Robust Quickest Change Detection for Unnormalized Models (Supplementary Material)

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

A LIKELIHOOD RATIO-BASED ROBUST CUSUM ALGORITHM

In this section, we review the result in Unnikrishnan et al. [2011] on classical robust quickest change detection. Let p_∞ and p_1 be the density functions of pre- and post-change distributions. If the post-change law is known, then given the data stream $\{X_n\}_{n\geq 1}$, the stopping rule of the likelihood ratio-based CUSUM algorithm is defined by

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

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

$$\begin{aligned} &\Lambda(0) = 0, \\ &\Lambda(n) \triangleq \left(\Lambda(n-1) + \log \frac{p_1(X_n)}{p_\infty(X_n)}\right)^+, \forall n \geq 1, \end{aligned} \tag{2}$$

which leads to a computationally efficient stopping scheme (if the densities p_1 and p_∞ are precisely known). In Moustakides [1986], it is shown that the CUSUM algorithm is exactly optimal, for every fixed constraint γ, for Lorden's problem. As pointed out in Lai [1998], the algorithm is also asymptotically optimal for Pollak's problem. In Lorden [1971] and Lai [1998], the asymptotic performance of the CUSUM algorithm is also characterized. Specifically, it is shown as $\gamma \to \infty$.

$$\mathcal{L}_{\texttt{WADD}}(T_{\texttt{CUSUM}}) \sim \mathcal{L}_{\texttt{CADD}}(T_{\texttt{CUSUM}}) \sim \frac{\log \gamma}{\mathbb{D}_{\texttt{KL}}(P_1 \| P_\infty)}. \tag{3}$$

Here $\mathbb{D}_{\texttt{KL}}(p_1 \| p_\infty)$ is the Kullback-Leibler divergence between the post-change density p_1) and pre-change distribution p_∞:

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

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 $c \mapsto h(c)$.

The CUSUM algorithm can successfully detect a change in law from p_1 to p_∞ because

$$\begin{aligned} &\int_x \log \frac{p_1(x}{p_\infty(x)} p_1(x) dx = \mathbb{D}_{\texttt{KL}}(P_1 \| P_\infty) > 0 \\ &\int_x \log \frac{p_1(x}{p_\infty(x)} p_\infty(x) dx = -\mathbb{D}_{\texttt{KL}}(P_\infty \| P_1) < 0. \end{aligned} \tag{4}$$

Thus, the mean of the increment of $\Lambda(n)$ in (2) before the change is negative, and after the change is positive.

If the post-change density p_1 is not known and assumed to belong to a family $\mathcal{G}_1$, then the test is designed using the least favorable distribution. Specifically, in Unnikrishnan et al. [2011], it is assumed that there is a density $q_1 \in \mathcal{G}_1$ such that for every $p_1 \in \mathcal{G}_1$,

$$\log \frac{q_1(X)}{p_\infty(X)}\bigg|_{X\sim q_1} \quad \prec \quad \log \frac{q_1(X)}{p_\infty(X)}\bigg|_{X\sim p_1}. \tag{5}$$

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

Here the notation $\prec$ is used to denote stochastic dominance: if W and Y are two random variables, then $W \prec Y$ if

$$P(Y \geq t) \geq P(W \geq t), \quad \text{for all } t \in (-\infty, \infty).$$

If such a density q_1 exists in the post-change family, then the robust CUSUM is defined as the CUSUM test with q_1 used as the post-change density. Such a test is exactly optimal for the problem of Lorden [1971] under additional assumptions on the smoothness of densities, and asymptotically optimal for the problem in Pollak [1985]. We refer the reader to Unnikrishnan et al. [2011] for a more precise optimality statement.

We note that in the literature on quickest change detection, the issue of the unknown post-change model has also been addressed by using a generalized likelihood ratio (GLR) test or a mixture-based test. While these tests have strong optimality properties, they are computationally even more expensive than the robust test described above; see Lorden [1971], Lai [1998], Tartakovsky et al. [2014].

As discussed in the introduction, the robust CUSUM algorithm discussed above may have two major drawbacks: 1) Due to the complicated characterization of the least favorable distribution q_1 (5), it may be hard to identify in high-dimensional models. 2) The robust CUSUM is a likelihood ratio-based test and is thus computationally expensive to implement for high-dimensional models.

In Section 4 of the main paper, we propose the RSCUSUM algorithm to mitigate these issues.

1. The RSCUSUM algorithm is based on Hyvärinen score (Hyvärinen [2005]) and is invariant to normalizing constants. This makes it computationally efficient for high-dimensional models which are often only learnable within a normalizing constant.
2. We defined the notion of least favorable distribution differently in our paper. For us, the least favorable distribution has the least Fisher divergence with respect to the pre-change model. We also provided an efficient computational method to identify the least favorable distribution.

B PROOFS

The theoretical analysis for delay and false alarms is analogous to that of analysis from Wu et al. [2023]. We give complete proofs here for completeness.

B.1 PROOF OF LEMMA 4.3

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

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

Observe that

$$h'(\lambda) \triangleq \frac{dh}{d\lambda}(\lambda) = \mathbb{E}_\infty[(S_{\mathrm{H}}(X, P_\infty) - S_{\mathrm{H}}(X, Q_1)) \exp(z_\lambda(X))].$$

Note that $h(0) = 0$, and $h'(0) = -\mathbb{D}_{\mathrm{F}}(P_\infty \| Q_1) < 0$. 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 $h(\lambda) < 0$.

Observe that

$$h''(\lambda) \triangleq \frac{d^2h}{d\lambda}(\lambda) = \mathbb{E}_\infty[(S_{\mathrm{H}}(X, P_\infty) - S_{\mathrm{H}}(X, Q_1))^2 \exp(z_\lambda(X))] \geq 0.$$

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 $S_{\mathrm{H}}(X, P_\infty) - S_{\mathrm{H}}(X, Q_1) = 0$ almost surely. This implies that $\mathbb{E}_\infty[(S_{\mathrm{H}}(X, P_\infty) - S_{\mathrm{H}}(X, Q_1))] = 0$ which in turn gives $-\mathbb{D}_{\mathrm{F}}(P_\infty \| Q_1) = 0$ and $P_\infty = Q_1$ almost everywhere, leading to a contradiction to the assumption $P_\infty \notin \mathcal{G}_1$. Thus, $h(\lambda)$ is *strictly convex* and $h'(\lambda)$ is *strictly increasing*.

Here, we recognize two cases: either 1) $h(\lambda)$ have at most one global minimum in $(0, \infty)$, or 2) it is strictly decreasing in $[0, \infty)$. We will 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 $P_\infty \left(S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, Q_1) \geq c\right) > 0$, for some $c > 0$. Otherwise, we have $P_\infty \left(S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, Q_1) \leq 0\right) = 1$. 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 $P_\infty(S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, Q_1) = 0) < 1$; otherwise $P_\infty(S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, Q_1) = 0) = 1$, and then $\mathbb{E}_\infty[S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, Q_1)] = -\mathbb{D}_{\text{F}}(P_\infty \| Q_1) = 0$, causing the same contradiction to $P_\infty \notin \mathcal{G}_1$.

- **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 $P_\infty \left(S_{\text{H}}(X, P_\infty) - S_{\text{H}}(X, Q_1) \leq 0\right) = 1$. Equivalently, all the increments of the RSCUSUM detection score are non-positive under the pre-change distribution, and $P_\infty(Z(n) = 0) = 1$ for all n. Accordingly, $\mathbb{E}_\infty[T_{\mathit{RSCUSUM}}] = +\infty$. When there occurs change (under measure Q_1), we also observe that RSCUSUM can get close to detecting the change point instantaneously as λ is chosen arbitrarily large. Obviously, this case is of no practical interest.

$\square$

B.2 PROOF OF THEOREM 4.4

Proof. We follow the proof of Lai [1998][Theorem 4] to conclude the result of Theorem 4.4. 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
$$\tilde{z}_\lambda(X) \triangleq z_\lambda(X) + \delta,$$
where
$$\delta \triangleq -\log\Big(\mathbb{E}_\infty\left[\exp(z_\lambda(X))\right]\Big).$$
Further define the sequence
$$\tilde{G}_n \triangleq \exp\Big(\sum_{k=1}^{n} \tilde{z}_\lambda(X_k)\Big), \ \forall n \geq 1.$$
Suppose $X_1, X_2, \dots$ are i.i.d according to P_∞ (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, under the measure P_∞, $\{\tilde{G}_n\}_{n\geq 1}$ is a non-negative martingale with the mean $\mathbb{E}_\infty[\tilde{G}_1] = 1$.
2. We next examine the new stopping rule
$$\tilde{T}_{\text{RSCUSUM}} = \inf\left\{n \geq 1 : \max_{1\leq k\leq n} \sum_{i=k}^{n} \tilde{z}_\lambda(X_i) \geq \tau\right\},$$
where $\tilde{z}_\lambda(X_i) = z_\lambda(X_i) + \delta$. By Inequality (10), we observe that $\delta \geq 0$. By Jensen's inequality,
$$\mathbb{E}_\infty[\exp(z_\lambda(X))] \geq \exp\left(\mathbb{E}_\infty[z_\lambda(X)]\right), \tag{6}$$
with equality holds if and only if $z_\lambda(X) = c$ almost surely, where c is some constant. Suppose the equality of Equation (6) holds, then
$$-\lambda \mathbb{D}_{\text{F}}(Q_1 || P_\infty) = \mathbb{E}_\infty[z_\lambda(X)] = c = \mathbb{E}_1[z_\lambda(X)] = \lambda \mathbb{D}_{\text{F}}(P_\infty || Q_1).$$

It follows that $0 \leq \mathbb{D}_{\mathrm{F}}(P_\infty||Q_1) = -\mathbb{D}_{\mathrm{F}}(Q_1||P_\infty) \leq 0$, which implies that $P_\infty \notin \mathcal{G}_1$ almost everywhere. This leads to a contradiction to the assumption $P_\infty \notin \mathcal{G}_1$. Thus, the inequality of Equation (6) is *strict*, and therefore $\delta < \lambda \mathbb{D}_{\mathrm{F}}(P_\infty||Q_1)$. Hence, $\tilde{T}_{\texttt{RSCUSUM}}$ 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 $\{\eta_k < \infty\}$,

$$P_\infty\left(\sum_{i=\eta_k+1}^{n} \tilde{z}_\lambda(X_i) \geq \tau \text{ for some } n > \eta_k \mid \mathcal{F}_{\eta_k}\right) \leq e^{-\tau}, \tag{7}$$

by Doob's submartingale inequality [Doob, 1953]. Let

$$M \triangleq \inf\left\{k \geq 0 : \eta_k < \infty \text{ and } \sum_{i=\eta_k+1}^{n} \tilde{z}_\lambda(X_i) \geq \tau \text{ for some } n > \eta_k\right\}. \tag{8}$$

Combining Inequality (7) and Definition (8),

$$P_\infty(M \geq k+1 \mid \mathcal{F}_{\eta_k}) = 1 - P_\infty\left(\sum_{i=\eta_k+1}^{n} \tilde{z}(X_i) \geq \tau \text{ for some } n > \eta_k \mid \mathcal{F}_{\eta_k}\right) \geq 1 - e^{-\tau}, \tag{9}$$

and

$$P_\infty(M > k) = \mathbb{E}_\infty[P_\infty(M \geq k+1 \mid \mathcal{F}_{\eta_k})\mathbb{I}_{\{M\geq k\}}] = \mathbb{E}_\infty[P_\infty(M \geq k+1 \mid \mathcal{F}_{\eta_k})]P_\infty(M > k-1). \tag{10}$$

Combining Equations (10) and (9),

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

Observe that

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

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

□

B.3 PROOF OF THEOREM 4.5

We first introduce a technical definition in order to apply Woodroofe [1982][Corollary 2.2.] to the proof of Theorem 4.5.

Definition B.1. 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 B.2. 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

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

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

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

R_τ is the excess of the random walk over a stopping threshold τ at the stopping time $T'_{\texttt{RSCUSUM}}$. Suppose the change point $\nu = 1$, then $X_1, X_2, \ldots,$ are i.i.d. following the distribution Q_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

$$\mu = \mathbb{E}_1[z_\lambda(X)] = \lambda(\mathbb{D}_{\text{F}}(P_1 \| P_\infty) - \mathbb{D}_{\text{F}}(P_1 \| Q_1)) > 0,$$

and

$$\sigma^2 = \text{Var}_1[z_\lambda(X)] = \mathbb{E}_1[z_\lambda(X)^2] - \left(\lambda(\mathbb{D}_{\text{F}}(P_1 \| P_\infty) - \mathbb{D}_{\text{F}}(P_1 \| Q_1))\right)^2.$$

Under the mild regularity conditions given by Hyvärinen [2005],

$$\begin{aligned} &\mathbb{E}_1[\mathcal{S}_{\text{H}}(X, P_\infty)]^2 < \infty, \text{ and} \\ &\mathbb{E}_1[\mathcal{S}_{\text{H}}(X, Q_1)]^2 < \infty. \end{aligned}$$

It implies that $\mathbb{E}_1[z_\lambda(X)^2] < \infty$ if λ is chosen appropriately, e.g. λ satisfy Inequality (14) 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, Q_1 must be non-arithmetic in order to have Hyvärinen scores well-defined. Hence, by Woodroofe [1982] Corollary 2.2.,

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

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

$$\mathbb{E}_1[T_{\texttt{RSCUSUM}}] \leq \mathbb{E}_1[T'_{\texttt{RSCUSUM}}] \leq \frac{\tau}{\mu} + \frac{\mu^2 + \sigma^2}{\mu^2}, \ \forall \tau \geq 0. \tag{11}$$

By Theorem 4, we select $\tau = \log \gamma$ to satisfy the constraint $\mathbb{E}_\infty[T_{\texttt{RSCUSUM}}] \geq \gamma > 0$. Plugging it back to Equation (11), we conclude that, as $\gamma \to \infty$,

$$\mathbb{E}_1[T_{\texttt{RSCUSUM}}] \sim \frac{\log \gamma}{\mu} = \frac{\log \gamma}{\lambda(\mathbb{D}_{\text{F}}(P_1 \| P_\infty) - \mathbb{D}_{\text{F}}(P_1 \| Q_1))},$$

to complete the proof.

Due to the stopping scheme of RSCUSUM, the expected time $\mathbb{E}_\nu[T_{\texttt{RSCUSUM}} - \nu | T_{\texttt{RSCUSUM}} \geq \nu]$ 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

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

Thus, we conclude that

$$\mathcal{L}_{\texttt{CADD}}(T_{\texttt{RSCUSUM}}) \sim \frac{\log \gamma}{\lambda(\mathbb{D}_{\text{F}}(P_1 \| P_\infty) - \mathbb{D}_{\text{F}}(P_1 \| Q_1))}.$$

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

B.4 SELECTION OF APPROPRIATE MULTIPLIER

It is worth noting that although results of our core results hold for a pre-selected λ that satisfied the condition discussed in Lemma 4.3. The effect of choosing any other λ' amounts to the scaling of all the increments of RSCUSUM 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 4.4 can be modified to be written as

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

for any $\lambda' > 0$. It is easy to see that this scaling will change the statement of Theorem 4.5 accordingly to

$$\mathbb{E}_1[T_{\text{RSCUSUM}}] \sim \frac{\max(\lambda, \lambda')}{\lambda} \frac{\log\gamma}{\lambda'(\mathbb{D}_{\text{F}}(P_1\|P_\infty) - \mathbb{D}_{\text{F}}(P_1\|Q_1))},$$

as $\gamma \to \infty$. In order to have the strongest results in Theorems 4.4 and 4.5, we must choose λ as close to λ^* as possible.

C EXPERIMENTAL DETAILS

C.1 SYNTHETIC DATASET

We consider the parametric family $\mathcal{P} = \{G_\theta :\ \theta \in \Theta\}$, and a set of basis elements $\mathcal{P}_m = \{P_1, \ldots, P_m\}, \forall P_i \in \mathcal{P}$. We set $m = 4$ for synthetic simulations. The uncertainty class of post-change distribution (pre-change distribution respectively) is given by

$$\mathcal{G}_1 = \left\{\sum_{i=1}^m \alpha_i P_i :\ \sum_{i=1}^m \alpha_i = 1,\ \forall\ \alpha_i \geq 0\right\},$$
$$\mathcal{G}_\infty = \{P_\infty :\ P_\infty \in \mathcal{P},\ P_\infty \notin \mathcal{G}_1\}.$$

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

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

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

For the scenario of MVN$_m$, we think the covariance matrix V is a constant for any distribution in the parametric family. The pre-change distribution $P_\infty = \mathcal{N}(\mu_*, V_*)$, where

$$\boldsymbol{\mu}_\star = (0, 0), \quad \text{and} \quad V_\star = \begin{pmatrix} 1, & 0.5 \\ 0.5, & 1 \end{pmatrix}.$$

The set $\mathcal{P}_m = \{\mathcal{N}(\boldsymbol{\mu}_j, V_j),\ j = 1, \ldots, m\}$, where

$$\boldsymbol{\mu}_j = (\epsilon_j, \epsilon_j), \quad \text{and} \quad V_j = \begin{pmatrix} 1, & 0.5 \\ 0.5, & 1 \end{pmatrix}.$$

We take the value of ϵ_1 ($\epsilon_j, j = 2, 3, 4$ respectively) as 0.5 ($0.6, 0.8, 1.0$ respectively) for P_1 ($P_j, j = 2, 3, 4$ respectively).

For the scenario of MVN$_c$, we consider both the mean and covariance matrix as the parameter. Again, we consider the pre-change distribution $P_\infty = \mathcal{N}(\mu_*, V_*)$, and the set $\mathcal{P}_m = \{\mathcal{N}(\boldsymbol{\mu}_j, V_i),\ j = 1, \ldots, m\}$. Here,

$$\boldsymbol{\mu}_j = (\epsilon_j, \epsilon_j), \quad \text{and} \quad V_j = \begin{pmatrix} 1, & 0.5 \\ 0.5, & 1 \end{pmatrix} \circ \exp(\delta_j),$$

where $\circ$ denotes the element-wise product and $\epsilon_{\log(\sigma^2)}$ denotes the element-wise perturbations of the covariance matrix. We take the value of δ_j (respectively $\delta_j, j = 2, 3, 4$) as 0.1 ($0.2, 0.8, 1.0$ respectively) for P_1 ($P_j, j = 2, 3, 4$ respectively). To make the perturbed covariance matrix positive-definite, we perturb the log of each component of the covariance matrix.

Table 1: EDD versus ARL for RSCUSUM and RCUSUM on Multivariate Gaussian Case

Perturbation/ARL		100	200	400	800	1500	3000
0.5	RSCUSUM	11.2552	12.6664	16.9057	20.3400	22.7026	27.3190
	RCUSUM	11.4017	12.8748	16.8437	20.2776	22.6781	27.2831
0.6	RSCUSUM	8.5636	9.5218	13.1102	15.2747	16.5815	19.8648
	RCUSUM	8.6460	9.5817	12.9797	15.2196	16.5526	19.7900
1	RSCUSUM	4.0894	4.5327	6.0542	7.1984	7.8237	9.4318
	RCUSUM	4.1259	4.5658	6.0447	7.1551	7.8026	9.3947
2	RSCUSUM	1.4053	1.6268	2.2620	2.7546	3.0592	3.6752
	RCUSUM	1.4290	1.6393	2.2516	2.7393	3.0481	3.6684

Exponential Family (EXP) We consider the Exponential family with the associated PDF given by

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

where $\theta = (\tau, \mu)$. The associated Hyvarinen score function is calculated by

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

where

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

$$\frac{\partial^2}{\partial x_i}\log P_\theta(X) = -\tau\left(12(x_i-\mu)^2 + 2\sum_{1\le i\le d, i\le j\le d}(x_j-\mu)^2\right).$$

We consider the pre-change distribution P_∞ with $\tau_\star = 1$ and $\mu_* = 0$. The post-change distribution basis elements are constructed with $\tau = \tau_* + \epsilon_j$ and $\mu = \mu_* + \delta_j$. Here, ϵ_j (δ_j respectively) denotes the perturbations of the scale parameter τ (the location parameter μ respectively) for each P_j, $j = 1, 2, 3, 4$. We take values of ϵ_j as $1.0, 2.0, 8.0, 10.0$, and values of δ_j as $0.01, 0.02, 0.08, 0.1$.

Gauss-Bernoulli Restricted Boltzmann Machine (RBM) As introduced in Subsection the main paper, we consider the RBM mode with the PDF given by $p_\theta(X) = \sum_{h\in\{0,1\}^{d_h}} p_\theta(X, H) = \frac{1}{Z_\theta}\exp\{-F_\theta(X)\}$, where $F_\theta(X)$ is the free energy given by

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

We compute the corresponding Hyvärinen score in a closed form

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

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

The pre-change distribution P_∞ is with the parameters $\mathbf{W} = \mathbf{W}_*$, $\mathbf{b} = \mathbf{b}_*$, and $\mathbf{c} = \mathbf{c}_*$, where each component of $\mathbf{W}_*$, $\mathbf{b}_*$, and $\mathbf{c}_*$ is randomly drawn from the standard Normal distribution $\mathcal{N}(0, 1)$. For the post-change distribution basis elements, we assign the parameters $\mathbf{W}_j = \mathbf{W}_* \oplus \epsilon_j$, $\mathbf{b}_j = \mathbf{b}_*$, and $\mathbf{c}_j = \mathbf{c}_*$. Here, we only consider shifts of weight matrix $\mathbf{W}$. We let ϵ_j take values from $0.001, 0.002, 0.008, 0.01$ for P_j, $j = 1, 2, 3, 4$.

References

Joseph L Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

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.

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

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.

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

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

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

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.

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