation involves paired observations (V, W) observed sequentially over time. The observations within each pair are correlated. An example of such an application is coal quality determinations V and W by two independent laboratories employing the same methodologies. Here V and W would be exchangeable, that is (V, W) and (W, V) would have the same joint distributions. Then the difference $Z = V - W$ should be symmetrically distributed around zero. A non-zero symmetry point indicates bias between the laboratories. Moreover, it is desired that the variance of Z remain at its current level. Ideally, the Z-values would be normally distributed, but this is not the case in our particular application. However, if V and W are identically distributed the differences should follow a symmetric distribution (be it normal or non-normal) around a zero median. This application is described in detail in Chapter 5.

The structure of this dissertation will now be outlined. Chapter 2 discusses the standard normal CUSUM procedure, our primary reference being Hawkins and Olwell (1998). We provide the theoretical background and illustrate its non-robustness to deviations from normality. This provides the justification for developing distribution-free CUSUM procedures.

In Chapter 3 we introduce the signed sequential rank CUSUM (SSR CUSUM) to detect a change in the median of a symmetric distribution. We study the in- and out-of-control properties by theoretical means and Monte Carlo simulation.

Chapter 4 takes a similar form to Chapter 3. We introduce the unsigned sequential rank CUSUM (USR CUSUM) to detect a change in the dispersion of a symmetric distribution and study its properties, again by theoretical calculations supplemented by Monte Carlo simulations.

In Chapter 5, the application of the CUSUM procedures is illustrated on some data sets from the fields of engineering and finance.

Finally, we discuss a number of matters for further research.

Chapter 2

Introduction to the CUSUM

This chapter provides an overview of the standard normal cumulative sum (CUSUM) procedure for data on the real line, concentrating on the theoretical background. The focus falls predominantly on procedures designed to detect changes in the mean and variance of the normal distribution.

2.1 The standard normal CUSUM for the mean

The standard normal CUSUM procedure has as its objective to detect a change in the mean of a normal distribution as soon as possible after a change has occurred. Let the random variables $X_1, X_2, \cdots, X_\tau, X_{\tau+1}, \cdots$ be independent. $X_1, X_2, \cdots, X_\tau$ are identically distributed $N(0,1)$ which is the in-control distribution. The out-of-control values $X_{\tau+1}, X_{\tau+2}, \cdots$ are $N(\delta, 1)$, $\delta > 0$, random variables. Thus τ denotes the changepoint which is fixed but unknown. $\tau = \infty$ indicates a sequence that remains forever in control, while $\tau = 0$ indicates a sequence that is out of control from the start.

The assumption of unit variance can be replaced by one in which the variance is any known value σ. Then X/σ has unit variance and the in- and out-of-control means are 0 and δ/σ, respectively. Thus, we may take the variance to be 1 without loss of generality, provided we interpret δ in units of the true underlying value of σ.

In order to derive a useful sequential procedure, we start by considering in-control random variables $X_1, \cdots, X_n$ with means $\mu_1, \cdots, \mu_n$ and unit variance and formulate the following hypotheses:

$$H_0 : \qquad \mu_1 = \cdots = \mu_n = 0$$

and, for $1 \le \tau \le n$,

$$\begin{aligned} H_\tau : \qquad & \mu_1 = \cdots = \mu_\tau = 0 \\ & \mu_{\tau+1} = \cdots = \mu_n = \delta. \end{aligned}$$

The likelihood functions under H_0 and H_τ are

$$\lambda_0 \;=\; \prod_{i=1}^{n} \phi(X_i)$$

and

$$\lambda_\tau \;=\; \prod_{i=1}^{\tau} \phi(X_i) \prod_{i=\tau+1}^{n} \phi(X_i - \delta),$$

respectively, where $\phi(\cdot)$ denotes the standard normal probability density function. The resulting log-likelihood ratio is

$$\begin{aligned} L_\tau &= \sum_{i=\tau+1}^{n} \log(\phi(X_i - \delta)/\phi(X_i)) \\ &= \delta\Big(\sum_{i=\tau+1}^{n} (X_i - \delta/2)\Big) \\ &= \delta\Big(\sum_{i=1}^{n} (X_i - \delta/2) - \sum_{i=1}^{\tau} (X_i - \delta/2)\Big) \\ &= \delta\big(S_n - S_\tau\big) \end{aligned}$$

where

$$S_k \;=\; \sum_{i=1}^{k} (X_i - \delta/2)$$

for $k \ge 1$. Then the maximised log-likelihood ratio over τ is

$$\max_{1 \le \tau \le n} L_\tau \;=\; \delta\big(S_n - \min_{1 \le k \le n} S_k\big).$$

H_0 is rejected when this maximised log-likelihood ratio, which we denote by $\widetilde{L}_n$, is sufficiently large.

When observations are made sequentially, it now seems reasonable to reject H_0 as soon as $\widetilde{L}_n$ becomes sufficiently large. For this, define

$$D_n = \begin{cases} S_n - \min_{1 \le k \le n} S_k, & n \ge 1 \\ 0, & n = 0 \end{cases} \tag{2.1}$$

and

$$T = \min\{n \ge 1 : D_n \ge h\}. \tag{2.2}$$

Then H_0 is rejected at observation number $n = T$. The "critical value" for rejection, h, is known as the control limit while T is the run length. We say that the CUSUM signals a change at time T. $T = \infty$ means that the CUSUM never signals. One interesting feature of this CUSUM process is that $\mathbb{E}_\tau[T]$, the average run length, is finite regardless of whether the process remains in control or not, that is, whether $\tau = \infty$ or $\tau < \infty$ – see for example Siegmund (1985, Chapter 2, Section 6). In particular this means that the type I error probability equals 1. Consequently, the in-control properties of the CUSUM are typically specified in terms of average run length rather than probability of falsely signalling a change in the mean. To design a CUSUM scheme we therefore first fix $\mathbb{E}_\infty[T]$, the in-control ARL, by choosing an appropriate value of the control limit h. The larger the required ARL, the larger h will be. The rate at which false signals occur is $1/\mathbb{E}_\infty[T]$ and it therefore seems reasonable to fix this rate at a small value. This rate plays the same role as a significance level does in an ordinary fixed sample hypothesis test.

Given δ and h, a measure of the performance of the CUSUM when the out-of-control mean μ is positive is the out-of-control ARL $\mathbb{E}_0[T|\mu]$. This is the average run length assuming that the process starts out of control with mean μ. Software and tables are widely available from which this value can easily be obtained, for example anyarl.exe from https://www.stat.umn.edu/cusum/software.htm – see Hawkins and Olwell (1998, Chapter 10).

For computational purposes, it is convenient to express D_n in the following recursive form,

$$D_n = \begin{cases} \max(0, D_{n-1} + X_n - \delta/2), & n \geq 1 \\ 0, & n = 0 \end{cases} \tag{2.3}$$

where the quantity $\delta/2$ is referred to as the reference value. That equations (2.3) and (2.1) are equivalent can be seen as follows. From (2.1) we obtain

$$\begin{aligned} D_{n+1} &= S_{n+1} - \min_{1 \leq k \leq n+1} S_k \\ &= S_{n+1} - \min(S_{n+1}, \min_{1 \leq k \leq n} S_k) \end{aligned}$$

and

$$\begin{aligned} D_{n+1} - D_n &= S_{n+1} - \min(S_{n+1}, \min_{1 \leq k \leq n} S_k) - (S_n - \min_{1 \leq k \leq n} S_k) \\ &= (X_{n+1} - \delta/2) - \min(S_{n+1}, \min_{1 \leq k \leq n} S_k) + \min_{1 \leq k \leq n} S_k \\ &= \begin{cases} X_{n+1} - \delta/2 & \text{if } S_{n+1} \geq \min_{1 \leq k \leq n} S_k \\ X_{n+1} - \delta/2 - (S_{n+1} - \min_{1 \leq k \leq n} S_k) & \text{if } S_{n+1} < \min_{1 \leq k \leq n} S_k. \end{cases} \end{aligned}$$

Consequently,

$$\begin{aligned} D_{n+1} &= \begin{cases} D_n + X_{n+1} - \delta/2 & \text{if } S_{n+1} \geq \min_{1 \leq k \leq n} S_k \\ D_n + X_{n+1} - \delta/2 - (S_{n+1} - \min_{1 \leq k \leq n} S_k) & \text{if } S_{n+1} < \min_{1 \leq k \leq n} S_k \end{cases} \\ &= \begin{cases} D_n + X_{n+1} - \delta/2 & \text{if } S_{n+1} \geq \min_{1 \leq k \leq n} S_k \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} D_n + X_{n+1} - \delta/2 & \text{if } D_n + X_{n+1} - \delta/2 > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \max(0, D_n + X_{n+1} - \delta/2). \end{aligned}$$

The representation (2.3) of D_n also makes it clear that the CUSUM is a Markov process.

Since

$$\mathbb{E}_0[X_n - \delta/2 | \mu = 0] \quad < \quad 0$$

the sequence $(D_n)_{n\geq 1}$ will have a downward drift with an elastic barrier at 0 when the process is in control. As a result, the CUSUM will reach the control limit h infrequently. On the other hand, when the process is out of control, namely when $\mu \geq \delta$, then

$$\mathbb{E}_0[X_n - \delta/2 | \mu] \quad > \quad 0,$$

and the CUSUM will exhibit a strong upward drift, which causes it to reach the control limit h rather quickly. The transition from a downward to an upward drift is the key feature of the CUSUM procedure which indicates that a change in distribution has possibly occurred.

It has yet to be explained what happens when $\mu < \delta$. The CUSUM is designed on the premises that changes in the mean larger than δ are considered substantial, that is when $\mu \geq \delta$ the process is out of control and the level of change in the mean is detrimental to the output that the process delivers. However, changes in the mean smaller than δ are deemed "acceptable" in an attempt to keep unproductive tinkering to the process to a minimum.

Once the CUSUM signals a change, it becomes important to estimate the time point when the mean possible changed. Assuming that the change is real, we wish to estimate τ. The changepoint estimate of Page (1954) is

$$\widehat{\tau} \quad = \quad \max\{1 \leq n \leq T-1 : D_n = 0\}, \tag{2.4}$$

while the maximum likelihood estimator is – see Samuel et al. (1998) –

$$\widehat{\tau} \quad = \quad \underset{1\leq t\leq T-1}{\arg\max} \left\{ (T-t) \left(\overline{\overline{X}}_{T,t} \right)^2 \right\} \tag{2.5}$$

where

$$\overline{\overline{X}}_{T,t} \quad = \quad \sum_{i=t+1}^{T} \overline{X}_i / (T-t).$$

Both changepoint estimators are evaluated by Pignatiello and Samuel (2001). They found that the estimator (2.5) performed better than the estimator of Page (1954) in both precision and accuracy. They conclude that the estimator due to Page (1954) generally underestimates the changepoint when the magnitude of the change exceeds that for which the CUSUM was designed, namely δ. The maximum likelihood estimator (2.5) is found to be less biased in this circumstance. For smaller changes there is little to choose between the estimators.

For illustrative purposes we show two typical CUSUM paths in Figure 2.1: one for which the reference value is $\delta/2$, the other with δ, where δ is the target value. We use out-of-control $N(\mu, 1)$ data for both CUSUM paths and specify the value of μ as 0.5. The CUSUM is designed according to the control limit 7.267 corresponding to an in-control nominal ARL 500 and target value $\delta = 0.5$. The control limit is obtained from anygeth.exe of Hawkins and Olwell (1998, Chapter 10). The first CUSUM signals a change at a short run length 29 and the changepoint estimate due to Page (1954) is $\widehat{\tau} = 2$. On the other hand, the second CUSUM signals a change at a much later time 74 with the corresponding changepoint estimate $\widehat{\tau} = 40$. Clearly, the CUSUM designed with reference value $\delta/2$ delivers more accurate and reliable results since a change is signalled much sooner.

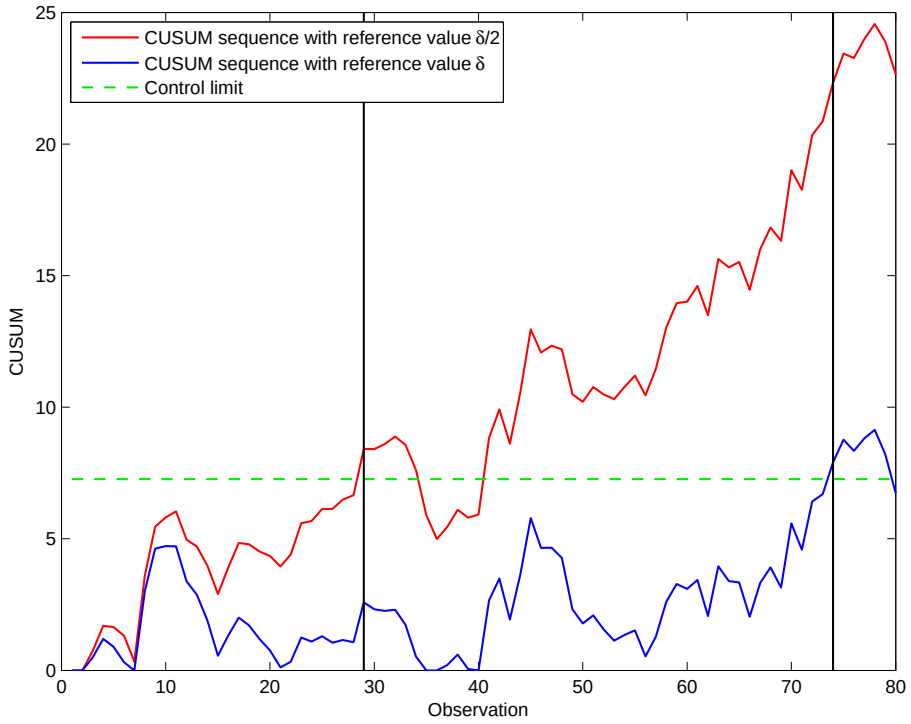

Figure 2.1: Two typical CUSUM sequences for in-control ARL 500 and target change size $\delta = 0.5$.

2.2 The standard normal CUSUM for variance

Assume an in-control distribution which is $N(0,1)$ and an out-of-control distribution which is $N(0,\sigma^2)$ for $\sigma > 1$. As before, the random variables $X_1, X_2, \cdots, X_\tau, X_{\tau+1}, \cdots$ are independent. τ again denotes the changepoint and is fixed but unknown.

Towards deriving a useful sequential procedure, we consider the in-control random variables $X_1, \cdots, X_n$ with zero means and variances $\sigma_1, \cdots, \sigma_n$ and formulate the hypotheses

$$H_0 : \quad \sigma_1 = \cdots = \sigma_n = 1$$

and for $1 \le \tau \le n$

$$\begin{aligned} H_\tau : \quad & \sigma_1 = \cdots = \sigma_\tau = 1 \\ & \sigma_{\tau+1} = \cdots = \sigma_n = \lambda. \end{aligned}$$

The log-likelihood ratio is

$$\begin{aligned} L_\tau &= \sum_{i=\tau+1}^{n} \log(\phi(X_i/\lambda)/(\lambda\phi(X_i))) \\ &= \sum_{i=\tau+1}^{n} \log((2\lambda^2\pi)^{-1/2} e^{-(X_i/\lambda)^2/2} / (2\pi)^{-1/2} e^{-X_i^2/2}) \\ &= \sum_{i=\tau+1}^{n} (X_i^2(1-\lambda^{-2})/2 + \log(\lambda^{-1})) \\ &= ((1-\lambda^{-2})/2) \sum_{i=\tau+1}^{n} (X_i^2 - 2\log(\lambda^{-1})/(\lambda^{-2}-1)) \\ &= ((1-\lambda^{-2})/2)\Big(\sum_{i=1}^{n} (X_i^2 - 2\log(\lambda^{-1})/(\lambda^{-2}-1)) - \sum_{i=1}^{\tau} (X_i^2 - 2\log(\lambda^{-1})/(\lambda^{-2}-1))\Big) \\ &= ((1-\lambda^{-2})/2)(S_n - S_\tau) \end{aligned}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution and

$$S_k = \sum_{i=1}^{k} (X_i^2 - 2\log(\lambda^{-1})/(\lambda^{-2}-1)), \quad k \ge 1.$$

Then the maximised log-likelihood ratio over τ is

$$\max_{1\le\tau\le n} L_\tau \;=\; ((1-\lambda^{-2})/2)(S_n - \min_{1\le k\le n} S_k).$$

H_0 is rejected when this maximised log-likelihood ratio, denoted by $\widetilde{L}_n$, is sufficiently large.

When observations are made sequentially, it again seems reasonable to reject H_0 when $\widetilde{L}_n$ is sufficiently large. Therefore, we define the run length T as in (2.2). We express D_n in recursive form, as before,

$$D_n \;=\; \begin{cases} \max(0, D_{n-1} + X_n^2 - \zeta), & n \ge 1 \\ 0, & n = 0 \end{cases} \tag{2.6}$$

where

$$\zeta \;=\; 2\log(\lambda^{-1})/(\lambda^{-2}-1) \text{ for } \lambda > 1.$$

Looking at (2.6) we see that this CUSUM has exactly the same form as the mean CUSUM, apart from the reference value and the presence of X_n^2 rather than X_n in the recursion formula. Therefore, the application of the variance CUSUM proceeds along the same lines as that of the mean CUSUM.

2.3 Non-robustness of the standard normal CUSUM

The question arises how the standard normal CUSUM performs when the underlying distribution deviates from normality. To answer this, we will compare the in-control ARL when the underlying distribution is non-normal with the nominal ARL when normality is assumed.

The simulation routine to estimate the in-control average run length proceeds as follows. The control limits h guaranteeing nominal in-control ARLs, denoted by ARL_0, of 125, 500 and 1000 at the typical choices $\delta = 0.5$ and 1 for a normal distribution, are applied – see Table 2.1. Data are then generated from a non-normal

distribution with zero mean and unit variance and the run length is found. This process is repeated 100 000 times, yielding an estimated ARL. We use the following symmetric distributions, standardised to have zero mean and unit variance:

1. Logistic distribution (light-tailed).
2. Student's t-distribution with 3 degrees of freedom (heavy-tailed).

Table 2.1 shows the standard normal CUSUM designs used in the simulations, with ARL_0 denoting the in-control average run length.

Nominal $\mathbf{ARL_0}$	$\boldsymbol{\delta}$	
	0.5	1.0
125	4.788	3.057
500	7.267	4.389
1000	8.585	5.071

Table 2.1: The standard normal CUSUM designs used in the simulation. The entries in the body of the table are the control limits h.

The simulation estimates obtained from 100 000 runs in each of the six designs are shown in Table 2.2.

Nominal $\mathbf{ARL_0}$	**Distribution**	$\boldsymbol{\delta}$	
		0.5	1.0
125	Logistic	127	115
	t_3	175	139
500	Logistic	491	406
	t_3	549	334
1000	Logistic	965	766
	t_3	932	494

Table 2.2: Simulation ARL_0 estimates for two distributions.

It is evident from Table 2.2 that only for the logistic distribution, at relatively small values of δ, are the simulation estimates of ARL_0 acceptable. In all other instances, the estimated ARL_0 results show unacceptably large deviations from the nominal ones.

The agreement in the case of the variance CUSUM is even worse. Table 2.4 shows the Monte Carlo simulation estimates obtained from 100 000 runs for each of the variance CUSUM designs in Table 2.3. The differences between nominal and simulated ARL_0 are unacceptably large in all cases considered.

Nominal $\mathbf{ARL_0}$	$\boldsymbol{\lambda}$	
	1.25	1.50
125	9.259	7.679
500	15.441	12.169
1000	18.892	14.562

Table 2.3: The standard normal CUSUM designs used in the simulation. The entries in the body of the table are the control limits h.

Nominal $\mathbf{ARL_0}$	**Distribution**	$\boldsymbol{\lambda}$	
		1.25	1.50
125	Logistic	80	73
	t_3	24	23
500	Logistic	218	182
	t_3	39	36
1000	Logistic	347	278
	t_3	47	44

Table 2.4: Simulation ARL_0 estimates for two distributions.

Evidently, the standard normal CUSUM procedure is not robust against even quite moderate deviations from normality. There are two approaches to dealing with this problem. Consider first the instance where the underlying distribution is known, but non-normal. In this case one can design an appropriate CUSUM using results from Chapter 6 of Hawkins and Olwell (1998). In particular, the appropriate control limits can be obtained using the Markov chain method. It is, however, rarely the case that the true underlying distribution is known precisely. The second approach, which we follow, involves the construction of CUSUM procedures that are distribution-free when the process is in control. The signed and unsigned sequential ranks of the observations are in fact distribution-free under the in-control assumption. Chapters 3 and 4 discuss the construction, implementation and evaluation of these sequential rank CUSUMs.

Chapter 3

A signed sequential rank CUSUM for location

This chapter introduces a CUSUM procedure based on signed sequential ranks with the aim of detecting a change in location of target size $\delta > 0$. The particular CUSUM procedure is designed specifically for symmetric distributions. The CUSUM is in fact distribution-free when the distribution is in control and is robust against outlier effects.

3.1 Design of the CUSUM

Assume that the i.i.d. random variables $X_1, X_2, \cdots, X_\tau$ follow a distribution which is symmetric around 0. The scale parameter, σ, may or may not be known. Suppose that a change in location of size $\delta > 0$ occurs at time $\tau < \infty$ (an upward change). The i.i.d. out-of-control random variables $X_{\tau+1}, X_{\tau+2}, \cdots$ have the same distribution as $X_1 + \delta$. $\tau = \infty$ indicates a sequence which remains in control. Define

$$s_i = \text{sign}(X_i) = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i = 0 \\ -1 & \text{if } X_i < 0 \end{cases}$$

and denote by R_i^+ the sequential rank of $|X_i|$ among $|X_1|, \cdots, |X_i|$, that is,

$$R_i^+ \quad = \quad \sum_{j=1}^{i} \mathbb{I}(|X_j| \leq |X_i|)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Consider the i.i.d. random variables $X_1, X_2, \cdots, X_n$. The corresponding sequential ranks of this sequence are independent and distributed according to

$$\mathbb{P}(R_i^+ = j) \quad = \quad 1/i \quad \text{for } j = 1, 2, \cdots, i \text{ and } i = 1, 2, \cdots, n. \tag{3.1}$$

A proof of this fact may be found in Barndorff-Nielsen (1963). Therefore, R_i^+ is uniformly distributed on the set $\{1, 2, \cdots, i\}$ no matter what the distribution underlying the data may be. The signed sequential rank of X_i among $X_1, X_2, \cdots, X_i$, is

$$V_i^+ \quad = \quad s_i \cdot R_i^+/(i+1).$$

Now, if $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with a continuous distribution symmetric around 0, then the following properties hold:

1. $V_1^+, V_2^+, \cdots, V_n^+$ are independent,
2. $F(-x) \cdot (1 - F(0)) = F(0) \cdot (1 - F(x))$ for all $x \geq 0$,
3. $|X_i|$ and $\text{sign}(X_i)$ are independent for $i = 1, 2, \cdots, n$,
4. R_i^+ and $\text{sign}(X_i)$ are independent for $i = 1, 2, \cdots, n$,

see Reynolds (1975, Theorem 2.1). Since X_i is symmetrically distributed around 0, we have

$$\mathbb{P}(s_i = 1) \quad = \quad \mathbb{P}(s_i = -1) \quad = \quad 1/2.$$

Then, by statement 4 above,

$$\mathbb{E}_\infty[V_i^+] \quad = \quad \mathbb{E}_\infty[s_i \cdot R_i^+/(i+1)] \quad = \quad 0.$$

The variance of V_i^+ is

$$\begin{aligned} \mathbb{V}\text{ar}_\infty[V_i^+] &= \mathbb{E}_\infty[(\text{sign}(X_i)(R_i^+/(i+1)))^2] \\ &= \mathbb{E}_\infty[(R_i^+/(i+1))^2] \\ &= \Big(\sum_{k=1}^{i} k^2\Big)/(i(1+i)^2) \\ &= (2i+1)/(6(i+1)), \end{aligned}$$

the next to last equality following from (3.1). Then

$$V_i \;=\; \sqrt{6(i+1)/(2i+1)}V_i^+ \tag{3.2}$$

is the standardised version of V_i^+. $V_1, V_2, \cdots, V_n$ are then independent with zero means and unit variances and, furthermore, V_i converges in distribution to a $U(-\sqrt{3}, \sqrt{3})$ random variable as $i \to \infty$.

At this point it is necessary to remark on the essential dissimilarities between the assumptions underlying our method and those made by Bakir and Reynolds (1979) and Bakir (2006). In the first instance, our method uses individual observations and not grouped observations. Secondly, our method assumes a known point of symmetry, or rather a known median, in contrast to the approach followed by Bakir (2006).

Suppose now that the median increases at time $\tau < \infty$. It is shown in Section 3.2.2 that the signed sequential ranks of $X_{\tau+1}, X_{\tau+2}, \cdots$ will then tend to be larger than those of $X_1, X_2, \cdots, X_\tau$ and consequently,

$$\mathbb{E}_\tau[V_i] \;>\; 0 \quad \text{for } i > \tau.$$

The latter feature provides the main motivation for the CUSUM based on signed sequential ranks (the SSR CUSUM). The upward SSR CUSUM is accordingly defined by

$$D_n \;=\; \begin{cases} \max(0, D_{n-1} + V_n - k), & n \geq 1 \\ 0, & n = 0, \end{cases} \tag{3.3}$$

where the reference value k is a non-negative constant. We apply the same sequential detection routine as in Chapter 2: the i.i.d. random variables $X_1, X_2, \cdots$ are replaced by their standardised signed sequential ranks $V_1, V_2, \cdots$ and a change is signalled as soon as $D_n \geq h$ where h is a positive control limit. To complete the design of an SSR CUSUM, we need to specify a value of k and find h that will guarantee a prescribed in-control ARL. These questions are discussed in Section 3.2.

In the event that an increase in the median is signalled, the changepoint τ can be estimated as the last observation where $D_n = 0$ – see (2.4). An analogue of the maximum likelihood estimator of τ is obtained upon replacing $X_1, X_2, \cdots$ by $V_1, V_2, \cdots$ in (2.5), that is,

$$\widehat{\tau} \;=\; \underset{1\le t\le T-1}{\arg\max} \left\{ (T-t)\left(\overline{\overline{V}}_{T,t}\right)^2 \right\}$$

where

$$\overline{\overline{V}}_{T,t} \;=\; \sum_{i=t+1}^{T} \bar{V}_i/(T-t).$$

However, this estimator has to date not been reviewed in academic literature and constitutes a possible matter for further research.

3.2 The in-control behaviour of the CUSUM

3.2.1 Determination of control limits

Our first order of business is to determine control limits h for given reference values k and specified in-control nominal ARLs. The distribution-free character of the signs and sequential ranks enables us to obtain control limits for the SSR CUSUM by Monte Carlo simulation using $U(-1, 1)$ random variables. Table 3.1 shows estimated Monte Carlo control limits for selected values of k and nominal ARL_0.

Nominal $\mathbf{ARL_0}$	**k**			
	0.125	0.250	0.375	0.500
100	6.000	4.500	3.485	2.750
200	7.950	5.650	4.300	3.380
300	9.170	6.350	4.775	3.695
400	10.095	6.851	5.095	3.945
500	10.850	7.267	5.420	4.145

Table 3.1: Estimated control limits for the upward SSR CUSUM.

To check the appropriateness of these estimated control limits, independent Monte Carlo simulations were done with 100 000 runs in each cell of Table 3.2. The estimated in-control ARL values found are shown in Table 3.2. The agreement between estimated and nominal values is excellent in all cases. Where discrepancies occur these err on the conservative side, that is estimated in-control ARL_0 values are slightly larger than the nominal values.

Nominal ARL_0	**k**			
	0.125	0.250	0.375	0.500
100	99	102	102	101
200	200	205	207	213
300	297	301	309	304
400	403	403	403	406
500	506	502	496	508

Table 3.2: Estimated ARL_0 results for combinations of nominal ARL and reference values for the upward SSR CUSUM.

We observe that the control limits in Table 3.1 for $k \leq 0.25$ are quite close to those of the standard normal CUSUM. On the other hand, for $k > 0.25$ the correspondence is poor, the estimated in-control ARLs being much larger than the nominal values. We provide two examples:

1. The control limit corresponding to an in-control nominal ARL 500 and reference value 0.25 for the SSR CUSUM is 7.267 (Table 3.1) which is equal to the corresponding control limit for the standard normal CUSUM (Table 2.1) for $\delta = 0.5$.
2. If we look at a reference value 0.5 for the SSR CUSUM, the control limit from Table 3.1 is 4.145 for a nominal in-control ARL 500. This control limit differs substantially from that of the standard normal CUSUM which is 4.389 (from Table 2.1).

These observations can be explained as follows.

Consider the standardised independent random variables $V_1, V_2, \cdots, V_n$ defined in (3.2) and set

$$S_n = (V_1 - k) + (V_2 - k) + \cdots + (V_n - k).$$

Then

$$\begin{aligned} \mathbb{E}[S_n + nk] &= 0, \\ \mathbb{V}\text{ar}[S_n + nk] &= n \end{aligned}$$

and

$$(S_n + nk)/\sqrt{n} \xrightarrow{\mathcal{D}} N(0,1) \text{ as } n \to \infty,$$

see Theorem 5.1 of Mason (1981). Consequently,

$$S_n/\sqrt{n} = (S_n + nk)/\sqrt{n} - \sqrt{n}k \xrightarrow{\mathcal{D}} N(c,1) \text{ as } n \to \infty$$

if $k = c/\sqrt{n}$. The partial sums S_n for the SSR CUSUM become normal and, therefore, upon comparing the result with the representation of D_n in (2.1), it is not difficult to imagine that the control limits of the SSR CUSUM will correspond closely to those of the standard normal CUSUM when k is small.

In contrast, consider what happens when k is large. The V_i are bounded and therefore the probability that it will exceed the reference value, k, is smaller than the corresponding probability that a normally distributed X_i, which is unbounded, will exceed it, in other words

$$\mathbb{P}(V_i \geq k) < \mathbb{P}(X_i \geq k).$$

In fact,

$$\mathbb{P}(V_i \geq k) = 0 \quad \text{for } k > \sqrt{3}.$$

Consequently, it will take the SSR CUSUM longer to cross the control limit than the normal CUSUM.

3.2.2 Specification of the reference value

In the normal CUSUM the reference value is specified as one half of the number of standard deviations that the target mean of the X_i is away from zero. The simplest approach towards specifying the reference value in the SSR CUSUM is to choose the reference value as a number of underlying standard deviations of the V_i away from zero. However, we would really like to relate k to the actual target change size in the median of the X_i. In general this is not possible because no firm assumptions about the functional form of the underlying density function have been made. Nevertheless, we can obtain some guidance by calculating the reference values that would be applicable to a medium- or light-tailed distribution such as the normal and a heavy-tailed distribution such as the t_3-distribution.

To arrive at an appropriate reference value we will assess the behaviour of $\mathbb{E}_\tau[V_{\tau+j}]/2$, $j \geq 1$ for a given τ. We assess this quantity on each of the following two bases:

A.1. Suppose that $\tau = 0$, i.e. the process starts out of control and we then compute $\lim_{j\to\infty} \mathbb{E}_0[V_j|\delta]$.

A.2. Fix $j \geq 1$ and compute $\lim_{\tau\to\infty} \mathbb{E}_\tau[V_{\tau+j}|\delta]$, i.e. we assume that the process runs in control for a long time and we then assess the expected change in V_i.

We begin with assessment A.1. Here $X_1, X_2, \cdots$ are i.i.d. with median δ and unit variance, that is, their common distribution function is $G(x) = F(x - \delta)$ where F denotes the in-control distribution function. We have, for $j \geq 1$,

$$R_j^+ \quad = \quad 1 + \sum_{i=1}^{j-1} \mathbb{I}(|X_i| < |X_j|)$$

with

$$\sum_{i=1}^{0} \mathbb{I}(|X_i| < |X_1|) \quad = \quad 0.$$

Then,

$$\begin{aligned}\mathbb{E}[s_j \cdot R_j^+] &= \mathbb{E}[s_j] + \mathbb{E}[s_j \cdot \sum_{i=1}^{j-1} \mathbb{I}(|X_i| < |X_j|)] \\ &= (2F(\delta) - 1) + (j-1)\,\mathbb{E}[s_j \cdot \mathbb{I}(|X_i| < |X_j|)] \\ &= (2F(\delta) - 1) + (j-1)\,\mathbb{E}[s_2 \cdot \mathbb{I}(|X_1| < |X_2|)].\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}[s_2 \cdot \mathbb{I}(|X_1| < |X_2|)] &= \mathbb{P}(|X_1| < |X_2|, X_2 > 0) - \mathbb{P}(|X_1| < |X_2|, X_2 < 0) \\ &= \mathbb{P}(|X_1| < X_2, X_2 > 0) - \mathbb{P}(|X_1| < -X_2, X_2 < 0) \\ &= \int_0^\infty [F(w-\delta) - F(-w-\delta)] f(w-\delta) dw \\ &\qquad - \int_{-\infty}^0 [F(-w-\delta) - F(w-\delta)] f(w-\delta) dw \\ &= \int_{-\infty}^\infty F(w-\delta) f(w-\delta) dw - \int_{-\infty}^\infty F(-w-\delta) f(w-\delta) dw \\ &= 1/2 - \int_{-\infty}^\infty F(-w-\delta) f(w-\delta) dw \\ &= 1/2 - \int_{-\infty}^\infty [1 - F(x+2\delta)] f(x) dx \\ &= \int_{-\infty}^\infty F(x+2\delta) f(x) dx - 1/2\end{aligned}$$

where the next to last equality is obtained by using the relation $F(x) = 1 - F(-x), x \geq 0$ and the substitution $x = w - \delta$. Therefore,

$$\begin{aligned}\mathbb{E}[s_j \cdot R_j^+/(j+1)|\delta] &= (2F(\delta) - 1)/(j+1) \\ &\qquad + ((j-1)/(j+1))\Big(\int_{-\infty}^\infty F(x+2\delta) f(x) dx - 1/2 \Big) \\ &\to \int_{-\infty}^\infty F(x+2\delta) f(x) dx - 1/2\end{aligned}$$

as $j \to \infty$ and

$$\mathbb{E}_0[V_j|\delta] \to \sqrt{3}\Big(\int_{-\infty}^\infty F(x+2\delta) f(x) dx - 1/2 \Big).$$

Thus, the reference value is

$$k := k(F, \delta) = \sqrt{3/4}\Big(\int_{-\infty}^\infty F(x+2\delta) f(x) dx - 1/2 \Big). \tag{3.4}$$

Table 3.3 gives the values of k for the normal and the t_3-distribution for the typical range of values of δ.

	δ		
Distribution	0.25	0.50	1.00
Normal	0.12	0.23	0.36
t_3	0.10	0.18	0.31

Table 3.3: Reference values from (3.4).

We next consider approach A.2. Here it is useful to introduce two sequences of i.i.d. random variables $Y_1, Y_2, \cdots$ distributed according to F and $W_1, W_2, \cdots$ distributed according to G. Set

$$X_i \stackrel{\mathcal{D}}{=} \begin{cases} Y_i & \text{for } 1 \le i \le \tau \\ W_j & \text{for } i = \tau + j, j \ge 1. \end{cases}$$

By definition, we have

$$\begin{aligned} R^+_{\tau+j} &= 1 + \sum_{i=1}^{\tau+j-1} \mathbb{I}(|X_i| < |X_{\tau+j}|) \\ &= 1 + \sum_{i=1}^{\tau} \mathbb{I}(|X_i| < |X_{\tau+j}|) + \sum_{i=\tau+1}^{\tau+j-1} \mathbb{I}(|X_i| < |X_{\tau+j}|) \\ &= 1 + \sum_{i=1}^{\tau} \mathbb{I}(|X_i| < |X_{\tau+j}|) + \sum_{i=1}^{j-1} \mathbb{I}(|X_{\tau+i}| < |X_{\tau+j}|) \\ &= 1 + \sum_{i=1}^{\tau} \mathbb{I}(|Y_i| < |W_j|) + \sum_{i=1}^{j-1} \mathbb{I}(|W_i| < |W_j|) \\ &= 1 + \tau\Big(\sum_{i=1}^{\tau} \mathbb{I}(|Y_i| < |W_j|) \Big)/\tau + \sum_{i=1}^{j-1} \mathbb{I}(|W_i| < |W_j|) \\ &= 1 + \tau \cdot F^+_\tau(|W_j|) + \sum_{i=1}^{j-1} \mathbb{I}(|W_i| < |W_j|) \end{aligned}$$

where $F^+_\tau(\cdot)$ denotes the empirical distribution function of $|Y_1|, \cdots, |Y_\tau|$. Then

$$\mathbb{E}_\tau[R^+_{\tau+j}] = \tau\, \mathbb{E}[F^+_\tau(|W_j|)] + (j+1)/2$$

because

$$\mathbb{E}[\mathbb{I}(|W_i| < |W_j|)] = \mathbb{P}(|W_i| < |W_j|) = 1/2.$$

Moreover, $F^+_\tau(y)$ converges to $F^+(y)$ almost surely as $\tau \to \infty$ for every y where

$$F^+(y) = \mathbb{P}(|Y| \le y).$$

Consequently, for every fixed $j \geq 1$, as $\tau \to \infty$,

$$\mathbb{E}_\tau[R^+_{\tau+j}/(\tau+j+1)] \;\to\; \mathbb{E}[F^+(|W|)]$$

which does not depend upon j. Therefore,

$$\begin{aligned}
\mathbb{E}_\tau[s_j \cdot R^+_{\tau+j}/(\tau+j+1)] &\to \mathbb{E}[\text{sign}(W)\cdot F^+(|W|)] \\
&= 2\,\mathbb{E}[\text{sign}(W)\cdot F(|W|)] - \mathbb{E}[\text{sign}(W)] \\
&= 2\Big(\int_{-\infty}^{\infty} \text{sign}(w) F(|w|) g(w) dw\Big) - (2F(\delta)-1) \\
&= 2\Big(\int_0^{\infty} F(w)(f(w-\delta) - f(-w-\delta)) dw - F(\delta) + 1/2\Big),
\end{aligned}$$

so that

$$\mathbb{E}_\tau[V_j|\delta] \;\to\; \sqrt{3}\Big(2\Big(\int_0^{\infty} F(w)(f(w-\delta) - f(-w-\delta)) dw - F(\delta) + 1/2\Big)\Big)$$

as $\tau \to \infty$, which gives the reference value

$$k \;:=\; k(F,\delta) \;=\; \sqrt{3}\Big(\int_0^{\infty} F(w)(f(w-\delta) - f(-w-\delta)) dw - F(\delta) + 1/2\Big). \quad (3.5)$$

Table 3.4 gives the values of k for the normal and the t_3-distribution for the typical range of values of δ.

	$\boldsymbol{\delta}$		
Distribution	0.25	0.50	1.00
Normal	0.12	0.24	0.45
t_3	0.10	0.20	0.37

Table 3.4: Reference values from (3.5).

We notice that the quantities in Tables 3.3 and 3.4 are indeed very similar, except at $\delta = 1$. In particular, if A.2. is judged to be appropriate, we see from Table 3.4 that the choice $k = \delta/2$ proposed at the beginning of this section seems to be quite appropriate.

3.3 Relative efficiency

To evaluate the relative efficiency of the SSR CUSUM with respect to the standard normal CUSUM, it is necessary that both CUSUMs be comparable when the process $X_1, X_2, \cdots$ is in control. This means that both CUSUMs must have the same in-control ARLs and use the appropriate reference values. For a target out-of-control mean of δ, the appropriate reference value for the normal CUSUM is $\delta/2$ (see Chapter 2, Section 2.1). In view of our findings in Section 3.2.2, we will use for the SSR CUSUM the reference values given in the row corresponding to the normal distribution in Table 3.4. We define the relative efficiency of the SSR CUSUM with respect to the standard normal CUSUM when the out-of-control mean is μ by

$$e \;:=\; e(\mu) \;=\; \mathbb{E}_0[T_N|\mu]/\mathbb{E}_0[T_S|\mu] \quad \text{for } \mu > 0 \tag{3.6}$$

where T_N and T_S denote respectively the run lengths of the normal and SSR CUSUMs. In other words, the relative efficiency is calculated under the assumption that the process is out of control from the start, that is, on the basis of A.1. (see Section 3.2.2), which is the usual assumption made in the literature. We do not use A.2. (see Section 3.2.2) due to the impracticalities presented by the choice of a large enough τ.

Given δ, the target out-of-control mean, and an in-control ARL, the numerator in (3.6) can be found using the existing software (anyarl.exe) from Hawkins and Olwell (1998) for any choice of the reference value $k = \delta/2$ and the out-of-control mean μ. For the SSR CUSUM, the appropriate control limits were again found by Monte Carlo simulation and are given in Table 3.5. Notice from Table 3.1 that the control limits in Table 3.5 at $k = 0.12, 0.24$ and 0.45 are close to those in Table 3.1 at $k = 0.125, 0.25$ and 0.5. The denominator in (3.6) was also found by Monte Carlo simulation for the corresponding control limits and reference values (100 000 runs in each instance). The values of the numerator and the denominator of the ratio e in (3.6) are shown in Table 3.6 and the values of e itself are shown in Table 3.7.

Nominal ARL_0	k		
	0.12	0.24	0.45
100	6.06	4.63	3.04
500	10.63	7.57	4.81

Table 3.5: Control limits for the upward SSR CUSUM corresponding to the appropriate reference values $0.12, 0.24$ and 0.45.

		Normal CUSUM			SSR CUSUM		
μ	Nominal ARL_0	k			k		
		0.125	0.25	0.50	0.12	0.24	0.45
0.25	**100**	31	31	35	32	35	39
	500	65	71	98	67	79	119
0.5	**100**	15	15	16	19	19	21
	500	28	26	31	33	32	43
1.0	**100**	8	7	6	12	11	11
	500	13	10	9	19	17	18

Table 3.6: Out-of-control ARLs obtained from the standard normal and SSR CUSUM for data from a $N(\mu, 1)$ distribution.

μ	Nominal ARL_0	δ		
		0.25	0.50	1.00
0.25	**100**	0.97	0.89	0.90
	500	0.97	0.90	0.82
0.5	**100**	0.79	0.79	0.76
	500	0.85	0.81	0.72
1.0	**100**	0.67	0.64	0.55
	5000	0.68	0.59	0.50

Table 3.7: Relative efficiency of the SSR CUSUM with respect to the standard normal CUSUM.

Inspection of Table 3.7 reveals that the SSR CUSUM has high relative efficiency at small values of δ and μ, but that its performance degenerates at larger values of these parameters. This pattern could be expected because the SSR CUSUM is based on the *locally* most powerful test of the hypothesis $\mu = 0$ against $\mu > 0$ in the logistic

distribution – see Section 3.4.3. Thus, since the logistic and normal densities are not very different, this high *local* power of the SSR CUSUM is not entirely unexpected. The comparison between the SSR and normal CUSUM above is, of course, not entirely fair, because in the normal CUSUM the underlying variance is assumed to be known whereas no such assumption is made in the SSR CUSUM.

In order to attain comparability with the SSR CUSUM we implement a modified version of the self-starting normal CUSUM (Hawkins and Olwell, 1998, Section 7.2) which assumes that the in-control mean is zero but does not require that the variance be known. We will use the Monte Carlo method to gauge the efficacy of the SSR CUSUM. We consider six designs:

1. target mean shift is $\mu = 0.25$, in-control ARL is 100,
2. target mean shift is $\mu = 0.50$, in-control ARL is 100,
3. target mean shift is $\mu = 1.00$, in-control ARL is 100,
4. target mean shift is $\mu = 0.25$, in-control ARL is 500,
5. target mean shift is $\mu = 0.50$, in-control ARL is 500,
6. target mean shift is $\mu = 1.00$, in-control ARL is 500.

For standard normal CUSUMs, which presume a known variance, the reference values are $\mu/2$ and the control limits guaranteeing the required in-control ARLs are 6.00, 4.49, 2.76, 10.738, 7.35 and 4.13, respectively. For a given in-control ARL, the self-starting CUSUM uses the same reference values and control limits as the standard normal CUSUM. However, the summand V_n in (3.3) is now

$$V_n = \Phi^{-1}(T_{n-1}(W_n)),$$

where $W_n = X_n/s_{n-1}$ for $n \geq 2$ with

$$s_n^2 = \sum_{i=1}^{n} X_i^2/n. \tag{3.7}$$

T_{n-1} denotes the cdf of the t-distribution with $n-1$ degrees of freedom and Φ the cdf of the standard normal distribution. For the SSR CUSUMs we use the reference values $k = 0.1$ and $k = 0.2$ in the first four designs and $k = 0.36$ (from Table 3.3) in the last two cases. The corresponding control limits from Table 3.1 are 6.39, 11.79, 5.04, 8.49, 3.63 and 5.68.

Normal self-starting CUSUMs are thought to perform best when an initial number, m, of in-control observations are available which, so to speak, "calibrate" the CUSUM. In the simulations we therefore estimate the ARLs under each of the six designs using (a) $m = 0$ initial in-control observations and (b) $m = 20$ initial in-control observations.

Each of the "SSR" and "normal" ARL estimates in Table 3.8 comes from 10 000 Monte Carlo trials. Clearly, the SSR CUSUM compares very favourably with the self-starting normal CUSUM. The results in Table 3.8 confirm that the performance of the normal self-starting CUSUM improves with a larger number of startup observations. However, the effect of this larger number on the SSR CUSUM is less noticeable.

	m = 0		**m = 20**	
Design	**normal**	**SSR**	**normal**	**SSR**
1	31	32	30	31
2	15	16	15	15
3	14	10	7	7
4	66	70	67	67
5	33	31	28	29
6	21	15	11	12

Table 3.8: Out-of-control ARLs of the upward normal self-starting and SSR CUSUMs.

3.4 Concluding remarks

3.4.1 Two-sided CUSUMs

It is generally desirable to detect either upward or downward changes in the median of the underlying distribution. Thus far, we have only considered upward changes. Denote by D_n^+ the CUSUM to detect an upward change in the median, i.e.,

$$D_n^+ = \begin{cases} \max(0, D_{n-1}^+ + V_n - k), & n \geq 1 \\ 0, & n = 0. \end{cases}$$

An upward change in distribution is signalled at the random time

$$T^{+} \quad = \quad \min\{n \geq 1 : D_n^{+} \geq h^{*}\}.$$

The CUSUM to detect downward changes is

$$D_n^{-} \quad = \quad \begin{cases} \min(0, D_{n-1}^{-} + V_n + k), & n \geq 1 \\ 0, & n = 0 \end{cases}$$

where k is the non-negative reference value. The downward CUSUM terminates at time

$$T^{-} \quad = \quad \min\{n \geq 1 : D_n^{-} \leq -h^{*}\}.$$

The sequential detection routine proceeds analogously to the SSR CUSUM designed to detect upward changes in the median. The run length, that is the number of observations required for the two-sided SSR CUSUM to terminate, is

$$T^{*} \quad = \quad \min\{T^{+}, T^{-}\}.$$

h^{*} is the control limit chosen to make both $E_{\infty}[T^{+}]$ and $\mathbb{E}_{\infty}[T^{-}]$ equal to *twice* the nominal value of $E_{\infty}[T^{*}]$ – see Hawkins and Olwell (1998, Chapter 3, p. 55).

3.4.2 Justification for using sequential ranks

To motivate the use of sequential rank statistics rather than ordinary rank statistics in our procedure, we consider the i.i.d. random variables $X_1, X_2, \cdots, X_n, X_{n+1}, \cdots$ that arrive sequentially from the output of the process. Consider now the fixed sample $X_1, \cdots, X_n$ and denote by $R_{n:i}$ the ordinary rank of X_j among $X_1, \cdots, X_n$, that is,

$$R_{n:i} \quad = \quad \sum_{j=1}^{n} \mathbb{I}(X_j \leq X_i).$$

For a sample size $n = 1$ the rank $R_{1,1} = 1$ always. When $n = 2$, the ordinary ranks $R_{2,1}$ and $R_{1,2}$ are either 1 or 2 depending on whether $X_1 < X_2$ or $X_1 > X_2$. For different sample sizes $n = 1, 2, \cdots, r$ we have the sequences of ordinary ranks:

$$\begin{aligned} n = 1 &: R_{1,1} \\ n = 2 &: R_{2,1}, R_{2,2} \\ n = 3 &: R_{3,1}, R_{3,2}, R_{3,3} \\ &\cdots \\ n = r &: R_{r,1}, R_{r,2}, R_{r,3}, \cdots, R_{r,r}. \end{aligned} \tag{3.8}$$

The sequential rank of X_i among $X_1, \cdots, X_i$ is $R_{i,i}$, thus the diagonal entries of the matrix (3.8) above.

It is a known fact that one can deduce uniquely the ordinary ranks from the sequential ranks on the diagonal. This fact is explained: let the random variables $X_1, X_2, \cdots$ arrive in a sequential manner over time. We have seen that $R_{1,1} = 1$ always. Upon arrival of X_2 one is then able to determine from the values of $R_{2,2}$ and $R_{1,1}$ the value of $R_{2,1}$ and similarly from the value of $R_{3,3}$ one can deduce the values of $R_{3,1}$ and $R_{3,2}$. For illustration we consider the following example.

Example: Suppose $X_1 = 5$ arrives from the process and the sequential rank is $R_{1,1} = 1$. Next, we measure $X_2 = 3$ for which the sequential rank is $R_{2,2} = 1$ such that it is clear that then $X_2 < X_1$ from which we deduce that $R_{2,1} = 2$. If we then measure the next reading $X_3 = 9$ with corresponding sequential rank $R_{3,3} = 3$ we may deduce from the fact that $R_{2,1} = 2$ and $R_{2,2} = 1$ that $R_{3,1} = 2, R_{3,2} = 1$ and $R_{3,3} = 3$, yielding complete information about the sequence of ordinary ranks.

Note that the sequential ranks $R_{11}, R_{22}, \cdots$ are independent, whilst the sequences of ordinary ranks $\{R_{11}\}, \{R_{21}, R_{22}\}, \{R_{31}, R_{32}, R_{33}\}, \cdots$ are not independent. It is possible to construct a test statistic based on the ordinary ranks. However, this approach leads to cumbersome computations and since we have seen that the sequential ranks provide the same information as ordinary ranks it is convenient to apply sequential ranks to the test procedure rather than ordinary ranks.

3.4.3 Derivation of the SSR statistic

In this section we motivate the use of the signed sequential rank statistic in (3.2). Let $X_1, \cdots, X_n$ be i.i.d. random variables symmetrically distributed around μ. We will derive the locally most powerful test for

$$H_0 : \quad \mu = 0$$

against

$$H_\mu : \quad \mu > 0.$$

Denote by F and f the distribution function and density function of X, respectively, when $\mu = 0$. Then the density of X under H_μ is $f(x-\mu)$. The locally most powerful (LMP) test of H_0 is based on the score statistic

$$\sum_{i=1}^{n} \frac{\partial}{\partial \mu} \log(f(X_i - \mu)) \quad = \quad \sum_{i=1}^{n} \Big(\frac{\partial}{\partial \mu} f(X_i - \mu) \Big) / f(X_i - \mu)$$

evaluated at $\mu = 0$ – see Rao (2002, pp. 453-456). Now,

$$\frac{\partial}{\partial \mu} f(X_i - \mu)\Big|_{\mu=0} \quad = \quad -f'(X_i)$$

and

$$f(X_i - \mu)\Big|_{\mu=0} \quad = \quad f(X_i)$$

so that the score statistic is

$$\sum_{i=1}^{n} H(X_i) \quad = \quad -\sum_{i=1}^{n} f'(X_i)/f(X_i).$$

Since $f'(X_i) = -f'(-X_i)$, we see that

$$-H(-X_i) \quad = \quad f'(-X_i)/f(-X_i) \quad = \quad -f'(X_i)/f(X_i) \quad = \quad H(X_i)$$

so that $H(X_i)$ is an odd function of X_i. Thus, since

$$X_i \quad = \quad s_i \cdot |X_i|,$$

we have

$$H(X_i) \;=\; s_i \cdot H(|X_i|),$$

so that the score statistic is

$$T_n \;:=\; \sum_{i=1}^{n} H(X_i) \;=\; \sum_{i=1}^{n} s_i \cdot H(|X_i|).$$

To find the LMP signed sequential rank test statistic, we project T_n into the set of linear signed sequential rank statistics using the projection lemma in Hájek et al. (1999, p. 59). In order to find the projection, we define $|X|_{i:j}$ for $j \le i$ to be the j^{th} order statistic among $|X_1|, |X_2|, \cdots, |X_i|$. The projection is

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}[T_n | s_i, R_i^+] \;&=\; \sum_{i=1}^{n} s_i \cdot \mathbb{E}[H(|X_i|) | R_i^+] \\
&=\; \sum_{i=1}^{n} s_i \cdot \mathbb{E}[H(|X|_{i:R_i^+}) | R_i^+] \\
&=\; \sum_{i=1}^{n} s_i \cdot \mathbb{E}[H(|X|_{i:j})]\Big|_{j=R_i^+} \qquad (3.9)
\end{aligned}$$

since $|X_i| = |X|_{i:R_i^+}$ and the sequential ranks $R_i^+, i = 1, \cdots, n$ and the order statistics $|X|_{i:j}, i = 1, \cdots, n$ are independent – see Theorem 1 in Hájek et al. (1999, Chapter 2, p. 37).

Consider, in particular, the case where f is the density of the logistic distribution,

$$f(x) \;=\; e^{-x}/(1+e^{-x})^2, \;\; -\infty < x < \infty$$

with distribution function

$$F(x) \;=\; 1/(1+e^{-x}), \;\; -\infty < x < \infty.$$

Then some calculation shows that

$$H(x) \;=\; 2F(x) - 1 \;=\; F^+(x),$$

the distribution function of $|X|$. Consequently, $H(|X_i|)$ is uniformly distributed.

Thus, in the projection (3.9),

$$\mathbb{E}[H(|X|_{i:j})] \;=\; \mathbb{E}[U_{i:j}] \;=\; j/(i+1), \tag{3.10}$$

where $U_1, U_2, \cdots, U_i$ are independent $U(0,1)$ random variables. Substituting (3.10) into (3.9), we see that the locally most powerful signed sequential rank statistic is

$$\sum_{i=1}^{n} s_i \cdot R_i^+/(i+1),$$

which is the SSR statistic. This is clearly reminiscent of the well-known Wilcoxon signed rank statistic.

3.4.4 Asymmetric distributions

One might speculate that the SSR CUSUM will not be robust against deviations from symmetry. For illustrative purposes, we simulate data from a Gumbel distribution standardised to have zero mean and unit variance. The Gumbel is a moderately skew distribution. Assuming that the mean and variance of the distribution remain fixed, we estimate the average run length in a Monte Carlo simulation with 100 000 runs in each cell. The results are shown in Table 3.9. We use the control limits from Table 3.1. In Table 3.9 we see that the estimated ARL_0 values are substantially smaller than the nominal ARL_0 in all instances. This indicates that the SSR CUSUM signals an upward change in the mean when, in fact, there is no such change.

Nominal ARL$_0$	**k**			
	0.125	0.250	0.375	0.500
100	61	66	72	80
200	103	116	133	157
300	137	157	184	217
400	167	196	231	279
500	194	232	289	341

Table 3.9: Estimated ARL_0 when applying the SSR CUSUM to a skew (Gumbel) distribution.

At first sight this may seem to be a defect of the SSR CUSUM. However, it indicates that this CUSUM is also able to detect the onset of skewness in the underlying distribution. From this point of view the apparent defect can be seen as an added positive feature of the SSR CUSUM.

Chapter 4

A sequential rank CUSUM for dispersion

This chapter introduces a CUSUM procedure based on unsigned sequential ranks for detecting a change in dispersion. Similar to the SSR CUSUM, this particular CUSUM is designed specifically for symmetric distributions, is distribution-free when the process is in control and is also robust against the effects of outliers. Moreover, the sequential rank CUSUM developed in this chapter does not require the existence of any moments of the underlying distribution. We therefore use the term "dispersion" rather than "variance" to describe variability.

4.1 Design of the CUSUM

Let the in-control random variables $X_1, X_2, \cdots, X_\tau$ be i.i.d. from a distribution which is symmetric around 0. We make no assumption regarding the dispersion parameter σ, which may or may not be known. Suppose there occurs after the finite time τ a change in the dispersion of the distribution of size $\lambda > 1$ (an upward change), i.e. the out-of-control random variables $X_{\tau+1}, X_{\tau+2}, \cdots$ have the same distribution as λX_1.

We will assume throughout that the median of the X_i is known. By subtracting the median from the X_i, we can take without loss of generality the median as zero. We make no assumption regarding the numerical value of the in-control dispersion parameter σ. The CUSUM procedure will only detect changes away from the current level of dispersion, whatever that value may be. We use the symbol λ to denote the factor by which σ changes. The in-control process X/σ has unit dispersion and the out-of-control process has then dispersion $\lambda > 1$. We may therefore take, without loss of generality, the in-control scale parameter σ as 1.

Consider the sequence of i.i.d. random variables $X_1, X_2, \cdots$ from the in-control process. Suppose that a change in dispersion occurs at time $\tau < \infty$ and consider the absolute values

$$|X_1|, \cdots, |X_\tau|, \lambda|X_{\tau+1}|, \lambda|X_{\tau+2}|, \cdots, \tag{4.1}$$

the logarithms of which are

$$\log|X_1|, \cdots, \log|X_\tau|, \log\lambda + \log|X_{\tau+1}|, \log\lambda + \log|X_{\tau+2}|, \cdots.$$

These logarithms have the same sequence of sequential ranks as the absolute values in (4.1). Therefore, detecting a change in dispersion can be formulated as detecting a change in location of the logarithms of the absolute values. Note that the $\log|X_i|$ will generally not be symmetrically distributed around 0 and, therefore, we cannot use the SSR CUSUM. However, we can use the unsigned sequential ranks

$$R_i^+ \quad = \quad \sum_{j=1}^{i} \mathbb{I}(|X_j| \le |X_i|)$$

to detect such a location change. Set

$$U_i \quad = \quad R_i^+/(i+1) \quad \text{for } i = 1, 2, \cdots, n.$$

When the process is in control the random variables $U_1, U_2, \cdots, U_n$ are uniformly distributed over $\{1/(1+n), 2/(1+n), \cdots, n/(1+n)\}$ and $U_1 \equiv 1/2, U_2, \cdots, U_n$ are independent according to Parent (1965). Define, for $i = 1, \cdots, n$,

$$V_i^+ \quad = \quad R_i^+/(i+1) - 1/2$$

which has zero mean and variance

$$\begin{aligned}\mathbb{V}\mathrm{ar}_\infty[V_i^+] &= \mathbb{E}_\infty[(R_i^+/(i+1) - 1/2)^2] \\ &= \Big(\sum_{k=1}^{i} k^2\Big)/(i(i+1)^2) - 1/4 \\ &= (i-1)/(12(i+1)).\end{aligned}$$

Then the standardised versions

$$V_i = \sqrt{12(i+1)/(i-1)}V_i^+$$

of V_i^+ are independent with zero expectation and unit variance when the process is in control.

If an increase in dispersion occurs at time $\tau < \infty$, the sequence of random variables $X_1, X_2, \cdots, X_\tau, X_{\tau+1}, \cdots$ are independent, but not necessarily identically distributed. It is shown in Section 4.2.2 that the sequential ranks of the out-of-control random variables $X_{\tau+1}, X_{\tau+2}, \cdots$ tend to be larger than those of the in-control random variables. As a result,

$$\mathbb{E}_\tau[V_i] > 0 \quad \text{for } i > \tau$$

and, therefore, the CUSUM based on V_i should be able to detect increases in dispersion. A similar argument shows that the CUSUM should also be able to detect decreases in dispersion.

Since we assume no specific knowledge of the in-control dispersion, an initial in-control sample of $m < \tau$ observations is required in order to establish a baseline. The upward CUSUM based on unsigned sequential ranks (the USR CUSUM) is defined as

$$D_n = \begin{cases}\max(0, D_{n-1} + V_n - k), & n > m \\ 0, & 0 \le n \le m\end{cases} \tag{4.2}$$

where k is the reference value, a non-negative constant. To complete the design of the USR CUSUM, the values of k and h that will guarantee a prespecified in-control ARL need to be specified. This is done in Section 4.2.

If a change is signalled, we estimate the changepoint as

$$\widehat{\tau} \;=\; \max\{m < n \leq T-1 : D_n = 0\},$$

as before.

4.2 The in-control behaviour of the CUSUM

4.2.1 Determination of control limits

First, we determine control limits h for given reference values k and in-control nominal ARLs. Due to the distribution-free character of the sequential ranks, the control limits can be obtained by Monte Carlo simulation using $U(-1, 1)$ random variables. Table 4.1 shows the estimated control limits for the upward USR CUSUM. We use an initial in-control sample of $m = 20$ observations in the simulations.

Nominal ARL_0	**k**			
	0.125	0.25	0.375	0.5
100	6.013	4.453	3.450	2.720
200	7.950	5.650	4.255	3.319
300	9.159	6.350	4.742	3.678
400	10.057	6.850	5.086	3.935
500	10.818	7.250	5.359	4.130

Table 4.1: Estimated control limits for the upward USR CUSUM.

To show that the estimated control limits are appropriate, we estimate in-control ARL values corresponding to the control limits in Table 4.1. These estimates were obtained from independent Monte Carlo simulations with 100 000 runs in each cell of Table 4.1. The initial sample consists of $m = 20$ observations. The estimated ARL_0 results are shown in Table 4.2. Again, it is worth mentioning that the control limits err on the conservative side in that we prefer an estimated in-control ARL_0 that is slightly larger than the specified nominal values, rather than smaller.

Nominal ARL_0	k			
	0.125	0.25	0.375	0.5
100	101	103	100	102
200	203	205	204	200
300	304	305	302	302
400	400	406	403	402
500	499	502	501	504

Table 4.2: Estimated ARL_0 results for combinations of nominal ARL and reference values for the upward USR CUSUM.

It remains to relate the reference value k to the target change size in dispersion λ. This is done in Section 4.2.2.

4.2.2 Specification of the reference value

This section provides a calculation for the reference value of the USR CUSUM in terms of the target change size λ in the dispersion of the X_i. As in the case of the SSR CUSUM, we can obtain some guidance regarding the appropriate reference value k by considering underlying normal and t_3-distributions.

As in Chapter 3, in order to find an appropriate reference value, we assess the behaviour of $\mathbb{E}_\tau[V_{\tau+j}]/2$, $j \geq 1$ for a given τ. We again consider two bases of calculation:

B.1. Fix $j \geq 1$ and compute $\lim_{\tau\to\infty} \mathbb{E}_\tau[V_{\tau+j}|\lambda]$, i.e. we assume that the process runs in control for a long time and we then assess the expected change in V_i.

B.2. Suppose that τ is positive and fixed, i.e. the process goes out of control at a fixed and positive time, and we then compute $\lim_{j\to\infty} \mathbb{E}_\tau[V_j|\lambda]$.

In assessment B.1. the random variables $X_1, X_2, \cdots, X_\tau$ are i.i.d. with zero median and unit dispersion. The out-of-control random variables $X_{\tau+1}, X_{\tau+2}, \cdots$ have distribution function $G(x) = F(x/\lambda)$ where F denotes the in-control distribution function. The out-of-control density function is assumed to exist and to be continuous and is $g(x) = f(x/\lambda)/\lambda$. It is again convenient to consider sequences of i.i.d.

random variables $Y_1, Y_2, \cdots$ distributed according to F and $W_1, W_2, \cdots$ distributed according to G. Set

$$X_i \;=\; \begin{cases} Y_i & \text{for } 1 \le i \le \tau \\ W_j & \text{for } i = \tau + j, j \ge 1. \end{cases}$$

Denote by R_j^+ the sequential rank of $|X_j|$ among $|X_1|, |X_2|, \cdots, |X_j|$, that is, for $j \ge 1$,

$$R_j^+ \;=\; 1 + \sum_{i=1}^{j-1} \mathbb{I}(|X_i| < |X_j|)$$

and

$$\sum_{i=1}^{0} \mathbb{I}(|X_i| < |X_1|) \;=\; 0.$$

By definition, we have, for a fixed $j \ge 1$,

$$\begin{aligned} R_{\tau+j}^+ &= 1 + \sum_{i=1}^{\tau+j-1} \mathbb{I}(|X_i| < |X_{\tau+j}|) \\ &= 1 + \sum_{i=1}^{\tau} \mathbb{I}(|X_i| < |X_{\tau+j}|) + \sum_{i=\tau+1}^{\tau+j-1} \mathbb{I}(|X_i| < |X_{\tau+j}|) \\ &= 1 + \sum_{i=1}^{\tau} \mathbb{I}(|X_i| < |X_{\tau+j}|) + \sum_{i=1}^{j-1} \mathbb{I}(|X_{\tau+i}| < |X_{\tau+j}|) \\ &= 1 + \sum_{i=1}^{\tau} \mathbb{I}(|Y_i| < |W_j|) + \sum_{i=1}^{j-1} \mathbb{I}(|W_i| < |W_j|) \\ &= 1 + \tau \Big(\sum_{i=1}^{\tau} \mathbb{I}(|Y_i| < |W_j|) \Big) / \tau + \sum_{i=1}^{j-1} \mathbb{I}(|W_i| < |W_j|) \\ &= 1 + \tau \cdot F_\tau^+(|W_j|) + \sum_{i=1}^{j-1} \mathbb{I}(|W_i| < |W_j|) \end{aligned}$$

where $F_\tau^+(\cdot)$ denotes the empirical distribution function of $|Y_1|, |Y_2|, \cdots, |Y_\tau|$. Then

$$\mathbb{E}_\tau[R_{\tau+j}^+] \;=\; \tau\, \mathbb{E}[F_\tau^+(|W_j|)] + (j+1)/2 \tag{4.3}$$

because

$$\mathbb{E}[\mathbb{I}(|W_i| < |W_j|)] \;=\; \mathbb{P}(|W_i| < |W_j|) \;=\; 1/2.$$

Moreover, $F_\tau^+(y)$ converges to $F^+(y)$ almost surely as $\tau \to \infty$ for every y where

$$F^+(y) \;=\; \mathbb{P}(|Y| \le y).$$

Then by using (4.3), we have that

$$\mathbb{E}_\tau[R^+_{\tau+j}/(\tau+j+1)] \;\to\; \mathbb{E}[F^+(|W|)]$$

as $\tau \to \infty$. The last equality does not depend upon j and consequently, for a fixed $j \ge 1$ and by letting $\tau \to \infty$,

$$\begin{aligned}
\mathbb{E}_\tau[R^+_{\tau+j}/(\tau+j+1)] \;&\to\; \mathbb{E}[F^+(|W|)] \\
&=\; \int_{-\infty}^{\infty}(2F(|w|)-1)g(w)dw \\
&=\; 2\int_{-\infty}^{\infty}F(|w|)g(w)dw - 1 \\
&=\; 2\Big(\int_0^{\infty}F(w)g(w)dw + \int_{-\infty}^{0}F(-w)g(w)dw\Big) - 1 \\
&=\; 4\int_0^{\infty}F(w)g(w)dw - 1 \\
&=\; 4\int_0^{\infty}F(\lambda y)f(y)dy - 1
\end{aligned}$$

upon setting in the next to last equality $w = \lambda y$ based on the construction $W_j = \lambda Y_j$ for $\lambda > 1$ and $j > \tau$. Therefore,

$$\mathbb{E}_\tau[V_j|\lambda] \;\to\; \sqrt{12}\Big(4\int_0^{\infty}F(\lambda y)f(y)dy - 3/2\Big) \tag{4.4}$$

as $\tau \to \infty$, yielding the reference value

$$k \;:=\; k(F,\lambda) \;=\; \sqrt{3}\Big(4\int_0^{\infty}F(\lambda y)f(y)dy - 3/2\Big). \tag{4.5}$$

Table 4.3 gives the value of k for the normal and the t_3-distribution for typical choices of λ.

Distribution	λ		
	1.25	1.375	1.50
Normal	0.12	0.17	0.22
t_3	0.10	0.15	0.19

Table 4.3: Reference values from (4.5).

We now consider basis B.2. From (4.3) we have for a fixed $\tau > 0$,

$$\mathbb{E}_\tau[R^+_{\tau+j}/(\tau + j + 1) - 1/2] \quad \to \quad 0$$

as $j \to \infty$. This result implies that after the onset of a change the upward USR CUSUM will eventually return to what seems to be an in-control state. This is to be expected, because if the USR CUSUM continues to run for a substantially long time after the onset of a change, the impact of the change will become increasingly minuscule. The out-of-control distribution in fact overrides the in-control distribution and the USR CUSUM takes on the new dispersion as the desired level.

4.3 Relative efficiency

To evaluate the relative efficiency of the USR CUSUM with respect to the standard normal CUSUM, it is necessary that both CUSUMs be comparable when the process $X_1, X_2, \cdots$ is in control. This means that both CUSUMs must have the same in-control ARLs and use the appropriate reference values. For a target out-of-control dispersion λ, the appropriate reference value for the normal CUSUM is ζ – see (2.6). In view of our findings in Section 4.2.2, we will use for the USR CUSUM the reference values given in the row corresponding to the normal distribution in Table 4.3. In light of (4.4), we define the relative efficiency of the USR CUSUM with respect to the standard normal CUSUM when the out-of-control variance is β^2 for finite and positive values of τ by

$$e \quad := \quad e(\beta) \quad = \quad \mathbb{E}_\tau[T_N|\beta]/\mathbb{E}_\tau[T_S|\beta] \quad \text{for } \beta > 1 \tag{4.6}$$

where T_N and T_S denote respectively the run lengths of the normal and USR CUSUMs. In other words, the relative efficiency is calculated under the assumption

that the process runs in control for a long time and then we assess the change (see assessment B.1., Section 4.2.2).

Given λ, the target out-of-control dispersion scale factor, and an in-control ARL, the numerator in (4.6) can be found using the existing software (anyarl.exe) from Hawkins and Olwell (1998) for any choice of the reference value ζ – see (2.6) – and the out-of-control dispersion β. For the USR CUSUM, the appropriate control limits were again found by Monte Carlo simulation and are given in Table 4.4. The denominator in (4.6) was also found by Monte Carlo simulation for the corresponding control limits and reference values (100 000 runs in each instance). The values of the numerator and the denominator of the ratio e in (4.6) are shown in Table 4.5 and the values of e itself are shown in Table 4.6.

Nominal ARL_0	**k**		
	0.12	0.17	0.22
100	6.08	5.45	4.83
500	10.96	9.53	8.11

Table 4.4: Control limits for the upward USR CUSUM corresponding to the appropriate reference values $0.12, 0.17$ and 0.22.

		Normal CUSUM			**USR CUSUM**		
$\boldsymbol{\beta}$	**Nominal ARL_0**	**k**			**k**		
		1.24	1.35	1.46	0.12	0.17	0.22
1.25	**100**	21	21	17	55	58	58
	500	41	42	43	316	384	381
1.375	**100**	14	14	14	40	43	44
	500	24	24	24	243	296	298
1.50	**100**	10	10	10	32	32	33
	500	17	16	16	179	220	229

Table 4.5: Out-of-control ARLs obtained from the standard normal and USR CUSUM for data from a $N(0, \beta^2)$ distribution.

β	Nominal $\mathbf{ARL_0}$	λ		
		1.25	1.375	1.50
1.25	**100**	0.38	0.36	0.29
	500	0.13	0.11	0.11
1.375	**100**	0.35	0.33	0.32
	500	0.10	0.08	0.08
1.50	**100**	0.31	0.31	0.30
	500	0.09	0.07	0.07

Table 4.6: Relative efficiency of the USR CUSUM with respect to the standard normal CUSUM.

All of the relative efficiencies in Table 4.6 are substantially smaller than 1, indicating that the standard normal variance CUSUM signals an upward change much earlier than the USR CUSUM. However, the standard normal CUSUM assumes a known underlying variance whereas the USR CUSUM does not. Thus the comparison in Table 4.6 is again not entirely fair. As in Section 3.3, in order to attain comparability with the USR CUSUM we implement a version of the self-starting normal CUSUM for a variance (Hawkins and Olwell, 1998, Section 7.3) which assumes that the in-control mean is zero but does not require knowledge of the in-control variance. We will again use the Monte Carlo method to gauge the efficacy of the USR CUSUM when the underlying distribution is normal. We consider four designs:

1. target scale factor is $\lambda = 1.5$, in-control ARL is 100,
2. target scale factor is $\lambda = 2.0$, in-control ARL is 100,
3. target scale factor is $\lambda = 1.5$, in-control ARL is 500,
4. target scale factor is $\lambda = 2.0$, in-control ARL is 500.

For normal variance CUSUMs, which presume a known variance, the reference values from the formula

$$\log(\lambda^2)/(1 - 1/\lambda^2)$$

are 1.46 for designs 1 and 3 and 1.85 for designs 2 and 4 – see (Hawkins and Olwell, 1998, Section 6.2.3) – and the control limits guaranteeing the required in-control ARLs are 7.007, 5.722, 12.169 and 9.743, respectively. For a given in-control ARL, this CUSUM uses the same reference values and control limits as the normal

variance CUSUM. However, the summand V_n in (4.2) is now

$$V_n \;=\; \Gamma^{-1}(F_{n-1}(W_n))$$

where, $W_n = X_n^2/s_{n-1}^2$ for $n \geq 2$ with s_{n-1}^2 given in (3.7), where F_{n-1} is the cdf of Fisher's F-distribution with 1 and $n-1$ degrees of freedom and where Γ is the cdf of a chi-square distribution with 1 degree of freedom.

For the sequential rank CUSUMs we use in the designs 1 and 3 the reference value $k = 0.2$ and we use $k = 0.36$ (from Table 4.3) in designs 2 and 4. The corresponding control limits are 5.08, 3.57, 8.68 and 5.59. In the simulations we estimate the ARLs for each of the four designs using $m = \tau = 10$ and $m = \tau = 20$ initial in-control observations. The CUSUMs were set to zero at $m = \tau$, i.e. we set $D_1 = \cdots = D_m = 0$.

Each of the "USR" and "normal" ARL estimates in Table 4.7 comes from 10 000 Monte Carlo trials. The columns with headings *eff* (efficiency) give the ratios of the normal ARL to USR ARL.

	m = 10			**m = 20**		
Design	**normal**	**USR**	***eff***	**normal**	**USR**	***eff***
1	40	44	0.91	21	27	0.78
2	16	24	0.67	7	12	0.58
3	297	361	0.82	161	216	0.75
4	177	223	0.79	45	74	0.61

Table 4.7: Out-of-control ARLs of the upward normal self-starting and USR CUSUMs.

Notice that the relative efficiencies (*eff*) in Table 4.7 are much higher than those in Table 4.6, which reflects the substantial effect that absence of knowledge regarding the in-control variance has on the performance of the normal variance CUSUM.

4.4 Concluding remarks

4.4.1 Two-sided CUSUMs

Two-sided CUSUMs can be constructed in the same manner as indicated in Section 3.4.1.

4.4.2 Asymmetric distributions

Whether the i.i.d. random variables $X_1, X_2, \cdots$ are skew or not, the sequential ranks of $|X_1|, |X_2|, \cdots$ are independent and follow a discrete uniform distribution. Thus, the USR CUSUM will be valid under the in-control situation regardless of whether the underlying distribution is symmetric or not. Table 4.8 shows that the estimated ARL_0 values are a good approximation to the nominal ARL_0 values, confirming that the USR CUSUM is also valid for this moderately skew in-control distribution.

Nominal ARL_0	**k**			
	0.125	0.250	0.375	0.500
100	101	101	101	98
200	204	200	200	200
300	303	304	303	307
400	401	398	404	398
500	508	500	497	503

Table 4.8: Estimated ARL_0 results for combinations of nominal ARL and reference value when applying the USR CUSUM to a Gumbel distribution.

The matter at hand is whether there is a reduction in the ARL when the process goes out of control. Consequently, we will evaluate the estimated ARL results of the USR CUSUM when the underlying distribution is skew with a non-unit dispersion $\lambda = 1.5$. The estimates were obtained in a Monte Carlo simulation with 100 000 runs in each cell. The results are shown in Table 4.9. We took $m = 20$ as the initial in-control sample to allow for the sequential ranks to reach a state of statistical control under the in-control distribution, a Gumbel with zero mean and unit dispersion.

We then generate data from a Gumbel distribution with dispersion $\lambda = 1.5$ after the changepoint $m = \tau = 20$. Clearly, the USR CUSUM results in a substantially reduced ARL when a change in dispersion occurred, which is desirable. The out-of-control ARL values in Table 4.9 agree very well with those of the USR CUSUM applied to data from a normal distribution – see Table 4.5.

$\boldsymbol{\lambda}$	**Nominal $\mathbf{ARL_0}$**	**k**			
		0.125	0.250	0.375	0.500
1.5	**100**	32	34	34	35
	500	178	204	282	258

Table 4.9: Estimated out-of-control ARLs of the USR CUSUM for data from a skew (Gumbel) distribution with dispersion $\lambda = 1.5$.

Chapter 5

Applications and data analysis

This chapter implements the SSR and USR CUSUM procedures in Chapters 3 and 4 to two engineering applications and one financial application. The applications include

1. the ash content of coal,
2. the calorific value (CV) of coal, and
3. the Dow Jones financial index.

5.1 Ash content of coal

Here we analyse a data set consisting of independent pairs of observations (V_i, W_i), $i \geq 1$ on the ash content of coal from two nominally identical laboratories. The measurements V_i and W_i, for $i \geq 1$, are made on a sample of coal split between the laboratories employing the same methodologies. Suppose the true value of the ash content of coal sample i is Q_i. Then the measurements are, respectively,

$$V_i = Q_i + \varepsilon_{V_i} \tag{5.1}$$

and

$$W_i = Q_i + \varepsilon_{W_i} \tag{5.2}$$

where ε_{V_i} and ε_{W_i} denote the measurement errors in each of the two laboratories. If the laboratories are indeed nominally identical, then the pairwise differences

$$Z_i = V_i - W_i$$

should be symmetrically distributed around zero. A non-zero median indicates the presence of bias between the laboratories, while asymmetry in the distribution of Z_i indicates that V_i and W_i are not identically distributed. The monitoring process should therefore be designed to detect a change in the median of Z_i away from zero, but also asymmetry in the distribution of Z_i. It is also necessary that the dispersion of the Z_i be monitored. If the dispersion should increase from its current value, it could possibly indicate procedural discrepancies between the laboratories.

Next, we discuss the application of the SSR and USR CUSUMs. First, consider detecting a change in the median of the measurements Z_i away from zero using the two-sided CUSUM discussed in Section 3.4.1. There is no expectation, a priori, that the tails of the distribution of the Z_i will be excessively heavy. We therefore opt for the reference values applicable to the normal distribution. We apply the SSR CUSUM with nominal ARL 250 and target change size $|\delta| = 0.5$ standard deviations. From Table 3.4 the appropriate reference value is 0.24 which is close to $\delta/2$. We therefore take the reference value $k = 0.25$ as acceptable. The corresponding control limits are ±7.267 from Table 3.1.

Simultaneously, we run the USR CUSUM to detect a change in the dispersion of the Z_i. For this we assume that the first 20 observations are in control. We apply the two-sided USR CUSUM with nominal ARL 250 and target change by a factor $\lambda = 1.5$. From Table 4.3 the appropriate reference value is $k = 0.22$. Application of linear interpolation to the control limits in Table 4.1 gives ±8.11 as the appropriate control limits.

The monitoring process stops once either of the SSR or USR CUSUMs signals a change. The two-sided USR CUSUM signals an upward change at time $i = 109$. Because of the time delay between laboratory observations and the CUSUM analysis, we actually have observations up to $i = 130$. From Figure 5.3 we see that the CUSUM has quickly returned to zero after the signal at time $i = 109$. This suggests strongly that this signal is a false alarm. Having decided, for the moment, that the signal at $i = 109$ is a false alarm, we continue the CUSUM plots. The two-sided SSR CUSUM

signals a downward change at time $i = 163$ – see Figure 5.2 – while the USR CUSUM has not again signalled up to that point. We again have observations up to $i = 180$ due to time delays. From Figure 5.2 we see that the CUSUM has quickly returned to zero after crossing the control limit at $i = 163$, which suggests strongly that the signal at $i = 163$ is a false alarm. This is further confirmed by the fact that the two-sample Wilcoxon rank sum statistic for the samples $Z_1, \cdots, Z_{163}$ and $Z_{164}, \cdots, Z_{180}$ has a p-value of 0.55.

Figure 5.1 shows a symmetry plot of observations $Z_1, \cdots, Z_{163}$. Except for the one "outlier" there is little reason to suspect asymmetry. Further confirmation comes from the Wilcoxon signed rank statistic which has a p-value of 0.45. The outlier corresponds to the observation at $i = 103$ which can be clearly seen in Figure 5.4 (this figure shows a time series plot of the set of Z_i up to $i = 250$). This outlier being a positive Z-value cannot be the cause of the signal at $i = 163$ which indicates a decrease in the median, however it may have possibly caused the USR CUSUM to signal at observation $i = 109$. This outlier could be due to a special cause (see Section 1.1) which should be investigated after the fact.

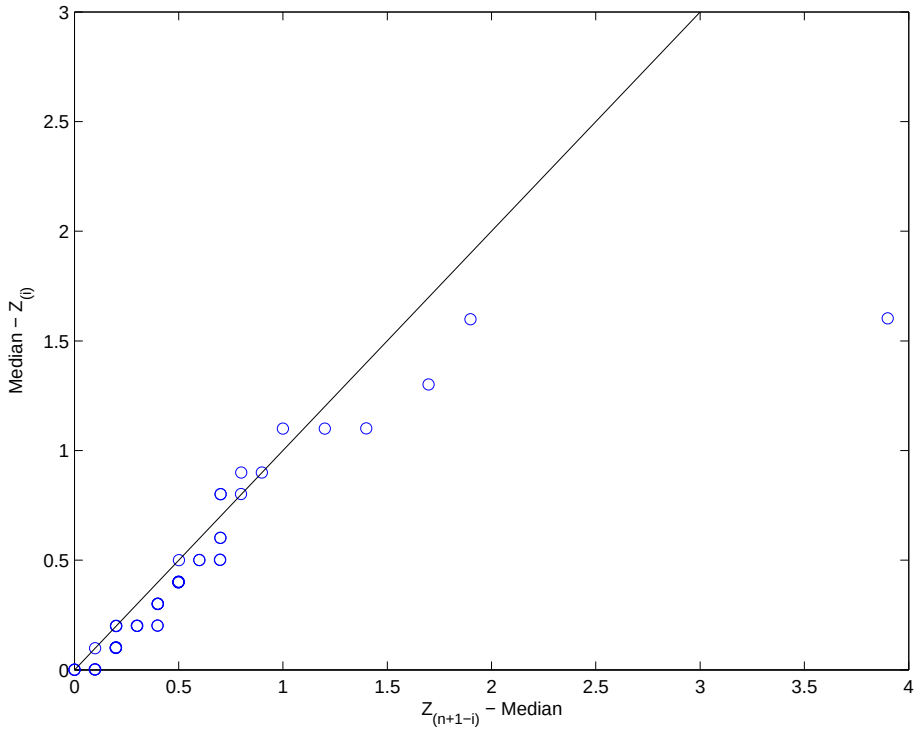

Figure 5.1: Symmetry plot of $Z_1, \cdots, Z_{163}$.

Having decided that the signal at $i = 163$ is a false alarm, we continue the CUSUM plot and see that there is another similar signal at time $i = 214$. After time $i = 214$, however, the CUSUM increases rapidly crossing the upper control limit at $i = 225$ and remains above it. The estimated changepoint for the increase in the median is $\widehat{\tau} = 214$. The USR CUSUM exhibits the same behaviour. The increase indicated by the USR CUSUM is due to the change in the median of the Z_i.

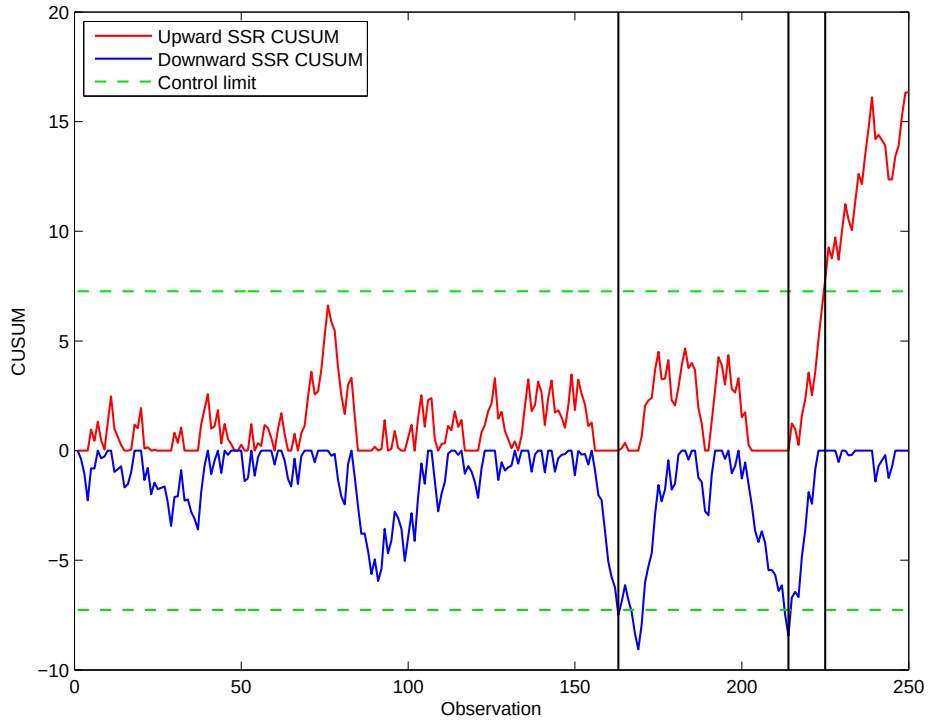

Figure 5.2: Two-sided SSR CUSUM for a change in the median of Z_i.

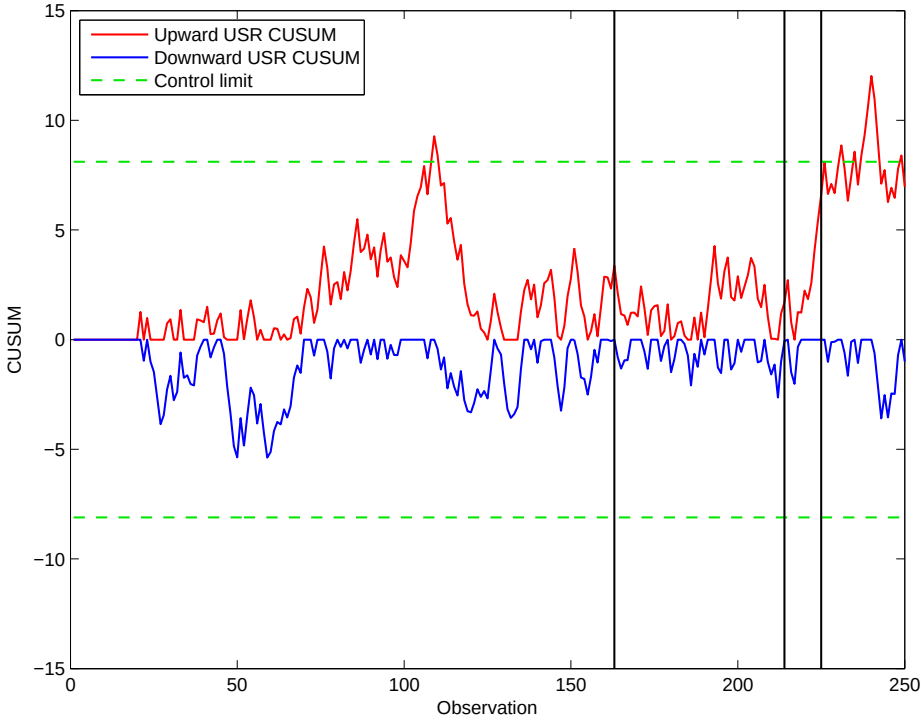

Figure 5.3: Two-sided USR CUSUM for a change in the dispersion of Z_i.

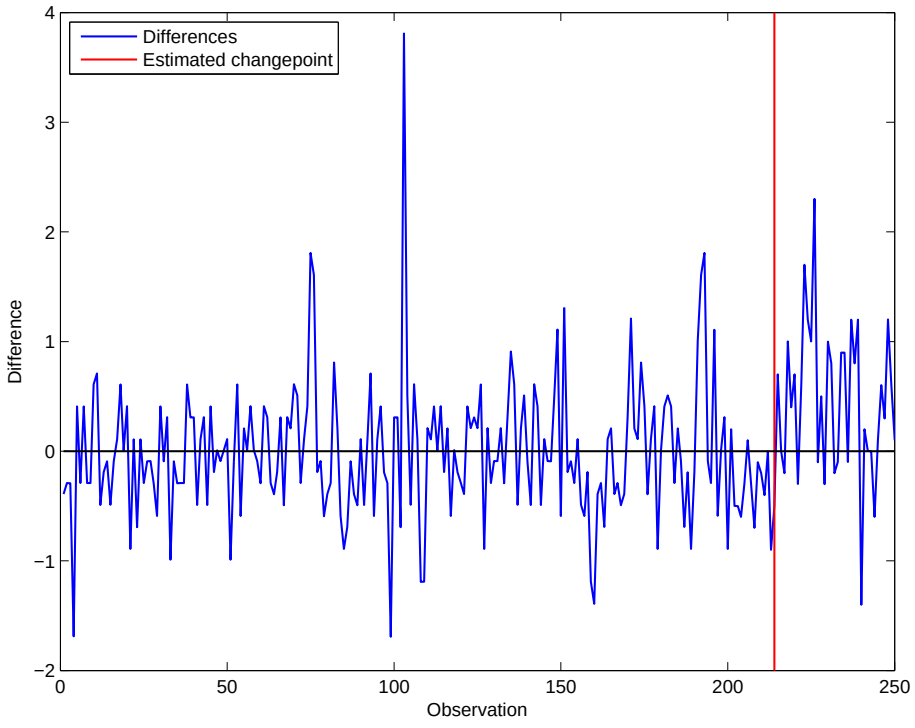

Figure 5.4: Time series plot of Z_i.

At this point it is again necessary to access the symmetry of the $Z_{164}, \cdots, Z_{214}$. The Wilcoxon signed rank statistic gives a p-value of 0.52 which suggests that there is little reason to suspect asymmetry. The mean and standard deviation of the Z_i for the segments $i = 1, \cdots, 214$ and $i = 215, \cdots, 250$ are shown in Table 5.1.

Segments	**164:214**	**215:250**
Mean	-0.01	0.46
Standard deviation	0.61	0.70

Table 5.1: Descriptive statistics for the segments $Z_{164}, \cdots, Z_{214}$ and $Z_{215}, \cdots, Z_{250}$.

From this table we see that there was an increase in the mean after time 214 which was signalled by the SSR CUSUM. The two-sample Wilcoxon rank sum statistic gives a p-value of 0.001 and there is enough evidence to suggest that the mean of Z_i after the estimated changepoint $\widehat{\tau} = 214$ is substantially different from zero. The standard deviation remains largely unchanged between the segments. Our overall conclusions are that the median remained at 0 up to observation $i = 214$ after which there was a sudden increase. There is little to no evidence suggesting a substantial deviation from the assumption of symmetry.

5.2 Calorific value of coal

The calorific value (CV) of coal is an indication of the quantity of heat produced by the combustion of coal at a constant pressure of 1013 mbar under normal conditions of 0°C. The calorific value is measured in kilojoules per kilogram (KJ/kg) (Mason and Gandhi, 1983). As in the case of the ash data set, the present data set consists of independent pairs of observations (V_i, W_i), $i \geq 1$ on the CV of coal from two nominally identical laboratories. The conceptual model in (5.1) and (5.2) again holds.

Figure 5.5 shows the two-sided SSR CUSUM (in-control ARL 250, reference value $k = 0.25$ and control limits ± 7.267), and signals an increase at observation $i = 57$. This increase seems sustained through observation $i = 80$.

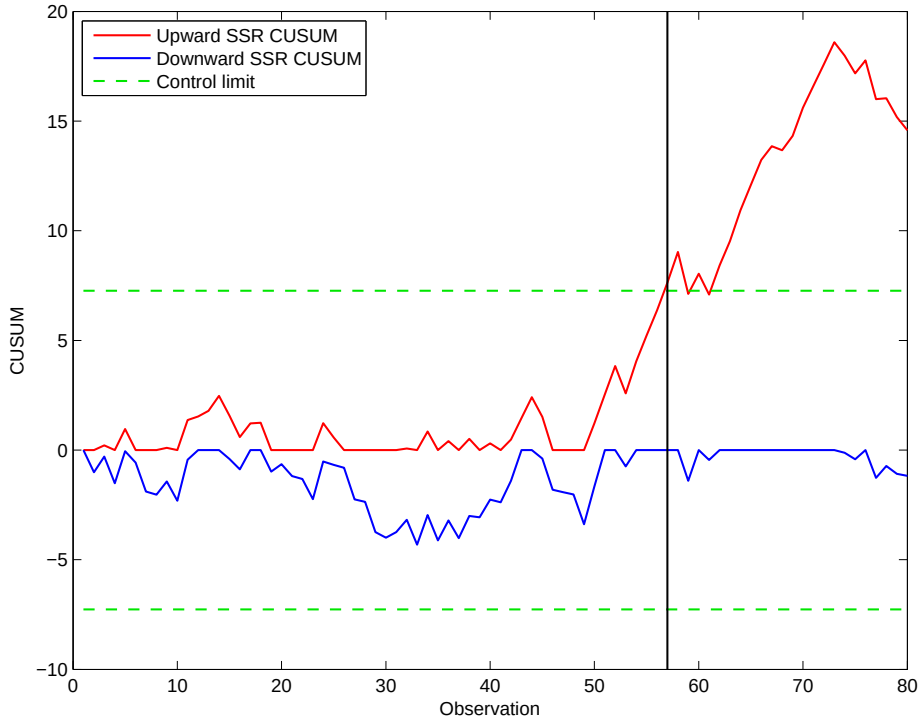

Figure 5.5: Two-sided SSR CUSUM for a change in the median of Z_i.

In Table 5.2 below a substantial increase in the mean appears to occur from the sample $Z_1, \cdots, Z_{49}$ to $Z_{50}, \cdots, Z_{80}$. The two-sample Wilcoxon rank sum statistic gives a p-value of 0 to four decimals confirming that there is enough evidence suggesting a substantial increase in the mean after the estimated changepoint $\widehat{\tau} = 49$.

Segments	**1:49**	**50:80**
Mean	-0.11	0.36
Standard deviation	0.38	0.54

Table 5.2: Descriptive statistics for the segments $Z_1, \cdots, Z_{49}$ and $Z_{50}, \cdots, Z_{80}$.

A question that remains is whether this putative change in the median is perhaps not due to asymmetry in the distribution of the Z_i. Figure 5.6 shows the symmetry plot of $Z_1, \cdots, Z_{49}$. The Wilcoxon signed rank statistic gives a p-value of 0.10 suggesting that there is little evidence that the distribution is asymmetric. We subtract the median 0.46 of Z_i for $i = 50, \cdots, 80$ from $Z_{50}, \cdots, Z_{80}$ to obtain $ZZ_{50}, \cdots, ZZ_{80}$ and show the symmetry plot in Figure 5.7. The Wilcoxon signed rank statistic gives a p-value of 0.49, again suggesting that there is little evidence that the distribution is asymmetric. However, the points on both symmetry plots seem to diverge away from the 45° symmetry-line indicating the presence of asymmetry.

Since the sample sizes are relatively small, it is not entirely surprising that the symmetry plots and the Wilcoxon signed rank statistic give contradictory results. We therefore conclude that the SSR CUSUM may have signalled either due to an increase in the median or due to the presence of asymmetry in the distribution of the Z_i. Inspection of Figure 5.8 reveals an increase in the median after the estimated changepoint $\widehat{\tau} = 49$.

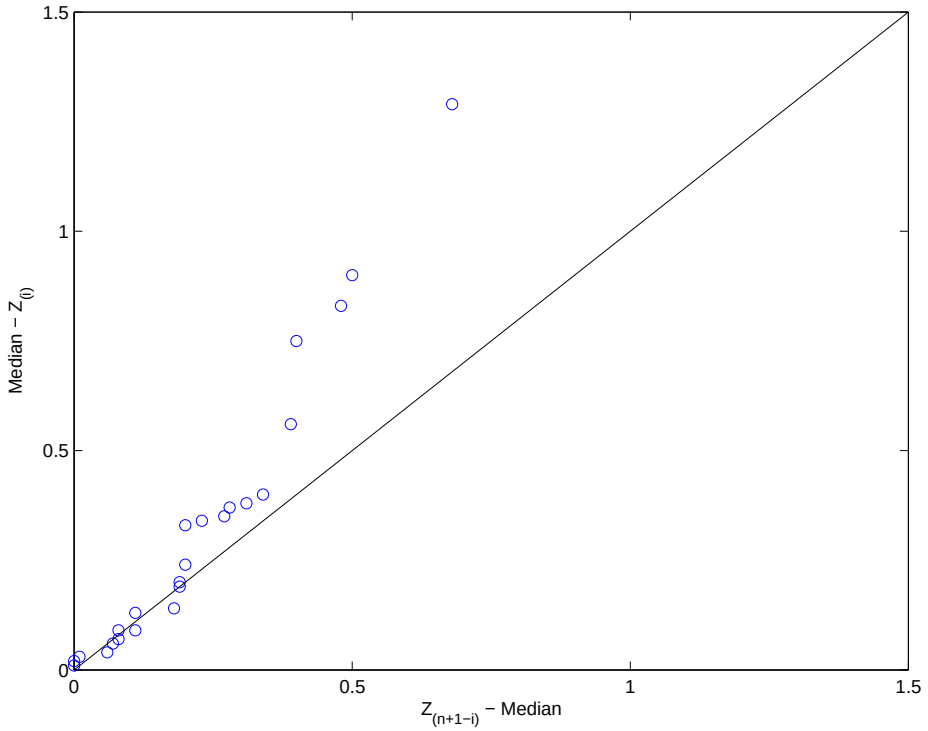

Figure 5.6: Symmetry plot of $Z_1, \cdots, Z_{49}$.

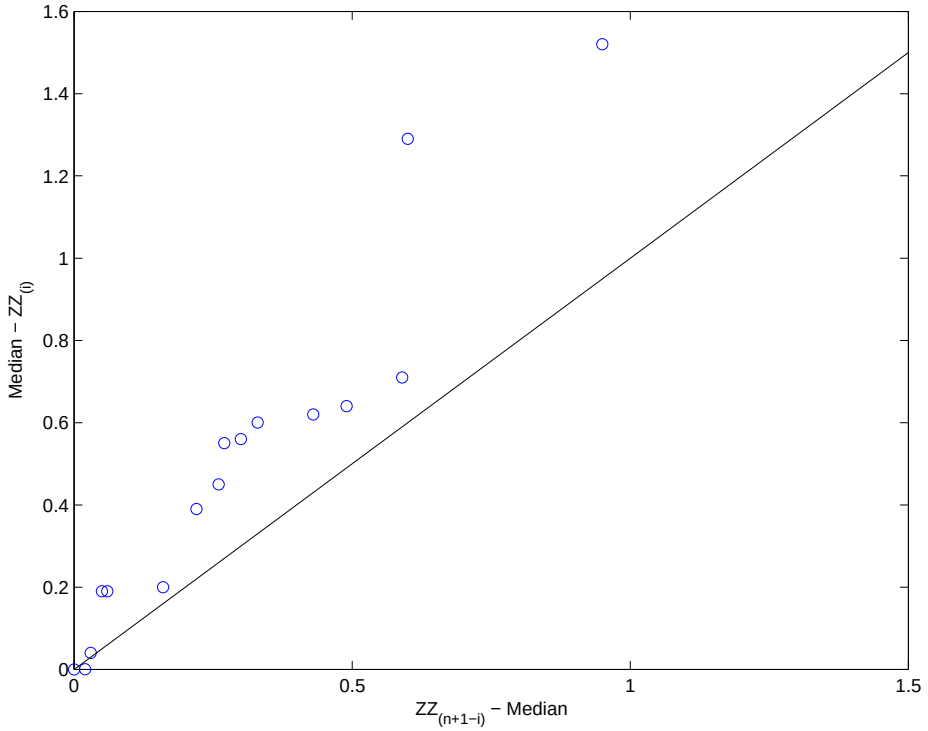

Figure 5.7: Symmetry plot of $ZZ_{50}, \cdots, ZZ_{80}$.

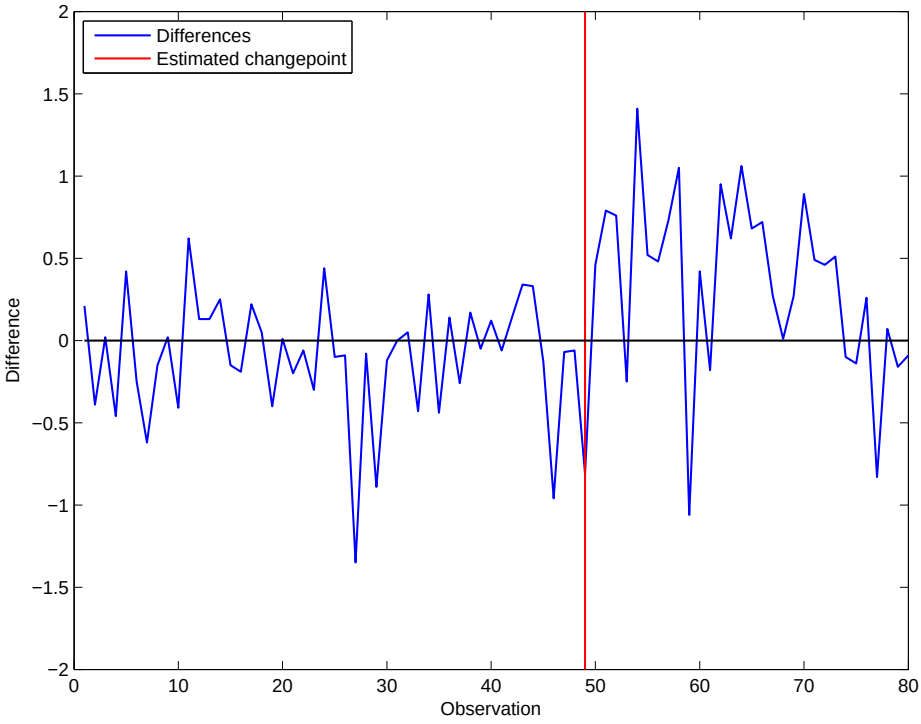

Figure 5.8: Time series plot of Z_i.

5.3 Dow Jones Index

Finally, we analyse the Dow Jones monthly index series, Y_i, $i \geq 1$. Investors are usually concerned with both the level and volatility of stock indices. The mean and the variance of the log returns

$$Z_i \;=\; \log(Y_{i+1}/Y_i)$$

are to be monitored. An increase in the mean of the log returns indicates that an investor might consider buying into the index, while a decrease suggests the opposite. The standard deviation of the Z_i is generally accepted as a measure of volatility, which represents risk in this context.

We apply the SSR CUSUM (in-control ARL 250, reference value $k = 0.25$ and control limits ± 7.267) and USR CUSUM (in-control ARL 250, reference value $k = 0.22$ and control limits ± 8.11) in a manner similar to the approach described in Sections 5.1 and 5.2. The USR CUSUM signals at observation $i = 57$ that a decrease in variance has occurred. The estimate of the changepoint from Figure 5.10 is $\widehat{\tau} = 30$. Up to this time the SSR CUSUM has not signalled a change, although the CUSUM has come close to both the lower and upper control limit – see Figure 5.9. After the signal we observe that the USR CUSUM sequence continues to decline below the lower control limit indicating that the decrease in the variance is persistent.

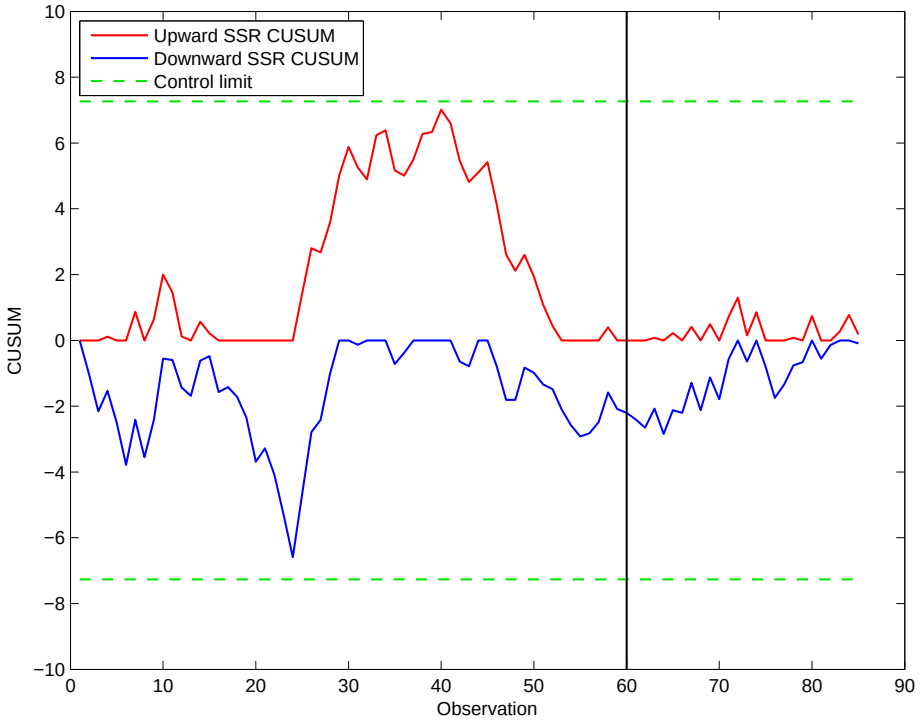

Figure 5.9: Two-sided SSR CUSUM for a change in the mean of Z_i.

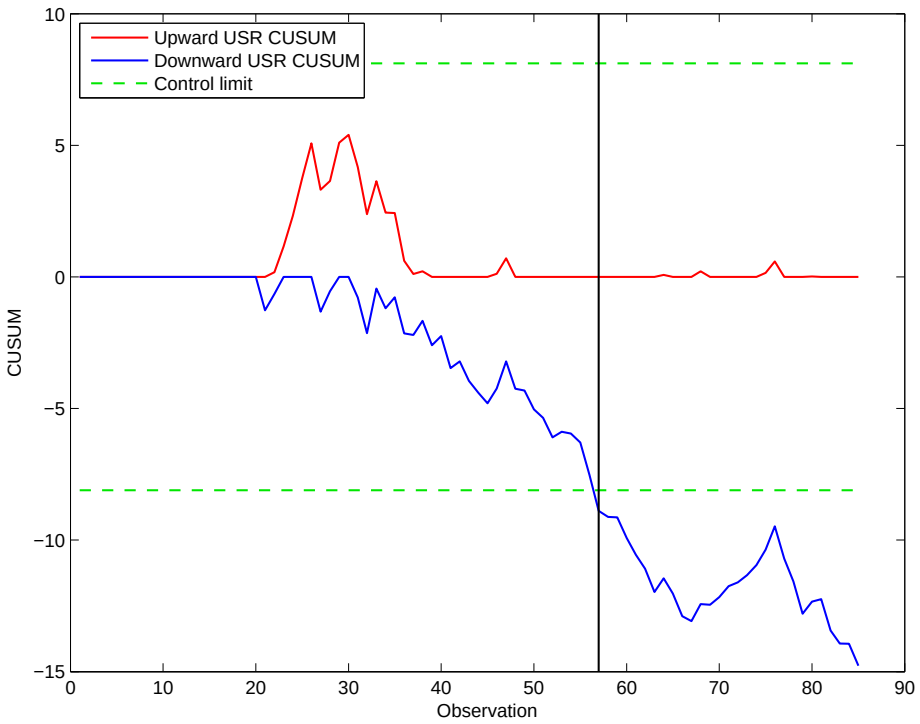

Figure 5.10: Two-sided USR CUSUM for a change in the variance of Z_i.

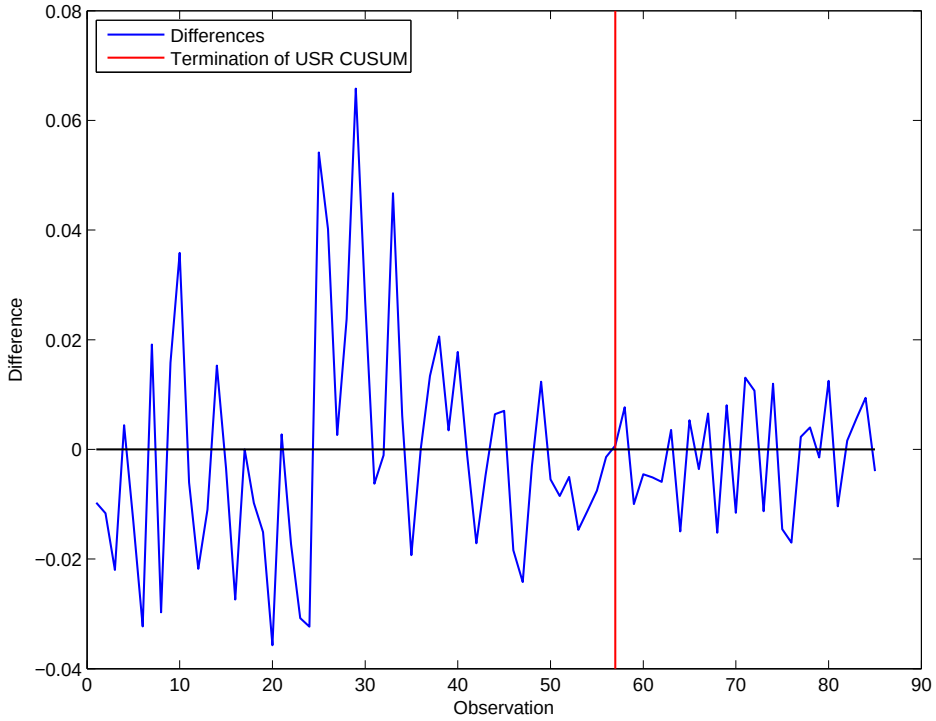

Figure 5.11: Time series plot of Z_i.

To better understand the behaviour of the two CUSUMs it is useful to look at the time series plot of the Z_i in Figure 5.11. The SSR CUSUMs seem to indicate a decrease followed by an increase in the median of the Z_i, however neither CUSUM exceeds the control limit. In the time series plot we see that initially the median of the Z_i is negative and then becomes positive for a relatively short time before settling down around zero. This explains to some extent the behaviour of the SSR CUSUM in Figure 5.9. However, what is clear from the time series plot in Figure 5.11 is that the overall variability has decreased and this is clearly signalled by the USR CUSUM.

An important assumption that we have made throughout is that the observations are independent. Figure 5.12 shows a correlogram of the first autocorrelations of the series $Z_1, \cdots, Z_{30}$ together with the conventional confidence limits $\pm 2/\sqrt{n}$. While this does not establish independence (unless the Z_i are normally distributed), it does indicate no serial correlation in the series. This apparent lack of correlation may at first sight be against the expectation from the literature (Campbell et al., 1996, p. 68) that financial returns generally exhibit autocorrelation. However, it is also known that the correlation decreases as the measurement interval increases. The one-month interval which we have used appears to be sufficient to preclude the presence of autocorrelation.

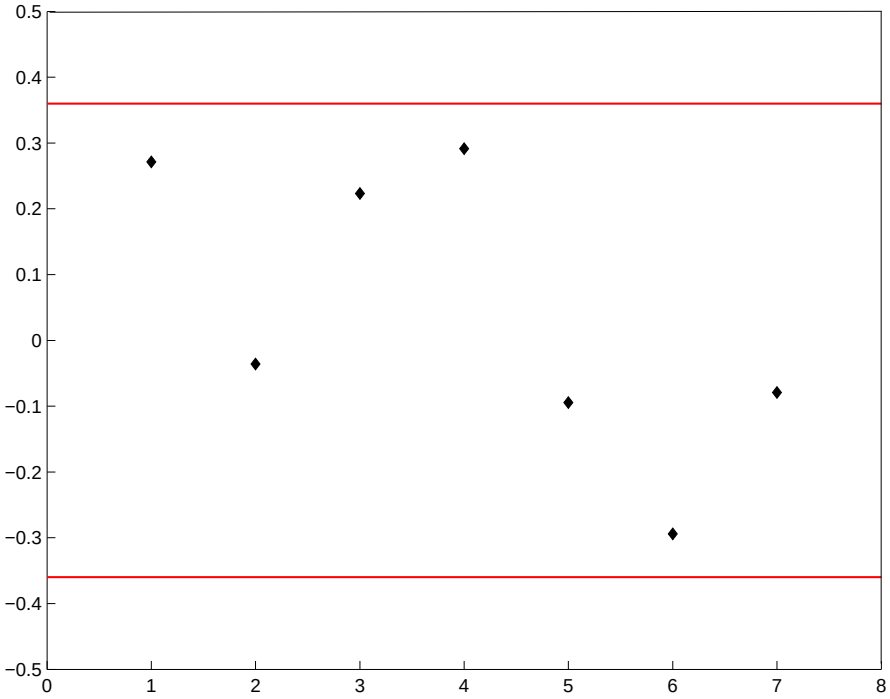

Figure 5.12: Correlogram of the first seven autocorrelations of the series $Z_1, \cdots, Z_{30}$.

Suggestions for further research

This chapter concludes the dissertation with a few remarks concerning future research in the field of sequential rank CUSUM procedures.

The first remark that comes to mind concerns the assumption of a known in-control median. In the analysis of the Dow Jones data, Section 5.3, the initial median is clearly not zero. Thus one would want a sequential rank procedure which does not assume a known initial median.

Following on this, the assumption of symmetry in the SSR CUSUM can also be problematic. Therefore, research should also be conducted into relaxing the symmetry assumption. Some work in this direction is due to McDonald (1990).

In Chapter 3, Section 3.1, we mentioned a sequential rank analogue of the Pignatiello and Samuel (2001) changepoint estimator. The properties of this estimator remain to be investigated.

Finally, one could consider the use of signed and unsigned sequential ranks in monitoring procedures other than the CUSUM, such as the Shiryaev-Roberts procedures (see Shiryaev (1963) and Roberts (1966)).

Bibliography

Bakir, S. T. (2006). Distribution-free quality control charts based on signed-rank-like statistics, *Communication in Statistics - Theory and Methods* **35**(4): 743–757.

Bakir, S. T. and Reynolds, M. R. (1979). A nonparametric procedure for process control based on within-group ranking, *Technometrics* **21**(2): 175–183.

Bandyopadhyay, U. and Mukherjee, A. (2007). Nonparametric partial sequential test for location shift at an unknown time point, *Sequential Analysis: Design Methods and Applications* **26**(1): 99–113.

Barndorff-Nielsen, O. (1963). On the limit behaviour of extreme order statistics, *The Annals of Mathematical Statistics* **34**(3): 992–1002.

Bhattacharya, P. K. and Frierson, D. J. (1981). A nonparametric control chart for detecting small disorders, *The Annals of Statistics* **9**(3): 544–554.

Brook, D. and Evans, D. A. (1972). An approach to the probability distribution of cusum run length, *Biometrika* **59**(3): 539–549.

Campbell, J., Lo, Y. and MacKinlay, A. (1996). *The Econometrics of Financial Markets*, Princeton: Princeton University Press.

Coleman, S. Y., Arunakumar, G., Foldvary, F. and Feltham, R. (2001). SPC as a tool for creating a successful business measurement framework, *Journal of Applied Statistics* **28**(3-4): 325–334.

Dudding, B. P. and Jennett, W. J. (1942). *Quality Control Charts: BS600R*, London: British Standards Institution.

Duncan, A. J. (1959). *Quality Control and Industrial Statistics*, Chicago: Irwin.

Ghosh, B. K. and Sen, P. K. (1991). *Handbook of Sequential Analysis*, New York: Marcel Dekker Inc.

Golosnoy, V. and Schmid, W. (2007). EWMA control charts for monitoring optimal portfolio weights, *Sequential Analysis: Design Methods and Applications* **26**(2): 195–224.

Hájek, J., Sidák, Z. and Sen, P. K. (1999). *Theory of Rank Tests*, Second edn, San Diego: Academic Press.

Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*, New York: Springer.

Hawkins, D. M., Qiu, P. H. and Kang, C. W. (2003). The changepoint model for statistical process control, *Journal of Quality Technology* **35**(4): 355–366.

Kahya, E. and Theodossiou, P. (1999). Predicting corporate financial distress: A time-series cusum methodology, *Review of Quantitative Finance and Accounting* **13**(4): 323–345.

Lam, K. and Yam, H. C. (1997). Cusum techniques for technical trading in financial markets, *Financial Engineering and the Japanese Markets* **4**(3): 257–274.

Lombard, F. (1983). Asymptotic distribution of rank statistics in the change-point problem, *South African Statistical Journal* **17**(1): 83–105.

Mason, D. M. (1981). On the use of a statistic based on sequential ranks to prove limit theorems for simple linear rank statistics, *The Annals of Statistics* **9**(2): 424–436.

Mason, D. M. and Gandhi, K. N. (1983). Formulas for calculating the calorific value of coal and coal chars: Development, tests and uses, *Fuel Processing Technology* **7**(1): 11–22.

McDonald, D. (1990). A cusum procedure based on sequential ranks, *Naval Research Logistics* **37**(5): 627–646.

Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*, Third edn, New York: John Wiley & Sons.

Mukherjee, A. (2009). Some rank-based two-phase procedures in sequential monitoring of exchange rate, *Sequential Analysis* **28**(2): 137–162.

Page, E. S. (1954). Continuous inspection schemes, *Biometrika* **41**(1): 100–115.

Parent, E. A. (1965). Sequential ranking procedures, *Technical Report 80*, Stanford University, Department of Statistics.

Pignatiello, J. J. and Samuel, T. R. (2001). Estimation of the change point of a normal process mean in SPC applications, *Journal of Quality Technology* **33**(1): 82–95.

Rao, C. (2002). *Linear Statistical Inference and Its Applications*, New York: Wiley.

Reynolds, M. R. (1975). A sequential signed-rank test for symmetry, *The Annals of Mathematical Statistics* **3**(2): 382–400.

Roberts, S. W. (1966). A comparison of some control chart procedures, *Technometrics* **8**(3): 411–430.

Ross, G. J. and Adams, N. M. (2012). Two nonparametric control charts for detecting arbitrary distribution changes, *Journal of Quality Technology* **44**(2): 102–116.

Samuel, T. R., Pignatiello, J. J. and Calvin, J. A. (1998). Identifying the time of a step change with $\overline{X}$ control charts, *Quality Engineering* **10**(3): 521–527.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*, New York: Macmillan.

Shiryaev, A. N. (1963). On optimum method in quickest detection problems, *Theory of Probability and its Applications* **8**(1): 22–46.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer.

Timmer, D. H., Pignatiello, J. J. and Longnecker, M. T. (2001). Applying an AR(1) cusum control chart to data from a chemical process, *Quality Engineering* **13**(1): 107–114.

Wald, A. (1947). *Sequential Analysis*, New York: Wiley.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance, *Journal of Quality Technology* **38**(2): 98–104.

Yi, G., Coleman, S. and Ren, Q. (2006). CUSUM method in predicting regime shifts and its performance in different stock markets allowing for transaction fees, *Journal of Applied Statistics* **33**(7): 647–664.