UNIVERSITY OF CALIFORNIA
RIVERSIDE

Cumulative Sum Algorithms Based on Nonparametric Kernel Density Estimators

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Tatevik Ambartsoumian

March 2013

Dissertation Committee:

Dr. Daniel R. Jeske, Chairperson
Dr. Keh-Shin Lii
Dr. Tae-Hwy Lee

The Dissertation of Tatevik Ambartsoumian is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my gratitude to my advisor, professor Daniel Jeske, for introducing me to the area of my research, his great enthusiasm, dedicated time and continuous support that led me through the challenging years of my research. I would also like to thank my committee members, professor Keh-Shin Lii and professor Tae-Hwy Lee, for their time and valuable suggestions.

In addition, I would like to express my appreciation of all the faculty in the Department of Statistics, at UCR, whose classes have lent me invaluable statistical knowledge and whose committed research and teaching work have served as a lifelong inspiration. Special thanks are due to Dr. Linda Penas for being a wonderful teaching assistant supervisor and a friend.

I would like to sincerely thank my classmates and good friends, Anne, Ashley, Gabe, Joyce and Rebecca for sharing the ups and downs of the graduate school, for their willingness to help and for the great memories together. With great pleasure, I thank my girl-friends, Anna, Ece, Sherry, Doxa and Christiana, for being my sisterhood and my major support team, for their open arms to give a hug and their open hearts to share happiness and victories. I am also grateful to my godparents, Garbis and Tagoohi, and my cousins, Lusine and Vard, for their warmth, support and caring and for making California feel like home.

Most importantly, I want to thank my parents, Qnar and Azat, for having taught me the indispensable value of hard work and for their endless, unconditional love. Last but not least, I want to thank my brother, Haik and my husband, Yannis, who encouraged, guided and inspired me at every step of this emotional journey and believed in me even when I doubted myself. Without them this work would have not been possible.

To Yannis

ABSTRACT OF THE DISSERTATION

Cumulative Sum Algorithms Based on Nonparametric Kernel Density Estimators

by

Tatevik Ambartsoumian

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, March 2013
Dr. Daniel R. Jeske, Chairperson

Cumulative sum (CUSUM) algorithms are used for monitoring in various applications, including manufacturing, network monitoring, financial markets, biosurveillance, and many more. A popular CUSUM technique for detecting a change in the in-control distribution of an independent data sequence is based on repeated use of the sequential probability ratio test (SPRT). Some optimality results have been derived for the SPRT-based CUSUM when the in-control and out-of-control distributions are fully known. We introduce an approximation formula for the threshold value of an SPRT-based CUSUM. Limited research has been performed on CUSUM techniques when the distributions are not fully specified. This research is concerned about how to use the CUSUM when the underlying in-control distribution is arbitrary and unknown, and the out-of-control density is either an additive or a multiplicative transformation of the in-control density. The proposed solution combines an adaptive nonparametric kernel density estimator derived from an in-control sample of observations with a smoothed bootstrap algorithm that enables the CUSUM to work effectively for reasonably sized sets of in-control data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In many real-life applications important characteristics of an underlying process are monitored over time. Continual surveillance of the process allows practitioners to detect changes in the stochastic structure of the data as fast as possible. Statistical methods designed to detect such changes in the data constitute the area of statistical process control (SPC), whereas the problems in which such statistical monitoring is necessary are generally referred to as change-point problems. Change-point problems and corresponding SPC methods have been studied extensively by multiple authors, including Zacks (1983), Kolmogorov et al. (1990), Montgomery (1991), Basseville and Nikiforov (1993), Yashchin (1993), Lai (1995), Brodsky and Darkhovsky (1993) and Pollak (1995).

Change-point problems can be classified according to multiple different criteria, such as the type of information available about the data (parametric, semi-parametric or nonparametric), the type of change (abrupt vs gradual), the time of change (fixed unknown vs random), the amount of change points (single vs multiple) and more (Brodsky and Darkhovsky (1993)).

In this work, we concentrate on the sequential change-point problem with a single change. The sequential change-point methods within the parametric framework, that is methods which operate under an explicitly specified statistical model, can be broadly classified into the following major schemes: Shewhart control charts, Cumulative Sum (CUSUM) charts, Exponentially-Weighted-Moving Average (EWMA) control charts, and methods based on a Bayesian approach, such as, for example, the Shiryaev-Roberts procedure. In simple terms, the basic mechanism common to all schemes is that at each step of the sequential monitoring, a quantity, often referred to as a tracking statistic is calculated. An alarm is generated by the method, indicating the detection of a change (false or real) if the tracking statistic crosses a predefined control limit. A useful review of the four major control chart schemes, related metrics and various optimality properties normal the normal model is given in Frisen (2003).

Below we briefly describe each of the four major schemes before moving to a more detailed review of the CUSUM scheme and its nonparametric alternatives, which are the main focus of our work. Shewhart charts were first introduced by Shewhart (1931), and remain popular even now due to their simplicity and practicality. A number of authors have studied, discussed and expanded the use of Shewhart charts (see, for example, Montgomery (1991), Woodall and Montgomery (1999)). The key idea, which is also one of the major restrictions of the Shewhart charts is the assumption that the sequence under study is normal with known parameters before and after the change. Another drawback of the Shewharts charts is that they are memoryless, as a result of which they are good at detecting large shifts but have been shown to be far less efficient in detecting small to moderate shifts (see for example Hawkins and Olwell (1998)). A method that has overcome this deficiency of the Shewhart chart is the CUSUM control chart. First introduced by Page (1954, 1955, 1957), the CUSUM has since gained large popularity in the SPC community (some references include Johnson and Leone (1962a,b,c) Bissell (1969), Hawkins and Olwell (1998)). We will be discussing the CUSUM method in more detail in sec-

tion 2.2, for now it suffices to say that efficient CUSUM charts have been designed for normal as well as other known distributional models. Certain optimality properties have been derived for Page's CUSUM by Lorden (1971), Moustakides (1986) and Ritov (1990). Due to their cumulative nature, the CUSUM charts are more efficient than the Shewhart charts in detecting small but persistent changes in a given sequence of observations. The Shiryaev-Roberts method proposed independently by Shiryaev (1963) and Roberts (1966) is often compared to the CUSUM. Both methods are derived from the log-likelihood ratio for testing the sequential hypothesis of change in a given sequence of iid observations. In the Shiryaev-Roberts approach the time of change is assumed to be a random variable and the method is developed by assigning a prior distribution to it. In the design of the CUSUM scheme, the worst change-time is assumed and a minimax solution is employed (Lai (1995), Frisen (2003), Kim et al. (2004)). Just as with the CUSUM method, strong optimality properties have also been derived for the Shiryaev-Roberts procedure in a parametric setting (see, for example Pollak (1995)). The exponentially weighted moving average (EWMA) method was proposed by Roberts (1959). Other authors who have contributed to the further development of EWMA charts include Hunter (1986), Crowder (1989), Montgomery (1991). The EWMA based charts similar to the CUSUM schemes have an advantage over the Shewhart control schemes in situations when shifts of relatively small sizes are to be detected. This is due to the fact that EWMA charts, as oppose to Shewhart charts take into account all observations and not just the last one. The main difference between the EWMA and the CUSUM chart is that the EWMA assigns larger weights to the most recent observations and lesser values (in a geometrically decreasing order) to the earlier observations, whereas the CUSUM gives equal weights to all observations within a variable length window, and zero weight otherwise. Montgomery (1991) discussed certain optimality properties of EWMA charts and provided recommendations for parameter values for increasing the efficiency of an EWMA chart in a specified context.

The main assumption in all of the above scenarios is that the distribution of the monitored process is known both before and after the change. Our objective is to design a robust nonparametric CUSUM procedure for detecting a change in a sequence of independent and identically distributed (iid) observations from an unknown continuous distribution. We assume that a set of in-control data is available from recent past but no other prior information is given about the unknown underlying distribution. The out-of-control state is assumed to be a result from either an additive or a multiplicative change in the process.

1.2 Outline of the Dissertation

In chapter 2, we introduce the most popular scenario discussed in the SPC literature, the so called conventional CUSUM algorithm designed for normal data. We then discuss the theory behind a more general CUSUM formula that can be applied to any distribution. We also explore some well known methods for finding the threshold value of the the conventional CUSUM and propose our own approximation formula for the general case. In chapter 3, we review the existing literature on general nonparametric change-point procedures, with emphasis on nonparametric CUSUM algorithms and propose two new nonparametric CUSUM methods. We discuss the concept of nonparametric density and distribution function estimation and explore relevant methods in the literature. We also introduce several benchmark nonparametric CUSUM procedures. In chapter 4, we conduct an extensive simulation analysis to compare and contrast our nonparametric CUSUMs with the benchmark algorithms. In addition, we propose a sanity test that can guide practitioners to a proper choice of the historical data depth used in our approaches. Finally, in chapter 5, as a future work, we present a real-life example from the area of computer network surveillance, where our proposed nonparametric CUSUMs can be extended for use in detecting faults in a timeslot heterogeneous data framework.

Chapter 2

Cusum for IID Data with Known Distribution

2.1 Conventional CUSUM

Suppose a sequence of in-control iid observations $\{X_i\}$ follows a normal distribution with a known mean μ_0 and a known standard deviation σ. Next suppose that at some unknown time the mean of the observations undergoes a shift of a fixed size, Δ, that is $\mu_1 = \mu_0 + \Delta$, while σ stays the same. Often Δ is measured in multiples of σ, that is $\Delta = \delta\sigma$. One of the standard applications in SPC is to detect the shift in the mean from the *in-control* value μ_0 to the *out-of-control* value μ_1. The "conventional" CUSUM, also known as, "tabular" or "classic" CUSUM that detects this change is defined as

$$\begin{aligned} C_0^+ &= 0 \\ C_i^+ &= \max(0, C_{i-1}^+ + X_i - (\mu_0 + K)). \end{aligned} \tag{2.1}$$

Here the $+$ sign indicates that the mean is shifted upward ($\mu_1 > \mu_0$). Equation (2.1) is referred to as a "one-sided upper" (conventional) CUSUM. Similarly, for the case when $\mu_1 < \mu_0$, that is $\mu_1 = \mu_0 - \Delta$, a "one-sided lower" (conventional) CUSUM is defined as follows

$$\begin{aligned} C_0^- &= 0 \\ C_i^- &= \max(0, C_{i-1}^- + (\mu_0 - K) - X_i). \end{aligned} \tag{2.2}$$

Parameter K in equations (2.1) and (2.2) is called a *reference* parameter and is generally set to $\Delta/2$. In each case, an alarm indicating an out-of-control state is raised when the corresponding CUSUM C^+ or C^- exceeds a predefined threshold. For controlling both an upward and downward shifts in the mean of a normal distribution, a "two-sided" CUSUM is formed by simultaneously running the two one-sided CUSUMs, C_i^+ and C_i^-, and an out-of-control situation is indicated when either one of them crosses the pre-specified threshold value (see, figure 2.1). The idea to detect a change in the mean of normal observations by using the cumulative sum of the deviations of the observations from a reference value belongs to Page (1954). Other useful references include Montgomery (1991) and Hawkins (1992).

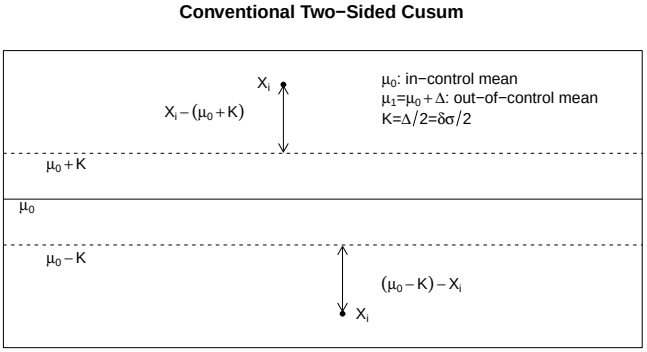

Figure 2.1: Two-Sided Conventional CUSUM with reference value K.

2.2 Optimal CUSUM

As we mentioned in section 1.1 the CUSUM control chart introduced by Page (1954) is a popular method for detecting small but persistent changes in the distribution of sequentially monitored iid observations $\{X_i\}_{i=1}^{\infty}$ with a known density function f_0. In simple terms the problem can be formulated as follows, at some unknown moment in time the distribution of the observations may change from f_0, referred to as the "in-control" density to an "out-of control" state, characterized by another known distribution, f_1. It is important to detect the transition from f_0 to f_1 as quickly as possible after it has occurred. It has been first observed by Page, and later repeatedly emphasized in the literature (Van Dobben De Bruyn (1968), Lorden (1971), Moustakides (1986), Hawkins (1992)) that Page's CUSUM can be viewed as a sequence of Sequential Probability Ratio Tests (SPRT)s each based on a score statistic, given by $Z_i = \log\left[f_1(X_i)/f_0(X_i)\right]$. For a detailed discussion of the SPRT one can refer to Wald (1947), Siegmund (1985). In what follows, we give a brief outline of the SPRT setup relevant to our problem. In the classical SPRT formulation, a simple null hypothesis H_0 is being tested against a simple alternative hypothesis, H_1. The test statistic is based on a log-likelihood ratio, given by $\log\Lambda_i = \sum_{j=1}^{i} Z_j$. The test results in three possible outcomes:

$$\text{Stop sampling and accept } H_0 \text{ if } \log\Lambda_i \leq \log A$$

$$\text{Stop sampling and accept } H_1 \text{ if } \log\Lambda_i \geq \log B$$

$$\text{Continue sampling if } \log A < \log\Lambda_i < \log B,$$

where A and B are predefined constans, determined from controlling the type I and type II errors. If we denote type I and type II errors by α and β correspondingly, the following approximations are available $A \approx \beta/(1-\alpha)$ and $B \approx (1-\beta)/\alpha$.

The SPRT is known for the following powerful optimality property; among all sequential tests with the same fixed α and β, it has the smallest expected number of samples/observations needed to make a decision (Wald (1947), Ferguson (1967)). As opposed to Wald's SPRT, in Page's CUSUM scheme, one never accepts H_0 when the evidence favors it, i.e. when the cumulative sum, $\sum_{j=1}^{i} Z_j < 0$; instead, when this happens the sum is reflected to 0, and the test restarts itself. Sampling continues until there is enough evidence to reject H_0. More formally, Page's CUSUM formula can be motivated through a log-likelihood ratio test approach as follows. In order to detect a potential shift of the distribution of $\{X_i\}_{i=1}^{\infty}$ from f_0 to f_1, at each time point i, we test the following hypothesis:

$$
\begin{aligned}
H_0: &\quad X_1, ..., X_i \sim f_0 \\
H_1: &\quad X_1, ..., X_{k-1} \sim f_0 \\
&\quad X_k, X_{k+1}, ..., X_i \sim f_1 \text{ for some k} \in [1, ..., i]
\end{aligned}
$$

Thus under the null hypothesis all observations follow f_0, implying no change in the distribution, whereas under the alternative hypothesis, a change occurs at time point k. We have

$$
\begin{aligned}
k = 1: &\quad X_1, ..., X_n \sim f_1 \\
k = 2: &\quad X_1 \sim f_0, \quad X_2, ..., X_i \sim f_1 \\
&\quad ... \\
k = i: &\quad X_1, ..., X_{i-1} \sim f_0, \quad X_i \sim f_1
\end{aligned}
$$

We reject H_0 and conclude that there has been a change, if L_0/L_1 is too small, where L_0 and L_1 are correspondingly the null and the general likelihood functions of our test, as shown below:

$$L_0 = \prod_{j=1}^{i} f_0(X_j)$$

$$L_1 = \max_{1\le k\le i+1}\left[\prod_{j=1}^{k-1} f_0(X_j)\prod_{j=k}^{i} f_1(X_j)\right],$$

where $\prod_{j=1}^{0} f_0(X_j) = 1$ and $\prod_{j=i+1}^{i} f_1(X_j) = 1$.

Then the likelihood ratio and the log-likelihood ratio statistics are given by

$$\frac{L_0}{L_1} = \frac{\prod_{j=1}^{i} f_0(X_j)}{\max_{1\le k\le i+1}\left[\prod_{j=1}^{k-1} f_0(X_j)\prod_{j=k}^{i} f_1(X_j)\right]}$$

$$= \frac{\prod_{j=1}^{i}\frac{f_0(X_j)}{f_1(X_j)}}{\max_{1\le k\le i+1}\left[\prod_{j=1}^{k-1}\frac{f_0(X_j)}{f_1(X_j)}\right]} = \frac{\prod_{j=1}^{i}\frac{f_0(X_j)}{f_1(X_j)}}{\max_{0\le k\le i}\left[\prod_{j=1}^{k}\frac{f_0(X_j)}{f_1(X_j)}\right]} \quad \left(s.t. \quad \prod_{j=1}^{0}\frac{f_0(X_j)}{f_1(X_j)} = 1\right)$$

and

$$\log\frac{L_0}{L_1} = \sum_{j=1}^{i}\log\frac{f_0(X_j)}{f_1(X_j)} - \log\max_{0\le k\le i}\left[\prod_{j=1}^{k}\frac{f_0(X_j)}{f_1(X_j)}\right]$$

$$= \sum_{j=1}^{i}\log\frac{f_0(X_j)}{f_1(X_j)} - \max_{0\le k\le i}\left[\sum_{j=1}^{k}\log\frac{f_0(X_j)}{f_1(X_j)}\right] \quad \left(\text{s.t.}\sum_{j=1}^{0}\log\frac{f_0(x_j)}{f_1(x_j)} = 0\right)$$

$$= -\sum_{j=1}^{i}\log\frac{f_1(X_j)}{f_0(X_j)} + \min_{1\le k\le i}\left[0, \sum_{j=1}^{k}\log\frac{f_1(X_j)}{f_0(X_j)}\right].$$

We reject H_0 if L_0/L_1 is too small or, equivalently, if -$\log L_0/L_1$ is too big. More formally, H_0 is rejected and an alarm indicating a change in the distribution is raised when

$$\sum_{j=1}^{i} \log \frac{f_1(X_j)}{f_0(X_j)} - \min_{1 \le k \le i} \left[0, \sum_{j=1}^{k} \log \frac{f_1(X_j)}{f_0(X_j)} \right] > H \tag{2.3}$$

for some predefined threshold value, H. A detailed discussion of available methods for determining H is given in section 2.4.

The CUSUM schemes were extensively studied the following years by multiple authors, including Van Dobben De Bruyn (1968), Kemp (1962), Johnson and Leone (1962a,b,c) and Bissell (1969). Lorden (1971) introduced a problem of "optimal stopping" in the context of a change-point problem and showed that Page's CUSUM (2.3) provided an asymptotically optimal solution to the problem based on a minimax criteria. Later, Moustakides (1986) generalized Lorden's result by proving an optimality of Page's CUSUM. Following the simple interpretation by Hawkins and Olwell (1998) of both results by Lorden and Moustakides, Page's CUSUM is optimal in the sense that it is the fastest procedure in detecting a change among the methods that have the same expected time until a false alarm.

The following recursive and computationally more suitable form of the above CUSUM scheme was also introduced by Page (1954), and has been commonly used till now:

$$\begin{aligned} S_i &= \max \left(0, S_{i-1} + \log \frac{f_1(X_i)}{f_0(X_i)} \right) \\ S_0 &= 0. \end{aligned} \tag{2.4}$$

As before, the CUSUM signals when $S_i > H$ for a predefined value of the threshold.

An induction proof of the equivalence between (2.3) and (2.4) has been shown in Montes De Oca (2008). For convenience, we also outline it below.

Proposition 1. *The two CUSUMs given by (2.3) and (2.4) are equivalent.*

Proof. Let $Z_i = \log(f_1(X_i)/f_0(X_i))$, then we have

$$L_i = \sum_{j=1}^{i} Z_j - \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j)$$

$$S_i = \max(0, S_{i-1} + Z_i), \quad S_0 = 0$$

We show that $S_i = L_i$ through induction. For $i = 1$, we have

$$S_1 = \max(0, Z_1) = \begin{cases} Z_1, & \text{if } Z_1 \ge 0, \\ 0, & \text{otherwise} \end{cases}$$

$$L_1 = Z_1 - \min(0, Z_1) = \begin{cases} Z_1, & \text{if } Z_1 \ge 0, \\ 0, & \text{otherwise} \end{cases}$$

Thus $S_1 = L_1$. Assume $S_i = L_i$ for $i \ge 1$. Show that $S_{i+1} = L_{i+1}$. For $i + 1$ we have

$$\begin{aligned} S_{i+1} &= \max(0, S_i + Z_{i+1}) \\ &= \max(0, L_i + Z_{i+1}) \\ &= \max\left(0, \sum_{j=1}^{i} Z_j - \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j) + Z_{i+1}\right) \\ &= \max\left(0, \sum_{j=1}^{i+1} Z_j - \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j)\right) \\ &= \begin{cases} \sum_{j=1}^{i+1} Z_j - \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j), & \text{if } \sum_{j=1}^{i+1} Z_j - \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j) \ge 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

$$
\begin{aligned}
L_{i+1} &= \sum_{j=1}^{i+1} Z_j - \min_{1\le k\le i+1}(0, \sum_{j=1}^{k} Z_j) \\
&= \sum_{j=1}^{i+1} Z_j - \min\left(\min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j), \sum_{j=1}^{i+1} Z_j\right) \\
&= \begin{cases} \sum_{j=1}^{i+1} Z_j - \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j), & \text{if } \sum_{j=1}^{i+1} Z_j \ge \min_{1\le k\le i}(0, \sum_{j=1}^{k} Z_j) \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

Hence $S_{i+1} = L_{i+1}$. □

2.3 Optimal CUSUM for a One-Parameter Exponential Family of Distributions

For the special case, when both f_0 and f_1 belong to a one-parameter exponential family of distributions with corresponding in-control and out-control parameters, θ_0 and θ_1, Hawkins and Olwell (1998) observed that CUSUM (2.4) conveniently depends only on the minimal sufficient statistic of the distribution and an appropriately defined reference value. Recall, that a density function belongs to a one-parameter exponential family of distributions if and only if it can be represented in the form $f(\mathbf{x}, \theta) = r(\mathbf{x}) \exp\{T(\mathbf{x})Q(\theta) + \beta(\theta)\}$ with $\mathbf{x} = (x_1, ..., x_n)$, where $r(\mathbf{x})$ and the minimal sufficient statistic $T(\mathbf{x})$ are real-valued functions on $\mathcal{R}_n$ (Rohatgi and Saleh (2001)). It is now easy to see that within this family, the increment $Z_i = \log[f_1(X_i)/f_0(X_i)]$ can be expanded into $Z_i = T(X_i)[Q(\theta_1) - Q(\theta_0)] + \beta(\theta_1) - \beta(\theta_0)$ and CUSUM (2.4) becomes

$$S_i = \max\{0, S_{i-1} + T(X_i)(Q(\theta_1) - Q(\theta_0)) + \beta(\theta_1) - \beta(\theta_0)\}$$

$$S_0 = 0.$$

Suppose $Q(\theta_1) - Q(\theta_0) > 0$, then dividing both sides of the equation by $Q(\theta_1) - Q(\theta_0)$ and defining $k = -(\beta(\theta_1) - \beta(\theta_0)/(Q(\theta_1) - Q(\theta_0))$, we derive the following simplified upper CUSUM formula

$$C_i^+ = \max\left\{0, C_{i-1}^+ + T(X_i) - k\right\} \tag{2.5}$$
$$C_0^+ = 0,$$

where $C_i^+ = S_i/[Q(\theta_1) - Q(\theta_0)]$, and k is the corresponding reference parameter. The $(+)$ sign indicates that the change is a result of an increase in the natural parameter value $Q(\theta)$. The procedure stops and alarms of the change when $C_i^+ > H^+$.

By a similar reasoning, it is easy to see that when $Q(\theta_1) - Q(\theta_0) < 0$, that is when the change is a result of a decrease in the natural parameter value, the lower CUSUM is of the form given below

$$C_i^- = \min\left\{0, C_{i-1}^- + T(X_i) - k\right\},$$
$$C_0^- = 0$$

with a stopping rule of $C_i^- < -H^-$. Often, the lower CUSUM is written with a reversed sign and using $\max$ to be consistent with the upper CUSUM scheme, that is

$$C_i^- = \max\left\{0, C_{i-1}^- + k - T(X_i)\right\}. \tag{2.6}$$
$$C_0^- = 0.$$

The stopping rule accordingly is $C_i^- > H^-$.

The conventional one-sided CUSUM schemes for normal iid variables defined in (2.1) and (2.2) can be easily motivated using (2.5) and (2.6) by rewriting the normal density as a function of μ_0

$$f(x, \sigma_0^2, \mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\{-\frac{x^2}{2\sigma_0^2} + x\frac{\mu}{\sigma_0^2} - \frac{\mu^2}{2\sigma_0^2}\},$$

and identifying the quantities $T(X) = X, Q(\mu_0) = \mu_0/\sigma^2, r(X) = 1/\sqrt{2\pi\sigma^2}\exp\{-X^2/(2\sigma^2)\})$ and $\beta(\mu_0) = -\mu_0^2/(2\sigma^2)$.

2.4 Determining H

2.4.1 Average Run Length (ARL)

A conventional way to determine the threshold value, H, in a CUSUM scheme (2.4) is to control its average run length (ARL) when the process is in control. The run length (RL) of the CUSUM is the number of observations that need to be accumulated before the CUSUM crosses H, provided it started from 0. There are cases, when it is more reasonable to start the CUSUM from a value greater than 0. This type of CUSUM is called a fast initial response (FIR) CUSUM, and was first suggested by Lucas and Crosier (1982). It is particularly beneficial to use FIR CUSUMs in cases when the monitoring begins after a change in the process has already occurred. Then the "head start" provided by the non-zero initial value of the CUSUM allows to detect the change faster than a CUSUM with a zero start. Details about the recommended head start values and the ARL properties of the FIR CUSUMs can be found in Hawkins (1992).

The RL of any CUSUM is a random variable itself, and although it is not always straightforward to find the distribution of the RL, it is often possible to calculate its descriptive statistics such as the mean, standard deviation and the quantiles. Multiple authors reported a high variability of the RL distribution (see, for example Woodall (1983), Hawkins (1992)).

Despite this obvious drawback, the ARL still serves as a helpful characteristic of the performance of the CUSUM and is traditionally used in quality control literature. It is customary to denote ARL_0 as the value of ARL when the process is in-control, and ARL_1 when the process is out-of-control.Ideally, if the process is in control, there should be no alarms at all, however due to natural variations in the process, the CUSUM of in-control observations may become large enough to cross the given threshold, producing a false alarm. Thus, when designing a CUSUM, the threshold H is chosen so that it guarantees a tolerable amount of false alarms in a given period of time, controlled by a fixed value of ARL_0. On the other hand, when a change does indeed occur, we want the CUSUM to detect is as quickly as possible, that is we want ARL_1 to be as small as possible for a given value of ARL_0.

2.4.2 Existing Methods to Determine H for the Conventional CUSUM

There is a direct dependence between the threshold value, the reference parameter and the ARL_0 of the conventional CUSUM (2.1). Thus, if any two of the three parameters are known, the other one is automatically determined (conditional on the underlying distribution). In particular, this interdependence allows the practitioners to find H for a fixed value of ARL_0. Below we outline several fundamental techniques existing in the literature for calculating the ARL for a given value of H in the context of the conventional CUSUM.

2.4.2.1 Integral Equations

As was mentioned earlier, one possible way to view the CUSUM is by treating it as a sequence of SPRT's. Therefore it is natural to expect that the ARL of a CUSUM could be determined by using the corresponding theory developed for the SPRTs. A characteristic, analogous to the ARL in SPRT is the average sample number (ASN), which is the number of observations sampled in a single test till a decision to accept H_0 or H_1 is made. The ASN for a Wald test

with given boundaries can be found directly by solving an integral equation (see, for example Page (1954), Goel and Rigdon (1991)). The ARL of a CUSUM correspondingly is determined by extending this integral equation for a single sequential test to a sequence of such tests. Major references include Page (1954), Van Dobben De Bruyn (1968), Montgomery (1991), Hawkins and Olwell (1998). Citing Van Dobben De Bruyn's interpretation of this integral equation, the ARL of a test, starting from point z is a sum of three components: the first being 1 (since one observation is always sampled before a decision is made), the second being equal to the product of the probability that the next observation will bring the CUSUM back to 0 and the ARL of the CUSUM with a starting value 0, and the final summand is the integral over the probabilities that the next observation will result in the CUSUM's value anywhere between 0 and H multiplied by the corresponding ARL. For a few families of distributions this integral equation can be solved analytically, which was demonstrated Van Dobben De Bruyn (1968). In most of the cases however closed form expressions are impossible to obtain and one needs to resort to numerical approximations. It is worth mentioning that this method allows to approximate the ARL of a CUSUM that starts from any state $i \in \{0, ..., H\}$ and so it is applicable to the FIR type schemes as well.

2.4.2.2 Markov Chain Approach

The idea of viewing the CUSUM statistic as a Markov process and using it for finding the ARL was briefly mentioned by Page (1954) as early as in 1954. However the approach gained popularity and practical value after the in-depth study by Brook and Evans (1972), who are considered to be the pioneers of the Markov chain approximation to the CUSUM. A simple and detailed interpretation of the approach can be found in Montgomery (1991). The idea is as follows. If we consider S_i to be a random variable indexed by n and treat the values it takes in the interval $(0, H)$ as states (with an initial state $S_0 = 0$ and an absorbing state corresponding

to $S_i \geq H$), then it is easy to observe that S_i has a Markov property. That is the conditional distribution of S_{i+1} given the previous i states $S_0, S_1, ...S_i$, depends only on the ith state, S_i. Regardless of whether the underlying distribution of the monitored variable is discrete or continuous, S_i is a discrete-time Markov process according to the integer index i. Moreover, if the variable being monitored is discrete, we get a discrete-state Markov process. If S_i takes continuous values on $(0, H)$, then a proper discretization of the underlying continuous distribution allows to still treat the CUSUM as discrete-state Markov process and apply the corresponding theory. The ARL can then be approximated by solving a system of equations involving the matrix of transition probabilities for the defined states. Similar to the integral equations, the equations based on the Markov-process approximation allow to calculate the ARL value starting from any point z, with z not necessarily equal to 0, which makes them practically suitable for the FIR schemes. Hawkins (1992) used the Markov chain approach to develop an analogous equation for the ARL, using more accurate approximations to the transition probabilities than those offered by Brook and Evans. His formula allows to calculate the ARL within 3% of the nominal value. Champ and Rigdon (1991) provide an in-depth comparison of the Markov Chain Approach and the Integral Equations for finding the ARL of a CUSUM.

2.4.3 Proposed Method to Determine H for the Optimal CUSUM

In this section, we propose an analytic approximation formula relating threshold H to the ARL of the optimal CUSUM scheme (2.4). A motivation for our formula has been an analogous result derived by Kim et al. (2007) who used the conventional CUSUM formulas (2.1), (2.2) in an autocorrelated data context. More specifically, for $S_i = \sum_{j=1}^{i} Z_j - \min_{0 \leq k \leq i} \left(\sum_{j=1}^{k} Z_j \right)$ with $Z_j = \log \left[f_1(X_j) / f_0(X_j) \right]$, we show that the formula in proposition 2 below holds for a general context of formula (2.4) with iid observations.

Proposition 2.

$$ARL_0 \approx \frac{\sigma_Z^2}{2d_Z^2}\left[\exp\left(-\frac{2d_Z(H+1.166\sigma_Z)}{\sigma_Z^2}\right)-1+\frac{2d_Z(H+1.166\sigma_Z)}{\sigma_Z^2}\right], \tag{2.7}$$

where $d_Z = E[Z]$ *and* $\sigma_Z^2 = Var[Z]$.

Proof. We start with introducing several relevant concepts from Kim's work. A *stopping time* for CUSUM S_i is defined as

$$T_Z = \min_{i=1,2,\ldots}\{i : S_i \geq H\}.$$

A *standardized* CUSUM, $C_i(t)$ for Z_i is defined as follows

$$C_i(t) = \frac{\sum_{j=1}^{\lfloor it \rfloor} Z_j - itd_Z}{\sigma_Z\sqrt{i}} \quad \text{for } t \in [0,1], \tag{2.8}$$

where $\lfloor q \rfloor$ is the greatest integer not exceeding q. Moreover, for each positive i, the random function $C_i(t)$ is a right-continuous function on $[0,1]$ that has left-hand limits. It is assumed that $\{Z_j\}_{j=1}^{\infty}$ satisfy the functional central limit theorem (FCLT) on $[0,1]$ (Billingsley (1968)), according to which

$$C_i(\cdot) \xrightarrow{d} W(\cdot) \quad \text{as i} \to \infty. \tag{2.9}$$

Here $\xrightarrow{d}$ denotes convergence in distribution, $\{W(t)\}: t \in [0,\infty)\}$ is a standard Brownian motion process such that for any $s, t \in [0,\infty)$, the random variables $W(s)$ and $W(t)$ are jointly normal with $E[W(s)] = E[W(t)] = 0$ and $Cov[W(s), W(t)] = \min\{s,t\}$.

A Brownian motion on $[0,\infty)$ with drift parameter d_Z and variance parameter σ_Z^2 is defined as

$$B(t) = d_Z t + \sigma_Z W(t) \quad \text{for } t \in [0,\infty) \tag{2.10}$$

so that for all $t \geq 0$, $E[B(t)] = d_Z t$ and $Var[B(t)] = \sigma_Z^2 t$.

Using equations (2.8), (2.9) and (2.10), we get

$$\begin{aligned} S_i &= \sum_{j=1}^{i} Z_j - \min_{0\le k\le i}\left(\sum_{j=1}^{k} Z_j\right) \\ &\overset{d}{\approx} C_i(1)\sigma_Z\sqrt{i} + iE(Z_j) - \inf_{0\le t\le 1}\{C_i(t)\sigma_Z\sqrt{i} + itE(Z_j)\} \\ &\overset{d}{\approx} W(1)\sigma_Z\sqrt{i} + id_Z - \inf_{0\le t\le 1}\{W(t)\sigma_Z\sqrt{i} + itd_Z\} \\ &\overset{d}{=} W(i)\sigma_Z + id_Z - \inf_{0\le u\le i}\{W(u)\sigma_Z + ud_Z\} \\ &= B(i) - \inf_{0\le u\le i}\{B(u)\}, \end{aligned} \tag{2.11}$$

where $\overset{d}{\approx}$ and $\overset{d}{=}$ indicate asymptotic and exact equality in distribution.

Let $\tilde{S}(t)=B(t)-\inf_{0\le u\le t}\{B(u)\}, t \ge 0$. The stopping time for $\tilde{S}_t$ is given by $T_{\tilde{S}} = \inf_{t\ge 0}\{t : \tilde{S}(t) \ge H\}$.

Using (2.11), Kim et al. (2005) showed that $T_Z \overset{d}{\approx} T_{\tilde{S}}$, and thus using results in Bagshaw and Johnson (1975) as well as in Siegmund (1985), we get

$E[T_Z] \approx E[T_{\tilde{S}}] =$

$$= \begin{cases} \left(\dfrac{H+1.166\sigma_Z}{\sigma_Z}\right)^2, & \text{if } d_Z = 0 \\ \dfrac{\sigma_Z^2}{2d_Z^2}\left[\exp\left(-\dfrac{2d_Z(H+1.166\sigma_Z)}{\sigma_Z^2}\right) - 1 + \dfrac{2d_Z(H+1.166\sigma_Z)}{\sigma_Z^2}\right], & \text{if } d_Z \ne 0 \end{cases}$$

When the process is in-control $d_Z \neq 0$, and therefore $ARL_0 = E[T_Z]$ is given by the bottom sub-equation above. □

For fixed d_Z, σ_Z, and ARL_0 one can find H from formula (2.7) through any available search algorithm. Below we show some results from applying (2.7) to calculate H by controlling ARL_0 at 200. For illustration, we consider gamma and Weibull distributions as in-control densities, and let the out-of-control distribution be a multiplicative change, that is $f_1(x) = f_0(x/c)/c$ for $c > 0$ $(c \neq 1)$. Note that for both gamma and Weibull densities, we can derive closed forms of d_Z and σ_Z^2 under H_0. For cases when these quantities can not be

determined analytically, numerical integration can be used. Suppose $\{X_i\}_{i=1}^{\infty}$ are iid observations from $Gamma(\alpha, \beta)$ with density function $f_0(x) = x^{\alpha-1}e^{-x/\beta}/(\Gamma(\alpha)\beta^{\alpha})$ with $\alpha, \beta > 0$. Then $Z = -\alpha \log c - (1/c - 1)(X/\beta)$, and under H_0, $d_Z = -\alpha(\log c - (1/c - 1)) < 0$ and $\sigma_Z^2 = \alpha(1/c - 1)^2 > 0$. Similarly, for iid observations $\{X_i\}_{i=1}^{\infty}$ from $Weibull(\alpha, \beta)$ with density $f_0(x) = x^{\alpha-1}e^{-x^{\alpha}/\beta}(\alpha/\beta)$ with $\alpha, \beta > 0$, we have $Z = -\alpha \log c - (1/c^{\alpha} - 1)(X/\beta)^{\alpha}$, and correspondingly $d_Z = -(\log c^{\alpha} - (1/c^{\alpha} - 1)) < 0$ and $\sigma_Z^2 = (1 - 1/c^{\alpha})^2 > 0$. For each of the two distributions under consideration, we vary $c \in \{0.9, 0.95, 0.975, 1.025, 1.05, 1.1\}$, find H through a search method using (2.7), then use this value, say H_{appr}, to evaluate the ARL_0 of CUSUM (2.4). For comparison, we also report a value of H calculated from Monte-Carlo simulations, say H_{mc}, that guarantees ARL_0 of 200 for each value of c. Note that d_Z and σ_Z and hence H_{appr} do not depend on the scale parameter β for either gamma or Weibull densities.

Our Monte-Carlo simulation analysis showed that for fixed values of α and c, the value of H_{mc} also does not depend on the scale parameter β for any of the two distributions, therefore without loss of generality, we set $\beta = 1$. Below, we report results for $Weibull(1, 1)$, $Weibull(3, 1)$ and $Gamma(3, 1)$ distributions. The plots of the corresponding densities are displayed in figure 2.2.

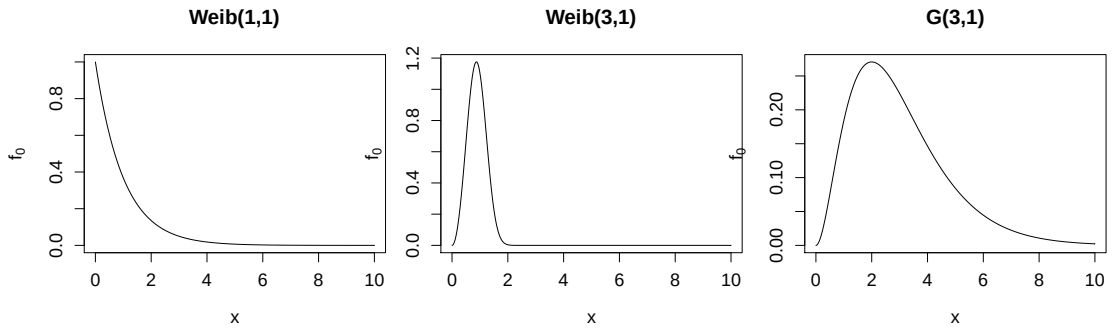

Figure 2.2: Densities of $Weibull(1, 1)$, $Weibull(3, 1)$ and $Gamma(3, 1)$ distributions.

Model		c					
		0.9	0.95	0.975	1.025	1.05	1.1
$Gamma(3,1)$	H_{appr}	1.70	0.96	0.52	0.49	0.88	1.42
	H_{mc}	1.62	0.94	0.51	0.49	0.88	1.47
	$\text{ARL}_{0_{appr}}$	223. 46	210.89	206.94	199.92	197.13	183.41
	$\text{ARL}_{0_{mc}}$	199.12	198.72	201.89	199.17	202.03	201.11
$Weibull(1,1)$	H_{appr}	1.14	0.60	0.31	0.3	0.55	0.94
	H_{mc}	1.08	0.59	0.31	0.3	0.55	0.97
	$\text{ARL}_{0_{appr}}$	228.37	206.85	201.06	203.49	199.47	191.41
	$\text{ARL}_{0_{mc}}$	199.15	201.07	200.78	202.17	200.21	198.61
$Weibull(3,1)$	H_{appr}	2.70	1.55	0.86	0.78	1.27	1.81
	H_{mc}	2.28	1.43	0.82	0.78	1.31	2.00
	$\text{ARL}_{0_{appr}}$	324.97	239.83	215.14	198.52	182.23	156.89
	$\text{ARL}_{0_{mc}}$	198.87	202.74	201.13	199.06	197.94	202.33

Table 2.1: Evaluation of ARL_0 using H_{appr}, for a multiplicative change alternative. Nominal $\text{ARL}_0 = 200$.

From table 2.1 we see that under all three scenarios, for small multiplicative changes $c = 0.975$ and $c = 1.025$, H_{appr} is almost equal to H_{mc}, and hence $\text{ARL}_{0_{appr}}$ is close to $\text{ARL}_{0_{mc}}$. As c moves away from 1 in any direction, H_{appr} and H_{mc} start diverging from each other resulting in discrepancy between $\text{ARL}_{0_{appr}}$ and $\text{ARL}_{0_{mc}}$. Note that in the Weibull example for fixed β, the difference between H_{appr} and H_{mc} (and correspondingly between $\text{ARL}_{0_{appr}}$ and $\text{ARL}_{0_{mc}}$) is more pronounced for $\alpha = 3$ compared to $\alpha = 1$, which can be explained by a higher variability of the RL distribution when $\alpha = 3$ (which in turn can be attributed to the fact that $Weibull(3,1)$ is more skewed than $Weibull(1,1)$). The Monte-Carlo error based on $10,000$ iterations varied between 1.67 and 2.78.

Based on the above, our recommendation is to use the analytical approximation formula (2.7) to get an adequate choice for a starting threshold value of the optimal CUSUM (2.4), and try to subsequently fine tune it depending on the application.

Chapter 3

Cusum for IID Data with Unknown Distribution

3.1 Literature Review

In section 2.2, we presented Page's CUSUM (2.4) and cited its optimality under the assumption that the underlying distributions of the monitored process before and after the change are known. In this section we will be discussing a more realistic setup when the densities are unknown, under the assumption that there is a finite amount of historical data available. Historical data is a sample of observations that has been accumulated from recent past during the normal operating conditions of the process. It is therefore free from anomalous events and thus can be safely used to represent the in-control state of the process.

Prior to describing our two new nonparametric CUSUM approaches, we present a review of the current literature on nonparametric change-point methods with emphasis on nonparametric CUSUM procedures. Brodsky and Darkhovsky (1993) provide a thorough overview of a number of existing nonparametric change-point procedures. Frisen (2003) is a good source

of references on distribution-free methods in general, including the paper on nonparametric EWMA method by Hackl and Ledolter (1991), and the paper on modified Shewhart and CUSUM charts by Liu (1995). In a series of papers by Gordon and Pollak (1994, 1995, 1997), the authors derive nonparametric algorithms based on the Shiryaev-Roberts method by replacing the observations with their corresponding signed ranks. Another comprehensive survey aimed toward univariate data monitoring is given by Chakraborti et al. (2001). The paper concentrates on the nonparametric equivalents of the "most common quality control charts", which include the Shewhart, the CUSUM and the EWMA charts. Nonparametric or distribution-free charts are formally defined by the authors as charts, whose in-control RL distribution is the same for every continuous distribution. Among the general strengths of the distribution-free methods highlighted in the paper are the flexibility of their usage with any continuous distribution, the simplicity of comparing different nonparametric charts on the basis of their RL and their improved performance over their parametric counterparts in the cases when the underlying distribution is non-normal or skewed or has heavy tails. In addition, Chakraborti et al. (2001) comment on papers emphasizing the necessity of developing and using nonparametric charts in applied scenarios.

Existing nonparametric CUSUM charts in the literature include the signed-rank based CUSUM, proposed by Bakir and Reynolds (1979) and a related sign statistic based CUSUM, suggested by Amin et al. (1995). Each of the methods is designed for detecting a shift in the location parameter of a process, and the corresponding tracking statistics are calculated on batches of observations. Batch processing is not well suited for our context, where monitoring individual observations is desired in order to keep detection delays as small as possible. A rank-based CUSUM chart, which is based on individual observations, is proposed by McDonald (1990). The author derives a distribution-free tracking statistic by calculating the sequential ranks of the incoming observations. McDonald's approach fits in our setting more naturally, and therefore

we will be considering it in more detail in our comparative analysis in section 3.5. In other related work, Yashchin (1992) and Chatterjee and Qiu (2009) use bootstrap procedures with historical data to establish CUSUM charts for arbitrary contexts. The CUSUM designs proposed in these two papers use the conventional CUSUM tracking statistic (2.1), whose form as we discussed in section 2.3 is only optimal in the context of normally distributed observations. Yashchin (1992) uses the bootstrap procedures to estimate the properties of this tracking statistic in a general setting, and then suitably adjusts the decision threshold. Chatterjee and Qiu (2009), on the other hand, try to replace the use of a single decision threshold with a sequence of control limits that are estimated from the bootstrap techniques.

In this dissertation we propose two nonparametric CUSUM methods, the nonparametric density estimation based CUSUM (NDEC) and the probability integral transformation based CUSUM (PITC). These two procedures differ from Yashchin (1992) and Chatterjee and Qiu (2009) in that our methods define tracking statistics that inherently adapt to the general context they need to operate in. Alternatively said, our strategy is not to adjust the threshold of a conventional CUSUM to fix it for general contexts, but instead to directly incorporate the context in the definition of the tracking statistic. In particular, the NDEC utilizes an estimate of the optimal CUSUM tracking statistic (2.4) and a bootstrap procedure to obtain a suitable decision threshold. The PITC first transforms the data using an estimate of the in-control cumulative distribution function, and then uses an estimate of an approximately optimal tracking statistic. Jeske et al. (2009) and Montes de Oca et al. (2010) also considered tracking statistics that automatically adapt to arbitrary contexts, and we include comparison with their methods in section 3.5.

3.2 Preliminaries

We develop our methodology within a framework of periodic cycles of data of finite length. A similar framework is discussed in Bhattacharya and Frierson (1981), where the authors describe an application in SPC in which, for the purpose of preventive maintenance, the process is periodically tuned after a fixed number of readings are taken, even if no anomaly has been observed. In this context, our major performance metric for each algorithm is the false alarm rate (FAR), the rate at which the algorithm erroneously alarms of a change when the process is in-control. In our comparative analysis, we estimate the FAR as the proportion of periodic cycles that falsely alarm when the process is in control. We measure the power of each algorithm in terms of the change detection rate or, equivalently, the true alarm rate (TAR), which we estimate by the proportion of cycles that alarm when the process is out-of-control. Along with TAR, we also keep track of the so called, detection delay (DD) and the average detection delay (ADD) characteristics. Given that the change has been detected, the DD is the number of observations collected from the moment the change has occurred till the time of the signal, the ADD is the average of the DD.

We assume that for each monitoring cycle of length n, denoted by $X_1, ..., X_n$, there are s cycles (each of length n) of historical observations available, denoted by $Y_1, ..., Y_N$, where $N = ns$. A sliding window mechanism can be used to update the historical data before each new cycle. This is done in two steps. First, when the end of the cycle is reached, a screening algorithm is run through the data to remove anomalous events, if there were any, to obtain up to n new in-control observations. The new observations are then added to the previously obtained in-control data, and at the same time an equal number of older observations are discarded. A detailed description of a rolling window mechanism and an implementation of a screening algorithm can be found in Montes de Oca et al. (2010).

3.3 Proposed Nonparametric CUSUM Algorithms

3.3.1 Nonparametric Density Estimation Based CUSUM (NDEC)

Our idea is to employ a properly chosen NDE method to estimate the true, unknown in-control density, f_0 based on the historical data. This estimate will subsequently replace f_0 in the optimal CUSUM formula (2.4). We make a reasonable assumption that the change results in either an additive or a multiplicative transformation of the in-control density. This allows us to simultaneously obtain the estimate of f_1 by estimating f_0. If we denote the nonparametric density estimates of f_0 and f_1 by $\hat{f}_0$ and $\hat{f}_1$, correspondingly, our proposed nonparametric extension of (2.4) is given by

$$\hat{S}_0 = 0 \quad \hat{S}_i = \max\left\{0, \hat{S}_{i-1} + \log\left[\frac{\hat{f}_1(X_i)}{\hat{f}_0(X_i)}\right]\right\}. \tag{3.1}$$

We determine H through bootstrapping, aiming to maintain the FAR during a monitoring cycle at a fixed level, γ, as outlined in algorithm 1 below. The smoothed bootstrap algorithm used at step 2 of algorithm 1 is tightly linked to the NDE method used to estimate $\hat{f}_0$ and will be discussed in section 3.4.2. Note that in our scheme H is dependent on the historical sample and it would need to be updated if the historical dataset is updated according to the s-cycle sliding window technique, mentioned earlier.

Algorithm 1 : Determining Threshold H in the NDEC.

1. Use $\{Y_i\}_{i=1}^{N}$ to estimate the in-control density, $\hat{f}_0$, and correspondingly, $\hat{f}_1$.
2. Simulate B_1 cycles of in-control observations from $\hat{f}_0$ via *smoothed bootstrap*.
3. Run the NDEC (3.1) along each simulated cycle and extract the maximum values.
4. Set H to be the $(1-\gamma)$ percentile of the ordered B_1 maximum values found in step 3.

3.3.2 Probability Integral Transformation Based CUSUM (PITC)

3.3.2.1 Problem Setup for Known Distributions

To motivate the PITC method, we consider the case when the underlying in-control distribution is known and given by $F_0(x)$ with a corresponding density function, $f_0(x)$. Under H_0, the transformed incoming observations $U_i = F_0(X_i)$ have a $U(0,1)$ distribution. To capture the out-of-control distribution, we hypothesize that under H_1, $U_i \sim Beta(a,b)$, for properly chosen values of a and b, that is we consider the following transformed test

$$H_0': \quad U_1, ..., U_n \sim U[0,1] = Beta(1,1)$$

$$H_1': \quad U_1, ..., U_{i-1} \sim Beta(1,1)$$

$$U_k, U_{k+1}, ..., U_i \sim Beta(a,b) \text{ for some k} \in \{1, ..., i\} \text{ and for specified values of} (a,b).$$

To identify the values of a and b, we match the first two moments of the $Beta(a,b)$ distribution to the first and second moments, m_1 and m_2, of the true out-of-control distribution of the transformed data. For the multiplicative change, we have $F_1(u) = F_0(u/c)$ and the out-of-control distribution function of $U = F_0(X)$ is given by $F_U(u) = F_0(F_0^{-1}(u)/c)$. For the additive change case, the out-of-control distribution function of U is $F_U(u) = F_0(F_0^{-1}(u) - K)$. The moment matching equations for the multiplicative change are then given by

$$m_1 = \int_0^1 [1 - F_0(F_0^{-1}(u)/c)]du, \tag{3.2a}$$

$$m_2 = \int_0^1 2u[1 - F_0(F_0^{-1}(u)/c)]du. \tag{3.2b}$$

Similarly, under an additive change, we have

$$m_1 = \int_0^1 [1 - F_0(F_0^{-1}(u) - K)]du, \tag{3.3a}$$

$$m_2 = \int_0^1 2u[1 - F_0(F_0^{-1}(u) - K)]du. \tag{3.3b}$$

In general, the solutions for a and b in terms of m_1 and m_2 are given by

$$a = \frac{m_1^2 - m_1 m_2}{m_2 - m_1^2}, \tag{3.4a}$$

$$b = \frac{(m_1 - m_2)(1 - m_1)}{m_2 - m_1^2}. \tag{3.4b}$$

Next, if we denote the density of the "best fit" Beta distribution describing the out-of-control behavior of the transformed data as g_1 and the corresponding in-control uniform distribution by g_0, then an approximately optimal CUSUM increment based on the transformed data is

$$\begin{aligned} \log\left[g_1(U_i)/g_0(U_i)\right] &= \log\left[\, U_i^{a-1}(1-U_i)^{b-1}/B(a,b)\,\right] \\ &= (a-1)\log[F_0(X_i)] + (b-1)\log[1 - F_0(X_i)] - \log B(a,b). \end{aligned} \tag{3.5}$$

The corresponding approximately optimal CUSUM is given by

$$S_i = \max[0, S_{i-1} + (a-1)\log[F_0(X_i)] + (b-1)\log[1 - F_0(X_i)] - \log B(a,b)] \tag{3.6}$$

$$S_0 = 0.$$

Proposition 3. *The increment of the approximately optimal CUSUM in (3.6) is equal in distribution to the increment of the optimal CUSUM* (2.4) *under H_0 for the Weibull distribution, when the alternative is a multiplicative change.*

Proof. The PDF of a Weibull distribution is $f_0(x)=(\alpha/\beta)(x/\beta)^{\alpha-1}e^{-(x/\beta)^\alpha}$ and the CDF is $F_0(x)=1-e^{-(x/\beta)^\alpha}$. For multiplicative change alternatives, the increment (2.4) is

$$\begin{aligned}\log\left[\frac{f_0(X_i/c)}{cf_0(X_i)}\right] &= \log(1/c^\alpha) + \log\left[\frac{e^{-(X_i/c\beta)^\alpha}}{e^{-(X_i/\beta)^\alpha}}\right] \\ &= \log(1/c^\alpha) - (1/c^\alpha - 1)(X_i/\beta)^\alpha.\end{aligned}$$

For the increment (3.5), $F_0^{-1}(u) = \beta[-\log(1-u)]^{1/\alpha}$ and $F_U(u) = 1-(1-u)^{1/c^\alpha}$. Using these functions in (3.2a) and (3.2b), we find that $m_1 = c^\alpha/(1+c^\alpha)$ and $m_2 = 2c^{2\alpha}/(1+c^\alpha)(1+2c^\alpha)$. Thus from (3.4a) and (3.4b) we have $a = 1$ and $b = 1/c^\alpha$. Substituting these values into (3.5) gives the same result as above. □

3.3.2.2 Proposal for Unknown Distributions

Here we return to the initial setup where the underlying in-control distribution F_0 is unknown. To derive the smooth estimate of (3.6), we approximate the unknown cdf F_0 with a kernel estimate $\hat{F}_0$, obtained from the historical data, $\{Y_i\}_{i=1}^N$. We discuss the construction of the kernel estimate of F_0 in subsection 3.4.3. For now, it suffices to mention that an almost sure convergence of $\hat{F}_0$ has been shown in Fernholz (1991). Values m_1 and m_2 (hence a and b) can be calculated numerically by replacing F_0 with $\hat{F}_0$. The proposed PITC tracking statistic is then

$$\hat{S}_i = \max[0, \hat{S}_{i-1} + (\hat{a}-1)\log[\hat{F}_0(X_i)] + (\hat{b}-1)\log[1-\hat{F}_0(X_i)] - \log B(\hat{a},\hat{b})] \quad (3.7)$$

$$\hat{S}_0 = 0.$$

As in the case with the NDEC, we determine the control limit H through Monte-Carlo simulation, using Fernholz's result that $\hat{F}_0(X_i)$ is approximately uniformly distributed. The steps are outlined in algorithm 2 below.

Algorithm 2 : Determining threshold H in the PITC

1. Use $\{Y_i\}_{i=1}^{N}$ to estimate the in-control distribution function $\hat{F}_0(x)$ and calculate $\hat{a}$ and $\hat{b}$.
2. Simulate B_1 cycles of size n, $U_1, ..., U_n$, from $U(0,1)$ distribution.
3. Run $\hat{S}_j = \max[0, \hat{S}_{j-1} + (\hat{a}-1)\log(U_j) + (\hat{b}-1)\log(1-U_j) - \log B(\hat{a}, \hat{b})]$ along each simulated cycle and extract the maximum values.
4. Set H to be the $(1-\gamma)$ percentile of the ordered B_1 maximum values found in step 3.

Note that as in the case with NDEC, H in the PITC also depends on the historical data, this time through $\hat{a}$ and $\hat{b}$. Therefore, if the historical data is updated via the rolling window mechanism, H will also need to be updated.

3.4 Nonparametric Density Estimation (NDE)

Nonparametric density estimation is the process of approximating the unknown density function from an available finite sample of observations which are assumed to have originated from this density. The term nonparametric in this setting is used to emphasize the fact that no known parametric family of distributions is used to describe the unknown density. In contrast to NDE methods, parametric density estimation methods assume that the distribution belongs to some known parametric family, and density estimation simply requires replacing the unknown parameters with corresponding estimates (e.g. MLEs).

The introduction to NDE is attributed to Fix and Hodges (1951). A helpful and practical reference to the most common NDE methods, describing their properties and applications is Silverman (1986). Asymptotic results and convergence properties of various NDE methods are derived in Devroye and Györfi (1985). Other useful references include Prakasa Rao (1983) and Wertz (1978).

3.4.1 Kernel Density Estimation (KDE)

In our work, we use a particular type of NDE, called kernel density estimation (KDE). Silverman (1986) presents an intuitive way to understand the idea behind the KDE starting from its simplest version, the so called naive estimator (NE). Finding the NE of f from a given sample $Y_1, ..., Y_N$ is equivalent to constructing a conventional histogram, whose bins are centered at Y_i's and have a width of a given size h. In this case, h is referred to as a *smoothing parameter* or a *bandwidth* of the resulting estimate. More formally, the naive estimator of the unknown density f at a given point x is defined as follows

$$\hat{f}(x) = \frac{\text{number of } (Y_1, ...Y_N) \text{ in } (x-h, x+h)}{2hN}. \tag{3.8}$$

A compact representation of (3.8) can be given as follows

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{n} w\left(\frac{x - Y_i}{h}\right), \tag{3.9}$$

where $w(x) = 0.5I(|x| < 1)$ is referred to as a "weight" function and $I(y)$ is the indicator function, such that $I(y \in D) = 1$ if $y \in D$ and 0 otherwise. A commonly used, descriptive interpretation of this definition is to imagine that a 'box' of width, $2h$, and height, $(2Nh)^{-1}$, is placed on each observation, Y_i and the value of $\hat{f}$ at a specified point x is found by summing the heights of such boxes at that point. An obvious and strong limitation of the NE is that it is not continuous. The KDE is an improved version of the NE that overcomes this drawback. It is defined as

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - Y_i}{h}\right), \tag{3.10}$$

where $K(x)$, referred to as the "kernel function", is such that $\int_{-\infty}^{\infty} K(x)dx = 1$. Since the latter is the only major condition imposed on the kernel, in practice K can be chosen so that it

has multiple convenient properties, including symmetry, continuity and differentiability. This is particularly important, as $\hat{f}$(x) automatically inherits all features of K. Often, the kernel itself is chosen to be a density function, such as, for example, the standard normal density. This automatically guarantees that $\hat{f}$ is a continuous and differentiable function. Moreover in that case, $\hat{f}$ will also be a density function. Analogous to the naive estimator, the value of the kernel estimator at point x can be viewed as a sum of 'bumps", centered at each sample point Y_i , the spread of which is controlled by the smoothing constant, h. In particular, if the standard normal density is used as a kernel, then $K((x - Y_i)/h)/h$ is a normal density function with mean Y_i and standard deviation h. Correspondingly, $\hat{f}(x)$ can be viewed as the average of n such normal density functions, evaluated at point x. It can be inferred from the definition that large values of h tend to oversmooth the density estimate, thus masking important details, while smaller values undersmooth and exaggerate certain details by making the estimate appear too "rough" and "bumpy", especially in the tails. Proper choice of the smoothing constant h in (3.10) is one of the most important and non-trivial aspects in the construction of the KDE. There are several different criteria that can help to determine h in $\hat{f}$. Since $\hat{f}$ is a random variable calculated from the sample $Y_1, ..., Y_N$, it is natural to discuss its sampling distribution along with such statistics as the sample mean, variance and the quantiles. One of the standard measures of discrepancy between $\hat{f}$ and f, due to Rosenblatt (1956), is the mean integrated square error (MISE), which is calculated as

$$\begin{aligned} MISE(\hat{f}) &= E \int \{\hat{f}(x) - f(x)\}^2 dx \\ &= \int \{E\hat{f}(x) - f(x)\}^2 dx + \int Var\hat{f}(x) dx. \end{aligned} \tag{3.11}$$

That is MISE is the sum of the integrated square of the bias and the integrated variance of $\hat{f}(x)$. It can be shown that $E\hat{f}(x) - f(x)$ does not depend on the sample size N explicitly and hence

just increasing the sample size may not necessarily help reduce the bias (Silverman (1986)). The MISE does however depend on the kernel function and on the smoothing parameter h, therefore for large sample sizes, the kernel function should be adjusted accordingly to obtain asymptotically unbiased estimates.

For the KDE given by (3.10), the MISE (3.11) is rarely found in a closed form. In a special case where the true density is known to be normal with mean μ and variance σ^2, and the standard normal density is used as the kernel K in (3.10), an analytic expression for MISE can be obtained, which is given below (Fryer (1976)):

$$\begin{aligned}(2\sqrt{\pi})MISE = n^{-1}\{h^{-1} - (\sigma^2 + h^2)^{-1/2}\} + \sigma^{-1} \\ + (\sigma^2 + h^2)^{-1/2} - 2\sqrt{2}(2\sigma^2 + h^2)^{-1/2}. \end{aligned} \tag{3.12}$$

The MISE in (3.12) can then be minimized over h to find the optimal value of the bandwidth. In general contexts, it is more feasible to derive the MISE asymptotically. For a symmetric kernel K that is a density function itself with mean 0 and a finite variance, $k_2 \neq 0$, Silverman (1986) has shown that for large values of n and small values of h, the following approximations are true:

$$\begin{aligned}\int \{bias_h(x)\}^2 &\approx 0.25h^4k_2^2 \int \{f''(x)\}^2 dx, \\ \int Var\hat{f}(x)dx &\approx n^{-1}h^{-1} \int \{K(x)\}^2 dx.\end{aligned}$$

From the above equations it becomes apparent that when minimizing (approximate) MISE the two of its summands are conflicting; large values of h will reduce the integrated variance at the same time increasing the integrated square of the bias, and vice-versa, choosing small values of h to minimize the bias will blow up the variance. This obvious trade-off between minimizing

the bias and the variance represents the major challenge in the proper choice of the smoothing constant h in practical situations. The value of h minimizing the corresponding approximate MISE can be derived (Parzen (1962)):

$$h_{opt} = k_2^{-2/5} \left\{ \int [K(t)]^2 dt \right\}^{1/5} \left\{ \int [f''(x)]^2 dx \right\}^{-1/5} n^{-1/5}. \tag{3.13}$$

Silverman (1986) mentions that the value of h_{opt} given by (3.13) is very close to the value of h derived by minimizing the MISE given by (3.11). The apparent problem of equation (3.13) is that the optimal value of h depends on the unknown density f. However it is still worth examining the behavior of the MISE when h_{opt} is used. The corresponding optimal MISE is given by

$$MISE(h_{opt}) \approx (5/4) k_2^{2/5} \left\{ \int [K(t)]^2 dt \right\}^{4/5} \left\{ \int [f''(x)]^2 dx \right\}^{1/5} n^{-4/5}. \tag{3.14}$$

From (3.14), the only term that can be controlled is the one containing the kernel (since f is unknown). So keeping everything else the same, one would want to choose K such that $k_2^{2/5} \left\{ \int [K(t)]^2 dt \right\}^{4/5}$ is minimized. It has been shown in the literature (Hodges and Lehmann (1956)) that the kernel satisfying this condition is the Epanechnikov kernel, given by

$$K_e(t) = \begin{cases} \dfrac{3}{4\sqrt{5}} \left(1 - \dfrac{t^2}{5} \right), & -\sqrt{5} \leq t \leq \sqrt{5}, \\ 0 & \text{otherwise.} \end{cases}$$

Even though $K_e(t)$ is considered as the optimal kernel in terms of minimizing $k_2^{4/5} \int [K(t)]^2 dt$, it has also been shown by multiple authors (Hodges and Lehmann (1956), Stone (1984), Silverman (1986)) that any other conveniently chosen kernel, such as the biweight, the triangular, the rectangular and the Gaussian kernels provide nearly the same result. Moreover, since the choice

of the kernel makes little difference in terms of minimizing the MISE, it is recommended to chose any kernel that is most appropriate for the specific context. We use the Gaussian kernel in our work.

A commonly used approach to determine the value of h_{opt} is by replacing the unknown f in (3.13) by some known reference distribution. In particular, if f is taken to be the normal density with variance σ^2 and mean 0, and if, in addition, the kernel K is the standard normal density, then the value of h_{opt} is easily obtained from (3.13) and is given by

$$h_{opt} = (4/3)^{(1/5)}\sigma n^{-1/5} \approx 1.06\sigma n^{-1/5}. \tag{3.15}$$

To use this value of the bandwidth in practice, one would estimate σ from the data and plug it in the KDE given by (3.10). Silverman (1986) has noted that a kernel density estimate based on h_{opt} given by (3.15) will perform fairly well for distributions, which are close to normal but it will oversmooth densities that are skewed or multimodal. Therefore, an alternative to (3.10) is to replace σ by a more robust measure, $A = \min(\sigma, IQR/1.34)$. That is, define the optimal bandwidth as

$$h_{opt} = (4/3)^{(1/5)}An^{-1/5}. \tag{3.16}$$

A further known modification of (3.16) is to reduce the factor $(4/3)^{1/5}$ to 0.9 (Silverman (1986)). Formulas 3.15 and 3.16 and their modifications are often referred to as *plug-in* formulas for h. In many scenarios, the practitioners are interested in fully automated algorithms for calculating the bandwidth. Two standard and well known procedures are the least squares cross-validation and the maximum likelihood cross-validation algorithms. The following references provide details on the two procedures, computational aspects as well as large-sample optimality properties of the resulting bandwidths, Rudemo (1982), Hall (1983), Bowman (1984), and

Bowman et al. (1984). A significant result is due to Stone (1984) who showed that under certain regularity conditions on f, the bandwidth chosen by the least-squares cross-validation is asymptotically optimal in the sense of minimizing the MISE. On the other hand, it is also important to mention that using the least-squares cross-validation based bandwidth with discretized data may produce poor results, and hence some caution should be applied (Hayfield and Racine (2008)).

3.4.2 Adaptive Kernel Density Estimation (AKDE)

One of the obvious drawbacks of the regular KDE is that it uses a fixed bandwidth value h along the entire sample. The disadvantage of using a constant bandwidth is that it may undersmooth the resulting estimate in areas where the data is sparse, for example in the tails of the distribution, and oversmooth in the areas with higher density (Simonoff (1996)). The problem becomes more apparent when the density being estimated has longer tails or is multimodal. A natural improvement over the kernel estimates with a fixed bandwidth is using kernels with flexible bandwidths, as has been demonstrated by multiple authors, including Abramson (1982), Fox and Long (1990), Salgado-Ugarte et al. (1993, 1995), Bowman and Azzalini (1997), Pagan and Ullah (1999), and Salgado-Ugarte and Perez-Hernandez (2003) Such methods are commonly referred to as adaptive kernel density estimates (AKDE)s. The variable or adaptive bandwidths used in these algorithms can be broadly divided into two main categories. Bandwidths in the first category are calculated so that they vary with the specific value x at which the density is being estimated. Among such methods are, for example, the nearest neighbor estimate and its generalized version, the kth nearest neighbour estimate by Silverman (1986). These methods are essentially improved versions of the naive estimator introduced above. Even though the estimate produced from both approaches has reduced undersmoothing in the tails, it still suffers from similar drawbacks as the naive estimator. In particular, it does not have continuous derivative everywhere and it has unsatisfactorily heavy tails. In addition, it does not

integrate to one and hence is not a probability density itself Silverman (1986). Bandwidths from the second category vary along the sample in a conceptually different way. They do not change with the point at which the density is estimated but rather take into account the configuration of the points in the sample from which the density is to be recovered. More specifically, in the parts of the distribution where the data is not dense, the individual kernels have larger window widths, which provide wider spread and result in an overall smoother estimate. And vice versa, in the denser parts of the distribution, where less smoothing is required, smaller values of individual bandwidth are utilized. Below we give a brief description of the AKDE method and its properties, by Silverman (1986).

$$\hat{f}_{adp}(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\hat{h}_j}K\left(\frac{x - Y_j}{\hat{h}_j}\right) \tag{3.17}$$

with $\hat{h}_j = \hat{h}\hat{\lambda}_j$. The *adaptive factors* $\hat{\lambda}_j$ depend on observations Y_j and are defined as

$$\hat{\lambda}_j = \{g/\hat{f}(Y_j)\}^{\alpha},, \tag{3.18}$$

where $\hat{f}$ is an initial, pilot estimate of f_0 found by any available estimation method, g is the geometric mean of $\hat{f}(Y_j)$, that is $g = \Pi_{i=1}^{N}(\hat{f}(Y_r))^{1/N}$, and $\alpha \in [0, 1]$ is a constant. We use a kernel density estimate based on the standard normal density and a fixed bandwidth to derive the pilot estimate of f_0. We use the rule-of-thumb formula (3.16) to estimate the fixed bandwidth $\hat{h}$. And finally, we set $\alpha = 0.5$ as recommended by Silverman (1986).

Smoothed Bootstrap: Recall that in algorithm 1, in subsection 3.3.1, we mentioned that in the stage of tuning the threshold H for the NDEC, the sample paths were generated by using a smoothed bootstrap algorithm by Silverman (1986). We delayed the introduction to the method till this section since it required knowledge of the relevant concepts from the AKDE method.

We are now ready to present the smoothed bootstrap method, which was implemented as follows: for each $i \in (1, ..., n)$, pick r_i randomly from $\{1, ..., N\}$, then let $\hat{X}_i = Y_{r_i} + (\hat{h}\hat{\lambda}_{r_i})\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$.

3.4.3 Smooth Estimation of a Cumulative Distribution Function (CDF)

The idea of using kernel estimation for approximating an unknown cumulative distribution function (CDF), $F(x)$ has been discussed in the literature by multiple authors, see for example Nadaraya (1964), Reiss (1981), Hill (1985). Assuming that $F(x)$ has a corresponding density $f(x)$, the kernel density estimate of $F(x)$ is taken to be the following statistic

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(u)du, \tag{3.19}$$

where $\hat{f}(x)$ is the KDE of the density function found by (3.10). The resulting estimate $\hat{F}(x)$ possesses several attractive features, including unbiasedness and consistency (Nadaraya (1964)) as well as an almost sure convergence to the uniform distribution (Fernholz (1991)). Clearly, different kernel estimation methods can be applied to approximate $f(x)$, such as the AKDE formula (3.17), thus controlling the overall performance of $\hat{F}(x)$.

Another straightforward approach to a smooth estimation of the unknown CDF is to use a linear interpolation of the empirical cumulative distribution function (ECDF), known as *ogive*, in the main part of the distribution and approximate it with exponential functions at the tails. We consider two cases; in the first, the underlying distribution of the observations is strictly positive, and in the second, it is defined on the whole real line. Let $x_{(1)}, x_{(2)}..., x_{(n-1)}, x_{(n)}$ be the ordered values of the sample of size n, whose ECDF we want to smooth. Note that the corresponding values of the ECDF are $\{y_1, ..., y_n\} = \{1/n, 2/n, ..., (1 - 1/n), 1\}$. For strictly positive distributions, let also $x_{(0)} = 0$ with the corresponding ECDF value of $y_0 = 0$.

Then, the smooth ECDF is defined as

$$\tilde{F}(x) = \begin{cases} m_i x + b_i, & \text{if } x_{(i-1)} \leq x < x_{(i)}, i = 1, ..., n-1 \\ 1 - e^{-\beta_1 x}, & \text{if } x \geq x_{(n-1)} \end{cases}$$

where

$$m_i = \frac{y_i - y_{i-1}}{x_{(i)} - x_{(i-1)}} = \frac{1}{n(x_{(i)} - x_{(i-1)})} \tag{3.20a}$$

$$b_i = y_i - m_i x_{(i)} \tag{3.20b}$$

$$\beta_1 = \frac{\log n}{x_{(n-1)}} \quad \text{(found from the condition } \tilde{F}(x_{(n-1)}) = y_{n-1}) \tag{3.20c}$$

The inverse of $\tilde{F}(x)$ is then found by simply inverting the linear and the exponential functions, as outlined below:

$$\tilde{F}^{-1}(y) = \begin{cases} \dfrac{y - b_i}{m_i}, & \text{if } y_{i-1} \leq y \leq y_i, i = 1, ..., n-1 \\ -\dfrac{\log(1-y)}{\beta_1}, & \text{if } y_{n-1} < y \leq 1 \end{cases}$$

where the parameters m_i, b_i and β_1 are found from (3.20a), (3.20b) and (3.20c) above.

For random variables with both positive and negative values, the ECDF is as follows

$$\tilde{F}(x) = \begin{cases} e^{\beta_2 (x - x_{(2)})}, & \text{if } x \leq x_{(1)} \\ m_i x + b_i, & \text{if } x_{(i-1)} \leq x \leq x_{(i)}, i = 2, ..., n-2 \\ 1 - e^{\beta_3 (x - x_{(n-2)})}, & \text{if } x \geq x_{(n-1)} \end{cases}$$

where

$$\beta_2 = -\frac{\log(n)}{x_{(1)} - x_{(2)}} > 0 \quad \text{(found from the condition } \tilde{F}(x_{(1)}) = 1/n) \tag{3.21a}$$

$$\beta_3 = \frac{\log(n)}{x_{(n-1)} - x_{(n-2)}} > 0 \quad \text{(found from the condition } \tilde{F}(x_{(n-1)}) = 1 - 1/n) \tag{3.21b}$$

The parameters m_i and b_i are found from (3.20a) and (3.20b), as before. The inverse of this approximation is again found by simple inversion.

$$\tilde{F}^{-1}(y) = \begin{cases} x_{(2)} + \dfrac{\log(y)}{\beta_2}, & \text{if } 0 \leq y \leq y_1 \\ \dfrac{y - b_i}{m_i}, & \text{if } y_{i-1} \leq y \leq y_i, i = 2, ..., n-2 \\ x_{(n-2)} - \dfrac{\log(1-y)}{\beta_3}, & \text{if } y_{n-1} \leq y \leq 1 \end{cases}$$

where the constants β_2, β_3, m_i and b_i are as above.

Recall that when we defined the PITC tracking statistic (3.7) in subsection 3.3.2, we mentioned that the corresponding values of a and b could be calculated numerically by replacing the unknown CDF with its smoothly approximated version. In table 3.1, we present comparison results for the a and b values in (3.7), calculated using the known CDF, $F_0(x)$ and the smoothed ECDF, $\tilde{F}_0(x)$, which has been estimated based on $N = 720$ in-control historical observations. We consider two underlying in-control densities, $W(1, 2)$ and $W(2, 1)$, and a multiplicative change alternative, with the out-of-control multiplicative factor $c \in \{1.01, 1.03, 1.07, 1.1, 1.3\}$. The results in both tables show that even for such a moderately sized set of historical observations, $N = 720$, the approximated values of a and b are almost identical to the true parameter values.

f_0	Parameters	c									
		0.70	0.90	0.93	0.97	0.99	1.01	1.03	1.07	1.1	1.3
$Weib(1,2)$	$\hat{a}$	0.999	0.999	0.998	0.999	1.002	1.001	1.001	0.999	0.999	0.997
	$\hat{b}$	1.423	1.248	1.111	1.077	1.011	0.989	0.970	0.932	0.907	0.768
	a	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	b	1.429	1.250	1.111	1.075	1.010	0.990	0.971	0.934	0.909	0.769
$Weib(2,1)$	$\hat{a}$	0.925	0.954	0.980	0.984	0.997	1.003	1.006	1.016	1.019	1.058
	$\hat{b}$	1.891	1.492	1.210	1.137	1.019	0.981	0.947	0.886	0.842	0.625
	a	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	b	2.039	1.563	1.235	1.156	1.020	0.980	0.943	0.873	0.826	0.592

Table 3.1: Comparison of the parameter values of the $Beta(a, b)$ distribution using a known CDF and a smoothed ECDF with a multiplicative change alternative.

3.5 Benchmark Nonparametric CUSUM Designs

3.5.1 Sequential Ranks CUSUM (SRC)

We start with the sequential ranks based CUSUM (SRC) algorithm proposed by McDonald (1990). To introduce the method, some general notation and definitions follow. Given a sequence of observations $X_1, X_2...$ from an unknown, in-control distribution F_0, the sequential rank of X_i is defined as $R_i = 1 + \sum_{j=1}^{i-1}(X_i - X_j)_+$, where x_+ is one for $x > 0$ and zero, otherwise. The SRC is of the form

$$T_0 = 0, \quad T_i = \max(0, T_{i-1} + R_i/(i+1) - k), \quad i = 1, 2, ... \tag{3.22}$$

Here k is a reference parameter fixed in advance. The quantities $R_i/(i+1)$ have been shown to be independent and discrete uniform on $\{1/(i+1), 2/(i+1), ..., i/(i+1)\}$. The CUSUM alarms as soon as T_i crosses a corresponding control limit, H. McDonald approximates the CUSUM T_i with a Markov chain defined on $[0, H)$ under H_0, and for a fixed k numerically determines the value of H by controlling the ARL_0 at a fixed level. We adapt McDonald's SRC

algorithm with a few modifications by calculating the threshold H to control the FAR instead of the ARL_0, as outlined in algorithm 3 below. We set $k = 0.5$. Since U_i's are discretely uniform on [1/(i+1), ... i/(ij+1)] under H_0, taking the reference value to be the mean of the measurements in the tracking statistic has some intuitive appeal. Besides, as we show in figures 3.1 and 3.2, $k = 0.5$ often provides the highest detection rate for a variety of shift and scale alternatives.

Algorithm 3 : Determining threshold H for SRC

---

1. For each $i = 1, ..., n$
   (a) simulate a discrete uniform variable U_i from $\{1/(i+1), ..., i/(i+1)\}$,
   (b) calculate the tracking statistic $T_i = \max(0, T_{i-1} + U_i - k)$
2. Extract the maximum value of $\{T_i\}_{i=1}^{n}$.
3. Repeat steps 1) and 2) B_1 times and set H to be the $(1-\gamma)$ percentile of the B_1 ordered maximum values.

One important requirement when using the SRC is that a significant amount of in-control observations should be accumulated before the change. Otherwise, if the change starts shortly after the monitoring begins, the out-of-control distribution will contaminate the in-control process and the SRC will not be able to properly differentiate between the in-control and the out-of-control states. Simply put, the SRC will treat the out-control distribution as in-control and not signal of a change. In the most extreme case, when the change occurs at the first observation, the true alarm rate will be equal to the false alarm rate. McDonald suggests to accumulate a sufficient amount of past observations before implementing the SRC. In addition, McDonald considers recalibration of the tracking statistic. Recalibration means that once an alarm (false or real) is raised, the SRC is reset to 0, and all observations preceding the alarm are discarded. While recalibration may be useful in certain applications, partial recalibration procedure seems more appropriate for our context. Our reasoning is such that in a real-life ap-

plication historical data should be accumulated up to a certain window size and then continue to be periodically refreshed through an appropriate mechanism. This process however should not be affected by false alarms, that is if the practitioner detects a false alarm during a cycle, the CUSUM should be reset to 0 but no parts of the historical data should be discarded. On the other hand, if the practitioner determines that the alarm indicates a true change, the process should be stopped, the cause of the change should be eliminated and observations during the window of the alarm should be removed from the historical data before resuming surveillance.

Figures 3.1-3.2 demonstrate simulation results that support the choice of $k = 0.5$ mentioned above. We used $N(0, 1)$ as a reference in-control distribution, set $\gamma = 0.1$ (for finding H) and considered three values for the additive change $\{0.1, 0.25, 0.5\}$, and three values for the multiplicative change $\{1.025, 1.075, 1.15\}$. We set the cycle length to $n = 300$ and inserted the change at the 150^{th} observation (as we show in section 4.2, for $n = 300$, the SRC is most efficient when the changepoint is 150). From figures 3.1 and 3.2, we see that for $k \in \{0.1, 0.2, ..., 0.8, 0.9\}$, TAR is the highest at $k = 0.5$, for all three choices of δ and is among the largest for all choices of c_1. The overall TAR values in figure 3.2 are noticeably lower compared to figure 3.1. We discuss this phenomena in more detail in section 4.2.

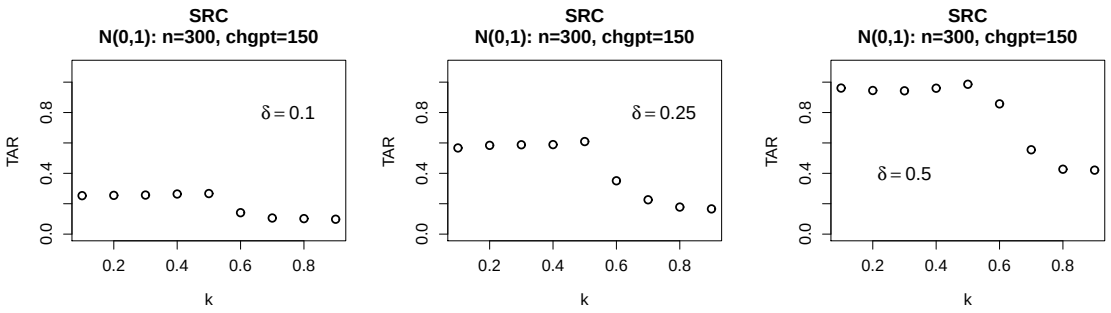

Figure 3.1: Choosing k in the SRC by maximizing the TAR.
Additive change of size δ is inserted at the 150^{th} observation.

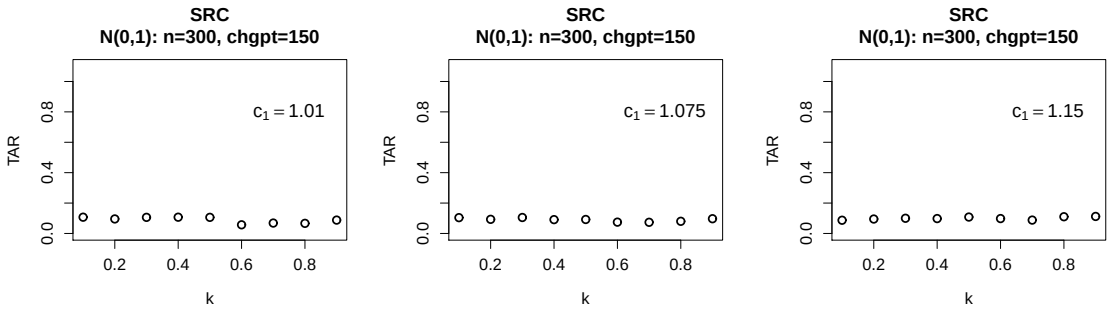

Figure 3.2: Choosing k in the SRC by maximizing the TAR.
Multiplicative change of size c_1 is inserted at the 150^{th} observation.

3.5.2 Transformed CUSUM (TC)

The Transformed CUSUM (TC), introduced by Jeske et al. (2009), is of a different nature. In their paper, the authors consider a scenario of monitoring data within a cycle that is divided into non-stationary timeslots. Observations within each timeslot are considered identically distributed. Possible correlation among the observations in each time-slot is eliminated through an application dependent transformation. A key mechanism in the implementation of TC is utilizing the PIT technique and using historical data to estimate the unknown distribution function. More formally, the one-sided TC tracking statistic for detecting upward elevated values is of the form

$$\hat{S}_i = \max[0, \hat{S}_{i-1} + \hat{F}_0(X_i) - \alpha], \quad \hat{S}_0 = 0. \tag{3.23}$$

where $\hat{F}_0(X_i)$ is the usual ECDF based on N in-control, historical observations, and $\alpha \in (0, 1)$ is an appropriately chosen constant. Under H_0, the transformed observations $\hat{F}_0(X_i)$, conditional on the historical data, are shown to be asymptotically discrete uniform, and therefore the tracking statistic (3.23) can be considered (asymptotically) distribution-free.

In Jeske et al. (2009), the value $\alpha = 0.9$ was motivated on the basis of wanting to detect large shifts. Thus, our first consideration for α is 0.9. One of the drawbacks of this approach in choosing α is its insensitivity to a specified targeted out-of-control distribution. Alternatively, α could be determined as the value that maximizes the TAR for a specified change. Note that setting $\alpha = 0.5$ makes the mean of the TC increment zero, under H_0. Therefore, we search for an appropriate value of α in $[0.5, 0.9]$, as outlined in algorithm 4 below.

Algorithm 4 : Determining an appropriate value of α in the TC

---

1. Simulate B_1 cycles $U_1, ..., U_n$ from $U[0, 1]$ distribution. Set $\alpha = 0.5$ and fix δ_α.
   (a) Run $S_i = \max[0, S_{i-1} + U_i - \alpha]$ along each cycle and extract the maximum values.
   (b) Set H to be the $(1 - \gamma)$ percentile of the B_1 maximum values found in step 2.
   (c) Simulate B_2 cycles of out-of-control data $X_1, ..., X_n$ from $f_1(x)$.
   (d) Run $S_i = \max[0, S_{i-1} + F_0(X_i) - \alpha]$ along each out-of-control cycle, stop the CUSUM as soon as S_i crosses H and record an alarm.
   (e) Evaluate the TAR based on the B_2 out-of-control cycles.
2. Update $\alpha = \alpha + \delta_\alpha$ and repeat steps 1(a)-1(e) till $\alpha = 0.9$. Choose α_0 to be the value of α that provides the highest TAR.

Note that the CUSUM at step 1(d) is based on a known F_0 and corresponds to CUSUM (3.23) with infinite amount of historical data. Figures 3.3-3.4 demonstrate simulation results based on algorithm 4, using $n = 300, \gamma = 0.1, B_1 = 10,000$ and $B_2 = 5,000$. We used $N(0, 1)$ as an in-control distribution and considered the values $\{0.1, 0.25, 0.5\}$ for the additive change, and the values $\{1.025, 1.075, 1.15\}$ for the multiplicative change. We inserted the change at the first observation (in contrast to the SRC, the TC performs adequately when the change starts at the beginning of a cycle). From figures 3.3 and 3.4, we see that the TC achieves the highest or close to the highest TAR at $\alpha = 0.5$ for additive changes, and at $\alpha = 0.9$ for multiplicative changes.

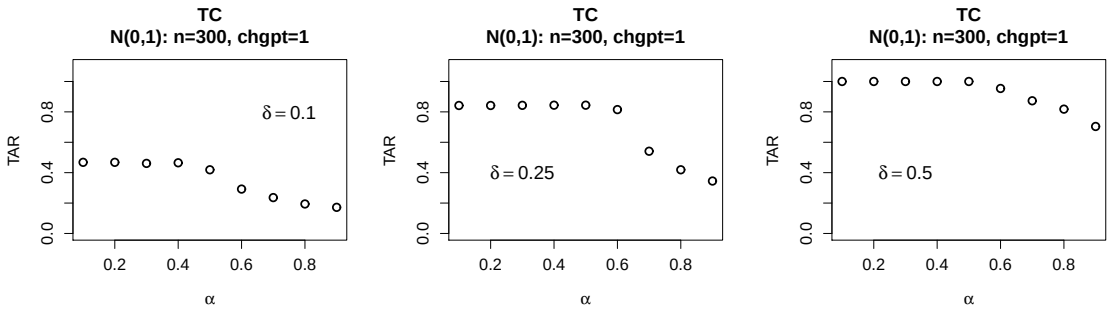

Figure 3.3: Choosing α in the TC by maximizing the TAR.
Additive change of size δ is inserted at the first observation.

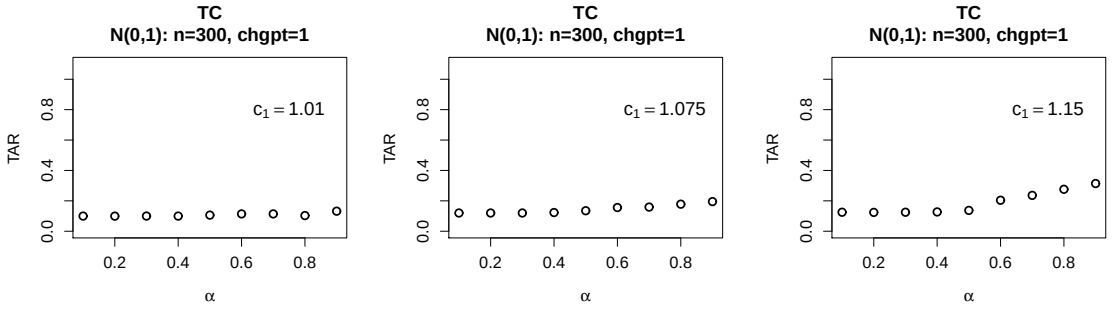

Figure 3.4: Choosing α in the TC by maximizing the TAR.
Multiplicative change of size c_1 is inserted at the first observation.

For a given value α, the control limit H is calculated through Monte Carlo simulations by controlling the FAR at a given level during the monitoring cycle, as outlined in algorithm 5 below. Moreover, since the distribution of the tracking statistic under H_0 depends only on the depth of the historical sample, the threshold H does not need to be re-calculated after the historical dataset is updated through the rolling window technique.

Algorithm 5 : Determining threshold H for the TC

1. Simulate B_1 cycles $U_1, ..., U_n$ from discrete uniform distribution over $[0, 1/N, ..., N]$.
2. Run $S_i = \max[0, S_{i-1} + U_i - \alpha]$ along each cycle and extract the maximum values.
3. Set H to be the $(1 - \gamma)$ percentile of the B_1 maximum values found in step 2.

Chapter 4

Comparative Analyses

In this section we conduct comparative analyses between the PITC and the NDEC and two benchmark algorithms, the Transformed CUSUM (TC) by Jeske et al. (2009) and the sequential-ranks CUSUM (SRC) by McDonald (1990). We consider three different distributions in our simulation study; normal, Student's-t (for heavy tails) and Weibull (for skewness) distributions. The layout of this section is as follows. In section 4.1, we evaluate the performance of all algorithms in terms of their conditional and unconditional FAR, and in section 4.2, we compare the methods based on their power. In section 4.3, we address the important implementation aspect of how to determine when the size of the historical data is sufficient for providing acceptable in-control FAR. We also discuss the computational time involved implementing each algorithm.

4.1 Size Comparison

We start with the case when the CUSUMs are tuned to detect an additive change, that is $f_1(x) = f_0(x - K)$. We then report the case with multiplicative change, that is $f_1(x) = f_0(x/c)/c$. In our simulation analysis, we measure the location shift of interest in

multiples of the in-control standard deviation σ, that is we set $K = \Delta\sigma$, for consistent comparisons of the CUSUMs across different distributions. However in applications, K could be alternatively chosen based on stated specification limits for the process under surveillance. To achieve better understanding on the dependence of our CUSUM schemes on the length of the cycle, we vary $n \in \{30, 60, 90, 120, 210, 300, 450, 600\}$. Also, for each fixed n and a given in-control σ, we consider $\Delta \in \{0.25, 0.5, 1\}$.

We emphasize that once H is determined for the SRC by algorithm 3, it automatically guarantees FAR of γ for every cycle, under H_0. In contrast, the NDEC, the PITC, and the TC depend on how well $\hat{F}_0(x)$ and/or $\hat{f}_0(x)$ estimate their population counterparts for a given historical dataset. Therefore we need to consider the FAR across multiple sets of of historical data. That is, we need to evaluate the conditional FAR for those algorithms. The major steps in evaluating conditional FAR are common for the NDEC, the PITC and the TC algorithms, and are outlined in algorithm 6.

Algorithm 6 : Evaluating FAR for the NDEC, PITC and the TC

---

1. Simulate s cycles of historical data from a specified in-control distribution and determine H for NDEC, PITC and TC, using algorithms 1, 2 and 5.
2. Simulate B_2 cycles of monitoring data from the same in-control distribution. Run each CUSUM through each cycle and report a false alarm if the CUSUM crosses H. Specifically:
   (a) for NDEC run CUSUM (3.1).
   (b) for PITC run CUSUM (3.7).
   (c) for TC run CUSUM (3.23).
3. Calculate the conditional FAR as the proportion of B_2 cycles that alarmed.
4. Repeat steps 1-3 M times.
5. Summarize the distribution of the conditional FAR estimates and calculate the mean over M values to obtain the unconditional FAR.

We consider three different distributions; normal with in-control parameters, μ_0 and σ, Student's t with mean μ_0 and degrees of freedom, ν, and Weibull with shape and scale parameters, α and β, accordingly. We show in a corollary to proposition 4 that for an additive change alternative, under the normal model, the tracking statistics of all CUSUMs in our comparison do not depend on the in-control parameters μ_0 and σ, when H_0 is true. Therefore, in our simulation study (for additive change alternatives) we use $N(0,1)$ as the in-control distribution. For the t-distribution, we show that there is only dependence on ν. However as ν increases, the performances of the CUSUMs get closer to the normal case, so we examined $\nu \in \{3,6\}$. And finally, for the Weibull model, we show that the tracking statistics of all CUSUMs under H_0 depends only on the shape parameter α. We consider three different values for $\alpha \in \{1,2,3\}$. For multiplicative change alternatives, it follows from the corollary to proposition 4 that under the normal distribution, the NDEC tracking statistic still does not depend on σ but it depends on μ_0. Under the t distribution, the NDEC tracking statistic depends on μ_0 and ν. And for the Weibull distribution, there is a dependence on α but there is no dependence on β, as in the additive change alternative. In our simulations, we set $B_1 = 10,000$, $B_2 = 5,000$, $M = 100$ and $\gamma = 0.1$, where appropriate.

Proposition 4. *The NDEC, PITC, TC and SRC tracking statistics are invariant to a linear transformation of the historical observations.*

Proof. We start with the NDEC tracking statistic. Recall that when tuning H, we simulate sample paths based on the smoothed bootstrap technique, as follows. For each $i \in (1,...,n)$, define r_i to be a random integer from $\{1,...,N\}$, then $\hat{X}_i = Y_{r_i} + (\hat{h}\hat{\lambda}_{r_i})\varepsilon_i$, where $\varepsilon_i \sim N(0,1)$, and the adaptive factors $\hat{\lambda}_j$s are given by (3.18). Now suppose the historical data is transformed according to $V_i = c_1 Y_i + c_2$, $i = 1,...,N$. Let $*$ denote calculations computed within this transformed setting. For example, the NDE of the in-control distribution is given by

$$\hat{f}_0^*(u) = \frac{1}{N\hat{h}^*}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j^*}\phi\left(\frac{u-V_j}{\hat{h}^*\hat{\lambda}_j^*}\right).$$

It is easy to verify that $\hat{h}^* = c_1\hat{h}$. It follows that the pilot estimate under the transformed setting is given by

$$\hat{f}^*(u) = \frac{1}{c_1 N\hat{h}}\sum_{r=1}^{N}\phi\left(\frac{u-V_r}{c_1\hat{h}}\right),$$

and thus

$$\hat{f}^*(V_j) = \frac{1}{c_1}\left\{\frac{1}{N\hat{h}}\sum_{r=1}^{N}\phi\left(\frac{Y_j-Y_r}{\hat{h}}\right)\right\} = \frac{1}{c_1}\hat{f}(Y_j).$$

Hence $\hat{\lambda}_j^* = \hat{\lambda}_j$, and as a result we have

$$\hat{f}_0^*(u) = \frac{1}{c_1}\frac{1}{N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{u-V_j}{c_1\hat{h}\hat{\lambda}_j}\right).$$

Considering that $\hat{X}_i^* = V_{r_i} + (\hat{h}^*\hat{\lambda}_{r_i}^*\varepsilon_i) = c_1\hat{X}_i + c_2$, we have

$$\hat{f}_0^*(\hat{X}_i^*) = \frac{1}{c_1}\frac{1}{N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{\hat{X}_i-Y_j}{\hat{h}\hat{\lambda}_j}\right) = \frac{1}{c_1}\hat{f}_0(\hat{X}_i).$$

First consider the case for additive change alternatives. If the additive shift in the untransformed setting was $K = \Delta\sigma$, then in the transformed setting it becomes $K^* = c_1\Delta\sigma = c_1 K$. Thus,

$$\hat{f}_1^*(u) = \hat{f}_0^*(u - c_1K) = \frac{1}{c_1 N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{u-c_1K-V_j}{c_1\hat{h}\hat{\lambda}_j}\right),$$

and so

$$\hat{f}_1^*(\hat{X}_i^*) = \frac{1}{c_1}\frac{1}{N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{c_1\hat{X}_i+c_2-c_1K-V_j}{\hat{h}\hat{\lambda}_j}\right) = \frac{1}{c_1}\hat{f}_0(\hat{X}_i-K).$$

Consequently, the conditional distribution (given the historical data) of the NDEC increment during the tuning phase is invariant to the transformation $c_1 Y_i + c_2, i = 1, ..., N$. Moreover, since $\hat{X}_i^* = c_1 \hat{X}_i + c_2$, in a similar way it can be shown that this invariance property holds during monitoring phase as well.

Now consider the case of multiplicative alternatives but suppose $c_2 = 0$. We have

$$\begin{aligned}\hat{f}_1^*(u) &= \frac{1}{c}\hat{f}_0^*\left(\frac{u}{c}\right) \\ &= \frac{1}{cc_1}\frac{1}{N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{\frac{u}{c}-V_j}{c_1\hat{h}\hat{\lambda}_j}\right),\end{aligned}$$

and so

$$\begin{aligned}\hat{f}_1^*(\hat{X}_i^*) &= \frac{1}{cc_1}\frac{1}{N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{\frac{c_1\hat{X}_i+c_2}{c}-V_j}{c_1\hat{h}\hat{\lambda}_j}\right) \\ &\overset{(c_2=0)}{=} \frac{1}{cc_1}\frac{1}{N\hat{h}}\sum_{j=1}^{N}\frac{1}{\hat{\lambda}_j}\phi\left(\frac{\frac{\hat{X}_i}{c}-Y_j}{\hat{h}\hat{\lambda}_j}\right) \\ &= \frac{1}{cc_1}\hat{f}_0^*\left(\frac{\hat{X}_i}{c_1}\right).\end{aligned}$$

Thus, for a multiplicative change alternative, the invariance is with respect to a linear transformation of the historical observations of a form $c_1 Y_i, i = 1, ..., N$. Also we note that in the in-control state $\hat{X}_i$ has the same distribution as X_i, and hence the invariance property of the NDEC increment holds during the in-control monitoring phase as well. □

Corollary 5. *The dependence of the CUSUM tracking statistics from the parameters of the underlying distributions.*

Proof. A straightforward application of proposition 4 allows us to prove the invariance of the in-control distribution of the NDEC tracking statistic to (μ_0, σ) under the normal model for an

additive change, and to σ for a multiplicative change; to any location shifts under the t model for an additive change; and to β under the Weibull model, both for additive and multiplicative changes. Specifically, the historical data in these three cases can be respectively represented as $Y_i \stackrel{d}{=} \sigma N(0,1) + \mu_0$, $Y_i \stackrel{d}{=} t(\nu) + \mu_0$, and $Y_i \stackrel{d}{=} \beta Weibull(\alpha, 1)$.

The SRC tracking statistic (3.22) is distribution-free by definition, and hence it is invariant to the linear transformation of the historical observations both under an additive and a multiplicative change alternatives. The PITC and the TC tracking statistics, correspondingly given by (3.7) and (3.23), are asymptotically distribution free under H_0, and hence one can argue their invariance to the corresponding transformations of the historical observations under either additive or multiplicative change alternatives. □

Clearly, the larger the size of the historical data for a given length of the monitoring cycle, the more precise are the nonparametric estimators of the density and the distribution functions used in the NDEC, the PITC and the TC. Correspondingly the conditional FARs will be closer to the nominal value. To properly compare the performance of the algorithms, for a fixed value of the cycle length n, we found the smallest value of the number of historical cycles, s, necessary to provide a satisfactory performance of the conditional FAR, for all distributions considered in the study. We used the following criteria for a satisfactory behavior of the conditional distribution of the FAR; for a nominal FAR level of 0.1, no more than 10% of the simulated historical datasets (10 out of 100) should have a FAR less than 0.075 or greater than 0.125, no more than 5% of the FAR's (5 out of 100 historical datasets) should be less than 0.065 or greater than 0.135, and finally to protect against outliers, an additional criteria that no dataset should have a FAR larger than 0.14 or less than 0.06 was used. A similar criteria can be used for any other desired nominal FAR γ. We refer to pairs (n, s), satisfying this criteria as *feasible* values.

4.1.1 Additive change

In table 4.1 below, we demonstrate how we arrived at the minimum value of s, which satisfies the feasibility criteria for the case when the change results from an additive transformation. For this illustration, we set $n = 300$ and considered the NDEC under the normal model. Each row of table 4.1 corresponds to a fixed number of cycles, s, and contains the proportion of the simulated historical datasets out of $M = 100$, whose FAR is below or above our specified break-points. Considering $s = 10$ we see that 67% of the historical datasets had FAR between 0.075 and 0.125. For 17% of the datasets the FAR was below 0.075, with the lowest values being between 0.04 and 0.06 recorded for 5% of the datasets. On the higher side, 16% of the datasets had FAR greater than or equal to 0.125, with 4 outliers greater than 0.14. These proportions did not satisfy our criteria defined above, so we increased s accordingly. The value $s = 30$, in this case, provided the minimum necessary performance of the FAR in terms of our defined criteria. For the last row with $s = \infty$, we ran the CUSUM using the true normal density f_0 instead of its nonparametric estimate. Table 4.1 illustrates the convergence of the conditional FAR to its nominal rate as the depth of the historical dataset increases.

Distribution of conditional FAR							
s	≤ 0.06	≤ 0.065	≤ 0.075	[0.075,0.125]	≥ 0.125	≥ 0.135	≥ 0.14
10	5	8	17	67	16	9	4
20	0	1	10	84	6	2	1
30	0	1	3	93	4	1	0
40	0	1	2	96	2	0	0
∞	0	0	0	100	0	0	0

Table 4.1: Choosing the minimal s for $n = 300$, under $N(0, 1)$, for an additive change alternative.

Analogously, we determined the appropriate values of s for the other targeted shifts, and then did a similar analysis for the t and the Weibull distributions. We then declared the

largest of these s values as the feasible value for $n = 300$ when using the NDEC. In a similar way, a feasible value of s for $n = 300$ was found for the PITC, $\text{TC}_{\alpha=0.5}$ and $\text{TC}_{\alpha=0.9}$. Table 4.2 shows the result of this study.

	s			
n	NDEC	PITC	$\text{TC}_{\alpha=0.5}$	$\text{TC}_{\alpha=0.9}$
30	120	120	120	100
60	75	75	80	70
90	65	65	70	60
120	55	55	60	45
210	40	40	40	30
300	35	35	35	20
450	25	25	30	18
600	20	20	20	15

Table 4.2: Feasible s values for fixed n values, for an additive change alternative.

For the pair $(n = 300, s = 35)$, the distributions of the conditional FAR of the NDEC, PITC and the TC methods under the $N(0, 1)$, $t(3)$, $t(6)$ and $Weibull(1, 1), Weibull(2, 1)$, $Weibull(3, 1)$ models are displayed in figures 4.1-4.6 in the form of boxplots. The structure of each boxplot is standard, with the left and the right sides of the rectangle indicating the 25^{th} and the 75^{th} quantiles respectively and the middle band representing the median. The endpoints of the whiskers are correspondingly the smallest and the largest FAR values observed in the set of 100 simulated historical samples. The three boxplots in each of the NDEC and the PITC figures correspond to the three different target shifts, $\Delta = 0.25, 0.5, 1$. The two boxplots in the TC figure represent the conditional distribution of the FAR for $\alpha = 0.9$ and for $\alpha = 0.5$, and are the same for any target shift Δ, because the TC tracking statistic does not require specifying a target shift. We note the similarity of the boxplots for each Δ in the NDEC and PITC plots, regardless of the underlying distribution. During our extensive simulations, we observed a near equivalence of the PITC and the NDEC results both in terms of the FAR and the TAR for all

distributions considered. It is also worth mentioning that the distribution of the FAR for the TC method based on α=0.9 has a noticeably smaller variance than the NDEC and PITC algorithms, whereas the spread is larger when α=0.5. While a tighter distribution of the conditional FAR around the nominal value is a desired feature, the omnibus nature of the TC method may result in a low detection rate of a true change. Our expectations are confirmed in the power analysis described in the next section. In addition to the conditional distribution of the FAR, the boxplots also report the FAR as the average over the 100 historical datasets. These numbers are shown above the boxplots and it can be seen that all of them are close to the nominal FAR of 0.1.

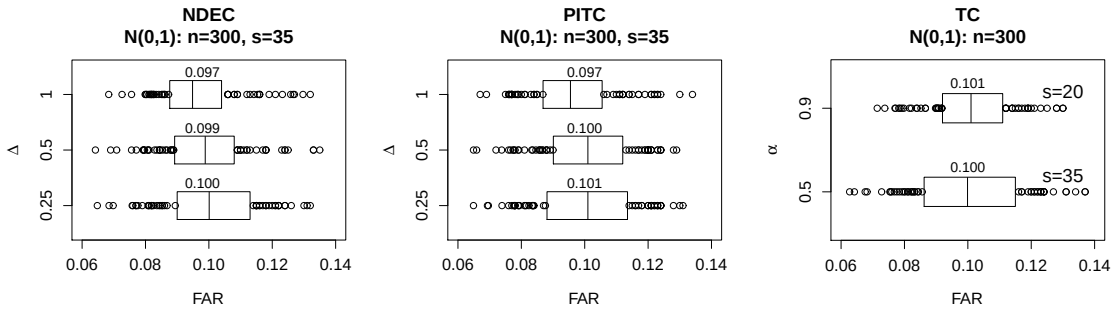

Figure 4.1: Conditional FAR distribution based on 100 historical datasets from $N(0, 1)$ for an additive change alternative. Nominal FAR is $\gamma = 0.1$.

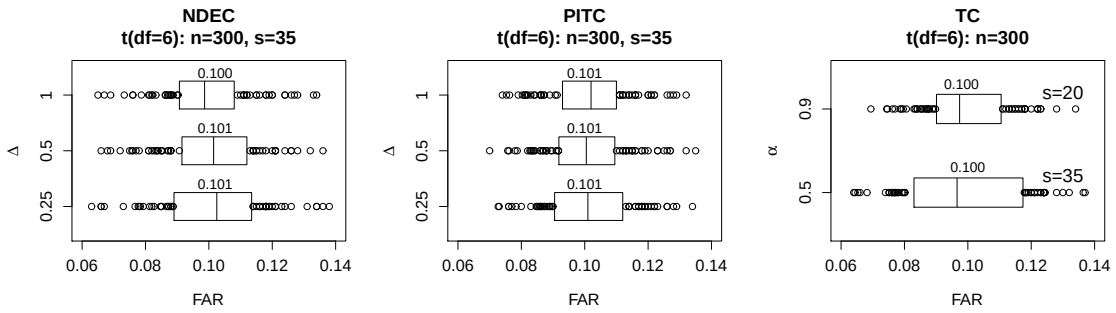

Figure 4.2: Conditional FAR distribution based on 100 historical datasets from $t(6)$ for an additive change alternative. Nominal FAR is $\gamma = 0.1$.

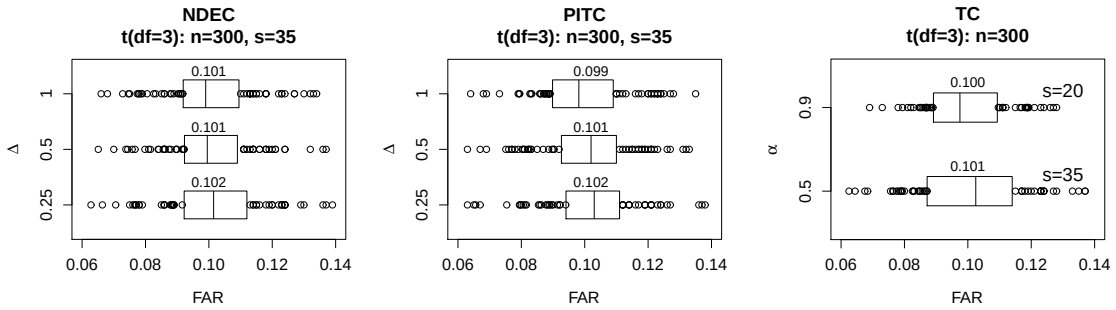

Figure 4.3: Conditional FAR distribution based on 100 historical datasets from $t(3)$ for an additive change alternative. Nominal FAR is $\gamma = 0.1$.

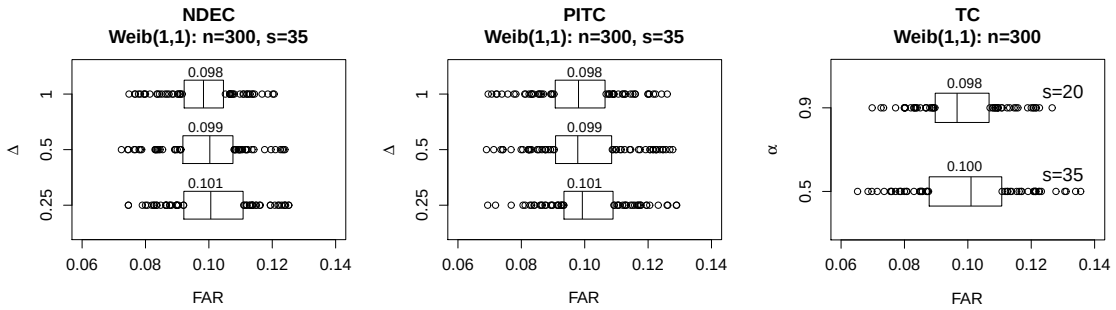

Figure 4.4: Conditional FAR distribution based on 100 historical datasets from $Weibull(1,1)$ for an additive change alternative. Nominal FAR is $\gamma = 0.1$.

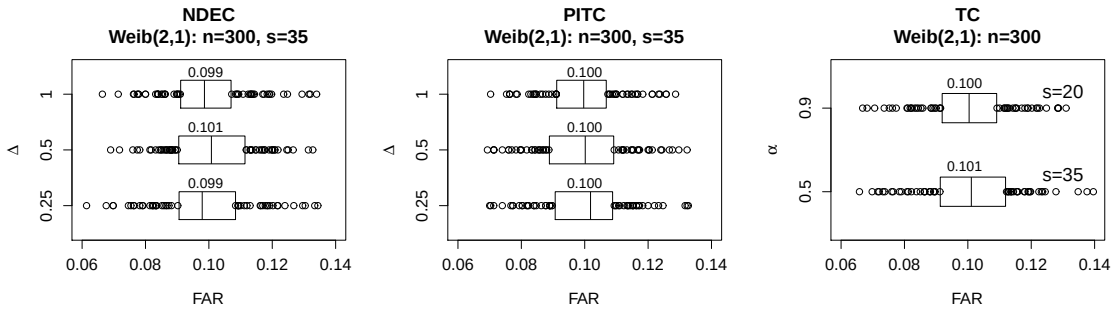

Figure 4.5: Conditional FAR distribution based on 100 historical datasets from $Weibull(2,1)$ for an additive change alternative. Nominal FAR is $\gamma = 0.1$.

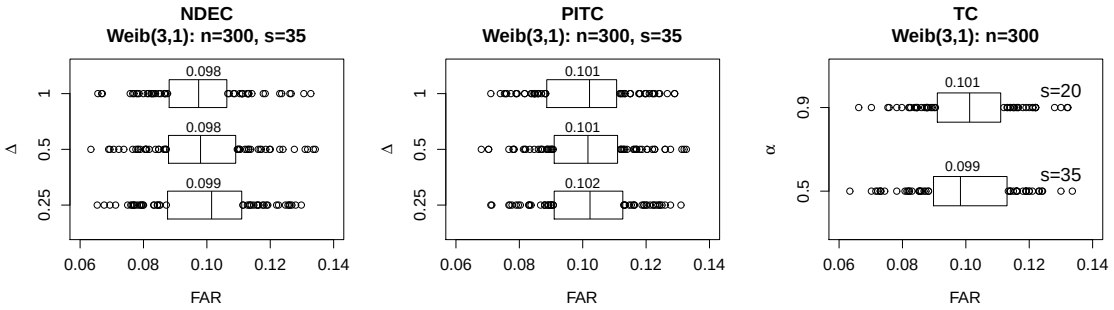

Figure 4.6: Conditional FAR distribution based on 100 historical datasets from $Weibull(3, 1)$ for an additive change alternative. Nominal FAR is $\gamma = 0.1$.

4.1.2 Multiplicative change

For the multiplicative change, $f_1(x) = f_0(x/c)/c$, we consider three different values of the target multiplicative factor $c \in \{1.025, 1.05, 1.1\}$. The feasible pairs corresponding to the multiplicative change are given in table 4.3. As with the additive change, we illustrate the plots of the conditional FAR values for $n = 300$, in figures 4.7-4.12. Since the TC tracking statistic does not depend on the specified out-of-control state, the boxplots corresponding to $\text{TC}_{\alpha=0.9}$ for the scale alternative are the same as the boxplots for $\text{TC}_{\alpha=0.9}$ for the additive alternative. As before, we observe that all methods achieve an unconditional FAR value close to $\gamma = 0.1$.

n	NDEC	PITC	$\text{TC}_{\alpha=0.9}$
30	180	180	100
60	120	120	70
90	95	95	60
120	85	85	45
210	70	70	30
300	60	60	20
450	45	45	18
600	35	35	15

Table 4.3: Feasible s values for fixed n values for a multiplicative change alternative.

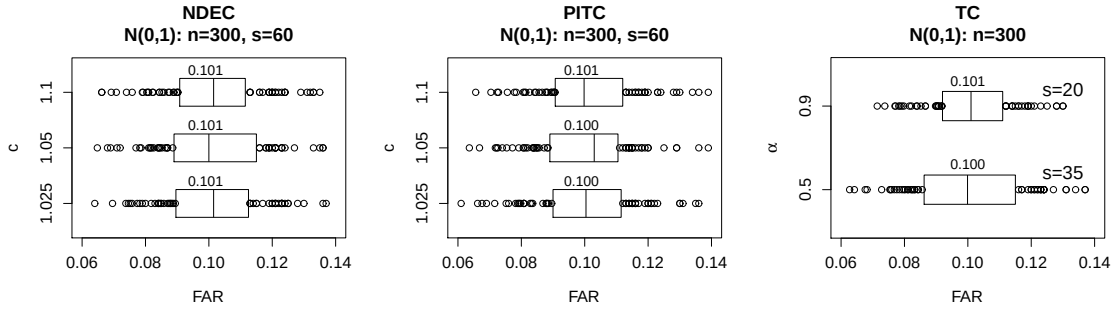

Figure 4.7: Conditional FAR distribution based on 100 historical datasets from $N(0,1)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

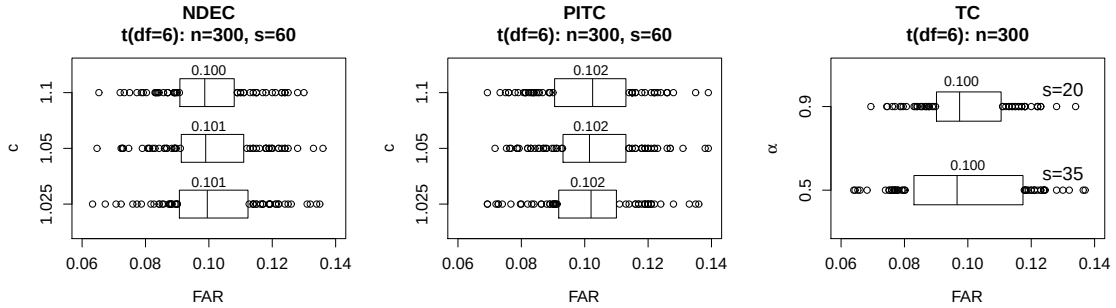

Figure 4.8: Conditional FAR distribution based on 100 historical datasets from $t(6)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

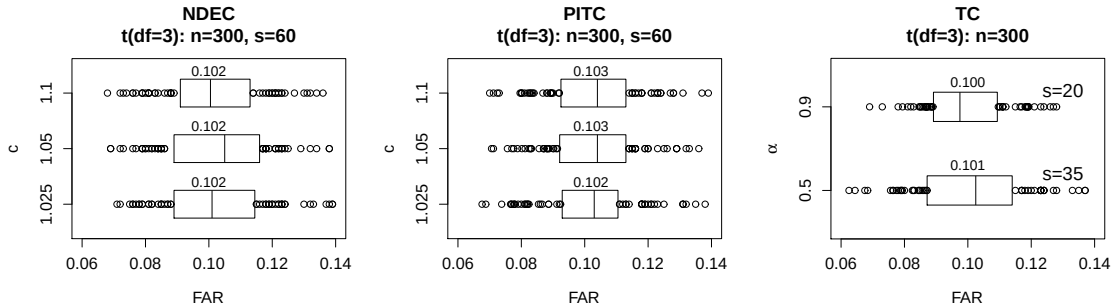

Figure 4.9: Conditional FAR distribution based on 100 historical datasets from $t(3)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

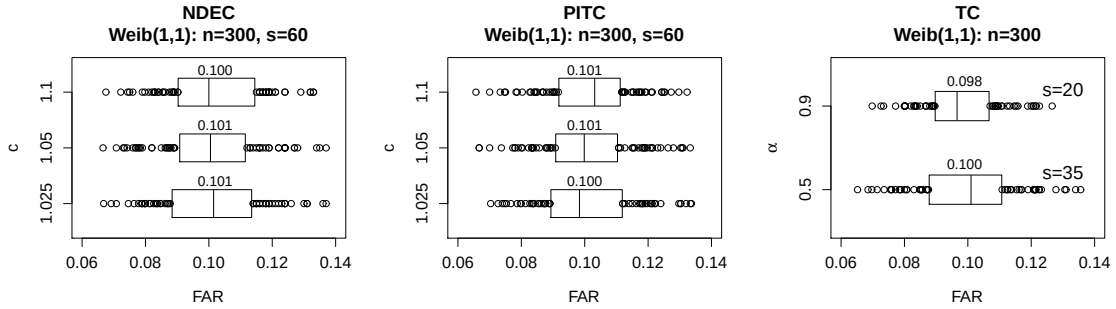

Figure 4.10: Conditional FAR distribution based on 100 historical datasets from $Weibull(1,1)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

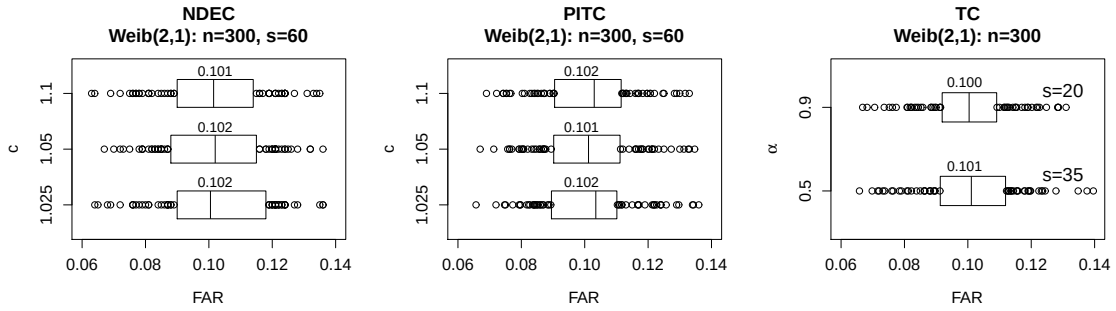

Figure 4.11: Conditional FAR distribution based on 100 historical datasets from $Weibull(2,1)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

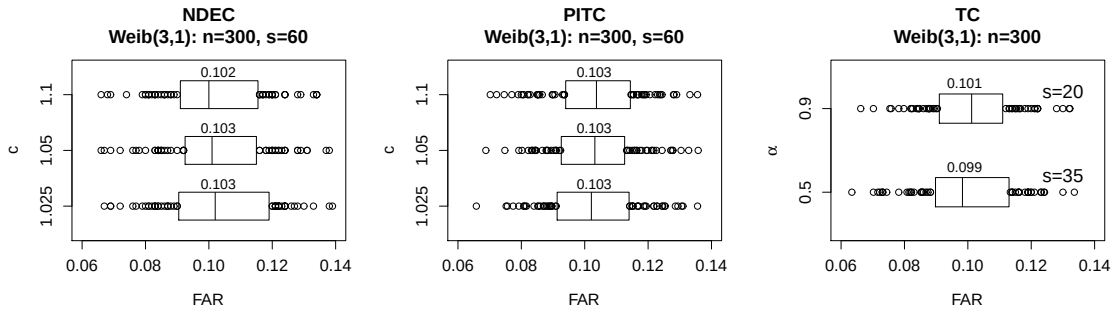

Figure 4.12: Conditional FAR distribution based on 100 historical datasets from $Weibull(3,1)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

4.2 Power Analysis

In this subsection, we compare the CUSUM algorithms on the basis of their effectiveness in detecting a true change. For this purpose, for each of the 100 in-control historical datasets used in the evaluation of the conditional FAR, we simulate B_2 monitoring cycles of data corresponding to a specified out-of-control state and calculate the conditional true alarm rate (TAR) and the conditional average detection delay (ADD). The conditional TAR is estimated through the proportion of the monitoring cycles during which the given CUSUM signals of a change. Given that the change has been detected, the detection delay (DD) is the number of observations collected from the moment the change occurred until the time of the signal (within each cycle), the ADD is the average of the DD over B_2 cycles. The unconditional TAR and ADD estimates are obtained by averaging the corresponding conditional values over the 100 historical samples. Corresponding characteristics of the SRC are derived by simulating B_2 monitoring cycles only once, since there is no dependence on the historical data. We use the unconditional TAR as a primary measure of the power of each CUSUM algorithm.

To gain insight on the performance of all CUSUMs with respect to the location of the changepoint, we investigate four different scenarios by inserting a specified fault at the first, $n/4$, $n/2$ and $3n/4$ time points of each of the B_2 monitoring cycles. Recall that the SRC does not perform adequately if the change starts immediately, and thus for the first scenario we compared only the NDEC, PITC and TC (tables 4.16 and 4.17). When implementing the SRC, we partially follow McDonald's recalibration approach in that we also reset the CUSUM to zero, in case of a false alarm. However, we do not discard the historical observations accumulated during the cycle in which a false alarm occurred. If the tracking statistic crosses the threshold after the changepoint, we stop the CUSUM, record a true alarm for the given cycle, then proceed to the same evaluation of a new cycle.

4.2.1 Additive change

We vary the size of the real change $\delta\sigma$ inserted at the specified locations of each monitoring cycle according to $\delta \in \{0.1, 0.25, 0.5, 0.75, 1\}$ for all distributions under consideration, that is for a given in-control density $f_0(x)$ with known mean, μ_0 and standard deviation, σ, we generate data for the monitoring cycles from $f_1(x) = f_0(x - \delta\sigma)$. Thus, for example for the normal distribution, we simulate observations before the changepoint from $N(\mu_0, \sigma)$, and starting with the change-point, from $N(\mu_0 + \delta\sigma, \sigma)$ (recall that in case of the normal distribution, without loss of generality, we can set $\mu_0 = 0$ and $\sigma = 1$). The NDEC and PITC methods are tuned to detect location shift of size $\Delta\sigma$, with $\Delta \in \{0.25, 0.5, 1\}$. Recall that the TC and the SRC methods do not depend on Δ. Again, we vary $n \in \{30, 60, 90, 120, 210, 300, 450, 600\}$. Results from the unconditional power analysis for $n = 300$ and $\Delta = 0.25$ under the normal distribution are given in table 4.4. Corresponding results for the t-distribution are given in tables 4.5 and 4.6, and for the Weibull distribution, in tables 4.7, 4.8 and 4.9. The first five columns in each table represent the ADD of each of the five CUSUMs in our study, while the last five columns contain the power, in terms of the TAR. The Monte Carlo error in all the reported results is negligibly small with standard errors in the range $[0.1, 0.84]$ for ADD, and in $[0, 0.043]$ for TAR, across all algorithms and settings.

First, we note that the results of the NDEC and the PITC both in terms of the TAR and the ADD are very similar to each other, regardless of the underlying distribution.We also see that for a given distribution, when all other parameters are fixed, the NDEC and the PITC are uniformly more powerful in terms of the TAR for all values of δ, relative to the TC and SRC. For example, if we refer to row with $\delta = 0.25$ corresponding to changepoint 75 in table 4.4, the TAR for the NDEC and the PITC is 95%, for $\text{TC}_{\alpha=0.5}$ it is 74%, for the SRC it is almost 59%, and finally for the $\text{TC}_{\alpha=0.9}$ the TAR is 27%. Since in table 4.4 the NDEC and the PITC

Change Point	δ	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	0.10	119.86	119.58	110.60	147.25	145.11	0.41	0.41	0.13	0.30	0.25
	0.25	88.37	87.58	108.20	107.10	127.68	0.95	0.95	0.27	0.74	0.59
	0.50	37.59	36.93	94.11	56.53	85.57	1.00	1.00	0.66	0.99	0.97
	0.75	23.09	22.35	62.79	38.75	52.81	1.00	1.00	0.96	1.00	1.00
	1.00	16.77	15.90	30.26	30.13	38.13	1.00	1.00	1.00	1.00	1.00
150	0.10	84.97	84.64	75.02	95.31	91.06	0.27	0.27	0.09	0.17	0.26
	0.25	74.43	73.88	73.74	86.28	86.39	0.83	0.83	0.19	0.61	0.63
	0.50	37.72	36.58	67.62	51.25	61.31	0.99	1.00	0.51	0.92	0.98
	0.75	22.90	22.18	52.06	35.15	40.05	1.00	1.00	0.88	1.00	1.00
	1.00	16.64	15.81	29.59	27.31	30.37	1.00	1.00	0.99	1.00	1.00
225	0.10	47.58	47.28	38.15	48.99	46.99	0.12	0.12	0.05	0.09	0.12
	0.25	45.77	45.86	38.41	45.29	45.03	0.46	0.47	0.10	0.38	0.32
	0.50	35.70	35.07	36.91	44.83	43.55	0.97	0.97	0.30	0.91	0.81
	0.75	22.92	22.16	32.79	33.47	36.06	1.00	1.00	0.65	0.99	0.99
	1.00	16.63	15.788	24.56	26.08	27.48	1.00	1.00	0.93	1.00	1.00

Table 4.4: Power when in-control distribution is $N(0,1)$, $n = 300, s = 35$.
Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$

have been tuned to detect a shift of 0.25σ, it should not be surprising that both our algorithms detect an actual shift of size 0.25σ more often than the SRC or the $TC_{\alpha=0.9}$. However, the better performance at all other δ values as well speaks of the superiority of our algorithms. Notice that although the TAR of $TC_{\alpha=0.5}$ is significantly higher than that of $TC_{\alpha=0.9}$ (0.74 vs 0.27), it is still lower than the TAR of the NDEC and the PITC (0.95). For $\delta = 0.1$, all methods have comparably lower detection rates but even in this case, the TARs of NDEC and the PITC are significantly higher from the rest. Actual changes of size $\delta \geq 0.25$ are naturally detected faster by all methods, with the PITC and the NDEC still having the lowest ADD, followed by the $TC_{\alpha=0.5}$, then the SRC and finally the $TC_{\alpha=0.9}$.

Continuing with the discussion of table 4.4, when an additive change of 0.25σ is inserted at the 150^{th} observation, we observe a decrease in the TAR values compared to the findings when the change is inserted at the 75^{th} observation for the NDEC, PITC and the TC. This can be explained by the shorter time left in a cycle to alarm compared to the case when the

changepoint is inserted earlier. The TAR values for the SRC, on the other hand, are comparable to and even slightly higher than those in the first part of the table. This is also intuitively justifiable considering the fact that the SRC needs a sufficient amount of observations before the change; having 150 versus 75 accumulated observations before the change in a cycle of length 300 provides an improved power for the SRC. The behavior of the ADD is similar to the results for changepoint 75; for $\delta = 0.1$ and $\delta = 0.25$, the $\text{TC}_{\alpha=0.9}$ has the lowest ADD value among all methods but starting with $\delta = 0.5$, the NDEC and the PITC catch up and then outperform the $\text{TC}_{\alpha=0.9}$.

To complete the discussion of the power analysis under the normal model, we look at the results corresponding to changepoint 225 in table 4.4. We observe a further decrease in the TAR values of the NDEC, PITC and the TC compared to the two previous cases, which is due to even a shorter time left in a cycle to alarm. The TAR values of the SRC are also lower compared to the results corresponding to changepoint 75 or 150. This shows that although more observations are accumulated before the change, the very little time left in a cycle to alarm causes the power to decrease. For $\delta = 0.1$, all methods have very low power, with the lowest TAR being 0.05, recorded for the $\text{TC}_{0.9}$. Such small TAR value(s) might seem confusing at first glance, since one would expect the TAR to be greater than the nominal FAR, $\gamma = 0.1$, for any $\delta \neq 0$. However, recall that when calculating power, we reset the CUSUMs to zero if the alarm occurs before the changepoint, i.e. if it is a false alarm, and therefore the TAR values for changepoints ≥ 1 may not necessarily be $\geq \gamma$. In the case, when the change is inserted at the first observation, we show in tables 4.16 and 4.17 that for all CUSUMs, the TAR> 0.1 for any $\delta > 0$. The power gradually increases with δ, and location shifts of size $\geq 0.25\sigma$ are detected with quite high rates. The behavior of the ADD is similar to the results corresponding to changepoint 75 and 150; for $\delta = 0.1$ and $\delta = 0.25$, the $\text{TC}_{\alpha=0.9}$ has the lowest ADD value among all methods but starting with $\delta = 0.5$, the NDEC and the PITC catch up and then outperform the $\text{TC}_{\alpha=0.9}$.

Tables 4.5, 4.6 and 4.7, 4.8, 4.9 contain results from a similar analysis, using $t(3)$, $t(6)$ and $Weibull(1,1)$, $Weibull(2,1)$, $Weibull(3,1)$ distributions. Element by element comparison of the corresponding tables shows the similarity of the findings with the normal case. As before, both the NDEC and the PITC have the highest TAR among all procedures considered, regardless of the location or the size of the change. For lower values of δ, the ADD of our two CUSUM procedures is comparable with the smallest ADD value, achieved by $TC_{\alpha=0.9}$. For $\delta \geq 0.5$ the NDEC and the PITC have significantly lower ADD than any of the other three methods. In addition, comparing tables 4.5, 4.6 corresponding to the $t(3)$ and the $t(6)$ distributions, respectively, we observe that the TAR values of all methods (except the $TC_{0.9}$) increase as the degrees of freedom decrease. This can be attributed to the fact that for heavier tail distributions, (positive) additive change results in more frequent occurrence of really large observations with respect to the in-control data, which are quickly caught by our CUSUMs.

Change		Unconditional ADD					Unconditional TAR				
Point	δ	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	0.10	107.63	106.19	105.81	127.02	131.36	0.49	0.50	0.18	0.42	0.33
	0.25	71.29	70.61	99.64	83.18	110.69	1.00	1.00	0.30	0.96	0.82
	0.50	28.45	26.82	87.78	39.69	57.90	1.00	1.00	0.64	1.00	0.99
	0.75	17.91	18.03	54.37	27.29	37.31	1.00	1.00	0.95	1.00	1.00
	1.00	10.55	11.92	23.21	18.52	29.02	1.00	1.00	1.00	1.00	1.00
150	0.10	79.71	78.14	78.36	81.29	91.98	0.39	0.40	0.12	0.31	0.34
	0.25	65.87	64.55	72.04	68.74	77.65	0.94	0.96	0.23	0.79	0.85
	0.50	27.30	29.09	63.91	34.32	43.75	1.00	1.00	0.58	0.98	1.00
	0.75	18.19	17.93	47.45	23.81	29.82	1.00	1.00	0.93	1.00	1.00
	1.00	11.21	11.07	21.93	19.16	23.68	1.00	1.00	1.00	1.00	1.00
225	0.10	37.95	36.35	34.10	38.22	47.84	0.21	0.22	0.06	0.14	0.16
	0.25	34.12	34.02	32.31	34.62	45.05	0.64	0.63	0.13	0.57	0.52
	0.50	28.38	27.86	29.52	30.33	37.88	1.00	1.00	0.32	1.00	0.98
	0.75	18.66	19.11	25.75	24.79	27.16	1.00	1.00	0.67	1.00	1.00
	1.00	12.17	11.46	21.83	20.05	21.83	1.00	1.00	0.98	1.00	1.00

Table 4.5: Power when in-control distribution is $t(3)$, $n = 300$, $s = 35$.
Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$.

Change Point	δ	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	0.10	118.27	118.12	107.12	145.77	142.48	0.43	0.42	0.15	0.36	0.26
	0.25	84.55	86.35	104.84	99.04	122.75	0.97	0.95	0.29	0.84	0.67
	0.50	33.77	35.60	91.27	51.69	76.48	1.00	1.00	0.67	0.99	0.90
	0.75	22.05	21.63	59.42	35.73	47.41	1.00	1.00	0.95	1.00	1.00
	1.00	16.01	15.61	28.35	28.13	35.01	1.00	1.00	1.00	1.00	1.00
150	0.10	84.69	83.01	76.11	95.59	90.75	0.30	0.31	0.10	0.26	0.27
	0.25	70.89	68.20	74.51	82.62	85.18	0.88	0.89	0.21	0.71	0.70
	0.50	33.62	32.19	66.82	46.88	55.92	1.00	1.00	0.53	0.96	0.91
	0.75	21.13	21.02	49.14	32.42	36.45	1.00	1.00	0.91	1.00	1.00
	1.00	15.93	14.22	26.66	25.55	27.98	1.00	1.00	1.00	1.00	1.00
225	0.10	47.26	46.91	37.80	49.30	47.87	0.15	0.14	0.05	0.11	0.13
	0.25	46.65	45.58	36.13	45.87	44.84	0.53	0.51	0.11	0.42	0.37
	0.50	32.92	34.14	34.41	42.62	43.77	0.99	0.98	0.31	0.94	0.88
	0.75	21.14	21.25	31.89	30.93	33.09	1.00	1.00	0.65	1.00	0.99
	1.00	15.89	15.52	23.42	24.35	25.71	1.00	1.00	0.95	1.00	1.00

Table 4.6: Power when in-control distribution is $t(6)$, $n = 300, s = 35$.
Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$.

Change Point	δ	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	0.10	103.11	101.23	112.26	139.16	137.89	0.54	0.53	0.10	0.49	0.41
	0.25	65.16	64.29	107.37	74.41	103.59	1.00	1.00	0.14	0.99	0.86
	0.50	24.32	24.10	94.23	42.32	56.46	1.00	1.00	0.27	1.00	0.99
	0.75	19.22	19.22	85.64	31.12	30.46	1.00	1.00	0.45	1.00	1.00
	1.00	13.40	13.39	67.89	20.19	23.78	1.00	1.00	0.72	1.00	1.00
150	0.10	72.18	70.88	78.08	86.80	91.71	0.45	0.46	0.06	0.35	0.41
	0.25	43.11	44.10	76.49	79.13	74.56	1.00	1.00	0.11	0.77	0.91
	0.50	22.98	24.45	74.30	43.76	41.69	1.00	1.00	0.19	0.96	1.00
	0.75	17.03	17.99	63.34	34.31	29.94	1.00	1.00	0.32	1.00	1.00
	1.00	14.19	15.06	48.49	21.29	24.36	1.00	1.00	0.63	1.00	1.00
225	0.10	42.24	43.51	40.09	46.99	46.94	0.33	0.32	0.05	0.19	0.21
	0.25	35.12	36.76	38.97	44.10	43.41	1.00	1.00	0.06	0.73	0.60
	0.50	26.34	26.88	37.82	38.32	37.26	1.00	1.00	0.07	1.00	0.99
	0.75	21.15	22.07	37.02	27.02	27.17	1.00	1.00	0.19	1.00	1.00
	1.00	14.50	14.96	35.14	20.18	22.45	1.00	1.00	0.43	1.00	1.00

Table 4.7: Power when in-control distribution is $Weibull(1, 1)$, $n = 300, s = 35$.
Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$.

Comparing tables 4.7, 4.8 and 4.9, we see that for all CUSUMs (except the $TC_{\alpha=0.9}$), the TAR values increase, as the Weibull shape parameter decreases (along with the skewness). A possible explanation is that between two (positively) skewed distributions, the one with less

Change Point	δ	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	0.10	127.19	126.30	112.49	146.52	143.97	0.38	0.40	0.11	0.34	0.25
	0.25	98.47	97.93	109.60	103.79	126.56	0.88	0.89	0.20	0.79	0.62
	0.50	48.21	47.59	96.24	55.53	83.22	1.00	1.00	0.47	1.00	0.97
	0.75	33.19	32.14	84.48	38.37	51.89	1.00	1.00	0.81	1.00	1.00
	1.00	20.86	19.31	54.84	29.36	37.78	1.00	1.00	0.96	1.00	1.00
150	0.10	84.53	85.75	75.57	96.21	91.56	0.32	0.33	0.07	0.28	0.26
	0.25	63.59	64.08	74.22	86.46	87.35	0.81	0.82	0.15	0.67	0.65
	0.50	32.74	31.80	71.21	51.05	60.03	1.00	1.00	0.35	0.99	0.99
	0.75	21.89	22.68	62.37	35.94	39.57	1.00	1.00	0.67	1.00	1.00
	1.00	16.68	15.92	46.14	25.88	29.70	1.00	1.00	0.93	1.00	1.00
225	0.10	47.11	46.93	39.42	51.72	47.57	0.26	0.25	0.06	0.14	0.12
	0.25	36.70	35.31	38.86	49.56	45.49	0.59	0.60	0.08	0.39	0.33
	0.50	28.51	28.39	37.62	46.38	44.58	0.99	0.99	0.18	0.91	0.83
	0.75	24.67	24.12	35.37	33.04	35.36	1.00	1.00	0.41	1.00	1.00
	1.00	17.08	16.83	31.43	24.69	27.01	1.00	1.00	0.74	1.00	1.00

Table 4.8: Power when in-control distribution is $Weibull(2, 1)$, $n = 300, s = 35$. Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$.

Change Point	δ	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	0.10	134.22	135.13	112.63	147.53	142.59	0.36	0.35	0.13	0.31	0.24
	0.25	102.46	102.90	110.47	108.98	128.54	0.85	0.85	0.25	0.72	0.60
	0.50	52.31	53.77	96.24	57.35	86.19	1.00	1.00	0.63	0.99	0.95
	0.75	35.89	36.18	67.28	39.22	54.17	1.00	1.00	0.94	1.00	1.00
	1.00	22.54	22.99	34.23	30.04	38.83	1.00	1.00	0.99	1.00	1.00
150	0.10	88.74	87.66	74.36	95.52	92.74	0.29	0.30	0.08	0.25	0.24
	0.25	69.90	67.14	73.97	86.47	88.12	0.79	0.78	0.17	0.63	0.61
	0.50	38.43	38.05	68.33	52.03	62.75	0.99	0.98	0.49	0.96	0.97
	0.75	29.82	28.13	54.05	35.00	40.76	1.00	1.00	0.85	1.00	1.00
	1.00	20.04	19.97	32.51	27.93	30.71	1.00	1.00	0.99	1.00	1.00
225	0.10	49.89	49.03	38.72	52.44	45.55	0.22	0.21	0.05	0.12	0.11
	0.25	36.77	35.64	37.88	49.90	44.28	0.54	0.54	0.10	0.29	0.31
	0.50	31.28	30.41	36.68	47.81	42.69	0.92	0.92	0.28	0.84	0.79
	0.75	27.55	27.21	33.20	35.76	36.65	1.00	1.00	0.61	1.00	0.99
	1.00	20.65	19.36	26.19	25.17	27.94	1.00	1.00	0.90	1.00	1.00

Table 4.9: Power when in-control distribution $Weibull(3, 1)$, $n = 300, s = 35$. Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$.

skewness has a heavier right tail, therefore an additive change for the latter case gives rise to relatively more (positive) high values, which are detected by our CUSUMs.

4.2.2 Multiplicative change.

For the multiplicative change, $f_1(x) = f_0(x/c)/c$, we consider three different values of the target multiplicative factor, $c \in \{1.025, 1.05, 1.1\}$. We let the size of the actual multiplicative factor, say, c_1, vary in $\{1.025, 1.05, 1.075, 1.1, 1.15\}$. The unconditional ADD and TAR values for $n = 300$ and c=1.05 under the normal model are given in table 4.10.

Change		ADD					TAR				
Point	c_1	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	1.025	134.17	132.80	122.28	148.61	149.69	0.21	0.23	0.09	0.09	0.10
	1.050	125.01	126.77	118.71	145.37	147.63	0.34	0.33	0.12	0.11	0.11
	1.075	112.89	110.03	115.33	143.72	147.12	0.51	0.52	0.15	0.13	0.11
	1.100	103.21	104.17	108.15	142.90	146.51	0.67	0.69	0.19	0.14	0.11
	1.150	89.18	91.22	105.07	140.08	145.02	0.91	0.92	0.25	0.15	0.12
150	1.025	87.19	88.27	85.76	87.94	88.68	0.17	0.18	0.10	0.08	0.09
	1.050	85.42	85.05	84.89	86.22	88.02	0.24	0.24	0.11	0.09	0.09
	1.075	84.70	84.16	83.57	84.39	87.06	0.35	0.36	0.12	0.11	0.10
	1.100	82.16	83.21	82.42	83.54	84.55	0.49	0.48	0.14	0.12	0.10
	1.150	81.71	80.39	81.65	82.45	82.74	0.77	0.77	0.18	0.13	0.11
225	1.025	42.27	44.19	38.89	41.31	45.13	0.12	0.13	0.05	0.03	0.05
	1.050	39.15	41.85	37.71	39.36	43.61	0.16	0.16	0.06	0.04	0.05
	1.075	36.20	37.71	35.90	39.22	42.78	0.28	0.29	0.07	0.06	0.06
	1.100	32.09	32.99	34.66	38.98	41.12	0.37	0.36	0.07	0.07	0.07
	1.150	29.11	30.14	34.42	37.88	38.42	0.49	0.51	0.09	0.07	0.08

Table 4.10: Power when in-control distribution $N(0, 1)$, $n = 300, s = 60$.
Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$.

First, we note that again, the NDEC and the PITC perform very similarly both in terms of TAR and ADD for all settings. From the first part of table 4.10 corresponding to changepoint 75, we see that for $c_1 < 1.05$ both the NDEC and the PITC have relatively low TAR values but for $c_1 \geq 1.05$ the TAR grows to reasonably high values. The TC and the SRC, on the other hand, have disappointingly low TAR values, even for relatively large values of c_1. Such low fault detection rates can be intuitively explained as follows. Since the multiplicative change results in heavier tails of the original distribution, the monitoring data contains both very

large and very small observations equally often. In case of the SRC, these outlier observations have either very high or very low sequential ranks, which in the long run neutralize each other by producing an almost equal number of positive and negative increments of roughly equal sizes in (3.22). Hence the SRC tracking statistic does not grow sufficiently large to cross the specified threshold. Similarly, in the TC tracking statistic (3.23), the positive increment values resulted from large observations are balanced by negative increment values produced by small observations (in the latter case, the TC is reset to 0), therefore the TC tracking statistic grows very slowly to cross the specified threshold. In terms of ADD, we observe a similar tendency as before, for smaller values of c_1, the TC has the smallest and the SRC has the largest ADD. For $c_1 \geq 1.05$, the ADD values of the NDEC and the PITC are comparable to the TC method. The TAR values corresponding to changepoints 150 and 225 are noticeably lower compared to changepoint 75, for all methods (for the SRC the TAR values for changepoints 75 and 150 are similar as in table 4.4). As before, we explain this by the shorter time left at the end of a cycle to alarm. At the same time, the NDEC and the PITC still maintain their superiority in terms of the TAR, and are comparable to the TC in terms of the low ADD values.

Similar results for $t(3)$ and $t(6)$ are shown in tables 4.11 and 4.12, and for $Weibull(1,1)$, $Weibull(2,1)$ and $Weibull(3,1)$ in tables 4.13, 4.14 and 4.15. Note that while the performance of the NDEC, the PITC and the TC is roughly the same for all distributions and under all considered scenarios, the SRC has improved performance under the Weibull distribution compared to the normal and the t distributions. In particular, if we compare the last columns of tables 4.10, 4.11 and 4.13, we observe a slight increase in TAR values of SRC in the last table. This can be attributed to the fact that $Weibull(1,1), Weibull(2,1)$ and $Weibull(3,1)$ are no longer symmetric distributions (as the normal or the t), and therefore we don't get the same effect of cancellation of the increment values using the SRC tracking statistic. The improved TAR values of SRC for changepoint 75 are still lower than the corresponding TAR values of the NDEC and

the PITC. However when the change starts at the 225^{th} observation, the TAR values of the SRC are nearly the same as the TAR values of the NDEC and the PITC. In addition, when comparing the TAR values in tables 4.13, 4.14 and 4.15, in contrast to the additive change results, the TAR values of all CUSUMs increase as the shape parameter increases. As before, this can be explained through the change in skewness of the Weibull distribution. The multiplicative change of a Weibull distribution with a larger shape parameter makes the long right tail of the density heavier, which results in a frequent occurrence of a outstandingly large observations. We also notice a slight change in the behavior of the TC algorithm; under Weibull distribution (and for a multiplicative change alternative), the $TC_{0.9}$ seems to be more efficient in terms of the TAR and the ADD than the $TC_{0.5}$.

Change		ADD					TAR				
Point	c_1	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	1.025	139.14	137.44	131.81	153.66	154.28	0.19	0.18	0.09	0.08	0.10
	1.050	133.11	132.98	129.37	149.18	152.75	0.25	0.26	0.12	0.09	0.10
	1.075	130.22	129.10	127.76	147.45	151.34	0.41	0.39	0.13	0.11	0.10
	1.100	122.98	120.47	123.75	146.83	150.86	0.52	0.51	0.14	0.11	0.11
	1.150	110.71	111.13	117.19	143.15	149.64	0.86	0.84	0.18	0.12	0.11
150	1.025	94.03	92.32	91.82	89.16	90.18	0.12	0.13	0.06	0.06	0.09
	1.050	91.74	90.45	90.38	89.02	89.25	0.17	0.16	0.07	0.07	0.09
	1.075	89.21	88.92	89.09	88.22	89.04	0.22	0.23	0.09	0.08	0.09
	1.100	87.43	87.18	88.62	87.57	88.95	0.34	0.35	0.11	0.09	0.10
	1.150	86.61	86.97	87.91	86.41	87.10	0.59	0.59	0.11	0.09	0.10
225	1.025	46.69	47.01	46.48	48.82	48.89	0.08	0.08	0.03	0.03	0.05
	1.050	44.16	44.78	44.15	47.90	48.56	0.13	0.12	0.04	0.03	0.05
	1.075	43.10	42.90	43.03	47.18	47.39	0.15	0.16	0.05	0.04	0.05
	1.100	41.28	41.85	42.91	46.69	47.12	0.29	0.28	0.06	0.04	0.06
	1.150	39.15	39.66	40.17	45.04	47.06	0.36	0.35	0.07	0.05	0.06

Table 4.11: Power when in-control distribution is $t(3)$, $n = 300$, $s = 60$.
Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$.

Change Point	c_1	ADD NDEC	ADD PITC	ADD $TC_{0.9}$	ADD $TC_{0.5}$	ADD SRC	TAR NDEC	TAR PITC	TAR $TC_{0.9}$	TAR $TC_{0.5}$	TAR SRC
75	1.025	136.88	135.91	128.51	150.96	151.39	0.20	0.19	0.09	0.09	0.10
	1.050	129.07	128.46	125.39	148.24	150.41	0.29	0.31	0.11	0.10	0.10
	1.075	116.75	117.43	121.05	147.33	150.02	0.47	0.48	0.14	0.11	0.11
	1.100	108.24	107.81	114.82	144.38	149.57	0.60	0.62	0.17	0.12	0.11
	1.150	92.73	94.59	109.44	143.75	148.32	0.87	0.88	0.22	0.12	0.12
150	1.025	90.26	89.02	87.48	88.49	88.17	0.14	0.13	0.07	0.07	0.09
	1.050	88.90	88.61	87.10	87.31	88.03	0.21	0.22	0.08	0.07	0.09
	1.075	86.33	85.75	86.24	86.98	87.54	0.28	0.29	0.10	0.09	0.10
	1.100	84.68	84.20	85.13	84.72	85.71	0.45	0.46	0.11	0.10	0.10
	1.150	83.41	83.93	83.26	83.91	83.99	0.71	0.72	0.13	0.11	0.10
225	1.025	43.70	44.19	43.73	46.44	47.20	0.11	0.10	0.05	0.04	0.05
	1.050	42.13	43.04	43.55	45.12	46.93	0.15	0.14	0.05	0.04	0.05
	1.075	39.85	39.76	42.19	44.13	45.31	0.22	0.23	0.06	0.05	0.06
	1.100	35.11	34.88	39.82	41.27	43.88	0.34	0.34	0.07	0.05	0.06
	1.150	32.38	30.56	37.67	40.89	42.64	0.45	0.47	0.08	0.06	0.07

Table 4.12: Power when in-control distribution is $t(6)$, $n = 300$, $s = 60$.
Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$.

Change Point	c_1	ADD NDEC	ADD PITC	ADD $TC_{0.9}$	ADD $TC_{0.5}$	ADD SRC	TAR NDEC	TAR PITC	TAR $TC_{0.9}$	TAR $TC_{0.5}$	TAR SRC
75	1.025	131.57	132.16	117.15	153.68	146.53	0.15	0.15	0.10	0.14	0.12
	1.050	126.71	126.19	115.04	151.72	139.13	0.22	0.23	0.12	0.18	0.16
	1.075	124.18	125.30	114.33	149.12	137.42	0.29	0.31	0.14	0.25	0.20
	1.100	121.03	121.98	112.01	143.87	136.61	0.44	0.45	0.16	0.30	0.21
	1.150	119.14	120.05	111.18	140.34	135.55	0.67	0.68	0.22	0.34	0.28
150	1.025	87.81	86.18	85.87	97.37	90.28	0.11	0.12	0.06	0.11	0.12
	1.050	86.29	86.04	85.04	95.47	90.15	0.15	0.15	0.08	0.14	0.14
	1.075	85.62	85.51	84.71	94.14	89.91	0.20	0.19	0.09	0.18	0.17
	1.100	84.33	83.87	84.12	93.22	88.42	0.29	0.28	0.10	0.24	0.22
	1.150	83.10	83.08	83.66	89.84	88.13	0.45	0.44	0.13	0.32	0.29
225	1.025	44.37	44.90	42.88	53.73	45.65	0.07	0.07	0.03	0.06	0.07
	1.050	41.31	42.03	40.62	51.16	44.58	0.09	0.09	0.04	0.08	0.08
	1.075	40.91	40.19	39.83	49.86	44.20	0.10	0.10	0.05	0.08	0.09
	1.100	40.12	39.76	39.41	45.51	43.21	0.12	0.11	0.06	0.10	0.10
	1.150	39.24	39.03	38.15	43.90	42.75	0.19	0.20	0.07	0.16	0.17

Table 4.13: Power when in-control distribution is $Weib(1, 1)$, $n = 300$, $s = 60$.
Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$.

Tables 4.16 and 4.17 contain power analysis results corresponding to the case when the change is inserted at the first observation under all distributions in our study and for both additive and multiplicative change alternatives. The columns in each table represent the un-

Change Point	c_1	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	1.025	127.67	126.89	114.18	152.49	150.66	0.23	0.23	0.10	0.19	0.15
	1.050	118.15	117.34	113.98	149.76	145.91	0.56	0.57	0.14	0.30	0.22
	1.075	109.21	111.03	112.33	141.48	142.11	0.76	0.75	0.21	0.39	0.32
	1.100	106.82	104.51	111.27	132.72	135.81	0.87	0.88	0.27	0.46	0.40
	1.150	66.40	64.75	101.12	128.55	129.34	0.99	0.99	0.43	0.68	0.61
150	1.025	85.28	87.92	78.78	98.64	92.95	0.16	0.15	0.07	0.16	0.14
	1.050	83.01	82.97	76.99	95.00	89.99	0.43	0.44	0.10	0.23	0.21
	1.075	81.77	86.10	76.68	94.93	91.91	0.60	0.62	0.15	0.38	0.32
	1.100	74.39	76.07	75.09	94.56	90.14	0.75	0.74	0.21	0.49	0.41
	1.150	61.72	60.58	71.00	93.42	86.22	0.95	0.96	0.34	0.70	0.63
225	1.025	45.68	46.84	40.93	48.93	47.29	0.09	0.08	0.04	0.07	0.08
	1.050	44.98	43.95	38.99	48.33	45.47	0.17	0.18	0.06	0.12	0.11
	1.075	43.34	43.60	38.13	45.42	43.99	0.28	0.29	0.08	0.17	0.15
	1.100	42.23	43.18	36.87	45.31	41.68	0.37	0.37	0.11	0.22	0.20
	1.150	42.09	42.27	36.54	42.11	40.80	0.65	0.66	0.19	0.33	0.32

Table 4.14: Power when in-control distribution is $Weibull(2, 1)$, $n = 300, s = 35$. Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$.

Change Point	c_1	Unconditional ADD					Unconditional TAR				
		NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	SRC
75	1.025	124.32	124.78	111.01	151.13	147.96	0.34	0.35	0.12	0.28	0.19
	1.050	114.22	113.82	107.97	141.75	142.50	0.72	0.73	0.23	0.41	0.32
	1.075	94.36	92.19	106.45	126.81	135.25	0.92	0.94	0.36	0.52	0.46
	1.100	72.77	73.06	101.20	108.85	128.27	0.98	0.98	0.49	0.74	0.62
	1.150	45.43	46.73	87.65	79.94	108.91	1.00	1.00	0.77	0.92	0.86
150	1.025	88.97	89.63	76.00	95.74	90.11	0.24	0.23	0.08	0.20	0.18
	1.050	85.17	86.98	74.55	94.21	89.60	0.57	0.55	0.13	0.37	0.31
	1.075	76.52	78.14	73.97	91.87	89.21	0.78	0.77	0.21	0.51	0.48
	1.100	67.83	66.17	71.73	85.14	86.41	0.94	0.95	0.35	0.73	0.63
	1.150	44.09	43.16	65.12	67.65	75.85	0.99	0.99	0.61	0.91	0.89
225	1.025	48.44	48.77	39.77	48.43	46.68	0.09	0.10	0.05	0.09	0.09
	1.050	47.88	47.09	39.45	48.21	46.43	0.26	0.25	0.07	0.18	0.16
	1.075	46.14	46.67	37.67	46.70	45.04	0.41	0.42	0.14	0.27	0.25
	1.100	45.55	45.21	37.29	44.59	43.10	0.58	0.59	0.19	0.39	0.33
	1.150	45.19	44.86	36.43	48.14	43.02	0.66	0.66	0.37	0.63	0.57

Table 4.15: Power when in-control distribution is $Weibull(3, 1)$, $n = 300, s = 35$. Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$.

conditional ADD and the unconditional TAR values of all CUSUM methods, besides the SRC (since the SRC does not perform adequately under this scenario). We note that the TAR values of all algorithms in this scenario are higher than the results corresponding to the cases when the

changepoint was inserted later during the cycle. This is quite expected as all CUSUMs in the study except the SRC do not require additional in-control data before the change starts during the monitoring cycle. Moreover, as we saw comparing results for changepoints 75, 150 and 225, for the NDEC, PITC and the TC procedures the earlier the change starts during the monitoring cycle, the more efficient these procedures are in detecting it. In particular, we note that, since no recallibration is applied in this scenario, all of the TAR values in tables 4.16 and 4.17 are at least as large as the nominal FAR $\gamma = 0.1$ (recall that in some cases above, the TAR was less than 0.1 due to recallibration).

4.3 Implementation Aspects

4.3.1 Sanity Test

In the simulation analysis in subsection 4.1 above, we illustrated an approach for determining the necessary number of historical cycles s for a given cycle length n in order to provide an adequate conditional FAR during the monitoring phase. This was a necessary input to the power study in order to make legitimate comparisons between CUSUM algorithms. In the simulation study we knew in advance the true f_0 and hence we were able to simulate monitoring data as well as replicates of historical datasets from the same known distribution. In a real-life application, the practitioner has a single dataset and is tasked with determining if the corresponding s is a feasible value for the purpose of providing a sufficient conditional FAR. Below we describe a sanity test that will allow the practitioner to carry out this task. Suppose $Y_1, ..., Y_N$ represent the historical data of size N from an unknown density f_0 available to the practitioner. Here, $N = ns$, where s is the number of cycles of historical data available to the practitioner. Our proposed sanity test consists of the steps outlined in algorithm 7.

Distribution		Unconditional ADD				Unconditional TAR			
	δ	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$
	0.10	162.42	160.73	152.22	195.28	0.50	0.52	0.16	0.45
	0.25	103.16	102.90	135.69	136.41	0.98	0.99	0.32	0.84
$N(0,1)$	0.50	46.06	48.54	126.37	88.95	1.00	1.00	0.69	1.00
	0.75	39.49	37.66	76.41	46.96	1.00	1.00	0.98	1.00
	1.00	20.93	22.48	37.25	37.89	1.00	1.00	1.00	1.00
	0.10	141.12	143.88	139.39	164.36	0.60	0.59	0.23	0.54
	0.25	92.24	93.57	124.82	112.37	1.00	1.00	0.41	1.00
$t(3)$	0.50	35.75	34.05	96.61	49.99	1.00	1.00	0.76	1.00
	0.75	26.27	28.31	63.77	33.19	1.00	1.00	1.00	1.00
	1.00	14.58	15.96	29.53	21.10	1.00	1.00	1.00	1.00
	0.10	159.44	157.03	148.59	187.56	0.56	0.55	0.17	0.49
	0.25	96.84	95.61	130.78	123.19	0.99	0.98	0.35	0.87
$t(6)$	0.50	41.09	42.17	118.47	63.86	1.00	1.00	0.72	1.00
	0.75	33.49	32.36	71.10	44.43	1.00	1.00	0.99	1.00
	1.00	18.56	19.28	34.94	35.61	1.00	1.00	1.00	1.00
	0.10	136.76	133.62	152.65	178.37	0.71	0.72	0.12	0.68
	0.25	85.53	84.90	144.91	86.54	1.00	1.00	0.18	0.97
$W(1,1)$	0.50	37.68	38.19	141.21	49.19	1.00	1.00	0.34	1.00
	0.75	29.32	30.84	128.43	36.64	1.00	1.00	0.59	1.00
	1.00	23.78	22.03	105.63	30.61	1.00	1.00	0.86	1.00
	0.10	198.77	201.04	150.57	202.77	0.56	0.54	0.14	0.47
	0.25	140.62	141.22	142.52	128.56	0.95	0.94	0.26	0.73
$W(2,1)$	0.50	62.56	60.88	134.73	67.58	1.00	1.00	0.56	0.96
	0.75	43.71	44.19	101.54	46.76	1.00	1.00	0.87	1.00
	1.00	35.45	34.50	56.78	36.60	1.00	1.00	0.97	1.00
	0.10	206.35	205.61	149.45	203.86	0.49	0.50	0.15	0.29
	0.25	147.43	150.06	139.48	133.32	0.89	0.91	0.28	0.64
$W(3,1)$	0.50	68.92	67.14	119.77	71.08	1.00	1.00	0.70	0.88
	0.75	51.13	53.25	74.54	48.38	1.00	1.00	0.94	0.99
	1.00	42.66	44.74	34.81	37.66	1.00	1.00	0.99	1.00

Table 4.16: Power when the changepoint$= 1$, $n = 300, s = 35$.
Target additive shift, $\Delta = 0.25\sigma$. Nominal FAR, $\gamma = 0.1$

In tables 4.18 and 4.19 below we show some results that support the usefulness of our proposed sanity test under different parametric models and cycle lengths, using the NDEC with an additive shift alternative. To begin, consider table 4.18 with $n = 30$. To derive the results in each row, we simulated a single historical dataset from a corresponding model with

Distribution		Unconditional ADD				Unconditional TAR			
	c	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$	NDEC	PITC	$TC_{0.9}$	$TC_{0.5}$
$N(0,1)$	1.025	187.44	185.23	174.13	215.67	0.29	0.31	0.13	0.11
	1.05	178.82	177.60	171.16	213.81	0.44	0.43	0.16	0.13
	1.075	161.19	160.58	167.72	210.32	0.57	0.59	0.18	0.15
	1.100	144.21	142.45	159.35	208.79	0.73	0.74	0.23	0.16
	1.150	139.67	138.24	151.19	205.20	1.00	1.00	0.26	0.17
$t(3)$	1.025	199.97	197.44	185.52	225.84	0.23	0.24	0.12	0.10
	1.050	184.82	185.69	182.58	222.78	0.31	0.32	0.13	0.10
	1.075	173.32	174.65	179.82	220.75	0.45	0.46	0.15	0.11
	1.100	164.46	162.93	176.11	219.62	0.58	0.59	0.18	0.12
	1.150	159.80	160.52	172.94	218.33	0.91	0.92	0.22	0.13
$t(6)$	1.025	191.51	193.06	178.66	219.61	0.26	0.27	0.12	0.11
	1.050	175.02	176.48	175.37	217.42	0.37	0.39	0.14	0.12
	1.075	167.32	165.73	172.71	215.26	0.51	0.52	0.17	0.12
	1.100	151.52	150.89	166.69	213.84	0.69	0.68	0.21	0.13
	1.150	147.16	145.73	161.94	210.13	0.97	0.98	0.24	0.14
$W(1,1)$	1.025	189.16	187.64	148.83	221.46	0.17	0.18	0.11	0.17
	1.050	178.98	174.10	148.20	218.92	0.23	0.22	0.14	0.23
	1.075	157.44	159.12	147.13	209.77	0.48	0.47	0.16	0.39
	1.100	142.63	141.03	145.84	206.68	0.62	0.60	0.21	0.51
	1.150	125.52	123.79	138.77	197.81	0.74	0.74	0.29	0.74
$W(2,1)$	1.025	184.18	182.93	154.60	218.89	0.29	0.31	0.12	0.22
	1.050	165.65	167.19	150.71	206.91	0.59	0.61	0.18	0.48
	1.075	148.13	147.24	149.73	190.47	0.83	0.82	0.26	0.71
	1.100	135.90	137.78	144.34	171.05	0.96	0.96	0.35	0.88
	1.150	95.31	94.07	141.97	132.25	1.00	1.00	0.53	0.94
$W(3,1)$	1.025	179.82	176.64	150.45	214.51	0.37	0.36	0.15	0.31
	1.050	157.24	155.18	146.88	188.86	0.76	0.75	0.25	0.69
	1.075	122.29	120.90	137.65	158.31	0.95	0.94	0.42	0.82
	1.100	88.13	89.52	128.68	129.67	1.00	1.00	0.56	0.93
	1.150	53.56	52.73	104.67	90.56	1.00	1.00	0.85	1.00

Table 4.17: Power when the changepoint$= 1$, $n = 300$, $s = 35$.
Target multiplicative factor, $c = 1.05$. Nominal FAR, $\gamma = 0.1$

the specified value s and followed the steps of the proposed sanity test. Thus the first four rows in table 4.18 correspond to four historical datasets from $N(0,1)$ with different depths and might be thought of as corresponding to four hypothetical practitioners. We see that the conditional FAR satisfies our feasibility criteria for $s \geq 100$. That is, only the third and fourth practitioners would declare s sufficient and both would be correct, since during our simulation analysis described in subsection 4.1, we have determined that for NDEC under $N(0,1)$ the feasible depth of the

Algorithm 7 : Sanity Test for Adequacy of the Historical Data Depth

1. Simulate 100 datasets of size N by re-sampling with replacement from $\{Y_i\}_{i=1}^{N}$.

2. For each of the 100 re-sampled historical datasets
   (a) determine the threshold H as previously described, depending on the CUSUM.
   (b) simulate $5,000$ cycles of length n by re-sampling with replacement from the original historical dataset $\{Y_i\}_{i=1}^{N}$ and run the corresponding CUSUM through each cycle.
   (c) estimate the conditional FAR as the proportion of $5,000$ cycles that had an alarm.

3. Construct a boxplot of the conditional FAR estimates and declare s to be sufficient if the boxplot meets the feasibility criteria described in section 4.1. If s is not declared sufficient, the practitioner will know that additional cycles of historical data are needed.

historical data when $n = 30$ is $s = 110$. Similarly under $t(3)$ and $n = 30$, we have determined that the minimum satisfactory historical data size is $s = 120$. Using rows 5-8 of table 4.18, we see again that all four practitioners would arrive at the correct conclusion concerning the sufficiency of their s value. And finally, we have found that $s = 100$ is required for $n = 30$ under $Weibull(1, 1)$, and rows 9-12 of table 4.18 show again that all four practitioners would reach the correct conclusion. Table 4.19 mirrors the same findings in table 4.18 for a larger cycle size $n = 300$. Namely, all four practitioners in each distribution scenario would reach the correct conclusion about the sufficiency of their s value.

4.3.2 Computational Time

Every R process in our simulation analysis was performed on a single core of a 4-core, 16 GB memory Intel Xeon CPU, at a processing speed of 2.4 GHz. Each column in table 4.20 indicates the computational time involved with the calculation of the corresponding quantity. We see that the preparation time for calculating the threshold H is the longest for the NDEC procedure, which is especially obvious for $n = 300$ (and $s = 35$); it takes about 3.5 hours to up-

Model	s	False Alarm Rate						
		≤0.06	≤0.065	≤0.075	[0.075,0.125]	≥0.125	≥0.135	≥0.14
$N(0,1)$	60	1	7	18	73	9	5	4
	90	0	0	7	87	6	2	1
	100	0	2	4	91	5	1	0
	120	0	2	4	94	2	0	0
$t(3)$	60	1	3	6	74	20	11	8
	90	2	2	3	85	12	4	2
	100	1	2	3	89	8	2	0
	120	0	1	2	93	5	2	0
$W(1,1)$	60	1	2	7	85	8	3	2
	90	1	2	5	90	5	2	1
	100	0	0	3	93	4	0	0
	120	0	0	2	97	1	0	0

Table 4.18: Choosing the minimal s for $n = 30$ based on NDEC
Target additive change, $\Delta = 0.5\sigma$. Nominal FAR, $\gamma = 0.1$

Model	s	False Alarm Rate						
		≤0.06	≤0.065	≤0.075	[0.075,0.125]	≥0.125	≥0.135	≥0.14
$N(0,1)$	10	11	16	27	67	6	4	2
	20	3	5	18	78	4	2	2
	30	0	2	5	92	3	1	0
	40	0	0	3	94	4	1	0
$t(3)$	10	3	4	5	58	37	27	20
	20	1	2	6	76	18	10	7
	30	0	0	2	88	10	0	0
	40	0	0	2	92	6	2	0
$W(1,1)$	10	2	4	9	81	10	4	2
	20	0	1	3	93	4	1	0
	30	0	0	1	96	3	0	0
	40	0	0	0	98	2	0	0

Table 4.19: Choosing the minimal s for $n = 300$ based on NDEC
Target additive change, $\Delta = 0.5\sigma$. Nominal FAR, $\gamma = 0.1$.

date H in NDEC, while it takes about 8 minutes for PITC and 5 minutes for TC to update their corresponding H. This is easily explainable, since at the stage of calculating H, the implementation of the NDEC involves a composite algorithm of NDE and simulation of cycles through the smoothed bootstrap, whereas both the PITC and the TC are based on uniform random variables. On the other hand, it is interesting to note that in the monitoring stage, the evaluation time of S_i for a given observation is the smallest for the NDEC compared to the PITC and the TC.

(n, s)	Preparation time (min), H NDEC	PITC	TC	Evaluation time (msec), S_i NDEC	PITC	TC
$(30, 120)$	6.03	1.5	0.5	1.2	3	3.4
$(300, 35)$	199	7.8	5	4.0	10	10.3

Table 4.20: Computational Time.

4.4 Summary and Recommendations

We proposed two nonparametric CUSUM algorithms, NDEC and PITC, that utilize historical data to detect a specified change in a finite sequence of iid observations from an unknown distribution. Our proposed procedures depend on several application-dependent parameters. The parameters K and c define the specified size of the additive and a multiplicative change that the practitioner wants to detect. The parameter n is the length of the periodic cycle in the application of interest. The parameter γ is the nominal level of false alarms that a practitioner is willing to tolerate during the monitoring cycle.

We suggested appropriately designed bootstrap algorithms for determining the alarm threshold in each of our methods. We developed a sanity test that can be used by a practitioner to determine if the depth of the historical data he/she has is sufficient for delivering an adequate conditional FAR Our size assessment simulation analysis showed that both the NDEC and the PITC can achieve the unconditional nominal FAR when a reasonably small amount of historical data is available, both for additive multiplicative changes.

We demonstrated that our CUSUM algorithms work successfully with different distributions, including symmetric (such as normal), heavy tail (such as Student's t) and skewed (such as Weibull) distributions. Our power study, using simulated out-of-control data, showed that the NDEC and the PITC are nearly equivalent to each other and have superior performance with re-

spect to the three benchmark procedures the $\text{TC}_{\alpha=0.9}$, the $\text{TC}_{\alpha=0.5}$ and the SRC, in terms of the TAR and the ADD. We also observed that the PITC algorithm outperforms the NDEC in terms of the computational time for H. Thus if the computational time is an issue, one might prefer using the PITC over the NDEC. However, both the NDEC and the PITC are eminently reasonable solutions for applications where nonparametric fault-detection algorithms are needed.

We also note that the SRC method does not require historical data, as oppose to the NDEC, PITC and the TC methods. Therefore our recommendation is to use a combination of two methods as follows: at a startup use SRC until sufficient historical in-control data has been collected, then switch over to use of either the NDEC or the PITC method to have greater power.

Chapter 5

Future Work

In this chapter, we propose an extension of the NDEC formula (3.1) to contexts with multiple timeslots, where the distribution of the incoming observations is the same within a timeslot but varies between different timeslots. We say that such data exhibits *structured non-stationarity*. A typical application where data exhibits structured non-stationarity is network monitoring, also known as, network surveillance systems. In section 5.2, we review relevant literature on CUSUM methods specifically designed for network monitoring. In section 5.3, we introduce a motivational example from network monitoring, and a benchmark nonparametric CUSUM procedure developed for such application. In section 5.4, we show how the NDEC (3.1) can be extended to a multiple timeslot scenario with known underlying densities, and then adapted to the case when the underlying densities are unknown. In section 5.5, we show a small simulation analysis where we assess the size of our proposed algorithm. And finally, in section 5.6, we summarize our findings and outline directions for future work.

5.1 Motivation

Network surveillance systems constantly monitor computer networks to detect *anomalous activities*. Network anomalies can arise due to various causes, including slow or failed components, network overload and network intrusions. Common variables being monitored in a network include traffic throughput, delay measurements, memory usage, active CPU time and response time. Network fault monitoring techniques can be divided into two classes: statistical and non-statistical. The non-statistical approaches are often based on trial and error and rely heavily on the expertise of the network engineers.

We want to explore the following issues that have not been fully addressed. First is the need for nonparametric techniques. As experiments have shown, even with relatively homogeneous time periods, the common parametric distribution families do not describe the variation in a data stream consistently enough to be reliable over long term use. Second is the need to handle time-varying distributions, since a variety of metrics used in the monitoring process often exhibit cyclical nature and trends that should be taken into account. And third is the need to be computationally efficient. Even in small networks the number of monitored metrics can exceed hundreds. The simultaneous and frequent measurements on each of these metrics is a challenging task.

5.2 Related Work

Hajji (2005) addresses the problem of *normal operation baselining* in network monitoring. It is assumed that each variable is a *switching* between different regimes, where each regime is a Gaussian distribution. This finite mixture model is of the form given below:

$$x_i = m_k + \varepsilon_k, \quad k = 1, ..., K$$

$$p(x_i) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp \frac{-(x_i - m_k)^2}{2\sigma_k^2},$$

where $\varepsilon \sim N(0, \sigma_k)$, m_k is the mean of the distribution, K is the number of regimes, and each regime has a mixing probability, π_k. The mains steps of the proposed method are as follows. An online adaptive algorithm produces the underlying parameter estimate, $\hat{\theta}^{n-1}$. As a new data point is sequentially added, the algorithm outputs a refined new estimate, $\hat{\theta}^n$. The *normal behavior* of the network is characterized through the random variable $\hat{\theta}^n - \hat{\theta}^{n-1}$. More specifically, under normal conditions, $\hat{\theta}^n - \hat{\theta}^{n-1}$ is expected to fluctuate around zero, and hence if the difference $\hat{\theta}^n - \hat{\theta}^{n-1}$ drifts systematically over long duration, then a change in the model has occurred. The mean value of this difference is a good indicator of the health of the network. An advantage of Hajji's work is the online parameter estimation. The limitations include the use of an explicit parametric model and not addressing possible non-stationarity of the data.

Hellerstein et al. (2001) describe an approach to detecting and predicting threshold violations in network monitoring process. The model of the *normal behavior* is built in two stages: (1) accounting for the nonstationarity in the mean, and (2) accounting for the time-serial dependencies. More explicitly, let S_{ijkl} be the random variable for observed values in the ith time-of-day, jth day of the week, kth month. Then the following ANOVA-type model accounts for the time-varying behavior of S_{ijkl} in stage 1, with overall mean μ:

$$ln(S_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijkl},$$

where α_i is the ith time-of-day effect, β_j is the jth day effect, γ_k is the kth month effect.

This model has shown to explain about 65% of the variability in the data. The remaining time-serial dependencies in the ε_{ijkl} sequence are removed through an *AR(2)* model in stage 2. The main advantage of this work is its accounting for non-stationarity of the data. The main limitation is the use of a parametric model.

Lambert and Liu (2006) describe a method for monitoring streams of network counts, with long-term trends, rough cyclical patterns, outliers, and missing data. Through detailed empirical analysis, the authors justify the use of negative binomial reference distributions that are parameterized by their means and variances to model the in-control behavior. The moments of the underlying (negative binomial) distribution are interpolated from values on a *coarse time grid*. These grid values capture the cyclical patterns in the data (for example, the grid may consist of 24 hourly values if there is no day of the week effect. Or it may have $24 \times 7 = 168$ hourly values if a day of the week effect is present). These grid values are subsequently updated by EWMA. The combination of the interpolation and EWMA smooth the grid values continuously over time and track both cyclical patterns and long-term trends. The main steps of the method are as follows. The p-values of counts are calculated to adjust for the time-of-day effects and long-term trends. The standardized p-values, $Z_t = \Phi^{-1}(p_t)$, are then plugged into classic EWMA, $S_t = (1-w)S_{t-1} + wZ_t, w \in (0,1]$. An alarm is raised when $S_t > L\sigma_w$, where L is a threshold parameter. The parameters L and w are chosen to control ARL_0 or the probability of false alarm. Important advantages of Lambert's work are handling of the discrete distributions (such as negative binomial distribution) and addressing the non-stationarity of the data. The limitations include assumption of a parametric model and a failure to account for any correlation in the data.

Jeske et al. (2009) and Montes de Oca et al. (2010), considered tracking statistics that automatically adapt to arbitrary contexts. We discuss one of these methods that is most relevant to our context in detail in section 5.3.

5.3 Example Application and Benchmark Procedures

Our research is an extension of the work in the two papers, Jeske et al. (2009) and Montes de Oca et al. (2010), which in turn has been motivated by a collaboration between the *Statistical Consulting Collaboratory* at UCR and *Integrien* (currently operating under *VMware Corporation*), a software company in Irvine, CA. The goal of *Integrien* is to find solutions to identify, predict and prevent problems in the information technology infrastructure. The objective of the research was to develop a robust and efficient change-point detection algorithm that could be applied to *Integrien's* data network, which exhibits structured non-stationary patterns in its traffic streams. More specifically, the distribution of incoming data in *Integrien's* applications has pronounced weekly cycles with heavier traffic during the weekdays compared to the weekends, in addition it varies within the hours of the day. This is well illustrated in figures 5.1 and 5.2 that display hourly estimates of the means and standard deviations of two particular metrics monitored daily (for 23 hours per day) on a server in a large network (Jeske et al. (2009)).

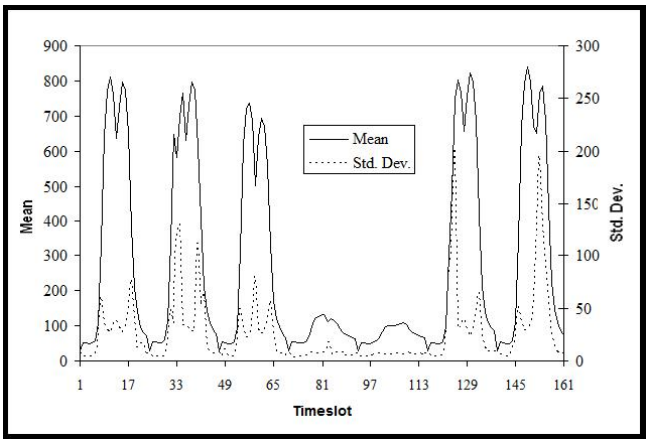

Figure 5.1: Mean and Standard Deviation for Live Sessions

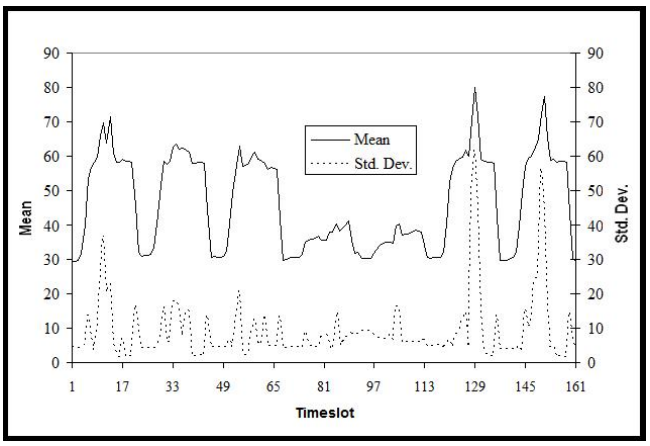

Figure 5.2: Mean and Standard Deviation for Oracle Sessions

Model for handling structured non-stationarity: Jeske et al. (2009) and Montes de Oca et al. (2010) proposed a general model for the data as follows. Each monitoring cycle (for example, a week) is divided into m timeslots (for example, hours), with the assumption that *suitably transformed observations* are independent, and within homogeneous timeslots are also identically distributed according to $\{F_j\}_{j=1}^{m}$. Jeske et al. (2009) illustrate an application-dependent transformation that breaks an existing autocorrelation in the data stream. Historical data available within each timeslot $j \in \{1, 2, ..., m\}$ is given by $\{Y_{jk}\}_{j=1,k=1}^{m,\ n_j}$, where n_j is the number of observations available for timeslot j. The historical data is used to estimate the underlying reference distributions $\{F_j\}_{j=1}^{\infty}$ through the ECDFs $\{\hat{F}_j\}_{j=1}^{\infty}$. For incoming data stream, $\{X_i\}$, the two-sided Transformed CUSUM (TC) designed for detecting upward and downward shifts of underlying distributions in a multiple timeslot scenario is defined as

$$T_i^+ = \max\left\{0, T_{i-1}^+ + \hat{F}_{\tau_i}(X_i) - \alpha\right\}, \quad T_0^+ = 0 \tag{5.1a}$$

$$T_i^- = \max\left\{0, T_{i-1}^- + (1-\alpha) - \hat{F}_{\tau_i}(X_i)\right\}, \quad T_0^- = 0. \tag{5.1b}$$

Here $\tau_i \in \{1, ..., m\}$ corresponds to the timeslot, which incoming observation X_i belongs to.

The advantages of the work Jeske et al. (2009) include the non-parametric extension of the classic CUSUM, the online screening of the data, which makes the algorithm fully automatic, and addressing the non-stationarity of the data through a defined timeslot mechanism. One of the potential disadvantages of the TC is that it is based on the conventional form of the CUSUM . We anticipate that applying an extended version of the optimal CUSUM formula to the same context may increase the efficiency of the algorithm in detecting a true change.

5.4 Proposed Extension of the CUSUM to Multiple Time-Slots

Following Jeske et al. (2009), we propose to split the monitoring cycle into heterogeneous timeslots, within which observations (following an application-dependent transformation) can be considered iid. We also assume that observations between different timeslots are independent. If we denote the in-control and out-of-control densities within timeslot j as f_{0j} and f_{1j} correspondingly, then the multiple timeslot extension of Page's optimal formula (2.3) is given by

$$\sum_{j=1}^{i} \log \frac{f_{1\tau_j}(X_j)}{f_{0\tau_j}(X_j)} - \min_{1\leq k\leq i} \left[0, \sum_{j=1}^{k} \log \frac{f_{1\tau_j}(X_j)}{f_{0\tau_j}(X_j)} \right] > H. \tag{5.2}$$

Below, we show that a recursive form similar to (2.4) can be derived for CUSUM (5.2) in a multiple timeslot scenario, as follows:

$$S_i = \max \left(0, S_{i-1} + \log \frac{f_{1\tau_i}(X_i)}{f_{0\tau_i}(X_i)} \right), \tag{5.3}$$

where as before τ_i is a mapping from i to the timeslot that X_i belongs to.

Proposition 6. *The two CUSUMs given by (5.2) and (5.3) are equivalent.*

Proof. For simplicity, without loss of generality, we assume $n_j = n$. Note the switch in the use of n; in the one timeslot context in chapter 3, n was used to denote the cycle length, whereas in

the current context of multiple heterogeneous timeslots, n indicates the length of a timeslot. We slightly change the notation in (5.3) by considering the following set-up:

- Timeslot 1: $X_{1,1}, ..., X_{1,n}$ with in-control and out-of-control densities, f_{01} and f_{11}.

  ....

- Timeslot m: $X_{m,1}, ..., X_{m,n}$ with in-control and out-of-control densities, f_{0m} and f_{1m}.

Let $Z_{ij} = \log\{f_{1i}(X_{i,j})/f_{0i}(X_{i,j})\}$ for $i = 1, ..., m$ and $j = 1, ..., n$. Then we have:

Timeslot 1: For $i = 1, j = 1$, we have

$$S_1 = \max(0, Z_{1,1}) = \begin{cases} Z_{1,1}, & \text{if } Z_{1,1} \geq 0, \\ 0, & \text{otherwise} \end{cases}$$

$$L_1 = Z_{1,1} - \min(0, Z_{1,1}) = \begin{cases} Z_{1,1}, & \text{if } Z_{1,1} \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, in proposition 1, we showed that $S_n = L_n$, where $S_n = \max(0, S_{n-1} + Z_{1,n})$ and $L_n = \sum_{j=1}^{n} Z_{1,j} - \min_{1\leq k\leq n}(0, \sum_{j=1}^{k} Z_{1,j})$.

Between timeslots 1 and 2: For $i = 2, j = 1$, we get

$$\begin{aligned} S_{n+1} &= \max(0, S_n + Z_{2,1}) \\ &= \max(0, L_n + Z_{2,1}) \\ &= \max\left(0, \sum_{j=1}^{n} Z_{1,j} - \min_{1\leq k\leq n}(0, \sum_{j=1}^{k} Z_{1,j}) + Z_{2,1}\right) \\ &= \begin{cases} \sum_{j=1}^{n} Z_{1,j} + Z_{2,1} - \min_{1\leq k\leq n}(0, \sum_{j=1}^{k} Z_{1,j}), & \text{if } \sum_{j=1}^{n} Z_{1,j} + Z_{2,1} - \min_{1\leq k\leq n}(0, \sum_{j=1}^{k} Z_{1,j}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

$$
\begin{aligned}
L_{n+1} &= \sum_{j=1}^{n} Z_{1,j} + Z_{2,1} - \min\{\min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}), \sum_{j=1}^{n} Z_{1,j} + Z_{2,1}\} \\
&= \begin{cases} \sum_{j=1}^{n} Z_{1,j} + Z_{2,1} - \min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}), & \text{if } \sum_{j=1}^{n} Z_{1,j} + Z_{2,1} \ge \min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}) \\ 0, & \text{otherwise} \end{cases} \\
&= S_{n+1}.
\end{aligned}
$$

Timeslot 2: For $i = 2, j = n$, we have

$$
\begin{aligned}
S_{2n} &= \max(0, S_{2n-1} + Z_{2,n}) = \max(0, L_{2n-1} + Z_{2,n}) \\
&= \max(0, \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n-1} Z_{2,j} - \min\{\min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}), ..., \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n-1} Z_{2,j}\} + Z_{2,n}) \\
&= \begin{cases} \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n} Z_{2,j} - \min\{\min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}), \sum_{j=1}^{n} Z_{1,j} + Z_{2,1}, ..., \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n-1} Z_{2,j}\}, \\ \qquad \text{if } \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n} Z_{2,j} \ge \min\{...\} \\ 0, \qquad \text{otherwise} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
L_{2n} &= \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n} Z_{2,j} - \min\{\min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}), \sum_{j=1}^{n} Z_{1,j} + Z_{2,1}, ..., \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n} Z_{2,j}\} \\
&= \begin{cases} \sum_{j=1}^{n} (Z_{1,j} + Z_{2,j}) - \min\{\min_{1\le k\le n}(0, \sum_{j=1}^{k} Z_{1,j}), \sum_{j=1}^{n} Z_{1,j} + Z_{2,1}, ..., \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n-1} Z_{2,j}\}, \\ \qquad \text{if } \sum_{j=1}^{n} Z_{1,j} + \sum_{j=1}^{n} Z_{2,j} \ge \min\{...\} \\ 0, \qquad \text{otherwise} \end{cases} \\
&= S_{2n}.
\end{aligned}
$$

□

For the case, when the timeslot densities are unknown, we propose to estimate them through NDE, as before. The resulting NDE based generalized CUSUM (NDEGC) is of the form

$$\hat{S}_i = \max\left(0, \hat{S}_{i-1} + \log\frac{\hat{f}_{1\tau_i}(X_i)}{\hat{f}_{0\tau_i}(X_i)}\right). \tag{5.4}$$

As before, we assume that $f_{1\tau_i}(X_i) = f_{0\tau_i}(X_i - K)$ or $f_{1\tau_i}(X_i) = f_{0\tau_i}(X_i/c)/c$ for specified values of K and c, therefore by obtaining $\hat{f}_{0\tau_i}(X_i)$, we simultaneously obtain $\hat{f}_{1\tau_i}(X_i)$. The NDEGC alarms when $\hat{S}_i > H$, where H is calculated through bootstrapping, as outlined in algorithm 8 below.

Algorithm 8 : Determining Threshold H in the NDEGC.

1. Use $\{Y_{jk}\}_{j=1,k=1}^{m,\ n_j}$ to estimate $\{\hat{f}_{0j}\}_{j=1}^{m}$, and correspondingly, $\{\hat{f}_{1j}\}_{j=1}^{m}$, within each timeslot.
2. Simulate B cycles of in-control observations from $\{\hat{f}_{0j}\}_{j=1}^{m}$ via smoothed bootstrap.
3. Run the NDEGC given by (5.4) along each simulated cycle and extract the maximums.
4. Set H to be the $(1-\gamma)$ percentile of the ordered B maximum values found in step 3.

5.5 Simulation Analysis

To test the performance of our NDEGC, we ran a Monte-Carlo simulation analysis by fitting Weibull densities to the timeslot distributions, using the parameter estimates corresponding to figure 5.1 in our motivational example in section 5.3. Following the setup of this example, we considered a week as a cycle with $m = 161$ hours, representing timeslots. Within each timeslot, observations have been collected every two minutes, resulting in $n_j = n = 30$ observations per timeslot. We used the same feasibility criteria defined in section 4.1 to guide

us in a proper choice of historical data depth, s_j, for timeslot j. For simplicity, we assumed that $s_j = s$ for all $j \in \{1, ..., m\}$. We used the method of moments to match parameters (α_j, β_j) of the Weibull distribution to the estimated means ans standard deviations, $(\bar{x}_j, s_j)$ from figure 5.1, for $j = 1, ..., m$. Thus, $\hat{\alpha}_j$ is the solution to the following equation

$$\frac{\Gamma(1 + 2/\alpha_j)}{\Gamma^2(1 + 1/\alpha_j)} - 1 - \frac{s_j^2}{\bar{x}_j^2} = 0,$$

and

$$\hat{\beta}_j = \frac{\bar{x}_j}{\Gamma(1 + 1/\hat{\alpha}_j)}.$$

Figure 5.3 below illustrates the conditional FAR for $n = 30, s = 30$ and $m = 161$ for the fitted Weibull densities, when we control FAR during the week at level $\gamma = 0.1$, and consider a multiplicative change of size $c = 1.05$. In our simulations, we set $B = 1,000$ both in the stage of calculation of H and in the evaluation of the conditional FAR.

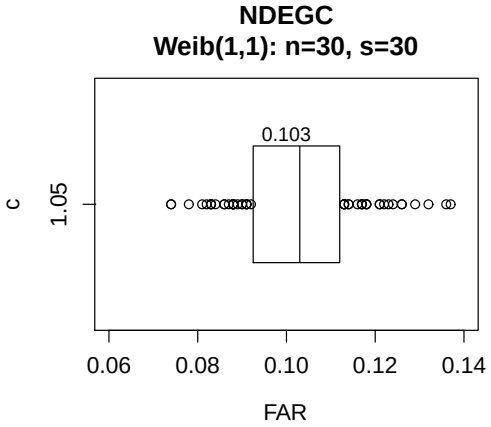

Figure 5.3: Conditional distribution of the FAR of NDEGC based on 100 historical datasets from $Weibull(1, 1)$ for a multiplicative change alternative. Nominal FAR is $\gamma = 0.1$.

We observe that the number of historical cycles needed to estimate the underlying unknown densities in each timeslot in order satisfy the feasibility criteria is significantly lower

compared to the case of one timeslot. Recall from section 4.1 that in a one timeslot scenario, for a cycle length $n = 30$, we needed $s = 120$ historical cycles to achieve the feasibility criteria, whereas in case of the multiple timeslots, for a timeslot with length $n = 30$, a value of $s = 30$ is satisfactory. We need to conduct more simulations in order to gain more insight in the nature of this phenomenon.

5.6 Summary and Proposed Direction

To summarize our findings from sections 5.4 and 5.5, it is feasible to extend our NDEC procedure to applications with multiple timeslot data such as the *Integrien's* data. Our simulation analysis suggests that the NDEC effectively controls the nominal FAR during a monitoring cycle that consists of heterogeneous timeslots. Moreover, our findings suggest that the required number of historical cycles needed to estimate the underlying unknown densities in each timeslot may be far smaller in the case of multiple timeslots than in the case of one timeslot. Our fault-detection simulation analysis using Weibull densities, showed that NDEC detects multiplicative changes of different sizes with outstandingly hight TAR and with very low ADD. These promising results combined with our findings in chapter 3 suggest that the NDEC might be highly competitive among other benchmark procedures designed for network surveillance, such as the TC. We need to investigate more scenarios, including other distributions and other types of out-of-control changes, such as, for example an additive change. We hope that our future simulation analyses will shed more light on the necessary amount of historical data needed for NDEC in a multiple timeslot scneario and, in particular, whether there is a relation between the size of the historical data in a multiple and a one timeslot cases.

Bibliography

I.S. Abramson. On bandwidth variation in kernel estimates-a square root law. *Annals of Statistics*, 10(4):1217 – 1223, 1982.

R. W. Amin, Jr. Reynolds, M. R., and S. T. Bakir. Nonparametric quality control charts based onthe sign statistic. *Communications in StatisticsTheory and Methods*, 24:1597 – 1623, 1995.

M. Bagshaw and R.A. Johnson. The effect of serial correlation on the performance of cusum tests ii. *Technometrics*, 17(1):73 – 80, 1975.

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

M. Basseville and I.V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

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

P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, NY, 1968.

A.F. Bissell. Cusum techniques for quality control (with discussion). *Applied Statistics*, 18:1 – 30, 1969.

A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353 – 360, 1984.

A.W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis:The Kernel Approach with S-Plus Illustrations*, volume 18. Oxford University Press, Oxford, UK, 1997.

A.W. Bowman, P. Hall, and D.M. Titterington. Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, 71:341 – 351, 1984.

B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, 1993.

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

S. Chakraborti, P. van der Laan, and S.T Bakir. Nonparametric control charts: An overview and some results. *Journal of Quality Technology*, 33(3):304 – 315, 2001.

C.W. Champ and S.E. Rigdon. A comparison of the markov chain and the integral equation approaches for evaluating the run length distribution of quality control charts. *Communications in Statistics - Simulation and Computation*, 20(1):191 – 204, 1991.

S. Chatterjee and P. Qiu. Distribution-free cumulative sum control charts using bootstrap-based control limits. *The Annals of Applied Statistics*, 3(1):349 – 369, 2009.

S.V. Crowder. Design of exponentially weighted moving average schemes. *Journal of Quality Technology*, 21:155 – 168, 1989.

L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, Inc., 1985.

W.D. Ewan and K.W. Kemp. Sampling inspection of continuous processes with no autocorrelation between successive results. *Biometrika*, 47:363 – 380, 1960.

T. S. Ferguson. *Mathematical Statistics*. Academic, 1967.

L.T. Fernholz. Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics*, 18(3):255 – 262, 1991.

E Fix and J.L. Hodges. Discriminatory analysis, nonparametric estimation: consistency properties. Report No. 4 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

J. Fox and J.S. Long. *Modern Methods of Data Analysis*. Sage Publications, Newbury Park, CA, 1990.

M. Frisen. Statistical surveillance. optimality and methods. *International Statistical Review*, 71 (2):403 – 434, 2003.

M.J. Fryer. Some errors associated with the nonparametric estimation of density functions. *Journal of the Institute of Mathematics and Its Applications*, 18:371 – 380, 1976.

M.A. Girshick and H. Rubin. A bayes approach to a quality control model. *Annals of Mathematical Statistics*, 23(1):114 – 126, 1952.

C.W. Goel and S. E. Rigdon. Determination of a.r.l. and a contour nomogram for cusum charts to control normal mean. *Technometrics*, 13:221 – 30, 1991.

L. Gordon and M. Pollak. An efficient sequential nonparametric scheme for detecting a change of distribution. *The Annals of Statistics*, 22(2):763 – 804, 1994.

L. Gordon and M. Pollak. A robust surveillance scheme for stochastically ordered alternatives. *The Annals of Statistics*, 23:1350 – 1375, 1995.

L. Gordon and M. Pollak. Average run length to false alarm for surveillance schemes designed with partially specified pre-change distribution. *The Annals of Statistics*, 25:1284 – 1310, 1997.

P. Hackl and J. Ledolter. A control chart based on ranks. *Journal of Quality Technology*, 23: 117–124, 1991.

H. Hajji. Statistical analysis of network traffic for adaptive faults detection. *IEEE Transactions on Neural Networks*, 2005.

P. Hall. Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, 11:1156–1174, 1983.

D. Hawkins. Evaluation of average run lengths of cumulative sum charts for an arbitrary data distribution. *Communications in Statistics - Simulation and Computation*, 21, 1992.

D. M. Hawkins and D. H. Olwell. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, 1998.

T. Hayfield and J.S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27, 2008.

J. Hellerstein, F. Zhang, and P. Shahabuddin. A statistical approach to predictive detection. *Elsevier Computer Networks*, 2001.

P.D. Hill. Kernel estimation of a distribution function. *Communications in Statistics, Theory and Methods*, 14:605–620, 1985.

J. L. Hodges and E. L. Lehmann. The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, 27:435–475, 1956.

J. S. Hunter. The exponentially weighted moving average. *Journal of Quality Technology*, 18: 203–210, 1986.

D.R. Jeske, V. Montes de Oca, W. Bischoff, and M. Marvasti. Cusum techniques for timeslot sequences with applications to network surveillance (2009), computational statistics and data analysis. *Computational Statistics and Data Analysis*, 53:4332 – 4344, 2009.

N.L. Johnson and F.C. Leone. Cumulative sum control charts: Mathematical principles applied to their construction and use. part i. *Industrial Quality Control*, 18(12):15 – 21, 1962a.

N.L. Johnson and F.C. Leone. Cumulative sum control charts: Mathematical principles applied to their construction and use. part ii. *Industrial Quality Control*, 19(1):29 – 36, 1962b.

N.L. Johnson and F.C. Leone. Cumulative sum control charts: Mathematical principles applied to their construction and use. part iii. *Industrial Quality Control*, 19(2):22 – 28, 1962c.

K.W. Kemp. The use of cumulative sums for sampling inspection schemes. *Applied Statistics*, 11:16 – 31, 1962.

K.W. Kemp. A simple procedure for determining upper and lower limits for the average sample run length of a cumulative sum scheme. *Journal of the Royal Statistical Society, Series B, Methodological.*, 29:263 – 265, 1969.

R.A. Khan. A sequential detection procedure and the related cusum procedure. *The Indian Journal of Statistics.*, 40(3/4):146 – 162, 1979.

H Kim, B.L. Rozovskii, and A.G. Tartakovsky. A nonparametric multichart cusum test for rapid detection of dos attacks in computer networks. *International Journal of Computing & Information Sciences*, 2(3):149 – 158, 2004.

S.-H. Kim, B.L. Nelson, and J.R. Wilson. Some almost-sure convergence properties useful in sequential analysis. *Sequential Analysis*, 24:411 – 419, 2005.

S.-H. Kim, C. Alexopoulos, K.-L. Tsui, and J.R. Wilson. A distribution-free tabular cusum chart for autocorrelated data. *IIE Transactions*, 39:317 – 330, 2007.

A.N. Kolmogorov, Y.V. Prokhorov, and A.N. Shiryaev. Probabilistic-statistical methods of detecting spontaneously occurring effects. *Proceedings of the Steklov Institute of Mathematics.*, pages 1 – 21, 1990.

T.L. Lai. Sequential changepoint detection in quality-control and dynamical-systems. *Journal of the Royal Statistical Society, series B.*, 57:613 – 658, 1995.

D. Lambert and C. Liu. Adaptive thresholds: Monitoring streams of network counts. *American Statistical Association*, 2006.

J.C. Liechty, D. K.J. Lin, and J.P. McDermott. Single-pass low-storage arbitrary quantile estimation for massive datasets. *Statistics and Computing*, 13:91 – 100, 2003.

R.Y. Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, 90:1380 – 1388, 1995.

G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42:1897 –1908, 1971.

J.M. Lucas and R.B. Crosier. Fast initial response for cusum quality control schemes: Give your cusum a head start. *Technometrics*, 24(3):199 –205, 1982.

D. McDonald. A cusum procedure based on sequential ranks. *Naval Research Logistics*, 37: 662 –646, 1990.

V. Montes De Oca. Nonparametric cusum algorithms with applications to network surveillance. In *PhD Dissertation, Department of Applied Statistics, UCR*, 2008.

V. Montes de Oca, D.R. Jeske, Q. Zhang, C. Rendon, and M. Marvasti. A cusum change-point detection algorithm for non-stationary sequences with application to network surveillance. *Journal of Software Systems*, 83:1288 –1297, 2010.

D.C. Montgomery. *Introduction to Statistical Process Control*. John Wiley & Sons, Inc., 6 edition, 1991.

G. V. Moustakides. Optimal stopping times for detecting changes in distributions. *Annals of Mathematical Statistics*, 14:1379 –1387, 1986.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Applications*, 10:186 –190, 1964.

A. Pagan and A. Ullah. *Nonparametric Econometrics*. Cambridge University Press, New York, 1999.

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

E.S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42: 523 –526, 1955.

E.S. Page. On problem in which a change in a parameter occurs at an unknown point. *Biometrika*, 44:248 –252, 1957.

E. A. Jr. Parent. Sequential ranking procedures. In *Doctoral Dissertation, Stanford University*, 1965.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065 –1076, 1962.

M. Pollak. Optimal detection of a change in distribution. *The Annals of Mathematical Statistics*, 13:206 –227, 1995.

B.L.S. Prakasa Rao. *Nonparametric Functional Estimation*. Academic Press, 1983.

R.D. Reiss. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8:116 – 119, 1981.

Y. Ritov. Decision theoretic optimality of the cusum procedure. *Annals of Statistics*, 18:1464 – 1469, 1990.

S.W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1:239 – 250, 1959.

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

K.V. Rohatgi and A.K. E. Saleh. *An introduction to Probability and Statistics*. John Wiley & Sons, Inc., 2 edition, 2001.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832 – 837, 1956.

M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9:65 – 78, 1982.

I.H. Salgado-Ugarte and M.A. Perez-Hernandez. Exploring the use of variable bandwidth kernel density estimators. *Stata Journal*, 3(2):133–147, 2003.

I.H. Salgado-Ugarte, M. Shimizu, and T. Taniuchi. Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin Reprints*, 3:155–173, 1993.

I.H. Salgado-Ugarte, M. Shimizu, and T. Taniuchi. Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin Reprints*, 5:172–190, 1995.

W.A. Shewhart. *Economic Control of quality of manufactured product*. D.van Nostrand, N.Y., 1931.

A.N. Shiryaev. On optimal methods in quickest detection problems. *Theory of Probability and Its Applications*, 8:22–46, 1963.

D. Siegmund. *Sequential Analysis*. Springer-Verlag, 1985.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1986.

J.S. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.

C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, 12:1285 – 1297, 1984.

C.S. Van Dobben De Bruyn. *Cumulative Sum Tests: Theory and Practice*. Griffin's Statistical Monographs and Courses, London: Griffin, 1968.

A. Wald. *Sequential Analysis*. John Wiley & Sons, Inc., 1947.

W. Wertz. *Statistical Density Estimation: A Survey*. Göttingen: Vandenhoeck and Ruprecht., 1978.

T.R. Willemain and G.C. Runger. Designing control charts using an empirical reference distribution. *Journal of Quality Technology*, 28:31 – 38, 1996.

W.H. Woodall. The distribution of the run length of one-sided cusum procedures for continuous random variables. *Technometrics*, 25:295 – 301, 1983.

W.H. Woodall. On the markov chain approach to the two-sided cusum procedure. *Technometrics*, 26:41 – 46, 1984.

W.H. Woodall and D.C. Montgomery. Research issues and ideas in statistical process control. *Journal of Quality Technology*, 29:172 – 183, 1999.

E. Yashchin. Analysis of cusum and other markov-type control schemes by using empirical distributions. *Technometrics*, 34:54 – 63, 1992.

E. Yashchin. Statistical control schemes-methods, applications and generalizations. *International Statistical Review*, 61:41 – 66, 1993.

S. Zacks. Survey of classical and bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation. *Recent Advances in Statistics*, pages 245 – 269, 1983.