entropy

MDPI

Review

Multi-Stream Quickest Change Detection: Foundations and Recent Advances

Topi Halme * and Visa Koivunen

Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland; visa.koivunen@aalto.fi
* Correspondence: topi.halme@aalto.fi

Abstract

This paper provides an overview of recent developments in quickest change detection (QCD) for high-dimensional multi-sensor systems, with an emphasis on settings involving structural constraints and limited sensing resources. Classical QCD methodologies, while well understood in low-dimensional and fully observed regimes, face significant challenges when extended to modern applications characterized by large-scale data, constrained sampling or communication, and heterogeneous signal structures. We review key approaches for handling high dimensionality, including methods that exploit sparsity, and other forms of signal heterogeneity. Additionally, we discuss sampling constraints, where observations must be selected or acquired sequentially under resource limitations. Multi-stream applications can require making multiple detections, for example when detecting changes separately in different streams. The underlying assumptions on probability models, the types of changes taking place, commonly used decision-making criteria, performance indices, and error types are described. We also briefly discuss the application of machine learning in cases where the underlying probability models are not known, or there is a need to select which sensors should monitor the phenomena because of the large scale of the system.

Keywords: quickest change detection; false discovery rate; adaptive sensing; high-dimensional inference

1. Introduction

In many modern applications, a fundamental problem is detecting and localizing rapid changes, anomalies, or adversarial events in monitored processes or phenomena. Multiple data streams may be monitored simultaneously by using multiple devices or sensors in distinct locations. Detecting a change in the statistical properties of a monitored process as quickly as possible is known as a quickest change detection (QCD) problem [1,2]. The general objective in a QCD problem is to construct procedures that detect the change with as minimal delay as possible, while controlling the rate of false alarms. A false positive is considered to occur when one prematurely makes a decision that a change has taken place before it actually happened. Importantly, the inference must be performed in real time, as in a variety of important applications the data are streaming, dynamic, and generated sequentially.

Historically, the change detection problem was first studied in the context of industrial quality control during the 1920s and 1930s. The first algorithms are due to Shewhart [3,4] and Page [5]. Shewhart proposed a scheme where an alarm is raised the first time an

Check for updates

Academic Editors: Onur Günlü, Holger Boche, Giuseppe Caire, Shlomo Shamai (Shitz) and Osvaldo Simeone

Received: 17 April 2026
Revised: 15 May 2026
Accepted: 15 May 2026
Published: 18 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

observation (or a function of it) falls outside pre-specified limits. As such, the decision to stop or continue is only based on the most recent data point and does not utilize previous observations. The resulting detection rule is easy to implement, but the sub-optimal use of all available data may make the detection of small changes very slow. Page [5] suggested using moving averages or cumulative sums of past statistics for more efficient detection. The algorithms of Shewhart and Page were not motivated by rigorous theoretical analysis, but by easy implementation and reasonable practical performance in quality control tasks. The first optimality considerations appeared in the 1950s and 60s after Wald's pioneering work on sequential analysis [6]. In a Bayesian context, the first optimal algorithms and theory are due to Shiryaev in a series of works in the 1960s [7,8]. Optimality in a non-Bayesian, or minimax, sense was first studied by Lorden in 1971 [9]. Since then, the general sequential change detection problem has been extended to many directions relevant to modern applications, including non-i.i.d. data models, uncertainty in the pre- and/or post-change distributions, high-dimensional problems, detection under sampling constraints, and decentralized systems, among others. Extensive general treatments of quickest change detection can be found in, e.g., the books by Hadjiliadis and Poor [1] and Tartakovsky et al. [2].

In contemporary applications of QCD algorithms, a defining characteristic is a large number of data streams that must be monitored simultaneously in real time. A large number of devices equipped with sensors and communications capabilities may be observing streaming data continuously, for example smartphones, smart buildings, distributed radars, and wireless base-stations, weather stations, seismic and environment monitoring, the Internet of Things, and sensors monitoring the state of our infrastructure (power grid, roads, and bridges), to name a few. Depending on the application, the task may be to detect changes in individual streams separately. Alternatively, the goal may be to detect a single change that affects one or more of the streams simultaneously or with a small delay due to the physics of the phenomena, such as radio wave propagation. In both cases, the large number of streams poses additional, but different, difficulties and opportunities. In particular, this overview article highlights three central challenges that arise specifically in the multi-stream settings, and discusses solutions proposed in the literature. The challenges are as follows:

1. **Changes with high-dimensional unknown parameters:** In practical applications of sequential change detection, the distribution of the data after the change is usually not exactly known. Instead, the post-change distribution of the data may belong to a family of distributions that depends on an unknown parameter vector $\boldsymbol{\theta}$. For example, each component of $\boldsymbol{\theta}$ could represent the post-change mean parameter in different data streams, or $\boldsymbol{\theta}$ could denote the indices of sensors that have been affected by the change. If the number of streams is large, estimating $\boldsymbol{\theta}$ accurately becomes increasingly difficult, which in turn makes detecting a change more challenging.
2. **Adaptive sampling under constraints:** When the number of sensors is large, it is often not feasible to receive data from all sensors at all times due to, e.g., power and bandwidth constraints. Naturally, a tradeoff exists between the proportion of data observed and the ensuing detection delay. However, if the observer, such as a common fusion center, can adaptively choose which sensors it wants to receive observations from at a given time instance, the impact of reduced communication detection efficiency can be mitigated. Designing sampling policies that most effectively sample the sensors specifically for change detection purposes is therefore an important topic with lots of recent and ongoing research.
3. **Adjusting for multiple comparisons:** In many applications, changes can happen at different times at different sensors, leading to many individual detections. In these

cases, conventional QCD performance metrics need to be adjusted to take into account the simultaneous tests. Multiple Hypothesis Testing (MHT) provides a framework for controlling the error rates in that setting.

This paper presents a brief overview of underlying models and methods in multi-stream QCD. The underlying assumptions on probability models, the types of changes taking place, commonly used decision-making criteria, performance indices, and error types are described. We focus on large-scale settings where the number of streams and the dimension of the probability model parameter vectors may be large. We will also briefly discuss the application of machine learning in cases where the underlying probability models are not known or there is a need to select which sensors should monitor the phenomena because of the large scale of the system. We will describe some timely and emerging applications of multi-stream QCD.

Notation. Time is indexed by $n \in \mathbb{N}$. In the single-stream setting, $X_1, X_2, \ldots$ denotes the observed data sequence, while in the multi-stream setting, $X_n = [X_n^{(1)}, \ldots, X_n^{(K)}]$ denotes the K-dimensional observation vector at time n, where K is the number of streams or sensors. The change point is denoted by ν, and in settings with streamwise changes, $\nu^{(k)}$ denotes the change point in stream k. Expectations and probabilities when change happens at time ν are denoted by $\mathbb{E}_\nu$ and $\mathbb{P}_\nu$, respectively, while $\mathbb{E}_\infty$ and $\mathbb{P}_\infty$ correspond to the no-change setting. When the post-change model depends on an unknown parameter $\boldsymbol{\theta}$, the corresponding expectation is denoted by $\mathbb{E}_\nu^{\boldsymbol{\theta}}$. Pre- and post-change distributions are denoted by f_0 and f_1 in the simple setting, and more generally by $f(\cdot \mid \boldsymbol{\theta})$ for a parametric family.

2. Quickest Change Detection: Basic Fundamentals

In this section, we briefly introduce the standard QCD problem in the single-stream i.i.d. setting. The objective is to detect an abrupt change, if one occurs, in the underlying probability model of a data sequence $X_1, X_2, \ldots$ as quickly as possible. Initially, the data $X_1, \ldots, X_{\nu-1}$ is assumed to be generated according to a probability density f_0, which usually represents a normal operational state of the system. A common example is that the data observed by a sensor contains Gaussian noise only. At an unknown time ν, the data distribution changes to some other distribution f_1, so that $X_\nu, X_{\nu+1}, \ldots \sim f_1$.

A change detection procedure is defined by a *stopping time* random variable T, which indicates the time the sampling is stopped and a change is declared.

Definition 1 (Stopping time)**.** *Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ denote the sigma-algebra generated by the first n observations. A random variable T is a stopping time with respect to a sequence $X_1, X_2, \ldots$ if for each n, the event $\{T = n\} \in \mathcal{F}_n$. Equivalently, the event $\{T = n\}$ is a function of only $(X_1, \ldots, X_n)$.*

A central modeling choice is whether the change point ν is considered to be a random variable with a prior distribution, or an unknown deterministic quantity. The former yields the Bayesian approach, while the latter corresponds to a non-Bayesian, or minimax, framework. The two paradigms are introduced next.

2.1. Classical Non-Bayesian Objectives

When ν is unknown, the usual QCD objective is to minimize some notion of average detection delay (ADD) while maintaining the average run length (ARL) to false alarm above a tolerable level γ. The ARL to false alarm is defined as

$$\mathrm{ARL}(T) = \mathbb{E}_\infty(T), \tag{1}$$

https://doi.org/10.3390/e28050566

meaning that when no change occurs, the constraint $\mathrm{ARL}(T) \geq \gamma$ implies that the test should take at least γ samples on average before (false positive) detection. In other words, it is the expected number of observations to an alarm assuming that there is no change. Average detection delay is usually measured in the worst case over values of ν. Lorden [9] proposed a "worst-worst-case" average detection delay (WADD) criterion, defined as

$$\mathrm{WADD}(T) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu[(T - \nu + 1)^+ | \mathcal{F}_{\nu-1}], \tag{2}$$

where a supremum is first taken over all possible realizations of pre-change data $\mathcal{F}_{\nu-1}$, and then all possible values of the change point. This leads to the following optimization problem:

$$\text{minimize} \;\; \mathrm{WADD}(T) \quad \text{subject to } \mathrm{ARL}(T) \geq \gamma. \tag{3}$$

In an alternate approach, Pollak [10] suggested taking the average instead of the supremum over all possible pre-change observations. He proposed using conditional average detection delay (CADD), the expected number of samples needed to detect the change after it occurs as a criterion:

$$\mathrm{CADD}(T) = \sup_{\nu \geq 1} \mathbb{E}_\nu(T - \nu | T \geq \nu). \tag{4}$$

It is easy to see that $\mathrm{WADD}(T) \geq \mathrm{CADD}(T)$ for all T.

Finding exactly optimal solutions to formulations like (3) for all values of γ is usually difficult, except in few simple cases. For this reason, a common consideration in the literature is asymptotic optimality, usually studied in the limit $\gamma \to \infty$. In the i.i.d. setting, a universal lower bound for CADD of any procedure as $\gamma \to \infty$ was derived by Lai [11] as

$$\inf_{T:\mathrm{ARL}(T) \geq \gamma} \mathrm{CADD}(T) \geq \frac{\log \gamma}{I}(1 + o(1)), \tag{5}$$

where I is the Kullback–Leibler (KL) divergence (relative entropy) between distributions f_1 and f_0:

$$I = \int f_1(x) \log \frac{f_1(x)}{f_0(x)} dx. \tag{6}$$

Earlier, a similar lower bound for WADD was derived by Lorden [9]. A change detection procedure T is said to be asymptotically (first-order) optimal if

$$\frac{\mathrm{CADD}(T)}{\inf_{\tau:\mathrm{ARL}(\tau) \geq \gamma} \mathrm{CADD}(\tau)} \to 1 \quad \text{as } \gamma \to \infty. \tag{7}$$

2.2. Fundamental Non-Bayesian QCD Algorithms

Next, we describe briefly two foundational algorithms for QCD in the non-Bayesian paradigm, i.e., the CuSum procedure and the Shiryaev–Roberts procedure.

CuSum procedure. The Cumulative Sum (CuSum) procedure was first introduced by Page [5]. The CuSum test tracks the drift of the cumulative log-likelihood-ratio sum $S_1, S_2, \ldots$:

$$S_n = \sum_{j=1}^{n} \log \frac{f_1(X_j)}{f_0(X_j)}. \tag{8}$$

It stops when the sum exceeds its past minimum by a sufficient amount. The test is defined by

$$W_n = S_n - \min_{k \le n} S_k, \tag{9}$$

$$T_b^{\mathrm{C}} = \inf\{n : W_n > b\}. \tag{10}$$

The CuSum statistic W_n can alternatively be written as

$$W_n = \max_{k \le n} \sum_{j=k}^{n} \log \frac{f_1(X_j)}{f_0(X_j)} \tag{11}$$

and updated recursively using the formula

$$W_n = \max\left\{0, W_{n-1} + \log \frac{f_1(X_n)}{f_0(X_n)}\right\}, \quad W_0 = 0. \tag{12}$$

The recursive update makes the CuSum test efficient to implement in practice, and provides a low complexity method for detecting changes in streaming data. Lorden [9] proved that as $\gamma \to \infty$, the CuSum procedure is asymptotically first-order optimal in terms of WADD. Later, Moustakides [12] showed that CuSum with the stopping threshold b, chosen such that $\mathrm{ARL}(T_b^{\mathrm{C}}) = \gamma$, is exactly optimal for all γ in the sense of (3). Finding a b such that the ARL constraint is satisfied with equality is generally not possible in closed form, but can be found, e.g., numerically via Monte Carlo simulations. A simple lower bound for $\mathrm{ARL}(T_b^{\mathrm{C}})$ as a function of b exists, and is given by (see e.g., [[2], Lemma 8.2.1.])

$$\mathrm{ARL}(T_b^{\mathrm{C}}) \ge e^b. \tag{13}$$

Choosing $b = \log \gamma$ guarantees $\mathrm{ARL}(T_b^{\mathrm{C}}) \ge \gamma$.

Shiryaev–Roberts procedure. The Shiryaev–Roberts (SR) procedure is defined by the statistic

$$R_n = \sum_{k=1}^{n} \prod_{j=k}^{n} \frac{f_1(X_j)}{f_0(X_j)} \tag{14}$$

and the stopping time

$$T_b^{\mathrm{SR}} = \inf\{n : R_n > b\}, \tag{15}$$

where b is a threshold. The difference between the SR and CuSum statistics is that instead of taking the maximum over $k \le n$ as in (11), the SR statistic takes the sum of likelihood ratios over $k \le n$. The SR statistic also has a recursive form, given by

$$R_n = (1 + R_{n-1}) \frac{f_1(X_n)}{f_0(X_n)}, \quad R_0 = 0. \tag{16}$$

The recursion stops and raises an alarm when R_n exceeds the threshold b. The procedure is also asymptotically optimal with respect to both WADD and CADD [13]. The same ARL lower bound (13) as for CuSum also applies to SR.

2.3. Bayesian Objective

In some settings, it may be possible to consider the change point ν as a random variable with a known distribution $\pi_n = \mathbb{P}(\nu = n)$. Optimal stopping times under this formulation were first studied by Shiryaev [7] under the loss function

$$B(T, \lambda) = \mathbb{P}^{\pi}(T < \nu) + \lambda \mathbb{E}^{\pi}[(T - \nu)^+], \tag{17}$$

where the notation $\mathbb{P}^\pi, \mathbb{E}^\pi$ is used to highlight that $\nu \sim \pi$. With an appropriate choice of λ, the objective (17) is equivalent to the following constrained optimization problem:

$$\text{minimize } \mathbb{E}^\pi[(T-\nu)^+] \quad \text{subject to } \mathbb{P}^\pi(T < \nu) \leq \alpha, \tag{18}$$

i.e., minimizing the average detection delay under a constraint that the probability of false alarm (PFA) $\mathbb{P}^\pi(T < \nu)$ is smaller than some α. When $\nu \sim \text{Geom}(\rho)$, Shiryaev [7] showed that the exactly optimal test for (17) is a threshold test on the posterior probability of change

$$p_n = \mathbb{P}^\pi(\nu \leq n | \mathcal{F}_n). \tag{19}$$

The probability p_n can be computed recursively as

$$p_n = \frac{\tilde{p}_{n-1} L(X_n)}{\tilde{p}_{n-1} L(X_n) + (1 - \tilde{p}_{n-1})}, \tag{20}$$

where $\tilde{p}_n = p_n + (1-p_n)\rho$ and $p_0 = \pi_0 = \rho$. Shiryaev's test T_b^S is then performed by comparing the probability p_n to a stopping threshold $b > 0$, i.e.,

$$T_b^S = \inf\{n : p_n \geq b\}. \tag{21}$$

As $\rho \to 0$, i.e., the prior distribution becomes increasingly uninformative, the limit of posterior odds $p_n/(1-p_n)$ converges to the Shiryaev–Roberts statistic (14). As for CuSum and SR, finding a threshold b such that the constraint in (18) is satisfied with equality is usually only possible numerically. However, by the law of iterated expectation,

$$\mathbb{P}^\pi(T_b^S < \nu) = \mathbb{E}^\pi(\mathbb{P}^\pi(T_b^S < \nu | \mathcal{F}_{T_b^S})) = \mathbb{E}^\pi(1 - p_{T_b^S}) \leq 1 - b, \tag{22}$$

where inequality follows from the definition of T_b^S (21). The probability of a false alarm constraint in (18) is then satisfied by setting $b = 1 - \alpha$.

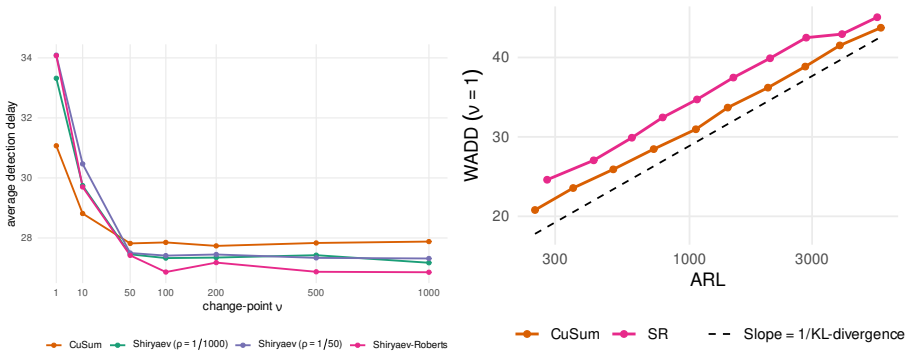
Shiryaev's test is also asymptotically optimal as a false alarm probability $\alpha \to 0$ in a large class of non-i.i.d. models and for general prior distributions π [14].

2.4. Comparison of Shiryaev's Test and CuSum

While the non-Bayesian and Bayesian procedures have very different notions of optimality, their empirical performance can be compared under common criteria. To illustrate this by a concrete example, we simulate a simple change-detection setting of a mean shift from 0 to $\mu = 0.5$ under Gaussian noise, so that $f_0 = \mathcal{N}(0,1), f_1 = \mathcal{N}(\mu, 1)$. Four algorithms are considered, the CuSum test (9)(T_b^C), SR procedure (15) (T_b^{SR}), and two versions of Shiryaev's test; one with parameter $\rho = 1/1000$ ($T_{b,1000}^S$) and the other with $\rho = 1/50$ ($T_{b,50}^S$). For each procedure T, the detection threshold b is set such that $\text{ARL}(T) = \mathbb{E}_\infty(T) = 1000$ for purposes of comparison, where the values are found by 5000 Monte Carlo simulations. Note that the Shiryaev procedures are typically calibrated to have a prespecified PFA $\mathbb{P}^\pi(T < \nu)$ rather than ARL. PFA depends on π, the assumed prior distribution of the change point ν, which may not be available. The ARL guarantee holds regardless of ν, even for the Bayesian procedures.

The average detection delay $\mathbb{E}_\nu(T - \nu | T \geq \nu)$ is evaluated for different values of ν in Figure 1a for all procedures. This delay metric is different than either CADD and WADD, which include taking the maximum over ν. For all tests, the largest detection delay occurs when the change happens early at time $\nu = 1$. This is intuitively explained by the fact that the test statistics that are typically updated using a recursive formula are guaranteed to be far away from the threshold if the change takes place in the very beginning of the

monitoring process. The optimality of CuSum in terms of WADD is observed from the figure, since its maximum delay over ν is the smallest. On the other hand, for larger values of the change point, the detection delay of Shiryaev–Roberts or the Bayesian procedures is smaller. The performance of the Shiryaev and SR procedures is similar, which is to be expected, as in the limit $\rho \to 0$, the Shiryaev procedure converges to SR. In Figure 1b, the WADD of CuSum and SR is plotted as function of $\log(\mathrm{ARL})$. The relationship is linear, and the slope given by the reciprocal of KL divergence $(\mu^2/2)^{-1}$ matches the asymptotic lower bound (5).



(**a**) (**b**)

Figure 1. Change point detection algorithms trade off between controlling false alarm rates (such as average run length (ARL) to false alarm) and average detection delay. (**a**) Average detection delay as a function of different values of the change point ν. CuSum procedure minimizes worst-case detection delay, but for many values of ν, the delay of Shiryaev–Roberts or Bayesian Shiryaev tests (calibrated to have the same ARL) is smaller. (**b**) Worst-case average detection delay for CuSum and SR (which happens when $\nu = 1$ as observed in the left panel) as a function of $\log(\mathrm{ARL})$. For both algorithms, the detection delay grows linearly with respect to log(ARL), and the slope is given by the reciprocal of KL divergence, in accordance with the asymptotic lower bound Equation (5).

2.5. Alternative Performance Criteria

In single-stream QCD, minimizing some notion of average detection delay while controlling the ARL to a false alarm (in non-Bayesian settings) or probability of a false alarm (in Bayesian settings) is the most common performance objective. However, in some applications, other performance criteria may also be of interest. For instance, if T represents the time when a machine part deemed to be faulty is replaced, it may make sense to penalize, for example, the average "size of miss" $\mathbb{E}^\pi|T-\nu|$. More generally, the objective trades off between early and late detection,

$$R(T,\lambda) = \mathbb{E}^\pi(\nu - T)^+ + \lambda \mathbb{E}^\pi(T-\nu)^+, \tag{23}$$

where λ is a tunable parameter that controls the relative importance of the two criteria. Under such a penalty, replacing the part before it is actually faulty is penalized less if the replacement is performed close to the eventual time of failure ν. It turns out, that under some conditions, the Shiryaev test T_b^S is also optimal for the risk $R(T,\lambda)$ [15].

In non-Bayesian settings, exactly optimal algorithms are known in relatively few cases besides the optimality of CuSum for Lorden's delay with i.i.d data. As an interesting example, Poor [16] considered an exponential delay penalty of the form

$$\frac{\beta^{(T-\nu+1)^+} - 1}{\beta - 1} \tag{24}$$

for some constant $\beta \neq 1$. If $\beta > 1$, each additional unit of delay is penalized in an exponentially increasing manner. It may be necessary to associate a higher cost on longer delays when the cost of undetected changes may be enormous. If $\beta < 1$, the cost of delay is saturating and sub-linear, with $\beta \to 1$ converging to the traditional linear detection penalty. For instance, in financial applications, a delayed detection of a shift in the market may incur an exponentially increasing penalty due to the compounding nature of financial gains and losses. Similarly, in applications where the change point triggers a cascade of events (e.g., healthcare, power grid, communication networks, etc.), the effects of delay may be multiplicative, motivating an exponential penalty. Poor showed that a modified version of the CuSum procedure with recursion

$$W_n = \max(0, W_{n-1}) + \ell(X_n) + \log \beta, \quad W_0 = 0 \tag{25}$$

minimizes the worst-case average of the exponential delay in (24) under an ARL constraint. The modified CuSum statistic (25) is closely related to the standard CuSum (12), with the difference being a supplementary constant $\log \beta$ added at each step. Recently, drawing in part on the results of [16], Tartakovsky et al. [17] used a very similar statistic for detection of transient changes, i.e., settings where the change is observable only for a limited period. An optimal procedure that maximizes the probability of detection within the transient period was derived.

3. Recent Advances in Multi-Stream Change Detection

In modern practical applications of sequential change detection, the observed data is usually high-dimensional, observed by a large number of sensors in distinct locations. For example, when a change in the environment or radio spectrum occurs, it may affect one or multiple sensors. The problem is significantly more demanding than single-stream QCD, because both the subset of affected sensors and the post-change distribution in those sensors may not be fully known or are subject to uncertainty. Additionally, the change may occur at different times in different sensors, depending on the physics of the monitored phenomenon or the topology of a networked system.

3.1. High-Dimensional Unknown Parameters

A relatively general formulation of many multi-stream change detection problems can be given as follows. The observations $\mathbf{X}_n = [X_n^{(1)}, \ldots, X_n^{(K)}]$ are K-dimensional random vectors independent over time. However, at the same time instance, the observations acquired by different sensors may be dependent. If the observations are acquired by sensors in distinct locations, it is reasonable to assume that sensor noises are independent, but the signals may be correlated. Let $f(\cdot|\boldsymbol{\theta}_n)$ denote a family of joint probability densities for $\mathbf{X}_n$ parametrized by a vector $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$ of length K. Then, $\mathbf{X}_n \sim f(\cdot|\boldsymbol{\theta}_n)$, where

$$\boldsymbol{\theta}_n = \begin{cases} \mathbf{0}, & n < \nu, \\ \boldsymbol{\theta}_1, & n \geq \nu, \end{cases} \tag{26}$$

meaning that the change manifests by altering a parameter from a known value $\mathbf{0}$ (assumed w.l.o.g.) to an unknown $\boldsymbol{\theta}_1$.

3.1.1. Binary $\boldsymbol{\theta}$

We begin with the simpler setting in which the post-change model is determined solely by the subset of affected sensors. In this case, the unknown parameter $\boldsymbol{\theta}$ is naturally represented as a binary vector, where each component indicates whether the corresponding stream is affected by the change. That is, $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta} = \{0,1\}^K$, and $\boldsymbol{\theta}^{(k)} = 1$ indicates that

the change affects sensor k. For example, if the components of X_n are independent also within the same time instance and have pre- and post-change distributions $f_0^{(k)}$ and $f_1^{(k)}$, respectively, the joint density can be written as

$$f(\mathbf{X}_n|\boldsymbol{\theta}_n) = \prod_{k=1}^{K}[\theta_n^{(k)} f_1^{(k)}(X_n^{(k)}) + (1-\theta_n^{(k)}) f_0^{(k)}(X_n^{(k)})]. \tag{27}$$

If $\boldsymbol{\theta}_1$ is known, the optimal procedure under Lorden's formulation is still the CuSum procedure, with the test statistic

$$W_n^{\boldsymbol{\theta}_1} = \max_{k\le n} \sum_{j=k}^{n} \log \frac{f(X_j|\boldsymbol{\theta}_1)}{f(X_j|\mathbf{0})}. \tag{28}$$

When $\boldsymbol{\theta}$ is unknown, two of the most common ways of handling distributional uncertainty are to either maximize or average over $\boldsymbol{\theta}$ in (28). The former leads to a generalized likelihood ratio (GLR) testing approach, in which the unknown nuisance parameters are estimated, while the latter yields a mixture test. When the post-change parameter space is finite, the two approaches have a lot in common, as demonstrated in [18]. Indeed, the GLR-CuSum statistic is given by

$$\max_{k\le n} \max_{\boldsymbol{\theta}\in\{0,1\}^K} \sum_{j=k}^{n} \log \frac{f(\mathbf{X}_j|\boldsymbol{\theta})}{f(\mathbf{X}_j|\mathbf{0})} \tag{29}$$

$$= \max_{\boldsymbol{\theta}\in\{0,1\}^K} W_n^{\boldsymbol{\theta}}, \tag{30}$$

where the second line follows from interchanging the maximizations, and $W_n^{\boldsymbol{\theta}}$ is the CuSum statistic in which the set of affected sensors is $\boldsymbol{\theta}$. This procedure is equivalent to running a CuSum test for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and declaring a change when one of them reaches a threshold b. More generally, one can specify a different threshold $b_{\boldsymbol{\theta}}$ for each $W_n^{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and consider the stopping time

$$T^G = \inf\left\{ n : \max_{\boldsymbol{\theta}\in\{0,1\}^K} (W_n^{\boldsymbol{\theta}} - b_{\boldsymbol{\theta}}) \ge 0 \right\}. \tag{31}$$

If $b_{\boldsymbol{\theta}}$ is further chosen to be of the form $b_{\boldsymbol{\theta}} = b - \log p_{\boldsymbol{\theta}}$, where $p_{\boldsymbol{\theta}} > 0$ and $\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p_{\boldsymbol{\theta}} = 1$, the stopping time can be written as

$$T^G = \inf\left\{ n : \max_{\boldsymbol{\theta}\in\{0,1\}^K} (W_n^{\boldsymbol{\theta}} + \log p_{\boldsymbol{\theta}}) \ge b \right\} \tag{32}$$

$$= \inf\left\{ n : \max_{\boldsymbol{\theta}\in\{0,1\}^K} p_{\boldsymbol{\theta}} e^{W_n^{\boldsymbol{\theta}}} \ge e^b \right\}. \tag{33}$$

Changing the maximization to a sum over $\boldsymbol{\theta}$ in (33), one immediately obtains a mixture-CuSum test

$$T^M = \inf\left\{ n : \sum_{\boldsymbol{\theta}\in\{0,1\}^K} p_{\boldsymbol{\theta}} e^{W_n^{\boldsymbol{\theta}}} \ge e^b \right\}, \tag{34}$$

where $p_{\boldsymbol{\theta}}$ can be interpreted as a prior distribution for $\boldsymbol{\theta}$. The GLR and mixture tests were analyzed in [18] and shown to be asymptotically optimal in a strong *second-order* sense, i.e., as $\gamma \to \infty$

$$\mathrm{WADD}(T^G) - \inf_{T:\mathrm{ARL}(T)\ge\gamma} \mathrm{WADD}(T) = O(1), \tag{35}$$

for any set of affected sensors $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and similarly for T^M.

The main drawback of the above procedures is that their implementation in principle requires running $|\boldsymbol{\Theta}|$ CuSum-recursions in parallel, and $|\boldsymbol{\Theta}|$ can be as large as 2^K if the set of affected sensors can be completely arbitrary. A significant reduction in computational complexity can be achieved by adaptive windowing [18]. A computationally simpler algorithm is the SUM-CuSum [19] for independent streams, given by

$$T^{\text{SUM}} = \inf\left\{ n : \sum_{k=1}^{K} W_n^{(k)} \geq b \right\}, \tag{36}$$

where $W^{(k)}$ is the CuSum-statistic using observations only from stream k (or equivalently (28) with $\boldsymbol{\theta} = e_k$, the k-th basis vector). The SUM-CuSum test declares a change when the sum of local sensor CuSums crosses a threshold, making it possible to perform most of the computations locally. The procedure was shown to be first-order asymptotically optimum in [19]. It also works well when the change does not necessarily occur at the same time in all affected streams. However, unlike (32) and (34), it is not second-order asymptotically optimal.

3.1.2. Continuous $\boldsymbol{\theta}$

In many applications, just knowing the indices of the affected sensors may not be enough to determine the full post-change distribution. For instance, the change may affect the sensors with different intensities based on, e.g., the distance between the sensor and the (unknown) signal source location. For example, the sensors may be monitoring the energy of radio frequency signals in a spectrum band that attenuate inversely proportionally to squared distance. In such cases, we can still adopt the general model from (26), but represent $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^K$ as a general K-dimensional real vector.

In principle, the general GLR and mixture-CuSum ideas from above can of course be extended to a continuous $\boldsymbol{\theta}$ without issue. However, the possibly high dimension K of the unknown parameter vector $\boldsymbol{\theta}$ presents additional challenges in the continuous case. Indeed, it follows from [20] that when $f(\cdot|\boldsymbol{\theta})$ is in the exponential family, and $\boldsymbol{\Theta}$ satisfies certain regularity conditions, the asymptotic detection delay of any procedure T as ARL $\gamma \to \infty$ satisfies the following minimax lower bound:

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} I(\boldsymbol{\theta}) \text{WADD}(T) \geq \log \gamma + \frac{K}{2} \log \log \gamma + O(1). \tag{37}$$

It is important to point out that in (37), the second-order delay term grows linearly with K. For moderate γ and large K, the influence of the second-order term can be greater than the first-order term.

Deriving procedures which are effective for high-dimensional $\boldsymbol{\theta}$ has attracted the most attention in the literature in the special case when $\boldsymbol{\theta}$ corresponds to the mean vector of a multivariate Gaussian distribution with diagonal covariance. The proposed ideas and the developed methods extend to more complicated models and different probability models.

Xie and Siegmund [21] developed a mixture procedure where each stream is hypothesized to be affected by a change with probability p. The stopping time is then given by the threshold rule

$$T^{XS} = \inf\left\{ n : \max_{0 \leq m \leq n} \sum_{k=1}^{K} \log\left(1 - p + p \exp\left[\max_{\boldsymbol{\theta}} \sum_{j=m}^{n} \log \frac{f(X_j^{(k)}|\boldsymbol{\theta})}{f(X_j^{(k)}|\mathbf{0})} \right] \right) \geq b \right\}. \tag{38}$$

The term inside the square brackets is the GLR statistic for stream k for observations between time instances m and n. Therefore, setting $p = 1$ recovers the standard GLR-

CuSum. The test is shown to be effective if the post-change vector $\boldsymbol{\theta}$ is sparse, i.e., only a small proportion of its components are non-zero. In the same vein, Chan [22] proposed detectability-score transformations of stream-wise GLR statistics. The central idea is to down-weight weak local evidence that would correspond to signals too weak to be reliably detected with the current sample size. This reduces the accumulation of noise from unaffected streams and yields an asymptotically near-optimal detection delay across a range of sparsity regimes. Asymptotic lower bounds for the detection delay in the limit $K \to \infty$ are also derived, which is different from the usual $\gamma \to \infty$ asymptotics in QCD. Chen et al. [23] propose an algorithm that simultaneously tracks various likelihood ratios corresponding to individual or multiple sensors, and uses hard thresholding to suppress likelihood ratios likely corresponding to noise in the sparse case. Using different levels for a hard threshold in parallel makes the test adaptable to different levels of sparsity in the post-change mean vector $\boldsymbol{\theta}$.

The usefulness of thresholding in detecting a change to a sparse $\boldsymbol{\theta}$ is related to the effectiveness of thresholding in sparse estimation; see, e.g., [24] and references therein. Indeed, a change detection problem with unknown parameters may be seen as a combination of an estimation and detection problem; the better $\boldsymbol{\theta}$ is known (estimated), the easier a change is to detect. Another popular technique used in high-dimensional problems is shrinkage estimation [25]. A natural way to incorporate a particular estimator to a change detection algorithm is to use a so-called adaptive change detection procedure [26–29]. For example, in the CuSum recursion (12), one can replace the unknown f_1 (here, $f(\cdot|\boldsymbol{\theta})$) has a plug-in $f(\mathbf{X}_n|\hat{\boldsymbol{\theta}}_n)$, where $\hat{\boldsymbol{\theta}}_n$ is some estimator of $\boldsymbol{\theta}$ based on the data up to time instance n). Xie et al. [29] proposed estimating $\boldsymbol{\theta}$ based on a sliding window of past w observations, i.e., $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(\mathbf{X}_{n-1}, \ldots, \mathbf{X}_{n-w})$. This yields the window-limited CuSum statistic

$$W_n^{\mathrm{WL}} = \max\{0, W_{n-1}\} + \log \frac{f(\mathbf{X}_n|\hat{\boldsymbol{\theta}}_n)}{f(\mathbf{X}_n|\mathbf{0})}, \quad W_0 = 0. \tag{39}$$

An important feature of this statistic is that it can be calculated recursively, which is not in general the case for, e.g., GLR type tests, as remarked in Section 3.1.1. Another advantage is that prior knowledge/structure (such as sparsity) about $\boldsymbol{\theta}$ can be easily incorporated by selecting the estimator $\hat{\boldsymbol{\theta}}$ accordingly. For example, shrinkage estimators reduce estimator variance by introducing some bias, often leading to reduced loss under an appropriate loss function. Wang and Mei [30] considered thresholding as well as linear shrinkage estimators of the form $a\hat{\boldsymbol{\theta}}^{\mathrm{ML}} + b$, where $\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$ is the maximum likelihood estimator. Later, it was demonstrated in [31] that using the James–Stein (JS) estimator [32] in place of the standard maximum likelihood estimator in the WL-CuSum test (39) can reduce detection delay for all values of $\boldsymbol{\theta}$ in the Gaussian case. This stems from the fact that the James–Stein estimator dominates the ML estimator in terms of the mean-square error when $K \geq 3$ [32]. The JS estimator is most effective when $\boldsymbol{\theta}$ is dense rather than sparse, i.e., when the change event causes subtle disturbances to many sensors. A simple way to create a procedure that is effective for both dense and sparse $\boldsymbol{\theta}$ simultaneously would be to run two versions of (39) in parallel: one with an estimator $\hat{\boldsymbol{\theta}}_n$ effective for dense changes (such as JS) and one for sparse changes (such as a hard-thresholding estimator [[24], Ch. 5]). Forming and analyzing such combined procedures in a principled manner is an interesting topic for future research.

3.1.3. Numerical Comparison of Methods for High-Dimensional Gaussian Mean Shifts

As an illustration of the properties of the methods discussed above, we present a brief numerical comparison. We consider a model with $K = 100$ data streams, where before the change,

$$f_0 = \mathcal{N}(\mathbf{0}, I_K),$$

and after the change,

$$f_1 = \mathcal{N}(\boldsymbol{\theta}, I_K).$$

Although simple, this setting is not unduly restrictive: more general Gaussian models with correlated components can often be reduced to this form by whitening, provided that the pre-change covariance matrix is known or can be estimated accurately.

We compare six procedures: T^{XS} in (38) from [21], Chan's detectability score procedure [22], the OCD procedure from [23], two versions of the WL-CuSum procedure (39) from [29] based on the ML and James–Stein estimators [31], and the classical GLR-CuSum procedure [11]. Since XS, Chan, and GLR do not admit recursive updates, we implement window-limited versions with window size 200. For the XS and Chan procedures, we set $p = 1/\sqrt{K} = 0.1$, as recommended in [22]. Parallel versions of the WL-CuSum tests are run (see [29]) with window sizes ranging from 30 to 80. The XS, Chan, and OCD procedures are implemented using the R package `ocd` [33]. The stopping thresholds of all procedures are set via 1000 Monte Carlo simulations such that $\mathrm{ARL}(T) \approx 1000$.

We consider changes affecting $L \in \{1, 25, 50, 100\}$ of the $K = 100$ streams. For a given value of L, the post-change vector $\boldsymbol{\theta}$ is generated at every MC iteration as follows: for $1 \le k \le K$, we first sample $\xi^{(k)} \sim \mathcal{N}(0.5, 1)\mathbf{1}\{k \le L\}$ and then set $\boldsymbol{\theta}^{(k)} = \xi^{(k)} / \sqrt{\sum(\xi^{(k)})^2}$. This normalization ensures that the KL divergence satisfies $I = \|\boldsymbol{\theta}\|^2/2 = 1/2$ at every iteration. The change point is fixed at $\nu = 100$. Figure 2 reports Monte Carlo estimates of the average detection delay $\mathbb{E}_\nu(T - \nu \mid T \ge \nu)$. The results show a clear dependence on sparsity. For sparse changes, the XS, Chan, and OCD procedures perform best. As the change becomes denser, their advantage diminishes, whereas the James–Stein WL-CuSum becomes increasingly competitive. By contrast, the ML-WL-CuSum and GLR procedures are comparatively insensitive to the number of affected streams. Importantly, for both very sparse and very dense $\boldsymbol{\theta}$, substantial performance gains over standard methods like GLR are available when the number of sensors is large.

Overall, the methods reviewed above incur different tradeoffs between generality, structural adaptivity, and computational complexity. GLR-based procedures are widely applicable, but computationally intensive. SUM-CuSum offers a simpler and recursively implementable alternative, at the expense of weaker asymptotic optimality properties. The methods by Xie and Siegmund [21] and Chan [22] are specifically designed for sparse changes, but work reasonably well in denser regimes as well. However, their computational complexity is similar to GLR-CuSum. The OCD procedure [23] can be updated recursively, and can adapt to different levels of sparsity. Finally, the WL-CuSum-based procedures [29,31] also admit recursive updates, and can adjust to a possible post-change structure with the choice of estimator.

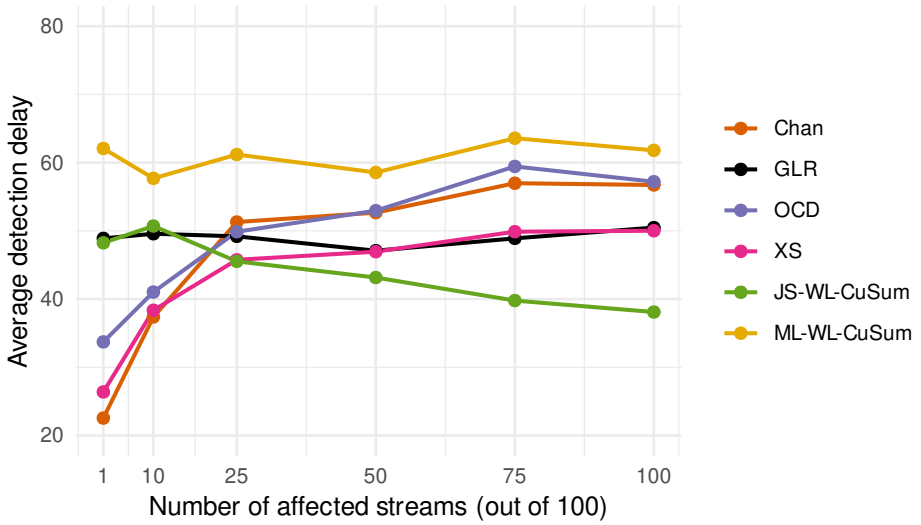


Figure 2. Average detection delays highlight a clear dependence on the sparsity of the change. For very sparse changes, XS [21], Chan [22], and OCD [23] procedures achieve the smallest delays, whereas their advantage diminishes as more streams are affected. In contrast, the James–Stein WL-CuSum [31] is ncreasingly effective in denser regimes, while the ML-WL-CuSum and GLR procedures remain comparatively stable across sparsity levels. Overall, it is observed that substantial gains over standard methods are possible in high-dimensional problems when the detection procedure is matched to the structure of the post-change signal.

3.2. QCD Under Sampling Constraints

In the previous subsection, we assumed that all sensors acquire data continuously at each time instance. In practice, it is often necessary to reduce the communication and sampling overhead by sampling only some subset of streams at a given time. By designing intelligent sensing policies and methods for sensor selection, significant savings in computational complexity and communications overhead may be achieved. Let us denote by $\mathcal{A}_n \subset \{1, \dots, K\}$ the indices of streams monitored at time n, meaning that, with some abuse of notation,

$$X_n^{(k)} = \begin{cases} X_n^{(k)} & \text{if } k \in \mathcal{A}_n \\ \emptyset & \text{otherwise,} \end{cases} \tag{40}$$

where $\emptyset$ denotes an empty symbol/missing observation. The sampling choice $\mathcal{A}_n$ can depend on previous data, i.e., $\mathcal{A}_n \in \mathcal{F}_{n-1}$. The size $\mathcal{A}_n$ is usually constrained, for example, $|\mathcal{A}_n| \leq L \; \forall n$ if at most $L < K$ of the sensors can be deployed at a given time. Under such a constraint, the typical QCD objective is now expanded to designing both a sampling policy and a detection rule.

Dragalin [34] considers the case $L = 1$ and proposes a scanning rule for stream selection in which a particular stream is sampled until its CuSum statistic hits either 0 or a stopping threshold b. If the statistic hits 0, the sampling moves to the next data stream. The procedure is shown to be asymptotically optimum under Pollak's CADD. Lai et al. [35] consider a closely related problem with $L = 1$ in a sequential hypothesis testing context. In an infinite sequence of streams, a signal is present independently in each stream with equal probability. The objective is to find an affected stream as quickly as possible, while sampling only a single stream at a time. It is shown that within a Bayesian formulation seeking minimal delay and false detection probability, successively running a CuSum test until either a reset and switch to another stream occurs, or detection in the current stream is exactly optimal. Recently, ref. [36] considered a similar rule in a QCD context, and showed

that such a cyclic procedure is asymptotically second-order optimal under Lorden's WADD. Moreover, the data distributions in different streams are allowed to be different.

The general idea in the above works may be extended to the case when $1 \leq L \leq K$ streams are observed at a time by, e.g., computing and updating a detection statistic for each stream, and sampling those with the L currently highest values. However, as observed in [37], especially if K is much larger than L, it may be beneficial to "increase exploration" by incrementing the detection statistics of all unobserved streams at each step by a small constant Δ. This way, the procedure cannot get stuck sampling a single stream for a long time. Determining an "optimal" value for Δ is not straightforward, and even the analysis of the procedure is challenging for $\Delta > 0$, but small values of $\Delta > 0$ are shown to perform better empirically than $\Delta = 0$ [38].

The above works assume that sampled data streams are independent, or at the very least that the change is observable from the marginal distributions of individual streams. If the data are acquired by distinct sensors, noise can be assumed to be independent while the signals may be correlated or otherwise dependent. More general observation models were studied recently in [39], where the change is assumed to affect the joint distribution of an unknown subset of sensors; for example the correlation structure. Conditions under which the cyclic, rotating order sampling policy remains asymptotically optimal were derived. Previously, in, e.g., [36], asymptotic optimality when only a single sensor observes the change was established with respect to a procedure that has access to all samples. In [39], asymptotic optimality is studied only with respect to procedures that satisfy the sampling constraint. In this class, the simple sampling policy of [36] is first-order asymptotically optimal even when an arbitrary number of sensors observe the change under the assumption of homogeneity of the statistical models.

The cyclic policies are computationally simple to implement and do not require storing a lot of past data. From a broader perspective, designing the sampling policy can be seen to be closely related to the multi-armed bandit (MAB) [40] problem. The core challenge in the MAB problem is the exploration vs. exploitation tradeoff: choosing the arm (stream) with the highest expected reward (exploitation) vs. choosing an arm (stream) that may have a higher reward than the current best arm (exploration). As an example, Restless Multiarm Bandit (RMAB) [41,42] and Bayesian Thompson sampling methods have been used for learning sensing policy and stream selection while balancing between exploitation and exploration in multi-sensor systems. In a change detection setting, however, the objective is not to minimize, e.g., cumulative regret, but for instance average detection delay. In case of long-term rewards, associating a per-action reward to each action (sampling choice) is therefore not obvious, although a large observed likelihood ratio could represent a reasonable proxy. In fact, from this perspective, the cyclic sampling policies discussed above can be seen as greedy (or follow-the-leader) selection policies. In general, myopic learning policies that maximize immediate or near-term rewards are more appropriate for stream or sensor selection in the QCD context.

A natural and fundamental algorithm for sampling that often works well in practice is Thompson sampling in a Bayesian framework. The idea is to define a distribution over possible environments, and at each time make the optimal action with respect to an environment drawn at random from the current posterior distribution. In a QCD context, this would mean specifying a (joint) prior distribution for the change points, computing the posterior probabilities of change having happened at each stream (i.e., the Shiryaev statistics (21)), and sampling according to these probabilities. Such a policy is studied in [43]. If specifying a prior distribution is infeasible, using the Shiryaev–Roberts statistics (14) is a logical substitute. A Thompson sampling-related sampling policy based on randomized SR statistics is proposed and analyzed in [44].

Finally, Veeravalli et al. [45] consider a generalized setup where possible sampling actions $\mathcal{A}_n$ do not have to correspond to subsets of sensors, but represent a general (finite) set of sampling actions. For example, different sampling actions could represent different ways of steering a sensor array. Depending on the true target parameters $\boldsymbol{\theta}$, some sampling actions a are more effective than others for distinguishing the change. The effectiveness is measured by the post-change KL divergence

$$I^a(\boldsymbol{\theta}) = \int f^a(x|\boldsymbol{\theta}) \log \frac{f^a(x|\boldsymbol{\theta})}{f^a(x|\mathbf{0})}, \tag{41}$$

where $f^a(\cdot|\boldsymbol{\theta})$ is the sampling distribution of the data under action a when the true parameter is $\boldsymbol{\theta}$. For example, if $\boldsymbol{\theta}$ is a binary vector as in (27), $\mathcal{A}_n$ is a subset of sensors, and the components of $\mathbf{X}_n$ are independent,

$$f^{\mathcal{A}_n}(\mathbf{X}_n|\boldsymbol{\theta}) = \prod_{k \in \mathcal{A}_n} [\theta_n^{(k)} f_1^{(k)}(X_n^{(k)}) + (1 - \theta_n^{(k)}) f_0^{(k)}(X_n^{(k)})]. \tag{42}$$

Naturally, the optimal action is $a^* = \arg\max_a I^a(\boldsymbol{\theta})$, but this depends on the unknown $\boldsymbol{\theta}$. As a solution, it is proposed in [45] to (1) estimate $\boldsymbol{\theta}$ with a maximum likelihood sliding window estimator $\hat{\boldsymbol{\theta}}_n$ based on past observations, and (2) select the next action $\mathcal{A}_n$ as either $\mathcal{A}_n = \arg\max_a I^a(\hat{\boldsymbol{\theta}}_n)$ or uniformly at random at deterministic exploration times. The form of the sampling policy goes all the way back to Chernoff and the sequential design of experiments in hypothesis testing [46]. Sampling policies based on Chernoff's rule are first-order asymptotically optimal across a wide range of controlled sensing tasks. However, its (empirical and theoretical) performance can be sub-optimal when the choice of an action can change the effect of future actions, as noted recently in [47].

3.3. QCD with False Discovery Rate Control

So far, we have considered settings where only a single change affecting the monitored system is detected. In many applications, changes can happen at different times at different sensors, possibly caused by the same or different phenomena. For example, in a sensor network, each sensor may be monitoring a different physical process, and a change in one process may or may not affect the others. When a change takes place in one stream, it can require an action to be taken only in that stream. In these cases, it may be necessary to detect a change point separately in each stream. Moreover, it is of interest to consider performance criteria that measure the performance across the streams as a whole, instead of separately optimizing the performance in each individual stream. For example, in a cognitive radio network [48], secondary users opportunistically sense the channels for available transmission opportunities. Even if the individual channels are independent, it is important to limit the total amount of interference or collisions caused by the SUs as a whole.

Let us denote the time of detection in the kth stream by $T^{(k)}$, and by $\nu^{(k)}$ the actual change point in stream k. If a QCD procedure that controls the ARL (1) individually is applied to each stream, one approximately obtains control over the average *number* of false-alarm events that can occur over all streams in a given time interval. In practice, it might be more useful to have control over the *proportion* of false detections among all detections. To this end, a natural criterion well-suited for large-scale applications is the false discovery rate (FDR) [49]. In a change-detection setting, the FDR at time n for a set

of stopping times $\mathcal{T} = \{T^{(1)}, \ldots, T^{(K)}\}$ is defined as the expected ratio of false detections $V(n)$ made by time n to all detections $R(n)$ by n:

$$\mathrm{FDR}_{\boldsymbol{\nu}}(\mathcal{T}, n) = \mathbb{E}_{\boldsymbol{\nu}}\left[\frac{V(n)}{\max\{1, R(n)\}}\right], \tag{43}$$

where

$$V(n) = \sum_{k=1}^{K} \mathbf{1}\left\{T^{(k)} < \nu^{(k)}, T^{(k)} \le n\right\} \tag{44}$$

$$R(n) = \sum_{k=1}^{K} \mathbf{1}\left\{T^{(k)} \le n\right\}. \tag{45}$$

The denominator in (43) includes the maximum only to prevent division by zero.

In the FDR literature, the typical starting point is a set of p-values corresponding to fixed-sample size hypothesis tests and the ubiquitous Benjamini–Hochberg procedure [49]. The BH-procedure was extended to sequential hypothesis tests by Bartroff and Song [50]. Utilizing some ideas from [50], Chen et al. [51] were the first to study multi-stream QCD procedures with FDR control. Under the Bayesian change point formulation (Section 2.3), a procedure that utilizes the posterior probabilities (19) as the test statistics in a sequential BH-type procedure is proposed. It is shown that the procedure controls the FDR averaged over the distribution of the change points $\pi(\boldsymbol{\nu})$ asymptotically as $K \to \infty$, i.e., for all fixed n and a user-defined FDR-threshold $0 < \alpha < 1$:

$$\mathrm{FDR}^{\pi}(\mathcal{T}, n) = \sum_{\boldsymbol{\nu}} \pi(\boldsymbol{\nu}) \mathrm{FDR}_{\boldsymbol{\nu}}(\mathcal{T}, n) \le \alpha + o(1). \tag{46}$$

Later, it was observed in [43] that in the Bayesian setting, it is in fact sufficient to use a single common stopping threshold for all streamwise tests, and such a procedure controls the FDR non-asymptotically. That is, (46) is achieved without the $o(1)$ term. In a simple setting, even an exactly optimal procedure for $\mathrm{FDR}^{\pi}(\mathcal{T}, \infty)$ can be derived [52] as a concatenation of Shiryaev tests (21) with specifically chosen stopping thresholds.

Under a non-Bayesian formulation, the natural objective in the spirit of Pollak's CADD (4) would be to control the worst-case FDR over all values of the change point, namely

$$\sup_{\boldsymbol{\nu}} \mathrm{FDR}_{\boldsymbol{\nu}}(\mathcal{T}, n) \le \alpha. \tag{47}$$

This condition is clearly stronger than in the Bayesian setting, since controlling the FDR for all individual $\boldsymbol{\nu}$ clearly implies that the FDR averaged over any distribution $\boldsymbol{\nu} \sim \pi$ is controlled as well. For example, (47) implies that $\mathrm{FDR}_{\infty}(\mathcal{T}, \infty) \le \alpha$, which in turn implies

$$P_{\infty}\left(\bigcup_{k=1}^{K} \{T^{(k)} < \infty\}\right) \le \alpha. \tag{48}$$

That is, in the case where no change ever occurs, the probability of any procedure declaring a change should be smaller than α. Such a condition is more stringent than the typical ARL used in single-stream change detection. Indeed, $P_{\infty}(T < \infty) < 1$ implies $\mathrm{ARL}(T) = \infty$. It is expected that a procedure that controls (47) will suffer large detection delays, at least in the worst case. Nonetheless, procedures that control (47) were developed in [53,54]. The methods consist of running a CuSum test in each stream with certain stopping thresholds that are increasing over time. As a result, the average detection delay can be large if the change happens late, since by the time the change happens, the stopping threshold has increased to a high level.

Recently, [55] consider a slightly more general version of (47), where FDR is controlled not only at any fixed n, but also any stopping time τ. For example, τ could be the time m changes have been detected. Since constructing a detection procedure with a small (or even finite) worst-case average detection delay is infeasible due to the reasons discussed above, a criterion called error over patience (EOP) is proposed:

$$\mathrm{EOP}(\mathcal{T}) = \sup_{\nu \geq 1} \sup_{\tau} \frac{\mathbb{E}(\mathrm{FDR}_\nu(\mathcal{T}, \tau))}{\mathbb{E}(\tau)}, \tag{49}$$

where the inner supremum is taken over all stopping times (Definition 1). In essence, the criterion allows the FDR to be larger if the process is terminated late, i.e., when τ is large. This way, control over the FDR does not need to be maintained for an indefinitely long time.

The definition of FDR we have adopted (43) measures the quality of detections made up to time n. In some applications, it could be of interest to consider the "quality" of the streams still active at time n. For example, the false non-discovery rate (FNR) at time n, defined as

$$\mathrm{FNR}_\nu(\mathcal{T}, n) = \mathbb{E}_\nu \left[\frac{\sum_{k=1}^{K} \mathbf{1}\left\{ \nu^{(k)} \leq n, T^{(k)} > n \right\}}{\max\{1, K - R(n)\}} \right], \tag{50}$$

measures the proportion of post-change streams among all streams where a change has not yet been detected. Discussions of many related performance metrics such as FNR and algorithms controlling them can be found in [56,57].

4. Future Outlook

In this brief overview, the discussion was mostly limited to parametric models. In modern applications, it may not always be possible to formulate the data distributions using parametric models. Many approaches that replace the likelihood ratio with some nonparametric data-driven statistic and empirical probability models have been proposed recently. For example, tests based on kernel minimum mean discrepancy (MMD) have been recently proposed [58,59] as interesting ideas. Many other ways of finding statistics that discriminate between the pre-change and post-change distributions in place of the likelihood ratio naturally exist as well; see, for example, [[60], Sec. IV-B]. Empirical likelihoods (ELs) and likelihood ratios provide a data driven, non-parametric approach while being able to take advantage of some benefits of likelihood-based sequential inference. Moreover, conformal p-values computed using training data from the pre-change probability model to indicate a change in distribution, outliers, or out of distribution data may be employed. They have found applications in sequential inference in single and multi-stream settings and multi-sensor systems performing spatio-temporal inference.

One recently proposed general framework for how statistics other than the likelihood ratio can be used for (single-stream) QCD is proposed by Shin et al. [61]. It is shown that the general CuSum and Shiryaev–Roberts procedures can also be used when the likelihood ratio is replaced by a so-called e-process. E-processes may be seen as a nonparametric generalization of the likelihood ratio.

Optimal detectors and estimators are derived based on a particular parametric signal model and strict assumptions on the underlying probability models. However, optimality is only achieved when the underlying assumptions hold. The performance of optimal procedures may deteriorate significantly, even for minor departures from the assumed model. Robust statistical methods are semi-parametric, namely taking advantage of parametric models but trade optimality for reliability. Statistically robust detectors are close-to-optimal in nominal conditions and highly reliable for real-life data, even if the assumptions are

only approximately valid. Robust procedures are an attractive approach in QCD problems where outliers may be present in the data, or the assumptions in probability models do not hold exactly; see, e.g., [62–64] and references therein. Robust procedures can distinguish among data from different populations and outliers. The parameters of the distribution may then be estimated from the population associated with the majority of the data, and the outliers and data from minority population will have a bounded influence.

Machine learning approaches provide an attractive approach for change point detection in the presence of modeling deficits. Probability models and their parameters, or the likelihood ratios directly [65], can then be learned from data [66,67]. In highly dynamic and high-dimensional scenarios, learning sensing policies allows for allocating sensing, computation, and inference resources to a subset of data streams where the change is likely to take place soon. Consequently, significant savings in computation and energy consumption may be achieved. Different variants and recent advances of reinforcement learning, including MAB and Partially Observable Markov Decision Processes (POMDP), provide a framework for learning optimized sensing policies. Statistical censoring methods can be used to reduce data transmissions and energy consumption in networked, multi-stream change point detection, since only informative decision statistics are transmitted. Only minor performance losses in inference are experienced even under rather strict communications rate constraints.

No matter the exact approach that is used in a non-parametric setting, the general challenges highlighted in this paper still persist. High dimensionality, sensing constraints, and multiplicity of decisions continue to complicate both algorithm design and performance analysis.

In a broader context, detecting rapid changes, disruptions, and adversarial events is a key capability in resilient systems that are able to identify such undesirable events, recover and reconfigure to continue normal operation, and provide desired trustworthy sequential inference performance with rigorous performance guarantees. Detecting changes is also crucial in creating situational awareness exploited by any adaptive or cognitive system.

5. Conclusions

This overview surveyed some key models and methods for multi-stream quickest change detection, with emphasis on three recurring challenges: high-dimensional unknown post-change parameters, multiple-hypothesis testing inspired error rate control, and adaptive sensing under resource constraints. Across these settings, a common message is that effective procedures must balance statistical optimality with computational and operational feasibility.

Several important challenges remain open. In high-dimensional settings, a central difficulty is to design procedures that retain strong optimality properties while remaining computationally feasible when the number of streams is very large. Beyond the parametric models emphasized here, there is also a growing need for methods that operate reliably when the pre- and post-change distributions are only partially specified or must be learned from data. In controlled sensing problems, an important direction is to develop adaptive sampling rules with rigorous performance guarantees under dependence, heterogeneity, and resource constraints. Finally, in settings with multiple streamwise detections, further work is needed to combine local detection efficiency with performance criteria that measure the global performance. A related challenge is to formulate performance criteria that are both achievable in practice and sufficiently stringent to control the error events and tradeoffs that matter most in applications.

https://doi.org/10.3390/e28050566

Author Contributions: Conceptualization, T.H. and V.K.; methodology, T.H. and V.K.; software, T.H.; validation, T.H.; formal analysis, T.H.; writing—original draft preparation, T.H.; writing—review and editing, T.H. and V.K.; visualization, T.H.; supervision, V.K.; project administration, V.K.; funding acquisition, V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported, in part, by Research Council of Finland grant 359094, Business Finland grant 6G-ISAC, and SoW 261.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The authors would like to thank H. Vincent Poor for insightful discussions on the topic of quickest change detection, and for the fruitful long-term collaboration on a variety of signal processing topics.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Poor, H.V.; Hadjiliadis, O. *Quickest Detection*; Cambridge University Press: Cambridge, UK , 2008.
2. Tartakovsky, A.; Nikiforov, I.; Basseville, M. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*; CRC Press: Boca Raton, FL, USA, 2014.
3. Shewhart, W.A. The application of statistics as an aid in maintaining quality of a manufactured product. *J. Am. Stat. Assoc.* **1925**, *20*, 546–548. [CrossRef]
4. Shewhart, W.A. Economic quality control of manufactured product. *Bell Syst. Tech. J.* **1930**, *9*, 364–389. [CrossRef]
5. Page, E.S. Continuous inspection schemes. *Biometrika* **1954**, *41*, 100–115. [CrossRef]
6. Wald, A. *Sequential Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1947.
7. Shiryaev, A.N. On optimum methods in quickest detection problems. *Theory Probab. Its Appl.* **1963**, *8*, 22–46. [CrossRef]
8. Shiryaev, A.N. Some exact formulas in a "disorder" problem. *Theory Probab. Its Appl.* **1965**, *10*, 348–354. [CrossRef]
9. Lorden, G. Procedures for reacting to a change in distribution. *Ann. Math. Stat.* **1971**, *42*, 1897–1908. [CrossRef]
10. Pollak, M. Optimal detection of a change in distribution. *Ann. Stat.* **1985**, *13*, 206–227. [CrossRef]
11. Lai, T.L. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inf. Theory* **1998**, *44*, 2917–2929. [CrossRef]
12. Moustakides, G.V. Optimal stopping times for detecting changes in distributions. *Ann. Stat.* **1986**, *14*, 1379–1387. [CrossRef]
13. Polunchenko, A.S.; Tartakovsky, A.G. On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *Ann. Stat.* **2010**, *38*, 3445–3457. [CrossRef]
14. Tartakovsky, A.G.; Veeravalli, V.V. General asymptotic Bayesian theory of quickest change detection. *Theory Probab. Its Appl.* **2005**, *49*, 458–497. [CrossRef]
15. Shiryaev, A.N. A remark on the quickest detection problems. *Stat. Decis.* **2004**, *22*, 79–82. [CrossRef]
16. Poor, H.V. Quickest detection with exponential penalty for delay. *Ann. Stat.* **1998**, *26*, 2179–2205. [CrossRef]
17. Tartakovsky, A.G.; Berenkov, N.R.; Kolessa, A.E.; Nikiforov, I.V. Optimal sequential detection of signals with unknown appearance and disappearance points in time. *IEEE Trans. Signal Process.* **2021**, *69*, 2653–2662. [CrossRef]
18. Fellouris, G.; Sokolov, G. Second-Order Asymptotic Optimality in Multisensor Sequential Change Detection. *IEEE Trans. Inf. Theory* **2016**, *62*, 3662–3675. [CrossRef]
19. Mei, Y. Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **2010**, *97*, 419–433. [CrossRef]
20. Siegmund, D.; Yakir, B. Minimax optimality of the Shiryayev–Roberts change-point detection rule. *J. Stat. Plan. Inference* **2008**, *138*, 2815–2825. [CrossRef]
21. Xie, Y.; Siegmund, D. Sequential multi-sensor change-point detection. *Ann. Stat.* **2013**, *41*, 670–692. [CrossRef]
22. Chan, H.P. Optimal sequential detection in multi-stream data. *Ann. Stat.* **2017**, *45*, 2736–2763. [CrossRef]
23. Chen, Y.; Wang, T.; Samworth, R.J. High-dimensional, multiscale online changepoint detection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2022**, *84*, 234–266. [CrossRef]
24. Candes, E.J. Modern statistical estimation via oracle inequalities. *Acta Numer.* **2006**, *15*, 257–325. [CrossRef]
25. Fourdrinier, D.; Strawderman, W.E.; Wells, M.T. *Shrinkage Estimation*; Springer: Berlin/Heidelberg, Germany, 2018.
26. Lorden, G.; Pollak, M. Nonanticipating estimation applied to sequential analysis and changepoint detection. *Ann. Stat.* **2005**, *33*, 1422–1454. [CrossRef]
27. Lorden, G.; Pollak, M. Sequential Change-Point Detection Procedures That are Nearly Optimal and Computationally Simple. *Seq. Anal.* **2008**, *27*, 476–512. [CrossRef]
28. Cao, Y.; Xie, L.; Xie, Y.; Xu, H. Sequential Change-Point Detection via Online Convex Optimization. *Entropy* **2018**, *20*. [CrossRef]

29. Xie, L.; Moustakides, G.V.; Xie, Y. Window-Limited CUSUM for Sequential Change Detection. *IEEE Trans. Inf. Theory* **2023**, *69*, 5990–6005. [CrossRef]
30. Wang, Y.; Mei, Y. Large-Scale Multi-Stream Quickest Change Detection via Shrinkage Post-Change Estimation. *IEEE Trans. Inf. Theory* **2015**, *61*, 6926–6938. [CrossRef]
31. Halme, T.; Veeravalli, V.V.; Koivunen, V. Quickest Change Detection for Multiple Data Streams Using the James–Stein Estimator. *IEEE Trans. Inf. Theory* **2025**, *71*, 7802–7814. [CrossRef]
32. James, W.; Stein, C. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Oakland, CA, USA, 1961; pp. 361–379.
33. Chen, Y.; Wang, T.; Samworth, R.J. *ocd: High-Dimensional Multiscale Online Changepoint Detection*, R package version 1.1, 2020 .
34. Dragalin, V. A simple and effective scanning rule for a multi-channel system. *Metrika* **1996**, *43*, 165–182. [CrossRef]
35. Lai, L.; Poor, H.V.; Xin, Y.; Georgiadis, G. Quickest Search Over Multiple Sequences. *IEEE Trans. Inf. Theory* **2011**, *57*, 5375–5386. [CrossRef]
36. Xu, Q.; Mei, Y.; Moustakides, G.V. Optimum Multi-Stream Sequential Change-Point Detection With Sampling Control. *IEEE Trans. Inf. Theory* **2021**, *67*, 7627–7636. [CrossRef]
37. Liu, K.; Mei, Y.; Shi, J. An adaptive sampling strategy for online high-dimensional process monitoring. *Technometrics* **2015**, *57*, 305–319. [CrossRef]
38. Xu, Q.; Mei, Y.; Shi, J. Impact of compensation coefficients on active sequential change point detection. *Seq. Anal.* **2025**, *44*, 153–177. [CrossRef] [PubMed]
39. Chaudhuri, A.; Fellouris, G.; Tajer, A. Round Robin Active Sequential Change Detection for Dependent Multi-Channel Data. *IEEE Trans. Inf. Theory* **2024**, *70*, 9327–9351. [CrossRef]
40. Lattimore, T.; Szepesvári, C. *Bandit Algorithms*; Cambridge University Press: Cambridge, UK, 2020.
41. Oksanen, J.; Koivunen, V. An order optimal policy for exploiting idle spectrum in cognitive radio networks. *IEEE Trans. Signal Process.* **2015**, *63*, 1214–1227. [CrossRef]
42. Lundén, J.; Koivunen, V.; Poor, H.V. Spectrum Exploration and Exploitation for Cognitive Radio: Recent Advances. *IEEE Signal Process. Mag.* **2015**, *32*, 123–140. [CrossRef]
43. Nitzan, E.; Halme, T.; Koivunen, V. Bayesian Methods for Multiple Change-Point Detection with Reduced Communication. *IEEE Trans. Signal Process.* **2020**, *68*, 4871–4886. [CrossRef]
44. Zhang, W.; Mei, Y. Bandit Change-Point Detection for Real-Time Monitoring High-Dimensional Data Under Sampling Control. *Technometrics* **2023**, *65*, 33–43. [CrossRef]
45. Veeravalli, V.V.; Fellouris, G.; Moustakides, G.V. Quickest Change Detection with Controlled Sensing. *IEEE J. Sel. Areas Inf. Theory* **2024**, *5*, 1–11. [CrossRef]
46. Chernoff, H. Sequential Design of Experiments. *Ann. Math. Stat.* **1959**, *30*, 755–770. [CrossRef]
47. Tajer, A.; Heydari, J.; Poor, H.V. Active Sampling for the Quickest Detection of Markov Networks. *IEEE Trans. Inf. Theory* **2022**, *68*, 2479–2508. [CrossRef]
48. Lai, L.; Fan, Y.; Poor, H.V. Quickest Detection in Cognitive Radio: A Sequential Change Detection Framework. In *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM)*; IEEE: Piscataway, NJ, USA, 2008; pp. 1–5.
49. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
50. Bartroff, J.; Song, J. Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates. *Seq. Anal.* **2020**, *39*, 65–91. [CrossRef]
51. Chen, J.; Zhang, W.; Poor, H.V. A False Discovery Rate Oriented Approach to Parallel Sequential Change Detection Problems. *IEEE Trans. Signal Process.* **2020**, *68*, 1823–1836. [CrossRef]
52. Halme, T.; Koivunen, V. Optimal Multi-Stream Quickest Detection with False Discovery Rate Control. In *Proceedings of the 2023 57th Asilomar Conference on Signals, Systems, and Computers*; IEEE: Piscataway, NJ, USA, 2023; pp. 877–881. [CrossRef]
53. Chen, J.; Zhang, W.; Poor, H.V. Non-Bayesian multiple change-point detection controlling false discovery rate. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*; IEEE: Piscataway, NJ, USA, 2016; pp. 31–35.
54. Nitzan, E.; Halme, T.; Poor, H.V.; Koivunen, V. Deterministic Multiple Change-Point Detection with Limited Communication. In *Proceedings of the 54th Annual Conference on Information Sciences and Systems (CISS)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
55. Dandapanthula, S.; Ramdas, A. Multiple testing in multi-stream sequential change detection. *arXiv* **2025**, arXiv:2501.04130. [CrossRef]
56. Chen, Y.; Li, X. Compound sequential change-point detection in parallel data streams. *Stat. Sin.* **2023**, *33*, 453–474. [CrossRef]
57. Lu, Z.; Chen, Y.; Li, X. Optimal Parallel Sequential Change Detection Under Generalized Performance Measures. *IEEE Trans. Signal Process.* **2022**, *70*, 5967–5981. [CrossRef]
58. Li, S.; Xie, Y.; Dai, H.; Song, L. Scan B-statistic for kernel change-point detection. *Seq. Anal.* **2019**, *38*, 503–544. [CrossRef]
59. Wei, S.; Xie, Y. Online kernel CUSUM for change-point detection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2026**, qkag020. [CrossRef]

60. Xie, L.; Zou, S.; Xie, Y.; Veeravalli, V.V. Sequential (quickest) change detection: Classical results and new directions. *IEEE J. Sel. Areas Inf. Theory* **2021**, *2*, 494–514. [CrossRef]
61. Shin, J.; Ramdas, A.; Rinaldo, A. E-detectors: A Nonparametric Framework for Sequential Change Detection. *N. Engl. J. Stat. Data Sci.* **2023**, *2*, 229–260. [CrossRef]
62. Unnikrishnan, J.; Veeravalli, V.V.; Meyn, S.P. Minimax robust quickest change detection. *IEEE Trans. Inf. Theory* **2011**, *57*, 1604–1614. [CrossRef]
63. Xie, L.; Liang, Y.; Veeravalli, V.V. Distributionally Robust Quickest Change Detection using Wasserstein Uncertainty Sets. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*; PMLR: Cambridge, MA, USA, 2024; pp. 1063–1071.
64. Hou, Y.; Bidkhori, H.; Banerjee, T. Robust Quickest Change Detection With Sampling Control. *IEEE Trans. Signal Process.* **2025**, *73*, 3588–3597. [CrossRef]
65. Moustakides, G.V.; Basioti, K. Training Neural Networks for Likelihood/Density Ratio Estimation. *arXiv* **2019**, arXiv:1911.00405. [CrossRef]
66. Gong, T.; Lee, J.; Cheng, X.; Xie, Y. Neural network-based CUSUM for online change-point detection. *arXiv* **2022**, arXiv:2210.17312.
67. Hushchyn, M.; Arzymatov, K.; Derkach, D. Online Neural Networks for Change-Point Detection. *Mach. Learn.* **2026**, *115*, 56. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.