Inductive Conformal Martingales for Change-Point Detection

Denis Volkhonskiy DVOLKHONSKIY@GMAIL.COM
Yandex School of Data Analysis, Moscow, Russia
Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Institute for Information Transmission Problems, Moscow, Russia

Evgeny Burnaev E.BURNAEV@SKOLTECH.RU
Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Institute for Information Transmission Problems, Moscow, Russia

Ilia Nouretdinov I.R.NOURETDINOV@RHUL.AC.UK
Information Security Group and Computer Learning Research Center, Department of Computer Science, Royal Holloway, University of London, London, UK

Alexander Gammerman A.GAMMERMAN@RHUL.AC.UK
Vladimir Vovk V.VOVK@RHUL.AC.UK
Computer Learning Research Center, Department of Computer Science, Royal Holloway, University of London, London, UK

Editors: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

We consider the problem of quickest change-point detection in data streams. Classical change-point detection procedures, such as CUSUM, Shiryaev-Roberts and Posterior Probability statistics, are optimal only if the change-point model is known, which is an unrealistic assumption in typical applied problems. Instead we propose a new method for change-point detection based on Inductive Conformal Martingales, which requires only the independence and identical distribution of observations. We compare the proposed approach to standard methods, as well as to change-point detection oracles, which model a typical practical situation when we have only imprecise (albeit parametric) information about pre- and post-change data distributions. Results of comparison provide evidence that change-point detection based on Inductive Conformal Martingales is an efficient tool, capable to work under quite general conditions unlike traditional approaches.

Keywords: Conformal prediction, nonconformity, anomaly detection, time-series, change-point detection, Exchangeability Martingales, Inductive Conformal Martingales, change-point detection oracles

1. Introduction

Conformal Martingales (martingales based on the conformal prediction framework, see Vovk et al. 2005, Section 7.1) are known as a valid tool for testing the exchangeability and i.i.d. assumptions. They were proposed in Vovk et al. (2003) and later generalized in Fedorova et al. (2012).

One of rather widespread examples of non-i.i.d. data is data with Change-Points (CPs) (see Basseville and Nikiforov 1993; Tartakovsky et al. 2014): we assume an on-line scheme of

observations, such that before some moment of time (change-point) observations are i.i.d., and after it observations are also i.i.d., but with some other distribution. Thus, overall observations are not i.i.d. This is the reason why application of Conformal Martingales (CMs) to CP detection is possible.

CP detection problems span many applied areas and include automatic video surveillance based on motion features (Pham et al., 2014), intrusion detection in computer networks (Tartakovsky et al., 2006), anomaly detection in data transmission networks (Casas et al., 2010), anomaly detection for malicious activity (Burnaev et al., 2015a,b; Burnaev and Smolyakov, 2016), change-point detection in software-intensive systems (Artemov et al., 2016; Artemov and Burnaev, 2016a,b), fault detection in vehicle control systems (Malladi and Speyer, 1999; Alestra et al., 2014), detection of onset of an epidemic (MacNeill and Mao, 1995), drinking water monitoring (Guépié et al., 2012) and many others.

Standard statistics for change-point detections, such as Cumulative Sum (CUSUM, Page 1954) and Shiryaev-Roberts (S-R, Shiryaev 1963, Roberts 1966), have very strong assumptions about data distributions both before and after the change-point. Usually in practice we do not know the change-point model.

The first attempt to use CMs for change-point detection was made in Ho (2005). However, only two different martingale tests were considered for CP detection.

CM is defined by two main components: a conformity measure (CM) and a betting function (Vovk et al., 2003; Fedorova et al., 2012). Nowadays there exist different approaches to define conformity measures and betting functions. Thus, a whole zoo of CMs for CP detection can be constructed.

Therefore, the goal of our work is to

- propose different versions of CMs for CP detection, based on available as well as newly designed conformity measures and betting functions, specially tailored for CP detection;
- perform extensive comparison of these CMs with classical CP detection procedures.

As classical CP detection procedures we consider CUSUM, Shiryaev-Roberts and Posterior Probability statistics. Also we perform comparison with CP detection oracles, which model a typical practical situation when we have only imprecise information about pre- and post-change data distributions. CP detection statistics, considered in the comparison, enjoy different information about statistical characteristics of data and CP models.

Comparison is performed on simulated data, corresponding to a classical CP model (Basseville and Nikiforov, 1993):

(a) i.i.d. Gaussian white noise signal,

(b) as a CP we consider change in the mean from zero initial level.

The results of our statistical analyses clearly show that in terms of mean time delay until CP detection for the same level of false alarms CMs are comparable with CP detection oracles and are not significantly worse than optimal CP detection statistics (requiring full information about CP model). At the same time, opposed to classical CP detection statistics, CP detection based on CMs is non-parametric and can be applied in the wild

without significant parameter tuning both in case of one-dimensional and multi-dimensional data streams.

The paper is organized as follows. In Section 2 we describe CMs. In Section 3 we consider quickest CP detection problem statement and describe optimal CP detection statistics, as well as CP detection approaches based on CMs, defined by different conformity measures and betting functions. In Section 4 we consider CP detection oracles. In Section 5 we describe a protocol of experiments and provide results of simulations. We list conclusions in Section 6.

2. Conformal Martingales

First we describe Conformal Prediction framework (Vovk et al., 2005), which can be regarded as a tool, satisfying some natural properties of validity, for measuring the strangeness of observations.

2.1. Non-Conformity measures and p-values

Let us denote by $z_1, \ldots, z_n, \ldots$ a sequence of observations, where each observation is represented as a vector in some vector space. Our first goal is to test whether the new observation z_n fits the previously observed observations $z_1, \ldots, z_{n-1}$. In other words, we want to measure how strange z_n is compared to other observations. For this purpose, we use the Conformal Prediction framework (Vovk et al., 2005). The first step is the definition of a *non-conformity measure*, which is a function

$$(z, S) \mapsto A(z, S),$$

mapping pairs (z, S) consisting of an observation z and a finite multiset S of observations to a real number $A(z, S)$ with the following meaning: the greater this value is, the stranger z is relative to S. As a simple example, one can consider the Nearest Neighbors conformity measure, where $A(z, S)$ is the average distance from z to its nearest neighbors in S.

The second step in the Conformal Prediction framework is the definition of the p-value for the observation z_n:

$$p_n = p(z_n, z_{n-1}, \ldots, z_1) = \frac{\#\{i = 1 \ldots n : \alpha_i > \alpha_n\} + U\#\{i = 1 \ldots n : \alpha_i = \alpha_n\}}{n}, \qquad (1)$$

where U is a random number in $[0, 1]$ independent of $z_1, z_2, \ldots$, and the *non-conformity scores* α_i (including $i = n$) are defined by

$$\alpha_i = A(z_i, \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}), \qquad (2)$$

i.e., the p-value for the observation z_n is defined, roughly, as the fraction of observations that have non-conformity scores greater than or equal to the non-conformity score α_n. Intuitively the smaller p-value is, the stranger the observation is.

Theorem 1 *If observations $z_1, \ldots, z_n, \ldots$ satisfy the i.i.d. assumption, the p-values $p_1, p_2, \ldots$ are independent and uniformly distributed in $[0, 1]$.*

The statement of Theorem 1 (proved in Vovk et al. 2003) provides grounds for CP detection:

- observations $z_1, \dots, z_{\theta-1} \sim f_0(z)$ are i.i.d.;
- $z_\theta, z_{\theta+1}, \dots \sim f_1(z)$ are also i.i.d.;
- in the case $\theta = 1$, all the observations are i.i.d., and therefore CMs couldn't be used for detecting a CP;
- since at $\theta \geq 2$ the distribution changes, the corresponding p-values $p_1, p_2, \dots, p_n, \dots$ are not i.i.d. uniform in $[0, 1]$ for $n \geq \theta$.

We use this fact for constructing CMs for CP detection.

2.2. Definition of Exchangeability Martingales

Given a sequence of random vectors $z_1, z_2, \dots$ taking values in some observation space $\mathbb{R}^d$, the joint probability distribution of $z_1, \dots, z_N$ for a finite N is *exchangeable* if it is invariant under any permutation of these random vectors. The joint distribution of the infinite sequence of random vectors $z_1, z_2, \dots$ is *exchangeable* if the marginal distribution of $z_1, \dots, z_N$ is exchangeable for every N. By de Finetti's theorem, every exchangeable distribution is a mixture of power distributions (i.e., distributions under which the sequence $z_1, z_2, \dots$ is i.i.d.).

A *test exchanegeability martingale* is a sequence of non-negative random variables $S_0 = 1, S_1, S_2, \dots$ such that

$$\mathbb{E}(S_{n+1} | S_1, \dots, S_n) = S_n, \quad n = 0, 1, 2, \dots,$$

where $\mathbb{E}$ is the expectation w.r.t. any exchangeable distribution (equaivalently, any power) on observations. According to Ville's inequality (Ville, 1939), in this case

$$\mathbb{P}(\exists n : S_n \geq C) \leq \frac{1}{C}, \ \forall C \geq 1$$

under any exchangeable distribution. If the final value of the martingale is large, we can reject the i.i.d. (equivalently, exchangeability) assumption with the corresponding probability. In the next section we define a way to transform p-values (1) into test exchangeability martingales. An *exchangeability martingale* is defined similarly but dropping the requirements that $S_0, S_1, \dots$ should be non-negative and that $S_0 = 1$.

2.3. Constructing Exchangeability Martingales from p-values

Given a sequence of p-values, we consider a martingale of the form

$$S_n = \prod_{i=1}^{n} g_i(p_i), \ n = 1, 2, \dots, \tag{3}$$

where each $g_i(p_i) = g_i(p_i \mid p_1, \dots, p_{i-1})$ is a *betting function* required to satisfy the condition $\int_0^1 g_i(p) dp = 1$. We can easily verify the martingale property under any exchangeable

distribution:

$$\mathbb{E}(S_{n+1}|S_0,\dots,S_n) = \int_0^1 \left\{\prod_{i=1}^n g_i(p_i)\right\} g_{n+1}(p)dp = \\ = \left\{\prod_{i=1}^n g_i(p_i)\right\} \int_0^1 g_{n+1}(p)dp = \prod_{i=1}^n g_i(p_i) = S_n.$$

Test exchangeability martingale of the form (3) are *conformal martingales*. (It is interesting whether there are any other test exchangeability martingales apart from the conformal martingales.)

The intuition behind the betting function is the following: we would like to penalize the fact that p-values are not uniformly distributed (cf. Theorem 1). In Section 3.6 we describe several betting functions along with their advantages and disadvantages.

3. Quickest Change-Point detection

3.1. Problem statement

We observe sequentially a series of independent observations whose distribution changes from $f_0(z)$ to $f_1(z)$ at some unknown point θ in time. Formally, $z_1, z_2, \dots, z_n, \dots$ are independent random variables such that $z_1, z_2, \dots, z_{\theta-1}$ are each distributed according to a distribution $f_0(z)$ and $z_\theta, z_{\theta+1}, \dots$ are each distributed according to a distribution $f_1(z)$, where $1 \le \theta \le \infty$ is unknown. The objective is to detect that a change has taken place "as soon as possible" after its occurrence, subject to a restriction on the rate of false detections.

Historically, the subject of change-point detection first began to emerge in the 1920-1930's motivated by considerations of quality control. When a process is "in control," observations are distributed according to $f_0(z)$. At an unknown point θ, the process jumps "out of control" and subsequent observations are distributed according to $f_1(z)$. We want to raise an alarm "as soon as possible" after the process jumps "out of control".

Current approaches to change-point detection were initiated by the pioneering work of Page (1954). In order to detect a change in a normal mean from μ_0 to $\mu_1 > \mu_0$ he proposed the following stopping rule τ: stop and declare the process to be "out of control" as soon as $C_n - \min_{1\le k\le n} C_k$ gets large, where $C_k = \sum_{i=1}^k (z_i - \mu^*)$ and $\mu_0 < \mu^* < \mu_1$ is suitably chosen. This and related procedures are known as CUSUM (cumulative sum) procedures (see Shiryaev 2010 for a survey).

There are different approaches how to formalize a restriction on false detections as well as to formalize the objective of detecting a change "as soon as possible" after its occurrence. The restriction on false detections is usually formalized either as a rate restriction on stopping rule τ, according to which we stop our observations and declare the process to be "out of control", or a probability restriction. The rate restriction is usually formalized by a requirement that $\mathbb{E}(\tau \mid \theta = \infty) \ge T$, the probability restriction is usually formalized by a requirement that $\mathbb{P}(\tau < \theta) \le \alpha$ for all θ. The objective of detecting a change "as soon as possible" after its occurrence is usually formalized in terms of functionals of $\tau - \theta$ (Shiryaev, 2010).

3.2. Optimal approaches to Change-Point detection

Let us describe main optimal statistics for CP detection. The main assumption here is that a known probability density of observations $f_0(z)$ changes to another known probability density of observations $f_1(z)$ at some unknown point θ. We denote by

$$L_n^\theta = \prod_{i=1}^{\theta-1} f_0(z_i) \prod_{i=\theta}^{n} f_1(z_i) \tag{4}$$

the likelihood of observations $z_1, \ldots, z_n$ when $\theta \in [1, n]$, and by

$$L_n = \prod_{i=1}^{n} f_0(z_i) \tag{5}$$

the likelihood of observations $z_1, \ldots, z_n$ without CP.

Shiryaev (1963) solved the CP detection problem in a Bayesian framework. As prior on θ the distribution Geometric(p) is used, i.e., $p(n) = \mathbb{P}(\theta = n) = p(1-p)^{n-1}$, $n = 1, 2, \ldots$. A loss function has the form $\mathbb{P}(\tau \le \theta) + c\mathbb{E}(\tau - \theta)^+$, where $(x)^+ = \max(x, 0)$ and $c > 0$. Shiryaev showed that it is optimal to stop observations as soon as the posterior probability of a change exceeds a fixed level h, i.e., $\tau_{\text{PP}} = \inf\{n : \varphi_n \ge h\}$, where

$$\varphi_n = \log\left[\frac{\sum_{\theta=1}^{n} L_n^\theta p(\theta)}{L_n (1-p)^n}\right]. \tag{6}$$

In the non-Bayesian (minimax) setting of the problem, the objective is to minimize the expected detection delay for some worst-case change-time distribution, subject to a cost or constraint on false alarms. Here the classical optimality result is due to Lorden, Ritov and Moustakides (Lorden, 1971; Moustakides, 1986; Ritov, 1990). They evaluate the speed of detection by $\sup_\theta \operatorname{ess\,sup}_\omega \mathbb{E}((\tau - \theta + 1)^+ \mid z_1, \ldots, z_{\theta-1})(\omega)$ under the restriction that the stopping rules τ must satisfy $\mathbb{E}(\tau \mid \theta = \infty) \ge T$. In fact from results of Lorden, Ritov and Moustakides it follows that Page's aforementioned stopping rule, which takes the form $\tau_{\text{CUSUM}} = \inf\{n : \gamma_n \ge h\}$ with

$$\gamma_n = \max_{\theta \in [1,n]} \log\left[\frac{L_n^\theta}{L_n}\right], \tag{7}$$

is optimal.

Pollak (1985; 1987) considered another non-Bayesian setting: the speed of detection is evaluated by $\sup_{1 \le \theta < \infty} \mathbb{E}(\tau - \theta \mid \tau \ge \theta)$ under the same restriction on the stopping rules, i.e. τ must satisfy $\mathbb{E}(\tau \mid \theta = \infty) \ge T$. Pollak proved that the so-called Shiryaev-Roberts statistics (Shiryaev, 1963; Roberts, 1966) is asymptotically ($T \to \infty$) minimax. The corresponding stopping rule has the form $\tau_{\text{S-R}} = \inf\{n : \psi_n \ge h\}$ with

$$\psi_n = \log\left[\frac{\sum_{\theta=1}^{n} L_n^\theta}{L_n}\right]. \tag{8}$$

As usual we select parameter h for stopping moments τ_{PP}, τ_{CUSUM} and $\tau_{\text{S-R}}$ in such a way that these stopping moments fulfill the corresponding restrictions on false detections.

The main disadvantage of statistics (6), (7) and (8) is that we should have full information about the CP model, in particular, we should know data distributions before and after the CP. In most of practical situations such assumptions are unrealizable.

3.3. Adaptation of Conformal Martingales for Change-Point detection problem

Let us describe a modification of Conformal Martingales tailored for the CP detection problem:

- Instead of CMs we use their computationally efficient modification that we call *Inductive Conformal Martingales* (ICMs). The main difference of ICMs from CMs is that to compute a non-conformity measure we use some fixed initial training set $\{z^*_{-(m-1)}, \dots, z^*_0\}$, i.e., each time we receive a new observation z_n we compute the non-conformity score according to the formula

$$\alpha_i = A(z_i, \{z^*_{-(m-1)}, \dots, z^*_0\})$$

  (cf. original CMs where α_i are defined by (2)). Intuitively, we fix some training set and evaluate to which extent new observations are strange w.r.t. this training set. This approach allows us to speed up computations without destroying the validity (see also section 3.4): one should not recompute all non-conformity scores at each iteration. Another advantage is the possibility of parallelization in the batch mode, i.e., when we receive observations in batches.

- One drawback of the original CMs, from the point of view of the performance measures adopted in this paper, is that in the case of i.i.d. observations CMs decrease to almost zero values with time. As a result, since CMs are represented as a product of betting functions (see (3)), it takes CMs a lot of time to recover from zero to some significant value when "strange" observations appear. In order to deal with this problem we introduce

$$C_n = \max\{0, C_{n-1} + \log(g_n(p_n))\}, \quad n = 1, 2, \dots, \tag{9}$$

  where $C_0 = 0$, p_n is a p-value, and g_n is a betting function. On each iteration we cut the logarithm of the martingale. This modification performs better in terms of the mean delay until CP detection.

The complete procedure is summarized in Algorithm 1. The stopping rule for CP detection has the form $\tau_{\text{CM}} = \inf\{n : C_n \geq h\}$, where C_n is the modification of the corresponding CM, calculated according to (9). Notice that

$$C_n = \log S_n - \min_{i=1,\dots,n} \log S_i.$$

An example of the martingale is given in Fig. 1. Here we consider observations from a normal distribution with a unit variance, such that at $\theta = 500$ its mean changes from 0 to 1. We use two non-conformity measures: 1 Nearest Neighbor Non-Conformity Measure (1NN NCM) and Likelihood Ratio Non-Conformity Measure (LR NCM), which are described in section 3.5.

3.4. Validity

Let us check empirically that our method is valid for small values of train set size m (the theoretical validity is lost because of the transition from S_n to C_n). For this purpose we

Input : Training set $\{z^*_{-(m-1)}, \dots, z^*_0\}$, data $\{z_1, z_2, \dots\}$, non-conformity measure A
Output: Inductive Conformal Martingale $(S_n)_{n\geq 1}$ and its modification $(C_n)_{n\geq 1}$
Randomly shuffle $z^*_1, \dots, z^*_m$ to induce exchangeability;
Initialize $S_0 = 1$;
for $n = 1, 2, \dots$ **do**
  observe new observation z_n
  calculate non-conformity score $\alpha_n = A(z_n, \{z^*_{-(m-1)}, \dots, z^*_0\})$
  calculate p-value $p_n = \frac{\#\{i=1\dots n : \alpha_i > \alpha_n\} + U\#\{i=1\dots n : \alpha_i = \alpha_n\}}{n}$, where $U \sim \text{Uniform}[0,1]$
  calculate new ICM value S_n according to (3) and calculate its modification C_n according to (9), where $g_n(p)$ is a betting function
end

Algorithm 1: Change-point detection with Inductive Conformal Martingale

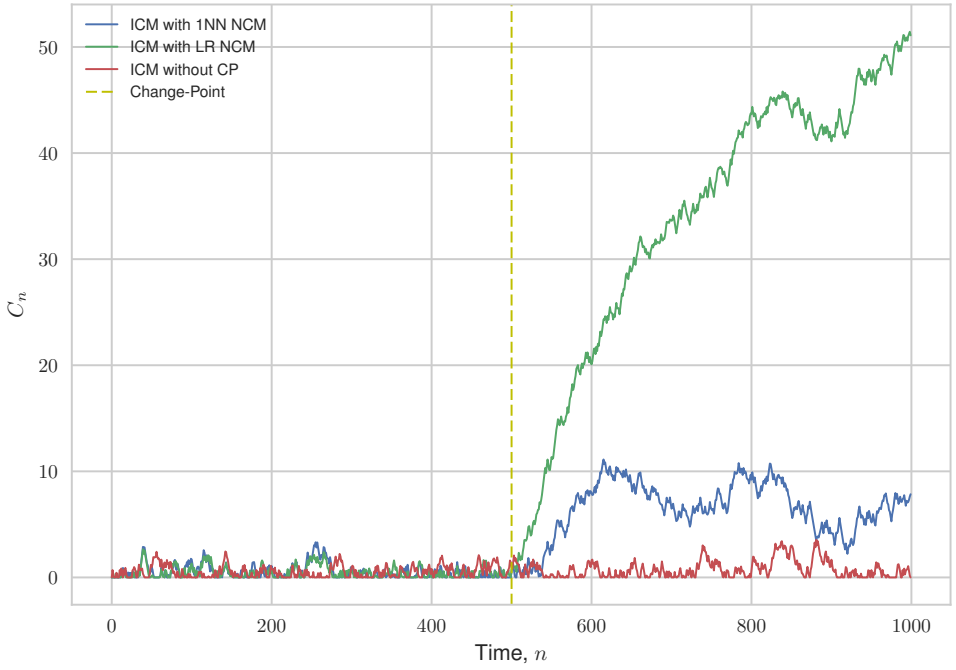

Figure 1: Example of the ICM in case of data with CP (at $\theta = 500$) and without CP

generate observations from $\mathcal{N}(\cdot \mid 0, 1)$ without CP and with CP (mean changes from 0 to 1 at $\theta = 500$). Here $\mathcal{N}(z \mid \mu, \sigma^2)$ is a value at point z of a normal density with mean μ and variance σ^2. We use k Nearest Neighbor non-conformity measure (see section 3.5 below). We plot ICM for train set sizes $m \in \{1, 2, 3, 4, 5\}$ in Fig. 2. Results of simulations, provided in Fig. 2, confirm the validity of our approach.

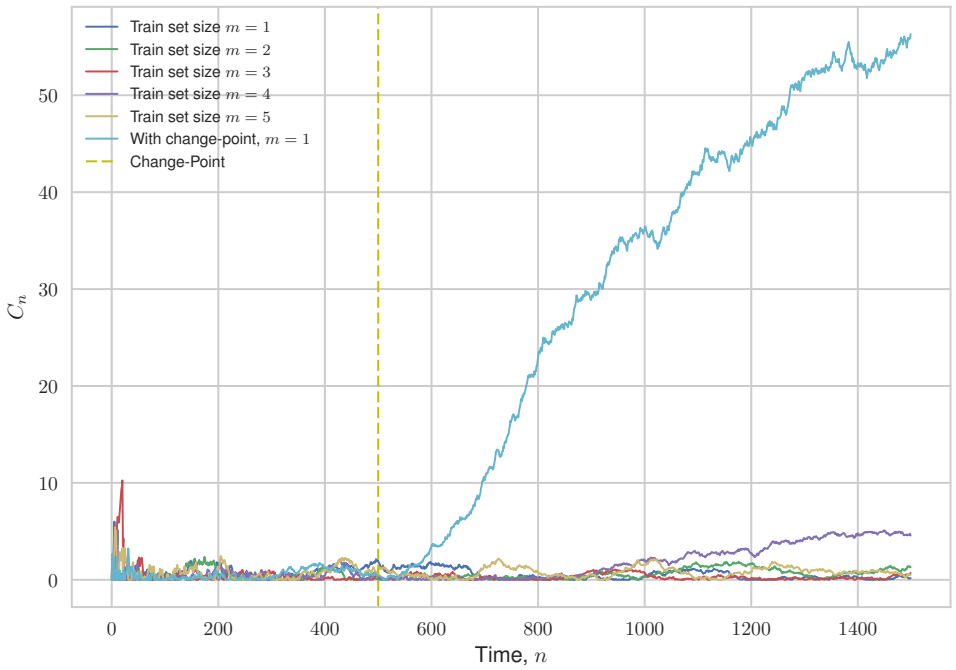

Figure 2: Validity Test of ICM: case of small train sets

3.5. Non-Conformity Measures

Let us describe non-conformity measures that we use:

- k Nearest Neighbors Non-Conformity Measure (kNN NCM). kNN NCM is computed as the average distance to k nearest neighbors. The advantage of this NCM is that it doesn't depend on any assumptions and can be used in a multi-dimensional case;

- Likelihood Ratio Non-Conformity Measure (LR NCM). One way or another the classical CP detection algorithms (see section 3.2) are based on a likelihood ratio. Thus it is worth to consider LR NCM. If we denote by $f_0(z)$ and $f_1(z)$ probability density functions before and after the CP, then a reasonable LR NCM would be

$$\alpha_n = \frac{f_1(z_n)}{f_0(z_n)}.$$

  However, we rarely know $f_i(z)$, $i = 0, 1$, exactly. Thus, we should somehow model this lack of information. We assume that $f_i(z)$, $i = 0, 1$, belong to some parametric class of densities, i.e., $f_i(z) = f(z \mid \mathbf{c}_i)$, where $\mathbf{c}_i \in \mathbf{C}$, $i = 0, 1$ are vectors of parameters. We estimate the value of $\mathbf{c}_0$ by some $\hat{\mathbf{c}}_0$ using the training set $\{z^*_{-(m-1)}, \ldots, z^*_0\}$. We also impose some prior $r(\mathbf{c}_1)$ on the parameter $\mathbf{c}_1$, i.e., we model the data distribution

after the CP by $\overline{f}_1(z) = \int f(z \mid \mathbf{c}_1) r(\mathbf{c}_1) d\mathbf{c}_1$. As a result LR NCM has the form

$$\alpha_n = \frac{\int f(z_n \mid \mathbf{c}_1) r(\mathbf{c}_1) d\mathbf{c}_1}{f(z_n \mid \hat{\mathbf{c}}_0)}. \tag{10}$$

E.g., in the one-dimensional case for $f(z \mid \mu_1) = \mathcal{N}(z \mid \mu_1, \sigma^2)$ and $r(\mu_1) = \mathcal{N}(z \mid \mu_r, \sigma_r^2)$ we get that

$$\alpha_n = \frac{\mathcal{N}(z_n \mid \mu_r, \sigma^2 + \sigma_r^2)}{\mathcal{N}(z_n \mid \hat{\mu}_0, \sigma^2)}, \tag{11}$$

where $\hat{\mu}_0 = \frac{1}{m} \sum_{i=1}^{m} z_{-m+i}$.

3.6. Betting Functions

Let us describe Betting Functions that we use:

- *Mixture Betting Function* was proposed in the very first work on testing exchangeability (Vovk et al., 2003). It doesn't depend on the previous p-values and has the form
$$g(p) = \int_0^1 \varepsilon p^{\varepsilon - 1} d\varepsilon.$$

- *Constant Betting Function.* We split the interval $[0, 1]$ into two parts at the point 0.5. We expect p-values to be small if observations are strange:
$$g(p) = \begin{cases} 1.5, & \text{if } p \in [0, 0.5), \\ 0.5, & \text{if } p \in [0.5, 1]. \end{cases}$$

- *Kernel Density Betting Function* has the form
$$g_n(p_n) = K_{p_{n-L}, \ldots, p_{n-1}}(p_n).$$
Here $K_{p_{n-L}, \ldots, p_{n-1}}(p)$ is a Parzen-Rosenblatt kernel density estimate (Rosenblatt et al., 1956) based on the previous p-values $\{p_{n-L}, \ldots, p_{n-1}\}$, L being a window size. We use a Gaussian kernel. Since p-values are in $[0, 1]$, then to reduce boundary effects we reflect the p-values to the left of zero and to the right of one, construct the density estimate, crop its support back to $[0, 1]$ and normalize. Fedorova et al. (2012) prove that such an approach provides an asymptotically better growth rate of the exchangeability martingale than any martingale with a fixed betting function. The corresponding martingale is also called the *plug-in martingale*. Let us note that for quicker CP detection we use not all p-values, but only last L of them: $\{p_{n-L}, \ldots, p_{n-1}\}$. Increasing L usually results in an increase of the mean delay, because after the CP we need to collect more observations to estimate the new distribution of p-values correctly.

- *Precomputed Kernel Density Betting Function.* To deal with the problem of high mean delay until CP detection, we propose to estimate the kernel density of p-values before constructing any martingale. For this purpose, we have to learn the betting function using some finite length realization of $z_1, z_2, \ldots, z_n, \ldots$, containing an example of a

typical CP, and some training set $\{z^*_{-(m-1)}, \ldots, z^*_0\}$. In other words, the realization should contain some CP with position and intensity resembling those of real CPs (say within the accuracy of order of magnitude) we are going to detect while applying the corresponding CM. Particular values of these parameters are specified in experimental Section 5. We compute p-values using (1) as in Algorithm 1. Using them we construct a kernel estimate of p-values density. Further we assume that for data of the same nature p-values will be distributed in a similar way, so we can use this precomputed kernel density betting function for new data realizations. Thus, thanks to the precomputed estimate we can

— Detect CP faster;

— Speed-up computations (we don't need to reconstruct density of p-values for each position of the sliding window).

4. Oracles for Change-Point detection

In the current section we describe Oracles for CP detection that we compare with CP detection based on Conformal Martingales.

4.1. Motivation to use Oracles

First we explain why we need to compare CP detection based on CMs with CP detection Oracles:

- Classical CP detection statistics are optimal in terms of the mean delay (subject to a restriction on the rate of false detections) if data distributions before and after the CP are known. There is no need for them to learn the distributions f_0 and f_1 before and after the CP.

- CMs are designed to solve another problem. As far as their validity is concerned, they assume nothing about the distributions f_i, $i = 0, 1$. They have to learn the distribution f_0 before the CP in order to detect a change.

- The profound difference between the classical setting and the adaptive setting dealt with in conformal prediction can be seen clearly if instead of the problem of quickest CP detection we consider the related problem of gambling (formalized by constructing a test martingale) against the null hypothesis (f_0 in the case of classical statistics and i.i.d. in the case of conformal prediction) in the presence of a CP. In the classical case, the growth rate of the optimal test martingale (likelihood ratio) will be exponential since the null hypothesis is simple, whereas in the i.i.d. case after an initial period of nearly exponential growth the growth rate will slow down as we start learning that f_1 is much closer to being the data-generating distribution than f_0 is.

- Thus for a fair comparison we should compare CP detection based on CMs not with CP detectors from Section 3.2, which are optimal under known f_0 and f_1, but with their modifications (oracles, defined in Section 4.2 below) that have plenty of information about pre- and post-change data distributions, but there is still some uncertainty;

the oracles only know the parametric models that f_0 and f_1 are coming from, and the task of competing with them making only a nonparametric assumption (i.i.d.) is challenging but not hopeless.

4.2. Description of Oracles

We assume that $f_i(z)$, $i = 0, 1$, belong to some parametric class of densities, i.e., $f_i(z) = f(z \mid \mathbf{c}_i)$, where $\mathbf{c}_i \in \mathbf{C}$, $i = 0, 1$ are vectors of parameters. We impose the same prior $q(\mathbf{c})$ on the parameters $\mathbf{c}_i$, $i = 0, 1$. Thus, instead of likelihood (5) of observations without CP we use

$$\overline{L}_n = \int \prod_{i=1}^{n} f(z_i \mid \mathbf{c}_0) q(\mathbf{c}_0) d\mathbf{c}_0, \tag{12}$$

and instead of likelihood (4) of observations $z_1, \dots, z_n$ with CP at $\theta \in [1, n]$ we use

$$\overline{L}_n^{\theta} = \int \prod_{i=1}^{\theta-1} f_0(z_i \mid \mathbf{c}_0) q(\mathbf{c}_0) d\mathbf{c}_0 \cdot \int \prod_{i=\theta}^{n} f_1(z_i \mid \mathbf{c}_1) q(\mathbf{c}_1) d\mathbf{c}_1. \tag{13}$$

Oracles are obtained from optimal statistics (6), (7) and (8) by using $\overline{L}_n$ from (12) instead of L_n from (5), and by using $\overline{L}_n^{\theta}$ from (13) instead of L_n^{θ} from (4) (cf. with section 2.4.2.1 and example 2.4.2 in Basseville and Nikiforov (1993)).

Let us consider a one-dimensional example. We set $f(z \mid \mu_i) = \mathcal{N}(z \mid \mu_i, 1)$, $i = 1, 2$ and $q(\mu) = \mathcal{N}(\mu \mid 0, 1)$, and we get that

$$\begin{aligned}
\overline{L}_n = \overline{L}_n(z_1, \dots, z_n) &= \int_{\mathbb{R}} \prod_{i=1}^{n} \mathcal{N}(z_i \mid \mu_0, 1) \mathcal{N}(\mu_0 \mid 0, 1) d\mu_0 \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \sqrt{\frac{2\pi}{n+1}} \exp\left\{ -\frac{n\left[\overline{z_{1,n}^2} - \frac{n}{n+1}(\overline{z_{1,n}})^2\right]}{2} \right\},
\end{aligned} \tag{14}$$

$$\begin{aligned}
\overline{L}_n^{\theta} = \overline{L}_n^{\theta}(z_1, \dots, z_n) &= \int_{\mathbb{R}} \prod_{i=1}^{\theta-1} \mathcal{N}(z_i \mid \mu_0, 1) \mathcal{N}(\mu_0 \mid 0, 1) d\mu_0 \int_{\mathbb{R}} \prod_{i=\theta}^{n} \mathcal{N}(z_i \mid \mu_1, 1) \mathcal{N}(\mu_1 \mid 0, 1) d\mu_1 \\
&= \frac{1}{\sqrt{(2\pi)^{n-2}\theta(n-\theta+2)}} \exp\left\{ -\frac{n\left[\overline{z_{1,n}^2} - \left\{\frac{(\theta-1)^2}{n\theta}(\overline{z_{1,\theta-1}})^2 + \frac{(n-\theta+1)^2}{n(n-\theta+2)}(\overline{z_{\theta,n}})^2\right\}\right]}{2} \right\},
\end{aligned} \tag{15}$$

where $\overline{z_{m,n}} = \frac{1}{n-m+1}\sum_{i=m}^{n} z_i$, $\overline{z_{m,n}^2} = \frac{1}{n-m+1}\sum_{i=m}^{n} z_i^2$. In such a way we model a situation when the Oracle does not know exact values of μ_i, $i = 1, 2$.

5. Experiments

In the current section we describe our experimental setup and provide results of experiments.

5.1. Experimental setup

We consider the following experimental setup:

- We use observations $\{z^*_{-(m-1)}, \dots, z^*_0\}$ as a training set for computation of non-conformity scores. We set $m = 200$ in all experiments.
- Observations $z^*_{-(m-1)}, \dots, z^*_0, z_1, \dots, z_{\theta-1}$ are generated from $f_0(z) \sim \mathcal{N}(z \mid 0, 1)$.
- Observations $z_\theta, z_{\theta+1}, \dots$ are generated from $f_1(z) \sim \mathcal{N}(z \mid \mu_1, 1)$. We consider $\mu_1 \in \{1, 1.5, 2\}$.

As performance characteristics we use:

- Mean delay until CP detection $\mathbb{E}_1(\tau - \theta \mid \tau > \theta)$,
- Probability of False Alarm (FA) $\mathbb{P}_0(\tau \leq \theta)$.

In all experiments using Monte-Carlo simulations we estimate dependency of the mean delay $\mathbb{E}_1(\tau - \theta \mid \tau > \theta)$ on the probability of the false alarm $\mathbb{P}_0(\tau \leq \theta)$.

For the LR NCM in (11) we set $\mu_r = 1$, $\sigma^2 = 1$ and $\sigma_r^2 = 1$, i.e.,

$$\alpha_n = \frac{\mathcal{N}(z_n \mid 1, 2)}{\mathcal{N}(z_n \mid \hat{\mu}_0, 1)},$$

where $\hat{\mu}_0 = \frac{1}{m} \sum_{i=1}^{m} z^*_{-m+i}$.

In the case of oracle detectors (see section 4.2) we use likelihoods from (14) and from (15) to obtain Posterior Oracle from the optimal statistics (6), CUSUM Oracle from the optimal statistics (7) and S-R Oracle from the optimal statistics (8). When calculating Posterior Probability statistics (6) and Posterior Oracle we set parameter p of the geometric distribution to $\frac{1}{100}$.

In experiments we consider all possible combinations of different types of Oracles, betting functions from section 3.6, non-conformity measures from section 3.5, as well as different values of $\mu_1 \in \{1, 1.5, 2\}$ and $\theta \in \{100, 200\}$. In the case of kNN NCM we set k to 7.

5.2. Refinement of the experimental setup

When applying Conformal Martingales both original and inductive versions can be used. First let us check that the inductive version is not worse than the original one. In our comparison we use a simple NCM: $\alpha_i = A(z_i, \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\} = \left| z_i - \frac{1}{n-1} \sum_{j \neq i} z_j \right|$. In Fig. 3 we plot estimated dependency of the mean delay on the probability of the false alarm for both ICM and CM with the constant betting function and different oracles. As we can see, there is almost no difference in the original and inductive versions. Later we consider only Inductive Conformal Martingales.

When calculating the Oracles we can either additionally use the train set $\{z^*_{-(m-1)}, \dots, z^*_0\}$ or not. Let us check how the addition of the train set influence results. The comparison is presented in Fig. 4. We can see that the results are practically the same. Later in the paper when calculating the Oracles we do not use the train set.

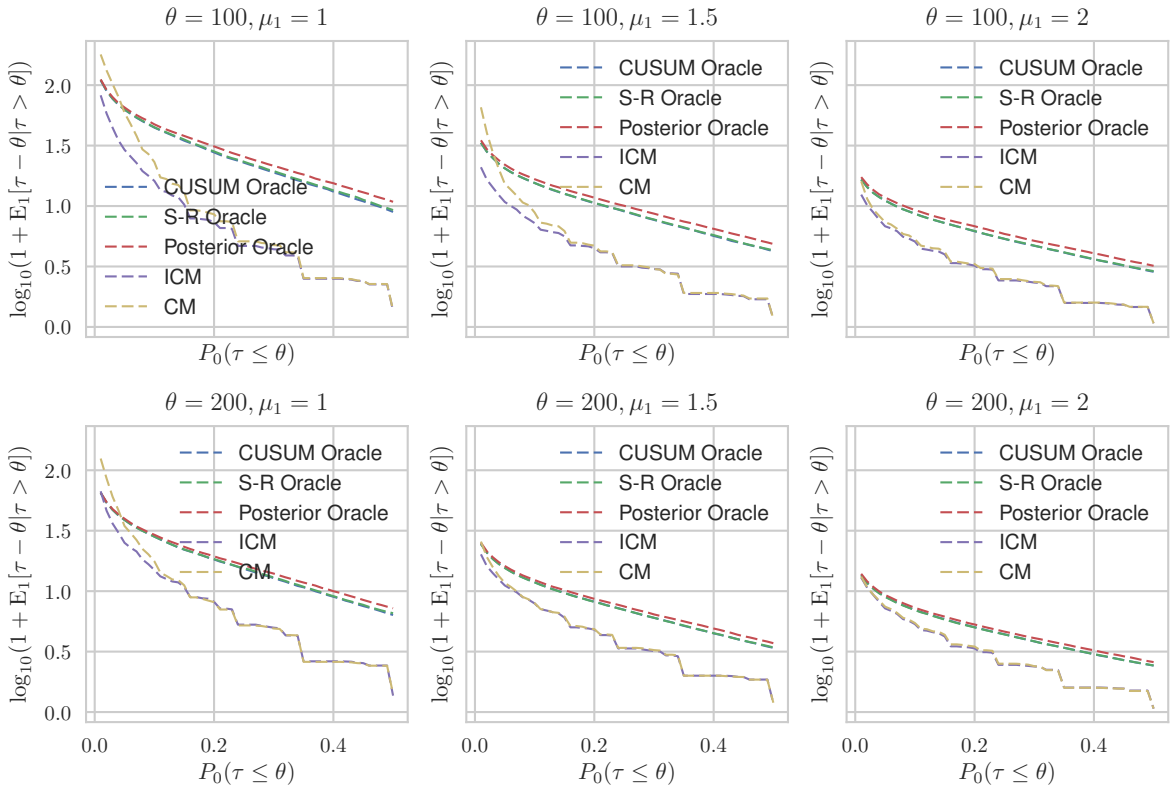

Figure 3: Comparison of ICM and CM for the constant betting function

Table 1: Comparison of ICM (Constant Betting Function) with Oracle by Mean Delay for different False Alarm probabilities

Param.\Probab. of FA	5%					10%				
	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle
$\theta = 100, \mu_1 = 1$	14.02	33.52	61.59	62.01	64.37	8.90	17.71	43.53	43.89	46.40
$\theta = 100, \mu_1 = 1.5$	7.08	12.51	19.51	19.51	20.98	4.79	7.79	14.50	14.51	15.67
$\theta = 100, \mu_1 = 2$	5.19	6.90	10.11	10.09	10.78	3.62	4.70	7.64	7.64	8.27
$\theta = 200, \mu_1 = 1$	13.22	31.33	37.78	37.80	38.73	8.33	17.17	27.24	27.24	28.25
$\theta = 200, \mu_1 = 1.5$	7.00	12.50	14.62	14.52	15.16	4.74	8.08	10.85	10.81	11.36
$\theta = 200, \mu_1 = 2$	5.13	7.12	8.02	7.98	8.30	3.59	4.85	6.00	5.97	6.28

5.3. Constant Betting Function

Results for Constant Betting Function are in Fig. 5. Here *SR* stands for S-R Oracle, *PP* — for Posterior Oracle, *CUSUM* — for CUSUM Oracle, *ICM 7 NN* — for ICM CP detector with $k = 7$ nearest neighbor NCM, *ICM LR* — for ICM CP detector with LR NCM. Mean delays for some values of false alarm probability are in Tab. 1.

5.4. Mixture Betting Function

Results for Mixture Betting Function are in Fig. 6. Mean delays for some values of false alarm probability are in Tab. 2.

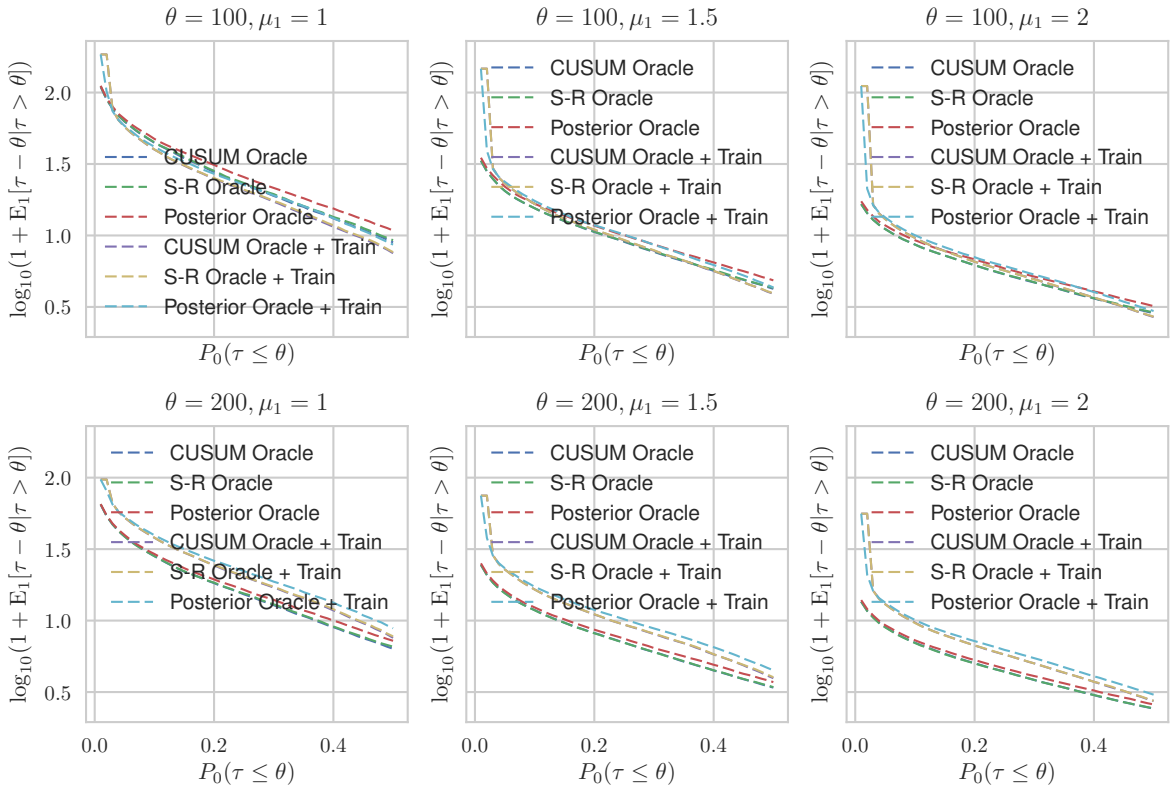

Figure 4: Comparison of Oracles with and without train set

Table 2: Comparison of ICM (Mixture Betting Function) with Oracle by Mean Delay for different False Alarm probabilities

Param.\Probab. of FA	5%					10%				
	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle
$\theta = 100, \mu_1 = 1$	132.58	193.27	61.59	62.01	64.37	66.34	124.34	43.53	43.89	46.40
$\theta = 100, \mu_1 = 1.5$	32.73	71.01	19.51	19.51	20.98	12.63	30.77	14.50	14.51	15.67
$\theta = 100, \mu_1 = 2$	11.37	16.60	10.11	10.09	10.78	5.45	7.57	7.64	7.64	8.27
$\theta = 200, \mu_1 = 1$	151.61	244.65	37.78	37.80	38.73	77.10	175.08	27.24	27.24	28.25
$\theta = 200, \mu_1 = 1.5$	29.50	65.29	14.62	14.52	15.16	16.56	32.13	10.85	10.81	11.36
$\theta = 200, \mu_1 = 2$	14.49	19.12	8.02	7.98	8.30	8.20	11.16	6.00	5.97	6.28

5.5. Kernel Betting Function

Results for Kernel Betting Function are in Fig. 7. Mean delays for some values of false alarm probability are in Tab. 3. We use a sliding window of size $L = 100$ to estimate density of p-values.

We can see, that results for the Kernel Betting Function is worse than for the Mixture Betting Function. The main reason is that it takes a long time for the martingale to grow sufficiently. In fact, before the change-point the distribution of p-values is uniform on $[0, 1]$. If for the current moment of time n it holds that $n - L \ge \theta$, the distribution of p-values p_s, $s \in [n - L, n]$ is also uniform. Thus, the martingale grows only when the change-point θ is

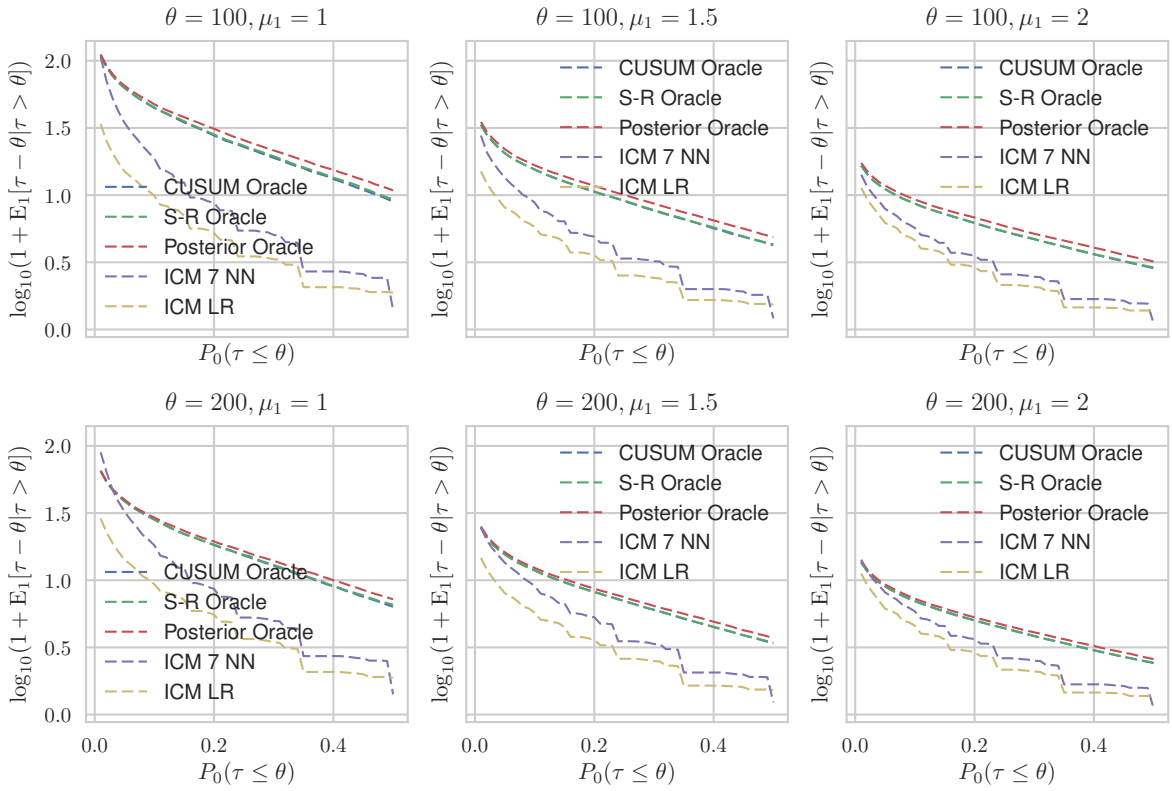

Figure 5: Constant Betting Function

Table 3: Comparison of ICM (Kernel Density Betting Function) with Oracle by Mean Delay for different False Alarm probabilities

Param.\Probab. of FA	5%					10%				
	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle
$\theta = 100, \mu_1 = 1$	33.10	65.26	61.59	62.01	64.37	22.92	38.70	43.53	43.89	46.40
$\theta = 100, \mu_1 = 1.5$	15.08	22.03	19.51	19.51	20.98	11.15	15.65	14.50	14.51	15.67
$\theta = 100, \mu_1 = 2$	9.04	11.62	10.11	10.09	10.78	6.66	8.55	7.64	7.64	8.27
$\theta = 200, \mu_1 = 1$	30.06	54.14	37.78	37.80	38.73	22.90	36.57	27.24	27.24	28.25
$\theta = 200, \mu_1 = 1.5$	15.44	22.02	14.62	14.52	15.16	12.08	17.13	10.85	10.81	11.36
$\theta = 200, \mu_1 = 2$	10.00	12.81	8.02	7.98	8.30	7.83	10.15	6.00	5.97	6.28

inside the interval $[n-L, n]$, p-values from which are used for density estimation. This is the reason why in section 3.6 we propose new Precomputed Kernel Density Betting Function.

5.6. Precomputed Kernel Betting Function

When learning the Precomputed Kernel Betting Function we use one realization $z_1, \ldots, z_n, \ldots$ of length 1000 with a CP at $\theta = 500$, such that $z_n \sim \mathcal{N}(\cdot \mid 0, 1)$ for $n < \theta$ and $z_n \sim \mathcal{N}(\cdot \mid 1, 1)$ for $n \ge \theta$ regardless of where the real CP is located and which amplitude it has.

Results for Precomputed Kernel Betting Function are in Fig. 8. Mean delays for some values of false alarm probability are in Tab. 4.

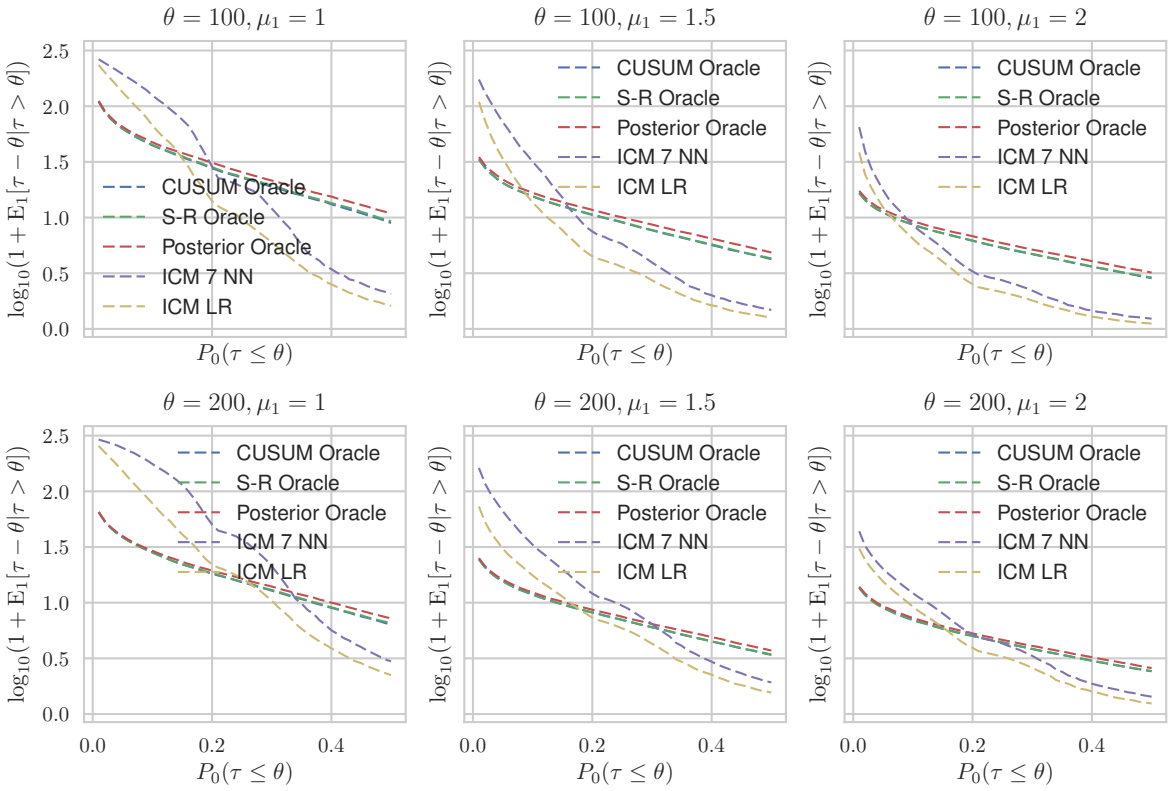

Figure 6: Mixture Betting Function

Table 4: Comparison of ICM (Precomputed Kernel Density Betting Function) with Oracle by Mean Delay for different False Alarm probabilities

Param.\Probab. of FA	5%					10%				
	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle	ICM LR	ICM kNN	CUSUM Oracle	S-R Oracle	Posterior Oracle
$\theta = 100, \mu_1 = 1$	15.20	34.41	61.59	62.01	64.37	10.08	20.27	43.53	43.89	46.40
$\theta = 100, \mu_1 = 1.5$	7.47	11.12	19.51	19.51	20.98	5.02	7.32	14.50	14.51	15.67
$\theta = 100, \mu_1 = 2$	4.95	6.22	10.11	10.09	10.78	3.28	4.11	7.64	7.64	8.27
$\theta = 200, \mu_1 = 1$	14.14	28.70	37.78	37.80	38.73	9.65	18.91	27.24	27.24	28.25
$\theta = 200, \mu_1 = 1.5$	7.24	10.80	14.62	14.52	15.16	4.92	7.39	10.85	10.81	11.36
$\theta = 200, \mu_1 = 2$	4.90	6.15	8.02	7.98	8.30	3.29	4.18	6.00	5.97	6.28

5.7. Comparison with Optimal detectors

We also compare CP detection based on CMs with optimal detectors: Cumulative Sum (CUSUM), Shiryaev-Roberts (S-R) and Posterior Probability statistics (PP). One can see from Tab. 5 and Fig. 9 that our results are comparable to results of the optimal methods. CMs perform a little bit worse, but we should notice that it requires fewer assumptions (does not know the true f_0 and f_1) and is more general (distribution-free).

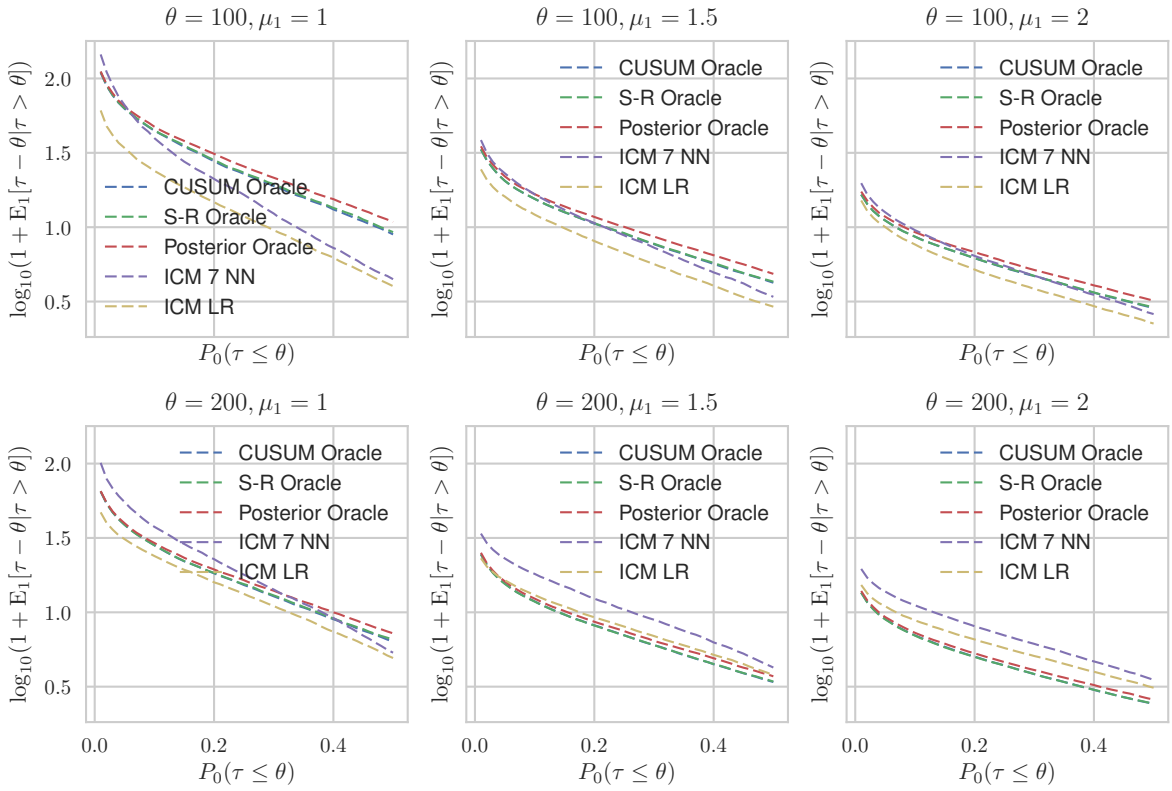

Figure 7: Kernel Density Betting Function

Table 5: Comparison of ICM (Precomputed Kernel Density Betting Function) with Optimal Detectors by Mean Delay for different False Alarm probabilities

Param.\Probab. of FA	5%					10%				
	ICM LR	ICM kNN	CUSUM	S-R	Posterior Prob.	ICM LR	ICM kNN	CUSUM	S-R	Posterior Prob.
$\theta = 100, \mu_1 = 1$	15.20	34.41	6.08	6.11	12.06	10.08	20.27	3.97	4.22	7.99
$\theta = 100, \mu_1 = 1.5$	7.47	11.12	3.42	3.60	7.11	5.02	7.32	2.19	2.43	4.67
$\theta = 100, \mu_1 = 2$	4.95	6.22	2.29	2.46	4.93	3.28	4.11	1.39	1.63	3.23
$\theta = 200, \mu_1 = 1$	14.14	28.70	6.19	6.22	12.55	9.65	18.91	4.07	4.19	8.38
$\theta = 200, \mu_1 = 1.5$	7.24	10.80	3.50	3.66	7.44	4.92	7.39	2.26	2.46	4.99
$\theta = 200, \mu_1 = 2$	4.90	6.15	2.33	2.48	5.22	3.29	4.18	1.46	1.64	3.44

6. Conclusion

In this paper we describe an adaptation of Conformal Martingales for change-point detection problem. We demonstrate the efficiency of this approach by comparing it with natural oracles, which are likelihood-based change-point detectors. Our results indicate that the efficiency of change-point detection based on conformal martingales in most of cases is comparable with that of oracle detectors.

We propose and compare several approaches to calculating a betting function (a function that transforms p-values into a martingale) and a non-conformity measure (a function that defines strangeness and, therefore, p-values). We get that the Precomputed Kernel Betting

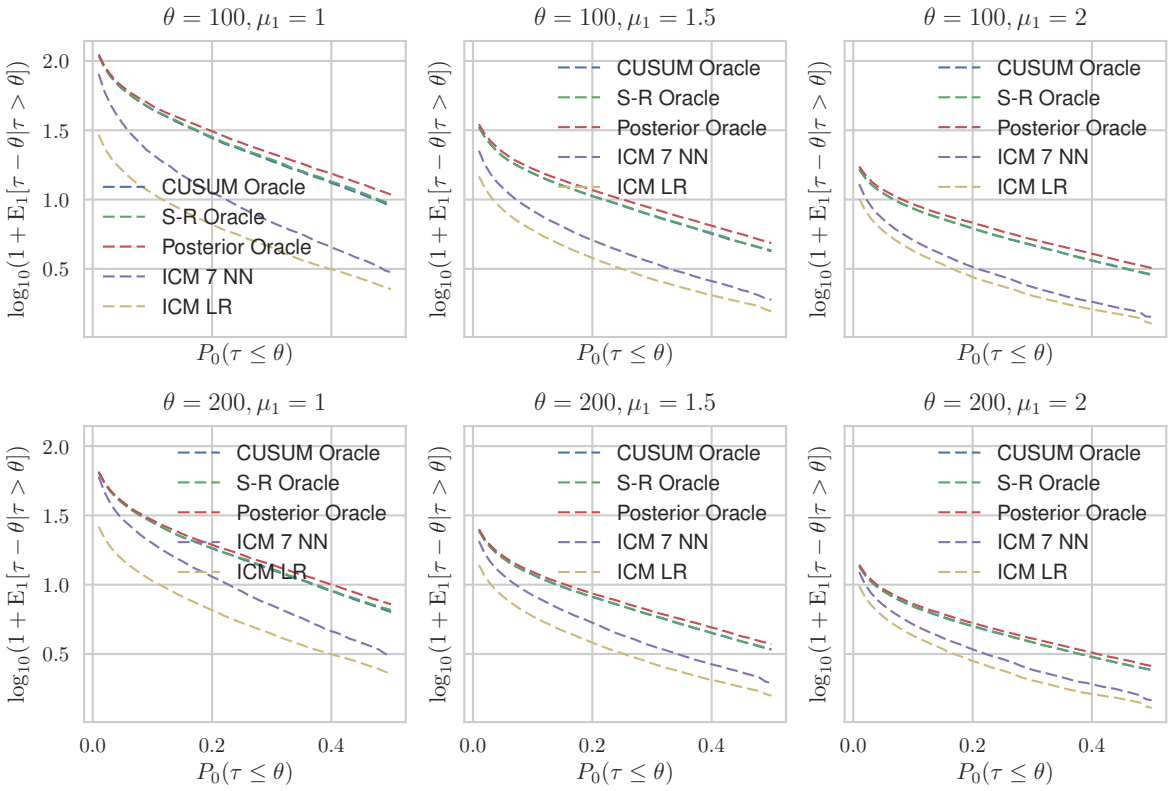

Figure 8: Precomputed Kernel Density Betting Function

Function provides the most efficient results and the Mixture Betting Function provides the worst results.

We also compare Inductive Conformal Martingales with methods that are optimal for known pre- and post-CP distributions, such as CUSUM, Shiryaev-Roberts and Posterior Probability statistics. Our results are worse but still they are comparable. Some deterioration is inevitable, of course, since CMs are distribution-free methods and, therefore, require much weaker assumptions.

Acknowledgments

We are grateful for the support from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671555 (ExCAPE project). The research presented in Section 5 of this paper was supported by the RFBR grants 16-01-00576 A and 16-29-09649 ofi_m. This work was also supported by the Russian Science Foundation grant (project 14-50-00150), the UK EPSRC grant (EP/K033344/1), and the Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway. We are indebted to Prof. Ilya Muchnik, School of Data Analysis, Yandex, and Royal Holloway, University of London, for the studentship support of one of the authors.

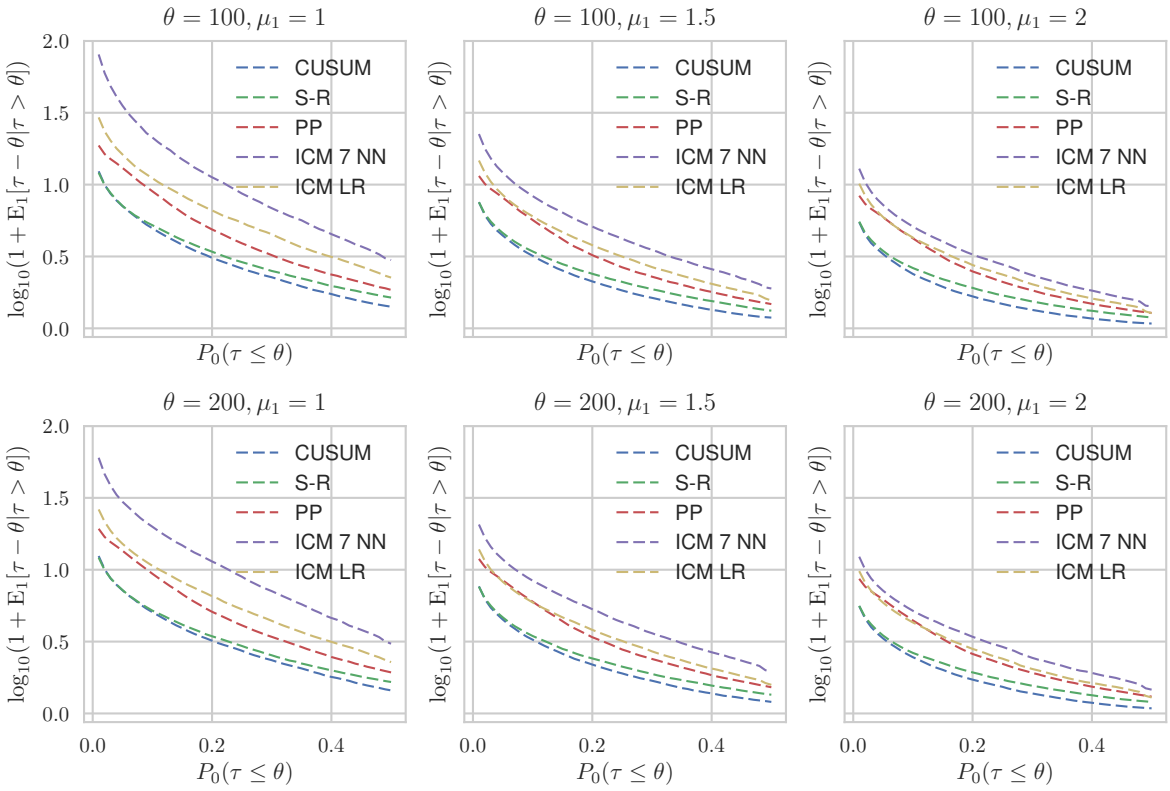

Figure 9: Comparison with Optimal detectors. Precomputed Kernel Density Betting Function

References

Stephane Alestra, Christophe Bordry, Christophe Brand, Evgeny Burnaev, Pavel Erofeev, Artem Papanov, and Cassiano Silveira-Freixo. Application of rare event anticipation techniques to aircraft health management. *Advanced Materials Research*, 1016:413–417, 2014.

Alexey Artemov and Evgeny Burnaev. Ensembles of detectors for online detection of transient changes. In *Proceedings of the Eighth International Conference on Machine Vision (ICMV)*, pages 1–5, 2016a.

Alexey Artemov and Evgeny Burnaev. Detecting performance degradation of software-intensive systems in the presence of trends and long-range dependence. In *Proceedings of the Sixteenth International Conference on Data Mining Workshops (ICDMW)*, pages 29–36. IEEE Conference Publications, 2016b.

Alexey Artemov, Evgeny Burnaev, and Andrey Lokot. Nonparametric decomposition of quasi-periodic time series for change-point detection. In *Proceedings of the Eighth International Conference on Machine Vision (ICMV)*, pages 1–5, 2016.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, Englewood Cliffs, 1993.

Evgeny Burnaev and Dmitry Smolyakov. One-class SVM with privileged information and its application to malware detection. In *Proceeding of the Sixteenth International Conference on Data Mining Workshops (ICDMW)*, pages 273–280. IEEE Conference Publications, 2016.

Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. Influence of resampling on accuracy of imbalanced classification. In *Proceedings of the Eighth International Conference on Machine Vision (ICMV)*, pages 1–5, 2015a.

Evgeny Burnaev, Pavel Erofeev, and Dmitry Smolyakov. Model selection for anomaly detection. In *Proceedings of the Eighth International Conference on Machine Vision (ICMV)*, pages 1–6, 2015b.

Pedro Casas, Sandrine Vaton, Lionel Fillatre, and Igor Nikiforov. Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements. *Computer Networks*, 54:1750–1766, 2010.

Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning (ICML)*, 2012.

Blaise Kévin Guépié, Lionel Fillatre, and Igor Nikiforov. Sequential detection of transient changes. *Sequential Analysis*, 31:528–547, 2012.

Shen-Shyang Ho. A martingale framework for concept change detection in time-varying data streams. In *Proceedings of the Twenty-Second International Conference on Machine learning (ICML)*, pages 321–327. ACM, 2005.

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

I. B. MacNeill and Y. Mao. Change-point analysis for mortality and morbidity rate. *Applied Change Point Problems in Statistics*, pages 37–55, 1995.

Durga P. Malladi and Jason L. Speyer. A generalized Shiryayev sequential probability ratio test for change detection and isolation. *IEEE Transactions on Automatic Control*, 44: 1522–1534, 1999.

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

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

Duc-Son Pham, Svetha Venkatesh, Mihai Lazarescu, and Saha Budhaditya. Anomaly detection in large-scale data stream networks. *Data Mining and Knowledge Discovery*, 28: 145–189, 2014.

Moshe Pollak. Optimal detection of a change in distribution. *Annals of Statistics*, 13: 206–227, 1985.

Moshe Pollak. Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics*, 15:749–779, 1987.

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

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

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

Albert N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8:22–46, 1963.

Albert N. Shiryaev. Quickest detection problems: fifty years later. *Sequential Analysis*, 29: 345–385, 2010.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press, Boca Raton, FL, 2014.

Alexander G. Tartakovsky, Boris L. Rozovskii, Rudolf B. Blažek, and Hongjoong Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54:3372–3381, 2006.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

Vladimir Vovk, Ilia Nouretdinov, and Alexander Gammerman. Testing exchangeability on-line. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, volume 12, pages 768–775, 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.