Score-based Quickest Change Detection for Unnormalized Models

Suya Wu[1] **Enmao Diao**[1] **Taposh Banerjee**[2] **Jie Ding**[3] **Vahid Tarokh**[1]
[1]Duke University [2]University of Pittsburgh [3]University of Minnesota Twin Cities

Abstract

Classical change detection algorithms typically require modeling pre-change and post-change distributions. The calculations may not be feasible for various machine learning models because of the complexity of computing the partition functions and normalized distributions. Additionally, these methods may suffer from a lack of robustness to model mismatch and noise. In this paper, we develop a new variant of the classical Cumulative Sum (CUSUM) change detection, namely Score-based CUSUM (SCUSUM), based on Fisher divergence and the Hyvärinen score. Our method allows the applications of the quickest change detection for unnormalized distributions. We provide a theoretical analysis of the detection delay given the constraints on false alarms. We prove the asymptotic optimality of the proposed method in some particular cases. We also provide numerical experiments to demonstrate our method's computation, performance, and robustness advantages. [1]

1 INTRODUCTION

Determining abrupt changes in the underlying distributions of online data streams as quickly as possible is an important problem commonly encountered in many applications. These algorithms typically rely on various pre-change and post-change data statistics, e.g., cumulative means. A false alarm occurs when a change has not happened but is declared by the change detection algorithm. However, reducing false alarms too strenuously can make a longer wait time between a change event and the time that a change is declared (often defined as the detection delay). A good algorithm provides a good balance between the false alarm probability and detection delay. The performance of the detection algorithm typically depends on the change point (Veeravalli and Banerjee, 2014). In this light, for a given algorithm, we are interested in the minimax objectives that evaluate the trade-off between the worst conditional detection delay and the probability of a false alarm (Pollak, 1985).

Unfortunately, most state-of-the-art methods require full knowledge of pre- and post-change distributions. This is not available in many modern machine learning applications when the data-generating distributions must be modeled using the available data. These models may be high-dimensional and, in some cases, may not lend themselves to explicit distributions. In some cases, such as energy-based models (LeCun et al., 2006), graphical models (Koller and Friedman, 2009), and score-based deep generative models (Song et al., 2020), the models are very expressive. They may be explicit to a normalization factor but computationally cumbersome to normalize. The likelihood of unnormalized models can be approximated by Monte Carlo-based methods (e.g., (Hinton, 2002) and the references therein). However, the performance of change detection may suffer from the underlying approximation errors. In particular, Chen and Zhang (2015) showed that the likelihood-based change detection algorithms for multivariate data are extremely sensitive through numerical simulations. For image datasets, Nalisnick et al. (2018) showed that the likelihood, learned from deep generative models (such as VAE-based or flow-based deep generative models), is not robust in the detection of distribution drifts.

This motivates our research in this paper, where we provide an online change detection scheme that one can use with unnormalized distributions. We are motivated by Hyvärinen and Dayan (2005), who established an empirical estimation procedure for unnormalized models. This is also known as score matching and can be used as a surrogate for maximum likelihood estimation (MLE). Score matching has been extended successfully to discrete (Lyu, 2012), non-parametric (Sriperumbudur et al., 2017), directional (Mardia et al., 2016) distribution estimations, and deep generative models Song et al. (2020). To present our approach, we first review the classic CUSUM detec-

[1]Our code is available at this URL.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

tion rule and develop an approach that replaces the negative log-likelihood terms in CUSUM with a multiple of the Hyvärinen score. We then mathematically analyze the new Score-based CUSUM (SCUSUM) algorithm. We summarize the main contributions of this work below.

- We propose SCUSUM, a new quickest change detection algorithm that applies to unnormalized models for pre- and post-change distributions.
- We provide a theoretical analysis of the performance of SCUSUM using Pollak's minimax objective (Pollak, 1985). Assume that the outcomes before (respectively after) the change point are drawn independent and identically distributed (*i.i.d.*) according to pre-change (respectively post-change) distribution. We prove that under no change assumption, the expected running length increases exponentially as a function of the stopping threshold (Theorem 3). Moreover, if a change occurs, we prove that the worst-case detection delay is a linear function of the stopping threshold (Theorem 4).
- We conduct extensive numerical experiments on synthetic data to demonstrate the performance of SCUSUM and compare it with the classical detection methods CUSUM (Page, 1955), Scan B-statistic (Li et al., 2019), and CALM-MMD (Cobb et al., 2022). Our method performs competitively with CUSUM in terms of empirical detection delay with respect to the expected run length to false alarms. SCUSUM outperforms Scan B-statistic and CALM-MMD in all non-Gaussian cases. Our experiments further illustrate the computational advantage of unnormalized models of SCUSUM over CUSUM.

2 RELATED WORK

Classical developments in the quickest change detection assumed those pre- and post-change distributions are explicitly known. In this case, if the outcomes before (respectively after) the change point are drawn *i.i.d.* according to pre-change (respectively post-change) distribution, Moustakides (1986) proved that the log-likelihood based CUSUM (described below) provides the optimal trade-off between worst-case detection delay and false alarm probability in the sense of Lorden (1971). Relaxing the independence assumption, Lai (1998) developed a window-limited generalized likelihood-based CUSUM and proved its *asymptotic* optimality in the sense of Pollak (1985). Another state-of-the-art likelihood-based approach is the Shiryaev–Roberts (SR) procedure and its extensions (Shiryaev, 1963; Roberts, 1966). These have been studied in both Bayesian and non-Bayesian settings (Pollak, 1985; Moustakides et al., 2011; Tartakovsky et al., 2012) and be optimal in a sense defined by Polunchenko and Tartakovsky (2010). For a more detailed discussion of the state-of-the-art theoretical results in this classical setting, we refer the reader to (Polunchenko and Tartakovsky, 2012; Veeravalli and Banerjee, 2014) and references therein.

Numerous recent advances have been made in the field (see Xie et al., 2021, and the references therein). For high-dimensional data streams, subspace dynamics has been investigated to change detection (Kawahara et al., 2007; Jiao et al., 2018; Xie et al., 2020; Alanqary et al., 2021). In contrast, Kernel-based nonparametric methods (Harchaoui et al., 2008; Li et al., 2015, 2019) have been proposed that employ higher-dimensional feature spaces. Other advances include *model-free* change detection methods such as graph-based (Sharpnack et al., 2013; Chen and Zhang, 2015), nearest neighbors based (Banerjee et al., 2018; Chen, 2019), and distance-based (Padilla et al., 2019; Cheng et al., 2020) methods. Most of these methods make less stringent assumptions on the pre- and post-change distributions than the classical approaches. In contrast, they do not lend themselves easily to theoretical analysis.

New applications of sequential change-point detection have also emerged beyond classical domains. A noteworthy example is in the field of continual learning or life-cycle modeling (Klaise et al., 2020). In this domain, attention has been paid to joint online training from streaming data and detecting changes in high-dimensional scenarios. For over-parameterized models (such as deep neural networks), Titsias et al. (2022) proposed an online change detection algorithm based on sequential learning (such as the deep neural network training process). Other likelihood-inspired empirical methods such as those proposed (Ren et al., 2019; Xiao et al., 2020; Kim et al., 2021) for out-of-distribution detection (OOD) have been developed for offline settings.

3 BACKGROUND

3.1 Proper Scoring Rules

Let X be a random variable with values in $\mathcal{X} \subseteq \mathbb{R}^d$, and let $\mathcal{P}$ be a family of distributions on $\mathcal{X}$. Let P and $Q \in \mathcal{P}$ denote the true data-generating distribution and a postulated distribution, and let p and q respectively denote their corresponding probability density functions (PDFs). Dawid and Musio (2014) introduced proper scoring rules as a unified framework to measure the quality of postulated models.

Definition 1 (Proper Scoring Rule)**.** A scoring rule (Dawid, 2007; Parry et al., 2012; Dawid and Musio, 2014) is a function $(X, Q) \mapsto \mathcal{S}(X, Q)$ that measures the quality of Q for modeling data X. It is said to be *proper* if for all $P \in \mathcal{P}$, the expected score $\mathbb{E}_{X\sim P}\mathcal{S}(X, Q)$ is minimized at $Q = P$ over all $Q \in \mathcal{P}$. Moreover, $\mathcal{S}$ is *strictly proper* with respect to $\mathcal{P}$, if for any $Q \in \mathcal{P}$ and $Q \neq P$, $\mathbb{E}_{X\sim P}[\mathcal{S}(X, Q)] > \mathbb{E}_{X\sim P}[\mathcal{S}(X, P)]$.

The logarithmic scoring rule is a well-known and widely applied example of a strictly proper scoring rule.

Definition 2 (Logarithmic Scoring Rule)**.** The logarithmic scoring rule (also called the log score) is

$$(X, Q) \mapsto \mathcal{S}_{\mathrm{L}}(X, Q) \triangleq -\log q(X),$$

where q is the PDF of Q.

Minimizing the log score is equivalent to maximum likelihood estimation (MLE) and minimizing the associated Kullback-Leibler (KL) divergence

$$\mathbb{D}_{\mathrm{KL}}(P\|Q) \triangleq \mathbb{E}_{X\sim P}\left[\log p(X) - \log q(X)\right].$$

Since $\mathbb{D}_{\mathrm{KL}}(P\|Q) > 0$ for any $Q \neq P$, the log score is *strictly proper*.

3.2 Fisher divergence and Hyvärinen score

Like before, we consider a family of distributions $\mathcal{P}$. However, we assume that any distribution $Q \in \mathcal{P}$ with the PDF $q(x)$ is potentially known only up to a normalizing constant. In other words, instead of $q(x)$, we are given $\tilde{q}(x)$ with

$$q(x) = \frac{\tilde{q}(x)}{\int_{x\in\mathcal{X}} \tilde{q}(x)dx}.$$

In many cases, the computation of the denominator (also known as the *normalizing factor* or the *partition function*) may be intractable. In fact, the number of points required for approximating the integral in the above may grow exponentially as a function of the dimension of the $\mathcal{X}$.

In addressing this issue, Hyvärinen and Dayan (2005) proposed a scale-invariant proper scoring function, referred to as the Hyvärinen score, which is closely connected to the Fisher divergence.

Definition 3 (Hyvärinen Score)**.** The Hyvärinen score (Hyvärinen and Dayan, 2005) is a mapping $(X, Q) \mapsto \mathcal{S}_{\mathrm{H}}(X, Q)$ given by

$$\mathcal{S}_{\mathrm{H}}(X, Q) \triangleq \frac{1}{2}\left\|\nabla_X \log q(X)\right\|_2^2 + \Delta_X \log q(X), \quad (1)$$

where ∇_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 $X = (x_1, \cdots, x_d)^\top$.

It is easy to see that the Hyvärinen score remains invariant if $\tilde{q}$ is used instead of q in Equation (1).

Under some mild regularity conditions on p and q, Hyvärinen and Dayan (2005) showed that

$$\begin{aligned}\mathbb{D}_{\mathrm{F}}(P\|Q) &\triangleq \mathbb{E}_{X\sim P}\left[\left\|\nabla_{\mathbf{x}} \log p(X) - \nabla_{\mathbf{x}} \log q(X)\right\|_2^2\right] \\ &= \mathbb{E}_{X\sim P}\left[\frac{1}{2}\left\|\nabla_{\mathbf{x}} \log p(X)\right\|_2^2 + \mathcal{S}_{\mathrm{H}}(X, Q)\right],\end{aligned}$$

where $\mathbb{D}_{\mathrm{F}}(P\|Q)$ is the classical *Fisher Divergence* from the distribution P to Q. Clearly, the invariance of the Hyvärinen score is inherited from the invariance of Fisher divergence with respect to the normalization factors. Additionally, it is easy to verify that $\mathbb{D}_{\mathrm{F}}(P\|Q) > 0$ for $Q \neq P$. It follows that the Hyvärinen score is *strictly proper*.

4 SCORE-BASED QUICKEST CHANGE DETECTION

4.1 Problem Background

Let $\{X_n\}_{n\geq 1}$ denote a sequence of independent random observations defined on the probability space $(\Omega, \mathcal{F}, P_\nu)$. Let $\mathcal{F}_n$ be the $\sigma-$algebra generated by random variables $\{X_n\}_{n\geq 1}$ and $\mathcal{F} = \sigma(\cup_{n\geq 1}\mathcal{F}_n)$, the $\sigma-$algebra generated by the union of sub-σ-algebras. We treat $\nu \geq 1$ as the time when an abrupt change has happened. Under P_ν, the observations $X_1, X_2, \ldots, X_{\nu-1}$ are *i.i.d.* according to a distribution P_∞, and $X_\nu, X_{\nu+1}, \ldots$ are *i.i.d.* according to another distribution P_1. We intuitively consider ν as the change point, P_∞ as the pre-change distribution, and P_1 as the post-change distribution. We write $\nu = \infty$ when no change ever happens, and $\nu = 1$ when all observations follow P_1. In the rest of the paper, we refer to the probability measure of the entire sequence $\{X_n\}_{n\geq 1}$ also by P_∞ when no change occurs. Similarly we refer to the law of $\{X_n\}_{n\geq 1}$ also by P_1 when $\nu = 1$. The differences will always be clear from the context.

We focus on the classical scenario where the pre- and post-change distributions, P_∞ and P_1, are known, and the change point ν is unknown but deterministic. We use $\mathbb{E}_\nu$ and Var_ν respectively to denote the expectation and the variance operator with the measure P_ν.

Any change detection scheme defines a stopping rule T with respect to the data stream $\{X_n\}_{n\geq 1}$. Clearly, for any n,

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

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.

Lorden (1971) defined the "double" worst averaged detection delay (WADD) as

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

where $(z)^+ \triangleq \max(z, 0)$ for any $z \in \mathbb{R}$. This leads to the minimax optimization problem

$$\text{minimize} \quad \mathcal{L}_{\mathtt{WADD}}(T) \quad \text{subject to} \quad \mathbb{E}_\infty[T] \geq \gamma, \quad (3)$$

overall stopping rules T (Lorden, 1971). Under the *i.i.d* assumptions for pre-change (respectively post-change) outcomes, Lorden (1971) showed that the likelihood-based CUSUM (Page, 1955) is an asymptotically optimal solution to the above optimization problem as $\gamma \to \infty$. Moustakides (1986) later proved that CUSUM provides an optimal solution to the above problem for any $\gamma > 0$.

Pollak (1985) proposed an alternative measure of detection delay. It replaces the double maximization of Lorden's problem with a single maximization over all possible $\nu \geq 1$. We define the worst conditional averaged detection delay (CADD) by

$$\mathcal{L}_{\texttt{CADD}}(T) \triangleq \sup_{\nu \geq 1} \mathbb{E}_{\nu}[T - \nu | T \geq \nu]. \tag{4}$$

Then, Pollak (1985) formulated an optimal stopping rule as a solution to

$$\text{minimize} \quad \mathcal{L}_{\texttt{CADD}}(T) \quad \text{subject to} \quad \mathbb{E}_{\infty}[T] \geq \gamma, \tag{5}$$

overall possible stopping rules T. When $\gamma \to \infty$, the Shiryaev-Roberts-Pollak procedure has been shown to be asymptotically optimal (Pollak, 1985) to the above problem. However, for any fixed $\gamma > 0$, the optimal point to Problem (5) remains unsolved. The worst CADD is referred to as Pollak's optimality criterion or Pollak's metric in the following. It is worth noting that $\mathcal{L}_{\texttt{WADD}}(T) \geq \mathcal{L}_{\texttt{CADD}}(T)$ for any stopping rule T.

4.2 The SCUSUM Algorithm

In this section, we first review the classic CUSUM rule and subsequently present our Score-based CUSUM (SCUSUM) algorithm. Following the scheme of CUSUM, the proposed method can be in a recursive form, which is not too demanding in computational and memory requirements for online implementation.

Given the data stream $\{X_n\}_{n \geq 1}$, the log score-based CUSUM rule is defined by,

$$T_{\texttt{CUSUM}} \triangleq \inf\left\{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log \frac{p_1(X_i)}{p_{\infty}(X_i)} \geq \tau \right\},$$

where $\tau > 0$ is referred to as the stopping threshold, and the infimum of the empty set is defined to be $+\infty$. The value of this threshold is clearly related to the false alarm probability. It is known (Lai, 1998) that $T_{\texttt{CUSUM}}$ can be written as

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

Here, $\Lambda(n)$ can be computed using the recursion

$$\begin{aligned} &\Lambda(0) = 0, \\ &\Lambda(n) \triangleq \left(\Lambda(n-1) + \log \frac{p_1(X_n)}{p_{\infty}(X_n)} \right)^{+}, \forall n \geq 1. \end{aligned} \tag{6}$$

It leads to a computationally convenient stopping scheme.

The results of (Lorden, 1971; Moustakides, 1986; Pollak, 1985) demonstrate that for any value of stopping threshold, the log-score based CUSUM achieves the optimality in solving Problem (3). It is also asymptotically optimal for Problem (5), e.g., as $\gamma \to \infty$,

$$\mathcal{L}_{\texttt{CADD}}(T_{\texttt{CUSUM}}) \sim \frac{\log \gamma}{\mathbb{D}_{\texttt{KL}}(P_1 \| P_{\infty})}. \tag{7}$$

Here, for two functions $c \mapsto g(c)$ and $c \mapsto h(c)$, the notation $g(c) \sim h(c)$ as $c \to c_0$ indicates that $g(c) = h(c)(1 + o(1))$ as $c \to c_0$.

Recall from Section 3 that the Hyvärinen score function is a surrogate of the log score function, which applies to unnormalized models. Motivated by this analogy, we consider replacing the log scores with the Hyvärinen scores in the classical CUSUM scheme. Next, we give the definition of the SCUSUM detection score and then explain the detection algorithm.

Let X represent a generic random variable. We define the instantaneous SCUSUM score function $X \mapsto z_{\lambda}(X)$ by

$$z_{\lambda}(X) \triangleq \lambda\big(\mathcal{S}_{\texttt{H}}(X, P_{\infty}) - \mathcal{S}_{\texttt{H}}(X, P_1)\big), \tag{8}$$

where $\lambda > 0$ is a multiplier, $\mathcal{S}_{\texttt{H}}(X, P_{\infty})$ and $\mathcal{S}_{\texttt{H}}(X, P_1)$ are respectively the Hyvärinen score functions of pre-change and post-change distributions. In Section 4.3, we will show that the multiplier λ needs to be chosen appropriately in the theoretical analysis of SCUSUM. Since λ needs to be pre-determined, we refer to it as a hyperparameter. In Section 5, we will discuss how to determine the value of λ for SCUSUM in practice.

Our proposed stopping rule is given by

$$T_{\texttt{SCUSUM}} \triangleq \inf\left\{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} z_{\lambda}(X_i) \geq \tau \right\}, \tag{9}$$

where $\tau > 0$ is a pre-selected stopping threshold whose value may be used to control the false alarm probability. As the stopping scheme in CUSUM, the stopping rule of SCUSUM can be written as

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

where $Z(n)$, referred to as the SCUSUM detection score, can be computed recursively by

$$\begin{aligned} &Z(0) = 0, \\ &Z(n) \triangleq (Z(n-1) + z_{\lambda}(X_n))^{+}, \ \forall n \geq 1. \end{aligned}$$

The SCUSUM algorithm is summarized in Algorithm 1.

Algorithm 1: SCUSUM Detection Algorithm

Input: The Hyvarinen score functions $\mathcal{S}_{\text{H}}(\cdot, P_\infty)$ and $\mathcal{S}_{\text{H}}(\cdot, P_1)$ of pre-change and post-change distributions, respectively.

Data: m previous observations $\mathbf{X}_{[-m+1,0]}$ and the online data stream $\{X_1, X_2, \ldots\}$

Initialization:

Set the current time $k = 0$, hyperparameter $\lambda > 0$, stopping threshold $\tau > 0$, and detection score $Z(0) = 0$

while $Z(k) < \tau$ **do**

$k = k + 1$

Update $z_\lambda(X_k) = \lambda(\mathcal{S}_{\text{H}}(X_k, P_\infty) - \mathcal{S}_{\text{H}}(X_k, P_1))$

Update $Z(k) = \max(Z(k-1) + z_\lambda(X_k), 0)$

Record the current time k as the stopping time $\hat{T}$

Locate the change point by $\hat{\nu} = \arg\min_{1 \le i \le k} Z(i)$

Output: $\hat{T}$ and $\hat{\nu}$

4.3 Theoretical Analysis

Using the same notations and setting of Subsection 4.1, we next provide a theoretical analysis of SCUSUM. The proofs of all the results are provided in the supplementary material. We first formalize an intuitive justification for the effectiveness of SCUSUM below.

Lemma 1 (Positive and Negative Drifts)**.** *Consider the instantaneous SCUSUM score function $X \mapsto z_\lambda(X)$ as defined in Equation (8). Then,*

$$\mathbb{E}_\infty [z_\lambda(X)] = -\lambda \mathbb{D}_{\text{F}}(P_\infty \| P_1) < 0, \text{ and}$$
$$\mathbb{E}_1 [z_\lambda(X)] = \lambda \mathbb{D}_{\text{F}}(P_1 \| P_\infty) > 0.$$

Lemma 1 shows that, prior to the change, the expected mean of instantaneous SCUSUM score $z_\lambda(X)$ is negative under the measurement of random observations. Consequently, the accumulated score has a negative drift at each time n prior to the change. Thus, the SCUSUM 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 SCUSUM 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, except in the degenerate case, where $P_\infty(S_{\text{H}}(X, P_1) - S_{\text{H}}(X, P_\infty) \le 0) = 1$, the value of λ cannot be arbitrarily large. It needs to satisfy the following key condition:

$$\mathbb{E}_\infty[\exp(z_\lambda(X))] \le 1. \tag{10}$$

We will present a technical lemma that guarantees the existence of such a λ to satisfy Inequality (10).

Lemma 2 (Existence of appropriate λ)**.** *There exists $\lambda > 0$ such that Inequality (10) holds. Moreover, either 1) there exists $\lambda^\star \in (0, \infty)$ such that the equality of (10) holds, or 2) for all $\lambda > 0$, the inequality of (10) is strict (This case is shown to be pathological in the supplementary material).*

From now on, we consider a fix $\lambda > 0$ that satisfies Inequality (10) 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

$$\tilde{h}(\lambda) \triangleq \frac{1}{m} \sum_{i=1}^{m} [\exp(z_\lambda(X_{i-m}))] - 1. \tag{11}$$

By Lemma 2 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 the performance of detection delay is not guaranteed. We discuss this situation further in Remark 1.

Theorem 3. *Consider the stopping rule T_{SCUSUM} defined in Equation (9). Then, for any $\tau > 0$,*

$$\mathbb{E}_\infty[T_{\text{SCUSUM}}] \ge e^\tau. \tag{12}$$

$\mathbb{E}_\infty[T_{\text{SCUSUM}}]$ is also referred to as the *Average Run Length* (ARL) (Page, 1955). Theorem 3 implies that the ARL increases at least exponentially as the stopping threshold increases. The following theorem gives the asymptotic performance of SCUSUM in terms of the detection delay under the control of the ARL.

Theorem 4. *Subject to $\mathbb{E}_\infty[T_{\text{SCUSUM}}] \ge \gamma > 0$, the stopping rule T_{SCUSUM} satisfies*

$$\begin{aligned} \mathcal{L}_{\text{WADD}}(T_{\text{SCUSUM}}) &\sim \mathcal{L}_{\text{CADD}}(T_{\text{SCUSUM}}) \\ &\sim \mathbb{E}_1[T_{\text{SCUSUM}}] \\ &\sim \frac{\log \gamma}{\lambda \mathbb{D}_{\text{F}}(P_1 \| P_\infty)}, \end{aligned} \tag{13}$$

as $\gamma \to \infty$.

The value $\mathbb{E}_1[T_{\text{SCUSUM}}]$ is also referred to as the *Expected Detection Delay* (EDD) in the literature. Theorems 3 and 4 imply that the EDD increases linearly as the stop threshold τ increases subject to a constraint on ARL.

Remark 1. It is worth noting that although results of our core results hold for a pre-selected λ that satisfied the Inequality (10), the effect of choosing any other λ' amounts

to the scaling of all the increments of SCUSUM by a constant factor of λ'/λ. This means that all of these results still hold adjusted for this scale factor. For instance, the result of Theorem 3 can be modified to be written as

$$\mathbb{E}_\infty[T_{\texttt{SCUSUM}}] \geq \exp\left\{\frac{\lambda\tau}{\max(\lambda,\lambda')}\right\},$$

for any $\lambda' > 0$. It is easy to see that this scaling will change the statement of Theorem 4 accordingly to

$$\mathbb{E}_1[T_{\texttt{SCUSUM}}] \sim \frac{\max(\lambda,\lambda')}{\lambda}\frac{\log\gamma}{\lambda'\mathbb{D}_{\texttt{F}}(P_1||P_\infty)},$$

as $\gamma \to \infty$. In order to have the strongest results in Theorems 3 and 4, we must choose λ as close to λ^* as possible.

In the end, we consider a special case where pre- and post-change distributions are both multivariate Normal distributions. In this case, SCUSUM attains the asymptotic optimality in the sense of Pollak's and Lorden's metrics.

Proposition 5 (Multivariate Normal Pre- and Post-change Distributions)**.** *Assume that $X_1, \cdots, X_{\nu-1} \sim N(\boldsymbol{\theta}_0, \Sigma)$, and $X_\nu, X_{\nu+1}, \cdots \sim N(\boldsymbol{\theta}_1, \Sigma)$. Suppose $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \Theta \subset \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ are known, and $\Sigma = \sigma_c\mathbf{I}_d$ where the scalar $\sigma_c > 0$. Then the stopping rule T_{SCUSUM} achieves the asymptotic optimality of Problem (2) and Problem (5) when $\gamma \to \infty$, namely, as $\gamma \to \infty$,*

$$\begin{aligned}\mathcal{L}_{WADD}(T_{SCUSUM}) &\sim \mathcal{L}_{CADD}(T_{SCUSUM}) \\ &\sim \frac{\log\gamma}{\mathbb{D}_{KL}(N(\boldsymbol{\theta}_1,\Sigma)\|N(\boldsymbol{\theta}_0,\Sigma))},\end{aligned} \tag{14}$$

under the constraint that $\mathbb{E}_\infty[T_{SCUSUM}] \geq \gamma > 0$.

Remark 2. We note that in the above Gaussian case (where the densities are normalized), whenever

$$\lambda\mathbb{D}_{\texttt{F}}(P_1||P_\infty) < \mathbb{D}_{\texttt{KL}}(P_1||P_\infty),$$

the performance of CUSUM is superior to that of SCUSUM. However, CUSUM is not readily applicable to unnormalized models. This is a small penalty that SCUSUM pays in order to unleash its computational advantages.

5 EXPERIMENTS

In this section, we conduct extensive numerical experiments on synthetic data to compare the performance of our method with various change point detection algorithms. We repeat each experiment for 100 trials. Further details of the experimental setup and results can be found in the supplementary material.

5.1 Experimental Setup

Dataset We simulate synthetic data streams from multivariate Normal distribution (MVN), a subfamily (Yu et al., 2016) of the exponential family (EXP), and the Gauss-Bernoulli Restricted Boltzmann Machine (GB-RBM) (LeCun et al., 2006). For the exponential family, we use the Hamiltonian Monte Carlo (HMC) to generate samples from the unnormalized models. We compute the normalizing constant by numerical integration to perform CUSUM based on log-likelihood. It is worth noting that this calculation is intractable when the dimension of EXP becomes large. The samples of GB-RBM are drawn using Gibbs sampling with 1000 iterations to ensure convergence.

Baseline We evaluate the performance in terms of empirical ARL and empirical CADD, where ARL and CADD are given by $\mathbb{E}_\infty[T]$ and $\mathbb{E}_\nu[T-\nu|T \geq \nu]$, respectively. When there is no change, we expect a large value of empirical ARL; when a change occurs, we expect a small value of empirical CADD. All the results of empirical CADD and empirical ARL are reported in a log scale. We do not provide the results of CUSUM for GB-RBM because the exact log-likelihood of GB-RBM is hard to compute. In all experiments, we set the change point as $\nu = 500$. To make sure the data stream is long enough for detection schemes, we fixed the total length as 10000. The values of ARL range from 500 to 20000. Their theoretical properties have been discussed in Section 4.

We compare the performance of SCUSUM with three other methods:

- CUSUM (Page, 1955). We consider the log score-based CUSUM as a baseline. The details were discussed in Subsection 4.2.
- Scan B-statistic (Li et al., 2015, 2019). The Scan B-statistic algorithm was motivated by the B-statistic (Zaremba et al., 2013). It is defined by the kernelized maximum mean discrepancy (MMD) between sliding bootstrap blocks of the data stream. The Scan B-statistic was proved to attain the asymptotic ARL at $\mathcal{O}(e^{\tau^2})$ (Li et al., 2015, Theorem 4), while the theoretical analysis of CADD was missing.
- CALM-MMD (Cobb et al., 2022). Cobb et al. (2022) proposed a dynamic threshold-selecting scheme, named CALM, which is applicable to most two-sample tests-like change detection methods. The CALM-MMD algorithm is returned by applying the CALM procedure to the kernelized two-sample MMD statistic (Gretton et al., 2012).

For CUSUM and SCUSUM, we follow Algorithm 1. For a fixed ARL, the stopping threshold is selected by $\tau = \log(\text{ARL})$ according to Equation (7). The choice of such a λ has been discussed in Subsection 4.2.

We implement the Scan B-statistic and CALM-MMD algorithms with the code released by Cobb et al. (2022). Both of

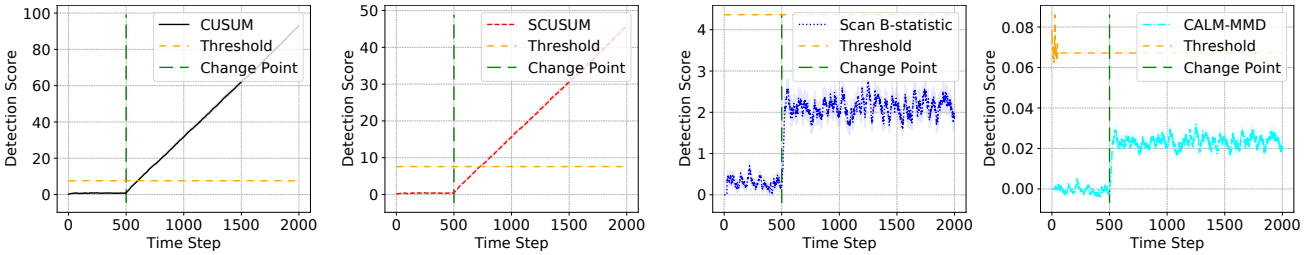


Figure 1: The results of Detection Score (before and after change) with MVN ($\epsilon_\mu = 0.3$) and ARL$= 2000$.

these are kernelized MMD-based methods where the Gaussian radial basis function (RBF) kernel, e.g. $k(x, x') = \exp(-\frac{1}{\phi^2}\|x, x'\|_2^2)$, is employed. The width of RBF is chosen by using the median heuristic, e.g., ϕ is taken to be the median of the pairwise distances between two samples. Their stopping thresholds are selected by past observations empirically, which can lead to significant miscalibration in practice, as shown by Cobb et al. (2022) and later in our numerical results.

Other than the evaluation of the trade-off between ARL and CADD, we also investigate the performance of quickest change detection in cases of slight changes, meaning that the pre- and post-change distributions are very close to each other. The closeness is measured by the magnitude of parameter drifts. Here, we run experiments by fixing the pre-change distribution and constructing the post-change distribution by perturbing the parameters of the pre-change distribution. For different families of distributions, we consider different magnitudes of perturbations.

5.2 Experimental Results

Detection Score We illustrate instantaneous detection scores at time steps in Figure 1. We control ARL to be fixed as 2000. The data streams are generated from bivariate Gaussian (MVN-ϵ_μ) with a mean drift $\epsilon_\mu = 0.3$ at time $t = 500$. We report the averaged detection scores, marked as solid lines, and standard errors, marked as shadow intervals. As presented in Figure 1, at the change point, both CUSUM and SCUSUM react immediately after the change occurs. In contrast, the detection scores of Scan B-statistic and CALM-MMD swing between the range of values 0 and 1. In this case, the two MMD-based methods fail in detection. In particular, the detection scores of CUSUM and SCUSUM monotonically increase after the change happens. However, the detection scores of Scan B-statistic and CALM-MMD maintain a stable level after the change happens. Therefore, the results demonstrate that Scan B-statistic and CALM-MMD may fail to reach the threshold even after a sufficient number of time steps.

Empirical CADD against ARL In Figure 2, we illustrate the empirical CADD against ARL in cases of bivariate Gaussian mean drifts (MVN-ϵ_μ), bivariate Gaussian co-

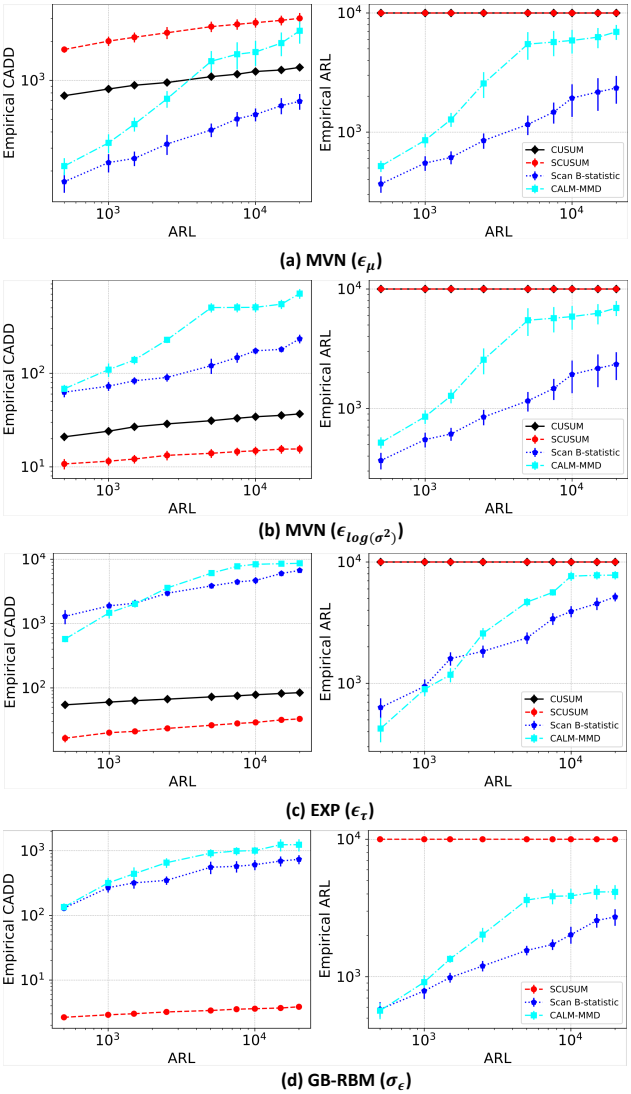


Figure 2: Empirical CADD against ARL and Empirical ARL against ARL for MVN ($\epsilon_\mu = 0.1$), MVN ($\epsilon_{\log(\sigma^2)} = 0.5$), EXP ($\epsilon_\tau = 1.0$), and GB-RBM ($\sigma_\epsilon = 0.05$).

variance drifts (MVN-$\epsilon_{\log(\sigma^2)}$), scale parameter drifts of an exponential family (EXP-ϵ_τ), and weight matrix drifts of the GB-RBM (GB-RBM-σ_ϵ), respectively. The notations ϵ_μ, $\epsilon_{\log(\sigma^2)}$, ϵ_τ, and σ_ϵ denote the magnitude of shits of the MVN mean, MVN covariance matrix, EXP scale parameter, and GB-RBM weight matrix, respectively. The results demonstrate that our proposed SCUSUM performs competitively with CUSUM in terms of empirical CADD against ARL. In particular, we see the red lines (SCUSUM) and the black lines (CUSUM) are in parallel, meaning that the empirical CADD of SCUSUM increases at a similar rate as that of CUSUM. Furthermore, SCUSUM can also outperforms CUSUM for a fixed ARL in Figures 2(b, c).

The right columns of Figure 2 illustrate empirical ARL against ARL when no change happens throughout all time steps. The results demonstrate that CUSUM and SCUSUM can successfully control the false alarm rate, while MMD-based methods fail to do so. For the Gaussian mean shifts, Scan B-statistic and CALM-MMD perform better than CUSUM and SCUSUM at low values of ARL. However, we point out that this gain is due to an out-of-control of false alarms, as illustrated in the right columns of Figure 2(a). Furthermore, MMD-based methods not only fail to control false alarms but also perform worse than CUSUM and SCUSUM, as illustrated in Figures 2(b-d).

Empirical CADD against Changes We investigate the performance of the detection methods in cases of slight changes in Figure 3, namely, pre- and post-change distributions are very close to each other. In the scenario of slight changes, CUSUM and SCUSUM perform better than MMD-based methods in Figures 3(b-d). In particular, CUSUM and SCUSUM have much smaller empirical CADD when the magnitude of changes increases. Although MMD-based methods perform better than CUSUM and SCUSUM in Figure 3(a), it is worth noting that it comes to the cost of out-of-control of false alarms as illustrated in Figure 2.

Computation We compare SCUSUM with other baselines in terms of computational costs by varying the dimensions of the EXP dataset. In Table 1, we demonstrate that when the dimension grows from $2D$ to $4D$, the run time needed for CUSUM grows significantly. It is due to the numerical integration of the exact log-likelihood calculation. Meanwhile, the run time of SCUSUM slightly grows due to the calculation of the Hyvärinen score. The run time of MMD-based methods stays consistent as the dimension grows. CALM-MMD requires a much longer run time due to its computation of candidate thresholds.

6 CONCLUSION

In this work, we proposed the SCUSUM algorithm to detect changes in unnormalized models. Our detection algorithm follows the classic CUSUM detection scheme, sharing its computational advantage of recursive implementation. We analyzed the asymptotic properties of SCUSUM in the sense of Pollak's optimality criterion. We also provided numerical results demonstrating significant performance gains and a reduction in computational complexity. Future work may relax the assumption of knowing the post-change distribution and data independence.

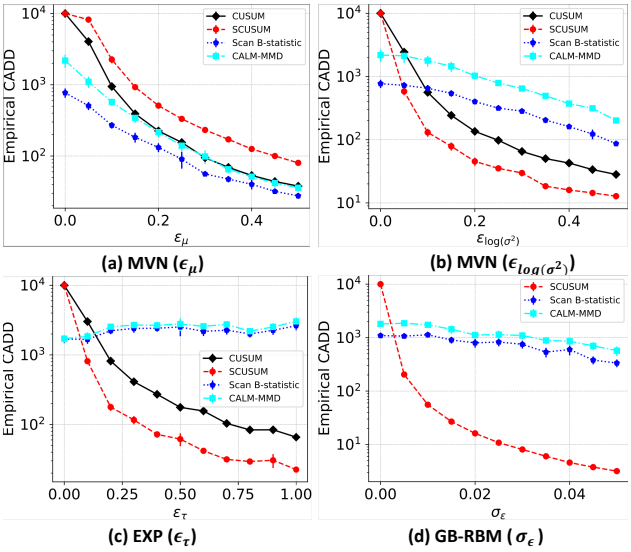


Figure 3: Empirical CADD against perturbations with ARL= 2000

Table 1: Running times (in seconds) of change detection of each trial (pre- and post-change distributions belong to the exponential family)

Detection Algorithms	$d = 1$	$d = 2$	$d = 3$	$d = 4$
CUSUM	2.4	2.9	294.8	66409.2
SCUSUM	2.2	9.1	21.0	38.4
Scan B-statistic	8.1	8.2	8.2	8.3
CALM-MMD	111.9	111.4	110.2	111.0

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

Alanqary, A., Alomar, A., and Shah, D. (2021). Change point detection via multivariate singular spectrum analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:23218–23230.

Banerjee, T., Firouzi, H., and Hero, A. O. (2018). Quickest detection for changes in maximal knn coherence of random matrices. *IEEE Trans. Signal Process.*, 66(17):4490–4503.

Chen, H. (2019). Sequential change-point detection based on nearest neighbors. *Ann. Stat.*, 47(3):1381–1407.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *Ann. Stat.*, 43(1):139–176.

Cheng, K. C., Aeron, S., Hughes, M. C., Hussey, E., and Miller, E. L. (2020). Optimal transport based change point detection and time series segment clustering. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Cobb, O., Van Looveren, A., and Klaise, J. (2022). Sequential multivariate change detection with calibrated and memoryless false detection rates. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 226–239. PMLR.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Ann. Inst. Stat. Math.*, 59(1):77–93.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2):169–183.

Doob, J. L. (1953). *Stochastic processes*, volume 7. Wiley New York.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The J. Mach. Learn. Res.*, 13(1):723–773.

Harchaoui, Z., Moulines, E., and Bach, F. (2008). Kernel change-point analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 21.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

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

Jiao, Y., Chen, Y., and Gu, Y. (2018). Subspace change-point detection: A new model and solution. *IEEE J. Sel. Top. Signal Process.*, 12(6):1224–1239.

Kawahara, Y., Yairi, T., and Machida, K. (2007). Change-point detection in time-series data based on subspace identification. In *IEEE International Conference on Data Mining (ICDM)*, pages 559–564. IEEE.

Kim, K., Shin, J., and Kim, H. (2021). Locally most powerful bayesian test for out-of-distribution detection using deep generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:14913–14924.

Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., and Coca, A. (2020). Monitoring and explainability of models in production. *arXiv preprint arXiv:2007.06299*.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. The MIT Press.

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

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. In *Predicting structured data*, volume 1. The MIT Press.

Li, S., Xie, Y., Dai, H., and Song, L. (2015). M-statistic for kernel change-point detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 28.

Li, S., Xie, Y., Dai, H., and Song, L. (2019). Scan b-statistic for kernel change-point detection. *Seq. Anal.*, 38(4):503–544.

Lorden, G. (1970). On excess over the boundary. *Ann. Math. Stat.*, 41(2):520–527.

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

Lyu, S. (2012). Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*.

Mardia, K. V., Kent, J. T., and Laha, A. K. (2016). Score matching estimators for directional distributions. *arXiv preprint arXiv:1604.08470*.

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

Moustakides, G. V., Polunchenko, A. S., and Tartakovsky, A. G. (2011). A numerical approach to performance analysis of quickest change-point detection procedures. *Stat. Sin.*, pages 571–596.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. (2018). Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*.

Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2019). Optimal nonparametric change point detection and localization. *arXiv preprint arXiv:1905.10019*.

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

Parry, M., Dawid, A. P., and Lauritzen, S. (2012). Proper local scoring rules. *Ann. Stat.*, 40(1):561–592.

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

Polunchenko, A. S. and Tartakovsky, A. G. (2010). On optimality of the shiryaev–roberts procedure for detecting a change in distribution. *Ann. Stat.*, 38(6):3445–3457.

Polunchenko, A. S. and Tartakovsky, A. G. (2012). State-of-the-art in sequential change-point detection. *Methodol. Comput. Appl.*, 14(3):649–684.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

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

Sharpnack, J., Singh, A., and Rinaldo, A. (2013). Changepoint detection over graphs with the spectral scan statistic. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 545–553. PMLR.

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.*, 18.

Tartakovsky, A. G., Pollak, M., and Polunchenko, A. S. (2012). Third-order asymptotic optimality of the generalized shiryaev–roberts changepoint detection procedures. *Theory Probab. Appl.*, 56(3):457–484.

Titsias, M. K., Sygnowski, J., and Chen, Y. (2022). Sequential changepoint detection in neural networks with checkpoints. *Stat. Comput.*, 32(2):1–19.

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

Woodroofe, M. (1982). *Nonlinear renewal theory in sequential analysis*. SIAM.

Xiao, Z., Yan, Q., and Amit, Y. (2020). Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:20685–20696.

Xie, L., Xie, Y., and Moustakides, G. V. (2020). Sequential subspace change point detection. *Seq. Anal.*, 39(3):307–335.

Xie, L., Zou, S., Xie, Y., and Veeravalli, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory (JSAIT)*, 2(2):494–514.

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

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems (NeurIPS)*, 26.

A THEORETICAL ANALYSIS

In this section, we give detailed proofs for the theoretical results presented in the main paper.

Assumption 1. $P_\infty \neq P_1$.

Assumption 2. The same mild regularity conditions[2] made by Hyvärinen and Dayan (2005) so that the Hyvärinen score is well-defined.

Assumption 3. The mild assumptions that the order of integrals and derivatives can be interchanged.

A.1 Proof of Lemma 1

Proof. Under some mild regularity conditions, Hyvärinen and Dayan (2005) proved that

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

Let $C(P)$ denote $\mathbb{E}_{X\sim P}\left[\frac{1}{2}\left\|\nabla_X \log p(X)\right\|_2^2\right]$ for any $P \in \mathcal{P}$, then

$$\mathbb{E}_\infty[\mathcal{S}_{\mathrm{H}}(X,P_\infty) - \mathcal{S}_{\mathrm{H}}(X,P_1)] = \mathbb{D}_{\mathrm{F}}(P_\infty\|P_\infty) - C(P_\infty) - \mathbb{D}_{\mathrm{F}}(P_\infty\|P_1) + C(P_\infty) = -\mathbb{D}_{\mathrm{F}}(P_\infty\|P_1),$$

and

$$\mathbb{E}_1[\mathcal{S}_{\mathrm{H}}(X,P_\infty) - \mathcal{S}_{\mathrm{H}}(X,P_1)] = \mathbb{D}_{\mathrm{F}}(P_1\|P_\infty) - C(P_1) - \mathbb{D}_{\mathrm{F}}(P_1\|P_1) + C(P_1) = \mathbb{D}_{\mathrm{F}}(P_1\|P_\infty).$$

Since $\lambda > 0$ is a constant with respect to P_1 and P_∞, the proof is complete. □

A.2 Proof of Lemma 2

Proof. Define the function $\lambda :\mapsto h(\lambda)$ given by

$$h(\lambda) \triangleq \mathbb{E}_\infty[\exp(z_\lambda(X))] - 1.$$

Observe that

$$h'(\lambda) \triangleq \frac{dh}{d\lambda}(\lambda) = \mathbb{E}_\infty[(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1))\exp(z_\lambda(X))].$$

Note that $h(0) = 0$, and $h'(0) = -\mathbb{D}_{\mathrm{F}}(P_\infty\|P_1) < 0$. Thus, there exists $\lambda > 0$ such that $h(\lambda) < 0$, and Inequality (10) is satisfied.

Next, we prove that either 1) there exists $\lambda^\star \in (0,\infty)$ such that $h(\lambda^\star) = 0$, or 2) for all $\lambda > 0$ we have $h(\lambda) < 1$.

Observe that

$$h''(\lambda) \triangleq \frac{d^2h}{d\lambda}(\lambda) = \mathbb{E}_\infty[(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1))^2\exp(z_\lambda(X))] \geq 0.$$

We claim that $h(\lambda)$ is *strictly convex*, namely $h''(\lambda) > 0$ for all $\lambda \in [0,\infty)$. Suppose $h''(\lambda) = 0$ for some $\lambda \geq 0$, we must have $S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1) = 0$ almost surely. This implies that $\mathbb{E}_\infty[(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1))] = 0$ which in turn gives $-\mathbb{D}_{\mathrm{F}}(P_\infty\|P_1) = 0$ and $P_\infty = P_1$ almost everywhere, leading to a contradiction to the assumption $P_\infty \neq P_1$. Thus, $h(\lambda)$ is *strictly convex* and $h'(\lambda)$ is *strictly increasing*.

It follows that either 1) $h(\lambda)$ have at most one global minimum in $(0,\infty)$, or 2) it is strictly decreasing in $[0,\infty)$. We recognize two cases, and we show that the second case is degenerate that is of no practical interest.

- **Case 1:** If the global minimum of $h(\lambda)$ is attained at $a \in (0,\infty)$, then $h'(a) = 0$. Since $h'(0) < 0$ and $h(0) = 0$, the global minimum $h(a) < 0$. Since $h'(\lambda)$ is *strictly increasing*, we can choose $b > a$ and conclude that $h'(\lambda) > h'(b) > h'(a) = 0$ for all $\lambda > b$. It follows that $\lim_{\lambda\to\infty} h(\lambda) = +\infty$. Combining this with the continuity of $h(\lambda)$, we conclude that $h(\lambda^*) = 0$ for some $\lambda^* \in (0,\infty)$ and any value of $\lambda \in (0,\lambda^*]$ satisfies Inequality (10).

  Note that in this case, we must have $P_\infty\left(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1) \geq c\right) > 0$, for some $c > 0$. Otherwise, we have $P_\infty\left(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1) \leq 0\right) = 1$. This implies that $P_\infty(z_\lambda(X) \leq 0) = 1$, or equivalently $\mathbb{E}_\infty[\exp(z_\lambda(X))] < 1$ for all $\lambda > 0$, and therefore leads to Case 2: $h(\lambda) < 0$ for all $\lambda > 0$. Here, $\mathbb{E}_\infty[\exp(z_\lambda(X))] \neq 1$ since $P_\infty(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1) = 0) < 1$; otherwise $P_\infty(S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1) = 0) = 1$, and then $\mathbb{E}_\infty[S_{\mathrm{H}}(X,P_\infty) - S_{\mathrm{H}}(X,P_1)] = -\mathbb{D}_{\mathrm{F}}(P_\infty\|P_1) = 0$, causing the same contradiction to $P_1 \neq P_\infty$.

[2] We refer the details to Hyvärinen and Dayan (2005).

- **Case 2:** If $h(\lambda)$ is strictly decreasing in $(0, \infty)$, then any $\lambda \in (0, \infty)$ satisfies Inequality (10). As discussed before, in this case, we must have $P_\infty\left(S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, P_1) \leq 0\right) = 1$. Equivalently, all the increments of the SCUSUM detection score are non-positive under the pre-change distribution, and $P_\infty(Z(n) = 0) = 1$ for all n. Accordingly, $\mathbb{E}_\infty[T_{SCUSUM}] = +\infty$. When there occurs change (under measure P_1), we also observe that SCUSUM can get close to detecting the change point instantaneously as λ is chosen arbitrarily large. Obviously, this case is of no practical interest.

□

A.3 Proof of Theorem 3

Proof. We follow the proof of Lai (1998, Theorem 4) to conclude the result of Theorem 3. A constructed martingale and Doob's submartingale inequality (Doob, 1953) are combined to finish the proof.

1. We first construct a non-negative martingale with mean 1 under the measure P_∞. Define a new instantaneous score function $X \mapsto \tilde{z}_\lambda(X)$ given by

$$\tilde{z}_\lambda(X) \stackrel{\Delta}{=} z_\lambda(X) + \delta,$$

where

$$\delta \stackrel{\Delta}{=} -\log\Big(\mathbb{E}_\infty\left[\exp(z_\lambda(X))\right]\Big).$$

Further define the sequence

$$\tilde{G}_n \stackrel{\Delta}{=} \exp\Big(\sum_{k=1}^{n} \tilde{z}_\lambda(X_k)\Big),\ \forall n \geq 1.$$

Suppose $X_1, X_2, \ldots$ are i.i.d according to P_∞ (when there is no change occurs). Then,

$$\mathbb{E}_\infty\left[\tilde{G}_{n+1} \mid \mathcal{F}_n\right] = \tilde{G}_n \mathbb{E}_\infty[\exp(\tilde{z}_\lambda(X_{n+1}))] = \tilde{G}_n e^{\delta} \mathbb{E}_\infty[\exp(z_\lambda(X_{n+1}))] = \tilde{G}_n,$$

and

$$\mathbb{E}_\infty[\tilde{G}_n] = \mathbb{E}_\infty\left[\exp\left(\sum_{i=1}^{n}(z_\lambda(X_i) + \delta)\right)\right] = e^{n\delta}\prod_{i=1}^{n}\mathbb{E}_\infty[\exp(z_\lambda(X_i))] = 1.$$

Thus, $\{\tilde{G}_n\}_{n\geq 1}$ is a non-negative martingale with the mean $\mathbb{E}_\infty[\tilde{G}_1] = 1$.

2. We next examine the new stopping rule

$$\tilde{T}_{\text{SCUSUM}} = \inf\left\{n \geq 1 : \max_{1\leq k\leq n}\sum_{i=k}^{n}\tilde{z}_\lambda(X_i) \geq \tau\right\},$$

where $\tilde{z}_\lambda(X_i) = z_\lambda(X_i) + \delta$. By Inequality (10), we observe that $\delta \geq 0$. By Jensen's inequality,

$$\mathbb{E}_\infty[\exp(z_\lambda(X))] \geq \exp\left(\mathbb{E}_\infty[z_\lambda(X)]\right), \tag{15}$$

with equality holds if and only if $z_\lambda(X) = c$ almost surely, where c is some constant. Suppose the equality of Equation (15) holds, then

$$-\lambda\mathbb{D}_{\text{F}}(P_1||P_\infty) = \mathbb{E}_\infty[z_\lambda(X)] = c = \mathbb{E}_1[z_\lambda(X)] = \lambda\mathbb{D}_{\text{F}}(P_\infty||P_1).$$

It follows that $0 \leq \mathbb{D}_{\text{F}}(P_\infty||P_1) = -\mathbb{D}_{\text{F}}(P_1||P_\infty) \leq 0$, which implies that $P_\infty = P_1$ almost everywhere. This leads to a contradiction to the assumption $P_\infty \neq P_1$. Thus, the inequality of Equation (15) is *strict*, and therefore $\delta < \lambda\mathbb{D}_{\text{F}}(P_\infty||P_1)$. Hence, $\tilde{T}_{\text{SCUSUM}}$ is not trivial.

Define a sequence of stopping times:

$$\begin{aligned}
&\eta_0 = 0,\\
&\eta_1 = \inf\left\{t : \sum_{i=1}^{t} \tilde{z}_\lambda(X_i) < 0\right\},\\
&\eta_{k+1} = \inf\left\{t > \eta_k : \sum_{i=\eta_k+1}^{t} \tilde{z}_\lambda(X_i) < 0\right\}, \text{ for } k \geq 1.
\end{aligned}$$

By previous discussion, $\{\tilde{G}_n\}_{n\geq 1}$ is a nonnegative martingale under P_∞ with mean 1. Then, for any k and on $\{\eta_k < \infty\}$,

$$P_\infty\left(\sum_{i=\eta_k+1}^{n} \tilde{z}_\lambda(X_i) \geq \tau \text{ for some } n > \eta_k \mid \mathcal{F}_{\eta_k}\right) \leq e^{-\tau}, \tag{16}$$

by Doob's submartingale inequality (Doob, 1953). Let

$$M \triangleq \inf\left\{k \geq 0 : \eta_k < \infty \text{ and } \sum_{i=\eta_k+1}^{n} \tilde{z}_\lambda(X_i) \geq \tau \text{ for some } n > \eta_k\right\}. \tag{17}$$

Combining Inequality (16) and Definition (17),

$$P_\infty(M \geq k+1 \mid \mathcal{F}_{\eta_k}) = 1 - P_\infty\left(\sum_{i=\eta_k+1}^{n} \tilde{z}(X_i) \geq \tau \text{ for some } n > \eta_k \mid \mathcal{F}_{\eta_k}\right) \geq 1 - e^{-\tau}, \tag{18}$$

and

$$P_\infty(M > k) = \mathbb{E}_\infty[P_\infty(M \geq k+1 \mid \mathcal{F}_{\eta_k})\mathbb{I}_{\{M\geq k\}}] = \mathbb{E}_\infty[P_\infty(M \geq k+1 \mid \mathcal{F}_{\eta_k})]P_\infty(M > k-1). \tag{19}$$

Combining Equations (19) and (18),

$$\mathbb{E}_\infty[M] = \sum_{k=0}^{\infty} P_\infty(M > k) \geq \sum_{k=0}^{\infty}(1 - e^{-\tau})^k = e^{\tau}.$$

Observe that

$$\tilde{T}_{\texttt{SCUSUM}} = \inf\left\{n \geq 1 : \sum_{i=\eta_k+1}^{n} \tilde{z}_\lambda(X_i) \geq \tau \text{ for some } \eta_k < n\right\} \geq M,$$

and $\tilde{T}_{\texttt{SCUSUM}} \leq T_{\texttt{SCUSUM}}$. We conclude that $\mathbb{E}_\infty[T_{\texttt{SCUSUM}}] \geq \mathbb{E}_\infty[\tilde{T}_{\texttt{SCUSUM}}] \geq \mathbb{E}_\infty[M] \geq e^{\tau}$.

□

A.4 Proof of Theorem 4

We first introduce a technical definition in order to apply (Woodroofe, 1982, Corollary 2.2.) to the proof of Theorem 4.

Definition 4. A distribution P on the Borel sets of $(-\infty, \infty)$ is said to be *arithmetic* if and only if it concentrates on a set of points of the form $\pm nd$, where $d > 0$ and $n = 1, 2, \ldots$.

Remark 3. Any probability measure that is absolutely continuous with respect to the Lebesgue measure is non-arithmetic.

Proof. Consider the random walk that is defined by

$$Z'(n) = \sum_{i=1}^{n} z_\lambda(X_i), \text{ for } n \geq 1.$$

We examine another stopping time that is given by

$$T'_{\texttt{SCUSUM}} \triangleq \inf\{n \geq 1 : Z'(n) \geq \tau\}.$$

Next, for any τ, define R_τ on $\{T'_{\texttt{SCUSUM}} < \infty\}$ by

$$R_\tau \triangleq Z'(T'_{\texttt{SCUSUM}}) - \tau.$$

R_τ is the excess of the random walk over a stopping threshold τ at the stopping time $T'_{\texttt{SCUSUM}}$. Suppose the change point $\nu = 1$, then $X_1, X_2, \ldots,$ are i.i.d. following the distribution P_1. Let μ and σ^2 respectively denote the mean $\mathbb{E}_1[z_\lambda(X)]$ and the variance $\text{Var}_1[z_\lambda(X)]$. Note that

$$\mu = \mathbb{E}_1[z_\lambda(X)] = \lambda \mathbb{D}_{\text{F}}(P_1 \| P_\infty) > 0,$$

and

$$\sigma^2 = \text{Var}_1[z_\lambda(X)] = \mathbb{E}_1[z_\lambda(X)^2] - \left(\lambda \mathbb{D}_{\text{F}}(P_1 \| P_\infty)\right)^2.$$

Under the mild regularity conditions given by Hyvärinen and Dayan (2005),

$$\mathbb{E}_1[\mathcal{S}_{\text{H}}(X, P_\infty)]^2 < \infty, \text{ and}$$
$$\mathbb{E}_1[\mathcal{S}_{\text{H}}(X, P_1)]^2 < \infty.$$

It implies that $\mathbb{E}_1[z_\lambda(X)^2] < \infty$ if λ is chosen appropriately, e.g. λ satisfy Inequality (10) and λ is not arbitrary large. Therefore, by Lorden (1970, Theorem 1),

$$\sup_{\tau \geq 0} \mathbb{E}_1[R_\tau] \leq \frac{\mathbb{E}_1[(z_\lambda(X)^+)^2]}{\mathbb{E}_1[z_\lambda(X)]} \leq \frac{\mu^2 + \sigma^2}{\mu},$$

where $z_\lambda(X)^+ = \max(z_\lambda(X), 0)$. Additionally, P_1 must be non-arithmetic in order to have Hyvärinen scores well-defined. Hence, by Woodroofe (1982, Corollary 2.2.),

$$\mathbb{E}_1[T'_{\texttt{SCUSUM}}] = \frac{\tau}{\mu} + \frac{\mathbb{E}_1[R_\tau]}{\mu} \leq \frac{\tau}{\mu} + \frac{\mu^2 + \sigma^2}{\mu^2}, \ \forall \tau \geq 0.$$

Observe that for any n, $Z'(n) \leq Z(n)$, and therefore $T_{\texttt{SCUSUM}} \leq T'_{\texttt{SCUSUM}}$. Thus,

$$\mathbb{E}_1[T_{\texttt{SCUSUM}}] \leq \mathbb{E}_1[T'_{\texttt{SCUSUM}}] \leq \frac{\tau}{\mu} + \frac{\mu^2 + \sigma^2}{\mu^2}, \ \forall \tau \geq 0. \tag{20}$$

By Theorem 4, we select $\tau = \log \gamma$ to satisfy the constraint $\mathbb{E}_\infty[T_{\texttt{SCUSUM}}] \geq \gamma > 0$. Plugging it back to Equation (20), we conclude that, as $\gamma \to \infty$,

$$\mathbb{E}_1[T_{\texttt{SCUSUM}}] \sim \frac{\log \gamma}{\mu} = \frac{\log \gamma}{\lambda \mathbb{D}_{\text{F}}(P_1 || P_\infty)},$$

to complete the proof.

Due to the stopping scheme of SCUSUM, the expected time $\mathbb{E}_\nu[T_{\texttt{SCUSUM}} - \nu | T_{\texttt{SCUSUM}} \geq \nu]$ is independent of the change point ν (This is obvious, and the same property for CUSUM has been shown by Xie et al. (2021)). Let $\nu = 1$, and we have

$$\mathcal{L}_{\texttt{CADD}}(T_{\texttt{SCUSUM}}) = \mathbb{E}_1[T_{\texttt{SCUSUM}}] - 1.$$

Thus, we conclude that

$$\mathcal{L}_{\texttt{CADD}}(T_{\texttt{SCUSUM}}) \sim \frac{\log \gamma}{\lambda \mathbb{D}_{\text{F}}(P_1 \| P_\infty)}.$$

Similar arguments applies for $\mathcal{L}_{\texttt{WADD}}(T_{\texttt{SCUSUM}})$. □

A.5 Proof of Proposition 5

Proof. By direct computation, it can see that

$$z_\lambda(X) = \lambda\left(-\frac{1}{2}(X-\boldsymbol{\theta}_0)^T\Sigma^{-2}(X-\boldsymbol{\theta}_0) + \frac{1}{2}(X-\boldsymbol{\theta}_1)^T\Sigma^{-2}(X-\boldsymbol{\theta}_1)\right),$$

where Σ^{-2} is a short notation for $\Sigma^{-1}\cdot\Sigma^{-1}$. Then

$$\mathbb{E}_\infty[\exp(z_\lambda(X))] = \int_{X\in\mathcal{X}} \frac{1}{\sqrt{2\pi}\det(\Sigma)}\exp\left(-\frac{1}{2}(X-\boldsymbol{\theta}_0)^T\Sigma^{-1}(X-\boldsymbol{\theta}_0) + \frac{\lambda}{2}(X-\boldsymbol{\theta}_0)^T\Sigma^{-2}(X-\boldsymbol{\theta}_0) - \frac{\lambda}{2}(X-\boldsymbol{\theta}_1)^T\Sigma^{-2}(X-\boldsymbol{\theta}_1)\right)dX.$$

The above integral can be calculated to be

$$\mathbb{E}_\infty[\exp(z_\lambda(X))] = \exp(-\lambda^2(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-3}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1) + \lambda(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-2}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)).$$

Clearly, $\mathbb{E}_\infty[\exp(z_\lambda(X))] = 1$ if

$$\lambda = \frac{(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-2}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)}{(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-3}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)}.$$

The Fisher divergence and KL divergence between two Normal distributions can be calculated by

$$\mathbb{D}_{\mathrm{F}}(\mathcal{N}(\boldsymbol{\theta}_1,\Sigma)||\mathcal{N}(\boldsymbol{\theta}_0,\Sigma)) = (\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-2}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1),$$

and

$$\mathbb{D}_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\theta}_1,\Sigma)||\mathcal{N}(\boldsymbol{\theta}_0,\Sigma)) = (\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-1}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1),$$

respectively. Thus

$$\frac{\lambda\mathbb{D}_{\mathrm{F}}(P_1||P_\infty)}{\mathbb{D}_{\mathrm{KL}}(P_1||P_\infty)} = \frac{[(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-2}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)]^2}{[(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-3}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)][(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)^T\Sigma^{-1}(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)]}.$$

Let $\{v_1, v_2, \cdots, v_d\}$ denote an orthonormal basis of eigenvectors of Σ, corresponding to its eigenvalues $\{\sigma_1, \sigma_2, \cdots, \sigma_d\}$. We can write $(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1)$ in this orthonormal basis as

$$(\boldsymbol{\theta}_0-\boldsymbol{\theta}_1) = \sum_{k=1}^d c_k v_k.$$

Then, it follows from direct calculations that

$$\frac{\lambda\mathbb{D}_{\mathrm{F}}(P_1||P_\infty)}{\mathbb{D}_{\mathrm{KL}}(P_1||P_\infty)} = \frac{(\sum_{k=1}^d \frac{c_k^2}{\sigma_k^2})^2}{(\sum_{k=1}^d \frac{c_k^2}{\sigma_k^3})(\sum_{k=1}^d \frac{c_k^2}{\sigma_k})}.$$

Applying the Cauchy-Schwarz inequality, we have

$$\lambda\mathbb{D}_{\mathrm{F}}(P_1||P_\infty) \le \mathbb{D}_{\mathrm{KL}}(P_1||P_\infty),$$

with equality if and only if all the eigenvalues σ_i, $i = 1, 2, \cdots, d$ for $c_i \neq 0$ are equal. In particular, in the case when Σ is a scalar matrix, $\lambda\mathbb{D}_{\mathrm{F}}(P_1||P_\infty) = \mathbb{D}_{\mathrm{KL}}(P_1||P_\infty)$, and thus CUSUM and SCUSUM both achieve the same asymptotic performance. □

B EXPERIMENTS

B.1 Experimental Setup

Multivariate Normal Distribution (MVN) We consider the multivariate normal distribution. Let $\boldsymbol{\mu}$ and Σ respectively denote the mean and the covariance matrix. The corresponding score function is calculated by

$$S_{\mathrm{H}}(X,P) = \frac{1}{2}(X-\boldsymbol{\mu})^T\Sigma^{-2}(X-\boldsymbol{\mu}) - \mathrm{tr}(\Sigma^{-1}),$$

where the operator $\mathrm{tr}(\cdot)$ takes the trace of matrix.

We consider the pre-change distribution with mean $\boldsymbol{\mu} = (0,0)^T$ and covariance matrix $\Sigma = \begin{pmatrix} 1, & 0.5 \\ 0.5, & 1 \end{pmatrix}$. For the post-change distribution, we first investigate the scenario of mean shifts by fixing the covariance matrix $\Sigma = \begin{pmatrix} 1, & 0.5 \\ 0.5, & 1 \end{pmatrix}$ and assigning post-change means $\boldsymbol{\mu} = (0,0)^T + \epsilon_\mu$, where $+$ here is element-wise plus and ϵ_μ is the perturbations of $\boldsymbol{\mu}$. We take values of ϵ_μ from 0 to 0.5 with step size 0.05. Next, we consider the case of covariance shifts. In this scenario, we fix the post-change mean as $\boldsymbol{\mu} = (0,0)^T$ and assign post-change covariance by $\Sigma = \begin{pmatrix} 1, & 0.5 \\ 0.5, & 1 \end{pmatrix} \circ \exp(\epsilon_{\log(\sigma^2)})$, where $\circ$ denotes the element-wise product and $\epsilon_{\log(\sigma^2)}$ denotes the element-wise perturbations of the covariance matrix. To make the perturbed covariance matrix positive-definite, we perturb the log of each component of the covariance matrix. We take the value of $\epsilon_{\log(\sigma^2)}$ vary from 0.05 to 0.5 by a step size 0.05.

Exponential Family (EXP) We consider a subfamily of the Exponential family belonging to pairwise interaction graphical models (Yu et al., 2016). Let P_τ and p_τ respectively represent the distribution and the associated PDF of the random variable X. The PDF is formulated as

$$p_\tau(X) = \frac{1}{Z} \exp\left\{ -\tau \left(\sum_{i=1}^{d} x_i^4 + \sum_{1 \le i \le d, i \le j \le d} x_i^2 x_j^2 \right) \right\},$$

where $\tau \in \mathcal{T} \subset \mathbb{R}^+$ is the model parameter and Z is the normalizing constant. The associated Hyvarinen score function is given by

$$S_{\mathrm{H}}(X, P_\tau) = \frac{1}{2} \sum_{i=1}^{d} \left(\frac{\partial}{\partial x_i} \log p_\tau(X) \right)^2 + \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i} \log p_\tau(X),$$

where

$$\frac{\partial}{\partial x_i} \log p_\tau(X) = -\tau \left(4x_i^3 + 2 \sum_{1 \le i \le d, i \le j \le d} x_i x_j^2 \right), \text{ and}$$
$$\frac{\partial^2}{\partial x_i} \log p_\tau(X) = -\tau \left(12x_i^2 + 2 \sum_{1 \le i \le d, i \le j \le d} x_j^2 \right).$$

We consider the pre-change distribution with $\tau = 1$ and post-change distribution with $\tau = 1 + \epsilon_\tau$, where ϵ_τ denotes the perturbations of the scale parameter τ. We take values of ϵ_τ from 0.1 to 2.0 by a step size 0.1.

Restricted Boltzmann Machine (RBM) The RBM (LeCun et al., 2006) is a generative graphical model defined on a bipartite graph of hidden and visible variables. We consider the Gauss-Bernoulli RBM (GB-RMB), which has binary-valued hidden variables $H = (h_1, \ldots, h_{d_h})^T \in \{0,1\}^{d_h}$, real-valued visible variables $X = (x_1, \ldots, x_{d_x})^T \in R^{d_x}$, and the joint PDF

$$p(X, H) = \frac{1}{Z} \exp\left\{ -\left(\frac{1}{2} \sum_{i=1}^{d_x} \sum_{j=1}^{d_h} \frac{x_i}{\sigma_i} W_{ij} h_j + \sum_{i=1}^{d_x} b_i x_i + \sum_{j=1}^{d_h} c_j h_j - \frac{1}{2} \sum_{i=1}^{d_x} \frac{x_i^2}{\sigma_i^2} \right) \right\},$$

where model parameters $\theta = (\mathbf{W}, \mathbf{b}, \mathbf{c})$ and Z is the normalizing constant. We set $\sigma_i = 1$ for all $i = 1, \ldots, d_x$.

Let P_θ and p_θ respectively represent the distribution and the associated PDF of the visible variable X. Its PDF can be written as $p_\theta(X) = \sum_{h \in \{0,1\}^{d_h}} p_\theta(X, H) = \frac{1}{Z} \exp\{-F_\theta(X)\}$, where $F_\theta(X)$ is the free energy given by

$$F_\theta(X) = \frac{1}{2} \sum_{i=1}^{d_x} (x_i - b_i)^2 - \sum_{j=1}^{d_h} \mathrm{Softplus}\left(\sum_{i=1}^{d_x} W_{ij} x_i + b_j \right).$$

The $\mathrm{Softplus}$ function is defined as $\mathrm{Softplus}(z) \triangleq \log(1 + \exp(z))$ with a default scale parameter $\beta = 1$.The corresponding

Hyvärinen score is given by

$$S_{\mathrm{H}}(X, P_\theta) = \sum_{i=1}^{d_x} \left[\frac{1}{2} \left(x_i - b_i + \sum_{j=1}^{d_h} W_{ij} \phi_j \right)^2 + \sum_{j=1}^{d_h} W_{ij}^2 \phi_j \left(1 - \phi_j\right) - 1 \right],$$

where $\phi_j \triangleq \mathrm{Sigmoid}(\sum_{i=1}^{d_x} W_{ij} x_i + b_j)$. The Sigmoid function is defined as $\mathrm{Sigmoid}(z) \triangleq (1 + \exp(-z))^{-1}$.

The pre-change distribution is with the parameters $\mathbf{W} = \mathbf{W}_0$, $\mathbf{b} = \mathbf{b}_0$, and $\mathbf{c} = \mathbf{c}_0$, where each component of $\mathbf{W}_0$, $\mathbf{b}_0$, and $\mathbf{c}_0$ is randomly drawn from the standard Normal distribution $\mathcal{N}(0, 1)$. For the post-change distribution, we assign the parameters $\mathbf{W} = \mathbf{W}_0 + \epsilon_{\mathbf{W}}$, $\mathbf{b} = \mathbf{b}_0$, and $\mathbf{c} = \mathbf{c}_0$. Here, we only consider the shift of weight matrix $\mathbf{W}$, denoted as $\epsilon_{\mathbf{W}}$. Each component of $\epsilon_{\mathbf{W}}$ is drawn from $\mathcal{N}(0, \sigma_\epsilon^2)$. We let σ_ϵ take values from 0.005 to 0.1 with step size 0.005.

B.2 Experimental Results

B.2.1 The Effect of Hyper-parameters

The implementation of SCUSUM requires a pre-selected multiplier λ. 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. This is formally claimed in Theorem 5, which states that the expected detection delay increases linearly with respect to the stopping threshold at the rate of $\frac{1}{\lambda \mathbb{D}_{\mathrm{F}}(P_1 \| P_\infty)}$. However, a larger value of λ also causes SCUSUM to stop prematurely when no change occurs, leading to a larger false alarm rate. Hence, instead of the trivial case discussed in the proof of Lemma 2, the value of λ cannot be arbitrarily large. The value λ needs to satisfy Inequality (10) in order to control the false alarm rate (by Theorem 3). Lemma 2 proved that there exists such a λ, and it can even make the equality of (10) hold. Therefore, it is possible to use m past observations to determine the value of λ to guarantee the theoretical performance of SCUSUM.

In practice, we choose λ as the positive root of the function $\lambda \to \tilde{h}(\lambda)$, given by

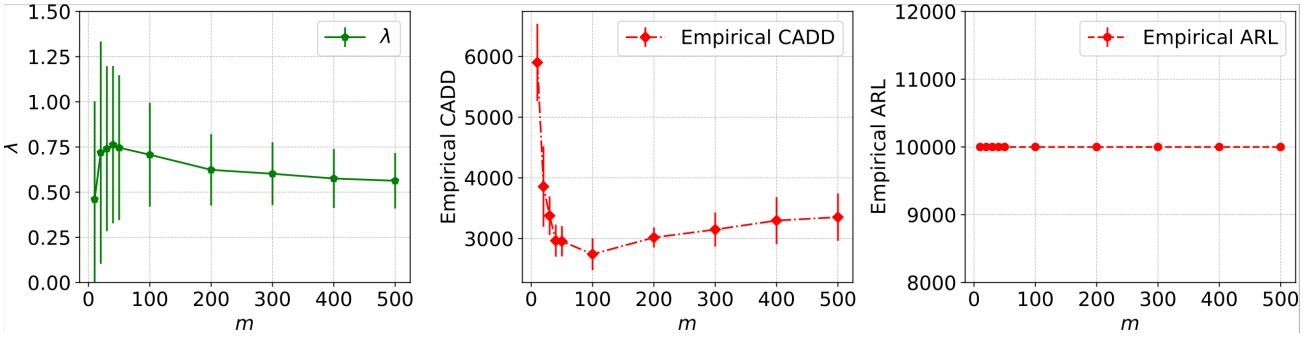
$$\tilde{h}(\lambda) = \frac{1}{m} \sum_{i=1}^{m} [\exp(z_\lambda(X_{i-m}))] - 1. \tag{21}$$

Different samples of past observations may determine different values of λ, which can cause the inconsistent performance of SCUSUM. We next investigate this problem through numerical simulations. In Figure 4 (a) to (d), the data streams are generated from MVNs with $\epsilon_\mu = 0.1$, MVNs with $\epsilon_{\log(\sigma^2)} = 0.5$, EXPs with $\epsilon_\tau = 1$, and GB-RBMs with $\sigma_\epsilon = 0.05$. The first columns of Figure 4 illustrate values of determined λ varying from the size of past observations. The second (and the third) columns of Figure 4 report the empirical CADD (respectively the empirical ARL) of SCUSUM varying from the size of past observations. We report all values in averages over 100 random runs with error bars.

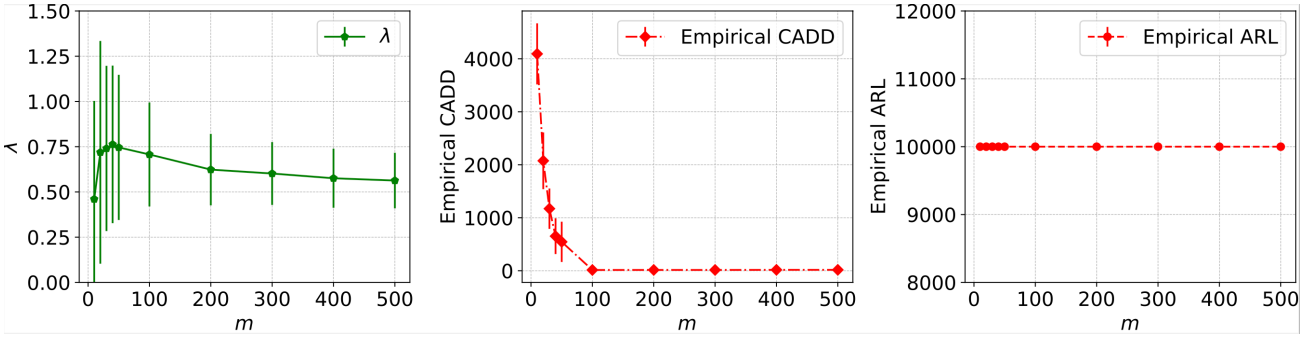
As Figure 4 demonstrates, as long as m is large enough, the value of λ is not too sensitive to different samples. In particular, when $m > 100$, we see small standard errors in Figue 4(a)-(c). Accordingly, the performance of SCUSUM in terms of the empirical CADD tends to be stable. Note that in the case of GB-RBM (as shown by Figure 4(d)), we take $\lambda = 1$ when $m < 300$. It is because we can not numerically find the positive root of Equation (11) given a small size of past observations. Finally, as shown in Column 3 of Figue 4, the empirical ARL is always under control.

B.2.2 Detection Scores

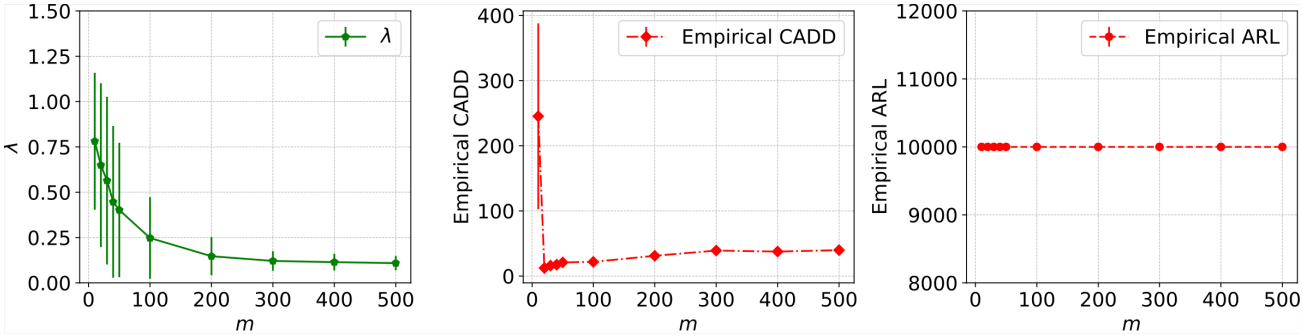
In this section, we add additional numerical results to illustrate instantaneous detection scores. We control ARL to be fixed as 2000. From Figure 5 to Figure 8, the data streams are generated from MVN-ϵ_μ, MVN-$\epsilon_{\log(\sigma^2)}$, EXP-$\epsilon_\tau$, and GB-RBM-$\sigma_\epsilon$, respectively. The change happens at time $t = 500$. We report the averaged detection scores, marked as solid lines, and standard error, marked as shadow intervals.



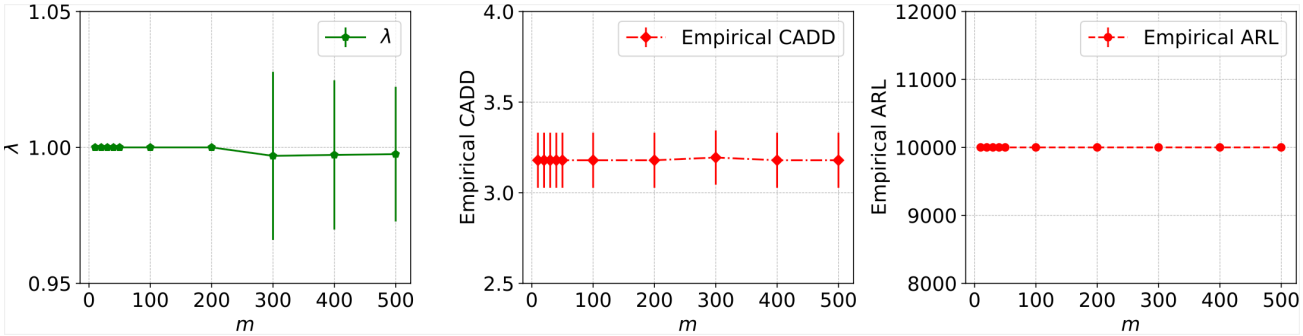
(a) MVN ($\epsilon_{\mu} = 0.1$)



(b) MVN ($\epsilon_{log(\sigma^2)} = 0.5$)



(c) EXP ($\epsilon_{\tau} = 1$)



(d) GB-RBM ($\sigma_{\epsilon} = 0.05$)

Figure 4: Column 1: λ versus m; Column 2: Empirical CADD versus m; Column 3: Empirical ARL versus m.

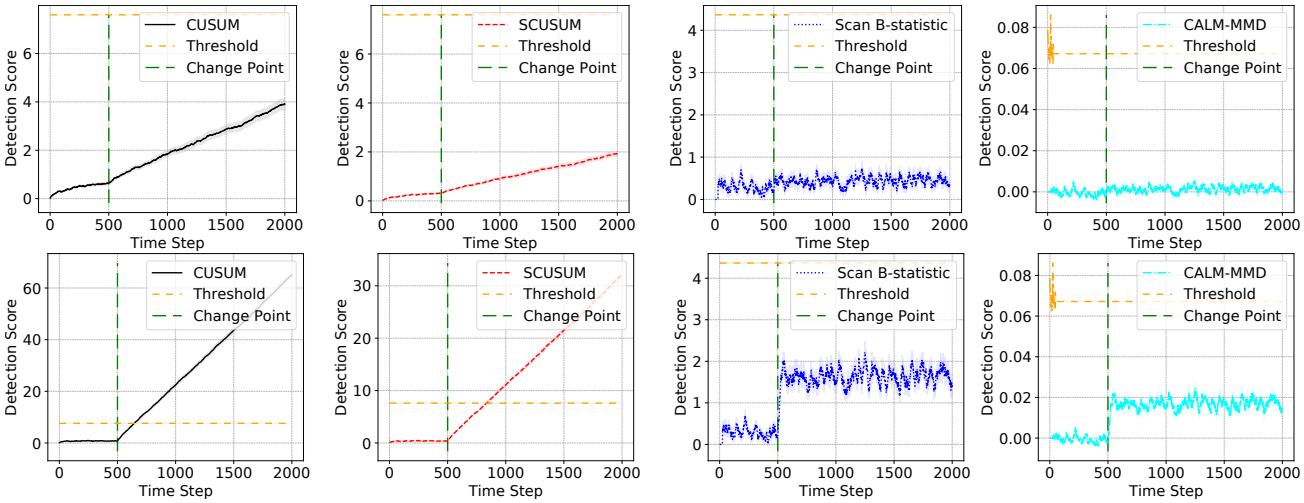


Figure 5: The results of Detection Score (before and after change) with MVN mean shifts (MVN-ϵ_μ) at $t = 500$ with ARL$=$ 2000. Top: $\epsilon_\mu = 0.05$; Bottom: $\epsilon_\mu = 0.25$.

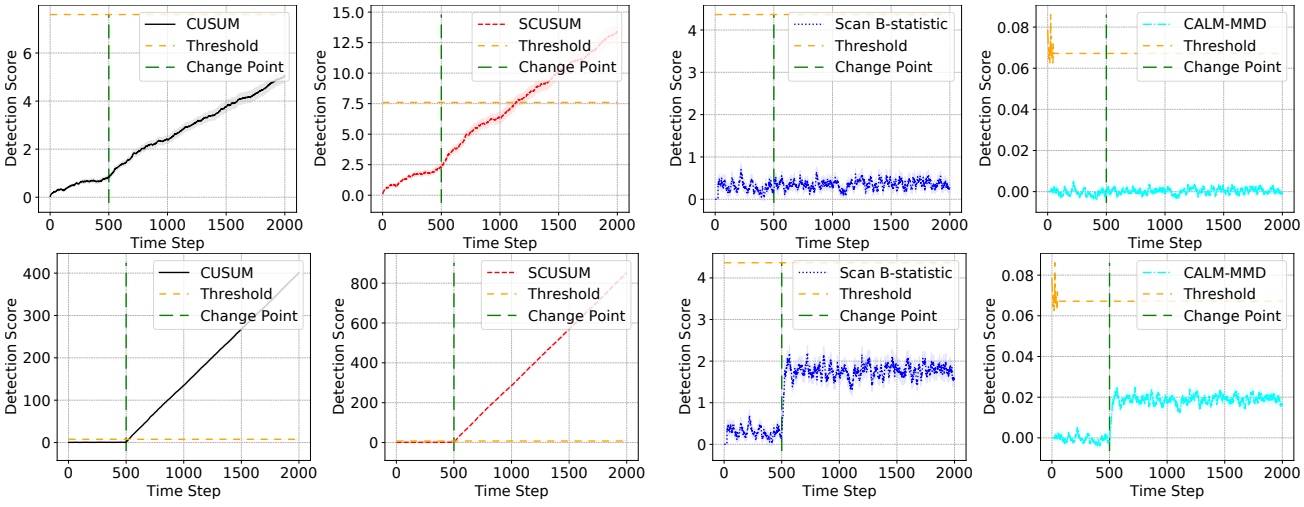


Figure 6: The results of Detection Score (before and after change) with MVN covariance shifts (MVN-$\epsilon_{\log(\sigma^2)}$) at $t = 500$ with ARL$=$ 2000. Top: $\epsilon_{\log(\sigma^2)} = 0.05$; Bottom: $\epsilon_{\log(\sigma^2)} = 0.5$.

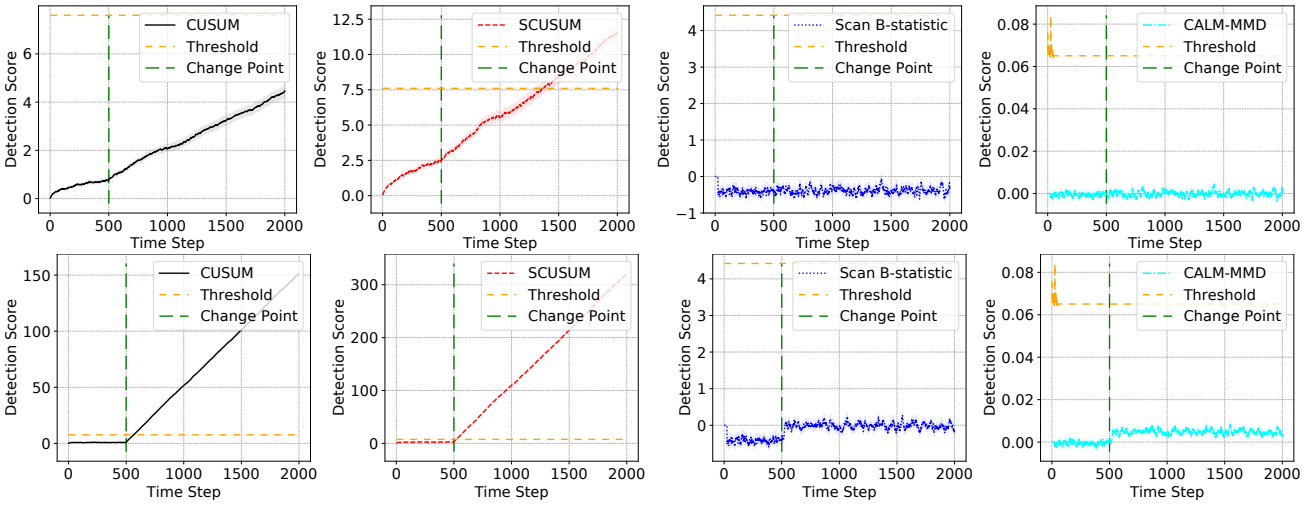


Figure 7: The results of Detection Score (before and after change) with the EXP scalar shifts (EXP-ϵ_τ) at $t = 500$ with ARL$= 2000$. Top: $\epsilon_\tau = 0.1$; Bottom: $\epsilon_\tau = 1.0$.

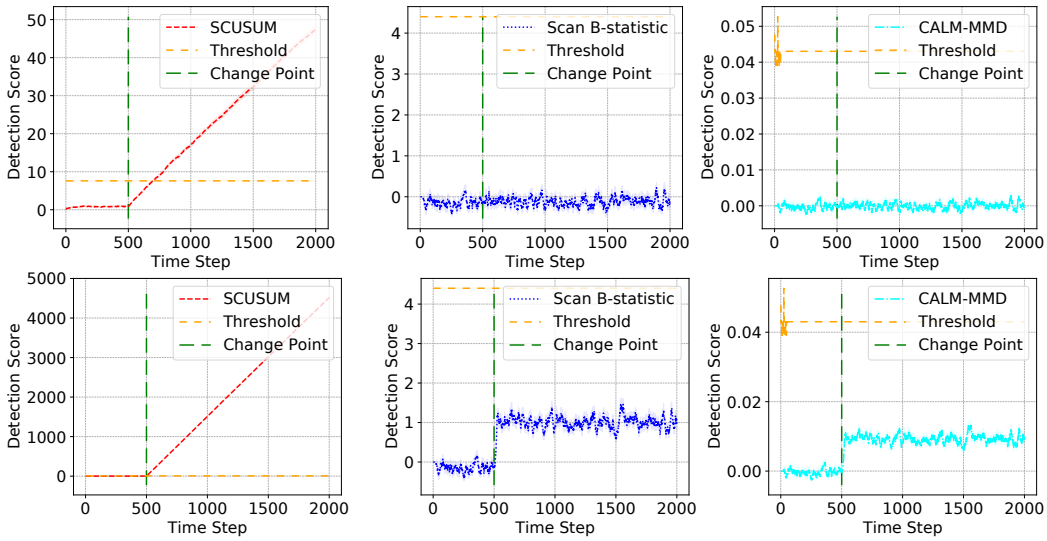


Figure 8: The results of Detection Score (before and after change) with GB-RBM weight matrix shifts (GB-RBM-σ_ϵ) at $t = 500$ with ARL$= 2000$. Top: $\sigma_\epsilon = 0.005$; Bottom: $\sigma_\epsilon = 0.05$.