Stochastic methods for measurement-based network control
Ellens, W.

Citation for published version (APA):
Part

I

Methods for

Change Point Detection
A review of some existing methods

Outline. This chapter describes and discusses several methods for change point detection. As explained in Section 1.1, a change point in a time series (a dataset with a time component) is a moment at which the underlying probability distribution changes. This means that we focus on data that evolve over time, showing sudden (sharp) and insistent changes. We discuss some known methods and concepts, which will be used in the following chapters of this part. The celebrated cumulative sum (CUSUM) method plays a leading role in this chapter, since it forms the basis for the new change point detection methods developed in Chapter 3.

The content of this chapter is as follows. Section 2.1 introduces the problem and the performance metrics (evaluation criteria). Section 2.2 discusses the CUSUM (cumulative sum) method. CUSUM is a parametric method (a method that assumes an underlying distribution which is known to the observer) which can be used to detect changes in mean, variance, or the full distribution. Section 2.3 describes two non-parametric methods (also known as distribution-free methods as no knowledge of the underlying distribution is assumed) to detect a change in mean. The chapter is concluded by a discussion section comparing the described detection methods (Section 2.4).
Chapter 2 A review of some existing methods

2.1 The change point detection problem

In this section we give a mathematical formulation of the change point detection problem. In order to be able to assess the performance of the methods described in the following sections, we also describe the evaluation criteria (the detection speed and the false alarm rate) in mathematical terms.

2.1.1 Problem description

Let us recapitulate the main terms and concepts as explained in Sections 1.1 and 1.3. We assume that the data are observed sequentially, i.e. we analyse the observations up to this moment, in every time step adding one more observation. At each time we ask ourselves, given the set of observations until that moment, whether a change point has occurred or not. In this way we hope to detect a potential change as soon as possible. In general, change point detection methods monitor some test statistic which is based on the observations issue an alarm if this test statistic exceeds a certain threshold, such that the calculated probability of not detecting a change point is kept below a certain predefined value $\alpha$, e.g. $\alpha = 5\%$. Statistically speaking, change point detection methods perform a hypothesis test for every time step. In this section we state this hypothesis test.

We consider a sequence of observations $X = (X_1, X_2, \ldots)$, during which potentially a change point occurs. In probabilistic terms such a change point, to be considered as a change in the statistical law of the underlying random variable, can be described as follows.

**Definition 2.1.** Suppose there is a $k > 0$ such that the $X_i$ are i.i.d. realisations of a random variable with density $f(\cdot)$ for $i = 1, \ldots, k - 1$, while the $X_i$ are i.i.d. with a different density $g(\cdot)$ for $i \geq k$. In this case we call $k$ a change point.

At a point in time $n = 1, 2, \ldots$ we check whether a change point has occurred at some time $k \leq n$, by evaluating $X_n := (X_1, \ldots, X_n)$; if not

---

1 There are also results for a change point detection in dependent data, see for example Chapter 3 and [27, Chapter II].
2.1 The change point detection problem

we continue by evaluating $X_{n+1} := (X_1, \ldots, X_{n+1})$, etc. In terms of hypothesis testing: at time $n$ we want to decide between two hypotheses:

- Under the *null-hypothesis* ($H_0$) the $X_i$ ($i = 1, \ldots, n$) are i.i.d. realisations of a random variable with density $f(\cdot)$.

- Under the alternative hypothesis ($H_1$) there is a $1 \leq k \leq n$ such that up to $k - 1$ the observations are i.i.d. samples from a distribution with density $f(\cdot)$, while from observation $k$ on they are i.i.d. with a *different* density $g(\cdot)$.

In other words: under the null-hypothesis there has *not* been a change point, while under the alternative hypothesis the process changes at some time $k$. As mentioned in Section 1.3 this setup is not a simple binary hypothesis testing problem, but a *multiple-hypotheses test*, as the alternative is essentially a *union* of hypotheses. More precisely: if $H_1(k)$ corresponds to having a change point at $k$, we can write $H_1$ as the union of the $H_1(k)$, with $k = 1, \ldots, n$.

### 2.1.2 Performance metrics

In statistics it is common to evaluate a hypothesis test by considering the probability of so-called type I and type II errors. As we perform a hypothesis test at each time step $n$ (as long as no change point has been detected), the performance of the test at time $n$ can be quantified by these error probabilities. They are defined as follows.

**Definition 2.2.**

- A *type I error* (or false positive/false alarm) occurs if we decide to reject $H_0$ while it is actually true, in other words, if we detect a change point that has not occurred. The *type I error probability* is the probability that we reject $H_0$ under $H_0$.

- A *type II error* (or false negative) occurs if we decide to accept $H_0$ while it is not true, in other words, if we miss a change point. The *type II error probability* is the probability that we accept $H_0$ under $H_1$. In applications the term detection probability is frequently used, which is the probability of a *true positive*, that is, the probability
Chapter 2 A review of some existing methods

to reject H₀ under H₁. It is thus equal to one minus the type II error probability.

Now consider a change point detection method for the sequence X as a whole, that is, a stopping rule that issues an alarm at time τ ≥ 1, defined as the first time that we decide to reject H₀. On the one hand τ should occur soon after the change point k, on the other hand, the rate of false alarms should be low. Mathematically, this can be formulated as keeping the distribution of τ − k stochastically small, given that the change point takes place at k (i.e. under H₁(k)), whereas the distribution of τ should be stochastically large in case there is no change point (i.e. under H₀). The criterion to manage the trade-off between detecting a change point early and to control the number of false alarms is translated in more formal terms as follows:

- minimise sup_{k≥1} E_k(τ − k + 1 | τ ≥ k) (meaning that the detection delay is as small as possible),

- while at the same time making sure that E₀(τ) tends to be ‘large’ (meaning that in case there is no change point, there is a strong tendency to not issue an alarm).

In this part E_k and P_k stand for expectation and probability, under H₀ for k = 0 and under H₁(k) for k ≥ 1.

2.2 A parametric method

The method described in this section, known as CUSUM, has been proposed to identify a change in distribution. The method assumes that the densities f(·) and g(·) are known. First, in Section 2.2.1 we introduce the method, roughly following the setup presented by Siegmund in [91, Chapter II.6]. Then, in Section 2.2.2, we describe the approach of Bucklew [28, Chapter VI.E] to approximate the probability of a false alarm using LD (large deviations) theory.

2.2.1 The CUSUM method

Let us consider the common likelihood ratio test for H₀ versus H₁(k) (with H₀ and H₁(k) as defined in Section 2.1.1). This test raises an alarm
if the *likelihood ratio*

\[
\bar{S}_k := \frac{P_k(X_n)}{P_0(X_n)} = \prod_{i=1}^{n} \frac{P_k(X_i)}{P_0(X_i)} = \prod_{i=k}^{n} \frac{g(X_i)}{f(X_i)}
\]

exceeds a certain value \( \bar{b} > 1 \). It turns out, though, that it is more practical to work with the corresponding *log-likelihood*:

\[
S_k := \log \bar{S}_k = \sum_{i=k}^{n} \log \frac{g(X_i)}{f(X_i)}.
\]

To deal with the fact that \( H_1 \) equals the union of the \( H_1(k) \), we have to verify whether there is a \( k \in \{1, \ldots, n\} \) such that \( S_k \) exceeds a certain threshold \( b > 0 \). As a result, the statistic for the *composite test* (that is, \( H_0 \) versus \( H_1 \)) is

\[
t_n := \max_{k \in \{1, \ldots, n\}} S_k.
\]

Using this test statistic, one could decide to reject \( H_0 \) at time \( n \) if \( t_n \geq b \), for some value \( b > 0 \) that needs to be selected (in a way that properly balances the false alarm rate and detection speed). This leads to the following detection method.

**Method 2.1.** For a given threshold \( b > 0 \) and a sequence \( X = (X_0, X_1, \ldots) \) the *CUSUM method* raises an alarm at time \( \tau \), defined as

\[
\tau := \inf \left\{ n : \max_{k \in \{1, \ldots, n\}} \sum_{i=k}^{n} \log \frac{g(X_i)}{f(X_i)} \geq b \right\}.
\]

Evidently, no alarm is raised if the sequence \((\max_{k \in \{1, \ldots, n\}} S_k)_{n \in \mathbb{N}}\) never exceeds \( b \). The interpretation is that we issue an alarm if, informally speaking, there is a point \( k \) such that is more likely that the observations from \( k \) on originate from a distribution with density \( g(\cdot) \), than from a distribution with density \( f(\cdot) \).
Chapter 2 A review of some existing methods

The test statistic $t_n$ can be rewritten as

$$t_n = T_n - \min_{k \in \{1, \ldots, n\}} T_{k-1}. \quad (2.1)$$

in terms of the cumulative sums $T_k := \sum_{i=1}^{k} Y_i$, with increments that are distributed as $Y_i := \log(g(X_i)/f(X_i))$. This alternative expression explains the name of the test. The statistic (2.1) can be seen as the height of the random walk $T_k$ (with $T_0 = 0$ and $T_k = T_{k-1} + Y_k$) with respect to the minimum that was achieved so far (among $T_0, \ldots, T_{k-1}$).

We continue by deriving expressions for $\mathbf{E}_0(\tau)$ (the expected time before an alarm is raised when no change occurs) and

$$\sup_{k \geq 1} \mathbf{E}_k(\tau - k + 1 \mid \tau \geq k)$$

(the expected time between the change point and its detection). For this derivation it is important to notice that $t_n$ has a regeneration point at $j$ if $T_j = \min_{k \in \{0, \ldots, j\}} T_k$. This means that the process $t_n$ ‘forgets’ its past after time $j$, i.e. for $n > j$ it holds that $t_n$ does not depend on the values $Y_i$ for $i \leq j$, more precisely

$$t_n = \hat{T}_n - \min_{k \in \{j+1, \ldots, n\}} \hat{T}_{k-1},$$

with $\hat{T}_k = \sum_{i=j+1}^{k} Y_i$. The concept of a regeneration point is illustrated in Figure 2.1. We see that the process $t_n$ is positive at time $j$ or otherwise $j$ is a minimum so far and thus a regeneration point (at which the process is ‘reset’).

We have that

$$\mathbf{E}_k(\tau - k + 1 \mid \tau \geq k) \leq \mathbf{E}_1(\tau), \quad (2.2)$$

with equality if $t_n$ has a regeneration point at $k - 1$. If $k - 1$ is not a regeneration point ($t_{k-1}$ is positive), $t_n$ (for $n \geq k$) becomes larger and $\tau$ smaller. It follows from (2.2) that

$$\sup_{k \geq 1} \mathbf{E}_k(\tau - k + 1 \mid \tau \geq k) = \mathbf{E}_1(\tau).$$
2.2 A parametric method

Hence, to assess the performance of the test, we only have to analyse $E_0(\tau)$ and $E_1(\tau)$.

These quantities can be evaluated as follows. Let us start by analysing $E_0(\tau)$. Define $M_0 := 0$ and define (recursively) $M_k$ as the first number $n$ after $M_{k-1}$ that $t_n$ either becomes negative (meaning that a minimum is achieved), or exceeds $b$ (meaning that an alarm is issued) and set $N_k := M_k - M_{k-1}$. Due to the underlying regenerative structure and the fact that all $X_i$ are i.i.d., the $N_k$ are i.i.d. as well (distributed as a random variable $N$); the number of $N_k$ needed to exceed $b$ is geometrically distributed with success probability $P_0(T_N \geq b)$. It is now an immediate consequence of Wald’s equation that

$$E_0(\tau) = \frac{E_0(N)}{P_0(T_N \geq b)};$$

and, analogously, $E_1(\tau) = E_1(N)/P_1(T_N \geq b)$.

These expressions nicely illustrate the trade-off between timely detection and an increased rate of false alarms. As both expressions are increasing in $b$, the lower $b$, the faster the test reacts to a change (the smaller $E_1(\tau)$), but the larger the number of false alarms (the smaller $E_0(\tau)$). One could for instance pick the lowest $b$ for which $E_0(\tau)$ remains above a given (large) threshold $B$.

In [91, Chapter II.6] approximations for $E_0(N)$, $P_0(T_N \geq b)$, $E_1(N)$ and $P_1(T_N \geq b)$ in terms of $b$, $f(\cdot)$ and $g(\cdot)$ are given by considering the sequential probability ratio test with a binary hypothesis ($H_0$ versus $H_1$).
Chapter 2 A review of some existing methods

H₁(1), test statistic Tₙ, lower bound 0 and upper bound b). As E₀(τ) and E₁(τ) are in general hard to compute in explicit terms, one could opt for approximating them. In our book chapter [1] we pointed out how to approximate the expectation E₀(τ) and the false alarm probability P₀(tₙ ≥ b), relying on a Brownian approximation (whose applicability is justified by the central limit theorem). In Section 2.2.2 we present an alternative approach (also discussed in the same book chapter) to approximate the false alarm probability based on LD asymptotics [28, 38, 69]. In Chapter 3 we extend this approach to dependent data sequences.

2.2.2 Large deviations approximation of \textsc{cusum}

We now discuss a way to approximate the false alarm probability, roughly following the setup of [28, Chapter VI.E]. A key step is that we scale the threshold b by n and focus on asymptotics of the probability of issuing a false alarm at time n, that is, P₀(tₙ ≥ nb) for large n. This probability can be rewritten as

\[ P₀(tₙ ≥ nb) = P₀(∃k ∈ \{1, \ldots, n\}: T_k ≥ nb). \]

Since, due to reversibility arguments, we have

\[ tₙ = Tₙ - \min_{k ∈ \{1, \ldots, n\}} T_{k-1} = \max_{k ∈ \{1, \ldots, n\}} (Tₙ - T_{k-1}) \]

\[ \frac{d}{k ∈ \{1, \ldots, n\}} \max T_{n-k+1} = \max_{k ∈ \{1, \ldots, n\}} T_k. \]

Due to \( n^{-1} \log n \to 0 \) and

\[ \max_{k ∈ \{1, \ldots, n\}} P₀(T_k ≥ nb) ≤ P₀(∃k ∈ \{1, \ldots, n\}: T_k ≥ nb) \]

\[ ≤ n \cdot \max_{k ∈ \{1, \ldots, n\}} P₀(T_k ≥ nb), \]

we have the following expression for the so-called decay rate

\[ \lim_{n → ∞} \frac{1}{n} \log P₀(tₙ ≥ nb) = \max_{λ ∈ (0,1]} \lim_{n → ∞} \frac{1}{n} \log P₀\left(\frac{T_{λn}}{n} ≥ b\right) \]

30
2.2 A parametric method

(realise that $\lambda n$ is not necessarily integer, so there is mild abuse of notation in the previous display); in words, this means that the decay rate of the union of all $n$ events coincides with the decay rate of the most likely event among these (the so-called principle of the largest term; see [44]). Relying on Cramér’s theorem [28, Chapter II.A], we can rewrite the above decay rate to

$$
\lim_{n \to \infty} \frac{1}{n} \log P_0(t_n \geq nb) = \max_{\lambda \in (0, 1]} \lim_{n \to \infty} \frac{\lambda}{\lambda n} \log P_0 \left( \frac{T_{\lambda n}}{\lambda n} \geq \frac{b}{\lambda} \right)
$$

$$
= \max_{\lambda \in (0, 1]} \left( -\lambda \sup_{\theta} \left( \frac{b}{\lambda} - \log M(\theta) \right) \right);
$$

here $M(\theta)$ is the moment generating function of $\log(g(X_i)/f(X_i))$ under $H_0$:

$$
M(\theta) := \mathbb{E}_0 \left( e^{\theta \log \frac{g(X_i)}{f(X_i)}} \right)
$$

$$
= \mathbb{E}_0 \left( (\frac{g(X_i)}{f(X_i)})^\theta \right) = \int_{-\infty}^{\infty} (g(x))^\theta (f(x))^{1-\theta} dx.
$$

We can then set the threshold $b$ such that the decay rate under study equals some predefined (negative) constant $-\gamma$ (where $\gamma > 0$).

In principle, however, there is no need to take a constant threshold $b$; we could pick a function $b(\lambda)$ instead. It can be seen that, in terms of minimising the type II error (i.e. not detecting a change point), it is optimal to choose the function $b(\lambda)$ such that

$$
\lim_{n \to \infty} \left( \frac{1}{n} \log P_0 \left( \frac{T_{\lambda n}}{n} \geq b(\lambda) \right) \right) = -\lambda \sup_{\theta} \left( \frac{\theta b(\lambda)}{\lambda} - \log M(\theta) \right) \tag{2.4}
$$

is constant in $\lambda \in (0, 1]$ (and equalling $-\gamma$). Intuitively, this choice entails that (in the limit) for any potential change point $\lambda n$, issuing an alarm (which is done if $T_n - T_{\lambda n-1}$ exceeds $nb(1 - \lambda + 1/n)$, because of the ‘time reversal’ in (2.3)) is essentially equally likely if there is actually no change point. Since the (limiting) false alarm probability only depends on the most probable of these events, in order to maximise the detection
Chapter 2 A review of some existing methods

probability while keeping the false alarm probability fixed, one should increase the thresholds for the other events such that their probabilities become equal to that of the most probable event.

As an illustration, let us identify the function \( b(\cdot) \) in a specific example. Let \( f(\cdot) \) correspond with the standard normal density, while \( g(\cdot) \) corresponds with a normal density with mean \( \mu \) and unit variance. It is seen that

\[
M(\theta) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta(x-\mu)^2}{2}} e^{-\frac{1-\theta}{2} x^2} = e^{-\frac{1}{2} \theta(1-\theta)\mu^2}.
\]

As a consequence, we need to derive \( b \) from

\[
\lim_{n \to \infty} \left( \frac{1}{n} \log P_0 \left( \frac{T_{\lambda n}}{n} \geq b(\lambda) \right) \right) = \theta_b b - \frac{\lambda}{2} \theta_b (1 - \theta_b) \mu^2 = \gamma,
\]

with \( \theta_b = b/(\lambda \mu^2) + \frac{1}{2} \), which amounts to solving a quadratic equation.

2.3 Two non-parametric methods

The CUSUM technique described in the previous section can be considered as an example of parametric statistics: it is assumed that the data stem from a type of probability distribution which is known in advance. This section is about non-parametric statistics, in which there are no a priori assumptions on data belonging to a particular distribution. Two non-parametric methods introduced in Section 2.3.1 and Section 2.3.2 both can be used to detect changes in the mean value. The first is a non-parametric version of CUSUM, we call the second the method of Brodsky-Darkhovsky. The latter method is used in Chapter 5.

2.3.1 A non-parametric version of CUSUM

A non-parametric alternative to the CUSUM technique that we described in Section 2.2.1 is to replace the increments \( Y_i = \log \left( f(X_i)/g(X_i) \right) \) by the \( X_i \) themselves, see e.g. 27 Chapter IV.2].
2.3 Two non-parametric methods

**Method 2.2.** *Non-parametric CUSUM* issues an alarm as soon as

\[
\tilde{U}_n := \max_{k \in \{1, \ldots, n\}} \sum_{i=k}^{n} X_i
\]

exceeds a given threshold \(c\).

The test statistic \(\tilde{U}_n\) can be written in terms of the cumulative sums \(\tilde{T}_k := \sum_{i=1}^{k} X_i\):

\[
\tilde{U}_n = \tilde{T}_n - \min_{k \in \{1, \ldots, n\}} \tilde{T}_{k-1}.
\]

The test is used to detect a change in mean from a distribution with \(\mu \leq 0\) to a distribution with \(\mu > 0\) (if the mean before the change is non-zero, it is subtracted from all observations). The idea behind the method is that the expectation of \(\tilde{T}_k\) is zero or less for all \(k \in \{1, \ldots, n\}\) if no change point has occurred, while the expectation of \(\tilde{T}_k\) grows with \(n\) if at time \(k\) a change point has taken place; therefore \(\tilde{T}_k\) must cross \(c\) for some time \(n\) (at time \(c/\mu\) on average).

### 2.3.2 The method of Brodsky-Darkhovsky

The second non-parametric method we describe is an algorithm studied by Brodsky and Darkhovsky in [27, Chapter IV.1]. At time \(n\) their method considers the observations \(X_{n-N+1}, X_{n-N+2}, \ldots, X_n\), where \(N\) is the so-called *window size*. In order to check whether a change in mean has occurred at time \(n-N+k+1\), the average over \(X_{n-N+1}, \ldots, X_{n-N+k}\) is compared with the average over \(X_{n-N+k+1}, \ldots, X_n\). If there is no change point, the difference tends to be close to 0. Therefore an alarm is raised if there is a \(k\) for which the difference exceeds some threshold \(c > 0\). Note that \(c\) should be smaller than the mean change, as otherwise issuing an
alarm remains rare, even if there is a change point. For \( k \) close to 1 or \( \frac{N-1}{2} \), one of the averages contains few values. Therefore, a parameter \( \tilde{\gamma} \in (0, \frac{1}{2}) \) is chosen and only values \( k \in \{\lceil \tilde{\gamma}N \rceil, \ldots, \lfloor (1 - \tilde{\gamma})N \rfloor \} \) are considered. The method, illustrated in Figure 2.2, is summarised as follows.

**Method 2.3.** Fix the window width \( N, \tilde{\gamma} \in (0, \frac{1}{2}) \) and threshold \( c > 0 \). Introduce

\[
U_n(k, N) := \frac{1}{k} \sum_{i=n-N+1}^{n-N+k} X_i - \frac{1}{N-k} \sum_{i=n-N+k+1}^{n} X_i.
\]

The method of Brodsky-Darkhovsky prescribes to issue an alarm if the test statistic

\[
U_n(N) := \max_{k \in \{\lceil \tilde{\gamma}N \rceil, \ldots, \lfloor (1 - \tilde{\gamma})N \rfloor \}} |U_n(k, N)|
\]

exceeds the threshold \( c \).

---

### 2.4 Conclusion

In this chapter we have presented several statistical methods for the detection of sudden and insistent changes (change points) in sequences of observations. The methods differ in the type of changes that can be detected (i.e. change in distribution or change in mean) and the requirements on the input (e.g. independence of observations). We conclude this chapter with a qualitative comparison of the presented methods.

**Cusum** \( [91] \) is designed to detect a change in distribution, obviously also covering changes in for example the mean or the variance. The method has two strong assumptions. Firstly, it assumes that the observations
2.4 Conclusion

Figure 2.2. An example where the method of Brodsky-Darkhovsky is used to detect a change in mean from $\mu_1 = 1$ to $\mu_2 = 2$ at time $k = 51$. The settings are $N = 40$, $\bar{\gamma} = 0.1$ and $c = 0.5$. A change point at $k^* = 50$ is detected at time $n^* = 53$. The difference between the sample averages $m_1 = \frac{1}{36} \sum_{i=14}^{49} X_i$ and $m_2 = \frac{1}{4} \sum_{i=50}^{53} X_i$ is $1.65 - 0.98 = 0.67$ which is larger than the threshold.

are independent and, secondly, it assumes that the distributions before and after the change are known (or estimated). The method further includes a threshold on the test statistic, which is (informally speaking) a threshold on the ratio between the maximum (over all points in time) of the probability that there is a change point at that specific time and the probability that there is no change point in the data. We have discussed a LD approach to approximate the false alarm probability as a function of the threshold. In Chapter 3 this LD approach is employed to develop change point detection methods for dependent sequences.

The method we described next is based on cusum, but for this method — referred to as non-parametric cusum [27, Chapter IV.2] — no distributions need to be estimated. This method aims at finding a change in mean and has one parameter only; the threshold imposed on the test statistic. An alarm is raised if the difference between the cumulative sum of the observed values and the minimum so far exceeds this threshold. See [1] for approximations of the performance metrics, which can be used to choose the threshold. In the corresponding derivations no independence is assumed; it is sufficient if a specific mixing condition applies.

A second distribution-free method to detect a change in mean is the method of Brodsky-Darkhovsky [27]. The method requires the user to set three parameters. Firstly, one has to decide on the window size and, secondly, on a parameter $\bar{\gamma}$. A window of observations is divided into
two intervals such that both of them contain at least a fraction $\tilde{\gamma}$ of the number of observations in the window. Thirdly, a threshold on the size of the mean change needs to be chosen; this threshold determines when an alarm is raised. Again, see [1] for approximations of the performance metrics. In this case, the approximations assume independence.

In [1] we have briefly discussed three additional methods for detecting a change in variance, which are not included in this chapter. Approximations for all of these methods assume the observed data to be normally distributed and rely on a test statistic that is a function of the sum of the squares of the observed values up to the potential change point $k$ and the sum of squares of the rest of the observations.