Robust Change Detection for Large-Scale Data Streams

Ruizhi Zhang
Department of Statistics, University of Nebraska-Lincoln,
Lincoln, Nebraska, USA

Yajun Mei
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, Georgia, USA

Jianjun Shi
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract: Robust change-point detection for large-scale data streams has many real-world applications in industrial quality control, signal detection, biosurveillance. Unfortunately, it is highly non-trivial to develop efficient schemes due to three challenges: (1) the unknown sparse subset of affected data streams, (2) the unexpected outliers, and (3) computational scalability for real-time monitoring and detection. In this article, we develop a family of efficient real-time robust detection schemes for monitoring large-scale independent data streams. For each data stream, we propose to construct a new local robust detection statistic called L_α-CUSUM statistic that can reduce the effect of outliers by using the Box-Cox transformation of the likelihood function. Then the global scheme will raise an alarm based upon the sum of the shrinkage transformation of these local L_α-CUSUM statistics so as to filter out unaffected data streams. In addition, we propose a new concept called *false alarm breakdown point* to measure the robustness of online monitoring schemes and propose a *worst-case detection efficiency score* to measure the detection efficiency when the data contain outliers. We then characterize the breakdown point and the efficiency score of our proposed schemes. Asymptotic analysis and numerical simulations are conducted to illustrate the robustness and efficiency of our proposed schemes.

Keywords: Robustness; Breakdown Point; Change Detection; Large-scale Data.

Subject Classifications: 62L15; 60G40.

Address correspondence to Ruizhi Zhang, Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, 68583 USA; E-mail: rzhang35@unl.edu

1. Introduction

Robust statistics have been extensively studied in the offline context when the entire data set is available for decision-making and is contaminated with outliers, e.g., robust estimation [1, 2], robust hypothesis testing [3, 4], and robust regression [5, 6]. Also, see the classical books, [7] or [8], for literature review. In this paper, we propose to develop robust methods in the context of sequential change-point detection when one is interested in detecting sparse, persistent smaller changes in large-scale data streams under the contamination of transient larger outliers. The problem of robust monitoring large-scale data streams in the presence of outliers occurs in many real-world applications such as industrial quality control, biosurveillance, key infrastructure, or internet traffic monitoring, in which sensors are deployed to constantly monitor the changing environment, see [9],[10],[11]. Unfortunately, it is highly non-trivial to develop efficient, robust real-time monitoring schemes or algorithms due to three challenges: (1) the sparsity, where only a few unknown data streams might be affected; (2) the robustness, where we are interested in detecting smaller persistent changes, not the larger transient outliers; and (3) the computational scalability, where the algorithms can be implemented recursively to make real-time decisions.

In the literature of sequential change-point detection for a large number of data streams, to the best of our knowledge, while the sparsity issue has been investigated, no research has been done on the robustness issue. To be more specific, the sparsity has been first addressed by [12] using a semi-Bayesian approach and later by [13] using shrinkage-estimation-based schemes. [14] developed asymptotic optimality theory for large-scale independent Gaussian data streams. Unfortunately, all these methods are sensitive to outliers since they are based on the likelihood function of specific parametric models (e.g. Gaussian) of the observations. Meanwhile, regarding the robustness issue, research is available for monitoring one- or low- dimensional streaming data such as rank-based method in [15, 16], kernel-based method in [17]. However, these nonparametric methodologies generally lose detection efficiency under specific parametric or semi-parametric models. By considering the worst-case of the outlier distribution, [18] formulated the problem of finding the optimal robust change detection procedure by solving a minimax problem. However, the resulting optimal test is based on the least-favorable-pair distributions of two uncertainty sets, which depends on the information of outliers. More importantly, it is unclear how to extend their method from monitoring a single data stream to monitoring multiple data streams when we also need to deal with the sparsity issue in which there is uncertainty on the subset of affected data streams.

In this paper, we develop efficient real-time monitoring schemes that are able to robustly detect smaller persistent changes in the presence of larger transient outliers when online monitoring of large-scale data streams. From the methodology viewpoint, our proposed schemes are semi-parametric and extend two contemporary concepts to the context of online monitoring of data streams: (i) L_q-likelihood [19, 20] for robustness, and (ii) the sum-shrinkage technique [21, 22] for sparsity. These allow us to develop statistically efficient and computationally simple schemes that can be implemented recursively over time for robust real-time monitoring of a large number of data streams. Moreover, we also extend the concept of breakdown in the offline robust statistics [23] to the sequential change-point detection context and conduct the false alarm breakdown point analysis, which turns out to be useful for the choices of tuning parameters in our proposed schemes.

We should point out that our contribution is not on the optimality theory but on the asymp-

totic properties of our proposed schemes that include the classical CUSUM-based procedures as a special case. Our research makes four contributions in the statistics field by combining robust statistics with sequential change-point detection for large-scale data streams. First, our proposed method is robust to infrequent outliers as well as the uncertainty of affected data streams. Second, our proposed method can be implemented recursively and distributed via parallel computing and thus is suitable for real-time monitoring over a long time period. Third, inspired by the concept of breakdown point [23] in the offline robust statistics, we propose a novel concept of false alarm breakdown point to quantify the robustness of any online monitoring schemes and show that our proposed schemes indeed have much larger false alarm breakdown point than the classical CUSUM-based schemes. Finally, from the mathematical viewpoint, we use Chebyshev's inequality to derive non-asymptotic lower bounds on the average run length of false alarm of our proposed methods. The non-asymptotic results hold regardless of dimensionality and allow us to provide a deep insight into the effect of high-dimensionality in change-point detection under the modern asymptotic regime when the dimension or the number of data streams goes to ∞.

The remainder of this article is organized as follows. In Section 2, we start with problem formulations and model assumptions. In Section 3, we introduce our proposed family of robust monitoring schemes. In Section 4, the properties of the detection efficiency of our proposed schemes and the guideline to choose tuning parameters in our proposed schemes are provided. Then, we investigate the robustness of our proposed methods by conducting breakdown point analysis in Section 5. Simulation results are presented in Section 6. In Section 7, we conclude our paper with a few remarks The proofs of our main theorems are postponed to Appendix.

2. Problem Formulation

Suppose we are monitoring K independent data streams in a system.

$$\begin{aligned} \text{Data Stream } 1: &\quad X_{1,1}, X_{1,2}, \cdots \\ \text{Data Stream } 2: &\quad X_{2,1}, X_{2,2}, \cdots \\ \cdots &\quad \cdots \\ \text{Data Stream } K: &\quad X_{K,1}, X_{K,2}, \cdots . \end{aligned} \tag{2.1}$$

Under the classical change-point detection model for monitoring multi-streams (e.g., [12, 14, 24, 21, 22]), one assumes that the data $X_{k,n}$'s are initially independent and identically distributed (i.i.d.) with probability density function (pdf) $f_0(x)$. At some unknown time $\nu \geq 1$, an undesired event occurs, and change the distributions of m out of K data streams, i.e., the affected local streams $X_{k,n}$'s have another distribution $f_1(x)$ when $n \geq \nu$. The objective is to raise an alarm as soon as possible once a change occurs. Here, we refer to this classical model as the idealized model.

In this paper, we investigate the change-point detection problem under Tukey-Huber's gross error model. As mentioned in the introduction, we want to raise an alarm as quickly as possible if there is a persistent distribution change on the data, but we prefer to take observations without any actions if there are only transient outliers. Mathematically, we assume the distribution of data $X_{k,n}$ might be changed from h_0 to h_1 at some change time ν, the h_0 and h_1 are the Tukey-Huber's gross error model of the mixture densities

$$h_0(x) = (1-\epsilon)f_0(x) + \epsilon g_0(x), \quad h_1(x) = (1-\epsilon)f_1(x) + \epsilon g_1(x), \tag{2.2}$$

where $\epsilon \in [0, 1)$ is referred to as the contamination/outlier ratio, g_0 and g_1 are the (unknown) outlier distributions. Denote by $\mathbf{P}_{h_0}^{(\infty)}$ and $\mathbf{E}_{h_0}^{(\infty)}$ the probability measure and expectation when the data $X_{k,n}$'s are i.i.d. with the density h_0 when no change occurs, and denote by $\mathbf{P}_{h_1}^{(\nu)}$ and $\mathbf{E}_{h_1}^{(\nu)}$ the same when the change occurs at time ν and m out of K streams $X_{k,n}$'s have the post-change distribution h_1.

As in the classical sequential change-point problem, a statistical procedure under our setting is defined as a stopping time T that represents the time when we raise an alarm to declare that a change has occurred. Here T is an integer-valued random variable, and the decision $\{T = t\}$ is based only on the observations in the first t time steps. To evaluate the performance of the detection procedure T under Tukey-Huber's gross error model when the outlier distributions g_0 and g_1 in (2.2) are unknown, we first assume the average run length to false alarm of the procedure T is controlled under the idealized model. That is, we assume that the procedure T is designed to satisfy the false alarm constraint

$$\mathbf{E}_{f_0}^{(\infty)}(T) \geq \gamma, \tag{2.3}$$

for some pre-specified value $\gamma > 0$. We then investigate the robustness and the detection efficiency of the monitoring procedure under the gross error model in (2.2).

First, we propose quantifying the robustness of a monitoring procedure T under the gross error model in (2.2) by borrowing the concept of breakdown point analysis from the offline robust statistics literature. To be more specific, we propose to define a new concept called false alarm breakdown point, which characterizes the minimal percentage of outliers that can make the false alarm rate under the gross error model $h_0 = (1 - \epsilon)f_0(x) + \epsilon g_0(x)$ very different from that under the idealized model f_0.

The false alarm breakdown point $\epsilon^*(T)$ of a family of monitoring schemes $T(b)$'s is defined as

$$\epsilon^*(T) = \inf\{\epsilon \geq 0 : \inf_{h_0 \in \hbar_{0,\epsilon}} \log(\mathbf{E}_{h_0}^{(\infty)} T(b_\gamma)) = o(\log \gamma)\}, \tag{2.4}$$

where $\mathbf{E}_{f_0}^{(\infty)}(T(b_\gamma)) \sim \gamma$ as $\gamma \to \infty$, and the set $\hbar_{0,\epsilon}$ is the ϵ-contaminated distribution density class of the idealized model $f_0(x)$ for given $\epsilon \in [0, 1)$:

$$\hbar_{0,\epsilon} = \{h | h = (1 - \epsilon)f_0 + \epsilon g, g \in \mathcal{G}\}, \tag{2.5}$$

and $\mathcal{G}$ denotes the class of all probability densities of the data $X_{k,n}$.

Roughly speaking, the false alarm breakdown point characterizes the minimal percentage of outliers that can make the designed average run length to false alarm γ unreliable. Thus, a scheme with larger breakdown points is more robust.

Second, we quantify the detection efficiency of the monitoring procedure T under the gross error model in (2.2). For that purpose, recall that under the Lorden's minimax criteria [25], the worst-case detection delay under h_1 is defined as

$$\mathbf{D}_{h_1}(T) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbf{E}_{h_1}^{(\nu)} \left((T - \nu + 1)^+ \middle| \mathcal{F}_{\nu-1} \right). \tag{2.6}$$

Here $\mathcal{F}_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]})$ denotes past global information at time ν. $X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1})$ is past local information for the k-th data stream. However, since the outlier

distribution g_1 is unknown, we propose to define two quantities on detection efficiency: one is *asymptotic efficiency score* defined by

$$\text{AE}(T, \epsilon; g_0, g_1) = \lim_{\gamma \to \infty} \frac{\log(\mathbf{E}_{h_0}^{(\infty)}(T(b_\gamma)))}{\mathbf{D}_{h_1}(T(b_\gamma))}, \tag{2.7}$$

and the other is the *worst-case asymptotic efficiency score* defined by

$$\text{WAE}(T, \epsilon) = \inf_{g_0 \in \mathcal{G}, g_1 \in \mathcal{G}} \text{AE}(T, \epsilon; g_0, g_1) = \lim_{\gamma \to \infty} \frac{\inf_{g_0 \in \mathcal{G}} \left[\log(\mathbf{E}_{h_0}^{(\infty)}(T(b_\gamma)))\right]}{\sup_{g_1 \in \mathcal{G}} [\mathbf{D}_{h_1}(T(b_\gamma))]}. \tag{2.8}$$

In both definitions, b_γ is a threshold of $T = T(b_\gamma)$ so that $\mathbf{E}_{f_0}^{(\infty)}(T(b_\gamma)) \sim \gamma$. Clearly, when the data contain outliers, the procedure with a larger asymptotic efficiency score implies more efficiency in detecting the persistent change. Note that the definition of the asymptotic efficiency in (2.7) depends on the outlier distributions g_0 and g_1, which are unknown in practice, but the worst-case detection efficiency $\text{WAE}(T, \epsilon)$ in (2.8) measures the worst case among all set $\mathcal{G}$ of outlier distributions g_0 and g_1. Note when we are monitoring a single data stream, i.e., the dimension $K = 1$, the optimal procedure that maximizes $\text{WAE}(T, \epsilon)$ is a CUSUM procedure constructed by a least-favorable-pair g_0^*, g_1^*, as shown in [18]. However, the problem of finding the optimal procedure that minimizes $\text{WAE}(T, \epsilon)$ becomes more complicated when the dimension K is large and the set of affected data streams is unknown.

In this paper, our objective is to develop a family of efficient, robust monitoring schemes that have a large breakdown point $\epsilon^*(T)$ in (2.4) and a large worst-case asymptotic efficiency score $\text{WAE}(T, \epsilon)$ in (2.8) subject to the constraints that this family of schemes satisfy the false alarm constraint in (2.3) under the idealized pre-change distribution f_0.

3. Our proposed method

In this section, we will present our proposed schemes. At the high-level, our proposed schemes include two components: (i) robust monitoring each local data stream individually in parallel, and then (ii) combining local detection statistics to make an online global-level decision. For the purpose of easy understanding, we split the presentation of our proposed schemes into two subsections, and each subsection focuses on one component of the proposed scheme.

3.1. Robust local statistics

For the k^{th} data stream, we propose to define a new local L_α-CUSUM statistic:

$$W_{\alpha,k,n} = \max\left(W_{\alpha,k,n-1} + \frac{[f_1(X_{k,n})]^\alpha - [f_0(X_{k,n})]^\alpha}{\alpha}, 0\right), \tag{3.1}$$

for $n \geq 1$, and $W_{\alpha,k,0} = 0$. Here $\alpha \geq 0$ is a tuning parameter that can control the tradeoff between statistical efficiency and robustness under the gross error model in (2.2) and its suitable choice will be discussed later.

The motivation of our L_α-CUSUM statistic in (3.1) is as follows. Recall that when locally monitoring the single k^{th} data stream $X_{k,n}$ with a possible local distribution change from f_0 to

f_1, the generalized likelihood ratio test becomes the classical CUSUM statistic $W^*_{k,n}$, which has a recursive form:

$$W^*_{k,n} = \max_{1\le\nu<\infty} \log \frac{\prod_{i=1}^{\nu-1} f_0(X_{k,i}) \prod_{i=\nu}^{n} f_1(X_{k,i})}{\prod_{i=1}^{n} f_0(X_{k,i})} = \max\Big(W^*_{k,n-1} + \log \frac{f_1(X_{k,n})}{f_0(X_{k,n})},\ 0\Big). \quad (3.2)$$

The CUSUM statistic enjoys nice optimality properties when all models are fully correctly specified [26], but unfortunately it is very sensitive to the outliers as in all other likelihood based methods in offline statistics. One recent idea in offline robust statistics is to replace the log-likelihood statistic $\log f(X)$ by L_α-likelihood statistic $([f(X)]^\alpha - 1)/\alpha$ for some $\alpha > 0$, see [19],[20]. At the high-level, L_α-likelihood statistic $\frac{[f(X)]^\alpha - 1}{\alpha}$ is always bounded below by $-1/\alpha$ whereas the log-likelihood statistic $\log f(X)$ could go to $-\infty$. Thus, the impact of outliers is bounded for the L_α-likelihood statistic but unbounded for the log-likelihood statistic. Moreover, as $\alpha \to 0$, the L_α-likelihood function converges to the log-likelihood statistic, and thus it keeps statistical efficiencies when α is small. Here we apply this idea to develop our L_α-CUSUM statistic. More rigorous robust properties will be discussed later in Section 5.

3.2. Efficient global monitoring statistics

With local L_α-CUSUM statistics $W_{\alpha,k,n}$ in (3.1) for each local stream, it is important to fuse these local statistics together smartly so as to address the sparsity issue. Here we propose to combine these local statistics together via the sum-shrinkage technique in [21], i.e., we raise a global-level alarm at time

$$N_\alpha(b) = \inf\left\{ n \ge 1 : \sum_{k=1}^{K} h(W_{\alpha,k,n}) \ge b \right\}, \quad (3.3)$$

where $h(\cdot) \ge 0$ are some suitable shrinkage transformation functions, and $b > 0$ is a pre-specified constant. Intuitively, the shrinkage functions $h(\cdot)$'s in (3.3) play the role of dimension reduction by automatically filtering out those non-changing local data streams and by keeping only those local streams that might provide information about the changing event. This will allow us to improve the detection power in the sparsity scenario when only a few local features are involved in the change.

For the purpose of illustration, here we focus on two kinds of shrinkage functions: one is the soft-thresholding function $h(x) = \max\{x - d, 0\}$, and the other is the order-thresholding function $h(x) = x\mathbf{1}\{x \ge w_{(r)}\}$, where $w_{(r)}$ is the r-th largest statistic of $w_1, \cdots, w_K$. Then the corresponding two global monitoring schemes are defined by

$$N^{(soft)}_\alpha(b,d) = \inf\left\{ n \ge 1 : \sum_{k=1}^{K} \max\{0, W_{\alpha,k,n} - d\} \ge b \right\}, \quad (3.4)$$

$$N^{(r)}_\alpha(b) = \inf\Big\{ n \ge 1 : \sum_{k=1}^{r} W_{\alpha,(k),n} \ge b \Big\}, \quad (3.5)$$

where $W_{\alpha,(1),n} \ge W_{\alpha,(2),n} \ge \ldots \ge W_{\alpha,(K),n}$ are the order statistics of the K local L_α-CUSUM statistics $W_{\alpha,1,n}, \ldots, W_{\alpha,K,n}$.

One can also consider other shrinkage functions such as the detectability score tansformation $h(x) = \log[1 - p_0 + 0.64p_0 \exp(x/2)]$ proposed in [14]. This yields another global monitoring scheme

$$N_{Chan,\alpha}(b, p_0) = \inf\left\{n \geq 1 : \sum_{k=1}^{K} \log\left[1 - p_0 + 0.64 * p_0 \exp(W_{\alpha,k,n}/2)\right] \geq b\right\}.$$

Our extensive numerical simulation experiences illustrate that for a given α, the scheme $N_{Chan,\alpha}(b)$ in (3.6) has the similar statistical/robustness properties to those schemes $N_\alpha^{(soft)}(b, d)$ and $N_\alpha^{(r)}(b)$ in (3.4) and (3.5) in many interesting sparse post-change scenarios when $p_0 = r/K$. This is because all these procedures utilize the same local L_α-CUSUM statistics $W_{\alpha,k,n}$ in (3.1) and aim to detect the same post-change scenarios (after regularization).

Besides these aforementioned shrinkage transformations, there are other approaches to combine the local detection statistics together to make a global alarm. Two popular approaches in the literature are the "MAX" and the "SUM" schemes, see [27] and [28]:

$$N_{\alpha,\max}(b) = \inf\left\{n \geq 1 : \max_{1\leq k\leq K} W_{\alpha,k,n} \geq b\right\}, \tag{3.6}$$

$$N_{\alpha,\text{sum}}(b) = \inf\left\{n \geq 1 : \sum_{k=1}^{K} W_{\alpha,k,n} \geq b\right\}. \tag{3.7}$$

On one hand, the "MAX" and the "SUM" schemes could be considered as the special cases of our proposed top-r based scheme $N_\alpha^{(r)}(b)$ in (3.5) when $r = 1$ and $r = K$ respectively. On the other hand, the "MAX" and "SUM" approaches are generally statistically inefficient unless in extreme cases of very few or many affected local data streams.

Note that there are three tuning parameters in our proposed schemes: (α, d, b) for the schemes $N_\alpha^{(soft)}(b, d)$ in (3.4) and (α, r, b) for the scheme $N_\alpha^{(r)}(b)$ in (3.5). It is natural to ask what are the "optimal" choices of these tuning parameters. It turns out that the most challenging one is the optimal choice of the common parameter α, which is related to the robustness from the gross error models in (2.2), and will be discussed in Section 5. Next, the "optimal" choice of the shrinkage parameter d or r mainly depends on the number of affected local data streams, see our asymptotic properties in the next section. Finally, the choice of the threshold b is straightforward for given two other parameters since it can be chosen to satisfy the false alarm constraint in (2.3).

4. Worst-case asymptotic efficiency score

In this section, we derive the worst-case asymptotic efficiency score (2.8) of our proposed schemes $N_\alpha^{(soft)}(b, d)$ in (3.4) and $N_\alpha^{(r)}(b)$ in (3.5). To see that, we first report two standard change-point detection properties of our proposed schemes: the ARL to false alarm and detection delay under the gross error model $h_i = (1-\epsilon)f_i + \epsilon g_i$, where the outlier distributions g_i are given and $i = 0, 1$. Then, we will look at the worst-case of the outlier distributions to derive the worst-case asymptotic efficiency. It is important to note that our proposed schemes do not involve the contamination ratio ϵ or the information of outliers ϵ, g. Finally, based on our detection delay analysis, we provide guidelines on how to choose the tuning parameters in our proposed schemes. The proofs of the theorems are presented in the Appendix.

Let us begin with the definition of the expectation of the L_α-likelihood ratio statistic $Y = ([f_1(X)]^\alpha - [f_0(X)]^\alpha)/\alpha$ when X is distributed according to $h_i = (1-\epsilon)f_i + \epsilon g_i$ for given ϵ, g_i, and $i = 0, 1$ Note that when $\alpha = 0$, the variable Y should be treated as the log-likelihood ratio $\log(f_1(X)/f_0(X))$.

Definition 4.1. Given $\epsilon \geq 0$ and $\alpha \geq 0$, for $i = 0, 1$, define

$$\begin{aligned} I_i(\epsilon, \alpha; g_i) &= \mathbf{E}_{h_i}\Big[\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\Big] \\ &= (1-\epsilon)\mathbf{E}_{f_i}\Big[\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\Big] + \epsilon\mathbf{E}_{g_i}\Big[\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\Big]. \end{aligned} \tag{4.1}$$

Note when $\epsilon = 0$, $I_i(\epsilon = 0, \alpha; g_i)$ does not depend on g_i. So we further denote $I_i(\alpha) := I_i(\epsilon = 0, \alpha; g_i)$ for simplification. It turns out that the ARL to false alarm and detection delay of our proposed schemes are depending on whether $I_i(\epsilon, \alpha; g_i) < 0$ or > 0. Next, let us summarize the false alarm properties of our proposed schemes under the gross error model $h_0 = (1-\epsilon)f_0 + \epsilon g_0$.

Theorem 4.1. *Assume $I_0(\epsilon, \alpha; g_0) < 0$, then there exists a unique positive constant $\lambda(\epsilon, \alpha; g_0)$ depends on $f_0, f_1, g_0, \alpha, \epsilon$ such that*

$$\mathbf{E}_{h_0} \exp\Big\{\lambda(\epsilon, \alpha; g_0)\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\Big\} = 1. \tag{4.2}$$

With the constant $\lambda(\epsilon, \alpha; g_0) > 0$ in (4.2), the ARL to false alarm of our proposed schemes, $N_\alpha^{(soft)}(b, d)$ in (3.4) and $N_\alpha^{(r)}(b)$ in (3.5), are given as follows under different sufficient conditions:
(a) When $\lambda(\epsilon, \alpha; g_0)b > K\exp\{-\lambda(\epsilon, \alpha; g_0)d\}$, we have

$$\mathbf{E}_{h_0}^{(\infty)}[N_\alpha^{(soft)}(b, d)] \geq \frac{1}{4}\exp\left(\Big[\sqrt{\lambda(\epsilon, \alpha; g_0)b} - \sqrt{K\exp\{-\lambda(\epsilon, \alpha; g_0)d\}}\Big]^2\right). \tag{4.3}$$

(b) When $\lambda(\epsilon, \alpha; g_0)b > K$, we have

$$\mathbf{E}_{h_0}^{(\infty)}[N_\alpha^{(r)}(b)] \geq \frac{1}{4}\exp\left(\Big[\sqrt{\lambda(\epsilon, \alpha; g_0)b} - \sqrt{K}\Big]^2\right). \tag{4.4}$$

Let us add some comments to better understand the theorem. First, the existence of the unique constant $\lambda(\epsilon, \alpha; g_0) > 0$ in (4.2) is based on the assumption that $I_0(\epsilon, \alpha; g_0) < 0$, see Appendix A2 of [29]. Moreover, when $\epsilon = 0$, both $I_0(0, \alpha; g_0)$ and $\lambda(0, \alpha; g_0)$ only depend on the idealized model $f_i(x)$ and α, but do not depend on the information of outliers, i.e., ϵ and g_i. For simplification, we denote $\lambda(\alpha) = \lambda(0, \alpha; g_0)$.

Second, our rigorous, non-asymptotic results in (4.3) and (4.4) hold no matter how large the number K of data streams is. This allows us to investigate the modern asymptotic regime when the dimension K goes to ∞.

Finally, the assumptions of $\lambda(\epsilon, \alpha; g_0)b > K\exp\{-\lambda(\epsilon, \alpha; g_0)d\}$ or $\lambda(\epsilon, \alpha; g_0)b > K$ essentially says that the global threshold b of our proposed schemes should be large enough if one wants to control the global false alarm rate when online monitoring large-scale streams. These results allow us to find a conservative threshold b so as to satisfy the false alarm constraint in (2.3), also see the details of parameter setting below.

Next, the following theorem summarizes the detection delays of our proposed schemes $N_\alpha^{(soft)}(b,d)$ in (3.4) and $N_\alpha^{(r)}(b)$ in (3.5) when m out of K features are affected by the occurring event for some given $1 \le m \le K$. The detailed proof of Theorem 4.2 will be presented in Appendix.

Theorem 4.2. *Suppose $I_1(\epsilon, \alpha; g_1) > 0$, and m out of K features are affected.*
(a) If $b/m + d$ goes to ∞, then the detection delay of $N_\alpha^{(soft)}(b,d)$ satisfies

$$\mathbf{D}_{h_1}(N_\alpha^{(soft)}(b,d)) \le (1+o(1))\frac{1}{I_1(\epsilon,\alpha;g_1)}\left(\frac{b}{m}+d\right), \tag{4.5}$$

(b) If $r \ge m$ and b/m goes to ∞, then the detection delay of $N_\alpha^{(r)}(b)$ satisfies

$$\mathbf{D}_{h_1}(N_\alpha^{(r)}(b)) \le (1+o(1))\frac{1}{I_1(\epsilon,\alpha;g_1)}\left(\frac{b}{m}\right), \tag{4.6}$$

where the $o(1)$ term does not depend on the dimension K, but might depend on m and α as well as the distributions h_1.

To simplify the notation, we use N_α to denote both the scheme $N_\alpha^{(soft)}(b,d)$ and scheme $N_\alpha^{(r)}(b)$. By Theorem 4.1 and Theorem 4.2, when K is fixed, if $I_0(\epsilon,\alpha;g_0) < 0$ and $I_1(\epsilon,\alpha;g_1) > 0$, we can get a natural lower bound of the asymptotic efficiency score (2.7) of our proposed schemes,

$$\mathrm{AE}(N_\alpha, \epsilon; g_0, g_1) \ge m\lambda(\epsilon,\alpha;g_0)I_1(\epsilon,\alpha;g_1). \tag{4.7}$$

However, if we can find outlier distributions g_0^*, g_1^* such that $I_0(\epsilon,\alpha;g_0^*) > 0$ and $I_1(\epsilon,\alpha;g_1^*) < 0$, we will get

$$\mathrm{AE}(N_\alpha, \epsilon; g_0^*, g_1^*) = 0, \tag{4.8}$$

which implies the procedure cannot detect the persistent change from f_0 to f_1 at all due to the contamination of outliers.

Now, we are ready to present the worst-case asymptotic efficiency score of our proposed scheme N_α. First, assume $I_0(\alpha) = \mathbf{E}_{f_0}\left[\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\right] < 0$ and $I_1(\alpha) = \mathbf{E}_{f_1}\left[\frac{[f_1(X)]^\alpha - [f_0(X)]^\alpha}{\alpha}\right] > 0$, denote

$$M^*(\alpha) = \operatorname*{ess\,sup}_x \left|\frac{[f_1(x)]^\alpha - [f_0(x)]^\alpha}{\alpha}\right|. \tag{4.9}$$

Then we have the following theorem:

Theorem 4.3. *For our proposed scheme $N_\alpha(b)$with $\alpha \ge 0$, suppose K is fixed and $b \to \infty$,*
(a) if $\epsilon < -I_0(\alpha)/[M^*(\alpha) - I_0(\alpha)]$ and $\epsilon < I_1(\alpha)/[M^*(\alpha) + I_1(\alpha)]$, we have

$$\mathit{WAE}(N_\alpha, \epsilon) \;\; \ge \;\; m\lambda^*(\epsilon,\alpha)\Big[(1-\epsilon)I_1(\alpha) - \epsilon M^*(\alpha)\Big] > 0, \tag{4.10}$$

where $\lambda^*(\epsilon,\alpha) = \inf_{g_0 \in \mathcal{G}} \lambda(\epsilon,\alpha;g_0) > 0$.
(b) Otherwise, WAE$(N_\alpha, \epsilon) = 0$.

Note if $\log(f_1(x)/f_0(x))$ is unbounded, we have $M^*(0) = +\infty$. Based on Theorem 4.3, for any $\epsilon > 0$, WAE$(N_{\alpha=0}, \epsilon) = 0$, which implies the CUSUM based method cannot detect the persistent change at all under any percentage of outliers. However, if both f_0, f_1 are bounded, for any $\alpha > 0$, we have $M^*(\alpha) < +\infty$. Thus, our proposed schemes N_α will always have a positive worst-case asymptotic efficiency score when the contamination ratio ϵ is small. This implies the detection efficiency of our proposed schemes under the gross error model.

Note that there are three tuning parameters in our proposed schemes: (α, d, b) for the schemes $N_\alpha^{(soft)}(b, d)$ in (3.4) and (α, r, b) for the scheme $N_\alpha^{(r)}(b)$ in (3.5). It is natural to ask what are the "optimal" choices of these tuning parameters. It turns out that Theorems 4.1 and 4.2 provide the optimal choices of (d, b) or (r, b) that asymptotically minimize the detection delay subject to the false alarm constraint γ in (2.3). Below we will report the corresponding results, and detailed proofs and descriptions are postponed to Appendix.

(1) The optimal choice of parameter α, which turns out to be the most challenging one, as it is related to the robustness from the gross error models in (2.2). We will discuss in more details in Section 5 through the concept of false alarm breakdown point. The result in Section 5 shows the optimal α_{opt} only depends on the distributions f_0, f_1 but independent of other parameters d, r, b, and outliers information ϵ, g.

(2) Given α_{opt}, the choice of the shrinkage parameter d or r mainly depends on the number m of affected local feature coefficients. If we want to minimize the detection delay subject to the false alarm constraint γ in (2.3), we can set $r = m$ for the scheme $N_{\alpha_{opt}}^{(r)}(b)$ in (3.5). The optimal choice of d for the proposed scheme $N_{\alpha_{opt}}^{(soft)}(b, d)$ in (3.4) is a little complicated, and given by

$$d_{opt} = \frac{1}{\lambda(\alpha_{opt})} \left(\log \frac{K}{m} + \log \frac{\log \gamma}{m} \right), \tag{4.11}$$

where $\lambda(\alpha_{opt})$ is defined in (4.2) and only depends on f_0, f_1 and α_{opt}.

(3) The choice of the threshold b is straightforward for given two other parameters, since it can be chosen to satisfy the false alarm constraint in (2.3) under the idealized distribution f_0. A choice of global detection threshold

$$b_\gamma = \frac{1}{\lambda(\alpha_{opt})} \left(\sqrt{\log(4\gamma)} + \sqrt{K \exp\{-\lambda(\alpha_{opt}) d_{opt}\}} \right)^2, \tag{4.12}$$

will guarantee that our proposed scheme $N_{\alpha_{opt}}^{(soft)}(b_\gamma, d_{opt})$ satisfies the global false alarm constraint γ in the idealized model as in (2.3).

Note that all these choices of parameters do not depend on the ϵ or g, and only depend on the idealized model f_0, f_1 and a prior knowledge on the number m of affected data streams.

5. Breakdown point analysis

In this section, we will investigate the robustness properties of our proposed schemes, $N_\alpha^{(soft)}(b, d)$ in (3.4) and $N_\alpha^{(r)}(b)$ in (3.5), through the false alarm breakdown point analysis. This will provide the guideline on how to choose the tuning parameter α, which controls the robustness of our proposed schemes.

In the classical offline robust statistics, the breakdown point is one of the most popular measures of robustness of statistical procedures. At a high-level, in the context of finite samples, the breakdown point is the smallest percentage of contaminations that may cause an estimator or statistical test to be really poor. Since the pioneering work of [23] for the asymptotic definition of breakdown point, much research has been done to investigate the breakdown point for different robust estimators or hypothesis testings in the offline statistics, see [30], [31]. To the best of our knowledge, no research has been done on the breakdown point analysis under the online monitoring or change-point context.

Given the importance of the system-wise false alarm rate for online monitoring large-scale data streams in real-world applications, here we focus on the breakdown point analysis for false alarms. Intuitively, for a family of procedures $T(b)$ that is robust, if it is designed to satisfy the false alarm constraint γ in (2.3) under the idealized model f_0, then its false alarm rate should not be too bad under the gross error model h_0 with some small amount of outliers. There are two specific technical issues that require further clarification. First, how bad is a "bad" false alarm rate? We propose to follow the sequential change-point detection literature to assess the false alarm rate by $\log \mathbf{E}_{h_0}^{(\infty)}(T(b))$ and deem the false alarm rate unacceptable if $\log \mathbf{E}_{h_0}^{(\infty)}(T(b))$ is much smaller than the designed level of $\log \gamma$, i.e., if $\log \mathbf{E}_{h_0}^{(\infty)}(T(b)) = o(\log \gamma)$. Second, what kind of the contamination function g in (2.5) should we consider in the gross error model? Here we propose to follow the offline robust statistics literature to consider the worst-case scenario in the ϵ-contaminated distribution class in [1] that includes any arbitrary contamination functions g's, which leads to the definition of the false alarm breakdown point in (2.4).

Now we are ready to conduct the false alarm breakdown point analysis for our proposed schemes $N_\alpha^{(soft)}(b,d)$ and $N_\alpha^{(r)}(b)$ with a given tuning parameter $\alpha \geq 0$. To do so, for the densities $f_0(x)$ and $f_1(x)$, and for any given $\alpha \geq 0$, we define an intrinsic bound

$$M(\alpha) = \operatorname*{ess\,sup}_x \frac{[f_1(x)]^\alpha - [f_0(x)]^\alpha}{\alpha}, \tag{5.1}$$

and the density power divergence between f_0 and f_1:

$$d_\alpha(f_0, f_1) \;=\; \int \Big\{ [f_1(x)]^{1+\alpha} - (1+\frac{1}{\alpha}) f_0(x)[f_1(x)]^\alpha + \frac{1}{\alpha}[f_0(x)]^{1+\alpha} \Big\} dx. \tag{5.2}$$

Note that $d_\alpha(f_0, f_1)$ was proposed in [2], which showed that it is always positive when f_1 and f_0 are different. Moreover, when $\alpha = 0$, $d_{\alpha=0}(f_0, f_1)$ becomes Kullback-Leibler information number $\int f_0(x) \log \frac{f_0(x)}{f_1(x)} dx$.

With these two new notations, the following theorem derives the false alarm breakdown point of our proposed schemes $N_\alpha^{(soft)}(b,d)$ and $N_\alpha^{(r)}(b)$ as a function of the tuning parameter α for a fixed soft-thresholding parameter d and r when online monitoring a given K number of data streams. Since they have the same breakdown point, to simplify the notation, we use N_α to denote both the scheme $N_\alpha^{(soft)}(b,d)$ and scheme $N_\alpha^{(r)}(b)$.

Theorem 5.1. *Suppose that $f_\theta(x) = f(x-\theta)$ is a location family of density function with continuous probability density function $f(x)$, and assume $f_{\theta_0}(x) - f_{\theta_1}(x)$ takes both positive and negative values for $x \in (-\infty, +\infty)$. For $\alpha \geq 0$, and any fixed d and K, the false alarm breakdown point of*

our proposed schemes N_α in (3.4) and (3.5) is the same and given by

$$\epsilon^*(N_\alpha) = \frac{d_\alpha(f_{\theta_0}, f_{\theta_1})}{d_\alpha(f_{\theta_0}, f_{\theta_1}) + (1+\alpha)M(\alpha)}, \tag{5.3}$$

where $M(\alpha)$ and $d_\alpha(f_{\theta_0}, f_{\theta_1})$ are defined in (5.1) and (5.2). In particular, $\epsilon^*(N_\alpha) = 0$ if $M(\alpha) = \infty$ and $d_\alpha(f_{\theta_0}, f_{\theta_1})$ is finite.

The proof of Theorem 5.1 requires the asymptotic properties of our proposed schemes $N_\alpha^{(soft)}(b, d)$ in (3.4) and $N_\alpha^{(r)}(b)$ in (3.5) under the assumption that ϵ and g are given, which has been studied in the previous section. The detailed proof of Theorem 5.1 will be presented in the supplementary materials.

Next, let us apply Theorem 5.1 to guide us to choose the optimal robustness parameter α. Since the false alarm breakdown point of our proposed schemes do not require any information about the contamination ratio ϵ and contamination distribution g, one nature idea is to maximize the false alarm breakdown point in (5.3):

$$\alpha_{opt} = \arg\max_{\alpha \geq 0} \frac{d_\alpha(f_{\theta_0}, f_{\theta_1})}{d_\alpha(f_{\theta_0}, f_{\theta_1}) + (1+\alpha)M(\alpha)} \tag{5.4}$$

As an illustration, let us see the results of (5.3) and (5.4) for widely used normal distributions, i.e., when f_θ is the pdf of $N(\theta, \sigma^2)$. In this case, when $\alpha = 0$, the density power divergence $d_{\alpha=0}(f_{\theta_0}, f_{\theta_1}) = \frac{1}{2\sigma^2}(\theta_1 - \theta_0)^2$ is finite, but the bound $M(\alpha = 0)$ in (5.1) becomes $+\infty$ since it is the supremum of the log-likelihood ratio $\log f_{\theta_1}(x) - \log f_{\theta_0}(x) = (\theta_1 - \theta_0)x - (\theta_1^2 - \theta_0^2)/2$ over $x \in (-\infty, \infty)$. Hence, $\epsilon^*(N_{\alpha=0}) = 0$. That is, the false alarm breakdown point of the baseline CUSUM-based scheme $N_{\alpha=0}$ is 0, i.e., any amount of outliers will deteriorate the false alarm rate of the classical CUSUM statistics-based schemes. This is consistent with the offline robust statistics literature that the likelihood-function based methods are very sensitive to model assumptions and are generally not robust.

Meanwhile, for any $\alpha > 0$, note that

$$\int_{-\infty}^{\infty} f_{\theta_0}(x)[f_{\theta_1}(x)]^\alpha dx = \frac{1}{(\sqrt{2\pi}\sigma)^\alpha \sqrt{1+\alpha}} \exp\left(-\frac{\alpha(\theta_1 - \theta_0)^2}{2(1+\alpha)\sigma^2}\right),$$

and thus it is not difficult to derive from (5.2) that

$$d_\alpha(f_{\theta_0}, f_{\theta_1}) = \frac{\sqrt{1+\alpha}}{\alpha(\sqrt{2\pi}\sigma)^\alpha}\left(1 - \exp(-\frac{\alpha(\theta_1 - \theta_0)^2}{2(1+\alpha)\sigma^2})\right). \tag{5.5}$$

Moreover, if we let $M(= 1/\sqrt{2\pi\sigma^2})$, then $|f_\theta(x)| \leq M$ for all x. By the definition in (5.1), we have $|M(\alpha)| \leq 2M^\alpha/\alpha$, which is finite for any $\alpha > 0$. This implies that for normal distributions, $\epsilon^*(N_\alpha) > 0$ for any $\alpha > 0$. Thus our proposed L_α-CUSUM based scheme with $\alpha > 0$ is much more robust than the classical CUSUM scheme.

To see the optimal choice of α based on (5.4), let us consider a concrete numerical example when $f_{\theta_0} \sim N(0, 1)$ and $f_{\theta_1} \sim N(1, 1)$. By (5.5), we can compute the value $d_\alpha(0, 1)$ for any $\alpha \geq 0$. While we do not have analytic formula for the upper bound $M(\alpha)$ in (5.1), its numerical value can be easily found by brute-force exhaustive search over the real line $x \in (-\infty, \infty)$. The result shows

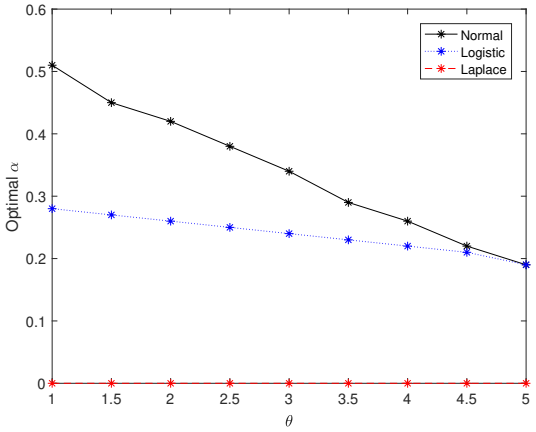

Figure 1: The value α_{opt} in (5.4) for different θ_1.

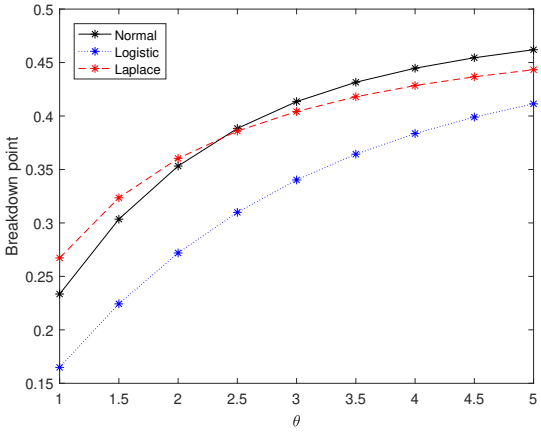

Figure 2: The false alarm breakdown point $\epsilon^*(N_\alpha)$ in (5.3) when $\alpha = \alpha_{opt}$ for different θ_1.

the false alarm breakdown point of our proposed scheme N_α will first increase and then decrease as α varies from 0 to 2., and yields the optimal choice of α_{opt} as 0.51, with corresponding breakdown point as 0.233. That means our proposed scheme with the choice of $\alpha = 0.51$ could tolerate 23.3% arbitrarily bad observations in terms of keeping the designed false alarm constraint stable.

Finally, we should emphasize that the optimal value α_{opt} in (5.4) and the false alarm breakdown point $\epsilon^*(N_\alpha)$ in (5.3) will generally depend on the change magnitude or signal-to-noise-ratio. To illustrate this, we consider three families: Normal, Laplace, and Logistic distributions with the scale parameter $\sigma = 1$. This yields three families of pdfs, $f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\theta)^2}{2})$, $\frac{1}{2}\exp(-|x-\theta|)$, or $\frac{\exp(-(x-\theta))}{(1+\exp(-(x-\theta)))^2}$. In each case, we assume that the pre-change parameter $\theta_0 = 0$, the designed post-change parameter θ_1 varies from 1 to 5. In Figures 1 and 2, we plot the optimal value α_{opt} and the corresponding false alarm breakdown point $\epsilon^*(N_\alpha)$ as a function of θ_1. Figure 2 implies that with the increasing of the post-change θ_1 or the signal-to-noise-ratio, our proposed robust schemes with optimal α can tolerate more outliers. Also it is interesting to see from Figure 1 that the optimal α_{opt} decreases for normal or logistic distribution as the post-change parameter θ_1 increases. A surprising result is that the optimal $\alpha_{opt} = 0$ for the Laplace distribution. This implies the classical CUSUM procedure for Laplace distribution is actually optimal in the sense of having the largest breakdown point. One possible explanation is that for the Laplace distribution, the log-likelihood ratio $\log(f_{\theta_1}(x)/f_{\theta_0}(x)) = -|x-\theta_1| + |x-\theta_0|$ takes values in the interval $[\theta_0 - \theta_1, \theta_1 - \theta_0]$ when $\theta_1 > \theta_0$. Thus the impact of outliers is directly controlled.

6. Numerical Simulations

In this section we conduct numerical simulation studies to illustrate the robustness and efficiency of our proposed schemes $N_\alpha^{(soft)}(b,d)$ and $N_\alpha^{(r)}(b)$.

In our simulation studies, we assume there are $K = 100$ independent data streams, and at some unknown time, $m = 10$ features are affected by the occurring event. Also the change is instantaneous if a stream is affected, and we do not know which subset of streams will be affected. We set $f_\theta =$ pdf of $N(\theta, 1)$, the pre-change parameter $\theta_0 = 0$, the post-change parameter $\theta_1 = 1$,

and the contamination densities g_0, g_1 are pdfs of $N(0, 3^2)$. Our proposed schemes $N_\alpha^{(soft)}(b, d)$ in (3.4) and $N_\alpha^{(r)}(b)$ in (3.5) are constructed by using the density function f_{θ_0} and f_{θ_1}.

In the first simulation study, we consider the idealized model when $\epsilon = 0$. In this case, for our proposed robust scheme $N_\alpha^{(soft)}(b, d)$ in (3.4), as shown in the previous section, the optimal choices of $\alpha_{opt} = 0.51$. By (4.11), if $\log(\gamma) << K$, then the corresponding optimal shrinkage parameters $d \approx \frac{1}{\lambda(\epsilon=0,\alpha=0.51)} \log \frac{K}{m} = 0.8915$ for $K = 100$ and $m = 10$, since $\lambda(\epsilon = 0, \alpha = 0.51) = 2.5829$. For our proposed robust scheme $N_\alpha^{(r)}(b)$ in (3.5), we choose $\alpha = \alpha_{opt} = 0.51$ and $r = 10$. For the baseline CUSUM-based scheme, i.e., $N_{\alpha=0}^{(soft)}(b, d)$ with $\alpha = 0$, we choose the shrinkage parameter $d = \frac{1}{\lambda(\epsilon=0,\alpha=0)} \log \frac{K}{m} = 2.3026$, since $\lambda(\epsilon = 0, \alpha = 0) = 1$.

In summary, we will compare the following different schemes.

- Our proposed scheme $N_\alpha^{(soft)}(b, d)$ in (3.4) with $\alpha_{opt} = 0.51$ and $d = 0.8915$.
- Our proposed scheme $N_\alpha^{(r)}(b)$ in (3.5) with $\alpha_{opt} = 0.51$ and $r = 10$.
- The baseline CUSUM-based scheme $N_{\alpha=0}^{(soft)}(b, d)$ with $d = 2.3026$.
- The MAX scheme $N_{\alpha=0.51,\max}(b)$ in (3.6);
- The SUM scheme $N_{\alpha=0.51,\text{sum}}(b)$ in (3.7);
- The method $N_{XS}(b, p_0 = 0.1)$ in [12] based on generalized likelihood ratio:

$$N_{XS}(b, p_0) = \inf\Big\{n \geq 1 : \max_{0\leq i<n} \sum_{k=1}^{K} \log(1 - p_0 + \quad p_0 \exp[\big(U_{k,n,i}^{+}\big)^2/2]) \geq b\Big\},$$

  where for all $1 \leq k \leq K, 0 \leq i < n$,

$$U_{k,n,i}^{+} = \max\big(0, \frac{1}{\sqrt{n-i}} \sum_{j=i+1}^{n} X_{k,j}\big).$$

- The method $N_{Chan,\alpha=0}(b, p_0 = 0.1)$ in [14] under the idealized model that is an extension of the SUM scheme in [28]:

$$N_{Chan,\alpha=0}(b, p_0) = \inf\Big\{n \geq 1 : \sum_{k=1}^{K} \log\big(1 - p_0 + 0.64 * p_0 \exp(W_{k,n}^{*}/2)\big) \geq b\Big\},$$

  where $W_{k,n}^{*}$ is the CUSUM statistics in (3.2).

- The method $N_{Chan,\alpha=0.51}(b, p_0 = 0.1)$ in (3.6) which is similar to $N_{Chan,\alpha=0}$ but replace the CUSUM statistic by our proposed L_α-CUSUM statistic.

For each of these schemes $T(b)$, we first find the appropriate values of the threshold b to satisfy the false alarm constraint $\gamma \approx 5000$ under the idealized model with $\epsilon = 0$ (within the range of sampling error). Next, using the obtained global threshold value b, we simulate the detection delay

Table 1: A comparison of the detection delays of 8 schemes with $\gamma = 5000$ under the idealized model. The smallest and largest standard errors of these 8 schemes are also reported under each post-change hypothesis based on 1000 repetitions in Monte Carlo simulations.

Gross error model with $\epsilon = 0$								
	# affected local data streams							
	1	3	8	10	15	20	50	100
Smallest standard error	0.29	0.12	0.05	0.04	0.03	0.03	0.01	0.00
Largest standard error	0.58	0.20	0.07	0.06	0.05	0.03	0.02	0.01
Our proposed robust scheme								
$N^{(soft)}_{\alpha=0.51}(b = 8.5, d = 0.8915)$	41.0	18.6	10.3	9.2	7.5	6.5	4.5	3.9
$N^{(r=10)}_{\alpha=0.51}(b = 17.19)$	40.6	18.5	10.3	9.2	7.7	6.9	5.3	4.8
Comparison of other methods								
$N^{(soft)}_{\alpha=0}(b = 21.52, d = 2.3026)$	33.6	15.2	8.4	7.5	6.1	5.3	3.7	3.0
$N_{\alpha=0.51,\max}(b = 4.3)$	27.7	19.6	16.2	15.6	14.8	14.2	12.7	11.9
$N_{\alpha=0.51,\text{sum}}(b = 36.85)$	63.7	26.9	12.5	10.5	7.8	6.4	3.3	2.0
$N_{Chan,\alpha=0.51}(b = 1.04, p_0 = 0.1)$	31.4	17.7	10.8	9.7	7.8	6.7	4.1	3.0
$N_{Chan,\alpha=0}(b = 21.6, p_0 = 0.1)$	32	15.2	11.2	7.5	5.3	4.2	3.3	2.3
$N_{XS}(b = 19.5, p_0 = 0.1)$	30.9	13.2	7.2	5.7	4.7	3.5	1.8	1.0

when the change-point occurs at time $\nu = 1$ under several different post-change scenarios, i.e., different number of affected sensors. All Monte Carlo simulations are based on 1000 repetitions.

Table 1 summarizes the detection delays of these 8 schemes under 9 different post-change hypothesis. Among all schemes, $N_{XS}(b, p_0)$ generally yields the smallest detection delay. However, we want to emphasize that it is computationally expensive. Specifically, even if we use a time window of size k as in [12] to speed up the implementation of $N_{XS}(b, p_0)$, at each time n, $O(Kk^2)$ computations are needed to get the global monitoring statistics, whereas our proposed scheme $N^{(soft)}_{\alpha}(b, d)$ only require $O(K)$ computations to get the global monitoring statistics.

Another interesting observation from Table 1 is that the detection delay of our proposed robust schemes $N^{(soft)}_{\alpha=0.51}(b, d)$ and $N^{(r)}_{\alpha=0.51}(b)$ are not too bad compared with the CUSUM-based scheme $N^{(soft)}_{\alpha=0}(b, d = 2.3026)$, and it just takes additional 1.7 time steps to raise a correct global alarm under the idealized model when $m = 10$ data streams are affected.

In the second simulation study, we will examine the detection efficiency of these schemes under the gross error model when $\epsilon = 0.1$. For each of these 8 schemes, we use the same threshold b obtained from the first simulation to guarantee these schemes satisfy the same false alarm constraint $\gamma = 5000$ under the idealized model. Then, we wilsimulate the in-control average run and the detection delay of these schemes when both the pre-change distribution and post-change distribution are the gross error model in (2.2) with $\epsilon = 0.1$, g_0, g_1 as pdfs of $N(0, 3^2)$. We then report the empirical version of the asympototic efficiency score in (2.7) of these schemes under 8 different post-change hypothesis in Table 2.

First, we can see our proposed scheme $N^{(soft)}_{\alpha=0.51}(b, d = 0.8915)$ and $N^{(r=10)}_{\alpha=0.51}(b = 18.7)$ have the largest detection efficiency score among all comparison methods when 10 data streams are affected. Moreover, by using our proposed L_α-CUSUM statistics with $\alpha = 0.51$, the method

Table 2: A comparison of the detection efficiency score of 8 schemes under the gross error model with $\epsilon = 0.1$ based on 1000 repetitions in Monte Carlo simulations. The threshold b is chosen to satisfy $\gamma = 5000$ in the idealized model.

Gross error model with $\epsilon = 0.1$								
	# affected local data streams							
	1	3	8	10	15	20	50	100
Our proposed robust scheme								
$N^{(soft)}_{\alpha=0.51}(b = 8.5, d = 0.8915)$	0.17	0.34	0.61	0.68	0.83	0.95	1.37	1.66
$N^{(r=10)}_{\alpha=0.51}(b = 17.09)$	0.17	0.35	0.62	0.68	0.82	0.92	1.2	1.35
Other methods for comparison								
$N^{(soft)}_{\alpha=0}(b = 21.52, d = 2.3026)$	0.27	0.32	0.4	0.43	0.48	0.53	0.7	0.8
$N_{\alpha=0.51,\max}(b = 4.3)$	0.3	0.36	0.44	0.46	0.49	0.51	0.58	0.62
$N_{\alpha=0.51,\text{sum}}(b = 36.85)$	0.12	0.25	0.5	0.58	0.77	0.94	1.73	2.88
$N_{Chan,\alpha=0}(b = 21.6, p_0 = 0.1)$	0.26	0.32	0.36	0.43	0.54	0.63	0.75	1.03
$N_{Chan,\alpha=0.51}(b = 1.04, p_0 = 0.1)$	0.21	0.37	0.59	0.65	0.81	0.94	1.53	2.19
$N_{XS}(b = 19.5, p_0 = 0.1)$	0.22	0.39	0.44	0.49	0.52	0.55	0.78	0.93

$N_{Chan,\alpha=0.51}(b, p_0 = 0.1)$ yields the similar detection efficiency to our proposed schemes. This illustrates that the improvement of L_α-CUSUM statistics is significant as compared to the baseline CUSUM statistics in the presence of outliers.

It is also interesting to note that the MAX-scheme $N_{\alpha=0.51,\max}(b)$ and the SUM-scheme $N_{\alpha=0.51,\text{sum}}(b)$ are designed for the case when $m = 1$ or $m = K$ features are affected, and Table 2 confirmed that their detection efficinecys are indeed the largest in their respective designed scenarios. However, when the number of affected features m is moderate and is arround 10, our proposed scheme $N^{(soft)}_{\alpha=0.51}(b, d)$ and $N^{(r)}_{\alpha=0.51}(b)$ have larger detection efficiency, which implies our proposed schemes with sum-shrinkage technique could be more robust to the number of affected features.

In the third experiment, we investigate the impact of contamination rate ϵ on the false alarms of different methods to illustrate the robustness of our proposed L_α-CUSUM statistics. Since the top-r scheme $N^{(r)}_{\alpha=0.51}(b)$, MAX-scheme $N_{\alpha=0.51,\max}(b)$, the SUM-scheme $N_{\alpha=0.51,\text{sum}}(b)$ and $N_{Chan,\alpha=0.51}(b, p_0)$ are all based on local L_α-CUSUM statistics, their robustness properties are similar to our proposed scheme $N^{(soft)}_{\alpha=0.51}(b, d)$. To highlight the robustness of our proposed L_α-CUSUM statistics, we only compare our proposed scheme $N^{(soft)}_{\alpha=0.51}(b, d)$ with other three schemes: $N^{(soft)}_{\alpha=0}(b, d)$, $N_{Chan,\alpha=0}(b, p_0)$, and $N_{XS}(b, p_0)$.

Figure 3 reports the curve of $\log \mathbf{E}^{(\infty)}_{h_0}(T)$ as the contamination ratio ϵ varies from 0.02 to 0.2 with stepsize 0.02. Clearly, all curves decrease with the increasing of contaminations, meaning that all schemes will raise false alarm more frequently when there are more outliers. However, the curves for the CUSUM or likelihood-ratio based methods decreased very quickly, whereas our proposed L_α-CUSUM statistics-based method with $\alpha_{opt} = 0.51$ decreases rather slowly. This suggests that our proposed scheme is more robust in the sense of keeping the designed ARL more stable with a small departure from the assumed model.

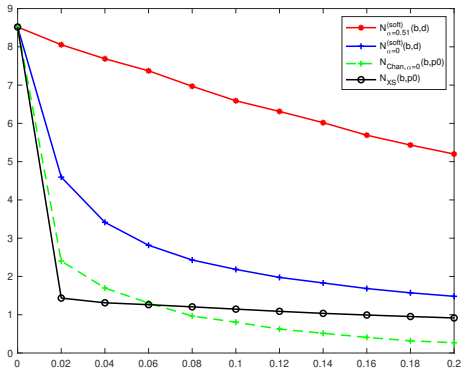

Figure 3: Each line represents $\log \mathbf{E}_{h_0}^{(\infty)}(T)$ of a scheme as a function of $\epsilon \in (0, 0.2)$.

7. Conclusion

In this paper, we study the problem of robust monitoring of large-scale data streams when the true observed data follow Huber's gross error model. We develop a family of efficient and robust detection schemes that can be implemented in real-time. From the worst-case detection efficiency point of view, we show our proposed methods can still have positive detection efficiency under a small proportion of arbitrary outliers. In contrast, the CUSUM-based methods lose all detection efficiency once the data include outliers. From the robustness point of view, we propose a new concept called false alarm breakdown point, which measures the stability of the designed false alarm constraint of any monitoring procedures under the effects of outliers. Our breakdown point analysis implies our proposed methods can have positive breakdown points. We also provide detailed guidelines on the choices of tuning parameters in our detection procedures. However, in this work, we focus on the problem of monitoring homogeneous independent data streams. It is of future interest to extend the problem to nonhomogeneous data streams with some correlation structures.

Appendix

In this online supplementary material, we provide detailed proofs to Theorems 4.1, 4.2, and Theorem 4.3, the optimal parameter choice in Section 4, and the proof of Theorem 5.1.

A. Proof of Theorem 4.1

(a) For any $x \geq 0$, by Chebyshev's inequality,

$$\begin{aligned}
\mathbf{E}_{h_0}^{(\infty)}[N_\alpha^{(soft)}(b,d)] &\geq x\mathbf{P}_{h_0}^{(\infty)}(N_\alpha^{(soft)}(b,d) \geq x) \\
&= x\left[1 - \mathbf{P}_{h_0}^{(\infty)}(N_\alpha^{(soft)}(b,d) < x)\right] \\
&= x\left[1 - \mathbf{P}_{h_0}^{(\infty)}(\sum_{k=1}^{K} \max\{0, W_{\alpha,k,n} - d\} \geq b) \text{ for some } 1 \leq n \leq x\right]
\end{aligned}$$

$$\geq \quad x\left[1 - x\mathbf{P}_{h_0}^{(\infty)}(\sum_{k=1}^{K}\max\{0, W_{\alpha,k}^* - d\} \geq b)\right], \tag{7.1}$$

where $W_{\alpha,k}^* = \limsup_{n\to\infty} W_{\alpha,k,n}$. We will show that $W_{\alpha,k}^*$ exists later, and when it does exist, it is clear that $W_{\alpha,k}^*$ are i.i.d. across different k under the pre-change measure $\mathbf{P}_{h_0}^{(\infty)}$. Now if we define the log-moment generating function of the $W_{\alpha,k}^*$'s

$$\psi_\alpha(\theta) = \log \mathbf{E}_{h_0}^{(\infty)} \exp\{\theta \max(0, W_{\alpha,k}^* - d)\} \tag{7.2}$$

for some $\theta \geq 0$, then another round application of Chebyshev's inequality yields

$$\begin{aligned}\exp(K\psi_\alpha(\theta)) &= \mathbf{E}_{h_0}^{(\infty)} \exp\{\theta \sum_{k=1}^{K}\max(0, W_{\alpha,k}^* - d)\} \\ &\geq e^{\theta b}\mathbf{P}_{h_0}^{(\infty)}(\sum_{k=1}^{K}\max\{0, W_{\alpha,k}^* - d\} \geq b)\end{aligned} \tag{7.3}$$

for $\theta > 0$. Combining (7.1) and (7.3) yields that

$$\mathbf{E}_{h_0}^{(\infty)}[N_\alpha^{(soft)}(b,d)] \quad \geq \quad x\left[1 - x\exp(-\theta b + K\psi_\alpha(\theta))\right] \tag{7.4}$$

for all $x \geq 0$. Since $x(1-xu)$ is maximized at $x = 1/(2u)$ with the maximum value $1/(4u)$. We conclude from (7.4) that

$$\mathbf{E}_{h_0}^{(\infty)}[N_\alpha^{(soft)}(b,d)] \quad \geq \quad \frac{1}{4}\exp\left(\theta b - K\psi_\alpha(\theta)\right). \tag{7.5}$$

for any $\theta > 0$ as long as $\psi_\alpha(\theta)$ in (7.2) is well-defined.

The remaining proof is to utilize the definition of $\lambda(\epsilon, \alpha; g_0) > 0$ in (4.2) to show that the upper limiting $W_{\alpha,k}^*$ of the proposed L_α-CUSUM statistics is well-defined and derive a careful analysis of $\psi_\alpha(\theta)$ in (7.2). When $\alpha = 0$, the L_α-CUSUM statistics become the classical CUSUM statistics, and the corresponding analysis is well-known, see [21]. Here our main insight is that our proposed L_α-CUSUM statistics $W_{\alpha,k,n}$ for detecting a change from $h_0(x)$ to $h_1(x)$ in (2.2) can be thought of as the classical CUSUM statistic for detecting a local change from $h_0(x)$ to another new density function $h_2(x)$. Hence, under the pre-change hypothesis of $h_0(\cdot)$, the false alarm properties of our proposed L_α-CUSUM statistics can be derived through those of the classical CUSUM statistics.

By the definition of $\lambda(\epsilon, \alpha; g_0) > 0$, if we define a new function

$$h_2(x) := \exp\left\{\lambda(\epsilon, \alpha; g_0)\left(\frac{(f_1(x))^\alpha - (f_0(x))^\alpha}{\alpha}\right)\right\} h_0(x), \tag{7.6}$$

then $h_2(x)$ is a well-defined probability density function. Then in the problem of detection a local change from $h_0(x)$ to $h_2(x)$, the local CUSUM statistics for the kth local data stream is defined recursively by

$$\begin{aligned}W'_{k,n} &= \max\{0, W'_{k,n-1} + \log\frac{h_2(X_{k,n})}{h_0(X_{k,n})}\} \\ &= \max\{0, W'_{k,n-1} + \lambda(\epsilon, \alpha; g_0)\frac{[f_1(X_{k,n})]^\alpha - [f_0(X_{k,n})]^\alpha}{\alpha}\}.\end{aligned}$$

Compared with our proposed L_α-CUSUM statistics $W_{\alpha,k,n}$, it is clear that $W'_{k,n} = \lambda(\epsilon,\alpha)W_{\alpha,k,n}$, and thus our proposed L_α-CUSUM statistics $W_{\alpha,k,n}$'s are equivalent to the standard CUSUM statistics $W'_{k,n}$ up to a positive constant $\lambda(\epsilon,\alpha;g_0)$. By the classical results on the CUSUM, see Appendix 2 on Page 245 of [32], as $n\to\infty$, $W'_{k,n}$ converges to a limit and thus $W_{\alpha,k,n}$ also converges to a limit, denoted by $W^*_{\alpha,k}$. Moreover, the tail probability of $W^*_{\alpha,k}$ satisfies

$$G(x) = \mathbf{P}^{(\infty)}_{\theta_0}(W^*_{\alpha,k}\ge x) = \mathbf{P}^{(\infty)}_{\theta_0}(\limsup_{n\to\infty} W'_{k,n}\ge \lambda(\epsilon,\alpha;g_0)x) \le e^{-\lambda(\epsilon,\alpha;g_0)x}. \tag{7.7}$$

Now we shall use (7.7) to derive information bound of $\psi_\alpha(\theta)$ in (7.2). In order to simplify our arguments, we abuse the notation and simply denote $\lambda(\epsilon,\alpha;g_0)$ by λ in the remaining proof of the theorem. By the definition of $\psi_{\alpha,k}(\theta)$ in (7.2) and the tail probability $G(x)$ in (7.7), for $\theta>0$,

$$\begin{aligned}
\psi_\alpha(\theta) &= \log[\mathbf{P}^{(\infty)}_{\theta_0}(W^*_{\alpha,k}\le d) - \int_d^\infty e^{\theta(x-d)}dG(x)] \qquad (7.8)\\
&= \log[1+\theta\int_d^\infty e^{\theta(x-d)}G(x)dx]\\
&\le \log[1+\theta\int_d^\infty e^{\theta(x-d)}e^{-\lambda x}dx]\\
&= \log\left(1+\frac{\theta}{\lambda-\theta}e^{-d\lambda}\right) \le \frac{\theta}{\lambda-\theta}e^{-d\lambda},
\end{aligned}$$

where the second equation is based on the integration by parts. Clearly, relation (7.8) holds for any $0<\theta<\lambda=\lambda(\epsilon,\alpha;g_0)$.

By (7.5) and (7.8), we have

$$\mathbf{E}^\infty_\epsilon N^{(soft)}_\alpha(b,d) \;\ge\; \frac{1}{4}\exp\left(\theta b - \frac{K\theta}{\lambda-\theta}e^{-d\lambda}\right) \tag{7.9}$$

for all $0<\theta<\lambda=\lambda(\epsilon,\alpha;g_0)$. When $\lambda b > K\exp\{-\lambda d\}$, relation (4.3) follows at once from (7.9) by letting $\theta = \sqrt{\lambda/b}\left(\sqrt{\lambda b}-\sqrt{K\exp\{-d\lambda\}}\right)\in(0,\lambda)$. This completes the proof of Theorem 4.1 (a).

(b) Note $N^{(soft)}_\alpha(b,d=0)\le N^{(r)}_\alpha(b)$ for any $b\ge 0$. Therefore, (4.4) can be derived directly from (4.3) by letting $d=0$ in (4.3).

B. Proof of Theorem 4.2

First, we will prove the part (a) of Theorem 4.2. To prove the detection delay bound (4.6) in Theorem 4.2, without loss of generality, assume the first m data streams are affected. Consider a new stopping time

$$T'(b,d) \;=\; \inf\{n\ge 1:\sum_{k=1}^m (W_{\alpha,k,n}-d)\ge b\} = \inf\{n\ge 1:\sum_{k=1}^m W_{\alpha,k,n}\ge b+md\}.$$

Clearly $N^{(soft)}_\alpha(b,d)\le T'(b,d)$, and thus

$$\mathbf{D}_{h_1}(N^{(soft)}_\alpha(b,d)) \;\le\; \mathbf{D}_{h_1}(T'(b,d)).$$

Next, by the recursive definition of $W_{\alpha,k,n}$ in (3.1), using the same approach in Theorem 2 of [25] that connects the recursive CUSUM-type scheme to the random walks, we have

$$\mathbf{D}_{h_1}(T'(b,d))) \;\leq\; \mathbf{E}_1 T''(b,d),$$

where $\mathbf{E}_1$ denotes the expectation when the change happen at time $\nu = 1$, and $T''(b,d)$ is the first passage time when the random walk with i.i.d. increment of mean $mI_1(\epsilon,\alpha;g_1)$ exceeds the bound $b+md$, and is defined as

$$T''(b,d) \;=\; \inf\{n \geq 1 : \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{[f_1(X_{k,i})]^\alpha - [f_0(X_{k,i})]^\alpha}{\alpha} \geq b + md\}.$$

By standard renewal theory, as $(\frac{b}{m}+d)\to\infty$, we have

$$\mathbf{E}_1 T''(b,d) \leq \frac{1+o(1)}{mI_1(\epsilon,\alpha;g_1)}\,(b+md)\,.$$

Relation (4.6) then follows at once from the above relations, which completes the proof of part (a) of Theorem 4.2.

To prove the part (b), we define another stopping time

$$\tau(b) := \inf\{n \geq 1 : \sum_{k=1}^{m} W_{\alpha,k,n} \geq b\}.$$

Note for the sorted statistics $W_{\alpha,(1),n} \geq W_{\alpha,(2),n} \geq \cdots \geq W_{\alpha,(K),n}$, we have $\sum_{k=1}^{m} W_{\alpha,k,n} \leq \sum_{k=1}^{m} W_{\alpha,(k),n}$. Thus,when $m \leq r$, $N_\alpha^{(r)}(b) \leq \tau(b)$. By standard renew theory, we have

$$\mathbf{D}_{h_1}(N_\alpha^{(r)}(b)) \;\leq\; \mathbf{D}_{h_1}(\tau(b)) \leq (1+o(1))\frac{b}{mI_\theta(\epsilon,\alpha)},$$

which completes the proof of part (b) of Theorem 4.2.

C. Proof of Theorem 4.3

Note if $\epsilon < -I_0(\alpha)/[M^*(\alpha) - I_0(\alpha)]$,

$$\begin{aligned}\sup_{g_0\in\mathcal{G}} I_0(\epsilon,\alpha;g_0) \;&=\; (1-\epsilon)I_0(\alpha) + \epsilon \sup_x (\frac{[f_1(x)]^\alpha - [f_0(x)]^\alpha}{\alpha}) \\ &\leq\; (1-\epsilon)I_0(\alpha) + \epsilon M^*(\alpha) < 0. \end{aligned} \tag{7.10}$$

Therefore, by Theorem 4.1, there exists a postive number $\lambda^*(\epsilon,\alpha) = \inf_{g_0\in\mathcal{G}} \lambda(\epsilon,\alpha;g_0) > 0$ such that

$$\lim_{b\to\infty} \frac{\inf_{g_0\in\mathcal{G}} \left[\log(\mathbf{E}_{h_0}^{(\infty)}(N_\alpha(b)))\right]}{b} \geq \lambda^*(\epsilon,\alpha).$$

Moreover, if $\epsilon < I_1(\alpha)/[M^*(\alpha) + I_1(\alpha)]$,

$$\begin{aligned} \inf_{g_1 \in \mathcal{G}} I_1(\epsilon, \alpha; g_1) &= (1-\epsilon) I_1(\alpha) + \epsilon \inf_x \left(\frac{[f_1(x)]^\alpha - [f_0(x)]^\alpha}{\alpha} \right) \\ &\geq (1-\epsilon) I_1(\alpha) - \epsilon M^*(\alpha) > 0. \end{aligned} \tag{7.11}$$

By Theorem 4.2, we have

$$\lim_{b \to \infty} \frac{\sup_{g_1 \in \mathcal{G}} [\mathbf{D}_{h_1}(N_\alpha(b))]}{b} \leq \frac{1}{m \inf_{g_1 \in \mathcal{G}} I_1(\epsilon, \alpha; g_1)} \leq \frac{1}{m[(1-\epsilon) I_1(\alpha) - \epsilon M^*(\alpha)]}.$$

Thus, by the definition of worst-case detection efficiency in (2.8), we have

$$\text{WAE}(N_\alpha, \epsilon) = \lim_{b \to \infty} \frac{\inf_{g_0 \in \mathcal{G}} \left[\log(\mathbf{E}_{h_0}^{(\infty)}(N_\alpha(b))) \right]}{\sup_{g_1 \in \mathcal{G}} [\mathbf{D}_{h_1}(N_\alpha(b))]} \geq m \lambda^*(\epsilon, \alpha) \Big[(1-\epsilon) I_1(\alpha) - \epsilon M^*(\alpha) \Big].$$

D. Parameter Setting in Section 4

The choice of $b = b_\gamma$ in (4.12) follows directly from Theorem 4.1 (a). To prove (4.11), we abuse the notation and use λ to denote $\lambda(\alpha)$ for simplification. By Theorem 4.2, the optimal d is the non-negative value that minimize the function

$$\ell(d) := \frac{b_\gamma}{m} + d = \frac{1}{\lambda m} (\sqrt{\log(4\gamma)} + \sqrt{K e^{-\lambda d}})^2 + d. \tag{7.12}$$

This is an elementary optimization problem, and the optimal d can be found by taking derivative of $\ell(d)$ with respect to d, since $\ell(d)$ is a convex function of d. To see this,

$$\begin{aligned} \ell'(d) &= -\frac{1}{m} (\sqrt{K e^{-\lambda d}} + \frac{\sqrt{\log(4\gamma)}}{2})^2 + 1 + \frac{\log(4\gamma)}{4m} \\ \ell''(d) &= \frac{\lambda}{m} (\sqrt{K e^{-\lambda d}} + \frac{\sqrt{\log(4\gamma)}}{2}) \sqrt{K e^{-\lambda d}} > 0. \end{aligned}$$

Thus $\ell(d)$ is a convex function on $[0, +\infty)$, and the optimal d_{opt} value can be found by setting $\ell'(d) = 0$:

$$\sqrt{K e^{-\lambda d}} = \sqrt{m + \frac{\log(4\gamma)}{4}} - \frac{1}{2} \sqrt{\log(4\gamma)}.$$

This gives an unique optimal value

$$\begin{aligned} d_{opt} &= \frac{1}{\lambda} \log \frac{K}{(\sqrt{m + \frac{1}{4}\log(4\gamma)} - \frac{1}{2}\sqrt{\log(4\gamma)})^2} \\ &= \frac{1}{\lambda} \left\{ \log \frac{\left[\sqrt{m + \frac{1}{4}\log(4\gamma)} + \frac{1}{2}\sqrt{\log(4\gamma)} \right]^2}{m} + \log \frac{K}{m} \right\}, \end{aligned} \tag{7.13}$$

which is equivalent to those in (4.11) under the assumption that $m = m(K) << \min(\log \gamma, K)$. Plugging $d = d_{opt}$ in (7.13) back to (4.12) yields the choice of b_γ.

D. Proof of Theorem 5.1

By Theorems 4.1 and 4.2, the false alarm breakdown point of our proposed method N_α can be found by finding the smallest ϵ value such that $I_0(\epsilon, \alpha; g_0) > 0$ for some distribution g_0, where $I_0(\epsilon, \alpha; g_0)$ is defined in (4.1). That is equivalent to

$$\epsilon^*(N_\alpha) = \inf\{\epsilon \geq 0 : \sup_{g_0} I_0(\epsilon, \alpha; g_0) > 0\}, \tag{7.14}$$

The remaining proof is based on a careful analysis of $I_0(\epsilon, \alpha; g_0)$ for any arbitrary outlier density function g_0. For any $h_0(x) = (1-\epsilon)f_0(x) + \epsilon g_0(x) \in \hbar_{0,\epsilon}$, by (4.1), we have

$$I_0(\epsilon, \alpha; g_0) \;\; = \;\; -\frac{1-\epsilon}{1+\alpha} d_\alpha(f_0, f_1) + \epsilon \int (\frac{[f_1(x)]^\alpha - [f_0(x)]^\alpha}{\alpha}) g(x) dx, \tag{7.15}$$

where $d_\alpha(f_0, f_1)$ is defined in (5.2) and is the density power divergence between f_0 and f_1 proposed by [2]. Here we use the fact that $\int [f_1(x)]^{1+\alpha} dx = \int [f_0(x)]^{1+\alpha} dx$ when $f_0(x)$ and $f_1(x)$ come from the same location family.

By the definition of $M(\alpha)$ in (5.1), it is clear from (7.15) that

$$\sup_{g_0} I_0(\epsilon, \alpha; g_0) = -\frac{1-\epsilon}{1+\alpha} d_\alpha(f_0, f_1) + \epsilon M(\alpha). \tag{7.16}$$

Therefore, by (7.14), if both $d_\alpha(f_0, f_1)$ and $M(\alpha)$ are finite, the false alarm breakdown point of N_α should be

$$\epsilon^*(N_\alpha) = \frac{d_\alpha(f_0, f_1)}{d_\alpha(f_0, f_1) + (1+\alpha) M(\alpha)}. \tag{7.17}$$

If $d_\alpha(f_0, f_1)$ is finite but $M(\alpha) = +\infty$, by (7.14) and (7.16), $\epsilon^*(N_\alpha) = 0$. If $d_\alpha(f_0, f_1) = +\infty$ but $M(\alpha)$ is finite, $\epsilon^*(N_\alpha) = 1$. If both $d_\alpha(f_0, f_1)$ and $M(\alpha)$ are $+\infty$ and $\frac{d_\alpha(f_0, f_1)}{M(\alpha)} = \rho$, by (7.14) and (7.16), we have $\epsilon^*(N_\alpha) = \frac{\rho}{\rho + (1+\alpha)}$ no matter ρ is finite or not. Therefore, for all cases, the false alarm breakdown point of N_α have the same expression in (7.17), which completes the proof of Theorem 5.1.

REFERENCES

[1] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[2] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[3] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[4] S. Heritier and E. Ronchetti, "Robust bounded-influence tests in general parametric models," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 897–904, 1994.

[5] V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, vol. 15, pp. 642–656, 1987.

[6] E. Cantoni and E. Ronchetti, "Robust inference for generalized linear models," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1022–1030, 2001.

[7] P. J. Huber and E. Ronchetti, *Robust Statistics*, 2nd ed. New York: Wiley, 2009.

[8] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 2011.

[9] G. Shmueli and H. Burkom, "Statistical challenges facing early outbreak detection in biosurveillance," *Technometrics*, vol. 52, no. 1, pp. 39–51, 2010.

[10] A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov, "Efficient computer network anomaly detection by changepoint detection methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 4–11, 2013.

[11] H. Yan, K. Paynabar, and J. Shi, "Image-based process monitoring using low-rank tensor decomposition," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 216–227, 2015.

[12] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[13] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6926–6938, 2015.

[14] H. P. Chan, "Optimal sequential detection in multi-stream data," *The Annals of Statistics*, vol. 45, no. 6, pp. 2736–2763, 2017.

[15] L. Gordon and M. Pollak, "An efficient sequential nonparametric scheme for detecting a change of distribution," *The Annals of Statistics*, vol. 22, pp. 763–804, 1994.

[16] ——, "A robust surveillance scheme for stochastically ordered alternatives," *The Annals of Statistics*, vol. 23, pp. 1350–1375, 1995.

[17] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2961–2974, 2005.

[18] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, "Minimax robust quickest change detection," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1604–1614, 2011.

[19] D. Ferrari and Y. Yang, "Maximum lq-likelihood estimation," *The Annals of Statistics*, vol. 38, no. 2, pp. 753–783, 2010.

[20] Y. Qin and C. E. Priebe, "Robust hypothesis testing via Lq-likelihood," *Statistica Sinica*, pp. 1793–1813, 2017.

[21] K. Liu, R. Zhang, and Y. Mei, "Scalable sum-shrinkage schemes for distribtued monitoring large-scale data streams," *Statistica Sinica*, vol. 29, no. 1, pp. 1–22, 2019.

[22] R. Zhang and Y. Mei, "Asymptotic statistical properties of communication-efficient quickest detection schemes in sensor networks," *Sequential Analysis*, vol. 37, no. 3, pp. 375–396, 2018.

[23] F. R. Hampel, "Contributions to the theory of robust estimation," Ph.D. dissertation, University of California Berkeley, 1968.

[24] G. Fellouris and G. Sokolov, "Second-order asymptotic optimality in multisensor sequential change detection," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3662–3675, 2016.

[25] G. Lorden, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[26] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[27] A. G. Tartakovsky and V. V. Veeravalli, "Asymptotically optimal quickest change detection in distributed sensor systems," *Sequential Analysis*, vol. 27, no. 4, pp. 441–475, 2008.

[28] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[29] A. Wald, *Sequential analysis*. Courier Corporation, 1973.

[30] W. S. Krasker and R. E. Welsch, "Efficient bounded-influence regression estimation," *Journal of the American statistical Association*, vol. 77, no. 379, pp. 595–604, 1982.

[31] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[32] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York, 1985.