View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Helsingin yliopiston digitaalinen arkisto

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Master's Thesis

Statistics

Quickest Detection under False Discovery Rate and Communication Constraints

Topi Halme

September 2021

Supervisors:
Prof. Visa Koivunen, Dr. Petteri Piiroinen, Dr. Eyal Nitzan

University of Helsinki

Master's Programme in Mathematics and Statistics

Faculty of Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Topi Halme

Työn nimi — Arbetets titel — Title

Quickest Detection under False Discovery Rate and Communication Constraints

Oppiaine — Läroämne — Subject		
Statistics		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	September 2021	75 pages

Tiivistelmä — Referat — Abstract

In a quickest detection problem, the objective is to detect abrupt changes in a stochastic sequence as quickly as possible, while limiting rate of false alarms. The development of algorithms that after each observation decide to either stop and declare a change as having happened, or to continue the monitoring process has been an active line of research in mathematical statistics. The algorithms seek to optimally balance the inherent trade-off between the average detection delay in declaring a change and the likelihood of declaring a change prematurely. Change-point detection methods have applications in numerous domains, including monitoring the environment or the radio spectrum, target detection, financial markets, and others.

Classical quickest detection theory focuses settings where only a single data stream is observed. In modern day applications facilitated by development of sensing technology, one may be tasked with monitoring multiple streams of data for changes simultaneously. Wireless sensor networks or mobile phones are examples of technology where devices can sense their local environment and transmit data in a sequential manner to some common fusion center (FC) or cloud for inference. When performing quickest detection tasks on multiple data streams in parallel, classical tools of quickest detection theory focusing on false alarm probability control may become insufficient. Instead, controlling the false discovery rate (FDR) has recently been proposed as a more useful and scalable error criterion. The FDR is the expected proportion of false discoveries (false alarms) among all discoveries.

In this thesis, novel methods and theory related to quickest detection in multiple parallel data streams are presented. The methods aim to minimize detection delay while controlling the FDR. In addition, scenarios where not all of the devices communicating with the FC can remain operational and transmitting to the FC at all times are considered. The FC must choose which subset of data streams it wants to receive observations from at a given time instant. Intelligently choosing which devices to turn on and off may extend the devices' battery life, which can be important in real-life applications, while affecting the detection performance only slightly. The performance of the proposed methods is demonstrated in numerical simulations to be superior to existing approaches.

Additionally, the topic of multiple hypothesis testing in spatial domains is briefly addressed. In a multiple hypothesis testing problem, one tests multiple null hypotheses at once while trying to control a suitable error criterion, such as the FDR. In a spatial multiple hypothesis problem each tested hypothesis corresponds to e.g. a geographical location, and the non-null hypotheses may appear in spatially localized clusters. It is demonstrated that implementing a Bayesian approach that accounts for the spatial dependency between the hypotheses can greatly improve testing accuracy.

Avainsanat — Nyckelord — Keywords
Quickest detection, Multiple hypothesis testing, False discovery rate, Sequential inference

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

List of Symbols

Abbrevations

ADD	Average detection delay
ANO	Average number of observations
ARL	Average run length
CADD	Conditional average detection delay
FDR	False discovery rate
FWER	Familywise error rate
PFA	Probability of false alarm
WADD	Worst-case average detection delay

Mathematical shorthands

$[K]$	$\{1, 2, ..., K\}$
$\lceil a \rceil$	Smallest integer greater than or equal to a
$\mathbf{1}_{\{A\}}$	Indicator function of event A
$\{X_n\}$	A sequence $X_1, X_2, ...$
$a \vee b$	Maximum of a and b
$a \wedge b$	Minimum of a and b
a^+	Positive part of a, i.e. $a \vee 0$
$h(\alpha) = o_\alpha(1)$	$\lim_{\alpha \to 0} h(\alpha) = 0$

Chapter 1

Introduction

Sequential inference and the detection of abrupt changes in time series data is a widely investigated topic in statistical literature. Its origins lie in the field of quality control [46], where a sudden change in some statistical property in a produced product might be an indication of e.g. a faulty machine that requires investigation. The time of this change is commonly called a change-point. Since its introduction, change-point detection problems have garnered interest in numerous fields, including intrusion detection in computer networks [73], radar [37], wireless communications [33, 45], financial markets [61], detection of signals in seismology [41] and many others.

The field of change-point detection can broadly be divided into two distinct categories depending on whether the data is processed online or offline. In offline change-point detection, the observer has the whole data sequence at hand, and the task is to estimate the time of change in an optimal manner and identify the pre- and post-change statistical models. In online change-point detection problems the statistician receives observations sequentially, and the objective is to identify the change in the stochastic process in real time with minimal delay, while avoiding false alarms, i.e. announcing a change prematurely. For example, in environmental monitoring it is of interest to detect possible hazards rapidly in order to take action and minimize damages. Real time decisions are also required in the monitoring of daily disease counts [65] or radio spectrum [7, 30], and navigation systems [20], to name a few. Online change-point detection is often also referred to as *sequential* or *quickest* change detection in the literature, and it is the focus of this thesis. Two comprehensive general references for theory and methods of quickest change detection, along with more extensive listings of application domains are [50] and [71].

In a classical quickest change detection task, the focus is on a single stream of data. On the other hand, in modern day challenges facilitated by quick development of sensing technology and a growing number of data sources, it is possible to track multiple data

streams for change-points in parallel. Wireless sensor networks (WSN) are a prime example of such technology, with cheap, energy-efficient sensors that can be deployed for various monitoring purposes. Smart phones of a large number of users equipped with a variety of sensors could be another source of multiple data streams. The sensors observe their local environment, and send data to a common Fusion Center (FC) or cloud for further analysis. Monitoring critical infrastructure such as bridges, power grid and oil networks, as well as remote environments in industrial or security applications are common tasks for which these networks are employed. In these scenarios the change-point would represent some abrupt change, anomaly or adversarial event in the underlying stochastic process of the monitored area, resulting from e.g. a leak in an oil pipe or a crack in the bridge structure. Internet of Things (IoT) systems are also central recent applications of such networks. An example of a quickest detection problem in IoT systems is radio spectrum occupancy monitoring, in which unlicensed secondary users wish to use the radio frequency spectrum for transmission in an opportunistic manner when it is idle. The secondary users have to detect the appearance of licensed, primary users and vacate the corresponding frequency bands as quickly as possible [33, 29]. If executed succesfully, effective utilization of available frequency bands would help in satisfying the demand for high-quality and high-data-rate wireless products in the future [7]. Depending on the scenario the change can be observed by all of the sensors simultaneously or only by a subset, if the real life phenomenon affects only a portion of the monitored area. For example, the state of the radio spectrum, air quality, and the weather are dynamic phenomena that vary locally. Alternatively, if the sensors are observing separate processes, their data streams acquired by the FC could be considered practically independent.

Another recently active sub-field of statistics that has been spurred by the increase of data sources is that of large-scale inference. In particular, large scale hypothesis testing has been an active area of study [16]. Classical frequentist testing theory of Fisher, Neyman and Pearson is in many ways constructed around the idea of Type I error rate control. When one is testing hundreds, if not thousands, of hypotheses simultaneously, the classical inference methods quickly become insufficient. The problem of multiple comparisons has been studied since the work by Tukey [74] in the 50's, but it was the introduction of the concept of false discovery rate (FDR) by Benjamini and Hochberg in their seminal 1995 paper [6] that really sparked interest in multiple testing research. The application that has arguably most motivated the research is analysis of DNA microarrays for identification of differentially expressed genes. At the same time, applications that require the use of wireless sensor networks often come in contact with large-scale hypothesis testing. For example, if the network is used for monitoring the presence of some physical world phenomenon, each sensor can be thought of as conducting a hypothesis test for the presence of the phenomenon at its location.

In this thesis, some novel methods and theory related change-point detection and

multiple hypothesis testing in sensor networks are presented. The author of this thesis has conducted research and authored and published peer reviewed papers [27, 28, 42, 43] as a member of a research group at the Department of Signal Processing, Aalto University. In these publications, methods for change-point detection in multiple data streams in parallel that aim to minimize detection delay while controlling the FDR have been proposed. Additionally, scenarios where not all of the devices communicating with the FC can be operational and communicating with the FC at all times have been considered. Instead, the FC must choose which subset of data streams it wants to receive observations from at a given time instant. This is important in practice, since in many applications the sensors may be battery operated with limited computational and communication capabilities and need to remain operational over long time periods. Intelligently controlling the on-off status of the observing device can greatly extend its life span while having only a minor impact on the performance of the inference task at hand. In addition to change-point detection, the problem of multiple hypothesis testing in a spatial domain is briefly addressed. The sensors acquiring the data are in distinct locations. Hence the non-null hypotheses (here representing sensors that are located inside a region of "signal") appear in spatially localized clusters. This work is still more in progress, but preliminary results are presented.

1.1 Contributions

The main contributions of the work presented in this thesis are summarized here. In conference paper [28], and its subsequent extension to a journal publication [42], novel methods for change-point detection in multiple data streams in parallel are presented. The methods handle communication limitations by at each time step monitoring the subset of sensors with the highest posterior probabilities of change-points having occurred. It is shown that the procedures control the FDR under a specified tolerated level, and that they are scalable in the sense that the detection delay and average number of data points used do not increase asymptotically with the number of sensors. Numerical simulations are conducted for validating the derived results, and for demonstrating that the suggested policy for choosing which sensors to monitor at a given time results in better detection performance than that of simpler alternatives. The author of this thesis contributed to the development of the proposed methods in approximately equal extent with the first author of [42]. Additionally, the author is responsible for deriving the proof of FDR control for the methods. The result also applies to previous work by other authors in the topic area, showing that the method presented [13] achieves non-asymptotic FDR control. This is a stronger result than the claim of asymptotic control that was shown in the original paper [13].

In conference publication [27], an approach for multiple hypothesis testing under spatial dependency is presented. The method is a product two popular paradigms of their own fields, the empirical Bayesian two-groups model of multiple testing, and latent Gaussian random fields of spatial statistics. The resulting approach is a flexible Bayesian framework in which the user can encode their own prior knowledge about the properties of the dependency in to the detection model. The Bayesian approach provides a simple way for the user to summarize their posterior beliefs in a way that is consistent with FDR control in light of the inferences. While this topic still requires further research, in numerical simulations it is illustrated that the taking the spatial dependency into account can significantly improve detection power. The author is primarily responsible for all of the contents of the publication, including designing the approach, conducting the simulations and the writing.

1.2 Structure of the thesis

To introduce the reader to the general topic area, we start from the basic definitions and foundations of change-point detection and large-scale testing, and work our way gradually toward the content in the mentioned publications. The rest of this thesis is structured as follows.

In Chapter 2, fundamental concepts and methods of change-point detection under both Bayesian and non-Bayesian frameworks are introduced. Additionally, a brief review of the important asymptotic results and the literature regarding data-efficient change-point detection is provided.

In Chapter 3, another building block of the subsequent content, multiple hypothesis testing, is introduced. After presenting the basic ideas, the connection of the purely frequentist and Bayesian interpretations of the concept of false discovery rate is discussed. Section 3.3 is devoted to the conference paper [27], where an approach for multiple hypothesis testing in spatial domains is presented. Chapter 3 is concluded by introducing the connection of change-point detection in sensor networks and multiple hypothesis testing.

Chapter 4 contains the contents of [28] and [42], where new methods for data-efficient change-point detection of multiple parallel data streams are developed and analyzed.

Chapter 5 concludes the thesis.

Chapter 2

Background on change-point detection

2.1 Fundamentals of change-point detection

The methods for change-point detection in multiple parallel data streams studied later in this thesis are founded on the theory of change-point detection in a single data stream. Hence, in this section, the conventional problem of sequential change-detection for a single stream of data is formulated. We observe a sequence of independent random variables $X_1, X_2, ... := \{X_n\}$ that follow a distribution f_0, until a change occurs at an unknown time $t \in \{0, 1, 2, ...\}$. After the change, the observations $X_t, X_{t+1}, ...$ are still independent, but follow a different distribution f_1. That is, conditional on t, $\{X_n\}$ is an independent sequence where $X_1, ..., X_{t-1} =: \mathbf{X}^{(t-1)}$ are i.i.d. with distribution f_0, and $X_t, X_{t+1}, ...$ i.i.d. with distribution f_1. An illustration of such data appears in Figure 2.1. The goal is to detect the change in distribution as soon as possible by observing the sequence, while avoiding false alarms.

First we introduce some important definitions. We consider the sequence $\{X_n\}$ to be defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ a σ-algebra and $\mathbb{P}$ a probability measure.

Definition 1 (Filtration)**.** *A filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$ on $(\Omega, \mathcal{F})$ is an increasing sequence of sub-σ-algebras of $\mathcal{F}$. That is*

$$\mathcal{F}_n \subseteq \mathcal{F}_m, \quad \textit{when } n < m.$$

A particularly useful object is the filtration $\{\mathcal{F}_n^X\}$ created by $\{X_n\}$, that is,

$$\mathcal{F}_n^X = \sigma(X_1, ..., X_n),$$

where $\sigma(X_1, ..., X_n)$ denotes the smallest σ-algebra with respect to which all of the random variables $(X_1, ..., X_n)$ are measurable. This filtration is sometimes referred to as the

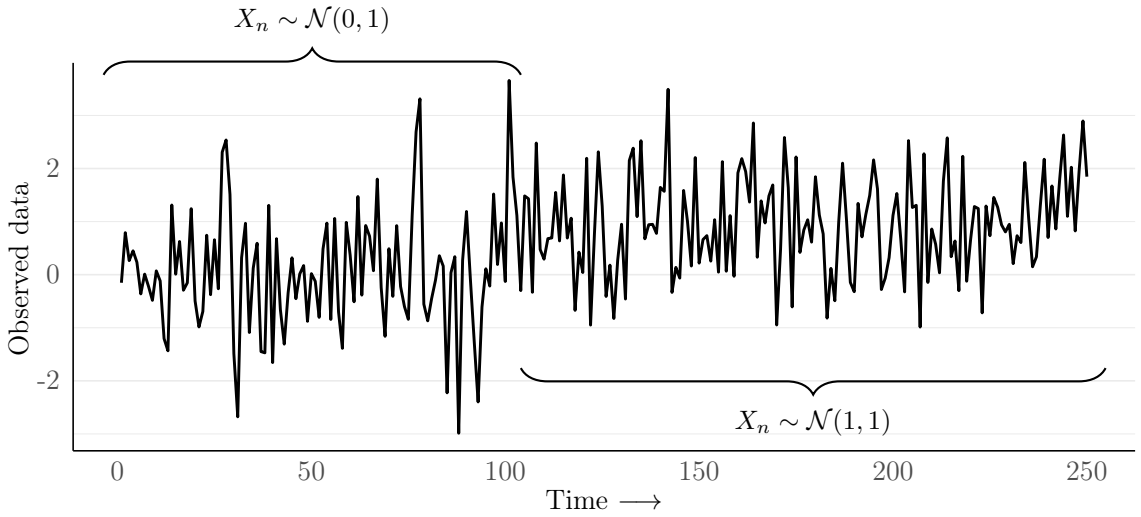

Figure 2.1: An example of sequence $\{X_n\}$ where the observations prior to the unobserved change-point at time 100 follow $\mathcal{N}(0,1)$, and $\mathcal{N}(1,1)$ afterwards.

natural filtration of $\mathcal{F}$ with respect to $\{X_n\}$. An informal interpretation for the nth element of the natural filtration $\mathcal{F}_n^X$ is that it contains any information that could be asked and answered for the considered random process at time n. For the rest of this section $\{\mathcal{F}_n\}$ is used when referring to $\{\mathcal{F}_n^X\}$.

The task of observing a sequence $\{X_n\}$ and "stopping" to declare that a change has occurred is achieved by means of a *stopping time* on the natural filtration $\{\mathcal{F}_n\}$.

Definition 2 (Stopping time)**.** *A random variable T defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\{0, 1, 2, ...\} \cup \infty$ is a stopping time with respect to a filtration $\{\mathcal{F}_n\}$ if*

$$\{T = n\} \in \mathcal{F}_n, \quad \textit{for all } n \geq 0.$$

Intuitively, the condition in the definition tells that the decision of whether to "stop" (and declare a change) at time n must be based only on information available at time n. An important class of stopping times are ones generated by threshold based rules

$$T' = \inf\{n : Y_n > s\},$$

where $\{Y_n\}$ is any sequence such that Y_k is $\mathcal{F}_k$-measurable[1] for all k and $s \in \mathbb{R}$ is a constant. That is, Y_n is any deterministic function of $\mathbf{X}^{(n)}$. To see that T' is a stopping time, observe that

$$\{T' = n\} = \{Y_1 \leq s, Y_2 \leq s, ..., Y_n > s\} \in \mathcal{F}_n.$$

[1]In probability jargon, the sequence $\{Y_n\}$ is adapted to $\{\mathcal{F}_n\}$

A central quantity in constructing effective stopping times will be the (log-)likelihood ratio of f_1 and f_0 for an observation X,

$$\ell(X) := \log L(X) := \log \frac{f_1(X)}{f_0(X)}.$$

The useful property of the likelihood ratio for sequential change-detection will become apparent after introducing the Kullback-Leibler (KL) divergence for distributions f_1 and f_0.

Definition 3 (KL Divergence)**.** *The KL divergence between two densities f_1 and f_0 is defined as*

$$D(f_1||f_0) := \int \log \frac{f_1(x)}{f_0(x)} f_1(x)dx = \mathbb{E}_{f_1}\ell(X).$$

A well known property of KL divergence is that $D(f_1||f_0) \geq 0$ with equality only if $f_1 = f_0$ almost surely. Since the detection problem becomes degenerate if $f_1 = f_0$ almost surely, we assume f_0 and f_1 to be such that $D(f_1||f_0) > 0$. Hence, for $n \geq t$ we have $\mathbb{E}[\ell(X_n)] = \mathbb{E}_{f_1}\ell(X) = D(f_1||f_0) > 0$. On the other hand, for $n < t$, $\mathbb{E}\ell(X_n) = -\mathbb{E}_{f_0}[\log(f_0(X)/f_1(X))] = -D(f_0||f_1) < 0$. That is, the sign of the expected value of the log-likelihood ratio is different for pre- and post-change observations.

The objective of the observer is to design a stopping rule such that $\{T = n\}$ claims that the change-point t has occurred within the first n observations, when T is the stopping time defined by the stopping rule. If, however, $T < n$, it is said that a *false alarm* has been raised. A good detection procedure should have small delay $T-t$ between the change-point and the time of its detection, and produce false alarms rarely. While the precise criteria for the detection speed and the rate of false alarms depend on additional assumptions, there is in general a trade-off between the two performance criteria. A procedure that is greedy in declaring the occurrence of a change can be prone to false alarms, and a very conservative detection strategy can be slow in detecting the change-point.

One can take either a frequentist or a Bayesian approach to change-point detection and consider t either as a deterministic but unknown quantity or a random variable, respectively. Both formulations are investigated in this thesis, beginning with the Bayesian viewpoint.

2.1.1 Bayesian quickest change detection

The case of Bayesian quickest change detection was first formulated by Shiryaev in [60]. In the Bayesian formulation, a prior distribution $\{p_n\}_{n=0}^{\infty}$ is specified for t, so that

$$p_n := \mathbb{P}(t = n), \quad n = 0, 1, 2, ...$$

represents the a priori knowledge regarding the change-point. Since T and t are both random variables in the Bayesian approach, the detection delay of a stopping time T can be measured with the average detection delay (ADD), defined as

$$\mathrm{ADD}(T) = \mathbb{E}\left[(T-t)^+\right],$$

where $x^+ = \max(x, 0)$. To measure the rate of false alarms, it is reasonable to use the probability of false alarm (PFA)

$$\mathrm{PFA}(T) = \mathbb{P}(T < t).$$

Given these definitions, the Bayesian quickest detection problem was formulated by Shiryaev as follows.

Definition 4 (Shiryaev's problem)**.** *For a given $\alpha > 0$, find a stopping time T that minimizes* $\mathrm{ADD}(T)$ *subject to PFA(T) $\leq \alpha$.*

The problem can be solved by considering the Lagrangian relaxation,

$$\inf_{T\in\mathcal{T}} \left\{\mathrm{PFA}(T) + c\cdot \mathrm{ADD}(T)\right\}, \tag{2.1}$$

where $\mathcal{T}$ is the set of all stopping times on $\{\mathcal{F}_n\}$. In (2.1), the constant c is the Lagrangian multiplier.

To proceed, we assume for convenience that the distribution of t is geometric with parameter ρ, i.e.

$$p_n = \rho(1-\rho)^{n-1}, \quad n = 1, 2, ...$$

The geometric distribution is the conventional assumption in Bayesian sequential change-point detection due to its memorylessness property[2] [1, 70]. That is, conditional on the change not having occurred within n, the probability of it occurring at $n+1$ is constant for all n.

An important quantity in the detection process will be the posterior probability of the change having happened at or prior to n, conditional on the observations up to n,

$$\pi_n := \mathbb{P}(t \leq n|\mathcal{F}_n). \tag{2.2}$$

Under the assumed Bayesian model, a convenient recursive update for π_n follows from Bayes rule as follows [60]:

$$\pi_n = \frac{d\mathbb{P}(\mathbf{X}^{(n)}|t\leq n)\mathbb{P}(t\leq n)}{d\mathbb{P}(\mathbf{X}^{(n-1)})} = \frac{d\mathbb{P}(X_n|t\leq n, \mathbf{X}^{(n-1)})d\mathbb{P}(\mathbf{X}^{(n-1)}|t\leq n)\mathbb{P}(t\leq n)}{d\mathbb{P}(\mathbf{X}^{(n)})}$$
$$= \frac{f_1(X_n)\left[(1-\pi_{n-1})\mathbb{P}(t=n|t\geq n)+\pi_{n-1}\right]}{f_1(X_n)\left[(1-\pi_{n-1})\mathbb{P}(t=n|t\geq n)+\pi_{n-1}\right] + f_0(X_n)(1-\pi_{n-1})(1-\mathbb{P}(t=n|t\geq n))}.$$

[2] A discrete random variable t is said to be memoryless if $\mathbb{P}(t > n+k|t\geq n) = P(t>k)$ for all k and n

Noting that as t is geometric, $\mathbb{P}(t = n|t \geq n) = \rho$ for all n, this expression simplifies to

$$\pi_n = \frac{L(X_n)\phi_n}{L(X_n)\phi_n + 1 - \phi_n}, \tag{2.3}$$

where $\phi_n := \pi_{n-1} + (1 - \pi_{n-1})\rho$ contains the components of the update that are independent of the observation X_n, and $\pi_0 = 0$.

The importance of the quantity π_n is apparent from the following result, originally derived in [60, Thm. 1]. The result establishes that the optimal stopping time for Shiryaev's problem is given by a threshold rule on the sequence $\{\pi_n\}$.

Theorem 1. *Assume t is geometrically distributed. Then, for an appropriately chosen threshold $A^* \in [0, 1]$, the stopping time*

$$T_{A^*} = \inf\{n \geq 0 | \pi_n \geq A^*\} \tag{2.4}$$

is Bayes optimal. That is, it solves (2.1).

Proof. See [60, Thm. 1] for the original proof. Alternatively, a proof via a convenient dynamic programming argument is given in [75]. □

The stopping time in (2.4), often referred to as the Shiryaev stopping time, is easy to implement for practical purposes, as the decision statistic is one-dimensional and admits a recursive form. However, observe that Theorem 1 does not provide the optimal threshold A^* for a given false alarm probability constraint α, it only tells that the optimal stopping rule is of this form for *some* threshold A^*. For practical use, a useful threshold is obtained from the following proposition [60]

Proposition 1. *Let $T_{1-\alpha}$ be a Shiryaev time from* (2.4) *with stopping threshold $1 - \alpha$. Then,*

$$\mathrm{PFA}(T_{1-\alpha}) \leq \alpha.$$

Proof. The proof follows from the law of total expectation and the definition of $T_{1-\alpha}$

$$\mathbb{P}(T_{1-\alpha} < t) = \mathbb{E}[\mathbb{P}(T_{1-\alpha} < t|\mathcal{F}_{T_{1-\alpha}})] = \mathbb{E}[1 - \pi_{T_{1-\alpha}}] = 1 - \pi_{T_{1-\alpha}} \leq \alpha.$$

□

Note that Proposition 1 is valid for any prior distribution, not just geometric. Whether $1 - \alpha$ is a good approximation for the optimal threshold depends on the amount the statistic π_n "overshoots" the threshold $1 - \alpha$ at the detection time $n = T_{1-\alpha}$. For i.i.d. observations and a geometric prior, accurate estimates of the overshoot are provided in [70]. Nonetheless for small values of ρ and $D(f_1||f_0)$ the overshoot can be neglected,

since the sequence $\{\pi_n\}$ evolves with small increments and large overshoots are unlikely to take place. Additionally, in Section 2.1.3 it is justified why this approximation error is negligible in the asymptotic limit PFA $\to 0$.

The Shiryaev algorithm is illustrated in Figure 2.2. The lines correspond to two different realizations of the the detection procedure T, generated by two different observation sequences $\{X_n\}$. Though t is random, in the illustration the same $t = 100$ applies for both realizations for brevity. The gray posterior probability trajectory results in a false alarm, as the detection threshold $1 - A$ is exceeded prior to t. The dark trajectory produces a correct detection, with detection delay $T - t$.

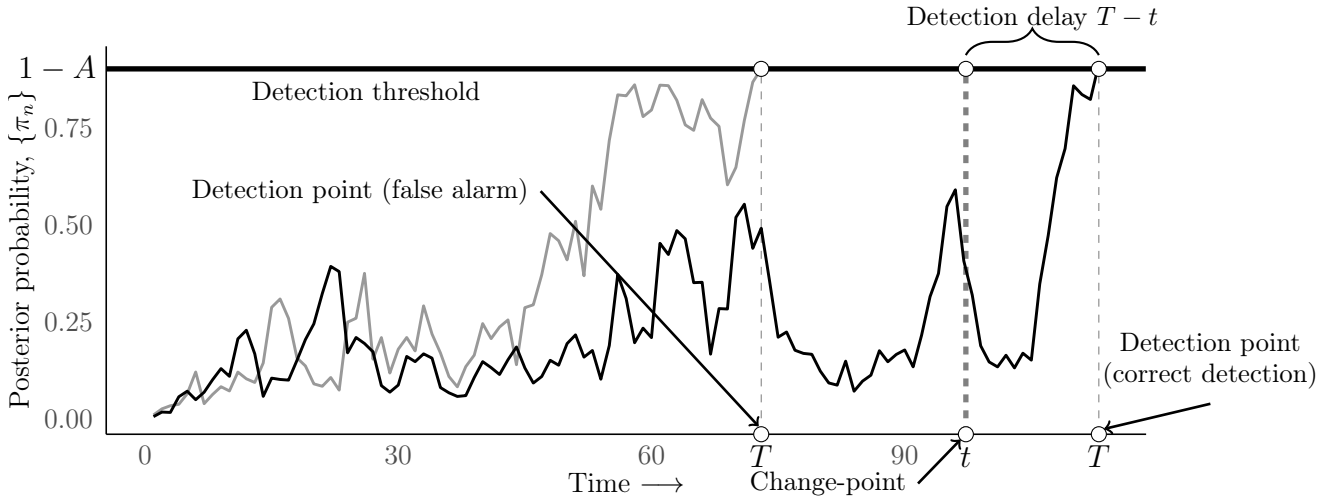

Figure 2.2: The Shiryaev algorithm declares a change-point the first time posterior probability surpasses the detection threshold. There are two possible outcomes: a false alarm (gray) or a correct detection (black).

2.1.2 Non-Bayesian sequential change-point detection

In the previous section, a convenient optimal procedure was found under the Bayesian formulation. The key assumption was that the change-point follows a geometric distribution. In this section no such assumption is made, and the change-point is treated as a deterministic quantity. In applications such as surveillance or inspection systems it is sometimes hard to define a prior distribution that would be logical and justified or to formulate the prior knowledge in a form of a probability distribution, making the frequentist approach necessary. While it requires slightly more complicated formulations and doesn't

admit any uniformly optimal procedures like the Shiryaev algorithm, several useful frequentist change-point detection formulations and solutions have been found. The most notable ones are due to Lorden [34] and Pollak [48]. A brief overview of these methods is provided in this subsection.

When the change-point t is considered to be a deterministic but unknown quantity, the probability $\mathbb{P}(T < t)$ can only be defined in the form of conditional probabilities $\mathbb{P}(T < t|t = k)$. The shorthand $\mathbb{P}_k(\cdot)$ will be used to denote the conditional measure $\mathbb{P}(\cdot|t = k)$, and $\mathbb{E}_k$ to denote the expectation with respect to $\mathbb{P}_k$. Additionally, $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$ are the probability measure and the corresponding expectation in the case when a change does not occur at all. To control the false alarm probability for all possible times of change, one would need to control $\sup_k \mathbb{P}_k(T < k)$. We will see later that this is a rather severe constraint for practical purposes. Instead, the rate of false alarms is usually measured by the average run length to false alarm (ARL), defined as $\mathrm{ARL}(T) = \mathbb{E}_\infty T$ [50]. In essence, ARL quantifies how long a particular stopping time can on average receive observations from the pre-change distribution without triggering a false alarm.

In the Bayesian formulation it was observed in eq. (2.3) that the data sequence $\{X_n\}$ influences the test statistic only through the likelihood ratios $L(X_n) = f_1(X_n)/f_0(X_n)$. The importance of the likelihood ratio in the frequentist approach is evident from the earlier observation that $\mathbb{E}_\infty \ell(X_1) < 0$ and $\mathbb{E}_1 \ell(X_1) > 0$. That is, prior to the change-point the log likelihood ratio has negative expectation, and positive afterwards. Since the observations are assumed i.i.d. conditional on the change-point, one possibility for detecting the change-point is to monitor the process generated by the sum of log likelihood ratios

$$S_n := \sum_{k=1}^{n} \ell(X_k)$$

for a change from negative to positive drift.

The Cumulative Sum (CuSum) algorithm, first introduced by Page in 1950's [46], is based around this concept. An intuitive idea is to keep track of the distance between the current value of S_n and the current smallest value of S_k for $k \leq n$. If the random walk has evolved far from its lowest point, it is considered as evidence of the process having positive drift in this interval, and thus of a change having taken place. Concretely, we define the statistic

$$W_n := S_n - \min_{1 \leq k \leq n} S_k,$$

and the associated stopping time

Definition 5 (CuSum algorithm [46])**.**

$$\tau_c := \inf\{n \geq 0 | W_n > b\}, \quad b \in \mathbb{R}^+. \tag{2.5}$$

Observe that $S_n - \min_{1\leq k\leq n} S_k = \max_{1\leq k\leq n+1} \sum_{j=k}^{n} \ell(X_j)$. From this a convenient recursion for W_n can be derived in the form of

$$W_n = \max\{0, W_{n-1} + \ell(X_n)\}, \quad W_0 = 0. \tag{2.6}$$

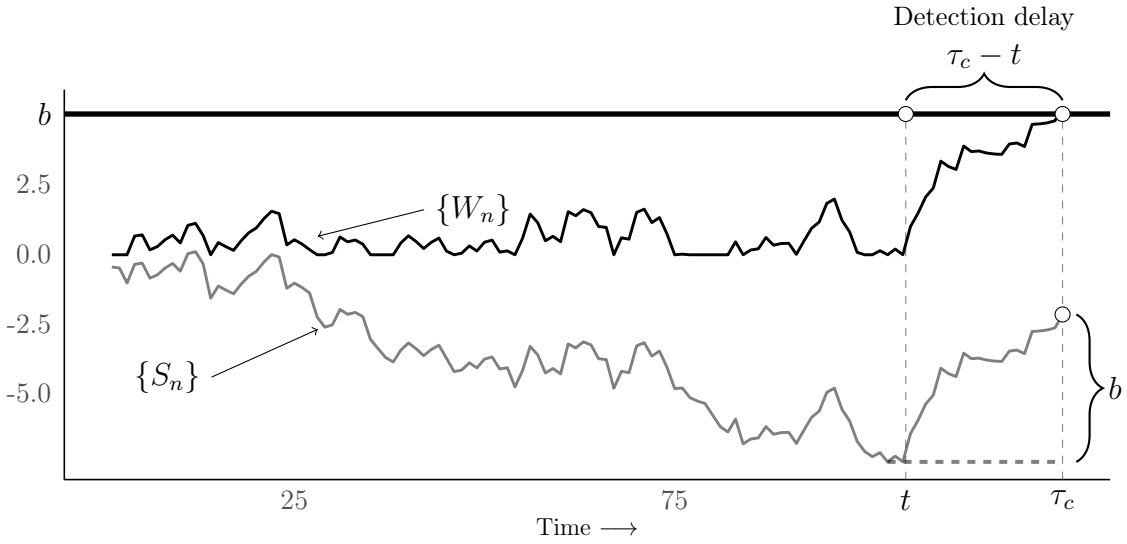

Figure 2.3: The CuSum statistic $\{W_n\}$ and the cumulative likelihood ratio sum $\{S_n\}$ computed for the data in Figure 2.1.

The CuSum algorithm, as well as the relationship between W_n and S_n, is illustrated in Figure 2.3. While the algorithm was initially developed using purely heuristics described above, it was later shown to have useful optimality properties. In [34], Lorden defined a quantity known as the worst case average detection delay (WADD) as[3]

$$\text{WADD}(\tau) := \sup_{k\geq 1} \operatorname{ess\,sup} \mathbb{E}_k \left((\tau - k)^+ | \mathcal{F}_{k-1} \right). \tag{2.7}$$

The "worst case" in WADD refers to the fact that the expectation in (2.7) is taken with respect to the worst possible (in terms of ensuing detection delay) pre-change-point observations $\mathcal{F}_{k-1}$. The problem considered by Lorden is the following

Definition 6 (Lorden's problem)**.** *For a given $\gamma > 0$, find a stopping time τ that minimizes* WADD(τ) *subject to* ARL$(\tau) \geq \gamma$.

[3]ess sup refers to the essential supremum, which is the least upper bound of the set of constants that bound the random variable with probability 1.

Lorden's initial result was that the CuSum algorithm is asymptotically optimal[4] for the problem in the limit $\gamma \to \infty$.

Later it was shown in [39] that τ_c is in fact exactly (not just asymptotically) optimal for Lorden's problem. Since WADD contains suprema with respect to both the change time and the pre-change observations, it is a rather pessimistic criterion for the detection delay. The other prominent criterion in the literature is the conditional average detection delay (CADD) [48]

$$\mathrm{CADD}(\tau) = \sup_{k \geq 1} \mathbb{E}_k(\tau - k | \tau \geq k).$$

To see that CADD is less pessimistic than WADD, observe that if τ is a stopping time on $\{\mathcal{F}_n\}$, then $\{\tau \geq k\} \in \mathcal{F}_{k-1}$. Therefore for all k,

$$\mathrm{WADD}(\tau) \geq \sup_{k \geq 1} \mathbb{E}_k((\tau - k)^+ | \tau \geq k) = \sup_{k \geq 1} \mathbb{E}_k(\tau - k | \tau \geq k) = \mathrm{CADD}(\tau).$$

It is shown in [48], that the optimal procedure for minimizing CADD while maintaining the ARL above a prespecified threshold is given by an algorithm based around a special case of the Shiryaev statistic of the Bayesian subsection. The statistic is known as the Shiryaev-Roberts statistic, given by

$$R_n = \sum_{k=1}^{n} \prod_{j=k}^{n} L(X_j).$$

With some algebraic manipulation, one can see that the recursion $R_n = (1+R_{n-1})L(X_n)$, $R_0 = 0$ applies. Since then, the Shiryaev-Roberts algorithm and its generalization the Shiryaev-Roberts-Pollak algorithm have been to shown to have multiple even stronger optimality properties in terms of CADD [72, 49].

Both of the non-Bayesian algorithms introduced impose a lower bound on the expected time to a false alarm $E_\infty(T)$. If we instead wanted to control the probability of false alarm similar to the Bayesian setup, we would need that $\mathbb{P}_k(T < k) \leq \alpha$ for all k. Since T is a stopping time, $\{T < k\} \in \mathcal{F}_{k-1}$ and hence

$$\mathbb{P}_k(T < k) = \mathbb{P}_{k+1}(T < k) = \ldots = \mathbb{P}_\infty(T < k).$$

Thus controlling $\mathbb{P}_k(T < k)$ for all k is equivalent to controlling $\mathbb{P}_\infty(T < \infty)$. It is easy to see that $\mathbb{P}_\infty(\tau_c < \infty) = 1$, when τ_c is the CuSum stopping time of (2.5) with any finite

[4] In this context, a stopping time τ said to be asymptotically optimal (in terms of a delay criterion D) if

$$\lim_{\gamma \to \infty} \frac{\mathrm{D}(\tau)}{\mathrm{D}(\tau^*)} \leq 1,$$

for all stopping times τ^* s.t. $\mathrm{ARL}(\tau^*) \geq \gamma$.

detection threshold b. The intuitive explanation is that under $\mathbb{P}_\infty$ the sequence $\{X_n\}$ is i.i.d. with density f_0, making the CuSum statistic in (2.6) a random walk that resets to 0 if it exits the interval $(0, b)$ through 0. After every reset, there is a constant non-zero probability that next exit of the random walk from the interval $(0, b)$ is through b, which would trigger the stopping time. Since there can be infinitely many resets, τ_c will almost surely trigger a false alarm under $\mathbb{P}_\infty$. Similar reasoning applies for the Shiryaev-Roberts and the Shiryaev-Roberts-Pollak algorithms.

The problem of controlling $\mathbb{P}_\infty(T < \infty)$ in the non-Bayesian context was studied in detail in [8]. It was shown, that the CuSum and Shiryaev-Roberts algorithms with thresholds that increase in time are asymptotically optimal for minimizing $\mathbb{E}_k(T-k|T \geq k)$ as $k \to \infty$ while controlling the false alarm probability. That is, instead comparing the CuSum statistic to a constant threshold b, the stopping rule needs to be of the form

$$\tau_{c,\infty} = \inf\{n \geq 0 : W_n > b(n)\},$$

where $b(n)$ is a sufficiently rapidly increasing function of n.

2.1.3 Asymptotic properties

So far the change-point detection algorithms introduced work by first constraining the some measure of the rate of false alarm alarms, and then minimizing the detection delay. Thus we obtain an upper bound for the rate of false alarms, but no estimates regarding the magnitude of delay have been given. Obtaining exact estimates is difficult, but analytical approximations for detection delay are available in the asymptotic domain as the rate of false alarms tends to zero. In the Bayesian problem formulation, important results are derived in [69]. The following theorem [69, Th. 4] shows, that for conditionally i.i.d. observations, the asymptotic detection delay of the Shiryaev stopping time depends essentially on two quantities, the KL divergence $D(f_1||f_0)$ and the geometric prior distribution parameter ρ. From hereafter we use notation $A_\alpha \simeq B_\alpha$ to mean

$$\lim_{\alpha \to 0} \frac{A_\alpha}{B_\alpha} = 1.$$

Theorem 2 ([69]). *Let the observations $X_1, X_2, ...$ be i.i.d. conditional on the change-point and the change-point t have a geometric distribution with parameter ρ. If $T_{1-\alpha}$ is the Shiryaev stopping time of* (2.4) *with threshold $1-\alpha$ and $D(f_1||f_0) < \infty$, then*

$$\mathrm{ADD}(T_{1-\alpha}) = \mathbb{E}(T_{1-\alpha} - t)^+ \simeq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}, \quad \text{as } \alpha \to 0. \tag{2.8}$$

In this thesis, a geometric prior is often assumed for simplicity. However, it is insightful to note that a result similar to Theorem 2 exists also for a general prior distribution [69, Th. 3]. Let us define

$$d := -\lim_{n\to\infty} \frac{\log \mathbb{P}(t > n)}{n}.$$

The parameter d is sometimes referred to as the exponential index of a random variable. In our case it describes the tail behavior of distribution of the true change-point. For so-called heavy-tailed distributions $d = 0$, and for distributions with light tails $d > 0$ [21, pp. 8]. If t is geometric, a simple calculation shows $d = |\log(1 - \rho)|$. Now, if the other conditions of Theorem 2 apply, but t obeys a general distribution with exponential index d and additionally $\mathbb{E}_{f_1}[\ell(X)]^2 < \infty$, Theorem 3 of [69] gives

$$\mathbb{E}(T_{A_\alpha} - t)^+ \simeq \frac{|\log \alpha|}{D(f_1||f_0) + d}, \quad \text{as } \alpha \to 0. \tag{2.9}$$

The results (2.8) and (2.9) imply that the detection delay decreases as KL divergence between f_1 and f_0 increases. This is not surprising, as a large KL divergence indicates that it is easier to distinguish between the pre- and post-change distribution. Additionally, from (2.8) we may observe that an inverse relationship exists between ρ and the expected delay. When ρ is large, the prior distribution is heavily concentrated on the small positive integers, providing a lot of information regarding the location of the change-point. This speeds up the detection process. As ρ decreases, the prior distribution becomes flatter and hence less informative. When the distribution decays slowly enough to belong in the heavy-tailed class ($d = 0$), it is seen from (2.9) that it has no influence on the asymptotic delay.

In the deterministic setting, asymptotic approximations are available in the regime ARL $\to \infty$. There is a vast literature regarding these approximations, starting from Lorden showing in [34] that for the CuSum algorithm

$$\mathrm{WADD}(\tau_c) \simeq \frac{\log b}{D(f_1||f_0)}, \quad \text{as } b \to \infty. \tag{2.10}$$

Note the apparent similarity between (2.9) and (2.10): in the asymptotic limit, the difference in delay between the optimal Bayesian and non-Bayesian procedures depends only on the tail behavior d of the prior distribution.

The asymptotic results shown in this section only characterize *rate* of growth of ADD as a function α. In particular, for the Bayesian case, we see from (2.8) that the delay grows linearly in $|\log \alpha|$, with slope determined by the KL divergence and the prior. For i.i.d. data, more accurate second and third-order asymptotic approximations have been

derived in e.g. [70], [72]. In this context a second order approximation is such that the true delay is within a constant of the approximation, and for a third order approximation this constant vanishes as $\alpha \to 0$. These results, however, will not be used in this thesis.

2.2 Data-efficient change-point detection

Having introduced the fundamental concepts and algorithms in sequential change-point detection, we now focus on a branch of the quickest detection literature that has been actively investigated in recent years. The problem of data-efficient sequential change detection is addressed. As mentioned in the introduction, surveillance and monitoring are key applications of sequential change-point detection theory. In many such cases, surveillance is only possible using inexpensive, long lifespan, battery powered, remote sensors with limited computation and communication capabilities, and no human intervention. Some examples are the Internet of Things, monitoring of the power grid, and environmental monitoring, e.g. the use of sensor networks for habitat monitoring of certain sea-birds [36]. These sea-birds choose a remote habitat precisely to avoid contact with humans and predators, making some kind of wireless sensor network the only feasible monitoring option.

In these applications, the sensors may need to remain operational for long periods of time. Thus it can be of interest to sometimes switch a sensor to a sleep mode or constrain its communication in order to save battery. This degrades the surveillance quality by increasing detection delay to some extent, since during these times no information is received from the sensor. However, designing a system that intelligently switches between the on and off modes based on the observed data can be valuable in managing the trade-off between the quality of statistical inference and energy efficiency. In this section, two central works in the data-efficient change-point detection literature are briefly reviewed. First, we consider a case with just one sensor introduced in [1], and in Section 2.2.1 a generalization to a network with more than one sensor, discussed in [52].

Mathematically, the problem of energy-efficient change-point detection was formulated in [1] as follows. Recall the classic Bayesian change-point detection formulation of Section 2.1, where we have sequence of random variables $\{X_n\}$ i.i.d. with density f_0 before the change-point t and i.i.d. with density f_1 after t. The change-point t is a geometric random variable with parameter ρ. In order to minimize sensor usage, at each time instant, a decision is made on whether to have the sensor turned on or off in the next time step, based on all available data at the current time instant. If a sensor is turned off for time step n, it means that no observation X_n is received. Formally, [1] introduces a control variable $S_n \in \{0, 1\}$, where $S_n = 1$ if the sensor is turned on at time n, so that X_n is available for decision making, and $S_n = 0$ otherwise. As S_n is determined by the user as

a function of the obtained information, we write

$$S_n = \mu(I_{n-1}),$$

for some control function μ, with $I_n = \sigma\left(\{X_j, S_j\}_{j=1}^n\right)$ being the σ-algebra generated by the pairs (X_j, S_j). It is understood that if $S_j = 0$, then the value of X_j does not exist in I_n.

For a stopping time T, detection delay and accuracy in this setting are still quantified by ADD and PFA respectively, as defined in Section 2.1. Additionally, data-efficiency is measured by the average number of observations (ANO) taken before the change-point, i.e.

$$\mathrm{ANO}(T, \mu) := \mathbb{E}\left[\sum_{n=1}^{t\wedge(T-1)} S_n\right].$$

Note that observations taken after the true change-point has occurred are not counted towards ANO. This is because a sampling policy should sample as frequently possible after the change to minimize delay. It is unnecessary observations taken before t that are to be avoided.

With these definitions, it is of interest to find a stopping time T^* and a sampling policy μ^* that solve the following optimization problem:

$$\begin{aligned} &\underset{T,\mu}{\text{minimize}} \quad \mathrm{ADD}(T, \mu) \\ &\text{subject to} \quad \mathrm{PFA}(T, \mu) \leq \alpha \ \text{ and } \mathrm{ANO}(T, \mu) \leq \beta. \end{aligned} \tag{2.11}$$

In (2.11) α is the false alarm probability constraint, and β is a constraint on the expected number of samples used in the process. Note that if $\beta = \infty$, the optimal sampling policy is trivial, and the classic Shiryaev problem of Definition 4 is recovered as a special case.

The optimal solution to (2.11) is derived in [1] via a dynamic programming argument. The optimal stopping time T^* is a Shiryaev type time $T^* = \inf\{n : \pi_n \geq A\}$, where π_n is again the posterior probability of change having occurred up to n, introduced in (3.12). The optimal sampling policy μ^* has more complex structure, and does not admit a simple representation. However, it is shown that for some $B < A$, a simple sampling policy μ' of the form

$$S_{n+1} = \mu'(I_n) := \mathbf{1}_{\{\pi_n \geq B\}},$$

where $\mathbf{1}_{\{E\}}$ is the indicator function of event E, coincides with the optimal policy for most practical configurations of the system parameters f_0, f_1, ρ, α and β. The thresholds A and B of the policy (T^*, μ') are chosen so that the PFA and ANO constraints are met. In words, the sampling policy μ' turns the sensor on only if the posterior probability of change is above a fixed threshold B. If $\pi_n < B$, the sensor is turned off. Note that

even when the sensor is turned off, the posterior probability π_n can still be updated. In particular, a simple computation shows that

$$\pi_{n+1} = \pi_n + (1 - \pi_n)\rho > \pi_n$$

when X_{n+1} is not received. This can be seen by setting $L(X) = 1$ in (2.3). Thus, when the sensor is turned off the sequence $\{\pi_n\}$ is monotonically increasing until it reaches B, and a new observation is taken. Once the posterior probability surpasses A, a change is declared.

It is shown in [1] that that the policy (T^*, μ') is asymptotically optimal in the limit $\alpha \to 0$. For moderate values of α and Gaussian observations, the ADD of the policy (T^*, μ') is within 10% of the classic Shiryaev procedure, while the number of observations used is reduced by more than 50%. In particular, it is shown that this data dependent sampling scheme is clearly superior to a naive system where the times the sensor is turned on and off are determined beforehand. This notion of choosing to activate a sensor only when π_n is sufficiently large will appear again in Chapter 4.

This problem formulation has been extended to non-Bayesian settings in [2], and to handle composite post-change distributions[5] in [3].

2.2.1 Data-efficient change-point detection in Sensor Networks

So far in this thesis change-point detection theory has been considered only from the perspective of observing only a single data stream $\{X_n\}$. We now turn to a more general scenario, where K data streams $\{X_n^{(1)}\}, \{X_n^{(2)}\}, ..., \{X_n^{(K)}\}$ are observed simultaneously. Each data stream can be thought of as corresponding to one sensor acquiring the data. The network structure is displayed in Figure 2.4. Thus, at each time instance we receive a vector $\mathbf{X}_n := (X_n^{(1)}, X_n^{(2)}, ..., X_n^{(K)})$ of observations. We assume that the data streams are mutually independent with pre- and post-change distributions f_0 and f_1. The change-point t is again a geometric random variable with parameter ρ. In this subsection, we assume that the change affects all sensors simultaneously, that is

$$X_n^{(k)} \sim \begin{cases} f_0, & \text{when } n < t \\ f_1, & \text{when } n \geq t \end{cases} \quad \text{for all } k = 1, 2, ..., K.$$

The ordinary Shiryaev's problem of Section 2.1.1 and its solution are essentially trivial to extend for multivariate independent observations. For independent streams, likelihood ratio between pre-change and post-change distributions takes the product form

$$L(\mathbf{X}_n) = \frac{\prod_{k=1}^K f_1(X_n^{(k)})}{\prod_{k=1}^K f_0(X_n^{(k)})}.$$

[5] A composite post-change distribution means that f_1 is only known up to some parameters.

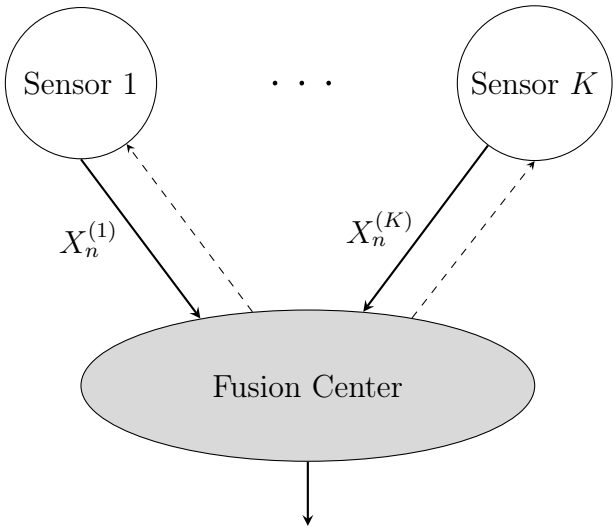

Figure 2.4: An illustration of the sensor network configuration considered in this section. Sensors send their observations to the FC. The FC sends a signal to the sensors it wants to receive an observation from in the next time step.

The rest of the treatment in Section 2.1.1 applies directly for independent multivariate observations by replacing X_n with $\mathbf{X}_n$.

Data-efficient change-point detection, on the other hand, does receive an additional degree of freedom in the multivariate case compared to the univariate. In the univariate case, the only options were to either have the single sensor on or off. When there are multiple sensors, it is possible to choose any number between 0 and K of the sensors to be active in a given time instance. Thus, at each time slot n, a vector of observations $\mathbf{X}_n^{(M_n)} = (X_n^{(1)}, X_n^{(2)}, ..., X_n^{(M_n)})$ is received, where M_n is the number of observations chosen. Note that as the sensors are exchangeable due to independence, it makes no difference which particular size M_k subset of the sensors is chosen. Write again $I_n = \sigma(\{\mathbf{X}_j^{(M_j)}\}_{j=1}^n)$ to represent all available information to the observer at time instance n. Since M_n is determined by the user, it is again of the form

$$M_n = \mu(I_{n-1}),$$

where μ is a control function, similar to the previous section. The only difference is that μ is now a map from the information set to the non-negative integers, rather than a binary function as in the previous section. The rest of the formulation is identical to the previous section. That is, we again want to find a stopping time T and a sampling policy

μ such that ADD is minimized given constraints on PFA and ANO. In this case, ANO is naturally defined as

$$\text{ANO} := \mathbb{E}\left[\sum_{n=1}^{t\wedge(T-1)} M_n\right].$$

A solution to this problem was studied in [52]. Similar to the single sensor case, the optimal stopping rule T^* is again a Shiryaev time $T^* = \inf\{n : \pi_n \geq A\}$ for some A. The optimal sampling policy, however, is more complex than eariler and not easily expressed in closed form. Nonetheless, in experiments the authors find that a numerically approximated optimal sampling policy chooses more sensors for time-step $n+1$ when π_n is near 0.5, compared to when π_n is near 0 or 1. The intuitive interpretation offered by the authors is that when the posterior probability of change π_n is around 0.5, there is the most uncertainty about whether a change has taken place or not. In these circumstances it is worth putting a lot of resources to the sensing, in order to resolve the uncertainty quickly one way or another. On the other hand, when π_n is for example very small, we have a strong indication of whether the change has taken place or not, and thus there is not as much need to use additional data.

Other than these two, several other works have considered discrete time single change-point detection in which only a part of the observations is available. Perhaps the most relevant to this section and the following chapters are [25] and [24], where quickest change detection problems with sampling right constraints were considered in the deterministic and the Bayesian frameworks, respectively. Quickest deterministic change-point detection over multiple data streams was considered in [26], where the observer can only observe one data stream at each time slot. Another way to reduce communication load between the sensors and the FC is via a censoring approach [35, 38], where a sensor sends its data to the FC only if it deems the observation to be informative. Therefore, instead of the FC choosing beforehand which sensors it wants to communicate with in a given timestep, the decision on whether to communicate is made by the individual sensors. The observation is sent if the likelihood ratio corresponding to the observation falls outside a predetermined "no-send region". This region is chosen such that required communication constraints are met.

Chapter 3

Multiple hypothesis testing

3.1 Fundamentals of multiple hypothesis testing

The problem of multiple hypothesis testing has been a popular topic in the statistical literature dating back to Tukey [74] in the 50's. More recently, It considers settings where a statistician is tasked with testing a large number of hypotheses simultaneously. In general, the hypotheses can be independent or dependent, and tested using data from a single or multiple data sets. In ordinary testing of a single hypothesis, the standard approach is to constrain the probability of a Type I error[1] below a certain threshold. However, when performing multiple tests at once, constraining the individual Type I errors independently may not lead to a desirable overall accuracy.

The following simple example illustrates the point. Suppose we have N hypotheses, for all of which the null hypothesis holds. If the Type I error probability is bounded by $\alpha > 0$ for each individual hypothesis, the probability of making at least one Type I error in the set is $1 - (1 - \alpha)^N$, which approaches 1 quickly for a constant α as N increases. As an example, for the canonical $\alpha = 0.05$, already $N = 50$ results in the probability of making at least one false rejection being greater than 90%. This probability of making at least one false rejection is known in the literature as family-wise error rate (FWER), which can be controlled using the Bonferroni correction [16, pp. 35]. The Bonferroni procedure tests each hypothesis at level α/N, which bounds the FWER below α. The drawback is that already for a moderate N the level α/N becomes very small, leading to minimal detection power[2].

Thus, when testing a large number of hypotheses, trying to avoid all false discoveries (false positives) may not be feasible, as it comes with a substantial cost on the detection

[1] Type I error: rejecting a null hypothesis when the null hypothesis is true.

[2] Detection power is the probability of rejecting a null hypothesis when the null hypothesis false.

	Accepted	Rejected	Total
True null hypotheses	$N_0 - V$	V	N_0
False null hypotheses	$N - N_0 - S$	S	$N - N_0$
	$N - R$	R	N

Table 3.1: A decision rule $\mathcal{I}$ has rejected a total of R out of N hypotheses, of which V were incorrect rejections and S correct rejections. The false discovery proportion is $V/(V+S) = V/R$.

power. Instead, we could allow for some false alarms as long as the *proportion of false discoveries among all discoveries* remains tolerable. Making a few incorrect rejections is tolerable if they come among dozens of correct rejections. This intuitive notion was introduced and formalized by Benjamini and Hochberg in 1995 [6] as false discovery rate (FDR). Note that in the example of the previous paragraph, all discoveries were false discoveries.

It is necessary to introduce some notation in order to describe the FDR concept. Let $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_N$ be the set of hypotheses tested. With a slight abuse of notation, $\mathcal{H}_i = 0$ is used to denote that the ith null hypothesis is true, and $\mathcal{H}_i = 1$ that it is false. Of the N null hypotheses, a total of $N_0 := \sum_{i=1}^{N} 1 - \mathcal{H}_i$ are true. We have data $\mathbf{Y} = (Y_1, Y_2, ..., Y_N)$ such that $Y_k \in \mathbb{R}$ is the data (often a test statistic) related to the kth hypothesis. A *multiple testing rule* $\mathcal{I} : \mathbb{R}^N \mapsto 2^N$ is a mapping of the data to a binary sequence of length N so that $\mathcal{I}(\mathbf{Y})_k = 1$ if the kth hypothesis is rejected, and zero otherwise.

In terms of this notation, the number of false rejections produced by the rule $\mathcal{I}$ is

$$V(\mathcal{I}) := \sum_{k:\mathcal{H}_k=0} \mathcal{I}(\mathbf{Y})_k$$

and the total number of rejections is

$$R(\mathcal{I}) := \sum_{k=1}^{N} \mathcal{I}(\mathbf{Y})_k.$$

The dependence on $\mathcal{I}$ is dropped from the notation when the rule in question is obvious. The notation is summarized in Table 3.1.

The false discovery proportion (FDP) is then defined as $\text{FDP}(\mathcal{I}) = V/R$, with the natural definition of FDP $= 0$ when $R = 0$, i.e. when no rejections are made. Observe that FDP is a random variable, since the output of $\mathcal{I}(Y)$ is a random variable. The FDR is given as the expectation of FDP, that is

Definition 7 (False discovery rate)**.**

$$\mathrm{FDR}(\mathcal{I}) = \mathbb{E}\left[\frac{V(\mathcal{I})}{R(\mathcal{I})}\right].$$

Some key properties of the FDR follow immediately from the definition. Note that if all the null hypotheses are true, i.e. $N_0 = N$, any discovery is a false discovery. Thus if $V = 0$ then FDP $= 0$, and if $V > 0$, then FDP $= 1$. Hence in this case FDR $= \mathbb{P}(V > 0)$. That is, when all hypotheses are true, FDR$(\mathcal{I})$ is equivalent to FWER$(\mathcal{I})$, as $\mathbb{P}(V > 0)$ is the definition of FWER.

If $N_0 < N$, the FDR is upper bounded by the FWER. This follows from observing that if $V > 0$ then $V/R \leq 1$, so that $V/R \leq \mathbf{1}_{\{V>0\}}$. Taking expectations from both sides yields FDR$(\mathcal{I}) \leq$ FWER$(\mathcal{I})$. Thus a procedure that controls the FWER also controls the FDR, but not vice versa. Heuristically, when the the number of false null hypotheses is large, there are more opportunities for correct rejections S, which inflate the denominator $R = V + S$ in the definition of FDR. Since FWER is independent of S, the difference between FDR and FWER is larger the more false null hypotheses there are.

In addition to the definition of FDR introduced here, other closely related variants variants have been introduced in the literature. Examples include pFDR $:= \mathbb{E}(V/R|R > 0)$ and mFDR $:= \mathbb{E}V/\mathbb{E}R$, see [66] for discussion of connections and interpretations these criteria.

3.2 Controlling the false discovery rate

Since the introduction of the FDR criterion to the statistical literature in 1995, numerous approaches have been developed for controlling the FDR in different circumstances. Still, to this day, arguably the most common procedure for controlling the FDR is the Benjamini-Hochberg (BH) procedure, introduced in the original paper [6]. In the paper it was assumed that the data $\mathbf{Y}$ are p-values[3] corresponding to the hypotheses. Moreover, the p-values generated by null hypotheses were assumed to independent. The algorithm in [6] is first presented in a general form, only assuming that the statistics follow a known cumulative distribution function F_0 independently when the null hypothesis is in place, and that small values of the data are critical for the null hypothesis. The most well known version of the BH-algorithm for conditionally independent p-values is then obtained from the general definition.

Since nothing else is known about the hypotheses, it is natural to consider a threshold based rule that rejects the kth hypothesis if $Y_k \leq t(\mathbf{Y})$, where $t(\mathbf{Y})$ is a data dependent

[3] p-value: probability of observing data at least as extreme as what was observed, conditional on the null hypothesis being correct.

threshold to control the FDR. Hence for the BH-algorithm $\mathcal{I}^{BH}$, the decision rule regarding the kth hypothesis is of the form,

$$\mathcal{I}^{BH}(\mathbf{Y})_k = \mathbf{1}_{\{Y_k \leq t(Y)\}}. \tag{3.1}$$

Since a higher threshold results in more rejections, in order to maximize detection power, we would like to choose the highest threshold $t(\mathbf{Y})$ that still controls the FDR below a user specified tolerable FDR level α. For the BH-procedure this threshold is given by

$$t(\mathbf{Y}) = \max\{t \in Y_1, ..., Y_N : \widehat{\text{FDP}}(t) \leq \alpha\}, \tag{3.2}$$

where $\widehat{\text{FDP}}(t)$ is an estimator of FDP if threshold t were to be used. The estimator is given by

$$\widehat{\text{FDP}}(t) = \frac{F_0(t)}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{Y_i \leq t\}}}. \tag{3.3}$$

The numerator of this estimator corresponds to the theoretical proportion of statistics generated under the null hypothesis that do not exceed the value t. The denominator is the empirical proportion of the observed data that does not t. Intuitively, if the denominator is large compared to the numerator, it implies that the data contains more extreme values than would be expected if all of the observations were generated under the null hypothesis. Thus we can expect some of these small values to correspond to false null hypotheses.

In the special case that $\mathbf{Y}$ contains independent p-values, $F_0(m) = m$, for $m \in [0, 1]$, since p-values are uniformly distributed under the null hypothesis (see Remark 1 below). Let $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(N)}$ be a rank-ordering of $\mathbf{Y}$. Then by definition

$$\frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\{Y_{(k)} \leq Y_{(m)}\}} \geq m/N,$$

and hence

$$\widehat{\text{FDP}}_{\text{p-values}}(Y_{(m)}) \leq \frac{N}{m} Y_{(m)}. \tag{3.4}$$

Then the correct threshold is found from (3.2) as

$$t(\mathbf{Y}) = \max\{Y_{(k)} \in \mathbf{Y} : Y_{(k)} \leq \frac{k}{N}\alpha\}. \tag{3.5}$$

Remark 1. In general, p-values might not be exactly uniformly distributed in $[0, 1]$. This is the case when, for example, the test statistic is discrete valued. For general experiments,

it is typically assumed (see e.g. [54]) that p-values are stochastically lower bounded by a uniform distribution on $[0, 1]$ under the null hypothesis, viz.

$$\mathbb{P}(Y_k \leq m) \leq m, \quad 0 \leq m \leq 1, \text{ when } \mathcal{H}_k = 0.$$

Note that the upper bound of (3.4) still applies under the more general condition.

For completeness, the FDR control property of the BH-algorithm is stated in a separate theorem. A rigorous proof of the FDR control of the BH-procedure is found in [6].

Theorem 3 (Benjamini-Hochberg)**.** *Let $Y_{(1)}, Y_{(2)}, ..., Y_{(N)}$ be p-values that under the null hypothesis are independent and stochastically lower bounded by a uniform distribution. For the BH-decision rule $\mathcal{I}_{BH}$ in* (3.1) *with threshold* (3.5) *it holds that*

$$\text{FDR}(\mathcal{I}_{BH}) \leq \frac{N_0}{N}\alpha.$$

In light of the approach and limitations of classical Fisherian null hypothesis testing, the implications of Theorem 3 are arguably quite striking. Under minimal assumptions, the result provides an upper bound for the expected value of the proportion of true null hypotheses among the set of rejected hypotheses. That is, something that could be loosely interpreted as resembling of an estimate of the probability that a rejected hypothesis is actually true. Conflating significance levels and p-values with probabilities that the null hypothesis is true is of course not recommended to any statistics practitioner, but the FDR bound seems to be a step in the direction of giving an estimate for such a probability. Still, having an algorithm that maintains FDR control at level α only means that the random variable defined as the ratio of false rejections to all rejections has expected value at most α, where the expectation is over the possible data generated by the true conditions. This is helpful, but perhaps still not quite as intuitive as one might hope for. Can *anything* be said about the probabilities that the given rejected hypotheses are true? Turns out, that taking a Bayesian viewpoint allows for such an interpretation.

3.2.1 A Bayesian perspective

The foundations of the concept false discovery rate are purely frequentist. Whether hypothesis $\mathcal{H}_k$ is true or not is an unknown deterministic quantity, and expectations in the definition of FDR are taken with respect to the data only. At the same time, multiple testing has a convenient interpretation in the Bayesian context. The so called *two-groups* model popularized by Efron [19] considers the data to be generated from the following

structure:

$$\begin{aligned} &Y_k|(\mathcal{H}_k = 0) \sim f_0 \\ &Y_k|(\mathcal{H}_k = 1) \sim f_1 \\ &\mathbb{P}(\mathcal{H}_k = 0) = \lambda_0 = 1 - \mathbb{P}(\mathcal{H}_k = 1). \end{aligned}$$

That is, the hypotheses are now random variables with some common prior probability λ_0 of being null. The marginal density $f(y)$ of a single observation y is then a mixture

$$f(y) = \lambda_0 f_0(y) + (1 - \lambda_0) f_1(y).$$

Suppose we now observe $Y_k = y_k$, and want to do inference regarding $\mathcal{H}_k$. Right away Bayes rule gives

$$\mathbb{P}(\mathcal{H}_k = 0 | Y_k = y_k) = \frac{\lambda_0 f_0(y_k)}{f(y_k)},$$

a direct estimate for the posterior probability that hypothesis $\mathcal{H}_k$ is null. Suppose then that the data are again p-values, and that F_0 and F represent the cumulative distribution functions of f_0 and f, respectively. Then for all hypotheses $\mathcal{H}_k$ with p-values in $[0, t]$, we get

$$\mathbb{P}(\mathcal{H}_{(m)} = 0 | Y_{(m)} \in [0, t]) = \frac{\lambda_0 F_0(t)}{F(t)}.$$

According to the assumptions of Theorem 3, we can assume F_0 to be known, but λ_0 and F are unknown. However, an upper bound for λ_0 is 1, and F can be estimated by the empirical cumulative distribution function. Then, a conservative estimate of the Bayes probability that a hypothesis with a p-value in $[0, t]$ is true reduces to [16, pp. 20]

$$\frac{F_0(t)}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{Y_i \leq t\}}},$$

which is the same as the $\widehat{\text{FDP}}$ estimator of (3.3) used in the BH-procedure. Thus an alternative interpretation for the hypotheses rejected by the BH-procedure is that for them the estimated Bayes probability of the null hypothesis being true is less than α.

A Bayesian approach to multiple testing provides more insight than just explaining the frequentist methods, of course. The next section presents one approach to multiple hypothesis testing utilizing the two-groups model in settings where there exists correlation between the observations. It is co-authored by the author of this thesis, and published in [27]. In particular, a case where the data points $\mathbf{Y}$ are received from sensors in distinct spatial locations that may or may not observe a spatially varying phenomenon of interest is investigated.

3.3 Multiple hypothesis testing in Sensor Networks

This section is a condensed version of a recent conference paper by the author of this thesis [27]. It presents an approach to multiple testing using data from different locations acquired by the a sensor network. The focus is in performing multiple hypothesis testing in a spatial domain instead of change-point detection like in other parts of this thesis.

As already highlighted in this thesis, large-scale sensor networks are an important new tool for monitoring spatially varying phenomena and fields. Prime examples of such phenomena include radio spectrum, seismic activity, pollution or emission levels, and monitoring of agricultural fields, smart buildings, security and surveillance applications or other inaccessible or hazardous environments. Massively connected and spatially distributed large-scale sensor systems are also a key technology in the Internet of Things. An important application of sensor networks is locating locally homogeneous regions that that are interesting or different in the field and would correspond to anomalies [44]. In this section, a scenario where one needs to make binary decisions about the prevalence of some phenomenon at different spatial locations is considered.

In distributed detection applications sensors generally have only limited computational and communications capabilities, as they need to remain operational over long time periods. Thus, the decision making and majority of the computation takes place in a Fusion Center or a cloud that receives information from all sensors and has sufficient computational resources. Furthermore, the limited bandwidth restricts the sensors to communicate condensed information to the FC about their observations, such as a p-value, z-score, likelihood ratio or some other sufficient statistic corresponding to the hypothesis tested locally. There are multiple ways in which these statistics can be generated by sensors and the sensors may have local computational capabilities, but these are not considered further here. The only assumption will be that the statistics follow some (often known) distribution under the null distribution, and some other (often unknown) distribution under the alternative. Thus, while in most of the examples the statistics are considered to be p-values, p-values are not in anyway fundamental to this problem. They merely represent a common example of a decision statistic with approximately the above mentioned theoretical properties. An example of a sensor network observing spatially varying phenomena is illustrated in Figure 3.1.

After receiving the test statistics, the FC is tasked with a multiple hypothesis testing problem. Hypotheses $\mathcal{H}_1, ..., \mathcal{H}_n$ associated with n sensors are tested simultaneously. For each i the null hypothesis is associated with noise-only case, i.e. that no signal is present at the location of sensor i. For each hypothesis $\mathcal{H}_i$ a corresponding test statistic y_i is received, and the task is to decide whether y_i was caused by signal ($\mathcal{H}_i = 1$) or noise ($\mathcal{H}_i = 0$). In contrast to the rest of the thesis, in this section it is assumed that data only exists from one "snapshot" in time, that is, a collection of observations acquired at the

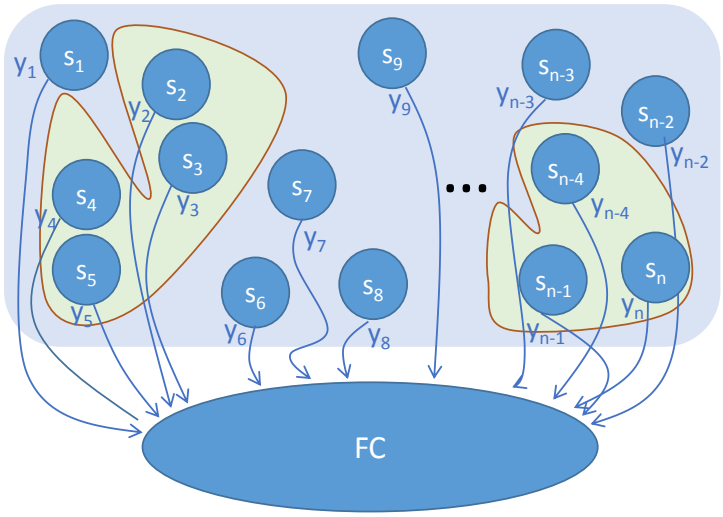

Figure 3.1: Illustration of the assumed system configuration. Circles represent sensors at known locations $\mathbf{S} = \mathbf{s}_1, ..., \mathbf{s}_n$. Each sensor $\mathbf{s}_i$ sends a local test statistic y_i to the FC. Green areas represent hypothetically interesting or different areas of the observed field where signal is present, i.e. the alternative hypotheses are in place. The FC is tasked with determining which of the sensors are located in these regions using only $\mathbf{S}$ and $\mathbf{y} = y_1, ..., y_n$ as information.

same time instance, and all inference is made using this one set of observations only.

The standard multiple hypothesis testing approaches like the Bonferroni and Benjamini-Hochberg methods do not utilize location information related to the test statistics, and can thus have very low power in spatial inference applications. Here we take an approach for multiple hypothesis testing that exploits the underlying spatial structure of the field and sensor location information in order to improve detection power. Spatial fields such as received signal power in radio spectrum or pollution levels in the ambient air tend to vary locally and smoothly. This will cause the significant hypotheses to appear in spatially localized clusters. A latent Gaussian process [55] is employed to locate these areas and in effect "relax the significance threshold" there, while making it more strict in areas where the test statistics seem to be mostly produced by noise. A flexible Bayesian framework is adopted by extending the prevalent two-groups model [18] introduced in Section 3.2.1 to account for spatial dependency. The approach is fundamentally non-parametric, but allows for principled inclusion of prior knowledge about the monitored phenomena.

There is recent research activity on the topic of spatial multiple hypothesis testing, for example, in [32], [4] and [5], under the assumption that the hypotheses are divided into groups *a priori* based on external information. In [14] and [63] discoveries are divided into disjoint clusters, and the focus of the inference is on the rate of falsely discovered

clusters. Other notable related works include [67] and [47], but in them more restricting assumptions about the type of field and hypotheses are made. The approach proposed in this thesis has the most in common with [68], [77] and [62] since they perform multiple hypothesis testing based only on a set of test statistics and locations information. However, none of these methods are designed for arbitrarily located sensors in a Euclidian space.

3.3.1 System model

We consider a set of test statistics $\mathbf{y} = y_1, ..., y_n$ obtained by the fusion center from n spatially distributed sensors at known, distinct locations $\mathbf{S} = \mathbf{s}_1, ..., \mathbf{s}_n$, where $\mathbf{s}_i$ are d-dimensional coordinate vectors $(s_i^1, ..., s_i^d)$, $d \leq 3$. A natural model for the test statistics is a mixture distribution,

$$p(y_i|\lambda_0) = \lambda_0 \cdot f_1(y_i) + (1 - \lambda_0) \cdot f_0(y_i), \tag{3.6}$$

where f_0 and f_1 denote the densities of the test statistics under null ($\mathcal{H}_i = 0$) and alternative ($\mathcal{H}_i = 1$) hypotheses respectively, and $\lambda_0 \in [0, 1]$ is the latent mixing proportion of null and alternate distributions. This formulation is the famous two-groups model for multiple testing, where λ_0 has an additional interpretation as the *a priori* probability of signal, discussed at length in e.g. [17].

In some cases f_0 can be taken to be the theoretical null distribution of the test statistic, most commonly uniform for p-values and standard Gaussian for z-scores. However, as demonstrated in [15], there are scenarios when the theoretical null model might not be accurate. In those cases the empirical null density can be estimated by fixing a parametric form and estimating the parameters [15]. Once the null density is established, there exist methods for inferring the alternative mixture component. Examples include the beta-uniform mixture model [51] for p-values and non-parametric recursion [40] for general finite mixture models. For sensor networks there might also be additional information or local training data available on the densities based on previous history. Much more discussion the estimation of mixture components can be found in [19]. Once f_0 and f_1 are estimated, we consider them as fixed. An illustration of the estimates using the approach in [51] appears in Figure 3.2.

The main quantities of interest are the posterior probabilities of signal $\mathbf{w} = w_1, ..., w_n$, which, conditional on the model $\mathcal{M}$ in (3.6) can be computed as

$$w_i = \mathbb{P}(\mathcal{H}_i = 1|\mathbf{y}, \mathcal{M}) = \frac{\lambda_0 \cdot f_1(y_i)}{\lambda_0 \cdot f_1(y_i) + (1 - \lambda_0) \cdot f_0(y_i)}. \tag{3.7}$$

A convenient property of these probabilities is that if we choose a subset of the test statistics $\mathbf{y}_1 \subset \mathbf{y}$ such that

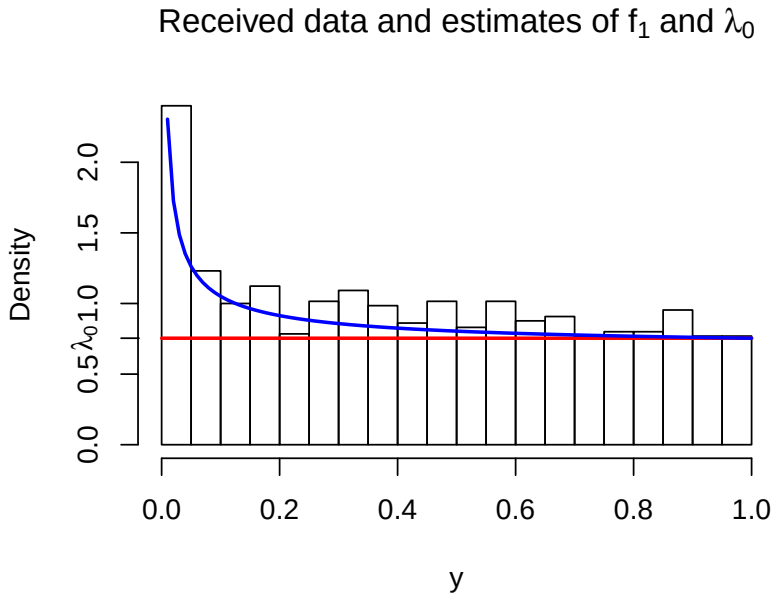

Figure 3.2: An example of how λ_0 and f_1 can be estimated empirically using the beta-uniform mixture model of [51]. The histogram contains the observed test statistics (here p-values), and under the assumption that the null statistics follow a uniform distribution it is possible to decompose the joint distribution to uniform and beta components. The height of the uniform component λ_0 serves also as an estimate for the overall proportion of null and non-null statistics.

$$\left\{ \sum_{i=1}^{n} \mathbf{1}_{\{y_i \in \mathbf{y}_1\}} \right\}^{-1} \sum_{i: y_i \in \mathbf{y}_1} (1 - w_i) \leq \alpha, \tag{3.8}$$

and reject the corresponding hypotheses, we obtain

$$\mathbb{E}\left[\frac{V}{R} \middle| \mathbf{y}, \mathcal{M} \right] \leq \alpha, \tag{3.9}$$

where V and R are the number of false rejections and all rejections, respectively. This

follows from

$$\begin{aligned}\mathbb{E}\left[\frac{V}{R}|\mathbf{y},\mathcal{M}\right] &= \left\{\sum_{i=1}^{n}\mathbf{1}_{\{y_i\in\mathbf{y}_1\}}\right\}^{-1}\mathbb{E}[V|\mathbf{y},\mathcal{M}] \\ &= \left\{\sum_{i=1}^{n}\mathbf{1}_{\{y_i\in\mathbf{y}_1\}}\right\}^{-1}\sum_{i:y_i\in\mathbf{y}_1}\mathbb{E}[\mathbf{1}_{\{\mathcal{H}_i=0\}}|\mathbf{y},\mathcal{M}] \\ &= \left\{\sum_{i=1}^{n}\mathbf{1}_{\{y_i\in\mathbf{y}_1\}}\right\}^{-1}\sum_{i:y_i\in\mathbf{y}_1}(1-w_i)\leq\alpha,\end{aligned}$$

where the first equality stems from the fact that after $\mathbf{y}_1$ is chosen R is a known constant, the second equality from the definition of V and the linearity of expectation, third from the definition of w_i in (3.7) and the inequality from (3.8). Observe that the quantity in (3.9) resembles the traditional FDR, but includes conditioning on the model $\mathcal{M}$. This conditioning on the model is explicitly emphasized due to the fact that whether the probabilities w_i and the resulting bound (3.9) are useful depend critically on the validity of the model we employ.

The basic empirical Bayes approach in the two-groups model is to estimate a common $c = \bar{c}$ for all sites over the whole data set, and do inference without taking advantage of the location information [16]. We incorporate this information by letting the latent mixture proportion c vary over the field. Each sensor location may have its own mixture proportions which are denoted by $\mathbf{c} = (c_1, ..., c_N)$. As can be seen from (3.7), a small c_i will require $f_1(y_i)$ to be very large compared to $f_0(y_i)$ in order for w_i to be large. That is, c_i controls the degree of extremity required from the decision statistic y_i to declare hypothesis $\mathcal{H}_i$ significant.

In the following we will find c_i by using its logit transformation $\beta_i := \phi(c_i) := \log(c_i/(1-c_i))$, which transforms β_i from $[0,1]$ to the whole real line. The choice of a suitable $\boldsymbol{\beta} := (\beta_i, ..., \beta_n)$ is essentially a non-parametric smoothing problem: we want the latent mixture proportions to fit to the observed data, and at the same time not vary too much between nearby locations. We take a Bayesian approach and assign a multivariate Gaussian prior distribution on $\boldsymbol{\beta}$ with a covariance matrix defined by the squared exponential covariance function. The squared exponential kernel is a popular choice for modeling smooth functions. Of course, other covariance kernels might certainly also be applicable. In general the covariance function should be chosen according to the characteristics of the exact application, but this issue is not investigated further here. Formally

[27],

$$p(\boldsymbol{\beta}|\sigma, l) = \mathcal{N}(\phi(\bar{c}), \Sigma) \tag{3.10}$$

$$\Sigma_{[i,j]} = \sigma^2 \exp\left(- \frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{2l^2} \right), \tag{3.11}$$

where σ^2 is the marginal variance and l specifies the length-scale of the spatial dependency. As the constant prior mean of $\boldsymbol{\beta}$ we use the logit transform of the average mixture proportion over the whole data set (see Figure 3.2). The selection of hyperparameters σ and l and their distribution $p(l, \sigma)$ will be addressed in the next subsection.

Combining the likelihood function (3.6) and the priors allows us to form the posterior distribution of $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto p(\mathbf{y}|\pi(\boldsymbol{\beta}))p(\boldsymbol{\beta}|l, \sigma)p(l, \sigma), \tag{3.12}$$

where $\pi(\boldsymbol{\beta}) = \exp(\boldsymbol{\beta})/(1 + \exp(\boldsymbol{\beta})) = \mathbf{c}$.

Unfortunately as the likelihood is in general not Gaussian, the distribution in (3.12) is analytically intractable. However, it is still possible to sample from it using Markov Chain Monte Carlo methods [9]. It is worth noting that constructing the covariance matrix with (3.11) is of complexity $\mathcal{O}(N^3)$, which can become an issue for very large scale networks [64]. In these cases approximate methods can be used, e.g. [58]. For the purposes of this work, we assume the monitored phenomena to be localized, meaning that even large scale problems can be solved by ignoring correlations between sufficiently far away sensors. This is the case for many physical phenomena such as the radio spectrum or air quality.

Once we have obtained an approximation for the posterior distribution of $\boldsymbol{\beta}$, we can collect the posterior signal probabilities $\mathbf{w}$ by integrating over the posterior,

$$\begin{aligned} w_i &= \mathbb{P}(\mathcal{H}_i = 1|\mathbf{y}, \mathcal{M}) = \int_{\beta_i} \mathbb{P}(\mathcal{H}_i = 1|\mathbf{y}, \beta_i) p(\beta_i|\mathbf{y}) d\beta_i \\ &= \int_{\beta_i} \frac{\pi(\beta_i) \cdot f_1(y_i)}{\pi(\beta_i) \cdot f_1(y_i) + (1 - \pi(\beta_i)) \cdot f_0(y_i)} p(\beta_i|\mathbf{y}) d\beta_i. \end{aligned}$$

Finally $\mathbf{w}$ can be plugged into (3.8) to search for the largest subset $\mathbf{y}_1$ such that the condition is fulfilled. As indicated earlier, it is not sensible to claim that that the conditional bound (3.9) that follows from (3.8) is equivalent to strong FDR control in the sense of the Benjamini-Hochberg algorithm, for example. While in the approach proposed in this thesis the decision rule is a simple function of the posterior probabilities, the posterior probabilities themselves are an extremely complicated function of all of the data, locations, covariance functions and prior distributions. Obtaining any useful analytic guarantees for such complex functions under the general conditions considered here is difficult, if not impossible. Still, the bound in (3.9) is not without merit. If the assumed model that

generated the posterior probabilities is valid for the problem, (3.8) and (3.9) show that the posterior probabilities can be used to form a decision rule that provides FDR control under this model.

3.3.2 Choosing the hyperparameters

The choice of the hyperparameters l and σ for the latent process in (3.10) is vital for the validity of the posterior inference. Short length scales l and large marginal variances σ^2 allowing for too much variation in the proportions between nearby locations will overfit to the data and deteriorate the error rate control. On the other hand, a very rigidly varying latent process induced by long length scales and small variances does not exploit the local spatial information sufficiently, resulting in a loss of detection power. Notably, choosing $l = \infty$ and $\sigma = 0$ reduces to the ordinary two groups model with a constant mixing proportion.

It is well established in spatial statistics literature that without prior information on l and σ, only the ratio l/σ is identifiable from the data, and the individual parameters are not [78]. Intuitively this means that we can not distinguish between a process with long length scale and high variance and a process with short length scale and low variance using just the data. Consequently, at least a weakly informative prior should be specified on one of the parameters. Defining such a prior can be difficult if the user has little knowledge about the process. In this particular application, we want to protect ourselves from overfitting, as it could invalidate the FDR control. Hence, we employ the Penalized Complexity (PC) prior derived in [22]. The PC prior penalizes deviance from a nominal baseline model, in this case the constant latent process corresponding to the ordinary two groups model. The magnitude of deviance is captured by the Kullback-Leibler divergence between the base model and the more flexible model where the latent variable is allowed to vary. This is a good foundation for FDR control, as the constant base model is a well studied model for multiple inference, and we will deviate from it only if the data strongly indicates so.

For a squared exponential kernel in d dimensions the PC prior is defined as [22],

$$p(l, \sigma) = \frac{d}{2}\tilde{\lambda}_1\tilde{\lambda}_2 l^{-d/2-1}\exp(-\tilde{\lambda}_1 l^{-d/2} - \tilde{\lambda}_2\sigma),$$
$$\tilde{\lambda}_1 = -\log(\alpha_1)l_0^{d/2} \quad \text{and} \quad \tilde{\lambda}_2 = -\frac{\log(\alpha_2)}{\sigma_0},$$

where $\alpha_1, \alpha_2, \sigma_0$ and l_0 are user defined constants implying tail prior probabilities $\mathbb{P}(l < l_0) = \alpha_1$ and $\mathbb{P}(\sigma > \sigma_0) = \alpha_2$. These allow the user to input their prior knowledge about the expected size of the monitored phenomena, with sensors $2l$ apart treated as practically independent.

3.3.3 Simulation examples

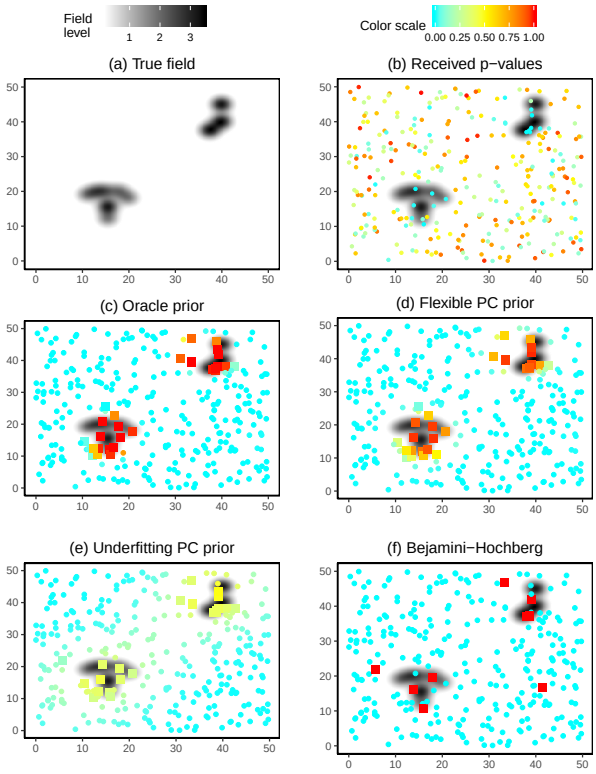

Figure 3.3: (a) True field level in the example. (b) Received p-values. (c-e) Posterior mean of $\mathbf{c}$ for different prior choices. Locations marked by squares are declared as significant for $\gamma = 0.2$. The smooth latent variable causes the discoveries to appear in clusters. The oracle prior and the flexibly chosen PC prior adapt to the field and identify the signal regions very reliably. The overly conservative PC prior does adjust the latent variable in approximately right areas, but not very accurately. (f) The standard Benjamini-Hochberg procedure applied to the pure p-values. Red squares denote sensors declared as significant for $\gamma = 0.2$. Lack of spatial information makes the BH procedure clearly the least proficient method.

In this subsection, a numerical experiment is conducted to demonstrate the advantage of taking the spatial dependency into account [27]. We consider a two dimensional field, where two signal sources cause changes in the field level monitored by the sensor network (Fig. 3.3(a)). Each of randomly placed $n = 300$ sensors in locations $\mathbf{S}$ acquire a noisy observation $x_i = \mu_i + \mathcal{N}(0, 1)$, where μ_i is the true field level at site $\mathbf{s}_i$. p-values y_i are computed and sent to the FC by performing a one-sided Z-test against the null hypothesis that $\mu_i = 0$. That is, $y_i = 1 - \Phi(x_i)$, where Φ is the CDF of the standard Gaussian. Fig.

3.3(b) illustrates these p-values in one exemplary simulation iteration. We estimate f_1 and $\bar{c}$ from the data using the beta-uniform mixture model proposed in [51]. The MCMC sampling is done using Stan software [10].

The model is tested with three different choices for the hyperparameters. First is an oracle setup, where we fix l and σ to near optimal values ($l = 5, \sigma = 10$) using our knowledge of the true field. The oracle model is compared with two different PC priors, one with extremely conservative constants causing underfitting ($\sigma_0 = 3, l_0 = 25$) and one that still promotes long length scales, but allows for more flexibility ($\sigma_0 = 10, l_0 = 10$). For the PC priors, α_1 and α_2 are set to .05. For reference, the standard Benjamini-Hochberg procedure is applied to the p-values. The capability of the methods to control the FDR at the tolerated level γ and the detection power in discovering true alternative hypotheses are considered. True positive rate (TPR), which is the proportion of true alternative hypotheses discovered, is used as a measure of detection power.

Figure 3.3(c)-(e) highlights the properties of different prior choices. The colors in the figures correspond to the posterior mean of $\mathbf{c}$, and the square shapes denote the hypotheses declared discoveries when $\alpha = 0.2$. The models with the oracle prior and the Flexible PC prior accurately discover the likely regions of signal. Notably, the Flexible PC model adapts to the data well despite conservative prior information. The underfitting PC prior does find some spatial structure, but the obtained structure is quite fuzzy. Consequently, as can be seen in Table I, the well chosen prior distributions are more powerful in locating true alternative hypotheses. However, even the badly misspecified conservative PC prior still controls the FDR below the limit.

Table 3.2: Simulation results averaged over 60 Monte Carlo trials. All models limit the false discovery rate below the specified limit of $\alpha = 0.2$. The models that adjust to the data accurately provide the best results. The Flexible PC model produces almost comparable power to the Oracle prior. The standard Benjamini-Hochberg method leads to a very low detection power, as expected.

Method	FDR	TPR
Oracle	.19	.74
Flexible PC	.17	.65
Underfitting PC	.17	.32
Benjamini-Hochberg	.19	.15

3.3.4 Discussion

The purpose of the work presented in this section and in the publication [27] was to study multiple hypothesis testing with false discovery rate control in spatial domains. Gaussian

processes are a workhorse of modeling in spatial statistics, and the two-groups model is a popular empirical Bayes approach for multiple testing, so a combination of these two approaches as outlined in this section is a natural starting point for the problem at hand, and to the best of my knowledge, had not explicitly been studied before, at least as of the publicatio of [27]. Spatial modeling using Gaussian processes and the two-groups model are products of two significant, disjoint lines of research, and the approach of this section presents some preliminary work in combining the two. At the same time, I think that this intersection could be worth studying, and that especially the Penalized Complexity approach for prior selection lends itself well to be combined with the two-groups model and the goal of FDR control.

Two of the more important things that require further investigation are the study of robustness for model misspecification and the choice of covariance functions and priors. For practical use, since the specific choice of these is dictated almost entirely by the application, the most relevant next step might be to focus on a specific application with an underlying physical phenomenon (e.g. radio spectrum occupancy monitoring), and construct and analyze methods with that specific application in mind. Further, combining the spatial focus of this section with the temporal focus of the rest of the thesis in order to perform spatio-temporal change-point detection could be an interesting direction.

3.4 Multiple hypothesis testing and change-point detection

In this subsection, the multiple hypothesis testing approach is combined with change-point detection. As an outcome, novel multiple change-point methods for multi-stream data that control the FDR are introduced in the chapter. Let us consider a setup where K independent data streams $\{X_n^{(k)}\}$ are observed simultaneously. In each data stream, a distinct change-point $t^{(k)}$ occurs. Therefore, a stopping rule $T^{(k)}$ must be defined for all $k \in [K]$, where we use $[K] := \{1, 2, ..., K\}$. In practical applications, the detection procedure must be stopped within some finite time interval. Thus we assume that there exists a deadline $N_{\max}$, such that if a change-point in the kth data stream has not been declared before time $N_{\max}$, we declare that there is no change-point in the kth data stream, and set $T^{(k)} = \infty$. Using the terminology of this chapter, a discovery occurs when a change-point is declared, i.e. on the event $\{T^{(k)} < \infty\}$. The total number of discoveries among the K detection tasks is then

$$R = \sum_{k=1}^{K} \mathbf{1}_{\{T^{(k)} < \infty\}}. \tag{3.13}$$

Naturally, a false discovery occurs on the event $\{T^{(k)} < t^{(k)}\}$, and the total number of false discoveries is

$$V = \sum_{k=1}^{K} \mathbf{1}_{\{T^{(k)} < t^{(k)}\}}. \tag{3.14}$$

To control the FDR, we then need to construct a set of stopping rules $\{T^{(k)}\}_{k=1}^{K}$ such that $\mathbb{E}[V/R] \leq \alpha$ for some user-specified α. This problem was recently first studied by Chen, Zhang and Poor in a set of papers [11, 12, 13] in both Bayesian and non-Bayesian change-point formulations. Their method for the Bayesian case, called MD-FDR, will be introduced as a special case of one of the method propose in the next chapter. In [13], it is shown that MD-FDR achieves control asymptotically as the number of data streams grows, meaning that FDR $\leq \alpha + o_K(1)$ where $o_K(1) \to 0$ as $K \to \infty$ as long as $N_{\max}$ scales with K sufficiently fast. In Chapter 4, among other things, it is shown that MD-FDR in fact achieves FDR control for any fixed K and $N_{\max}$.

Chapter 4

Multiple change-point detection under communication and FDR constraints

In this chapter, data-efficient Bayesian methods for change-point detection in multiple parallel data streams are derived an analyzed. This chapter is a condensed version of the publication [42] co-authored by the author of this thesis. Preliminary ideas and results were presented in [28]. The proposed methods are based on the fundamentals of Bayesian change-point detection and ideas of data-efficient change detection from Chapter 2, and the concept of false discovery rate from Chapter 3.

The problem of rapidly detecting change-points in multiple data streams is considered. In particular, a fusion center (FC) receives independent data streams from multiple sensors in a large-scale sensor network. We assume that the sensors acquire observations from their environment all the time. However, each sensor communicates its observation to the FC only if the FC decides to actively monitor and receive an observation from this sensor. This can be implemented, for example, by sending a control signal from the FC to the sensor. Due to energy considerations similar to those discussed in Section 2.2 and a potentially massive number of sensors in the network, at a given time slot the FC monitors only a subset of the active data streams. A data stream is called active, if no change-point has yet been declared in it. The size of this subset is selected to be a fixed proportion of the number of active data streams. The FC performs computations based on the received data and chooses which particular sensors monitor at each time slot. Each data stream may or may not have a change-point.

The main contributions of this work are:

1. A Bayesian sequential procedure, named the sequential maximum a-posteriori probability (S-MAP) procedure, is proposed. This procedure detects the change-points in all of the data streams, while controlling the FDR. The proposed procedure is based on sequentially updating the sensors' posterior probabilities of change-points having

occurred. Then, at each time slot we choose to monitor a subset of the sensors with the highest posterior probabilities within the allowed proportion. This approach aims to minimize the time between change-point occurrence and its declaration by monitoring the sensors for which change-point occurrence is most probable given the data. The S-MAP procedure satisfies the same Type I error constraints as in [13] and extends this work to communication constrained scenarios. The FDR control of the S-MAP procedure is established using analytical tools.

2. We develop an improved S-MAP (IS-MAP) procedure that outperforms the S-MAP procedure in the sense that it has lower ADD and required average number of observations (ANO). The decrease of the ADD and ANO is obtained by reducing the detection threshold values of the IS-MAP procedure compared to the S-MAP procedure. The IS-MAP procedure has higher FDR than the S-MAP procedure. However, it is proved analytically that the FDR of the IS-MAP procedure is still controlled under the desired level despite its lower detection threshold values.

3. The asymptotic ADD behavior of the S-MAP and the IS-MAP procedures is established analytically for a geometric prior distribution of the change-points. It is shown that for any proportion value, both detection procedures are scalable in the sense that their asymptotic ADD does not increase with the number of data streams. In addition, the asymptotic ADD improvement that is obtained by using the IS-MAP procedure in comparison to the S-MAP procedure is characterized quantitatively.

4. Deeper asymptotic analysis of the IS-MAP procedure is performed in terms of ADD and ANO. This analysis characterizes the tradeoff between reducing the ADD and reducing the ANO communicated until change-points are declared. The proposed ADD-ANO tradeoff analysis can be useful for developing distributed statistical inference procedures using large-scale sensor networks in limited communication capability scenarios.

5. We conduct simulations in order to evaluate the performance and to verify the established theoretical properties of the S-MAP and the IS-MAP procedures.

Next, the detection problem at hand is formulated.

4.1 Problem formulation

An FC can monitor $K \geq 2$ discrete time data streams denoted by $\{X_n^{(k)}\}_{n=1}^{\infty}$, $k \in [K]$. For the kth data stream there is a change-point, $t^{(k)} \geq 1$, $\forall k \in [K]$. A data stream is called active if a change-point has not been declared for this data stream. After a change-point

is declared, the corresponding data stream is stopped and is no longer monitored by the FC.

Let $(\Omega, \mathcal{F})$ denote a measurable space with sample space, Ω, and σ-algebra, $\mathcal{F}$. For this measurable space there exists a family of probability measures, $\{\mu^{(k)},\ k \in [K]\}$. Expectation and probability with respect to $\{\mu^{(k)},\ k \in [K]\}$ are denoted by $\mathbb{E}_\mu[\cdot]$ and $\mathbb{P}_\mu(\cdot)$, respectively. In the kth data stream under μ_k, the change-point, $t^{(k)}$, is a random variable with a known prior distribution. We allow the case in which a data stream may not have a change-point so $t^{(k)} = \infty$ occurs with a known probability that may not be equal to zero. The prior distributions of the different change-points are not necessarily the same. Given $t^{(k)}$, we assume that $\{X_n^{(k)}\}_{n=1}^{t^{(k)}-1}$ are independent and identically distributed (i.i.d.) with known probability density $f_0^{(k)}$ and $\{X_n^{(k)}\}_{n=t^{(k)}}^{\infty}$ are i.i.d. with another known probability density $f_1^{(k)}$ and statistically independent of $\{X_n^{(k)}\}_{n=1}^{t^{(k)}-1}$. We assume that the data streams are mutually statistically independent. This assumption holds, for example, when data streams are communicated by sensors with large displacements. Moreover, the sensors experience statistically independent observation noises.In some applications, the data streams independence assumption may not be fully valid. However, this assumption allows for analytic derivation of performance guarantees for multiple change-point detection procedures that can be used as benchmarks for more complex dependence structures among the data streams.

Due to communication limitations, at a given time slot we choose a subset of data streams to observe among the active data streams. Let $K_n \in \mathbb{N}$ denote the number of active sensors at time slot n. We set a fixed proportion value $q \in [0, 1]$ and observe $\lceil qK_n \rceil \in \mathbb{N}$ of the active data streams, where $\lceil \cdot \rceil$ is the ceiling operator. The actual data vectors that are sequentially observed by the FC are denoted by $\{Y_n\}_{n=1}^{\infty}$. The subset of sensor indices that are monitored at each time slot is denoted by $s_n \subset [K],\ n = 1, 2, \ldots$. The FC is able to choose s_n according to some sampling strategy and based on all the availabl data up to the current time slot. The filtration at time slot n, $\mathcal{F}_n := \sigma(\{Y_m, s_m\}_{m=1}^{n})$, is the σ-algebra generated by the pairs $\{Y_1, s_1\}, \ldots, \{Y_n, s_n\}$. In addition, we define the filtration of all the data as $\mathcal{F}_\infty := \sigma(\{Y_m, s_m\}_{m=1}^{\infty})$.

Recall from Section 2.1.1 that the posterior probability of the change-point having occurred is a central quantity in Bayesian change-point detection. The posterior probability of the event $\{t^{(k)} \leq n\}$ using the data up to time slot n is defined as

$$\pi_n^{(k)} := \mathbb{P}_\mu(t^{(k)} \leq n | \mathcal{F}_n),\ n = 1, 2, \ldots,$$

Under the assumed Bayesian model, by using Bayes' rule we can recursively compute $\pi_n^{(k)}$ as follows:

$$\pi_n^{(k)} = \begin{cases} \frac{L(f_1^{(k)}, f_0^{(k)}, X_n^{(k)})\phi_n^{(k)}}{L(f_1^{(k)}, f_0^{(k)}, X_n^{(k)})\phi_n^{(k)} + 1 - \phi_n^{(k)}},\ k \in s_n \\ \phi_n^{(k)},\ k \notin s_n \end{cases}, \tag{4.1}$$

where $L(f_1^{(k)}, f_0^{(k)}, X_n^{(k)}) := f_1(X)/f_0(X)$, and $\phi_n^{(k)} := \pi_{n-1}^{(k)} + \rho_n^{(k)}(1 - \pi_{n-1}^{(k)})$ in which $\rho_n^{(k)} := \mathbb{P}_\mu(t^{(k)} = n | t^{(k)} \geq n)$ depends on the prior distribution of the kth change-point. In case $k \in s_n$, then at time slot n an observation is received from sensor k and $\pi_n^{(k)}$ is computed using the observations received before time slot n, the prior distribution of $t^{(k)}$, and the new observation, $X_n^{(k)}$. The posterior update in (4.1) for the case $k \notin s_n$ corresponds to the case in which at time slot n we do not receive an observation from sensor k. In this case, $\pi_n^{(k)}$ is computed using only the observations received before time slot n and the prior distribution of $t^{(k)}$.

In the considered problem, we have to define multiple stopping rules $T^{(k)}$, $k \in [K]$, where the event $\{T^{(k)} \leq n\}$ is measurable w.r.t. $\mathcal{F}_n$. In practice, the detection procedure must be stopped at within some finite time interval. Thus, we allow the existence of a deadline $N_{\max}$ for the change-points detection. If a change-point in the kth data stream has not been declared before time slot $N_{\max}$, we declare that there is no change-point in the kth data stream and set $T^{(k)} = \infty$. The FDR criterion is defined as follows

$$\text{FDR} := \mathbb{E}_\mu \left[\frac{V}{R \vee 1} \right], \tag{4.2}$$

where V and R are as in (3.14) and (3.13).

The detection delay ADD for the kth data stream is defined as

$$\text{ADD}_k := \mathbb{E}_\mu [0 \vee (T^{(k)} - t^{(k)})], \tag{4.3}$$

where $\infty - \infty := 0$. We define the overall ADD as

$$\text{ADD} := \frac{1}{K} \sum_{k=1}^{K} \text{ADD}_k. \tag{4.4}$$

Assume that at time slot n, we have K_n active data streams. Then, we observe $\lceil qK_n \rceil$ of them. We define the average number of observations (ANO)

$$\text{ANO} := \mathbb{E}_\mu \left[\frac{1}{K} \sum_{n=1}^{(N_{\max}-1) \wedge T^{(k,\sup)}} \lceil qK_n \rceil \right], \tag{4.5}$$

where and $T^{(k,\sup)} := \sup_{k \in [K]} T^{(k)}$. The ANO definition is related the ANO definitions appearing in Section 2.2. However, as opposed to the works [1] and [52], no explicit constraint on the ANO is assumed.

One would be interested in minimizing the ADD under upper bound constraints on the FDR and the ANO. However, to the best of our knowledge no tractable solution to

such an optimization problem has been found yet. In fact, there is currently no tractable solution for minimization of only the ADD under upper bound constraint on the FDR [13]. Therefore, suboptimal procedures that satisfy the communication constraints, control the FDR, and attain low ADD and ANO are developed. In the following section, a procedure called S-MAP (sequential maximum a posteriori) is introduced. S-MAP is a Bayesian multiple change-point detection procedure that controls the FDR and satisfies the limitation on the proportion of sensors communicating their data streams to the FC.

4.2 S-MAP detection procedure

In this section, a Bayesian detection procedure that is tasked to eventually discover all the random change-points that occur in the monitored environment is derived. At a given time slot, each sensor is considered individually and its posterior probability from (4.1) is evaluated using the recursive formula from (4.1). At time slot n, there are K_n active data streams of which we observe only a subset of size $\lceil qK_n \rceil \in \mathbb{N}$. The developed S-MAP procedure extends the MD-FDR method from [13] by proposing a rule for choosing the subset of $\lceil qK_n \rceil$ data streams to observe. Hence, the method allows for saving energy and makes the inference scalable for a large number of sensors or data streams. In the S-MAP procedure, we use the posterior probability from (4.1) as a test statistic, rather than a transformation of the posterior probability used in [13]. However, the methods are equivalent for $q = 1$ where all the active data streams are monitored in parallel.

Under the communication limitations, among the K_n active data streams, we choose to observe the $\lceil qK_n \rceil$ data streams with the highest posterior probabilities of a change-point having occurred. The motivation for the S-MAP approach is that we are interested in minimizing the time between the occurrence of a change-point and its declaration using the sequentially updated posterior probabilities. Another reason for sampling the active data streams with highest posterior probabilities is that we often have the most uncertainty regarding the change-points in these data streams and sampling them is essential in order to get more information. This intuition stems from the sampling policies discussed in Section 2.2. In the following, we describe the proposed S-MAP procedure.

We construct a descending set of K thresholds $Q_r,\ r \in [K]$,

$$Q_r = 1 - \frac{r\alpha}{K},\ r \in [K]. \tag{4.6}$$

The different thresholds are chosen using a similar approach as the p-values thresholds from [6] to obtain FDR control. In particular, the thresholds from (4.6) guarantee that the detection on the kth data stream that samples until $\pi_n^{(k)} \geq Q_r$ has a Type I error probability that is smaller than or equal to $\frac{r\alpha}{K}$, where $\alpha \in (0, 1)$ is the predefined FDR

tolerance level. Formally,

$$\mathbb{P}_\mu(\exists n < t^{(k)} \text{ s.t. } \pi_n^{(k)} \geq Q_r) \leq \frac{r}{K}\alpha. \tag{4.7}$$

The requirements in (4.7) are used in [13, Theorem 1] to show the FDR control of the MD-FDR procedure. It will be shown in Section 4.3 that (4.7) may be too conservative.

The proposed S-MAP detection procedure is divided into sampling stages. Each sampling stage may take several time slots. In the beginning of a sampling stage, we gather all the active data streams and obtain observations from a subset of them, according to the S-MAP approach. This process is repeated at each time slot sequentially until the deadline is reached or in case at least one active data stream posterior probability exceeds its corresponding threshold. If the latter happens, then changes are declared for some of the active data streams, which are then eliminated from the active data streams set.

Let I_j denote the set of indices of active data streams with cardinality $|I_j|$ at the beginning of the jth sampling stage and let n_j denote the time slot at the end of the jth sampling stage. Note that $I_1 = [K]$ and $n_0 = 0$. The jth stage of sampling is described as follows:

1. Sample the $\lceil q|I_j| \rceil$ data streams with the currently highest posterior probabilities.
2. Update the posterior probabilities of the sensors with active data streams using (4.1).
3. Sort the updated posterior probabilities in ascending order as $\pi_n^{(i(n,l))}$, where $i(n,l)$ denotes the index of the lth ordered posterior probability at time slot n.
4. Repeat this process until time slot n_j in which at least one of the posterior probabilities is higher than its corresponding threshold or in case the deadline, $N_{\max}$, is reached, i.e. $n_j = N_{\max} \wedge \min\{n > n_{j-1} : \exists l \in [|I_j|], \pi_n^{(i(n,l))} \geq Q_{K-l+1}\}$.
   (a) If $n_j < N_{\max}$: Declare change-points for the data streams $i(n_j, l_j), i(n_j, l_j + 1), \ldots, i(n_j, |I_j|)$, where $l_j = \min\{l \in [|I_j|] : \pi_{n_j}^{(i(n_j,l))} \geq Q_{K-l+1}\}$ and remove these data streams from the set of active data streams. Update I_{j+1} to be the set of indices of the remaining active data streams. Stop the procedure if $|I_{j+1}| = 0$.
   (b) Otherwise $n_j = N_{\max}$: Declare that all the active data streams have no change-points and stop the procedure.

In the following theorem, it is shown that the FDR of the S-MAP procedure is controlled to remain under the prespecified upper bound.

Theorem 4. *For upper bound constraint $\alpha \in (0,1)$ and deadline $N_{\max}$, the S-MAP procedure satisfies*

$$\mathrm{FDR} \leq \alpha.$$

Proof. The number of change-points declared, R, is known given the filtration of all the data up to the deadline, $\mathcal{F}_{N_{\max}}$. Thus, using the law of total expectation, we can rewrite the FDR from (4.2) as

$$\mathrm{FDR} = \mathbb{E}_\mu \left[\frac{\mathbb{E}_\mu[V|\mathcal{F}_{N_{\max}}]}{R \vee 1} \right]. \tag{4.8}$$

Recall that V is the number of false discoveries, i.e. the size of the subset of $[K]$ s.t. $T^{(k)} < t^{(k)}$ and $T^{(k)} < N_{\max}$. Thus, V can be written as

$$\begin{aligned} V &= \sum_{k=1}^{K} \mathbf{1}_{\{T^{(k)} < t^{(k)}\} \cap \{T^{(k)} < N_{\max}\}} \\ &= \sum_{k=1}^{K} \mathbf{1}_{\{T^{(k)} < t^{(k)}\}} \mathbf{1}_{\{T^{(k)} < N_{\max}\}} \\ &= \sum_{k:T^{(k)} < N_{\max}} \mathbf{1}_{\{T^{(k)} < t^{(k)}\}}, \end{aligned} \tag{4.9}$$

By substituting the last row of (4.9) in the term $\mathbb{E}_\mu[V|\mathcal{F}_{N_{\max}}]$ and using the linearity of the expectation operator and the definitions of V and R, we obtain

$$\begin{aligned} \mathbb{E}_\mu[V|\mathcal{F}_{N_{\max}}] &= \sum_{k:T^{(k)} < N_{\max}} \mathbb{E}_\mu(\mathbf{1}_{T^{(k)} < t^{(k)}}|\mathcal{F}_{N_{\max}}) \\ &= \sum_{k:T^{(k)} < N_{\max}} \mathbb{P}_\mu(T^{(k)} < t^{(k)}|\mathcal{F}_{T^{(k)}}) \\ &= \sum_{k:T^{(k)} < N_{\max}} (1 - \pi^{(k)}_{T^{(k)}}) \leq R(1 - Q_K) = R\alpha. \end{aligned} \tag{4.10}$$

The second equality is due to the fact that the stopping times, $\{T^{(k)}\}_{k\in[K]}$, are known given $\mathcal{F}_{N_{\max}}$, since we stop sampling the data stream k at $T^{(k)}$, and since the change-points are statistically independent. The inequality in (4.10) is obtained using the definition of R and since at time slot $T^{(k)}$, satisfying $T^{(k)} < N_{\max}$, the event $\{\pi^{(k)}_{T^{(k)}} \geq Q_K = 1 - \alpha\}$ occurs. This is because Q_K is the smallest threshold from (4.6). By substituting the last term in (4.10) into (4.8), one obtains

$$\mathrm{FDR} \leq \mathbb{E}_\mu \left[\frac{R\alpha}{R \vee 1} \right] \leq \alpha. \tag{4.11}$$

□

It should be noted that Theorem 4 is valid for $N_{\max} = \infty$ and also for any sampling strategy and not just for the maximum a-posteriori probability (MAP) sampling. For $q = 1$, the proof of Theorem 4 can be viewed as an alternative method to show the FDR control of the MD-FDR procedure from [13, Theorem 1]. Moreover, Theorem 4 extends the result from [13, Theorem 1] since it is shown that for $N_{\max} < \infty$ the FDR is controlled under level α without the addition of an asymptotically vanishing term. It can be observed from (4.10)-(4.11) that the thresholds from (4.6) may be too restrictive (too high) and lower thresholds are sufficient to guarantee FDR control. In the following section, we propose an alternative detection procedure that is less restrictive than the S-MAP procedure in terms of FDR control.

4.3 Improved S-MAP procedure

In this section, the Improved S-MAP (IS-MAP) detection procedure is proposed. It is similar to the S-MAP procedure except that its threshold values are uniformly lower than the thresholds of the S-MAP procedure. Since the IS-MAP procedure uses lower threshold values, then for a fixed proportion, q, the ADD and ANO will decrease compared to the S-MAP procedure, i.e. the ADD and ANO performance will improve. Moreover, using the lower thresholds, we prove that we can still control the FDR under the desired level, α. In the IS-MAP procedure, we choose a single threshold,

$$Q = 1 - \alpha, \tag{4.12}$$

for all the data streams and follow the same steps as the S-MAP procedure. This setup is different than detecting each change-point separately, as it will be shown that the FDR is controlled and not just the individual false alarm probability for each change-point detection. Since the thresholds of the IS-MAP procedure are all equal to Q, its jth sampling stage can be written in a more compact form than the corresponding sampling stage of the S-MAP procedure. Recall that I_j stands for the set of indices of active data streams at the beginning of the jth sampling stage and n_j stands for the time slot at the end of the jth sampling stage. The jth stage of sampling in the IS-MAP procedure is described as follows:

1. Sample the $\lceil q|I_j| \rceil$ data streams with highest posterior probabilities.
2. Update the posterior probabilities of the sensors with active data streams using (4.1).
3. Repeat this process until time slot n_j in which at least one of the posterior probabilities is higher than the threshold Q or in case the deadline, $N_{\max}$, is reached, i.e. $n_j = N_{\max} \wedge \min\{n > n_{j-1} : \exists k \in I_j, \pi_n^{(k)} \geq Q\}$.

(a) If $n_j < N_{\max}$: Declare change-points for all the data streams with indices in I_j whose posterior probabilities are higher than or equal to Q and remove these data streams from the set of active data streams. Update I_{j+1} to be the set of indices of the remaining active data streams. Stop the procedure if $|I_{j+1}| = 0$.

(b) Otherwise $n_j = N_{\max}$: Declare that all the still active data streams have no change-points and stop the procedure.

In the following theorem, it is shown that the FDR of the IS-MAP procedure satisfies the desired upper bound constraint.

Theorem 5. *For upper bound constraint $\alpha \in (0, 1)$ and deadline, N_{max}, the IS-MAP procedure satisfies*

$$\text{FDR} \leq \alpha.$$

Proof. The proof is similar to the proof of Theorem 4 except that the thresholds from (4.6) are replaced by the single threshold from (4.12). □

Theorem 5 is valid for $N_{\max} = \infty$ and also for any sampling strategy, similarly to Theorem 4. In the following section, we analyze the ADD and ANO of the S-MAP and the IS-MAP procedures in the asymptotic regime.

4.4 Performance of the S-MAP and the IS-MAP procedures

In this section, asymptotic lower and upper bounds on the ADD of the S-MAP and the IS-MAP procedures are derived as $\alpha \to 0$ for a fixed number of data streams K. Then, we characterize the behavior of these bounds as $K \to \infty$. Furthermore, the asymptotic ANO behavior of the procedures is studied. Since the analysis is asymptotic, we remove the deadline assumption in the S-MAP and the IS-MAP procedures. However, the results are informative and useful also for sufficiently high value of the deadline $N_{\max}$ for deciding about the changes, as will be demonstrated in the simulations. For simplicity of the analysis and in order to gain more insights, it is assumed that the change-points are i.i.d. and that the data streams have the same pre-change and post-change probability densities, f_0 and f_1, respectively. In addition, it is assumed that the prior distribution of each change-point obeys a geometric distribution with common parameter $\rho \in (0, 1)$. The results can be extended to non-identical change-point distributions and different pre/post-change distributions of the data streams.

4.4.1 ADD analysis of the S-MAP and the IS-MAP procedures

Under communication limitations, the FC observes a subsequence of the complete observation sequence from each sensor. According to the MAP approach, the time slots during which a data stream is sampled are stochastic. Whether a given sensor is sampled is determined online based on the sampling proportion, q, and the posterior probability values of the active sensors at each time slot. Therefore, it is difficult to characterize the subsequence of observations acquired from each sensor, especially when the number of data streams is large. Without such characterization it is not feasible to accurately analyze the ADD of the S-MAP and the IS-MAP procedures. However, asymptotic bounds on the ADD of these procedures can be derived. For the derivations, single change-point detection with a geometric prior as described in Section 2.1.1 is considered first. Thus, we consider an observation sequence $\{X_n\}_{n=1}^{\infty}$ with geometrically distributed change-point t and stopping rule of the form

$$T = \inf\{n \in \mathbb{N} : \pi_n \geq 1 - \eta\}, \eta \in (0, 1). \tag{4.13}$$

It is assumed that only a *subsequence* of the complete observation sequence is obtained. It is shown in [24] that for any subsequence of observations, the ADD of the stopping rule in the form of (4.13) as $\eta \to 0$ satisfies

$$\text{ADD} \geq \frac{|\log \eta|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\eta(1)) \tag{4.14}$$

and

$$\text{ADD} \leq \frac{|\log \eta|}{|\log(1-\rho)|}(1 + o_\eta(1)). \tag{4.15}$$

The asymptotic ADD lower bound from (4.14) is the same as the one in (2.8), and is attained when the complete observation sequence is available. The asymptotic ADD upper bound from (4.15) is attained when we do not take observations at all, and the stopping rule is based only on the prior.

In the following theorem, using (4.14) and (4.15) we derive asymptotic lower and upper bounds on the ADDs of the S-MAP and the IS-MAP procedures as $\alpha \to 0$. These ADD bounds do not require any assumptions on the subsequence of observations obtained from each sensor.

Theorem 6. *For $\alpha \to 0$ and any proportion of observed sensors, q, the following bounds are obtained*

$$ADD_{S\text{-}MAP} \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)), \tag{4.16}$$

$$ADD_{\text{S-MAP}} \leq \frac{\log K - \frac{1}{K}\log K! + |\log \alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)), \tag{4.17}$$

$$ADD_{\text{IS-MAP}} \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)), \tag{4.18}$$

and

$$ADD_{\text{IS-MAP}} \leq \frac{|\log \alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)). \tag{4.19}$$

Proof. The proof is given in Appendix A. □

For any fixed proportion, q, of observed data streams and for sufficiently small $\alpha \neq 0$ the bounds from (4.16)-(4.19) hold. The behavior of these bounds as K increases towards ∞ is characterized in order to investigate the scalability of the S-MAP and the IS-MAP procedures, as the number of data streams increases. Let

$$\text{ADD}_{\text{LB}} := \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}$$

denote the asymptotic ADD lower bound (LB) for both the S-MAP and the IS-MAP procedures. It can be seen that this lower bound is a finite constant w.r.t. K.

The asymptotic ADD upper bound (UB) for the S-MAP procedure is denoted by

$$\text{ADD}_{\text{S-MAP,UB}} := \frac{\log K - \frac{1}{K}\log K! + |\log \alpha|}{|\log(1-\rho)|}. \tag{4.20}$$

Consider the sequence $\{\log K - \frac{1}{K}\log K!\}_{K=1}^{\infty}$. Using [59, Eq. (5)] and Stirling's approximation (see e.g. [59, 56]) and applying some algebraic manipulations, it can be verified that this sequence is monotonically increasing and converges to 1. Thus, it is obtained that $\text{ADD}_{\text{S-MAP,UB}}$ is monotonically increasing with K and converges to a finite constant, i.e.

$$\lim_{K\to\infty} \text{ADD}_{\text{S-MAP,UB}} = \frac{1 + |\log \alpha|}{|\log(1-\rho)|}. \tag{4.21}$$

In a similar manner to (4.20), the asymptotic upper bound for IS-MAP is defined as

$$\text{ADD}_{\text{IS-MAP,UB}} := \frac{|\log \alpha|}{|\log(1-\rho)|} \tag{4.22}$$

The upper bound in (4.22) is a finite constant w.r.t. K.

The sequence $\{\log K - \frac{1}{K}\log K!\}_{K=1}^{\infty}$ is nonnegative and thus,

$$\text{ADD}_{\text{IS-MAP,UB}} \leq \text{ADD}_{\text{S-MAP,UB}}. \tag{4.23}$$

In addition, by comparing (4.22) to (4.20) and using (4.21), it is obtained that

$$\lim_{K\to\infty} \frac{\mathrm{ADD}_{\text{IS-MAP,UB}}}{\mathrm{ADD}_{\text{S-MAP,UB}}} = \frac{|\log\alpha|}{1+|\log\alpha|} < 1. \tag{4.24}$$

The results in (4.23) and (4.24) demonstrate the asymptotic ADD improvement obtained by using the IS-MAP procedure instead of the S-MAP procedure.

4.4.2 Detailed ADD analysis of the IS-MAP procedure

In this subsection, the focus is on the IS-MAP procedure in order to obtain a deeper asymptotic analysis of its ADD. Similar results for the S-MAP procedure can be derived in the same way. First, a single stopping rule, T, from (4.13) is considered. For the ADD of T, a tighter upper bound than (4.15) can be derived under some assumptions on the subsequence of observations obtained for the detection. Let us denote by $\{X_{V_n}\}_{n=1}^{\infty}$ the subsequence of the complete observation sequence, where $V_0 := 0$ and $V_1, V_2, \ldots$ are the discrete time slots in which observations are acquired for the detection of the single change-point, t, using the stopping rule, T. Equivalently, the complete observation sequence is sampled with intervals

$$\zeta_n := V_n - V_{n-1} \geq 1, n \in \mathbb{N}. \tag{4.25}$$

In addition, we define

$$\zeta^{(N)} := \frac{1}{N}\sum_{n=1}^{N} \zeta_n = \frac{V_N}{N}, \tag{4.26}$$

which is the average length of intervals in which we sample N observations from the observation sequence, the stopping rule,

$$\Gamma := \inf\{n \in \mathbb{N} : \pi_{V_n} \geq 1-\eta\}, \tag{4.27}$$

and the random change-point,

$$\gamma := \inf\{n \in \mathbb{N} : V_n \geq t\}. \tag{4.28}$$

The stopping rule and change-point from (4.27) and (4.28), respectively, represent the case in which we only count time slots where observations are obtained. In this analysis, the time slots $V_1, V_2, ...$ are considered to be deterministic. For the derivation of a tighter asymptotic upper bound on the ADD of the stopping rule, T, the only assumptions made are that the intervals are bounded, i.e. there exists $1 \leq \mathcal{B} < \infty$ s.t.

$$\zeta_n \leq \mathcal{B}, \ \forall n \in \mathbb{N}, \tag{4.29}$$

there exists $\zeta \in [1, \mathcal{B}]$ s.t.

$$\lim_{N\to\infty} \zeta^{(N)} = \zeta, \tag{4.30}$$

and

$$\mathbb{E}_\mu[\zeta^{(\Gamma)}(0 \vee (\Gamma - \gamma))] = \zeta \mathbb{E}_\mu[0 \vee (\Gamma - \gamma)](1 + o_\eta(1)). \tag{4.31}$$

From (4.25)-(4.26), $\zeta^{(\Gamma)} = \frac{V_\Gamma}{\Gamma}$, $\zeta^{(\gamma)} = \frac{V_\gamma}{\gamma}$, $\zeta_\gamma = V_\gamma - V_{\gamma-1}$. The specific value of ζ may be unknown. The assumption in (4.31) essentially requires that $\Gamma \to \infty$ as $\eta \to 0$. In the following proposition, we derive an asymptotic ADD upper bound for the stopping rule, T, which is tighter than (4.15).

Proposition 2. *Assume that* (4.29)-(4.31) *are satisfied. Then, as* $\eta \to 0$ *the ADD of the stopping rule* T *from* (4.13) *satisfies*

$$ADD \leq \frac{|\log \eta|}{\frac{1}{\zeta} D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\eta(1)). \tag{4.32}$$

Proof. The proof is given in Appendix B. □

It should be noted that a special case of (4.32) with $\zeta_n = \zeta < \infty, n \in \mathbb{N}$, was proved in [24].

Assume that $q > 0$ and that each stopping rule in the IS-MAP procedure satisfies the ADD upper bound in (4.32) with $\zeta = g_k < \infty$, i.e.

$$\mathrm{ADD}_{\text{IS-MAP},k} \leq \frac{|\log \alpha|}{\frac{1}{g_k} D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)), \tag{4.33}$$

$\forall k \in [K]$. The term g_k is the limiting average of time intervals between each sample of the kth data stream. In addition, assume that

$$\sup_{k\in[K]} g_k < \infty. \tag{4.34}$$

The assumptions in (4.33)-(4.34) are reasonable, because according to the MAP approach the subset of active data streams with highest posterior probabilities is sampled. For any $\alpha > 0$, as long as a data stream is not sampled its posterior probability is monotonically increasing in accordance with the recursive posterior update (4.1). Thus, after a finite number of time slots this data stream will have a sufficiently high posterior probability s.t. it will be sampled, i.e. this posterior probability will be among the highest posterior probabilities within the allowed proportion. Alternatively, this data stream posterior probability will cross the threshold $Q = 1 - \alpha$ and a change-point will be declared. Thus,

in a similar manner to the derivation of the upper bound in (4.19), a tighter asymptotic ADD upper bound is obtained for the IS-MAP procedure, given by

$$\mathrm{ADD}_{\text{IS-MAP}} \leq \frac{|\log\alpha|}{\frac{1}{\sup\limits_{k\in[K]} g_k} D(f_1||f_0) + |\log(1-\rho)|}(1+o_\alpha(1)).$$

In order to shed some light on the IS-MAP ADD dependence on the proportion q, a simple multiple change-point detection procedure that will be used as a benchmark is considered. In the following, it is assumed for simplicity that $\frac{1}{q}, Kq \in \mathbb{N}$. In the simple procedure, the set of data streams is divided into $\frac{1}{q}$ subsets. The mth subset includes Kq data streams with indices $(m-1)Kq+1, (m-1)Kq+2, \ldots, mKq$ and these data streams are sampled periodically at time slots $m, m+\frac{1}{q}, m+\frac{2}{q}, \ldots$, as long as they are active, for each $m = 1, \ldots, \frac{1}{q}$. For each data stream the stopping rule is in the posterior probability threshold form of (4.13) with threshold $Q = 1-\alpha$ as in (4.12). Before any change-points are declared, or equivalently as $\alpha \to 0$, the simple procedure satisfies the same communication constraint as the IS-MAP procedure. According to Proposition 2 and the ADD definition from (4.4), it is obtained that as $\alpha \to 0$, the ADD of the simple procedure satisfies

$$\mathrm{ADD}_{\text{simple}} \leq \frac{|\log\alpha|}{qD(f_1||f_0) + |\log(1-\rho)|}(1+o_\alpha(1)). \tag{4.35}$$

It can be seen that as q increases the ADD upper bound from (4.35) decreases.

For a fixed proportion of the active sensors sampled at each time slot q, it is expected that the IS-MAP procedure will outperform the simple procedure in terms of ADD. The reason is described as follows. In order to detect changes quickly, it is only important for the procedure to monitor a particular data stream after a change has occurred in it. How frequently a data stream is sampled pre-change has no influence on the detection delay, since the change cannot be observed from the pre-change observations. Therefore, an "oracle" procedure with access to unobservable information should always sample the active sensors for which the change-point has already occurred but has not yet been detected (or at least as many of such streams that fit under the communication constraint). Since we of course don't know whether a change has taken place in a given sensor or not, in IS-MAP the posterior probabilities are used as a proxy for that information. By choosing to monitor the sensors with the highest posterior probabilities, IS-MAP aims to reduce the times that a data stream in pre-change state is monitored in favor of a stream in post-change state. In contrast, the simple procedure that uses a predetermined sampling schedule makes no effort in trying to sample the data streams that are in post-change state. In Section 4.5, it is demonstrated in simulations that the IS-MAP procedure has

lower ADD than the simple procedure. Moreover, the ADD of the IS-MAP procedure is compared to the asymptotic upper bound in (4.35). It will be shown that this upper bound holds and is useful in describing the performance of the IS-MAP procedure.

4.4.3 ANO analysis of the IS-MAP procedure

A low proportion q of active sensors monitored by the FC may result in high ADD. However, an advantage of monitoring only a small subset of sensors is that the ANO, or equivalently the communication burden, for the detection task may decrease. In the following, the asymptotic ANO behavior of the IS-MAP procedure is studied. Since the imposed deadline for decision making was removed, the ANO from (4.5) can be rewritten as

$$\begin{aligned} \mathrm{ANO} &= \frac{1}{K}\mathbb{E}_\mu\Bigg[\sum_{n=1}^{T^{(k,\inf)}} \lceil qK_n \rceil \Bigg] + \frac{1}{K}\mathbb{E}_\mu\Bigg[\sum_{n=T^{(k,\inf)}+1}^{T^{(k,\sup)}} \lceil qK_n \rceil \Bigg] \\ &= q\mathbb{E}_\mu[T^{(k,\inf)}] + \frac{1}{K}\mathbb{E}_\mu\Bigg[\sum_{n=T^{(k,\inf)}+1}^{T^{(k,\sup)}} \lceil qK_n \rceil \Bigg], \end{aligned} \tag{4.36}$$

where $T^{(k,\inf)} := \inf_{k\in[K]} T^{(k)}$. The second equality in (4.36) is obtained since up to $T^{(k,\inf)}$ the number of active data streams satisfies $K_n = K$ and under the assumption that $Kq \in \mathbb{N}$. It is assumed that

$$\frac{1}{K}\mathbb{E}_\mu\Bigg[\sum_{n=T^{(k,\inf)}+1}^{T^{(k,\sup)}} \lceil qK_n \rceil \Bigg] = o_\alpha(|\log\alpha|). \tag{4.37}$$

This assumption essentially requires that as $\alpha \to 0$ the difference between the smallest and largest stopping times of the IS-MAP procedure will remain finite or increase slower than $|\log\alpha|$. In the following proposition, bounds on the asymptotic ANO of the IS-MAP procedure are derived.

Proposition 3. *Assume that $Kq \in \mathbb{N}$ and (4.37) are satisfied. Then, as $\alpha \to 0$ the ANO of the IS-MAP procedure satisfies*

$$\mathrm{ANO}_{\textit{IS-MAP}} \geq \frac{q|\log\alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)) \tag{4.38}$$

and

$$\mathrm{ANO}_{\textit{IS-MAP}} \leq \frac{q|\log\alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)). \tag{4.39}$$

Proof. By Theorem 6, as $\alpha \to 0$ and for any proportion, q, the ADD of the kth stopping rule of the IS-MAP procedure satisfies

$$\mathrm{ADD}_{\text{IS-MAP},k} \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)) \tag{4.40}$$

and

$$\mathrm{ADD}_{\text{IS-MAP},k} \leq \frac{|\log \alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)), \tag{4.41}$$

$\forall k \in [K]$. Using (4.3), it can be verified that

$$\mathbb{E}_\mu[T^{(k)}_{\text{IS-MAP}}] = \mathrm{ADD}_{\text{IS-MAP},k}(1 + o_\alpha(1)),\ \forall k \in [K]. \tag{4.42}$$

By substituting (4.42) in (4.40) and (4.41), one obtains

$$\mathbb{E}_\mu[T^{(k)}_{\text{IS-MAP}}] \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)) \tag{4.43}$$

and

$$\mathbb{E}_\mu[T^{(k)}_{\text{IS-MAP}}] \leq \frac{|\log \alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)), \tag{4.44}$$

respectively for all $k \in [K]$. Since (4.43)-(4.44) hold for all $k \in [K]$, one obtains

$$\mathbb{E}_\mu[T^{(k,\inf)}_{\text{IS-MAP}}] \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)) \tag{4.45}$$

and

$$\mathbb{E}_\mu[T^{(k,\inf)}_{\text{IS-MAP}}] \leq \frac{|\log \alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)). \tag{4.46}$$

From (4.36), (4.37), and (4.45), one obtains (4.38). Similarly, from (4.36), (4.37), and (4.46), one obtains (4.39). □

It can be seen from (4.38)-(4.39) that the asymptotic lower and upper bounds on the ANO of the IS-MAP procedure are linearly increasing with q. Thus, from (4.35) and (4.38)-(4.39) it can be seen that there is a tradeoff between the ADD and the ANO that depends on the chosen proportion, q. The ADD-ANO tradeoff will be investigated in more detail in Section 4.5. Finally, it can be seen that the asymptotic ANO bounds from (4.38)-(4.39) are independent of the number of data streams, K. Therefore, the IS-MAP procedure is scalable in the sense that its asymptotic ANO does not increase without bound with the number of data streams. The ANO of the S-MAP procedure is higher than the ANO of the IS-MAP procedure. However, similarly to Proposition 3, it can be shown that the S-MAP procedure is also scalable in terms of ANO as K increases.

4.5 Performance evaluation via simulations

In this section, the performance of the proposed S-MAP and IS-MAP procedures is evaluated in terms of FDR, ADD, and ANO. In addition, the analytical results from Sections 4.2-4.4 are verified in the simulations. The simulation results are based on 1,000 Monte Carlo runs. The deadline for the detection procedures is $N_{\max}$ =10,000. In all of the following simulations, all the finite change-points were eventually detected and there were no missed detections, where missed detection happens if a stopping time is set to be ∞, while the change-point is finite.

For comparison purposes, the fully parallel procedure from [12], named D-FDR, that observes all the data streams all the time is implemented. The FDR control of the D-FDR procedure is established in [12]. In this procedure, the following test statistic is used

$$G_n^{(k)} = \sum_{m=1}^{\infty} P_\mu(t^{(k)} = m) \prod_{i=m}^{n} \frac{f_1^{(k)}(X_i^{(k)})}{f_0^{(k)}(X_i^{(k)})},\ n \in \mathbb{N}, \tag{4.47}$$

where $G_0^{(k)} := 1$. This test statistic is the average likelihood ratio (ALR) between the hypotheses that the change occurs at $t^{(k)} = m < \infty$ and that no change takes place, $t^{(k)} = \infty$. A recursive method for computing the ALR from (4.47) can be found in [12]. For $q = 1$, the D-FDR procedure is similar to the S-MAP procedure except that it uses the ALR test statistic, rather than the posterior probability test statistic. The thresholds are set to

$$Q_r = \frac{K}{r\alpha},\ r \in [K],$$

in order to guarantee the same false positive constraints as in (4.7). Assume that for the kth data stream, the corresponding threshold is $Q_{r_k} = \frac{K}{r_k\alpha}$, $r_k \in [K]$. It is shown in [69] that in this case, using the ALR test statistic with the threshold Q_{r_k} is equivalent to using the posterior probability test statistic with the threshold

$$Q_{r_k}^* = 1 - p(t^{(k)} \geq n+1)\frac{r_k\alpha}{K}. \tag{4.48}$$

Thus, from (4.6), (4.12), and (4.48), the posterior probability thresholds of the D-FDR procedure are higher than the posterior probability thresholds of the S-MAP and the IS-MAP procedures. Consequently for $q = 1$, the ADD and ANO of the S-MAP and the IS-MAP procedures will be lower than those obtained by the D-FDR procedure.

In addition to D-FDR, three other procedures for are implemented and evaluated for comparison purposes. The procedures control the FDR under the tolerated level in accordance with the proof of Theorem 5, which is valid for any sampling strategy. The first procedure is the "simple procedure" described in Subsection 4.4.2. In this procedure,

given the proportion q the data streams are sampled periodically with a fixed period $\frac{1}{q}$. The second procedure is a "random sampling procedure". This procedure is an alternative for the IS-MAP procedure from Section 4.3, where the method of choosing the subset of sensors to monitor is different. In the random sampling procedure, at each time slot a subset of sensors to monitor is chosen randomly with uniform probability within the allowed proportion. The third procedure is a hybrid of the IS-MAP and the random sampling procedures. In the hybrid procedure, the subset of active data streams sampled is chosen according to the MAP sampling or according to the random sampling with equal probability 0.5 at each time. The random sampling and hybrid procedures are implemented in order to verify that the MAP approach for choosing the subset of sensors to monitor, as used in the IS-MAP procedure, improves the ADD performance compared to randomly choosing this subset.

4.5.1 Gaussian distribution scenario

Gaussian pre- and post-change distributions with a change in the mean are considered, so that $f_0 = \mathcal{N}(0,1)$ and $f_1 = \mathcal{N}(1,1)$. The true change-points are generated independently for each sensor from a geometric distribution with parameter $\rho = 0.01$ and it is assumed that this parameter is known when applying the procedure. The FDR upper bound is set at $\alpha = 0.1$. For $K = 50, 100, 200, 400, 800$, the FDR control of the proposed S-MAP and IS-MAP procedures is investigated with sampling proportions $q = 0.5, 1$. The proportion $q = 1$ corresponds to the parallel versions of the S-MAP and the IS-MAP procedures that observe all of the active data streams at each time slot. The resulting maximum observed average FDP values (over different K) for each procedure are .031 for S-MAP $q = 1$ and $q = 0.5$, and .060 for IS-MAP $q = 0.5$ and $q = 1$. Consequently, the considered procedures control the FDR under the upper bound $\alpha = 0.1$, corroborating the analytical results. The S-MAP FDR values are lower than the IS-MAP FDR values, since the S-MAP procedure is more conservative and uses higher thresholds than the IS-MAP procedure. For both the S-MAP and the IS-MAP procedures there is still a gap between the FDR values and the upper bound $\alpha = 0.1$. This result follows from the choices of thresholds in (4.6) and (4.12) for the S-MAP and the IS-MAP procedures, respectively, that neglect the overshoot in the stopping rule, mentioned on page 12.

In the left panel of Fig. 4.1, the ADD for the D-FDR, S-MAP, simple, random sampling, hybrid, and IS-MAP procedures are evaluated with with different values of q as a function of K. It can be seen that the S-MAP and IS-MAP procedures have an approximately constant ADD as K increases, which verifies the analytical results in Section 4.4. The parallel version of the IS-MAP procedure, i.e. for $q = 1$, has the lowest ADD. Moreover, it can be seen that the IS-MAP procedure with $q = 0.5$ outperforms the parallel version of the S-MAP procedure and the D-FDR procedure. These results demonstrate

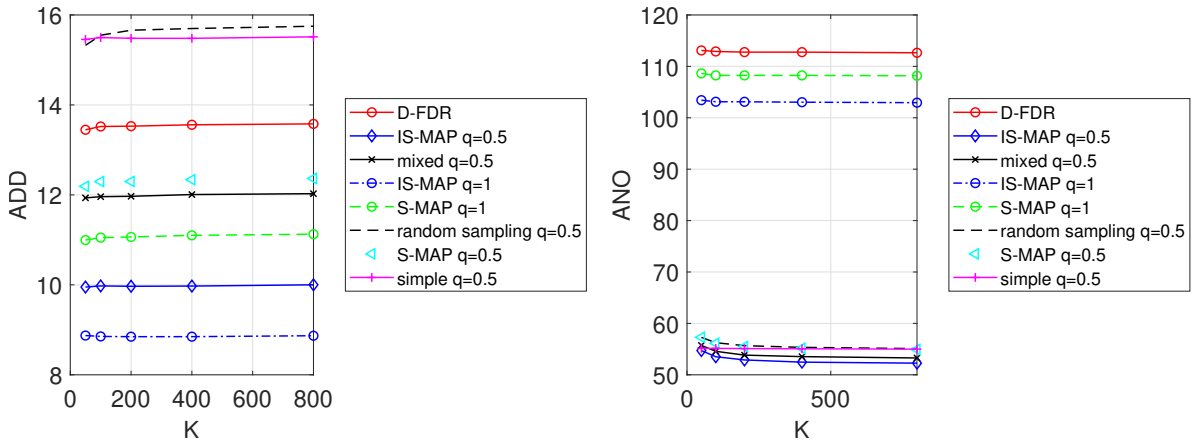

Figure 4.1: Left panel: ADD evaluated as a function of the number of sensors K for different methods. For S-MAP and IS-MAP, the delay is approximately constant as a function of K. IS-MAP achieves smallest ADD compared to alternatives. Right panel: ANO for the different methods as a function of K. For all approaches, the ANO is approximately constant in K, but depends heavily on q.

the advantage of using the IS-MAP procedure instead of the S-MAP or the D-FDR procedures in terms of ADD. The simple and random sampling procedures with $q = 0.5$ have the highest ADDs implying that the MAP approach provides better performance than periodic sampling or random sampling of the subset of active data streams. The hybrid procedure with $q = 0.5$ has higher ADD than IS-MAP procedure with $q = 0.5$. This result shows that combining the MAP sampling with random sampling does not reduce the detection delay.

In the right panel of Fig. 4.1, the ANO versus K for the same procedures is evaluated. It can be seen that the IS-MAP procedure with $q = 0.5$ has the lowest and the D-FDR has the highest ANO. In addition, it can be seen that for all the considered procedures, the ANO is approximately a constant w.r.t. K. However, the ANOs of parallel procedures that monitor all the active data streams are significantly higher than the ANOs of communication-limited procedures with $q = 0.5$.

In the left panel of Fig. 4.2, the ADDs of the S-MAP, hybrid, and IS-MAP procedures for $K = 300$ are plotted versus the proportion value q. It can be seen that for any of the considered proportions, the IS-MAP procedure achieves the lowest ADD. For all the procedures the ADD monotonically decreases as the proportion, q, increases. As $q \to 1$, the mixed procedure ADD approaches the ADD of the IS-MAP procedures since for $q = 1$ the procedures coincide. However, for $q < 1$ the IS-MAP procedure outperforms the mixed procedure, which demonstrates the advantage of MAP sampling compared to random sampling. In the right panel, the ANOs of the procedures are plotted against

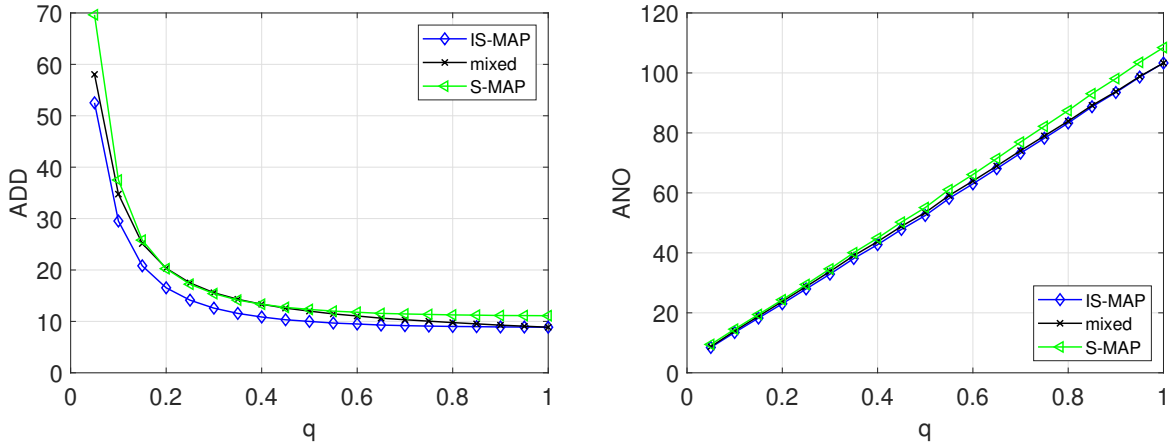

Figure 4.2: Left panel: ADD evaluated as a function of q for IS-MAP, S-MAP and the hybrid procedures. IS-MAP provides the smallest detection delay for all q. Right panel: ANO evaluated as a function of K for the same procdures. ANO increases linearly in q, as was analyzed in Subsection 4.4.3.

q. The IS-MAP procedure achieves the lowest ANO. For all the procedures the ANO increases approximately linearly as the proportion, q, increases. This result validates the ANO analysis in Subsection 4.4.3 for the IS-MAP procedure.

In the left panel of Fig. 4.3, the asymptotic ADD of the IS-MAP is examined by plotting it against small values of α. The ADD of the IS-MAP procedure is compared to the asymptotic LB and asymptotic UB from (4.18) and (4.19), respectively, for different q. In addition, the approximate upper bound on the ADD of the IS-MAP procedure from (4.35) is evaluated. The bound depends on the proportion parameter q and decreases as q increases. It is seen that the ADD of the IS-MAP procedure increases as $|\log \alpha|$ increases, which shows the tradeoff between achieving low ADD and requiring strict upper bound on the FDR. The asymptotic ADD bounds from (4.18)-(4.19) hold for any of the considered values of $|\log \alpha|$ and q. The approximate ADD UB from (4.35) holds for $q = 0.25, 0.5$ and for any of the considered values of $|\log \alpha|$. For sufficiently high value of $|\log \alpha|$ the approximate UB for $q = 0.75$ holds as well. This figure shows that for sufficiently low values of α the asymptotic ADD bounds from (4.18)-(4.19) and the approximate ADD UB from (4.35) are informative for performance analysis of the IS-MAP procedure. In the right panel, we examine the ANO of the IS-MAP procedure versus $|\log \alpha|$ for small values of the FDR upper bound α. The ANO of the IS-MAP procedure is compared to the asymptotic LB and asymptotic UB from (4.38) and (4.39), respectively, for $q = 0.25, 0.5, 0.75$. It can be seen that these asymptotic bounds hold for any of the considered values of $|\log \alpha|$ and q. The ANO of the IS-MAP procedure increases as $|\log \alpha|$ increases, which shows that requiring a strict upper bound on the FDR implies higher ANO for the multiple

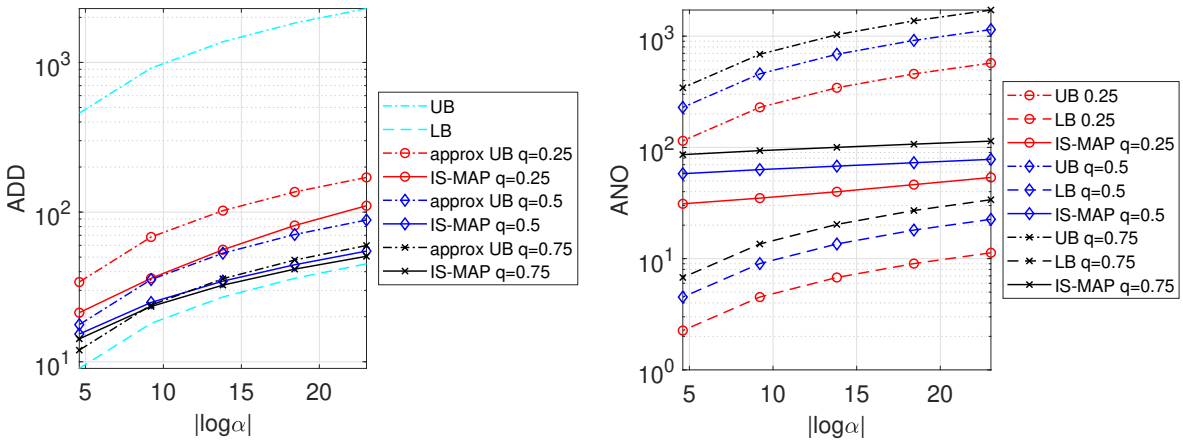

Figure 4.3: Left panel: ADD of the IS-MAP for different q as a function of $|\log \alpha|$. The dashed lines denote the ADD upper bounds derived in 4.4.2. For all q, and α sufficiently small, the observed ADD falls below the corresponding upper bound, confirming the results. Right panel: ANO for different q as a function of $|\log \alpha$. The observed ANO lies between the asymptotic ANO bounds of eqs. (4.38) and (4.39) for all q.

change-point detection task.

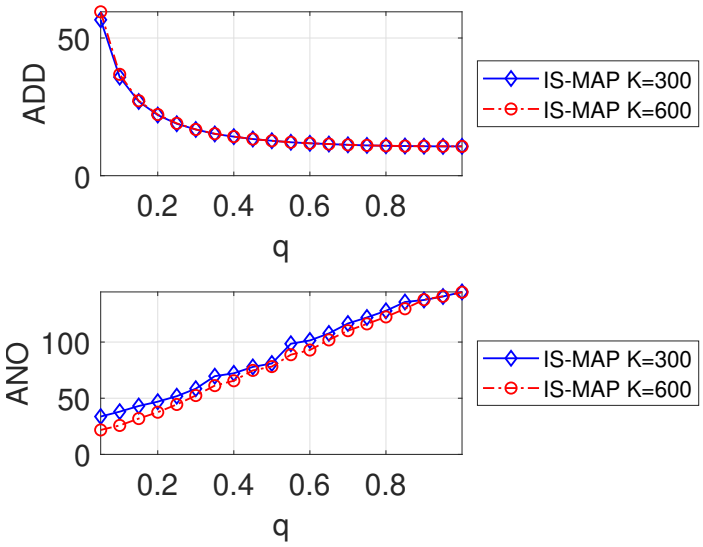

Figure 4.4: Upper: ADD of the IS-MAP for $K = 300, 600$, when the the prior distributions of the change-points and the post-change distributions vary from sensor to sensor. The ADD decreases in q, and for these system and is not affected by the number of sensors in the network K. Lower: ANO of the IS-MAP as a function of q in this setup. ANO increases approximately linearly in q.

Finally, we consider a scenario with different prior distributions of the change-points and different post-change distributions of the data streams. In the kth data stream, there is no change-point with probability p_∞ and with probability $1-p_\infty$ the kth change-point prior distribution is geometric with parameter ρ_k. All the data streams are assumed to obey $f_0 = \mathcal{N}(0,1)$ distribution before the change and $f_1^{(k)} = \mathcal{N}(\mu_k, 1)$ distribution after the change, where μ_k is the mean of the kth data stream after the change. In this case, the posterior probability of each data stream is calculated according to the standard recursive update (4.1), where

$$\rho_n^{(k)} = P_\mu(t^{(k)} = n | t^{(k)} \geq n) = \rho_k \frac{(1-p_\infty)(1-\rho_k)^{n-1}}{p_\infty + (1-p_\infty)(1-\rho_k)^{n-1}},\ k \in [K], n \in \mathbb{N}.$$

We set $p_\infty = 0.01$ and for the K data streams we set $\rho_k = 0.01, b_k = 2$ for $k = 1, \ldots, \frac{K}{2}$ and $\rho_k = 0.05, b_k = 1$ for $k = \frac{K}{2}+1, \ldots, K$. Under the FDR upper bound constraint, $\alpha = 0.1$, the FDR of the IS-MAP procedure for $K = 300$ and $K = 600$ versus the proportion values $\{q = 0.05m\}_{m=1}^{20}$ is evaluated. The resulting minimum and maximum estimated FDR values are .045 and 0.047, respectively. Thus, the IS-MAP procedure controls the FDR under the tolerated level in accordance with Theorem 5 in Section 4.3. In the left plot of Fig. 4.4, the ADD of the IS-MAP procedure for $K = 300$ and $K = 600$ versus the sampling proportion q is shown. It can be seen that the ADD of the IS-MAP procedure decreases as q increases and that for these system parameters the ADD does not change significantly between $K = 300$ and $K = 600$. In the right plot of Fig. 4.4, the ANO of the IS-MAP procedure for $K = 300$ and $K = 600$ versus the sampling proportion q is displayed. It can be seen that the ANO of the IS-MAP procedure increases as q increases. The increase is approximately linear and there is no significant difference between the ANOs for $K = 300$ and $K = 600$. The results in Fig. 4.4 show that the ADD and ANO behavior of the IS-MAP procedure validate the analysis in Section 4.4 despite the different prior distributions of the change-points and different post-change distributions of the data streams.

4.6 Conclusion

In this chapter, which follows the publication [42] co-authored by the author of this thesis, methods for Bayesian multiple change-point detection in a sensor network with communication constraints are developed. The S-MAP detection procedure was proposed. In S-MAP, sensors with the highest posterior probabilities of change-points having occurred are chosen for sampling at each time step. In addition, an improved procedure named IS-MAP was proposed. IS-MAP requires lower stopping thresholds than the S-MAP procedure and hence attains lower ADD and ANO. It was proven that both proposed procedures

control the FDR at a predefined level and achieve ADD and ANO that asymptotically remain constant as the number of sensors in the network increases. The dependence of the IS-MAP procedure ADD and ANO on the chosen proportion of sensors monitored was characterized in the asymptotic regime.

In the simulations, Gaussian distributed observations with known mean and variance. The change was associated with a rapid change in the mean. We compared the proposed S-MAP and IS-MAP procedures to procedures that use random and periodic sampling of the active data streams. In all the simulations, the IS-MAP procedure achieved the best performance in terms of ADD and ANO. These results show the advantage of using low detection thresholds together with the MAP sampling approach in multiple change-point detection.

Topics for future research include, among other things, the derivation of novel procedures with FDR control capabilities for multiple change-point detection under non-parametric models [31], in case spatial information is available [76, 53], and in case each data stream can have multiple change-points [23, 57].

Chapter 5

Conclusion

In this thesis, theory and methods for sequential and multiple statistical inference and detection were studied. The fundamentals of sequential change-point detection were covered in Chapter 2. Both the Bayesian and minimax formulations were introduced. Additionally, some methods for data-efficient change-point detection were presented. Chapter 3 introduced the fundamentals of multiple hypothesis testing. Moreover, a Bayesian method for multiple testing in spatial domains was presented. In Chapter 4 the ideas of data-efficient change-point detection and multiple hypothesis testing were fused. A novel method that performs data-efficient sequential change-point detection in multiple parallel data streams was presented and analyzed. The method guarantees FDR control, scales well as the number of data streams increases, and provides competitive detection delay.

Appendix A

Proof of Theorem 6

In this appendix, asymptotic lower and upper bounds on the ADD of the S-MAP and the IS-MAP procedures are derived. For any data stream, the lowest possible threshold of the S-MAP procedure from (4.6) is $Q_K = 1 - \alpha$, i.e. change-point cannot be declared before the posterior probability is higher than or equal to $1 - \alpha$. Thus, from (4.14)

$$\mathrm{ADD}_{\text{S-MAP},k} \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)), \tag{A.1}$$

$\forall k \in [K]$. It can be seen that the asymptotic lower bound in (A.1) is independent of k. Thus, by substituting (A.1) in (4.4), we obtain (4.16).

According to the S-MAP procedure we can find a threshold for the kth data stream, $Q_{r_k} = 1 - \frac{r_k \alpha}{K},\ r_k \in [K]$, which is different from the thresholds of the other data streams. For this threshold, the change of the kth data stream is declared at the first time slot in which this threshold is exceeded or even before this threshold is exceeded. Thus, from (4.15),

$$\mathrm{ADD}_{\text{S-MAP},k} \leq \frac{\log \frac{K}{r_k \alpha}}{|\log(1-\rho)|}(1 + o_\alpha(1)), \forall k \in [K]. \tag{A.2}$$

By substituting (A.2) in (4.4), we obtain

$$\mathrm{ADD}_{\text{S-MAP}} \leq \left(\frac{1}{K} \sum_{k=1}^{K} \frac{\log \frac{K}{r_k \alpha}}{|\log(1-\rho)|} \right) (1 + o_\alpha(1)). \tag{A.3}$$

Since the thresholds are different, we obtain

$$\sum_{k=1}^{K} \log r_k = \sum_{k=1}^{K} \log k = \log K!. \tag{A.4}$$

By substituting (A.4) into (A.3) and reordering, (4.17) is obtained.

In the IS-MAP procedure, for any data stream the threshold is $Q = 1 - \alpha$ from (4.12). Thus, using (4.14) and (4.15), one obtains

$$\mathrm{ADD}_{\text{IS-MAP},k} \geq \frac{|\log \alpha|}{D(f_1||f_0) + |\log(1-\rho)|}(1 + o_\alpha(1)) \tag{A.5}$$

and

$$\mathrm{ADD}_{\text{IS-MAP},k} \leq \frac{|\log \alpha|}{|\log(1-\rho)|}(1 + o_\alpha(1)), \tag{A.6}$$

respectively, $\forall k \in [K]$. The asymptotic lower and upper bounds in (A.5) and (A.6), respectively, are independent of k and thus, by substituting (A.5) and (A.6) in (4.4), we obtain (4.18) and (4.19), respectively.

Appendix B

Proof of Proposition 2

In this appendix, the asymptotic ADD upper bound from (4.32) on page 53 is derived under the assumption that (4.29)-(4.31) are satisfied. Using the definition of γ from (4.28), it can be seen that the prior distribution of $\gamma \in \mathbb{N}$ is

$$\begin{aligned}\mathbb{P}_\mu(\gamma = m) &= \mathbb{P}_\mu(V_{m-1} < t \leq V_m) \\ &= \mathbb{P}_\mu(t \leq V_m) - \mathbb{P}_\mu(t \leq V_{m-1}).\end{aligned} \tag{B.1}$$

Under the geometric prior assumption on t we obtain

$$\mathbb{P}_\mu(t \leq m) = 1 - (1-\rho)^m, m \in \mathbb{N}. \tag{B.2}$$

By substituting (B.2) in (B.1), one obtains

$$\mathbb{P}_\mu(\gamma = m) = (1-\rho)^{V_{m-1}} - (1-\rho)^{V_m}. \tag{B.3}$$

Using (B.3), we obtain

$$\begin{aligned}\lim_{m\to\infty} \frac{-\log \mathbb{P}_\mu(\gamma \geq m+1)}{m} &= \lim_{m\to\infty} \frac{-\log((1-\rho)^{V_m})}{m} \\ &= \left(\lim_{m\to\infty} \frac{V_m}{m}\right) |\log(1-\rho)| \\ &= \zeta |\log(1-\rho)|,\end{aligned} \tag{B.4}$$

where the third equality is obtained by substituting (4.26) and (4.30) into the second equality. Using the definition of γ from (4.28), we obtain that on $\{\gamma = n\}$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=n}^{n+N-1} \log L(X_{V_i}) = D(f_1||f_0) \tag{B.5}$$

almost surely. From the definitions of the stopping rule, Γ, and the change-point, γ, in (4.27) and (4.28), respectively, and from (B.4) and (B.5), it can be seen that the detection of γ using Γ based on the sequence $\{X_{V_n}\}_{n=1}^{\infty}$ is a Bayesian change-point detection procedure that satisfies the conditions of Theorem 3 in [70]. Thus, using this theorem, we obtain the following asymptotic upper bound on the ADD of Γ:

$$\mathbb{E}_{\mu}[0 \vee (\Gamma - \gamma)] \leq \frac{|\log \eta|}{D(f_1||f_0) + \zeta|\log(1-\rho)|}(1 + o_{\eta}(1)). \tag{B.6}$$

Next, the stopping rule

$$T^* = \inf\{V_n, n \in \mathbb{N} : \pi_{V_n} \geq 1 - \eta\} = V_{\Gamma}. \tag{B.7}$$

is considered. In a similar manner to T, the stopping rule T^* uses the same posterior update, but can only take values from the subsequence $\{V_n\}_{n=1}^{\infty}$ rather than $\mathbb{N}$. Therefore, $T \leq T^*$ and consequently

$$T - t \leq T^* - t = V_{\Gamma} - V_{\gamma} + V_{\gamma} - t, \tag{B.8}$$

where the equality follows from (B.7). From (4.25) and (4.28) we obtain that

$$V_{\gamma} - t \leq \zeta_{\gamma} - 1. \tag{B.9}$$

In addition, using (4.26) we can write

$$V_{\Gamma} = \Gamma\zeta^{(\Gamma)} \text{ and } V_{\gamma} = \gamma\zeta^{(\gamma)}. \tag{B.10}$$

By substituting (B.9)-(B.10) into the right hand side of (B.8), one obtains

$$\begin{aligned} T - t &\leq \zeta^{(\Gamma)}(\Gamma - \gamma) + \gamma(\zeta^{(\Gamma)} - \zeta^{(\gamma)}) + \zeta_{\gamma} - 1 \\ &\leq \zeta^{(\Gamma)}(\Gamma - \gamma) + |\gamma(\zeta^{(\Gamma)} - \zeta^{(\gamma)}) + \zeta_{\gamma} - 1|. \end{aligned} \tag{B.11}$$

Using (4.25) and (4.29), we obtain

$$1 \leq \zeta^{(N)} \leq \mathcal{B}, \ \forall N \in \mathbb{N}. \tag{B.12}$$

Substituting (4.29) and (B.12) in (B.11), one obtains

$$T - t \leq \zeta^{(\Gamma)}(\Gamma - \gamma) + \gamma(\mathcal{B} - 1) + \mathcal{B} - 1. \tag{B.13}$$

From (B.3), it can be verified that

$$\mathbb{E}_{\mu}[\gamma] \leq \frac{1}{\rho}. \tag{B.14}$$

By using (4.31), (B.6), (B.13) and (B.14), we obtain that the ADD of T satisfies

$$\text{ADD} \leq \frac{\zeta|\log \eta|}{D(f_1||f_0) + \zeta|\log(1-\rho)|}(1 + o_{\eta}(1))$$

and consequently (4.32) is obtained.

Bibliography

[1] T. Banerjee and V. V. Veeravalli. Data-efficient quickest change detection with on–off observation control. *Sequential Analysis*, 31(1):40–77, 2012.

[2] T. Banerjee and V. V. Veeravalli. Data-efficient quickest change detection in minimax settings. *IEEE Trans. Inf. Theory*, 59(10):6917–6931, Oct. 2013.

[3] T. Banerjee and V. V. Veeravalli. Data-efficient minimax quickest change detection with composite post-change distribution. *IEEE Trans. Inf. Theory*, 61(9):5172–5184, Sep. 2015.

[4] R.F. Barber and Ramdas A. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1247–1268, 2017.

[5] Y. Benjamini and R. Heller. False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[7] Ezio Biglieri, Andrea J Goldsmith, Larry J Greenstein, H Vincent Poor, and Narayan B Mandayam. *Principles of cognitive radio*. Cambridge University Press, 2013.

[8] A. A. Borovkov. Asymptotically optimal solutions in the change-point problem. *Theory of Probability & Its Applications*, 43(4):539–561, 1999.

[9] S. Brooks, A. Gelman, G. Jones, and X. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[10] B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[11] J. Chen, W. Zhang, and H. V. Poor. Non-Bayesian multiple change-point detection controlling false discovery rate. In *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, pages 31–35, July 2016.

[12] J. Chen, W. Zhang, and H. V. Poor. On parallel sequential change detection controlling false discovery rate. In *Proc. of the 50th Asilomar Conference on Signals, Systems and Computers*, pages 107–111, Nov. 2016.

[13] J. Chen, W. Zhang, and H. V. Poor. A false discovery rate oriented approach to parallel sequential change detection problems. *IEEE Trans. Signal Process.*, 68:1823–1836, 2020.

[14] A. Chouldechova. *False discovery rate control for spatial data*. PhD thesis, 2014.

[15] B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[16] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1. Cambridge University Press, 2012.

[17] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[18] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[19] Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[20] Daniel Egea-Roca, Gonzalo Seco-Granados, and José A López-Salcedo. Comprehensive overview of quickest detection theory and its application to gnss threat detection. *Gyroscopy and Navigation*, 8(1):1–14, 2017.

[21] Serguei Foss, Dmitry Korshunov, and Stanley Zachary. *An introduction to heavy-tailed and subexponential distributions*. Springer Series in Operations Research and Financial Engineering. Springer, 2011.

[22] G. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, pages 1–8, 2018.

[23] Axel Gandy and F Din-Houn Lau. Non-restarting cumulative sum charts and control of the false discovery rate. *Biometrika*, 100(1):261–268, 2013.

[24] J. Geng, E. Bayraktar, and L. Lai. Bayesian quickest change-point detection with sampling right constraints. *IEEE Trans. Inf. Theory*, 60(10):6474–6490, Oct. 2014.

[25] J. Geng and L. Lai. Non-Bayesian quickest change detection with stochastic sample right constraints. *IEEE Trans. Signal Process.*, 61(20):5090–5102, Oct. 2013.

[26] J. Geng and L. Lai. Quickest change-point detection over multiple data streams via sequential observations. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4404–4408, Apr. 2018.

[27] T. Halme, M. Golz, and V. Koivunen. Bayesian multiple hypothesis testing for distributed detection in sensor networks. In *Proc. of the IEEE Data Science Workshop (DSW)*, pages 105–109, June 2019.

[28] T. Halme, E. Nitzan, H. V. Poor, and V. Koivunen. Bayesian multiple change-point detection with limited communication. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5490–5494, May 2020.

[29] ArunKumar Jayaprakasam and Vinod Sharma. Cooperative robust sequential detection algorithms for spectrum sensing in cognitive radio. In *2009 International Conference on Ultra Modern Telecommunications & Workshops*, pages 1–8. IEEE, 2009.

[30] L. Lai, Y. Fan, and H. V. Poor. Quickest detection in cognitive radio: A sequential change detection framework. In *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–5, 2008.

[31] T. S. Lau, W. P. Tay, and V. V. Veeravalli. A binning approach to quickest change detection with unknown post-change distribution. *IEEE Trans. Signal Process.*, 67(3):609–621, Feb. 2019.

[32] A. Li and R.F. Barber. Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74, 2019.

[33] Husheng Li, Huaiyu Dai, and Chengzhi Li. Collaborative quickest spectrum sensing via random broadcast in cognitive radio systems. *IEEE Transactions on Wireless Communications*, 9(7):2338–2348, 2010.

[34] G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

[35] Jarmo Lundén, Visa Koivunen, Anu Huttunen, and H Vincent Poor. Collaborative cyclostationary spectrum sensing for cognitive radio systems. *IEEE Transactions on Signal Processing*, 57(11):4182–4195, 2009.

[36] Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, and John Anderson. Wireless sensor networks for habitat monitoring. In *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, WSNA '02. Association for Computing Machinery, 2002.

[37] M. Marcus and P. Swerling. Sequential detection in radar with multiple resolution elements. *IRE Transactions on Information Theory*, 8(3):237–245, 1962.

[38] Yajun Mei. Quickest detection in censoring sensor networks. pages 2148–2152, 2011.

[39] G. V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986.

[40] M.A. Newton. On a nonparametric recursive estimator of the mixing distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 306–322, 2002.

[41] Igor V Nikiforov and Ivan N Tikhonov. Application of change detection theory to seismic signal processing. In *Detection of Abrupt Changes in Signals and Dynamical Systems*, pages 355–373. Springer, 1985.

[42] E. Nitzan, T. Halme, and V. Koivunen. Bayesian methods for multiple change-point detection with reduced communication. *IEEE Transactions on Signal Processing*, 68:4871–4886, 2020.

[43] E. Nitzan, T. Halme, H. V. Poor, and V. Koivunen. Deterministic multiple change-point detection with limited communication. In *Proc. of the 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, Mar. 2020.

[44] R. Nowak and U. Mitra. Boundary estimation in sensor networks: Theory and methods. In *Information processing in sensor networks*, pages 80–95. Springer, 2003.

[45] T. Oskiper and H.V. Poor. Online activity detection in a multiuser environment using the matrix cusum algorithm. *IEEE Transactions on Information Theory*, 48(2):477–493, 2002.

[46] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[47] M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004.

[48] M. Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1):206–227, 1985.

[49] Moshe Pollak and Alexander G. Tartakovsky. Optimality properties of the shiryaev-roberts procedure. *Statistica Sinica*, 19(4):1729–1739, 2009.

[50] H. V. Poor and O. Hadjiliadis. *Quickest Detection*. Cambridge University Press, 2008.

[51] S. Pounds and S.W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.

[52] K. Premkumar and A. Kumar. Optimal sleep-wake scheduling for quickest intrusion detection using wireless sensor networks. In *Proc. of the 27th IEEE Conference on Computer Communications (INFOCOM)*, pages 1400–1408, Apr. 2008.

[53] Vasanthan Raghavan and Venugopal V Veeravalli. Quickest change detection of a Markov process across a sensor array. *IEEE Trans. Inf. Theory*, 56(4):1961–1981, 2010.

[54] Aaditya K Ramdas, Rina F Barber, Martin J Wainwright, Michael I Jordan, et al. A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics*, 47(5):2790–2821, 2019.

[55] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[56] Herbert Robbins. A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29, 1955.

[57] G. Rovatsos, G. V. Moustakides, and V. V. Veeravalli. Quickest detection of a dynamic anomaly in a sensor network. In *Proc. of the 53rd Asilomar Conference on Signals, Systems and Computers*, pages 98–102, 2019.

[58] H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[59] József Sándor and Lokenath Debnath. On certain inequalities involving the constant e and their applications. *Journal of Mathematical Analysis and Applications*, 249(2):569–582, 2000.

[60] Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

[61] Albert N Shiryaev. Quickest detection problems in the technical analysis of the financial data. pages 487–521. Springer, 2002.

[62] H. Shu, B. Nan, and R. Koeppe. Multiple testing for neuroimaging via hidden markov random field. *Biometrics*, 71(3):741–750, 2015.

[63] D.O. Siegmund, N.R. Zhang, and B. Yakir. False discovery rate for scanning statistics. *Biometrika*, 98(4):979–985, 2011.

[64] D. Simpson, F. Lindgren, and H. Rue. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23(1):65–74, 2012.

[65] Ross S. Sparks, Tim Keighley, and David Muscatello. Early warning cusum plans for surveillance of negative binomial daily disease counts. *Journal of Applied Statistics*, 37(11):1911–1929, 2010.

[66] John D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013 – 2035, 2003.

[67] W. Sun, B. Reich, T.T. Cai, M. Guindani, and A. Schwartzman. False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):59–83, 2015.

[68] W. Tansey, O. Koyejo, R.A. Poldrack, and J.G. Scott. False discovery rate smoothing. *Journal of the American Statistical Association*, 113(523):1156–1171, 2018.

[69] A. G. Tartakovsky. Asymptotic optimality in Bayesian changepoint detection problems under global false alarm probability constraint. *Theory of Probability & Its Applications*, 53(3):443–466, 2009.

[70] A. G. Tartakovsky and V. V. Veeravalli. General asymptotic Bayesian theory of quickest change detection. *Theory of Probability & Its Applications*, 49(3):458–497, 2005.

[71] Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

[72] Alexander G Tartakovsky, Moshe Pollak, and Aleksey S Polunchenko. Third-order asymptotic optimality of the generalized shiryaev–roberts changepoint detection procedures. *Theory of Probability & Its Applications*, 56(3):457–484, 2012.

[73] Alexander G. Tartakovsky, Boris L. Rozovskii, Rudolf B. Blazek, and Hongjoong Kim. Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, 3(3):252–293, 2006.

[74] John Wilder Tukey. *The Problem of Multiple Comparisons: Introduction and Parts A, B, and C*.

[75] Venugopal V Veeravalli and Taposh Banerjee. Quickest change detection. 3:209–255, 2014.

[76] Andi Wang, Xiaochen Xian, Fugee Tsung, and Kaibo Liu. A spatial-adaptive sampling procedure for online monitoring of big data streams. *Journal of Quality Technology*, 50(4):329–343, 2018.

[77] C. Zhang, J. Fan, and T. Yu. Multiple testing via fdr l for large-scale imaging data. *Ann. Statist.*, 39(1):613–642, 02 2011.

[78] H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.