Empirical Self-Starting CUSUM Charts for Location and Scale

Abstract

From a parametric stand point, during the estimation phase of statistical process control, the underlying distribution of a process and its corresponding parameters are assessed. However, if no a priori information of the process is available, and samples are limited, performance during the monitoring phase of the process is affected. To address this problem, self-starting control charts were created. Nevertheless, most of these approaches assume an underlying distribution. This paper proposes a distribution free method based on a running empirical distribution function to create self-starting CUSUM charts for location and scale. Performance is compared with the corresponding parametric approach that assumes normality. Asymptotic properties of the proposed statistic are studied, and performance is evaluated using extensive Monte Carlo simulations over Gamma, t-student and Normal observations. First, it is shown that asymptotic performance is the same as the parametric counterpart under the in-control hypothesis. Then, when dealing with limited sample size, the proposed method performs better, in terms of ARL, when the underlying distribution is not Normal, and relatively similar when it actually is Normal. Practitioners will find this approach useful when limited observations are available and a specific parametric model cannot be assumed.

Keywords: nonparametric, empirical distribution, SPC, Monte Carlo, asymptotic properties

technometrics tex template (do not remove)

1 Introduction

Statistical process control can be traced back to Shewhart (1922, 1931) as a strategy to monitor the quality of a process. Shewhart understood that a constant quality doesn't exist, only variable quality that could be predicted between certain limits, later known as control limits. This way, if measurements from a process fall within these limits, it is said that the process is under control, i.e., its performance is predictable. This is the basic principle of what it is now called statistical process control (SPC). As a consequence, to implement quality controls on a product, it is required to estimate the corresponding control limits from a sample, otherwise a priori information is needed. However, prior information might not be available, and samples might not be as large as desired to have decent estimation of a population distribution. This situation invites nonparametric procedures that might work when no other option is available.

During a Phase I of an SPC implementation, from a parametric perspective, when a sample is not large enough to estimate the parameters of a process, but certain distribution can be assumed, it might be possible to implement an approach developed by Hawkins (1987), the first self-starting control chart. A self-starting control chart does not requires a Phase I, or estimation phase, to be implemented. Hawkins (1987) proposed a scheme where unknown parameters are estimated with running estimators that improve as observations are incorporated. He integrates in his method a series of transformations over independent and identically distributed (i.i.d.) normal observations $\mathrm{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown, to obtain i.i.d. $\mathrm{N}(0, 1)$ statistics that can be analyzed with traditional SPC methods that assume normality and known parameters. In his paper, Hawkins uses a CUSUM approach to evaluate the proposed statistic. He also suggested the use of a Shewhart $\overline{X}$-chart to address situations with isolated changes.

Several control charts were developed following Hawkins' strategy by making different assumptions about the distribution of the process. For instance. Quesenberry (1991c,a,b) continued the work of Hawkins by proposing similar schemes when the knowledge about parameters from a normal distribution is partial, or when observations come from a Binomial or Poisson process. Zou et al. (2007a) worked with linear profiles. Capizzi and Masarotto (2012) implemented a CUSCORE approach. Sullivan and Jones (2002a) developed a MEWMA to address vectors of observations. All of these and others followed the self-starting scheme by Hawkins.

A different strategy to create self-starting charts was proposed by Hawkins et al. (2003). Using an approach similar to the generalized likelihood ratio test, a change-point approach was developed to create control charts based on the recursive use of test statistics. By using test statistics to evaluate unknown parameters, the approach is capable of creating new monitoring techniques that can be considered as self-starting. As seen in the following section, many other developments followed the same idea. Hawkins and Zamba (2005) and Zhou et al. (2009) are only a few examples of the flexibility provided by the approach. This method showed relatively good performance to detect sustained changes. However, the increase in power comes with an increase in computational complexity and a dependance on simulated data to estimate the quantiles needed to obtain control limits during the development.

Besides Zhou et al. (2009), most of previously mentioned approaches, even though they are self-starting, require knowledge about the distribution function. Ignorance is assumed to be limited only to the parameters. In practice, this is not always the case, and a distribution-free approach might become useful. On one hand, there are several nonparametric approaches like the ones presented by Das (2008) and Das and Bhattacharya (2008) that require large amounts of data to work in practice. On the other hand, powerful approaches like Zhou et al. (2009), Ross et al. (2011), or Ross and Adams (2012) depend on high computational times that can be controlled by using moving windows of observations, which add (in many cases) an unwanted risk of not detecting a shift when it happens.

To address the issue of power to detect shifts in a process without requiring large samples or an elevated number of computations, a method is developed by adapting the concept of the empirical distribution function to create a running estimator of the true distribution that can be used to construct a nonparametric Hawkins-type self-starting control chart. Similar to Hawkins (1987), the proposed statistic can be used to monitor location and scale shifts in series of independent and identically distributed observations where knowledge about parameters and the true underlying distribution function does not exist.

The following sections of the paper are organized as follows: Section 2 reviews the literature of the self-starting control charts; Section 3 describes the proposed method and its assumptions without concerns about the theoretical basis, and presents two numerical examples to illustrate the corresponding implementation process. Section 4 summarizes the theory behind the method by showing the mathematical properties of the empirical distribution and how it is adapted to create the proposed statistic used in a tabular CUSUM chart. Performance over limited samples is discussed in Section 5. Finally, Section 6 concludes with some final comments over the applicability of the proposed method and a description of future works.

2 Background

2.1 Previous work on self-starting control charts

Self-starting control charts, developed at first by Hawkins (1987), are an alternative to traditional control charts used without the initial parameter's estimation, phase I of SPC, whose performance (in terms of Type I Errors) remains stable. Jensen et al. (2006) presented a literature review of the effects of parameter estimation in control chart's performance and they stated that when it is prohibitive to have a large number of samples for parameter's estimation it is recommended to use approaches that avoid distinction between the estimation Phase and the monitoring Phase, such as a self-starting approach. Jones et al. (2001) and Jones et al. (2004) described the effect of parameter's estimation in the ARL of EWMA and CUSUM control charts, respectively.

Because of its performance and applicability several self-starting control charts have been developed for different situations, in particular to monitor a series of normal observations. Zantek (2006) designed a self-starting CUSUM using Q statistics to detect shifts in the mean of independent normal observations, designing CUSUM charts to obtain the best performance which can be used in start-up and short run processes. Li et al. (2010) created a self-starting control chart based on a likelihood ratio test and an EWMA, SSELR chart, to detect changes in both mean and variance in a series of normal samples. This control chart is sensitive to different types of shift even for a decrease in variability; and can be adapted to be used with a sample size of one. Li and Wang (2010) proposed a self-starting control chart to detect changes in the mean of normal observations, integrating CUSUM of Q chart and an adaptive reference value for a better performance. Capizzi and Masarotto (2012) developed an adaptive CUSCORE control chart (ACUSCORE) for monitoring location in normal observations by using adaptive EWMA.

Monitoring the fraction of conforming units was analyzed under this approach by Zhang et al. (2012), who developed a self-starting cumulative count-of-conforming chart for group inspection (CCCg) by integrating the self-starting approach and an approximately average run length. Their results allow to use this control chart for monitoring a process as soon as possible, with at least 25 observations.

Multivariate self-starting control charts have been created under different schemes to monitor multivariate normal process as can be seen in Sullivan and Jones (2002b), Zantek et al. (2006), Li et al. (2014), and Hawkins and Maboudou-Tchao (2007). Also, for monitoring profiles, Zou et al. (2007b) designed a self-starting EWMA control chart for the detection of shifts in slope and standard deviation in linear profiles with normal errors based on recursive residuals.

Functional changes in a time series can be found from a nonparametric perspective since Page (1955), who proposed a test based on the sign of $(X_i - \mu_0)$, when the mean μ_0 is known. Later, with the development of a CUSUM chart and the asymptotic distribution of a CUSUM statistic, Pettitt (1980) extended Page's work.

Bhattacharyya and Johnson (1968) developed several tests for a retrospective analysis of change-point based on transformations, including a test based on the sign test, the Wilcoxon rank test, and one sample normal scores test. Pettitt (1979) created tests for changes based on Bernoulli observations, Binomial random variables, and the two sample Mann-Whitney U statistic. By using aligned rank transformations, Sen (1977) proposed an asymptotically distribution free procedure to test changes in location; and extended this work to detect a change in the slope of a regression line in Sen (1980). Theoretical results about the asymptotic properties of cumulative sums of sequential ranks can be found in Bhattacharya and Frierson Jr (1981). Darkhovski (1994) used a Kolmgorov-Smirnov type statistic to test functional changes in a series of independent observations. Although this test can be used for any type of changes in the distribution, it's not as powerful as a test developed for an specific kind of change.

Empirical likelihood function was used by Zou et al. (2007b) to develop a test and an estimator of change-point, which have the asymptotic behavior of a parametric test with known probability distribution. Liu et al. (2008) used these results to detect changes in linear regression models.

Rafajłowicz et al. (2010) proposed a nonparametric control chart by using moving windows to monitor changes for observations centered on the last observation in control available. Although the use of a moving window reduces the computational complexity of the control chart, if a change-point is not detected when time windows pass over it, it might not be detected later. Mukherjee and Bandyopadhyay (2011) presented a method to detect changes in the trend of observations based on rank transformations using sequential testings.

Nonparametric self-starting control charts can be found in the literature. Liu et al. (2013) created a nonparametric self-starting control chart using sequential ranks in an adaptive nonparametric EWMA (ANE) control chart, for monitoring shifts in location, whereas control limits are obtained using a regression. Their results show robust performance over different distributions but when dealing with skewed distribution and small shifts its performance is not as good as the nonparametric control chart based on the Mann-Whitney two sample test developed by Hawkins and Deng (2010). Liu et al. (2014) provide a self-starting nonparametric EWMA control chart using sampling intervals to monitor shifts in location, which has better performance that the ANE control chart.

Overviews of nonparametric change-point methods can be read in Wolfe and Schechtman (1984), and Brodsky and Darkhovsky (1993). Amiri and Allahyari (2012) presented a literature review in change-point analysis and developed a classification scheme and a gap analysis for future research. A review of theoretical results and applications in univariate and multivariate change-point analysis from a classical parametric approach can be found in Chen and Gupta (2011).

3 Empirical self-starting CUSUM Chart

3.1 The Method

To monitor location or scale changes in a series of independent and identically distributed (i.i.d.) observations without assuming a distribution function, the following method is proposed:

Let $X_1, ..., X_n$ be a series of independent observations where n stands for the last observation of the series. k is defined as the last known in-control observation, hence $k \leq n$. If there is a change-point, it is located at moment $\tau > k$, $X_i \sim F(x)$ for $i = 1, ..., \tau$, and $X_i \sim F(x + a)$ or $X_i \sim F(x/a)$ for $i = \tau + 1, ..., n$. An observation X_i, where $i > k$, is monitored using the following functions:

$$P_i = \frac{1}{i} \left(\sum_{j=1}^{i} I(X_j \leq X_i) - 0.5 \right) \tag{1}$$

where $I(X_j \leq X_i)$ is the indicator function

$$I(X_j \leq X_i) = \begin{cases} 1 & X_j \leq X_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

P_i corresponds to the evaluation of a given observation X_i in an empirical distribution function (with a correction factor of $0.5/i$) obtained from observations 1 to i.

P_i is transformed into an approximate $N(0, 1)$ using

$$Z_i = \Phi^{-1}(P_i) \tag{3}$$

where operator Φ^{-1} stands for the inverse cumulative standard normal distribution.

As seen in Section 4, series $\{Z_i : i = 1, ..., n\}$ consist of mutually independent asymptotically standard normal random variables under the in-control assumption. When a sustained shift occurs in location after moment τ, Z_i values get biased to positive or negative values, depending on the direction of the change, due to an inertia created by the amount of in-control values used in the evaluation of new observations. Hence, any control chart developed to monitor location changes in normal observations is suited as a tool for analysis, mostly if the chart considers the history of the process. For the purpose of this research, a CUSUM approach is used to evaluate performance. The corresponding CUSUM chart is

$$L_i^+ = \max\left(0, L_{i-1}^+ + Z_i - k\right) \tag{4}$$

$$L_i^- = \max\left(0, L_{i-1}^- - Z_i - k\right) \tag{5}$$

where

$$L_0^+ = L_0^- = 0 \tag{6}$$

k can be interpreted as an allowance. It is often defined as a standardized mid-range between the CUSUM target value (or known process mean) and the out-of-control mean desired to detect quickly. Values of 0.25 and 0.5 are usually selected to detect shifts in a process of 0.5 and 1 standard deviation, respectively.

To monitor scale changes, as presented in Hawkins (1987), the following standardized quantity is used

$$V_i = \frac{\sqrt{|Z_i|} - 0.822}{0.349} \tag{7}$$

Then, the self-starting CUSUM chart for scale is

$$S_i^+ = \max\left(0, S_{i-1}^+ + V_i - k\right) \tag{8}$$

$$S_i^- = \max\left(0, S_{i-1}^- - V_i - k\right) \tag{9}$$

where

$$S_0^+ = S_0^- = 0 \tag{10}$$

A shift in a series is signaled when L_i^+, L_i^-, S_i^+, or S_i^- are bigger or equal than a value h. $h > 0$ is usually set to obtain a desired in-control ARL.

3.2 Numerical example

To illustrate the application of the empirical SSCUSUM, data from a hard-brake process and Dow Jones montly increments is used as real life applications of the proposed control charts for location and scale. Data was obtained from Montgomery (2008) and Yahoo! (2014) respectively.

First, a hard-brake process is used in conjunction with photolithography in order to manufacture semiconductors. It is desired to monitor changes in location of the flow width attached to this process. Twenty-five samples of wafers were taken of five observations each. Individual observations are plotted and studied using the proposed empirical self-starting control chart for location. A change was detected at observation 193 (see Figure 1). Ross and Adams (2012) analyzed this data set finding a signal of a change at moment 199, estimating that the true moment of change ocurred between 186 and 187.

Second, montly increments of Dow Jones close price (adjusted for dividends and splits) were evaluated using the proposed empirical SSCUSUM for scale changes. Data from March 2003 to November 2014 was use in the analysis. Original observations and the evaluation of the empirical SSCUSUM for scale can be seen in Figure 2. It can be seen that a shift in scale was signaled at moment 59, which corresponds to February 1, 2008. The change detected agrees with a period of economic change well known as the U.S. subprime mortage crisis and the financial crisis of 2007-2008. Since Dow Jones increments continues to be monitored after the change, the chart moves back to control due to the increment of out-of-control observations in the reference set. These results agree with the ones obtained using Hawkins' SSCUSUM, but without making assumptions about the distribution of the data, and, as seen in Sections 4 and 5, the statistical confidence is guarantee.

Finally, to show how calculations were made, the first 10 evaluations of the Dow Jones index are presented in Table 1 in some detail. Both statistics for location and scale were used during the monitoring. As seen in the table, monitoring begins at $i = 2$. Increments of Dow Jones can only be determined after the first measure, and non-trivial evaluations of the empirical distribution are obtained after the second one. Calculations can be made by hand, however, the use software such as a general purpose spreadsheet application is recommended. Because of this, it is believed that the procedure can be readily applied with relatively little training and investment.

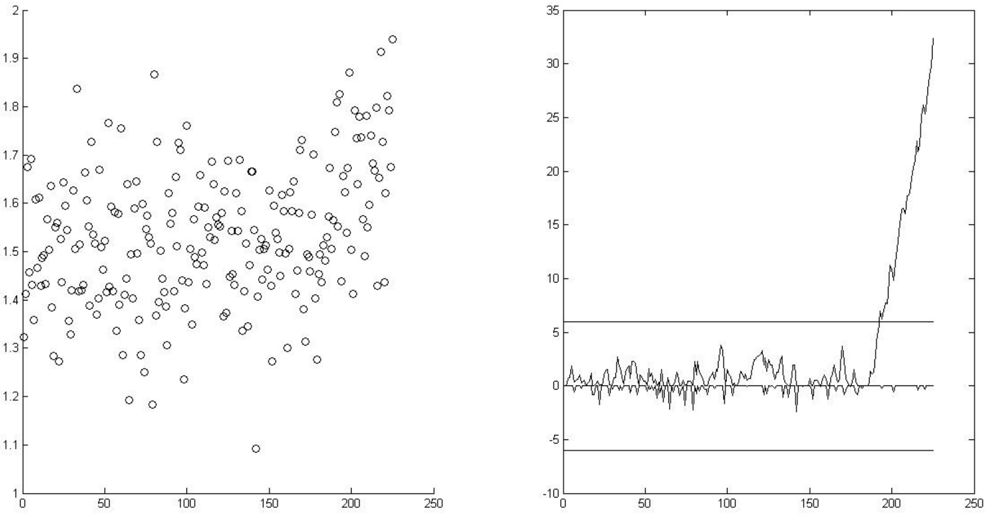

Figure 1: Data from the Hard-Bake Process and the corresponding CUSUM for location

4 Theoretical framework

4.1 Empirical Distribution Function

Definition 1. *Let $X_1, X_2, ..., X_n$ be a random sample with probability function $F(x)$. The empirical distribution function $S(x)$ (e.d.f.) is a function that estimates the true distribution function $F(x)$:*

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{11}$$

where I is an indicator function.

$S_n(x)$ is a maximum likelihood estimator of $F(x)$. Given a value of x

$$E\{S_n(x)\} = F(x) \tag{12}$$

$$Var\{S_n(x)\} = \frac{1}{n}F(x)(1 - F(x)) \tag{13}$$

which implies that

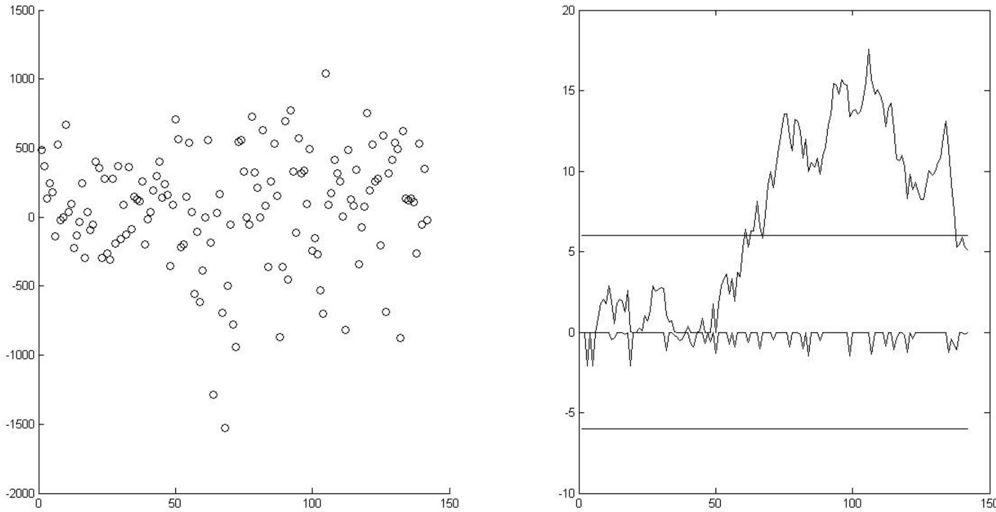

Figure 2: Data from the Dow Jones index monthly increments and the corresponding CUSUM for scale

$$S_n(x) \xrightarrow{p} F(x) \tag{14}$$

By applying the Strong Law of Large numbers, the Glivenko-Cantelli Theorem (Wellner, 1977) states that

$$\lim_{n\to\infty} \sup |S_n(x) - F(x)| \xrightarrow{a.s.} 0 \tag{15}$$

Equation (12) implies that $S_n(x)$ is an unbiased estimator of $F(x)$, the true cumulative distribution function. Also, Equation (13) indicates that the variance of $S_n(x)$ tends to zero when n tends to infinite, which implies that $S_n(x)$ is consistent, and, as indicated in Equation (14), $S_n(x)$ converges in probability to $F(x)$. A stronger result is shown in Equation (15); this means that $S_n(x)$ converges uniformly to $F(x)$. These properties are easily verified by understanding that $S_n(x)$ is a function of a sum of Bernoulli random variables, and the corresponding relationship with the Strong Law of Large Numbers. As a consequence, it can be stated that e.d.f is a excellent statistic to be used in place of the true distribution function when the latter is not known.

Table 1: First 10 iterations when evaluating Dow Jones index monthly increments of close prices using the Empirical SSCUSUM for location and scale

i	Date	Adj. Dow Jones	Increments (X_i)	$R_i(X_i)$	$S_i(X_i)$	U_i	V_i	L_i^-	L_i^+	S_i^-	S_i^+
0	Mar 3, 2003	7,992.13	-	-	-	-	-	-	-	-	-
1	Apr 1, 2003	8,480.09	487.96	-	-	-	-	-	-	-	-
2	May 1, 2003	8,850.26	370.17	2	0.50	0.00	-2.36	0.00	0.00	-2.11	0.00
3	Jun 2, 2003	8,985.44	135.18	2	0.38	-0.32	-0.74	-0.07	0.00	0.00	0.00
4	Jul 1, 2003	9,233.80	248.36	3	0.50	0.00	-2.36	0.00	0.00	-2.11	0.00
5	Aug 1, 2003	9,415.82	182.02	3	0.42	-0.21	-1.04	0.00	0.00	0.00	0.00
6	Sep 2, 2003	9,275.06	-140.76	1	0.07	-1.47	1.11	-1.22	0.00	0.00	0.86
7	Oct 1, 2003	9,801.12	526.06	8	0.94	1.53	1.19	0.00	1.28	0.00	1.81
8	Nov 3, 2003	9,782.46	-18.66	2	0.17	-0.97	0.46	-0.72	0.07	0.00	2.02
9	Dec 1, 2003	10,453.92	671.46	10	0.95	1.64	1.32	0.00	1.46	0.00	3.09

4.2 Sequence of Independent Statistics

The process to adapt the e.d.f. to create a series of independent and asymptotically normal random variables to use in a CUSUM monitoring is described in the following.

Definition 2. *Given a series of i.i.d. $X_1, X_2, ..., X_n$ with continuous probability function $F(x)$, $R_i(X)$ is defined as a running rank statistic where*

$$R_i(x) = \sum_{j=1}^{i} I(X_j \leq x) \tag{16}$$

Proposition 1. *Let $X_1, X_2, ..., X_n$ be series of i.i.d. observations, then $R_1(X_1)$, $R_2(X_2)$, ..., $R_n(X_n)$ are independent.*

Proof. Let us define $R_i = R_i(X_i)$. Then, because $X_1, X_2, \ldots, X_n$ are i. i. d., it is well known that the rank of X_n, relative to $X_1, X_2, \ldots, X_{n-1}$, has the probability distribution

$$P\left(R_n = r\right) = \begin{cases} \frac{1}{n} & r = 1, 2, ..., n \\ 0 & otherwise \end{cases} \tag{17}$$

Because the probability distribution of R_n is independent of variables $R_1, R_2, \ldots R_{n-1}$; then R_n is independent of $R_1, R_2 \ldots, R_{n-1}$ and vice versa.

□

The following theorems will be used to prove Proposition 2. Proof of these two theorems is found in Casella and Berger (2002) and Mood et al. (1974) respectively.

Theorem 1. *Let $X_1, X_2, ..., X_n$ be independent random vectors. Let $g_i(x_i)$ be a function only of $x_i, i = 1, 2, ..., n$. Then the random variables $U_i = g_i(X_i), i = 1, 2, ..., n$ are mutually independent.*

Theorem 2. *Probability Integral Transformation. If X is a random variable with continuous cumulative distribution function $F_X(x)$, then $U = F_X(X)$ is uniformly distributed over the interval $(0, 1)$. Conversely, if U is uniformly distributed over the interval $(0, 1)$, then the inverse function $F_x^{-1}(u)$ has cumulative distribution function $F_X(\cdot)$.*

Proposition 2. *Let $X_1, X_2, ..., X_n$ be a sequence of i.i.d. random variables. Define*

$$P_i = \frac{(R_i(X_i) - 0.5)}{i} \tag{18}$$

$$Z_i = \Phi^{-1}(P_i) \tag{19}$$

where Φ^{-1} stands for the inverse of the standard normal cumulative distribution function. Then,

1. *Series $\{P_i : 1, \ldots, n\}$ are mutually independent random variables.*

2. *$\{Z_i : 1, \ldots, n\}$ are independent asymptotically standard normal random variables.*

Proof. Since Proposition 1 states that $\{R_i : i = 1, 2, \ldots\}$ are mutually independent, it follows by Theorem 1 that $\{P_i : i = 1, 2, \ldots\}$ are also mutually independent random variables. It can be seen that definitions of P_i in Proposition 2 and Equation (1) are equivalent.

To prove the second part, we'll use the fact that $R_i(X_i) = R_{i-1}(X_i) + 1$. Hence,

$$\begin{aligned} P_i &= \frac{(R_{i-1}(X_i) + 1 - 0.5)}{i} \\ &= \frac{\left(\sum_{j=1}^{i-1} I(X_j \leq X_i) + 0.5\right)}{i - 1 + 1} \\ &= \frac{(i-1)S_{i-1}(X_i)}{i} + \frac{0.5}{i} \end{aligned}$$

where $S_{i-1}(\cdot)$ follows Definition 1, and $i - 1$ indicates that only the previous observations 1 to $i - 1$ are used to estimate $F_X(\cdot)$, the distribution function of X. This implies that, the sequence of functions $\{P_i - S_{i-1}(X_i)\}$ converges uniformly to the function 0, i.e., $\{P_i - S_{i-1}(X_i)\} \to 0$ uniformly. Now, since the sequence $S_{i-1}(X_i) \to F_X(X_i)$ uniformly and almost surely (a.s.), then $P_i \to F_X(X_i)$ a.s. uniformly. As a consequence, by Theorem 2, P_i follows asymptotically a Uniform $(0, 1)$ distribution, which means that, also by Theorem 2, $Z_i = \Phi^{-1}(P_i) \sim N(0, 1)$ as $i \to \infty$, which concludes the proof. □

Since random variables Z_i meet, asymptotically, the corresponding assumptions of a CUSUM chart, it seems to make sense to use a CUSUM chart to perform a monitoring of the independent random variables Z_i in the same manner as done by Hawkins (1987) and stated in Section 3.

4.3 Remarks

To avoid estimated probabilities of 1, which leads to undesired infinite values when evaluated in a normal distribution, observations obtained from the empirical distribution are shifted by subtracting 0.5 from the numerator, as seen in P_i and defined in Proposition 2. 0.5 assures balanced tail probabilities, but points are not evenly spaced, it creates an artificial bias of $0.5/i$ but achieves the same variance as the empirical distribution. Due to this fact, readers might prefer the use of $R_i(X_i)/(i+1)$. However, they should be aware that this latter formulation, even though it has a smaller standard error and evenly spaced points over the Uniform distribution, has a bigger bias when estimating upper tail probabilities. This situation is reflected as a slower approximation to the limiting normal distribution. and Monte Carlo simulation showed that the estimated ARL differs more from the one expected over normal observations. The true Type I Error probability is reduced, which is good, but the power to detect changes is reduced as well, which is not good. For this reason, the approach selected to estimate cumulative probabilities was $(R_i(X_i) - 0.5)/i$.

5 Performance Results

To evaluate performance, extensive Monte Carlo simulations were used. Observations were generated from $N(0, 1)$, $G(0.5, 1)$, and $t(3)$ distributions. Here, 0.5 stands for the shape parameter and 1 for the scale parameter of a Gamma distribution. To evaluate the Empirical

SSCUSUM for location changes, if $H(x)$ is the function used to represent the in-control behavior, $H(x+\delta)$ was used to represent the out-of-control situation. The value of δ is given in standard deviations. A Gamma distribution was used to evaluate a skewed distribution, $t(3)$ was used for symmetric but heavy tailed distributions, and the Normal was tested to assess how much performance is compromised when the data is actually Normal, and the nonparametric approach is used instead of the parametric one. To perform simulation, τ in-control observations were generated prior to monitoring. From time $\tau+1$ observations were monitored until an out-of-control signal is given by the corresponding control chart. To measure performance, the average run length (ARL) after the change-point τ was used. For example, in Table 2, in-control observations follow a normal with mean zero and standard deviation one. After τ in-control observations, the mean changes, and the run length is the number of observation since the shift until an observation surpasses the corresponding control limits. This was repeated 10,000 times, and the average run length is calculated to estimate the ARL of the given scenario.

As seen in Table 2, self-starting CUSUM is an efficient tool to monitor normal observations. Tables 2a and 2b show similar results, with a slight advantage going to Hawkins' SSCUSUM. The fact that the Empirical SSCUSUM has similar performance as Hawkins' chart indicates that the use of the empirical distribution provides a relatively good approximation of standard normal distribution, and the approach can be used even in situations when normality is the true underlying unknown distribution without a major compromise in power to detect changes. As in Hawkins (1987), a history of 75 in-control observations seems an adequate number to start the monitoring. ARLs are so similar between methods, that maximum difference in performance found was 5.4, when only 25 in-control observations existed and there was no change in location.

The empirical control chart is a distribution-free method under the null hypothesis of no change. Therefore as seen in Tables 2 to 5, in-control ARL remains almost the same (differences are due to the simulation error). This means, that the probability of Type I Error is controlled over any distribution, and the analysts will always have the same confidence with their decisions. In-control ARL of Hawkins' SSCUSUM oscillates. Under a t distribution with 3 degrees of freedom, in-control ARL is smaller in most cases, and relatively bigger when dealing with a Gamma. As seen in Table 3, even though the in-control ARL is smaller, out-of-control ARL is bigger. By using the empirical distribution, not only the Type I Error is under control, but also, the power to detect shifts is improved.

Due to lack of symmetry in Gamma distributions, it is also necessary to evaluate performance for downward changes. Results are given in tables 4 and 5. As can be seen, out-of control ARL of the parametric approach that assumes normality underperforms in almost all scenarios. For a change of $\delta = -0.5$, the out-of-control ARL of the nonparametric approach is almost 10 times smaller.

Table 2: ARL_δ performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually $\mathrm{N}(\delta, 1)$

(a) Empirical SSCUSUM

$\mathbf{ARL}_\delta$	δ				
τ	**0**	**0.5**	**1**	**1.5**	**2**
25	130.0	67.1	15.7	8.1	6.0
50	128.0	41.9	11.1	6.7	5.2
75	129.2	32.6	10.2	6.4	4.8
100	127.0	28.8	9.8	6.2	4.7
500	128.5	21.9	8.9	5.7	4.3

(b) Hawkins' SSCUSUM

$\mathbf{ARL}_\delta$	δ				
τ	**0**	**0.5**	**1**	**1.5**	**2**
25	125.4	61.7	14.6	7.2	5.1
50	126.4	39.4	10.6	6.3	4.5
75	126.0	31.3	9.9	6.0	4.4
100	125.2	28.5	9.5	5.9	4.3
500	125.6	22.0	8.9	5.6	4.1

Table 3: ARL_δ performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually t with 3 degrees of freedom

(a) Empirical SSCUSUM

$\mathbf{ARL}_\delta$	δ				
τ	**0**	**0.5**	**1**	**1.5**	**2**
25	131.8	46.0	10.8	6.7	5.5
50	130.2	26.1	8.4	5.8	4.8
75	131.9	21.0	8.0	5.5	4.6
100	126.3	18.4	7.7	5.4	4.4
500	126.3	15.8	7.2	5.0	4.2

(b) Hawkins' SSCUSUM

$\mathbf{ARL}_\delta$	δ				
τ	**0**	**0.5**	**1**	**1.5**	**2**
25	107.1	59.1	17.8	10.2	6.5
50	112.9	45.8	13.4	7.3	5.0
75	120.8	38.7	13.0	6.1	4.6
100	122.2	34.4	10.8	6.0	4.6
500	140.0	26.5	8.7	5.9	4.2

Similar results are found when the approach is evaluated to monitor changes in scale, i.e., when the monitored statistic is the one from equation (7). The in-control distribution is $H(x;\theta_0)$ and the out-of-control is $H(x;\theta_1)$. Tables 6 to 8 show this performance. For

Table 4: ARL$_\delta$ performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually Gamma $(0.5) + \delta : \delta \geq 0$

(a) Empirical SSCUSUM

ARL$_\delta$	δ				
τ	**0**	**0.5**	**1**	**1.5**	**2**
25	130.0	24.2	10.1	7.4	6.3
50	130.5	12.8	7.8	6.3	5.5
75	128.0	11.4	7.4	6.0	5.3
100	126.5	10.8	7.2	5.9	5.2
500	124.2	9.8	6.8	5.6	4.9

(b) Hawkins' SSCUSUM

ARL$_\delta$	δ				
τ	**0**	**0.5**	**1**	**1.5**	**2**
25	126.2	76.1	31.7	12.8	7.2
50	127.1	61.9	20.1	7.5	5.1
75	137.9	56.2	14.5	7.0	4.8
100	139.5	52.5	12.8	6.8	4.8
500	155.2	33.5	10.6	6.3	4.5

Table 5: ARL$_\delta$ performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually Gamma $(0.5) + \delta : \delta \leq 0$

(a) Empirical SSCUSUM

ARL$_\delta$	δ				
τ	**0**	**-0.5**	**-1**	**-1.5**	**-2**
25	130.0	7.7	6.0	5.4	5.0
50	130.5	6.2	4.9	4.5	4.3
75	128.0	5.3	4.3	3.9	3.7
100	126.5	4.9	4.1	3.7	3.5
500	124.2	4.3	3.6	3.4	3.2

(b) Hawkins' SSCUSUM

ARL$_\delta$	δ				
τ	**0**	**-0.5**	**-1**	**-1.5**	**-2**
25	126.2	74.6	12.5	6.9	5.1
50	127.1	42.2	9.9	6.3	4.7
75	137.9	32.2	9.6	6.1	4.6
100	139.5	28.3	9.4	6.0	4.5
500	155.2	21.5	9.0	5.8	4.4

instance, Table 6 indicates that the Empirical SSCUSUM and Hawkins' SSCUSUM have similar in-control ARLs for normal distribution.

Tables 7 and 8 show the same in-control ARL, even though distributions are symmetric and skewed with heavy tails, respectively. Nevertheless, Hawkins' SSCUSUM in-control ARL dimished, indicating a larger amount of false alarms if used in practice. Out-of-control ARLs of the Empirical SSCUSUM for scale give relatively good results, mostly if at least 75 in-control observations are used to start the monitoring. Although it is difficult to compare the relative performance in Tables 7 and 8 because of the drastic difference in the actual α, the power to detect changes in variance using the Empirical SSCUSUM is most of the time better when dealing with the Gamma scenarios, even when only 25 in-control observations exist. As a general rule, the proposed nonparametric control chart, when dealing with changes in scale, seems to perform better when at least 75 observations are used to start a monitoring, in agreement with the recommendations given by Hawkins (1987) for the SSCUSM.

Additional simulation results, not shown here due to space constraints, show that the Empirical SSCUSUM for location is relatively robust to changes in variance. However, the Empirical SSCUSUM for scale changes is sensitive to both location and scale changes. These results are similar to the ones found by Hawkins (1987) with the SSCUSUM for normal observations. Hence, in practice, on one hand, if a signal is detected using the Empirical SSCUSUM for location, it is more likely due to a change in location. On the other hand, if a signal is triggered in the chart for scale changes, it might be due to shift in the location or scale of observations. This implies that a change in location would be detected by both charts, and a change in scale is more likely to be detected only by the Empirical SSCUSUM for scale.

Table 6: ARL_{σ_1/σ_0} performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually Normal

(a) Empirical SSCUSUM

ARL_{σ_1/σ_0}	σ_1/σ_0						
τ	**0.51**	**0.64**	**0.8**	**1**	**1.25**	**1.56**	**1.995**
25	48.0	84.4	118.8	127.4	100.2	55.6	25.4
50	26.5	57.5	108.3	125.3	76.5	30.2	15.3
75	20.3	45.4	103.9	125.2	64.0	24.6	13.0
100	18.4	38.4	96.1	127.3	58.3	21.6	12.3
500	15.3	26.5	71.7	127.4	41.5	16.9	10.2

(b) Hawkins' SSCUSUM

ARL_{σ_1/σ_0}	σ_1/σ_0						
τ	**0.51**	**0.64**	**0.8**	**1**	**1.25**	**1.56**	**1.995**
25	28.1	64.3	113.9	123.3	89.5	45.6	20.5
50	18.8	43.4	103.7	125.9	69.5	26.4	12.7
75	17.1	35.8	98.0	124.1	58.7	21.5	11.2
100	16.4	32.7	89.6	125.0	54.9	19.5	10.4
500	15.1	25.9	68.9	122.9	40.8	16.1	9.3

Table 7: ARL_{σ_1/σ_0} performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually t(3)

(a) Empirical SSCUSUM

ARL_{σ_1/σ_0}	σ_1/σ_0						
τ	**0.51**	**0.64**	**0.8**	**1**	**1.25**	**1.56**	**1.995**
25	59.0	91.5	117.6	129.5	109.0	75.5	43.1
50	34.5	68.1	112.0	126.9	92.7	49.1	23.3
75	26.4	56.0	106.5	127.5	85.8	38.4	19.2
100	23.1	48.9	99.8	125.5	78.6	33.7	17.4
500	18.2	32.1	79.8	126.7	59.3	25.0	14.5

(b) Hawkins' SSCUSUM

ARL_{σ_1/σ_0}	σ_1/σ_0						
τ	**0.51**	**0.64**	**0.8**	**1**	**1.25**	**1.56**	**1.995**
25	17.0	25.1	36.5	47.2	49.9	44.6	32.6
50	13.2	19.7	30.5	44.2	48.9	41.1	26.4
75	12.1	17.6	27.9	42.9	49.2	38.9	24.4
100	11.6	16.4	26.3	41.5	49.1	38.4	22.4
500	10.3	13.7	21.1	37.1	50.0	35.2	19.7

Table 8: ARL_{σ_1/σ_0} performance of the empirical SSCUSUM and the traditional SSCUSUM from Hawkins when readings are actually $\text{Gamma}(0.5, 1)$

(a) Empirical SSCUSUM

ARL_{σ_1/σ_0}	σ_1/σ_0						
τ	**0.51**	**0.64**	**0.8**	**1**	**1.25**	**1.56**	**1.995**
25	20.0	31.3	71.0	127.5	38.1	16.5	10.9
50	14.2	17.0	42.4	130.7	18.3	10.2	7.7
75	13.6	14.7	32.1	127.8	14.7	8.8	7.0
100	13.3	13.6	28.1	125.7	13.0	8.2	6.6
500	12.7	11.6	20.6	125.6	9.4	6.4	5.3

(b) Hawkins' SSCUSUM

ARL_{σ_1/σ_0}	σ_1/σ_0						
τ	**0.51**	**0.64**	**0.8**	**1**	**1.25**	**1.56**	**1.995**
25	17.3	33.0	60.1	82.7	87.1	70.3	47.0
50	13.5	23.2	49.5	83.4	88.2	64.0	33.0
75	12.7	20.3	43.5	82.0	92.3	58.7	26.1
100	12.3	19.1	39.5	82.0	92.5	54.7	22.6
500	11.4	16.5	30.9	76.4	92.0	36.7	15.8

6 Conclusions

This paper introduced a nonparametric monitoring procedure for location and scale in a series of i.i.d. observations based on the recursive application of the empirical distribution function to create a statistic that is mutually independent over time under the in-control hypothesis. The limiting distribution of the proposed statistic is standard normal, and a CUSUM is used for monitoring. The approach works with individual observations, and it is distribution-free under the null hypothesis. When dealing with normal observations, the proposed method presents a similar behavior to the self-starting CUSUM chart of Hawkins (1987), but power to detect changes increases relative to the parametric approach for non-normal distributions. Due to its construct, the procedure is considered to be robust against individual outliers. Hence, it is not recommended if isolated changes are a concern. The method is best used to address sustained changes. In addition, the proposed statistic, due to its approximation to the standard normal distribution, can be monitored using other approaches beside the CUSUM evaluated in this research. Practitioners are encouraged to test the approach using other schemes to monitor shifts over time, like EWMA or CUSCORE. Finally, due to the self-starting nature of the statistic, procedures to increase the inertia of in-control observations might be developed to reduce the natural bias that this type of self-starting procedures have.

References

Amiri, A. and Allahyari, S. (2012). Change point estimation methods for control chart postsignal diagnostics: a literature review. *Quality and Reliability Engineering International*, 28(7):673–685.

Bhattacharya, P. and Frierson Jr, D. (1981). A nonparametric control chart for detecting small disorders. *The Annals of Statistics*, pages 544–554.

Bhattacharyya, G. and Johnson, R. A. (1968). Nonparametric tests for shift at an unknown time point. *The Annals of Mathematical Statistics*, pages 1731–1743.

Brodsky, E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems*. Number 243. Springer.

Capizzi, G. and Masarotto, G. (2012). An enhanced control chart for start-up processes and short runs. *Quality Technology & Quantitative Management*, 9(2):189–202.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer.

Darkhovski, B. S. (1994). Nonparametric methods in change-point problems: A general approach and some concrete algorithms. *Lecture Notes-Monograph Series*, pages 99–107.

Das, N. (2008). Non-parametric control chart for controlling variability based on rank test. *Economic Quality Control*, 23(2):227–242.

Das, N. and Bhattacharya, A. (2008). A new non-parametric control chart for controlling variability. *Quality Technology and Quantitative Management*, 5(4):351–361.

Hawkins, D. (1987). Self-starting CUSUM charts for location and scale. *The Statistician*, page 299–316.

Hawkins, D. M. and Deng, Q. (2010). A nonparametric change-point control chart. *Journal of Quality Technology*, 42(2):165–173.

Hawkins, D. M. and Maboudou-Tchao, E. M. (2007). Self-starting multivariate exponentially weighted moving average control charting. *Technometrics*, 49(2):199–209.

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

Hawkins, D. M. and Zamba, K. (2005). A change-point model for a shift in variance. *Journal of Quality Technology*, 37(1):21–31.

Jensen, W. A., Jones-Farmer, L. A., Champ, C. W., Woodall, W. H., et al. (2006). Effects of parameter estimation on control chart properties: a literature review. *Journal of Quality Technology*, 38(4):349–364.

Jones, L. A., Champ, C. W., and Rigdon, S. E. (2001). The performance of exponentially weighted moving average charts with estimated parameters. *Technometrics*, 43(2).

Jones, L. A., Champ, C. W., and Rigdon, S. E. (2004). The run length distribution of the cusum with estimated parameters. *Journal of Quality Technology*, 36(1):95–108.

Li, Y., Liu, Y., Zou, C., and Jiang, W. (2014). A self-starting control chart for high-dimensional short-run processes. *International Journal of Production Research*, 52(2):445–461.

Li, Z. and Wang, Z. (2010). Adaptive cusum of the q chart. *International Journal of Production Research*, 48(5):1287–1301.

Li, Z., Zhang, J., and Wang, Z. (2010). Self-starting control chart for simultaneously monitoring process mean and variance. *International Journal of Production Research*, 48(15):4537–4553.

Liu, L., Chen, B., Zhang, J., and Zi, X. (2014). Adaptive phase ii nonparametric ewma control chart with variable sampling interval. *Quality and Reliability Engineering International*.

Liu, L., Zi, X., Zhang, J., and Wang, Z. (2013). A sequential rank-based nonparametric adaptive ewma control chart. *Communications in Statistics-Simulation and Computation*, 42(4):841–859.

Liu, Y., Zou, C., and Zhang, R. (2008). Empirical likelihood ratio test for a change-point in linear regression model. *Communications in Statistics—Theory and Methods*, 37(16):2551–2563.

Montgomery, D. C. (2008). *Introduction to Statistical Quality Control*. Wiley, Hoboken, N.J, 6 edition edition.

Mood, A. M., Graybill, F., and Boes, D. (1974). Introduction to the theory of. *Statistics*, page 133.

Mukherjee, A. and Bandyopadhyay, U. (2011). Some partially sequential nonparametric tests for detecting linear trend. *Journal of Statistical Planning and Inference*, 141(8):2645–2655.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, pages 523–527.

Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Applied statistics*, pages 126–135.

Pettitt, A. (1980). A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika*, 67(1):79–84.

Quesenberry, C. P. (1991a). SPC q charts for a binomial parameter p: short or long runs. *Journal of Quality Technology*, 23(3):239–246.

Quesenberry, C. P. (1991b). SPC q charts for a poisson parameter $łambda$: short or lung runs. *Journal of quality technology*, 23(4):296–303.

Quesenberry, C. P. (1991c). SPC q charts for start-up processes and short or long runs. *Journal of Quality Technology*, 23(3):213–224.

Rafajłowicz, E., Pawlak, M., and Steland, A. (2010). Nonparametric sequential change-point detection by a vertically trimmed box method. *Information Theory, IEEE Transactions on*, 56(7):3621–3634.

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

Ross, G. J., Tasoulis, D. K., and Adams, N. M. (2011). Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4):379–389.

Sen, P. K. (1977). Tied-down wiener process approximations for aligned rank order processes and some applications. *The Annals of Statistics*, pages 1107–1123.

Sen, P. K. (1980). Asymptotic theory of some tests for a possible change in the regression slope occurring at an unknown time point. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 52(2):203–218.

Shewhart, W. A. (1922). On the measurement of a physical quantity whose magnitude is influenced by primary causes beyond the control of the observer and on the method of determining the relation between two such quantities. *Proceedings of the National Academy of Sciences of the United States of America*, 8(8):248–251. ArticleType: primary_article / Full publication date: Aug. 15, 1922 / Copyright © 1922 National Academy of Sciences.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. D. Van Nostrand Company, Inc.

Sullivan, J. H. and Jones, L. A. (2002a). A self-starting control chart for multivariate individual observations. *Technometrics*, 44(1):24–33.

Sullivan, J. H. and Jones, L. A. (2002b). A self-starting control chart for multivariate individual observations. *Technometrics*, 44(1):24–33.

Wellner, J. A. (1977). A glivenko-cantelli theorem and strong laws of large numbers for functions of order statistics. *The Annals of Statistics*, 5(3):473–480.

Wolfe, D. A. and Schechtman, E. (1984). Nonparametric statistical procedures for the changepoint problem. *Journal of Statistical Planning and Inference*, 9(3):389–396.

Yahoo! (2014). Yahoo finance - business finance, stock market, quotes, news.

Zantek, P. F. (2006). Design of cumulative sum schemes for start-up processes and short runs. *Journal of quality technology*, 38(4):365–375.

Zantek, P. F., Wright†, G. P., and Plante, R. D. (2006). A self-starting procedure for monitoring process quality in multistage manufacturing systems. *IIE Transactions*, 38(4):293–308.

Zhang, C. W., Xie, M., and Jin, T. (2012). An improved self-starting cumulative count of conforming chart for monitoring high-quality processes under group inspection. *International Journal of Production Research*, 50(23):7026–7043.

Zhou, C., Zou, C., Zhang, Y., and Wang, Z. (2009). Nonparametric control chart based on change-point model. *Statistical Papers*, 50(1):13–28.

Zou, C., Zhou, C., Wang, Z., and Tsung, F. (2007a). A self-starting control chart for linear profiles. *Journal of Quality Technology*, 39(4):364–375.

Zou, C., Zhou, C., Wang, Z., and Tsung, F. (2007b). A self-starting control chart for linear profiles. *Journal of Quality Technology*, 39(4):364–375.