Robust Quickest Change Detection for Unnormalized Models

Suya Wu[1] **Enmao Diao**[1] **Taposh Banerjee**[2] **Jie Ding**[3] **Vahid Tarokh**[1]

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
[2]Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15213, USA
[3]School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA

Abstract

Detecting an abrupt and persistent change in the underlying distribution of online data streams is an important problem in many applications. This paper proposes a new robust score-based algorithm called RSCUSUM, which can be applied to unnormalized models and addresses the issue of unknown post-change distributions. RSCUSUM replaces the Kullback-Leibler divergence with the Fisher divergence between pre- and post-change distributions for computational efficiency in unnormalized statistical models and introduces a notion of the "least favorable" distribution for robust change detection. The algorithm and its theoretical analysis are demonstrated through simulation studies.

1 INTRODUCTION

In the problem of quickest change detection, the objective is to detect an abrupt change in the statistical properties of an observed stochastic process. This change in the distribution has to be detected with the minimum possible delay, subject to a constraint on the rate of false alarms. This problem has applications in sensor networks, cyber-physical systems, biology, and neuroscience; see Veeravalli and Banerjee [2014], Basseville et al. [1993], Poor and Hadjiliadis [2008], Tartakovsky et al. [2014].

When the pre- and post-change distribution of the data is known, a typical optimal algorithm in the literature is a stopping rule. A sequence of statistics is calculated using the likelihood ratio of the observations, and a change is declared when the sequence of statistics crosses a pre-designed threshold. The threshold is chosen to meet a constraint on false alarms; see Shiryaev [1963], Lorden [1971], Pollak [1985], Moustakides [1986], Lai [1998], Tartakovsky and Veeravalli [2005]. The three most important algorithms in the literature are the Shiryaev algorithm (Shiryaev [1963], Tartakovsky and Veeravalli [2005]), the cumulative sum (CUSUM) algorithm (Page [1955], Lorden [1971], Moustakides [1986], Lai [1998]), and the Shiryaev-Roberts algorithm (Roberts [1966], Pollak [1985]).

The main challenge in implementing a change detection algorithm in practice is that the pre- and post-change distributions are not precisely known. This challenge is amplified when the data is high-dimensional. Specifically, in several machine learning applications, the data models may not lend themselves to explicit distributions. For example, energy-based models (LeCun et al. [2006]) capture dependencies between observed and latent variables based on their associated energy (an unnormalized probability), and score-based deep generative models Song et al. [2020] generate high-quality images by learning the score function (the gradient of the log density function). These models can be computationally cumbersome to normalize themselves as probabilistic density functions. Thus, optimal algorithms from the change detection literature, which are likelihood ratio-based tests, are computationally expensive to implement.

This issue is partially addressed in Wu et al. [2023] where the authors have proposed the SCUSUM algorithm, a Hyvärinen score-based (Hyvärinen [2005]) modification of the CUSUM algorithm for quickest change detection. It is shown in Wu et al. [2023] that the SCUSUM algorithm is consistent and the authors also provide expressions for the average detection delay and the mean time to a false alarm. The Hyvärinen score is invariant to scale and hence can be applied to unnormalized models. This makes the SCUSUM algorithm highly efficient as compared to the classical CUSUM algorithm for high-dimensional models.

The main drawback of the SCUSUM algorithm is that its effectiveness is contingent on knowing the precise post-change unnormalized model, i.e., knowing the post-change model within a normalizing constant. In practice, due to a limited amount of training data, the post-change model

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

can only be learned within an uncertainty class. To detect the change effectively, an algorithm must be robust against these modeling uncertainties. The SCUSUM algorithm is not robust in this sense. Specifically, if not carefully designed, the SCUSUM algorithm can fail to detect several (in fact, infinitely many) post-change scenarios.

In this paper, we propose a robust score-based variant of the CUSUM algorithm for the quickest change detection. We refer to our algorithm as the RSCUSUM algorithm. Under the assumption that the post-change uncertainty class is convex and compact, we show that the RSCUSUM algorithm is robust, i.e., can consistently detect changes for every possible post-change model. This consistency is achieved by designing the RSCUSUM algorithm using the *least favorable* distribution from the post-change class.

The problem of optimal robust quickest change detection is studied in Unnikrishnan et al. [2011]. In a minimax setting, the optimal algorithm is the CUSUM algorithm designed using the least favorable distribution. The robust CUSUM test in Unnikrishnan et al. [2011] may suffer from two drawbacks: 1) It is a likelihood ratio-based test and hence may not be amenable to implementation in high-dimensional models. 2) The notion of least favorable distribution is defined using *stochastic boundedness*, which may be difficult to verify for high-dimensional data.

In contrast with the work in Unnikrishnan et al. [2011], we define the notion of least favorable distribution using Fisher divergence and provide a method to effectively identify the least favorable distribution for the post-change model.

1.1 OUR CONTRIBUTIONS

We now summarize our contributions in this paper.

• We propose a new robust score-based quickest change detection algorithm that can be applied to unnormalized models, namely, statistical models whose density involves an unknown normalizing constant. Specifically, we use the Hyvärinen score (Hyvärinen [2005]) to propose a robust score-based variant of the SCUSUM algorithm from Wu et al. [2023], which we refer to as RSCUSUM. In this variant and its subsequent theory, the role of Kullback-Leibler divergence in classical change detection is replaced with the Fisher divergence between the pre-and post-change distributions. Please see Section 3.

• Our developed RSCUSUM algorithm can address unknown post-change models. Specifically, assuming that the post-change law belongs to a known family of distributions that is convex and compact, we identify a least favorable distribution that is closest in terms of Fisher divergence from the pre-change family. We then show that the RSCUSUM algorithm can consistently detect each post-change distribution from the family, and is robust in this sense. Please see Section 4.

• We provide an effective method to identify the least favorable post-change distribution in a post-change family. This is in contrast to the setup in Unnikrishnan et al. [2011] where a stochastic boundedness characterization makes it harder to identify the least favorable distribution. Please see Section 5.

• From a theoretical perspective, unlike the CUSUM algorithm that leverages the fact that the likelihood ratios form a martingale under the pre-change model Lai [1998], Woodroofe [1982], the RSCUSUM algorithm is a score-based algorithm where cumulative scores do not enjoy a standard martingale characterization. Our analysis of the delay and false alarm analysis for RSCUSUM is based on new analysis techniques. Pleas see Section 4.

• We demonstrate the effectiveness of the RSCUSUM algorithm through simulation studies on Gaussian and Gauss-Bernoulli Restricted Boltzmann Machine (RBM) models. Please see Section 6.

2 PROBLEM FORMULATION

Let $\{X_n\}_{n\geq 1}$ denote a sequence of independent random variables defined on the probability space $(\Omega, \mathcal{F}, P_\nu)$. Let $\mathcal{F}_n$ be the $\sigma-$algebra generated by random variables $X_1,\ X_2,\ \ldots,\ X_n$, and let $\mathcal{F} = \sigma(\cup_{n\geq 1}\mathcal{F}_n)$ be the $\sigma-$algebra generated by the union of sub-σ-algebras. Under P_ν, $X_1,\ X_2,\ \ldots,\ X_{\nu-1}$ are i.i.d. according to a density p_∞ and $X_\nu,\ X_{\nu+1},\ \ldots$ are i.i.d. according to a density p_1. We think of ν as the change point, p_∞ as the pre-change density, and p_1 as the post-change density. We use $\mathbb{E}_\nu$ and Var_ν to denote the expectation and the variance associated with the measure P_ν, respectively. Thus, ν is seen as an unknown constant and we have an entire family $\{P_\nu\}_{1\leq\nu\leq\infty}$ of change-point models, one for each possible change point. We use P_∞ to denote the measure under which there is no change, with $\mathbb{E}_\infty$ denoting the corresponding expectation.

A change detection algorithm is a stopping time T with respect to the data stream $\{X_n\}_{n\geq 1}$:

$$\{T \leq n\} \in \mathcal{F}_n.$$

If $T \geq \nu$, we have made a *delayed detection*; otherwise, a *false alarm* has happened. Our goal is to find a stopping time T to optimize the trade-off between well-defined metrics on delay and false alarm. We consider two minimax problem formulations to find the best stopping rule.

To measure the detection performance of a stopping rule, we use the following minimax metric (Lorden [1971]), the worst-case averaged detection delay (WADD):

$$\mathcal{L}_{\texttt{WADD}}(T) \triangleq \sup_{\nu\geq 1} \text{ess sup}\, \mathbb{E}_\nu[(T-\nu+1)^+|\mathcal{F}_\nu],$$

where $(y)^+ \triangleq \max(y, 0)$ for any $y \in \mathbb{R}$. Here $\operatorname{ess\,sup}$ is the essential supremum, i.e., the supremum outside a set of measure zero. We also consider the version of minimax metric introduced in Pollak [1985], the worst conditional averaged detection delay (CADD):

$$\mathcal{L}_{\texttt{CADD}}(T) \triangleq \sup_{\nu \geq 1} \mathbb{E}_\nu[T - \nu | T \geq \nu].$$

For false alarms, we consider the *average running length* (ARL), which is defined as the mean time to false alarm:

$$\text{ARL} \triangleq \mathbb{E}_\infty[T].$$

We now formulate a robust quickest change detection problem; see Unnikrishnan et al. [2011]. We assume that pre- and post-change distributions are not precisely known. However, each is known within an uncertainty class:

$$P_\infty \in \mathcal{G}_\infty$$
$$P_1 \in \mathcal{G}_1.$$

For simplicity, in this paper, we will assume that the pre-change class is a singleton:

$$\mathcal{G}_\infty = \{P_\infty\}.$$

Our proposed method can also be extended to the case of composite $\mathcal{G}_\infty$. The objective is to find a stopping rule to solve the following problem:

$$\min_T \sup_{P_1 \in \mathcal{G}_1} \mathcal{L}_{\texttt{WADD}}(T) \quad \text{subject to} \quad \mathbb{E}_\infty[T] \geq \gamma, \quad (1)$$

where γ is a constraint on the ARL. The delay $\mathcal{L}_{\texttt{WADD}}$ in the above problem is a function of the true post-change law P_1 and should be designated as $\mathcal{L}^{P_1}_{\texttt{WADD}}$. We will, however, suppress this notation and simply refer to $\mathcal{L}^{P_1}_{\texttt{WADD}}$ by $\mathcal{L}_{\texttt{WADD}}$. Thus, the goal in this problem is to find a stopping time T to minimize the worst-case detection delay, subject to a constraint γ on $\mathbb{E}_\infty[T]$.

We are also interested in the version with the minimax metric introduced in Pollak [1985]:

$$\min_T \sup_{P_1 \in \mathcal{G}_1} \mathcal{L}_{\texttt{CADD}}(T) \quad \text{subject to} \quad \mathbb{E}_\infty[T] \geq \gamma. \quad (2)$$

If the post-change family is also singleton, $\mathcal{G}_1 = \{P_1\}$, then the above formulations are the classical minimax formulations from the quickest change detection literature; see Veeravalli and Banerjee [2014], Tartakovsky et al. [2014], Poor and Hadjiliadis [2008]. The optimal algorithm (exactly optimal for (1) and asymptotically optimal for (2)) is the CUSUM algorithm given by

$$T_{\texttt{CUSUM}} = \inf\{n \geq 1 : \Lambda(n) \geq \tau\},$$

where $\Lambda(n)$ is defined using the recursion

$$\Lambda(0) = 0,$$
$$\Lambda(n) \triangleq \left(\Lambda(n-1) + \log \frac{p_1(X_n)}{p_\infty(X_n)}\right)^+, \forall n \geq 1, \quad (3)$$

which leads to a computationally convenient stopping scheme. We recall that here p_1 is the post-change density and p_∞ is the pre-change density.

In Lorden [1971] and Lai [1998], the asymptotic performance of the CUSUM algorithm is also characterized. Specifically, it is shown as $\gamma \to \infty$,

$$\mathcal{L}_{\texttt{WADD}}(T_{\texttt{CUSUM}}) \sim \mathcal{L}_{\texttt{CADD}}(T_{\texttt{CUSUM}}) \sim \frac{\log \gamma}{\mathbb{D}_{\texttt{KL}}(P_1 \| P_\infty)}.$$

Here $\mathbb{D}_{\texttt{KL}}(P_1 \| P_\infty)$ is the Kullback-Leibler divergence between the post-change distribution and pre-change distribution:

$$\mathbb{D}_{\texttt{KL}}(P_1 \| P_\infty) = \int p_1(x) \log \frac{p_1(x)}{p_\infty(x)} dx,$$

and the notation $g(c) \sim h(c)$ as $c \to c_0$ indicates that $\frac{g(c)}{h(c)} \to 1$ as $c \to c_0$ for any two functions $c \mapsto g(c)$ and $c \mapsto h(c)$.

Since the CUSUM algorithm uses likelihood ratio to compute its statistic, it is not amenable to implementation for high-dimensional models (see Wu et al. [2023]), where often the densities p_1 or p_∞ are only known within a normalizing constant.

3 ROBUST QUICKEST CHANGE DETECTION FOR UNNORMALIZED MODELS

In this section, we propose a robust score-based CUSUM (RSCUSUM) algorithm. We first review the SCUCUM algorithm proposed by Wu et al. [2023] to address the issues with likelihood ratio-based CUSUM for unnormalized models. The SCUSUM is defined based on Hyvärinen Score (Hyvärinen [2005]), which circumvents the computation issue of the normalization constant. Similar to the schemes of SCUSUM, we use the Hyvärinen score and propose a robust variant that releases the knowledge of the true post-change distribution, where we assume the true post-change distribution is unknown but its uncertainty class is known.

Recall from Section 2 that under the measure P_∞, there is no change, and the density for each random variable is p_∞. In the rest of the paper, we refer to the probability measure of X_1 under P_∞, also by P_∞. Similarly, we refer to the law of X_1 under P_1 also by P_1. The differences will always be clear from the context.

We provide the definition of the Hyvärinen Score below.

Definition 3.1 (Hyvärinen Score)**.** The Hyvärinen score of any measure P (with density p) is a mapping $(X, P) \mapsto \mathcal{S}_{\texttt{H}}(X, P)$ given by

$$\mathcal{S}_{\texttt{H}}(X, P) \triangleq \frac{1}{2} \|\nabla_X \log p(X)\|_2^2 + \Delta_X \log p(X)$$

whenever it can be well defined. Here, $\|\cdot\|_2$ denotes the Euclidean norm, ∇_X and $\Delta_X = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ respectively denote the gradient and the Laplacian operators acting on $X = (x_1, \cdots, x_d)^\top$.

By using the Hyvärinen Score in our algorithm, the role of Kullback-Leibler divergence in the theoretical analysis of the algorithm is replaced by the Fisher divergence.

Definition 3.2 (Fisher Divergence)**.** The Fisher divergence between two probability measures P to Q (with densities p and q) is defined by

$$\mathbb{D}_{\mathrm{F}}(P\|Q) \triangleq \mathbb{E}_{X\sim P}\left[\|\nabla_X \log p(X) - \nabla_X \log q(X)\|_2^2\right],$$

whenever the integral is well defined.

Clearly, $\nabla_X \log p(X)$, $\nabla_X \log q(X)$, and $\Delta_X \log q(X)$ remain invariant if p and q are scaled by any positive constant with respect to X. Hence, the Fisher divergence and the Hyvärinen Score remain *scale-variant* concerning an arbitrary constant scaling of density functions.

The SCUSUM [Wu et al., 2023] assumes that the true pre- and post-chagne distributions P_∞ and P_1 are known. It defines the detection score by

$$\tilde{z}_\lambda(X) \triangleq \lambda\big(\mathcal{S}_{\mathrm{H}}(X, P_\infty) - \mathcal{S}_{\mathrm{H}}(X, P_1)\big). \tag{4}$$

However, it is impractical, in particular for online data streams, to know the true post-change distribution. We assume that pre-change data is available. This data and a model class $\mathcal{G}_\infty$ are used to model/learn the pre-change distribution P_∞. The post-change distribution P_1 is assumed to be modeled by an unknown element of a parametric family $\mathcal{G}_1 = \{G_\theta : \theta \in \Theta_1\}$. We note that our framework readily extends to the case of non-parametric families but for simplicity, we present our results only in the parametric case.

We define the notion of least favorable distribution. This approach to defining the least favorable distribution for the quickest change detection is novel.

Definition 3.3 (Least Favorable Distribution (LFD))**.** Assume that the family $\mathcal{G}_1 = \{G_\theta : \theta \in \Theta_1\}$ is convex and compact. We define

$$Q_1 = \arg\min_{G_\theta \in \mathcal{G}_1} \mathbb{D}_F(G_\theta \| P_\infty). \tag{5}$$

The existence of Q_1 is guaranteed by the compactness of $\mathcal{G}_1$ and the continuity of the Fisher divergence as a function of its arguments. Thus, Q_1 is the closest element of $\mathcal{G}_1$ to P_∞ in the Fisher-divergence sense.

Given the pre-change law P_∞ (with density p_∞), we now use Q_1 and its density q_1 to design the RSCUSUM algorithm. We define the instantaneous RSCUSUM score function $X \mapsto z_\lambda(X)$ by

$$z_\lambda(X) \triangleq \lambda\big(\mathcal{S}_{\mathrm{H}}(X, P_\infty) - \mathcal{S}_{\mathrm{H}}(X, Q_1)\big), \tag{6}$$

where $\lambda > 0$ is a pre-selected multiplier, $\mathcal{S}_{\mathrm{H}}(X, P_\infty)$ and $\mathcal{S}_{\mathrm{H}}(X, Q_1)$ are respectively the Hyvärinen score functions of P_∞ and Q_1. If the post-change model is precisely known, then the Q_1 in the above equation will be replaced by the known post-change law and RSCUSUM is identical to SCUSUM [Wu et al., 2023]. In Section 4, we will provide more discussion on the role of λ in the RSCUSUM algorithm.

Our proposed stopping rule is given by

$$T_{\mathtt{RSCUSUM}} = \inf\{n \geq 1 : Z(n) \geq \tau\}, \tag{7}$$

where $\tau > 0$ is a stopping threshold that is pre-selected to control false alarms, and $Z(n)$ can be computed recursively:

$$Z(0) = 0,$$
$$Z(n) \triangleq (Z(n-1) + z_\lambda(X_n))^+, \ \forall n \geq 1.$$

The statistic $Z(n)$ is referred to as the detection score of RSCUSUM at time n. The RSCUSUM algorithm is summarized in Algorithm 1.

Algorithm 1: RSCUSUM Detection Algorithm

Input: Hyvarinen score functions $\mathcal{S}_{\mathrm{H}}(\cdot, P_\infty)$ and $\mathcal{S}_{\mathrm{H}}(\cdot, Q_1)$ of pre-change distribution and least favorable distribution in $\mathcal{G}_1$, respectively.
Data: m previous observations $\mathbf{X}_{[-m+1,0]}$ and the online data stream $\{X_n\}_{n\geq 1}$
Initialization:
  Current time $k = 0$, $\lambda > 0$, $\tau > 0$, and $Z(0) = 0$
while $Z(k) < \tau$ **do**
  $k = k + 1$
  Update $z_\lambda(X_k) = \lambda(\mathcal{S}_{\mathrm{H}}(X_k, P_\infty) - \mathcal{S}_{\mathrm{H}}(X_k, Q_1))$
  Update $Z(k) = \max(Z(k-1) + z_\lambda(X_k), 0)$
Record the current time k as the stopping time $T_{\mathtt{RSCUSUM}}$
Output: $T_{\mathtt{RSCUSUM}}$

4 DELAY AND FALSE ALARM ANALYSIS OF THE RSCUSUM ALGORITHM

In this section, we provide delay and false alarm analysis of the RSCUSUM algorithm. We introduce two assumptions: 1) $P_\infty \notin \mathcal{G}_1$, and 2) the same mild regularity conditions introduced in Hyvärinen [2005] so that the Hyvärinen score is well-defined.

We first prove an important lemma for our problem. If the Fisher divergence is seen as a measure of distance between two probability measures, then the following lemma provides a reverse triangle inequality for this distance, under

the mild assumption that the order of integrals and derivatives can be interchanged.

Lemma 4.1. *Let P_∞ be the pre-change distribution, $Q_1 \in \mathcal{G}_1$ be the least-favorable distribution (as defined in Equation 5), and $Q_2 \in \mathcal{G}_1$ be any other post-change distribution. Then*

$$\mathbb{D}_F\left(Q_1 \| P_\infty\right) \leq \mathbb{D}_F\left(Q_2 \| P_\infty\right) - \mathbb{D}_F\left(Q_2 \| Q_1\right).$$

Proof. Consider a convex set of densities

$$\left\{y \mapsto q_\xi(x) : q_\xi(x) = \xi q_1(x) + (1-\xi) q_2(x), \xi \in [0,1]\right\},$$

where q_1 and q_2 are densities of Q_1 and Q_2, respectively. Let Q_ξ denote the distribution characterized by density q_ξ. We note that $Q_\xi \in \mathcal{G}_1$ due to the convexity assumption on $\mathcal{G}_1$. We use $\mathcal{L}(\xi)$ to denote the Fisher divergence $\mathbb{D}_{\mathrm{F}}\left(Q_\xi \| P_\infty\right)$, and

$$\begin{aligned}\mathcal{L}(\xi) &= \int \left\|\nabla \log q_\xi - \nabla \log p_\infty\right\|^2 q_\xi dx \\ &= \int \left\|\nabla \log\big(\xi q_1 + (1-\xi) q_2\big) - \nabla \log p_\infty\right\|^2 \\ &\qquad \big(\xi q_1 + (1-\xi) q_2\big) dx.\end{aligned}$$

Clearly $\mathcal{L}(\xi)$ is minimized at $\xi = 1$, and $\frac{\partial \mathcal{L}(\xi)}{\partial \xi} \mid_{\xi = 1^-} \leq 0$. Let $\mathcal{L}'(\xi) = \frac{\partial \mathcal{L}(\xi)}{\partial \xi}$, we have

$$\begin{aligned}\mathcal{L}'(\xi) &= \int \big(q_1 - q_2\big) \left\|\nabla \log q_\xi - \nabla \log p_\infty\right\|^2 dx \\ &\quad + \int 2 q_\xi \nabla \left(\frac{q_1 - q_2}{q_\xi}\right)^T \big(\nabla \log q_\xi - \nabla \log p_\infty\big) dx.\end{aligned}$$

This implies

$$\begin{aligned}\mathcal{L}'(1^-) &= \int \big(q_1 - q_2\big) \left\|\nabla \log q_1 - \nabla \log p_\infty\right\|^2 dx \\ &\quad + \int 2 q_1 \nabla \left(\frac{q_1 - q_2}{q_1}\right)^T \big(\nabla \log q_1 - \nabla \log p_\infty\big) dx \\ &= \mathbb{D}_{\mathrm{F}}\left(Q_1 \| P_\infty\right) - \int \underbrace{q_2 \left\|\nabla \log q_1 - \nabla \log p_\infty\right\|^2}_{\text{term 1}} \\ &\qquad + \underbrace{2 q_1 \nabla \left(\frac{q_1 - q_2}{q_1}\right)^T \big(\nabla \log q_1 - \nabla \log p_\infty\big)}_{\text{term 2}} dx.\end{aligned} \tag{8}$$

For term 1, we have

$$\begin{aligned}&q_2 \left\|\nabla \log q_1 - \nabla \log p_\infty\right\|^2 \\ &= q_2 \left\|\nabla \log q_1 - \nabla \log q_2\right\|^2 + q_2 \left\|\nabla \log q_2 - \nabla \log p_\infty\right\|^2 \\ &\quad + \underbrace{2 q_2 \big(\nabla \log q_1 - \nabla \log q_2\big)^T \big(\nabla \log q_2 - \nabla \log p_\infty\big)}_{\text{term 1(a)}}.\end{aligned} \tag{9}$$

We note that,

$$\int_x q_2 \left\|\nabla \log q_1 - \nabla \log q_2\right\|^2 dx = \mathbb{D}_F(Q_2 \| Q_1), \tag{10}$$

$$\int_x q_2 \left\|\nabla \log q_2 - \nabla \log p_\infty\right\|^2 dx = \mathbb{D}_F(Q_2 \| P_\infty). \tag{11}$$

For term 2, we note that

$$\nabla \left(\frac{q_1 - q_2}{q_1}\right) = \frac{q_2}{q_1} \big(\nabla \log q_1 - \nabla \log q_2\big).$$

Therefore,

$$\begin{aligned}&2 q_1 \nabla \left(\frac{q_1 - q_2}{q_1}\right)^T \big(\nabla \log q_1 - \nabla \log p_\infty\big) \\ &= 2 q_2 \big(\nabla \log q_1 - \nabla \log q_2\big)^T \big(\nabla \log q_1 - \nabla \log p_\infty\big).\end{aligned} \tag{12}$$

Combining the last term in Equation (9) with Equation (12),

$$\begin{aligned}&- \text{term 1(a)} + \text{term 2} \\ &= 2 q_2 \big(\nabla \log q_1 - \nabla \log q_2\big)^T \\ &\qquad \big(\nabla \log q_1 - \nabla \log p_\infty - \nabla \log q_2 + \nabla \log p_\infty\big) \\ &= 2 q_2 \left\|\nabla \log q_1 - \nabla \log q_2\right\|^2.\end{aligned} \tag{13}$$

Plugging Equations (10), (11), and (13) into Equation (8),

$$\mathcal{L}'(1^-) = \mathbb{D}_{\mathrm{F}}\left(Q_1 \| P_\infty\right) + \mathbb{D}_{\mathrm{F}}\left(Q_2 \| Q_1\right) - \mathbb{D}_{\mathrm{F}}\left(Q_2 \| P_\infty\right).$$

The results follows since $\frac{\partial \mathcal{L}(\xi)}{\partial \xi} \mid_{\xi = 1^-} \leq 0$. □

We now use Lemma 4.1 to prove a result on the consistency of our proposed RSCUSUM algorithm.

Lemma 4.2 (Positive and Negative Drifts)**.** *Consider the instantaneous RSCUSUM score function $X \mapsto z_\lambda(X)$ as defined in Equation (6). Recall that $P_1 \in \mathcal{G}_1$ is the true (but unknown) post distribution. Then,*

$$\begin{aligned}&\mathbb{E}_\infty\left[z_\lambda(X)\right] = -\lambda \mathbb{D}_F(P_\infty \| Q_1) < 0, \textit{ and} \\ &\mathbb{E}_1\left[z_\lambda(X)\right] \geq \lambda \mathbb{D}_F(Q_1 \| P_\infty) > 0.\end{aligned}$$

Proof. Under some mild regularity conditions, Hyvärinen [2005] proved that

$$\mathbb{D}_{\mathrm{F}}(P \| Q) = \mathbb{E}_{X \sim P}\left[\frac{1}{2} \left\|\nabla_X \log p(X)\right\|_2^2 + \mathcal{S}_{\mathrm{H}}(X, Q)\right].$$

We use C_P to denote the term $\mathbb{E}_{X \sim P}\left[\frac{1}{2} \left\|\nabla_X \log p(X)\right\|_2^2\right]$. Then

$$\begin{aligned}&\mathbb{E}_\infty[\mathcal{S}_{\mathrm{H}}(X, P_\infty) - \mathcal{S}_{\mathrm{H}}(X, Q_1)] \\ &\quad = \mathbb{D}_{\mathrm{F}}(P_\infty \| P_\infty) - C_{P_\infty} - \mathbb{D}_{\mathrm{F}}(P_\infty \| Q_1) + C_{P_\infty} \\ &\quad = -\mathbb{D}_{\mathrm{F}}(P_\infty \| Q_1),\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_1[\mathcal{S}_{\mathrm{H}}(X,P_\infty)-\mathcal{S}_{\mathrm{H}}(X,Q_1)]\\&\quad=\mathbb{D}_{\mathrm{F}}(P_1\|P_\infty)-C_{P_1}-\mathbb{D}_{\mathrm{F}}(P_1\|Q_1)+C_{P_1}\\&\quad\geq\mathbb{D}_{\mathrm{F}}(Q_1\|P_\infty),\end{aligned}$$

where we applied Lemma 4.1.

Since $\lambda > 0$, the results follow. □

Lemma 4.2 shows that, prior to the change, the expected mean of instantaneous RSCUSUM score $z_\lambda(X)$ is negative. Consequently, the accumulated score has a negative drift at each time n prior to the change. Thus, the RSCUSUM detection score $Z(n)$ is pushed toward zero before the change point. This intuitively makes a false alarm unlikely. In contrast, after the change, the instantaneous score has a positive mean, and the accumulated score has a positive drift. Thus, the RSCUSUM detection score will increase toward infinity and leads to a change detection event.

Next, we discuss the values of the multiplier λ in the theoretical analysis. Obviously, with a fixed stopping threshold, a larger value of λ results in a smaller detection delay because the increment of the SCUSUM detection score is large, and the threshold can be easily reached. However, a larger value of λ also causes SCUSUM to stop prematurely when no change occurs, leading to a larger false alarm probability. Hence, the value of λ cannot be arbitrarily large (except in the degenerate case where $P_\infty(S_{\mathrm{H}}(X,Q_1) - S_{\mathrm{H}}(X,P_\infty) \leq 0) = 1$). It needs to satisfy the following key condition:

$$\mathbb{E}_\infty[\exp(z_\lambda(X))] \leq 1. \tag{14}$$

We will present a technical lemma that guarantees the existence of such a λ to satisfy inequality (14).

Lemma 4.3 (Existence of appropriate λ)**.** *There exists $\lambda > 0$ such that Inequality (14) holds. Moreover, either 1) there exists $\lambda^\star \in (0,\infty)$ such that the equality of (14) holds, or 2) for all $\lambda > 0$, the inequality of (14) is strict. As noted in Wu et al. [2023], the second case is of no practical interest.*

Proof. We give proof in the supplementary material. □

From now on, we consider a fix $\lambda > 0$ that satisfies Inequality (14) to present our core results. In practice, it is possible to use m past samples $\mathbf{X}_{[-m+1,0]}$ to determine the value of λ. In particular, λ can be chosen as the positive root of the function $\lambda \to \tilde{h}(\lambda)$ given by

$$\tilde{h}(\lambda) \triangleq \frac{1}{m}\sum_{i=1}^{m}[\exp(z_\lambda(X_{i-m}))] - 1.$$

By Lemma 4.3 and its related technical discussions, the above equation has a root greater than zero with a high probability if m is sufficiently large. In the case that λ is not chosen properly, the algorithm remains implementable but optimal performance of detection delay is not guaranteed. We discuss this situation further in the supplementary material.

Theorem 4.4. *Consider the stopping rule $T_{\texttt{RSCUSUM}}$ defined in Equation (7). Then, for any $\tau > 0$,*

$$\mathbb{E}_\infty[T_{\texttt{RSCUSUM}}] \geq e^\tau.$$

To satisfy the constraint of $\mathbb{E}_\infty[T_{\texttt{RSCUSUM}}] \geq \gamma$, it is enough to set the threshold $\tau = \log\gamma$.

Proof. We give proof in the supplementary material. □

Theorem 4.4 implies that the ARL increases at least exponentially as the stopping threshold increases.

The following theorem gives the asymptotic performance of the RSCUSUM algorithm in terms of the detection delay under the control of the ARL.

Theorem 4.5. *Subject to $\mathbb{E}_\infty[T_{\texttt{RSCUSUM}}] \geq \gamma > 0$, the stopping rule $T_{\texttt{RSCUSUM}}$ satisfies*

$$\begin{aligned}\mathcal{L}_{\texttt{WADD}}(T_{\texttt{RSCUSUM}}) \sim \mathcal{L}_{\texttt{CADD}}(T_{\texttt{RSCUSUM}}) &\sim \mathbb{E}_1[T_{\texttt{RSCUSUM}}]\\&\sim \frac{\log\gamma}{\lambda(\mathbb{D}_F(P_1\|P_\infty)-\mathbb{D}_F(P_1\|Q_1))}\\&\lesssim \frac{\log\gamma}{\lambda\mathbb{D}_F(Q_1\|P_\infty)}, \quad \textit{as } \gamma\to\infty.\end{aligned}$$

Proof. We give proof in the supplementary material. □

In the above theorem, we have used the notation $g(c) \lesssim h(c)$ as $c \to c_0$ to indicate that $\limsup \frac{g(c)}{h(c)} \leq 1$ as $c \to c_0$ for any two functions $c \mapsto g(c)$ and $c \mapsto h(c)$.

Theorems 4.4 and 4.5 imply that the *expected detection delay* (EDD) increases linearly as the stop threshold τ increases subject to a constraint on ARL.

5 IDENTIFICATION OF THE LEAST FAVORABLE DISTRIBUTION

Consider a general parametric distribution family $\mathcal{P}$ defined on $\mathcal{X}$. We use $\mathcal{P}_m$ to denote a set of a finite number of distributions belonging to $\mathcal{P}$, namely

$$\mathcal{P}_m = \{P_i,\ i = 1,\ldots,m:\ P_i \in \mathcal{P}\},\ m \in \mathbb{N}^+.$$

We use p_i to denote the density of each distribution $P_i,\ i = 1,\ldots,m$. Then, we define a convex set of densities

$$\mathcal{A}_m \triangleq \left\{x \mapsto \sum_{i=1}^{m}\alpha_i p_i(x) : \sum_{i=1}^{m}\alpha_i = 1, \alpha_i \geq 0\right\}. \tag{15}$$

We further define a set of functions

$$\mathcal{B}_m \triangleq \Big\{ x \mapsto \sum_{i=1}^{m} \beta_i(x) \nabla_x \log p_i(x) : \sum_{i=1}^{m} \beta_i(x) = 1,\ \beta_i(x) \geq 0,\ p_i \in \mathcal{P}_m \Big\}. \quad (16)$$

Consider the pre-change distribution P_∞ (with density p_∞) such that $P_\infty \in \mathcal{P}$ and $P_\infty \notin \mathcal{A}_m$. We use $\mathbb{E}_\infty$ to denote its corresponding expectation with p_∞. Next, we provide a result to identify the LFD in $\mathcal{A}_m$ in terms of the Fisher-divergence (as defined in Definition 3.3).

Theorem 5.1. *Assume that there exists an element $P_0 \in \mathcal{A}_m$ (with density p_0) such that*

$$\mathbb{E}_{p_0}\Big\{ \|\nabla_x \log p_0(X) - \nabla_x \log p_\infty(X)\|_2^2 \Big\} = \min_{p \in \mathcal{A}_m, \phi \in \mathcal{B}_m} \mathbb{E}_p \Big\{ \|\phi(X) - \nabla_x \log p_\infty(X)\|_2^2 \Big\}. \quad (17)$$

Then, we have

$$\mathbb{E}_{p_0}\Big\{ \|\nabla_x \log p_0(X) - \nabla_x \log p_\infty(X)\|_2^2 \Big\} = \min_{p \in \mathcal{A}_m} \mathbb{E}_p \Big\{ \|\nabla_x \log p(X) - \nabla_x \log p_\infty(X)\|_2^2 \Big\}.$$

Proof. For any $p \in \mathcal{A}_m$, there exist w_i such that $p = \sum_{i=1}^m w_i p_i$, where $w_i \geq 0$ and $\sum_{i=1}^m w_i = 1$. Direct calculations give

$$\begin{aligned}
&\mathbb{E}_p\Big\{ \|\nabla_x \log p(X) - \nabla_x \log p_\infty(X)\|_2^2 \Big\} \\
&= \mathbb{E}_p\Big\{ \Big\| \frac{\nabla_x p(X)}{p(X)} - \nabla_x \log p_\infty(X) \Big\|_2^2 \Big\} \\
&= \mathbb{E}_p\Big\{ \Big\| \frac{\sum_{i=1}^m w_i \nabla_x p_i(X)}{\sum_{i=1}^m w_i p_i(X)} - \nabla_x \log p_\infty(X) \Big\|_2^2 \Big\} \\
&= \mathbb{E}_p\Big\{ \Big\| \sum_{i=1}^m u_i(X) \nabla_x \log p_i(X) - \nabla_x \log p_\infty(X) \Big\|_2^2 \Big\},
\end{aligned}$$

where $u_i(X) = \frac{w_i p_i(X)}{\sum_{i=1}^m w_i p_i(X)}$ for all $i = 1, \ldots, m$, and $\sum_{i=1}^m u_i(X) = 1$. Clearly $\nabla_x \log u_i(x) - \nabla_x \log u_j(x) = \nabla_x \log p_i(x) - \nabla_x \log p_j(x)$ for all $1 \leq i, j \leq m$.

Using Condition (17), the quantity above is minimized at $p = p_0$, which concludes the proof. □

Theorem 5.1 provides an efficient way to identify the LFD in a convex set with only knowledge of the gradient of the log density functions.

Next, we provide a method to find the LFD in a class of Gaussian mixture models.

Theorem 5.2. *Let G_θ denote the d-dimensional Gaussian distribution centered at $\theta \in \mathbb{R}^d$ with a constant covariance matrix $V \in \mathbb{R}^{d \times d}$. Let the set $\Theta_1 \subseteq \mathbb{R}^d$ be compact and convex. Consider the pre-change distribution G_{θ_*} and post-change distribution class $\mathcal{G}_1$ defined as all Gaussian mixture models given by the convex hull of $\{G_\theta : \theta \in \Theta_1\}$. For any vector $v \in \mathbb{R}^d$, let $\|v\|_V = (v^T V^{-2} v)^{1/2}$. Assume that $\theta_* \notin \Theta_1$, and $\theta_0 \in \Theta_1$ is the closest to θ_* under the $\|\cdot\|_V$ norm, namely $\|\theta_0 - \theta_*\|_V = \min_{\theta \in \Theta_1} \|\theta - \theta_*\|_V$. Then, G_{θ_0} is the closest to G_{θ_*} among $\mathcal{G}_1$ under the Fisher divergence.*

Proof. Let g_{θ_0} and g_{θ_*} denote the densities of G_{θ_0} and G_{θ_*}, respectively. Clearly,

$$\min_{g_\theta \in \mathcal{G}_1} \mathbb{E}_{g_\theta}\Big\{ \|\nabla_x \log g_\theta(X) - \nabla_x \log g_{\theta_*}(X)\|_2^2 \Big\} \leq \mathbb{E}_{g_{\theta_0}}\Big\{ \|\nabla_x \log g_{\theta_0}(X) - \nabla_x \log g_{\theta_*}(X)\|_2^2 \Big\}$$

We will prove the equality by proving the reverse inequality. To this end, consider an arbitrary element of $\mathcal{G}_1$. By definition of convex hull, this element can be written as $G_1 = \sum_{i=1}^m w_i G_{\theta_i}(X)$ for some $m \geq 1$, $w_i \geq 0, i = 1, \cdots, m$ with $\sum_{i=1}^m w_i = 1$ and $\theta_i \in \Theta_1$ for $i = 1, \cdots, m$. As proved in the above theorem

$$\mathbb{E}_{g_1}\Big\{ \|\nabla_x \log g_1(X) - \nabla_x \log g_{\theta_*}(X)\|_2^2 \Big\} = \mathbb{E}_{g_1}\Big\{ \Big\| \sum_{i=1}^m \beta_i(X) \nabla_x \log g_{\theta_i}(X) - \nabla_x \log g_{\theta_*}(X) \Big\|_2^2 \Big\},$$

where $\beta_i(X) = \frac{w_i g_{\theta_i}(X)}{\sum_{i=1}^m w_i g_{\theta_i}(X)}$ for all $i = 1, \ldots, m$.

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{g_1}\Big\{ \|\nabla_x \log g_1(X) - \nabla_x \log g_{\theta_*}(X)\|_2^2 \Big\} \\
&= \mathbb{E}_{g_1} \Big\| \sum_{i=1}^m \beta_i(X)(X - \theta_i) - (X - \theta_*) \Big\|_V^2 \\
&= \mathbb{E}_{g_1} \Big\| \sum_{i=1}^m \beta_i(X)(\theta_* - \theta_i) \Big\|_V^2.
\end{aligned}$$

Using the assumption that $\|\theta_0 - \theta_*\|_V = \min_{\theta \in \Theta_1} \|\theta - \theta_*\|_V$, we have

$$\begin{aligned}
&= \mathbb{E}_{g_1} \Big\| \sum_{i=1}^m \beta_i(X)(\theta_* - \theta_i) \Big\|_V^2 \\
&= \mathbb{E}_{g_1} \Big\| \theta_* - \sum_{i=1}^m \beta_i(X) \theta_i \Big\|_V^2 \geq \mathbb{E}_{g_\theta} \|\theta_* - \theta_0\|_V^2 \\
&= \mathbb{E}_{g_{\theta_0}}\Big\{ \|\nabla_x \log g_{\theta_0}(X) - \nabla_x \log g_{\theta_*}(X)\|_2^2 \Big\}.
\end{aligned}$$

This concludes the proof. □

For a general parametric family of potential post-change distributions, it may be difficult to identify the LFD. In Section 6.1, we propose a method to find the LFD in parameter space.

6 NUMERICAL RESULTS

In this section, we present numerical results for both synthetic and real data demonstrating the robustness of RSCUSUM. Specifically, we identify the LFD in $\mathcal{G}_1$ defined as convex hull of given distributions $P_i(x)$, $i = 1, 2, \cdots, m$. To this end, we minimize the Fisher divergence over the set $\mathcal{B}_m$ defined in Equation (16) and invoke Theorem 5.1. In general, we can then estimate the $\nabla_x \log p_0(x)$ for LFD by $\sum_{i=1}^m \beta_i(x) \nabla_x \log p_i(x)$.

6.1 EXAMPLE OF THE LEAST FAVORABLE DISTRIBUTION

We consider the parametric family $\mathcal{P}$ as the multivariate Normal distribution (MVN), a subfamily [Yu et al., 2016] of the exponential family (EXP), and the Gauss-Bernoulli Restricted Boltzmann Machine (RBM) [LeCun et al., 2006]. For example, in the case of MVN,

$$\mathcal{G}_\infty = \{\mathcal{N}(\mu_*, V_*)\},$$
$$\mathcal{G}_1 = \left\{ \sum_{i=1}^m \alpha_i \mathcal{N}(\boldsymbol{\mu}_i, V_i) : \sum_{i=1}^m \alpha_i = 1, \ \forall\, \alpha_i \geq 0 \right\}.$$

Here the pre-change distribution $P_\infty = \mathcal{N}(\mu_*, V_*)$ and the uncertainty class $\mathcal{G}_1$ is constructed from a finite basis $\mathcal{P}_m = \{\mathcal{N}(\boldsymbol{\mu}_i, V_i), \ i = 1, \ldots, m\}$ (see Equation (15)). Each basis element P_i is parameterized by the corresponding vector $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, V_i)$. Without loss of generality, we assume $\boldsymbol{\theta}_1$ to be the closest to $\boldsymbol{\theta}_* = (\boldsymbol{\mu}_\star, V_\star)$ in L_2 (Euclidean) norm.

By Theorem 5.1, it is sufficient to find P_0 such that Condition (17) holds. Any $\phi(x) \in \mathcal{B}_m$ is characterized by coefficients $\beta_j(\cdot)$, $j = 1, \ldots, m$ (see Equation (16)).

We use a neural network $\mathrm{Softmax}_j \circ f_{NN}(x)$ to estimate $\beta_j(\cdot)$, specifically,

$$\beta_j(x) = \mathrm{Softmax}_j \circ f_{NN}(x),$$

where f_{NN} is given by the feature extractor part of a multi-layer neural network corresponding to hidden layer sizes $[128 - 64 - m]$, with $\mathrm{Softmax}$ the last layer all ReLU activation functions in hidden layers. Note that $\mathrm{Softmax}_j$ denotes the j-th element of the Softmax function. The use of Softmax function ensures $\sum_{i=1}^m \beta_i(x) = 1$ and $\beta_i(x) \geq 0$ for all $1 \leq i \leq m$.

To identify P_0, we learn f_{NN} by minimizing the following loss function over the training sample $X_1, \cdots, X_N \sim P$:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \left\| \sum_{j=1}^m \beta_j(X_i) \nabla \log p_j(X_i) - \nabla \log p_\infty(X_i) \right\|_2^2,$$

where P is updated at each epoch based on the learned coefficients $\beta_i(x)$ by

$$\nabla_x \log p(x) = \sum_{i=1}^m \beta_i(x) \nabla_x \log p_i(x).$$

To generate samples from the unnormalized density function $\nabla_x \log p(x)$, standard Markov Chain Monte Carlo (MCMC) techniques (such as MALA) are employed. Furthermore, the neural network is trained using the Adam optimizer.

In Table 1, we report the average value $\frac{1}{M} \sum_{i=1}^M \beta_j(Y_i)$ over the test sample $Y_1, \cdots Y_M \sim P$ respectively in cases where the basis elements of $\mathcal{P}_m$ are MVN_m (with mean shifts), MVN_c (with covariance shifts), EXP, and RBMs. Details of P_∞ and basis elements of $\mathcal{P}_m$ are given in the Supplementary Material. In all cases the average value of $\beta_1(y)$ (respectively $\beta_j(y)$, $j = 2, 3, 4$) is extremely close to 1 (respectively to 0). This gives strong evidence that the LFD is achieved by one of the basis $\mathcal{P}_m$, and Theorem 5.1 can be invoked to give the LFD.

j	1	2	3	4
MVN_m	1.00e+00	4.90e-09	2.43e-11	6.29e-12
MVN_c	9.99e-01	7.47e-06	3.23e-08	3.55e-08
EXP	9.99e-01	2.84e-05	1.37e-09	1.01e-09
RBM	1.00e+00	3.18e-33	0.00e+00	0.00e+00

Table 1: Empirical average values of $\beta_j(x)$ over 10000 test sample for MVN, EXP, and RBM models.

6.2 SYNTHETIC DATA

As in Subsection 6.1, we simulate synthetic data streams from MVNs and RBMs to evaluate the performance of RSCUSUM. The LFD in the uncertainty class is identified as in Subsection 6.1. We also report the performance of the SCUSUM (which is not robust) Wu et al. [2023] for arbitrary *wrong* distributions in the uncertain class.

We consider a change detection scenario where the pre- and post-distributions are modeled by MVN (respectively RBM) models with $m = 4$. Both P_∞ and the elements of the uncertainty class are created according to detailed descriptions in the supplementary material. We use Gibbs sampling method with 1000 iterations for RBMs. In each trial, we treat one of $P_i \in \mathcal{P}_m$, $i = 1, 2, 3, 4$ as the *true* post-change distribution. For each trial, we perform the experiment for 1000 runs.

In all experiments, we set the change point as $\nu = 50$, and we set the total length of each data stream as 10000 to assure the generated data stream is long enough for detection. We evaluate the detection delay for ARL values ranging from 100 to 3000.

In Figure 1(a) and (b), we respectively report the detection scores versus time in cases for MVN_m and RBM experiments. The results demonstrate that the average increment of detection scores is positive for RSUCUM, while negative for the non-robust SCUSUM. This means that a non-robust CUSUM fails to detect this post-change scenario but the RSCUSUM algorithms detects it.

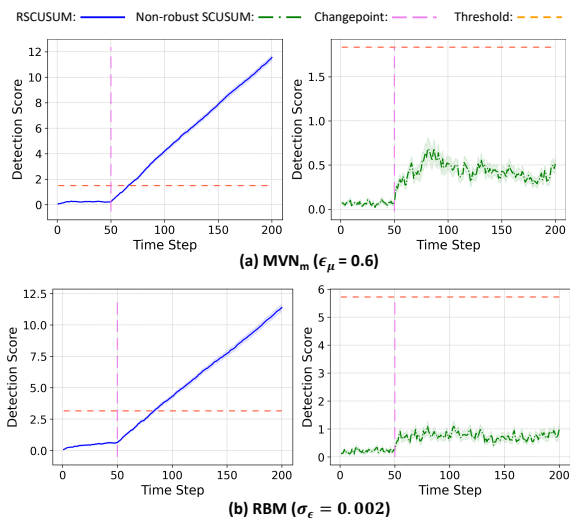

Figure 1: Detection score versus Time.

In Figure 2(a) and (b), we respectively demonstrate the empirical EDD against log-scaled ARL for both MVN_m and RBM experiments. The results demonstrate that RSCUSUM is robust and performs competitively in terms of detection delay. In particular, we observe that the EDD of RSCUSUM (subplot in left rows) increases at a linear rate for all cases, while some EDD of non-robust SCUSUM (subplot in right rows) may increase at an exponential rate (compare the y-axis labels for the plots).

7 CONCLUSIONS

In this work, we proposed the RSCUSUM algorithm, a robust score-based algorithm for quickest change detection when the post-change distribution is not precisely known. We defined the least favorable distribution in the sense of Fisher divergence. Using asymptotic analysis, we also analyzed the delay and false alarms of RSCUSUM in the sense of Lorden's and Pollak's metrics. We provided both theoretical and algorithmic methods for computing the least favorable distribution for unnormalized models. Numerical

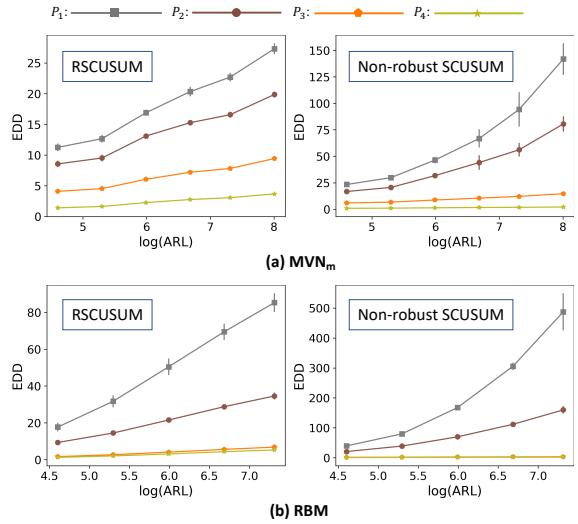

Figure 2: EDD versus log-scaled ARL.

simulations were provided to demonstrate the performance of our robust algorithm.

Acknowledgements

Suya Wu and Vahid Tarokh were supported in part by Air Force Research Lab Award under grant number FA-8750-20-2-0504. Jie Ding was supported in part by the Office of Naval Research under grant number N00014-21-1-2590. Taposh Banerjee was supported in part by the U.S. Army Research Lab under grant W911NF2120295.

References

Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6(4), 2005.

Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inf. Theory*, 44(7):2917–2929, 1998.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. In *Predicting structured data*, volume 1. The MIT Press, 2006.

Gary Lorden. Procedures for reacting to a change in distribution. *Ann. Math. Stat.*, pages 1897–1908, 1971.

George V Moustakides. Optimal stopping times for detecting changes in distributions. *Ann. Stat.*, 14(4):1379–1387, 1986.

ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

Moshe Pollak. Optimal detection of a change in distribution. *Ann. Stat.*, pages 206–227, 1985.

H Vincent Poor and Olympia Hadjiliadis. *Quickest detection*. Cambridge University Press, 2008.

SW Roberts. A comparison of some control chart procedures. *Technometrics*, 8(3):411–430, 1966.

Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory Probab. Appl.*, 8(1):22–46, 1963.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

Alexander G Tartakovsky and Venugopal V Veeravalli. General asymptotic bayesian theory of quickest change detection. *Theory of Probability & Its Applications*, 49 (3):458–497, 2005.

Jayakrishnan Unnikrishnan, Venugopal V Veeravalli, and Sean P Meyn. Minimax robust quickest change detection. *IEEE Transactions on Information Theory*, 57(3):1604–1614, 2011.

Venugopal V Veeravalli and Taposh Banerjee. Quickest change detection. In *Academic press library in signal processing*, volume 3, pages 209–255. Elsevier, 2014.

Michael Woodroofe. *Nonlinear renewal theory in sequential analysis*. SIAM, 1982.

Suya Wu, Enmao Diao, Taposh Banerjee, Jie Ding, and Vahid Tarokh. Score-based change point detection for unnormalized models. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Ming Yu, Mladen Kolar, and Varun Gupta. Statistical inference for pairwise graphical models using score matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.