ROBUST REAL-TIME MONITORING OF HIGH-DIMENSIONAL DATA STREAMS

BY RUIZHI ZHANG AND YAJUN MEI AND JIANJUN SHI

Georgia Institute of Technology

Robust real-time monitoring of high-dimensional data streams has many important real-world applications such as industrial quality control, signal detection, biosurveillance, but unfortunately it is highly non-trivial to develop efficient schemes due to two challenges: (1) the unknown sparse number or subset of affected data streams and (2) the uncertainty of model specification for high-dimensional data. In this article, motivated by the detection of smaller persistent changes in the presence of larger transient outliers, we develop a family of efficient real-time robust detection schemes for high-dimensional data streams through monitoring feature spaces such as PCA or wavelet coefficients when the feature coefficients are from Tukey-Huber’s gross error models with outliers. We propose to construct a new local detection statistic for each feature called $L_\alpha$-CUSUM statistic that can reduce the effect of outliers by using the Box-Cox transformation of the likelihood function, and then raise a global alarm based upon the sum of the soft-thresholding transformation of these local $L_\alpha$-CUSUM statistics so that to filter out unaffected features. In addition, we propose a new concept called false alarm breakdown point to measure the robustness of online monitoring schemes, and also characterize the breakdown point of our proposed schemes. Asymptotic analysis, extensive numerical simulations and case study of nonlinear profile monitoring are conducted to illustrate the robustness and usefulness of our proposed schemes.

1. Introduction. Robust statistics have been extensively studied in the offline context when the full data set is available for decision making and is contaminated with outliers, e.g., robust estimation (Huber, 1964; Basu et al., 1998), robust hypothesis testing (Huber, 1965; Heritier and Ronchetti, 1994), and robust regression (Yohai, 1987; Cantoni and Ronchetti, 2001). Also see the classical books, Huber and Ronchetti (2009) or Hampel et al. (2011), for literature review. In this paper, we propose to develop robust methods in the context of online monitoring when one is interested in detecting sparse persistent smaller changes in high-dimensional streaming data under the contamination of transient larger outliers.

A concrete motivating example of our research is profile monitoring in a progressive forming process, see Figure 1 for illustration. A progressive forming process has a set of dies installed

Keywords and phrases: Change-point, CUSUM, real-time monitoring, robustness, quickest detection, sparsity.
within one stamping press. The part is transferred from one die station to the next die station sequentially and each die station has a formed part processed in previous die station. During this process, the forming force measured by the tonnage sensor installed in the linkage of press is the summation of all forming forces generated in each die. The forming force is measured as a profile or functional data that consists of $2^{11} = 2048$ measurements points. As a work piece passes through the die stations, a fault in any die station might change the forming force (e.g. tonnage profiles). Figure 2 plots some typical patterns of the profile data under the normal condition as well as under two faulty conditions: fault #1 (the smaller change) caused by the malfunction of a part transferred in the forming station, and fault #2 (the larger change) due to missing operations in the pre-forming station. In practice, it is difficult to detect the smaller fault #1 condition since the difference between the fault #1 profile and the normal profile is sparse and small in magnitude. However, if this fault is neglected and the faulty condition remains uncovered, it will lead to persistent quality issues of formed parts, and further damage die. Meanwhile, the larger fault #2 can be observed easily due to the large difference from the normal profile. On one hand, line workers generally will be able to fix the corresponding root cause in the pre-forming station. On the other hand, the workers are generally unable to check whether it will affect the down-stream stations or not, and thus it may or may not lead the fault #1 condition. Hence, when monitoring high-dimensional data streams, it is highly desirable to develop effective methodologies to detect those smaller but persistent changes in the presence of infrequent larger changes which can be thought as outliers, and might or might not related to the smaller persistent changes.

In general, the problem of robust monitoring high-dimensional data 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 Shmueli and Burkom (2010); Tartakovsky, Polunchenko and Sokolov (2013); Yan, Paynabar and Shi (2015). Unfortunately, it is highly non-trivial to develop efficient robust monitoring schemes or algorithms due to two challenges: (1) the sparsity, where only a few unknown local components or features of data might be affected, but we do not know which local components or features are affected; and (2) the robustness, where we are interested in detecting smaller persistent changes, not the transient outliers.

In the sequential change-point literature for high-dimensional data, 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 Xie and Siegmund (2013) using a semi-Bayesian approach, and later by Wang and Mei (2015) using shrinkage-estimation-based schemes. Chan (2017) 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-dimensional streaming data: rank-based method in Gordon and Pollak (1994, 1995), kernel-based method in Desobry, Davy and Doncarli (2005), or least-favorable-distribution method in Unnikrishnan, Veeravalli and Meyn (2011). Unfortunately it is unclear how to extend these existing robust methods from one-dimension to high-dimension when we also need to deal with the sparsity issue in which there is uncertainty on the subset of affected local components or features.

In this paper, we develop efficient robust real-time monitoring schemes that are able to robustly detect smaller persistent changes in the presence of transient outliers when online monitoring of high-dimensional steaming data. From the methodology viewpoint, our proposed schemes are semi-parametric, and extend two contemporary concepts to the context of online monitoring of high-dimensional data streams: (i) $L_q$-likelihood in Ferrari and Yang (2010); Qin and Priebe (2017) for robustness, and (ii) the sum-shrinkage technique in Liu, Zhang and Mei (2017) for sparsity. These allow us to develop statistical efficient and computationally simple schemes that can be implemented recursively over time for robust real-time monitoring of high-dimensional data streams. Moreover, we also extend the concept of breakdown in the offline robust statistics (Hampel, 1968) to the sequential change-point detection context, and conduct the false alarm breakdown point analysis, which turns out to be useful for tuning parameters in our proposed schemes.

Our research makes four contributions in the statistics field by combining robust statistics with
sequential change-point detection for high-dimensional streaming data. First, our proposed method is robust with respect to infrequent outliers as well as the uncertainty of affected components of the data. Second, our proposed method can be implemented recursively and distributed via parallel computing, and thus is suitable for real-time monitoring over long time period for high-dimensional data. Third, inspired by the concept of breakdown point (Hampel, 1968) 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 scheme is indeed has 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 low bounds on the average run length of false alarm for our proposed method. The non-asymptotic results hold regardless of dimensionality, and allow us to provide a deep insight on the effect of high-dimensionality in the context of change-point detection under the modern asymptotic regime when the dimension or the number of data streams goes to $\infty$.

The remainder of this article is organized as follows. In Section 2, we start with the modern assumptions and present our proposed scheme in three steps. Then we provide the theoretical properties of our proposed scheme in Section 3. In Section 4, we introduce the concept of false alarm breakdown point and propose the general method to choose the robust tuning parameter $\alpha$. Simulation and case study results are presented in Section 5 and Section 6 respectively. The proofs of our main theorems are postponed to the Appendix.

2. Our proposed scheme. Suppose we are monitoring a sequence of high-dimension streaming data, $\{Y_n\}$, over time step $n = 1, 2, \cdots$, where the data might be corrupted with transient outliers. 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 no persistent distribution changes or if there are only transient outliers.

In this section, we will present the description of our proposed scheme, and then develop its asymptotic properties in next section, with the focus on the effect of the high-dimensionality in the context of change-point detection. At the high-level, our proposed scheme includes three components: (i) modeling extracted features, (ii) monitoring each local feature individually in parallel, and then (iii) combines local detection statistics together to make an online global-level decision. For the purpose of easy understanding, we split the presentation of our proposed scheme into three subsections, and each subsection focuses on each component of the proposed scheme.
2.1. Data and model. In many real-world applications such as profile monitoring in Figure 1, each raw data is independent over time, but local coordinates of each high-dimensional data can be dependent. In such a case, a standard technique is to extract independent features from the historical in-control data using principal component analysis (PCA), wavelets, tensor-decomposition, etc., and then monitor the feature coefficients instead of raw data themselves, see Jin and Shi (1999); Chang and Yadama (2010); Yan, Paynabar and Shi (2015); Paynabar, Zou and Qiu (2016). In the context of off-line estimation or prediction, one can focus on a few important features for the purpose of dimension reduction. However, a new challenge in the monitoring context is that we do not know which features might be affected by the change, and thus one often needs to monitor a relatively large number of features, see Wang, Mei and Paynabar (2018); Zhang, Mei and Shi (2018).

For each high-dimensional raw data $Y_n$, denote the corresponding $K$-dimensional feature coefficients as $X_n = (X_{1,n}, \cdots, X_{K,n})^T$. We assume that the local features are independent, and we have sufficient historical in-control data to model the pre-change cumulative density function (cdf) $F_k$ of the $k^{th}$ feature $X_{k,n}$’s. Without loss of generality, we assume that the $X_{k,n}$’s have the identical distribution, say, with the same probability density function (pdf) $f_{\theta_0} = \text{pdf of } N(0,1)$, under the in-control state, as we can consider the transformation $\Phi^{-1}(F_k(\cdot))$, where $\Phi$ is the cdf of the standard normal distribution, to standardize or normalized the in-control data if needed, see Efron (2012). Furthermore, as in our motivating example of profile monitoring in Figure 1, we further assume the $X_{k,n}$’s will have pdf $g$ when the raw data involves larger transient changes or outliers, and will have pdf $f_\theta$ when the raw data involves a smaller persistent change, where the unknown post-change parameter $\theta \geq \theta_1$ for some known value $\theta_1 > 0$.

Mathematically, recall the Tukey-Huber’s gross error model of the two-component mixture densities

$$h_\epsilon(x) = (1 - \epsilon)f_\theta(x) + \epsilon g(x),$$

where $\epsilon \in [0,1)$ is referred to as the contamination/outlier ratio and $g$ is the (unknown) outlier distributions. Then we model the $X_{k,n}$’s as the following change-point Tukey-Huber’s gross error model: for some unknown change time $\nu = 1, 2, \cdots$, all $X_{k,n}$’s are independent and identically distributed (i.i.d.) with $h_{\theta_0}(x)$ in (1) when $n \leq \nu - 1$, but $m$ out of $K$ local streams $X_{k,n}$’s have another distribution $h_\epsilon(x)$ in (1) when $n \geq \nu$, where the post-change parameter $\theta \geq \theta_1$, and $\theta_1 - \theta_0$ is the smallest meaningful magnitude of the change, which is pre-specified.
In the sequential change-point problem, at each and every time step, we need to test the null hypothesis
\[ H_0 : \nu = \infty \quad (i.e., \ no \ persistent \ change \ occurs) \]
against a composite alternative hypothesis
\[ H_1 : \nu = 1, 2, \cdots \quad (i.e., \ a \ persistent \ change \ occurs \ at \ some \ finite \ time). \]

The statistical procedure in the sequential change-point problem is often 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. Denote by \( P^{(\infty)}_{\theta_0} \) and \( E^{(\infty)}_{\theta_0} \) the probability measure and expectation when the data \( X_{k,n} \)'s are i.i.d. with density \( h_{\theta_0} \), and denote by \( P^{(\nu)}_{\theta} \) and \( E^{(\nu)}_{\theta} \) the same when the change occurs at time \( \nu \) and \( m \) out of \( K \) streams \( X_{k,n} \)'s have the post-change distribution \( h_{\theta} \). Under the standard minimax formulation for online change-point detection (Lorden, 1971), the performance of a stopping time \( T \) is evaluated by the average run length to false alarm (ARLFA), \( E^{(\infty)}_{\theta_0}(T) \) and the worst-case detection delay

\[ D_{\epsilon,\theta}(T) = \sup_{\nu \geq 1} \text{ess sup} E^{(\nu)}_{\theta} \left( (T - \nu + 1) + |F_{\nu-1} - 1| \right). \]

Here \( F_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]}) \) denotes past global information at time \( \nu \), \( X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1}) \) is past local information for the \( k \)-th feature.

An efficient detection procedure \( T \) should have small detection delay \( D_{\epsilon,\theta}(T) \) subject to the false alarm constraint

\[ E^{(\infty)}_{\theta_0}(T) \geq \gamma \]

for some pre-specified large constant \( \gamma > 0 \).

We should acknowledge that this is the standard formulation for monitoring of one- or low-dimensional data, and many classical procedures have been developed such as Page’s CUSUM procedure (Page, 1954), Shiryaev-Roberts procedure (Shiryaev, 1963; Roberts, 1966), window-limited procedures (Lai, 1995) and scan statistics (Glaz et al., 2001). Also some fundamental optimality results for one-dimensional data were established in Shiryaev (1963); Lorden (1971); Pollak (1985, 1987); Moustakides (1986); Ritov (1990); Lai (1995), etc. For a review, see the books such as Basseville and Nikiforov (1993); Poor and Hadjiliadis (2009); Tartakovsky, Nikiforov and Basseville (2014).
Note that here we do not aim to develop optimality theorem for monitoring of high-dimensional data, which is still an open problem in a general setting. Our main objective is to develop an efficient and robust scheme, and then to investigate its statistical properties, which shed the new light of the effect of the dimensionality $K$ on the high-dimensional change-point detection problem.

### 2.2. Robust local statistics.

To develop real-time robust monitoring schemes, we propose to borrow the parallel computing technique to monitor each local feature individually, and then use the sum-shrinkage technique to combine the local monitoring statistics together to make a global decision. For that purpose, it is crucial to have an efficient local monitoring statistic that is robust to outliers. To do so, for the $k^{th}$ local feature, we propose to define a new local $L_\alpha$-CUSUM statistic:

$$ W_{\alpha,k,n} = \max \left( W_{\alpha,k,n-1} + \frac{[f_{\theta_1}(X_{k,n})]^{\alpha} - [f_{\theta_0}(X_{k,n})]^{\alpha}}{\alpha}, 0 \right), $$

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 (1) and its suitable choice will be discussed later.

The motivation of our $L_\alpha$-CUSUM statistic in (4) is as follows. Recall that when locally monitoring the single $k^{th}$ data stream $X_{k,n}$ with a possible local distribution change from $f_{\theta_0}$ to $f_{\theta_1}$, the generalized likelihood ratio test becomes the classical CUSUM statistic $W_{k,n}^\ast$, which has a recursive form:

$$ W_{k,n}^\ast = \max \left( W_{k,n-1}^\ast + \log \frac{f_{\theta_1}(X_{k,n})}{f_{\theta_0}(X_{k,n})}, 0 \right). $$

The CUSUM statistic enjoys nice optimality properties when all models are fully correctly specified (Moustakides, 1986), 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 $L_\alpha$-likelihood function $([f(X)]^\alpha - 1)/\alpha$ for some $\alpha > 0$, see Ferrari and Yang (2010), Qin and Priebe (2017). At the high-level, $L_\alpha$-likelihood function is bounded below by $-1/\alpha$ when $f(X) \to 0$ for outliers, and thus become more much robust to outliers as compared to the log-likelihood statistics. Moreover, as $\alpha \to 0$, the $L_\alpha$-likelihood function converges to the log-likelihood statistic, and thus it keeps statistical efficiencies when $\alpha$ is small. Here we apply this idea to develop $L_\alpha$-CUSUM statistics that turns out to be robust to outliers. More rigorous robust properties will be discussed later in Section 4.
2.3. Efficient global monitoring statistics. With local $L_\alpha$-CUSUM statistics $W_{\alpha,k,n}$ in (4) for each local feature, 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 and raise a global-level alarm at time

$$N_\alpha(b,d) = \inf \left\{ n : \sum_{k=1}^{K} \max\{0, W_{\alpha,k,n} - d \} \geq b \right\},$$

for some pre-specified constants $b, d > 0$ whose appropriate choices will be discussed later.

Note that our proposed scheme $N_\alpha(b,d)$ in (6) uses the soft-thresholding transformation, $h(W) = \max\{0, W - d\}$, to filter out those non-changing local features, and keep only those local features 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, also see Liu, Zhang and Mei (2017) for more discussions.

It is useful to compare our proposed scheme $N_\alpha(b,d)$ in (6) with other existing methods from the spatial-temporal detection viewpoint. In the literature, many existing change-point schemes are developed by looking at the time domain first, and then searching the spatial domain over different features for possible feature changes, see Xie and Siegmund (2013); Wang and Mei (2015). Unfortunately, such approach is often computationally expensive and cannot be implemented online for real-time monitoring due to lack of recursive forms. Here our proposed method (6) switches the order of spatial and temporal domains by parallel searching for local changes for each and every possible local changes, yielding computationally simple schemes that can be implemented recursively for real-time monitoring.

We should also mention that besides the soft-thresholding transformation, 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 Tartakovsky and Veeravalli (2008) and Mei (2010):

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

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

Unfortunately, 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, $\alpha$, $d$ and $b$ in our proposed scheme $N_\alpha(b,d)$ in (6) and $L_\alpha$-CUSUM statistic $W_{\alpha,k,n}$ in (4), and it is useful to discuss what are the “optimal” choices of these tuning parameters. The most challenging one is the optimal choice of $\alpha$, which is related to the robustness from the gross error models in (1), and will be discussed in Section 4 through developing a new concept of false alarm breakdown point. Meanwhile, the “optimal” choice of the shrinkage parameter $d$ mainly depends on the spatial sparsity of the change on the $K$ local features, or the number $m$ of affected local feature coefficients, which will be discussed in the next section when we derive the asymptotic properties of our proposed scheme $N_\alpha(b,d)$ in (6). Finally, for given $\alpha$ and $d$, the choice of the threshold $b$ is straightforward, as it can be chosen to satisfy the false alarm constraint in (3).

3. Theoretical properties. In this section, we investigate the statistical properties of our proposed scheme $N_\alpha(b,d)$ in (6) in the modern asymptotic setting when the dimension $K$ goes to $\infty$, which shed light on the suitable choice of tuning parameters when monitoring high-dimensional data streams. It is important to note that the definition of our proposed scheme $N_\alpha(b,d)$ in (6) does not involve the contamination ratio $\epsilon$ or the probability density distribution of outlier $g$, but its statistical properties will depend on $\epsilon$ or $g$ in the gross error model in (1). Hence, in this section and only in this section, we assume that $\epsilon$ and $g$ are given, as our focus is to investigate the statistical properties of our proposed schemes.

For that purpose, let us first introduce two technical assumptions on the $L_\alpha$-likelihood ratio statistic $Y = ([f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha)/\alpha$ when $X$ is distributed according to $h_{\theta_0}$ or $h_{\theta_1}$ under the gross error model in (1). Note that when $\alpha = 0$, the variable $Y$ should be treated as the log-likelihood ratio $\log(f_{\theta_1}(X)/f_{\theta_0}(X))$.

The first assumption on $Y$ is related to the detection delay properties of our proposed schemes:

Assumption 3.1. Given $\theta \geq \theta_1$, $\epsilon \geq 0$ and $\alpha \geq 0$, assume

$$I_\theta(\epsilon, \alpha) = \mathbf{E}_{h_{\theta}} \left[ \frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha} \right]$$

$$= (1 - \epsilon)\mathbf{E}_{f_{\theta}} \left[ \frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha} \right] + \epsilon \mathbf{E}_g \left[ \frac{[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha} \right]$$

is positive, where $\mathbf{E}_{h_{\theta}}, \mathbf{E}_{f_{\theta}}$ and $\mathbf{E}_g$ denote the expectations when the density function of $X$ is $h_{\theta}, f_{\theta}$ and $g$, respectively.
We should mention that this assumption is very wild for small $\epsilon, \alpha > 0$. To see this, when $\epsilon = \alpha = 0$ and $\theta = \theta_1$, $I_{\theta}(\epsilon, \alpha)$ in the assumption becomes the well-known Kullback-Leibler information number

\begin{equation}
I_{\theta=\theta_1}(\epsilon = 0, \alpha = 0) = \mathbb{E}_{f_{\theta_1}} \log(f_{\theta_1}(X)/f_{\theta_0}(X)) = I(f_{\theta_1}, f_{\theta_0}),
\end{equation}

which is always positive unless $f_{\theta_0} = f_{\theta_1}$. Since all functions are continuous with respect to $\alpha$ and $\epsilon$, it is reasonable to assume that $I_{\theta}(\epsilon, \alpha)$ are also positive for small $\epsilon, \alpha > 0$. Indeed, if $f_{\theta}$ belongs to a one-parameter exponential family

\begin{equation}
f_{\theta}(x) = \exp(\theta x - b(\theta)),
\end{equation}

where $b(\theta)$ is strictly convex on $\mathbb{R}$, then it is straightforward to show that $I_{\theta}(\epsilon = 0, \alpha = 0)$ would be an increasing function of $\theta$. This implies $I_{\theta}(\epsilon = 0, \alpha = 0) \geq I_{\theta=\theta_1}(\epsilon = 0, \alpha = 0) = I(f_{\theta_1}, f_{\theta_0}) > 0$ for all $\theta \geq \theta_1$. Thus, $I_{\theta}(\epsilon, \alpha) > 0$ for small $\epsilon, \alpha > 0$, and Assumption 3.1 holds.

The second assumption on $Y$ is related to the false alarm rate of our proposed schemes, and involves some basic probability knowledge on the moment generating function (MGF). For a random variable $Y$ with pdf $s(y)$, recall that the MGF is given by $\varphi(\lambda) = \mathbb{E}(e^{\lambda Y}) = \int e^{\lambda y} s(y) dy$ when well-defined. A nice property of MGF is that $\varphi(\lambda)$ is a convex function of $\lambda$ with $\varphi(0) = 1$. An important corollary is that there often exists another non-zero constant $\lambda^*$ such that $\varphi(\lambda^*) = 1$, and $\lambda^* > 0$ if and only if $\mathbb{E}(Y) < 0$, see Lemma 7.1 in the Appendix. Our second assumption essentially says that this is the case under the pre-change hypothesis, and is rigorously stated as follows.

**Assumption 3.2.** Given $\epsilon \geq 0$ and $\alpha \geq 0$, assume there exists a number $\lambda(\epsilon, \alpha) > 0$ such that

\begin{equation}
1 = \mathbb{E}_{b_{\theta_0}} \exp \left\{ \frac{\lambda(\epsilon, \alpha)[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha} \right\} \\
= (1 - \epsilon)\mathbb{E}_{f_{\theta_0}} \exp \left\{ \frac{\lambda(\epsilon, \alpha)[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha} \right\} + \epsilon \mathbb{E}_Y \exp \left\{ \frac{\lambda(\epsilon, \alpha)[f_{\theta_1}(X)]^\alpha - [f_{\theta_0}(X)]^\alpha}{\alpha} \right\}.
\end{equation}

We should mention that Assumption 3.2 is reasonable at least when $\epsilon$ and $\alpha$ are small. To see this, note that when $\alpha = 0$ and $\epsilon = 0$, for $Y = \log(f_{\theta_1}(X)/f_{\theta_0}(X))$, we have $\mathbb{E}_{f_{\theta_0}}(e^Y) = 1$ and thus $\lambda(\epsilon = 0, \alpha = 0) = 1$ in Assumption 3.2. Therefore, $\lambda(\epsilon, \alpha)$ should be in the neighborhood of 1 and thus are positive when $\epsilon$ and $\alpha$ are small.

With Assumptions 3.1 and 3.2, we are able to present the properties of our proposed scheme $N_\alpha(b, d)$ in (6) in the following subsections. Subsection 3.1 discusses the false alarm properties,
whereas subsection 3.2 investigates the detection delay properties including the robustness regarding on the number of affected local data streams.

3.1. False alarm analysis. In this subsection, we analyze the global false alarm rate of our proposed scheme $N_\alpha(b,d)$ in (6) for online monitoring $K$ independent features under the gross error model in (1), no matter how large $K$ is. The classical techniques in sequential change-point detection for one-dimensional data are based on the change of measure arguments and then use renewal theory to conduct overshoot analysis under the asymptotic setting as the global threshold $b$ goes to $\infty$. Unfortunately such renewal-theory-based analysis often yields poor approximations when the dimension $K$ is moderately large, since the overshoot constant generally increases exponentially as a function of the dimension $K$. Moreover, they cannot be extended to the modern asymptotic regime when the number $K$ of local data streams goes to $\infty$. In other words, these classical techniques are unable to provide deep insight on the effects of the dimension $K$.

Here we present an alternative approach that is based on Chebyshev’s inequality and can provide useful information bounds on the global false alarm rate regardless of how large the number $K$ of features is.

**Theorem 3.1.** Given that Assumption 3.2 holds for $\epsilon \geq 0$ and $\alpha \geq 0$, i.e., $\lambda(\epsilon, \alpha) > 0$. If $\lambda(\epsilon, \alpha)b > K\exp\{-\lambda(\epsilon, \alpha)d\}$, then the average run length to false alarm of our proposed scheme $N_\alpha(b,d)$ in (6) satisfies

$$E_\epsilon^{(\infty)}[N_\alpha(b,d)] \geq \frac{1}{4}\exp\left(\left[\sqrt{\lambda(\epsilon, \alpha)b} - \sqrt{K\exp\{-\lambda(\epsilon, \alpha)d\}}\right]^2\right).$$  

The detailed proof of Theorem 3.1 will be postponed in subsection 7.1, and here let us add some comments to better understand the theorem. First, our rigorous, non-asymptotic result in (13) holds no matter how large the number $K$ of features is. This allows us to investigate the modern asymptotic regime when the dimension $K$ goes to $\infty$.

Second, the assumption of $\lambda(\epsilon, \alpha)b > K\exp\{-\lambda(\epsilon, \alpha)d\}$ essentially says that the global threshold $b$ of our proposed scheme $N_\alpha(b,d)$ in (6) should be large enough if one wants to control the global false alarm rate when online monitoring large-scale streams. In particular, in order to satisfy the false alarm constraint $\gamma$ in (3), it is natural to set the right-hand side of (13) to $\gamma$. This yields a conservative choice of $b$ that satisfies $\sqrt{\lambda(\epsilon, \alpha)b} = \sqrt{K\exp\{-\lambda(\epsilon, \alpha)d\}} + \sqrt{\log(4\gamma)}$. Such a choice of $b$ will automatically satisfy the key assumption of $\lambda(\epsilon, \alpha)b > K\exp\{-\lambda(\epsilon, \alpha)d\}$ in the theorem.
Third, when $\epsilon = \alpha = 0$, we have $\lambda(\epsilon = 0, \alpha = 0) = 1$, and our lower bound (13) is similar, though slightly looser, as compared to those results in equation (3.17) of Liu, Zhang and Mei (2017), whose arguments are heuristic under a more refined assumption on some tail distributions (see $G(x)$ defined in (36) below). Here we provide a rigorous mathematical statement in Theorem 3.1 with fewer assumptions, though the price we pay is that the corresponding lower bound is a little loose.

Finally, it turns out that our lower bound (13) provides the correct first-order term of the classical CUSUM procedure when online monitoring $K = 1$ data stream under the idealized model. In that case, we have $\epsilon = \alpha = d = 0$, and the classical CUSUM procedure is the special case of our procedure $N_{\alpha=0}(b,d = 0)$. Since $\lambda(\epsilon = 0, \alpha = 0) = 1$, our lower bound (13) shows that for any $b > 1$,

$$\liminf_{b \to \infty} \frac{\log E_{\epsilon,\theta}^{(\infty)}[N_{\alpha=0}(b,d = 0)]}{b} \geq 1.$$  \hskip1em (14)

Meanwhile, as the classical CUSUM procedure, it is well-known from the classical renewal-theory-based techniques that $\lim_{b \to \infty} \frac{\log E_{\epsilon,\theta}^{(\infty)}[N_{\alpha=0}(b,d = 0)]}{b} = 1$, see Lorden (1971). Hence, our lower bound (13) provides the correct first-order term for $\log E_{\epsilon,\theta}^{(\infty)}[N_{\alpha}(b,d)]$ under the one-dimensional case as $b \to \infty$. As a result, we feel our lower bound in (13) is not bad in the modern asymptotic regime when the dimension $K$ goes to $\infty$.

3.2. Detection delay analysis. In this subsection, we provide the detection delays of our proposed scheme $N_{\alpha}(b,d)$ in (6) under the gross error model $h_{\theta}$ in (1) when $m$ out of $K$ features are affected by the occurring event for some given $1 \leq m \leq K$. In particular, note our proposed scheme $N_{\alpha}(b,d)$ in (6) only use the information of the pre-change parameter $\theta_0$, the minimal magnitude of the change parameter $\theta_1$ and tuning parameters $\alpha, b, d$, we will investigate its detection delay properties when the true post-change parameter $\theta$ is not less than $\theta_1$. The following theorem presents the detection delay properties, and the proof will be postponed in Section 7.

**Theorem 3.2.** Suppose Assumption 3.1 of $I_\theta(\epsilon, \alpha) > 0$ in (9) holds, and assume $m$ out of $K$ features are affected. If $b/m + d$ goes to $\infty$, then the detection delay of $N_{\alpha}(b,d)$ satisfies

$$D_{\epsilon,\theta}(N_{\alpha}(b,d)) \leq (1 + o(1)) \frac{1}{I_\theta(\epsilon, \alpha)} \left( \frac{b}{m} + d \right),$$  \hskip1em (15)

where the $o(1)$ term does not depend on the dimension $K$, and might depend on $m$ and $\alpha$ as well as the distributions $h_{\theta}$.
Theorem 3.2 characterizes the detection delay of our proposed scheme $N_\alpha(b,d)$ in (6), which is constructed by using the density function of $f_{\theta_0}$ and $f_{\theta_1}$, under the gross error model when the true post-change parameter $\theta \geq \theta_1$. As we can see, the upper bound of the detection delay depends on the value of $I_\theta(\epsilon, \alpha)$, which might have different properties depending on whether $\alpha > 0$ (Our proposed $L_\alpha$-CUSUM) or $\alpha = 0$ (Classical CUSUM).

As a concrete example, assume $f_\theta$ is the pdf of the normal distribution $N(\theta, 1)$, $\theta_0 = 0$, $\theta_1 = 1$, we can get

$$I_\theta(\epsilon = 0, \alpha) = \begin{cases} \frac{1}{\alpha \sqrt{1+\alpha}} \left( \frac{1}{\sqrt{2\pi}} \right)^\alpha \left( e^{-\frac{\alpha(\theta-1)^2}{2(1+\alpha)}} - e^{-\frac{\alpha \theta^2}{2(1+\alpha)}} \right), & \text{if } \alpha > 0 \\ \theta - 1/2, & \text{if } \alpha = 0. \end{cases}$$

In this case, when $\alpha = 0$, $I_\theta(\epsilon = 0, \alpha = 0)$ is a monotonic increasing function of $\theta$, which implies the detection delay of the scheme $N_{\alpha=0}(b,d)$ for $\theta \geq \theta_1$ is maximized when $\theta = \theta_1$ (the designed minimal magnitude of the change). However, such property may no longer hold when $\alpha > 0$. Figure 3 plots the curve $I_\theta(0, \alpha)$ as a function of $\theta$ for two different choices of $\alpha = 0.21$ and $0.51$. Both functions $I_\theta(0, \alpha)$ are highly nonlinear: they first increase and then decrease. This implies for robust change-point detection in the present of transient outliers, it will be difficult to detect both smaller changes and very larger changes: the former is consistent with the classical result with $\alpha = 0$, and the latter is a new phenomena as the larger change might be regarded as outliers. This is the price we paid for robust detection in the present of transient outliers. This phenomena is also observed when monitor the dependent data streams under the hidden Markov models (Fuh and Mei, 2015).

So far Theorems 3.1 and 3.2 investigate the statistical properties of our proposed scheme $N_\alpha(b,d)$ in (6) without considering the false alarm constraint $\gamma$ in (3). Let us now investigate the detection delay properties of our proposed scheme $N_\alpha(b,d)$ in (6) under the gross error model in (1), subject to the false alarm constraint $\gamma$ in (3). The following corollary characterizes such detection delay properties under the asymptotic regime when the false alarm constraint $\gamma = \gamma(K) \to \infty$ as the dimension $K \to \infty$ whereas the number $m$ of affected features $m = m(K)$ may or may not go to $\infty$. It also includes the suitable choices of the soft-threshold parameter $d$ and the global detection threshold $b$.

**Corollary 3.1.** Under the assumptions of Theorems 3.1 and 3.2, for a given $\alpha \geq 0$ and given $d \geq 0$, a choice of global detection threshold

$$b_\gamma = \frac{1}{\lambda(\epsilon, \alpha)} \left( \sqrt{\log(4\gamma)} + \sqrt{K \exp\left\{ -\lambda(\epsilon, \alpha)d \right\}} \right)^2,$$
will guarantee that our proposed scheme $N_{\alpha}(b,d)$ satisfies the global false alarm constraint $\gamma$ in (3). Moreover, in the asymptotic regime when the false alarm constraint $\gamma = \gamma(K) \to \infty$ and $m = m(K) << \min(\log \gamma, K)$ as the dimension $K \to \infty$, with $b = b_\gamma$ in (16), a first-order optimal choice of the soft-thresholding parameter $d$ that minimizes the upper bound of detection delay in (15) is

$\begin{align}
d_{opt} &= \frac{1}{\lambda(\epsilon, \alpha)} \left\{ \log \frac{K}{m} + \log \frac{\log \gamma}{m} \right\},
\end{align}$

and the detection delay of the corresponding optimized scheme $N_{\alpha}(b_\gamma, d_{opt})$ in (6) satisfies

$\begin{align}
D_{\epsilon, \theta}(N_{\alpha}(b_\gamma, d_{opt})) \leq \frac{1 + o(1)}{\lambda(\epsilon, \alpha) I_\theta(\epsilon, \alpha)} \left\{ \frac{\log \gamma}{m} + \log \frac{\log \gamma}{m} + \log \frac{K}{m} \right\}.
\end{align}$

Note that on the right-hand side of (18), the dominant order is $\max(\log \frac{\gamma}{m}, \log \frac{K}{m})$, and the second term of $\log \frac{\log \gamma}{m}$ might be negligible. However, we decide to keep it in Corollary 3.1, since this term will help us to compare with some classical results. As research is rather limited in the sequential change-point detection literature in the modern asymptotic regime when the number $K$ of data streams goes to $\infty$. If we compare the optimal soft-thresholding parameter $d_{opt}$ in (17) with the minimum detection delay in (18), the effects of the dimension $K$ are the same, but the effects of the false alarm constraint $\gamma$ are different. Thus, different asymptotic scenarios may arise depending on the asymptotic orders of $\log \frac{K}{m}, \log \frac{\gamma}{m}$ and $\log \frac{\log \gamma}{m}$, and below we consider several extreme cases.

First, let us consider the extreme case when $\log \frac{K}{m} << \log \frac{\log \gamma}{m}$, i.e., $K << \log \gamma$. This is consistent with the classical asymptotic regime when $K$ is fixed and the false alarm constraint $\gamma$ goes to $\infty$. 

**Fig 3.** The value of $I_{\theta}(0, \alpha)$ with two choices of $\alpha = 0.21$ and $\alpha = 0.51$.

**Fig 4.** Search for the optimal $\alpha$.
In this case, for our proposed scheme, the minimum detection delay in (18) is of order $\frac{\log \gamma}{m}$. To be more concrete for the idealized model with $\epsilon = 0, \alpha = 0$, $\lambda(\epsilon = 0, \alpha = 0) = 1$, if the true post-change parameter $\theta = \theta_1$, then $I_{b_{\theta_1}}(\epsilon = 0, \alpha = 0) = I(f_{\theta_1}, f_{\theta_0})$, which is the Kullback-Leibler divergence. Hence based on the Corollary 3.1, the delay of $N_{\alpha=0}(b_{\gamma}, d_{opt})$ would be bounded above by $\frac{1+O(1)}{R(f_{\theta_1}, f_{\theta_0})} \frac{\log \gamma}{m}$. Meanwhile, under the idealized model, for any scheme $T$ satisfying the false alarm constraint $\gamma$ in (3), it is well-known that $D_{\epsilon=0}(T) \geq \frac{1+O(1)}{R(f_{\theta_1}, f_{\theta_0})} \frac{\log \gamma}{m}$ as $\gamma$ goes to $\infty$, see Mei (2010). This suggests that our proposed scheme with $\alpha = 0$ attains the classical asymptotic lower bound under the idealized model with $\epsilon = 0$ and the true post-change parameter $\theta = \theta_1$, in the classic asymptotic regime of $K << \log \gamma$.

Second, let us consider another extreme case when $\log K \gg \log \gamma$, or equivalently, when $\log \gamma << m \log K$. This may occur when the number $m$ of affected data streams is fixed and $\log \gamma = o(\log K)$, i.e., the false alarm constraint $\gamma$ is relatively small as compared to $K$. In this case, both the optimal soft-thresholding parameter $d_{opt}$ in (17) and the minimum detection delay in (18) are of order $\log \frac{K}{m}$, and the impact of the false alarm constraint $\gamma$ is negligible. In other words, our proposed scheme need to take at most $O(\log K)$ observations to detect the sparse post-change scenario when only $m$ out of $K$ data streams are affected. This is consistent with the modern asymptotic regime results in the off-line high-dimensional sparse estimation that $O(\log K)$ observations can fully recover the $K$-dimensional sparse signal, see Candes and Tao (2007).

Third, the other extreme case is when both $\log \frac{K}{m}$ and $\log \frac{\log \gamma}{m}$ have the same order. This can occur if $m = K^{1-\beta}$ and $\log \gamma = K^\zeta$ for some $0 < \beta, \zeta < 1$, which was first investigated in Chan (2017) under the idealized model for Gaussian data. It is interesting to compare our results with those in Chan (2017). Under the idealized model with $\epsilon = 0$, the optimal choice of $\alpha = 0$, and thus our results in Corollary 3.1 showed that the detection delay of our proposed scheme is of order $K^{\zeta+\beta-1} + (\xi + 2\beta - 1) \log K$, which is actually of order $\log K$ if $\frac{1-\zeta}{2} < \beta < 1 - \zeta$ but of order $K^{\zeta+\beta-1}$ if $\zeta + \beta > 1$. These two cases are exactly the assumptions in Theorems 1 and 4 of Chan (2017). While the assumption of $m << \min(\log \gamma, K)$ in Corollary 3.1 corresponds to $\zeta + \beta > 1$, in which our detection delay bound is identical to the optimal detection bound in Chan (2017), it is not difficult to see that the proof of Corollary 3.1 can be extended to the case of $\frac{1-\zeta}{2} < \beta < 1 - \zeta$, in which our results are only slightly weaker than that of Chan (2017) in the sense that the order is the same but our constant coefficient is larger. The latter is understandable because Chan (2017) used the Guassian assumptions extensively to conduct a more careful detection delay analysis than our
results in (15), and his results are refiner for Gaussian data under the idealized model. Meanwhile, our results are more general as they are applicable to any distributions and the gross error models. More importantly, our results give an simpler and more intuitive explanation on those assumptions in the theorems of Chan (2017), and provide a deeper insight of online monitoring large-scale data streams under general settings.

Fourth, from the detection delay point of view, Corollary 3.1 seems to suggest that an ideal choice of $\alpha$ is to maximize $\lambda(\epsilon, \alpha)I_{\theta_1}(\epsilon, \alpha)$ for each and every $\theta \geq \theta_1$, which is impossible. Here we follow the standard change-point or statistical process control (SPC) literature to tune the $\alpha$ value on the boundary $\theta = \theta_1$ as it is often easier to detect smaller changes than larger changes. In this case, we can define an optimal choice of $\alpha$ as the one that maximizes $\lambda(\epsilon, \alpha)I_{\theta_1}(\epsilon, \alpha)$. For the purpose of better illustration, we treat $\alpha = 0$ as the baseline since it corresponds to the classical CUSUM scheme that is optimal under the idealized model. Then relation (18) inspires us to define the asymptotic efficiency improvement of the proposed scheme $N_{\alpha}(b, d)$ with $\alpha \geq 0$ as compared to the baseline scheme $N_{\alpha=0}(b, d)$ as

$$e(\epsilon, \alpha) = \frac{\lambda(\epsilon, \alpha)I_{\theta_1}(\epsilon, \alpha)}{\lambda(\epsilon, \alpha = 0)I_{\theta_1}(\epsilon, \alpha = 0)} - 1$$

Hence, the oracle optimal choice of $\alpha$ can be defined by maximizing the efficiency improvement $e(\epsilon, \alpha)$. That is

$$\alpha_{oracle}(\epsilon) = \arg \max_{\alpha \geq 0}[\lambda(\epsilon, \alpha)I_{\theta_1}(\epsilon, \alpha)] = \arg \max_{\alpha \geq 0}[e(\epsilon, \alpha)]$$

It is non-trivial to derive the theoretical properties of $\alpha_{oracle}$ as a function of $\epsilon$, as it will depend on the relationships between $f_{\theta_0}, f_{\theta_1}$ and the contamination density $g$. But the good news is that the numerical values of $\alpha_{oracle}$ can be found fairly easy. The main tool is the Monte Carlo integration and grid search, and our key idea to simplify computational complexity is to run Monte Carlo simulation once to compute $\lambda(\epsilon, \alpha)$ in (9) and $I_{\theta_1}(\epsilon, \alpha)$ in (12) simultaneously for many possible combinations of $(\epsilon, \alpha)$.

As an illustration, we consider a concrete example when $f_{\theta_0}$ is the pdf of $N(0, 1)$, $f_{\theta_1}$ is the pdf of $N(1, 1)$, $g$ is the pdf of $N(0, 3^2)$. Figure 4 plots $e(\epsilon, \alpha)$ as a function of the tuning parameter $\alpha$ for several fixed $\epsilon$. From Figure 4, it is clear that when $\epsilon = 0$, the $e(\epsilon = 0, \alpha)$ curve (red curve) is linearly decreasing as a function of $\alpha \geq 0$, and thus the optimal choice of $\alpha$ is 0 for $\epsilon = 0$. This is consistent with the optimality properties of the CUSUM statistic under the idealized model.
without outliers. Meanwhile, for any other contamination rate $\epsilon > 0$, the $e(\epsilon, \alpha)$ curve is first increasing and then decreasing as $\alpha$ increases. Thus the optimal choice of $\alpha_{\text{oracle}}$ is often positive when $\epsilon > 0$. For instance, when $\epsilon = 0.1$, Figure 4 (blue curve) shows that $\alpha_{\text{oracle}}(\epsilon = 0.1) \approx 0.21$, and $e(\epsilon = 0.1, \alpha = 0.21) \approx 0.63$. This suggests that our proposed $L_\alpha$-CUSUM based scheme with $\alpha = 0.21$ will be 63% more efficient than the baseline CUSUM based scheme under the gross error model when there are 10% outliers. Figure 5 shows the efficiency improvement of our proposed $L_\alpha$-CUSUM based scheme with $\alpha = 0.21$ under different contamination ratio $\epsilon$ from 0 to 0.15. From the plot, we can see that as compared to the classical CUSUM based method, our proposed $L_\alpha$-CUSUM based scheme with $\alpha = 0.21$ will gain $40\% \sim 70\%$ more efficiency when the contamination ratio $\epsilon \in [2\%, 15\%]$, and the price we pay is to lose $5\%$ efficiency under the idealized model with $\epsilon = 0$.

Note the oracle optimal choice of $\alpha_{\text{oracle}}(\epsilon)$ in (20) requires the full information of the outliers $\epsilon$ and $g$, which may be unknown in practice. In the next section, we will investigate the robustness property of our proposed scheme and provide a practical way to choose $\alpha$, which does not rely on any information of outliers.

4. Breakdown point analysis. 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. For instance, when estimating parameters of a distribution, the breakdown point of the sample mean is 0 since a single outlier
can completely change the value of the sample mean, whereas the breakdown point of the sample median is $1/2$. This suggests that the sample median is more robust than the sample mean.

Since the pioneering work of Hampel (1968) 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 Krasker and Welsch (1982), Rousseeuw (1984). 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 $\gamma$ in (3) under the idealized model with $\epsilon = 0$, then its false alarm rate should not be too bad under the gross error model 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 $\log \mathbb{E}^{(\infty)}_{\theta_0}(T(b))$ and deem the false alarm rate unacceptable if $\log \mathbb{E}^{(\infty)}_{\theta_0}(T(b))$ is much smaller than the designed level of $\log \gamma$, i.e., if $\log \mathbb{E}^{(\infty)}_{\theta_0}(T(b)) = o(\log \gamma)$. Second, what kind of the contamination function $g$ in (22) should we consider in the gross error model? In the previous subsection we investigate the asymptotic properties of our proposed schemes when the contamination distribution $g$ is given. However, this is unsuitable for breakdown point analysis. Here we propose to follow the offline robust statistics literature to consider the $\epsilon$-contaminated distribution class in Huber (1964) that includes any arbitrary contamination functions $g$’s.

To be more rigorous, in and only in this section, we define $\mathbb{E}^{(\infty)}_f$ as the expectation when the observations are i.i.d with pdf $f$, we propose to define the false alarm breakdown point of a family of schemes $T(b)$ as follows.

**Definition 4.1.** Given a family of schemes $T(b)$ with $b = b_\gamma$ satisfying the false alarm constraint $\gamma$ under the idealized model with $\epsilon = 0$, i.e., $\mathbb{E}^{(\infty)}_{f_{\theta_0}}(T(b)) = (1 + o(1)) \gamma$, as $\gamma \to \infty$. The false alarm breakdown point $\epsilon^*(T)$ of $T(b)$’s is defined as

\begin{equation}
\epsilon^*(T) = \inf \{ \epsilon \geq 0 : \inf_{h_0 \in \mathcal{H}_0, \epsilon} \log \mathbb{E}^{(\infty)}_{h_0}(T(b)) = o(\log \gamma) \},
\end{equation}

where the set $\mathcal{H}_{0, \epsilon}$ is the $\epsilon$-contaminated distribution density class of the idealized model $f_{\theta_0}(x)$ for
given $\epsilon \in [0, 1)$, and is defined as

$$h_{0,\epsilon} = \{h|h = (1 - \epsilon)f_{\theta_0} + \epsilon g, g \in G\},$$

and $G$ denotes the class of all probability densities on the data $X_{k,n}$’s.

Now we are ready to conduct the false alarm breakdown point analysis for our proposed scheme $N_\alpha(b,d)$ in (6) with a given tuning parameter $\alpha \geq 0$. To do so, for the densities $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$, and for any given $\alpha \geq 0$, we define an intrinsic bound

$$M(\alpha) = \text{ess sup}_x \frac{[f_{\theta_1}(x)]^\alpha - [f_{\theta_0}(x)]^\alpha}{\alpha},$$

and the density power divergence between $f_{\theta_0}$ and $f_{\theta_1}$:

$$d_\alpha(\theta_0, \theta_1) = \int \left\{ [f_{\theta_1}(x)]^{1+\alpha} - (1 + \frac{1}{\alpha})f_{\theta_0}(x)[f_{\theta_1}(x)]^\alpha + \frac{1}{\alpha} [f_{\theta_0}(x)]^{1+\alpha} \right\} dx.$$ 

Note that $d_\alpha(f_{\theta_0}, f_{\theta_1})$ was proposed in Basu et al. (1998), which showed that it is always positive when $f_{\theta_1}$ and $f_{\theta_0}$ are different. Moreover, when $\alpha = 0$, $d_{\alpha=0}(\theta_0, \theta_1)$ becomes Kullback-Leibler information number $I(f_{\theta_0}, f_{\theta_1}) = \int f_{\theta_0}(x) \log \frac{f_{\theta_0}(x)}{f_{\theta_1}(x)} dx$.

With these two new notations, the following theorem derives the false alarm breakdown point of our proposed schemes $N_\alpha(b,d)$ as a function of the tuning parameter $\alpha$ for a fixed soft-thresholding parameter $d$ when online monitoring a given $K$ number of data streams.

**Theorem 4.1.** 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 scheme $N_\alpha(b,d)$ is given by

$$\epsilon^*(N_\alpha) = \frac{d_\alpha(\theta_0, \theta_1)}{d_{\alpha=0}(\theta_0, \theta_1) + (1 + \alpha)M(\alpha)},$$

where $M(\alpha)$ and $d_\alpha(\theta_0, \theta_1)$ are defined in (23) and (24). In particular, $\epsilon^*(N_\alpha) = 0$ if $M(\alpha) = \infty$ and $d_\alpha(\theta_0, \theta_1)$ is finite.

The proof of Theorem 4.1 will be presented in subsection 7.4. Here let us apply the results for widely used normal distributions, i.e., when $f_\theta$ is the pdf of $N(\theta, \sigma^2)$. In this case, when $\alpha = 0$, the density power divergence $d_{\alpha=0}(\theta_0, \theta_1) = \frac{1}{2\sigma^2} (\theta_1 - \theta_0)^2$ is finite, but the bound $M(\alpha = 0)$ in
(23) 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

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

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

and thus it is not difficult from (24) to show that,

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

Moreover, if we let $M(= 1/\sqrt{2\pi\sigma^2})$, then $|f_{\theta}(x)| \leq M$ for all $x$. By the definition in (23), 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 $L_\alpha$-CUSUM based scheme with $\alpha > 0$ is much more robust than the classical CUSUM scheme.

Note the false alarm breakdown point of our proposed scheme does not require any information about the contamination ratio $\epsilon$ and contamination distribution $g$. Therefore, we proposed to choose the optimal robustness parameter $\alpha$ which maximizes the false alarm breakdown point in (25). That is

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

To be more specific, let us use the same example when $f_{\theta_0} \sim N(0,1)$ and $f_{\theta_1} \sim N(1,1)$. By (27), 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 (23), its numerical value can be easily found by brute-force exhaustive search over the real line $x \in (-\infty, \infty)$. Figure 6 shows the false alarm breakdown point of our proposed scheme $N_\alpha(b,d)$ when $\alpha$ varies from 0 to 2. We can see clearly the breakdown point will first increase and then decrease, which yields the optimal choice of $\alpha_{\text{opt}}$ 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.

It is interesting to compare the performance of the two choices of $\alpha_{oracle}$ in (20) and $\alpha_{opt}$ in (28). By the previous subsection, when $\epsilon = 0.1$ and contamination distribution is $N(0, 3^2)$, we get $\alpha_{oracle} = 0.21$ with the efficiency improvement as 63%. If we use $\alpha_{opt} = 0.51$, we will get the corresponding efficiency improvement as 55%, which makes sense because $\alpha_{opt}$ uses the full information of the outliers. However, from Theorem 4.1 and Figure 6, we can get the false alarm breakdown point of our proposed scheme with the choice of $\alpha_{oracle} = 0.21$ is 0.217, which implies $\alpha_{oracle}$ can tolerate less arbitrarily contaminations than the choice of $\alpha_{opt}$. In the next section, we will also compare the performance of the two choices of $\alpha$ by conducting simulation studies.

5. Numerical Simulations. In this section we conduct extensive numerical simulation studies to illustrate the robustness and efficiency of our proposed scheme $N_\alpha(b, d)$ in (6).

In our simulation studies, we assume that there are $K = 100$ independent features, and at some unknown time, $m = 10$ features are affected by the occurring event. Also the change is instantaneous if a feature is affected, and we do not know which subset of features will be affected. In our simulations below, we set $f_\theta = \text{pdf of } N(\theta, 1)$. Then pre-change parameter $\theta_0 = 0$, the minimal magnitude of the change $\theta_1 = 1$, and the contamination density $g = \text{pdf of } N(0, 3^2)$. Our proposed scheme $N_\alpha(b, d)$ in (6) is constructed by using the density function $f_{\theta_0}$ and $f_{\theta_1}$.

We conduct four different simulation studies based on the gross error model in (1) with different values of the contamination rate $\epsilon$. In the first one, we consider the case when the true post-change parameter $\theta = \theta_1 = 1$, $\epsilon = 0.1$, and the objective is to illustrate that with optimized tuning parameters, our proposed robust scheme $N_\alpha(b, d)$ in (6) will have better detection performance than the other comparison methods in the presence of outliers. In the second one, we consider the case when $\theta = \theta_1 = 1$, $\epsilon = 0$ to demonstrate that our proposed robust scheme in the first experiment does not lose much efficiency under the idealized model. In the third simulation study, we illustrates that the false alarm rate of our proposed robust scheme indeed is more stable as compared to those CUSUM- or likelihood-ratio- based methods as the contamination rate $\epsilon$ in (1) varies. In the last simulation study, we investigate the sensitivity of our proposed scheme $N_\alpha(b, d)$ when the true post-change parameter $\theta$ is greater than $\theta_1$. The detailed simulation results under these three simulation studies are presented below.
In our first simulation study, we consider the case when \( \epsilon = 0.1 \), e.g., 10\% of data are from the outlier distribution \( N(0, 3^2) \). In this case, for our proposed robust scheme \( N_\alpha(b, d) \) in (6), as shown in previous sections, the two optimal choices of \( \alpha \) are \( \alpha_{\text{oracle}}(\epsilon = 0.1) = 0.21 \) and \( \alpha_{\text{opt}} = 0.51 \). By (17), if \( \log(\gamma) << K \), then the corresponding optimal shrinkage parameters \( d \approx \frac{1}{\lambda(\epsilon = 0.1, \alpha = 0.21)} \log \frac{K}{m} = 1.6831 \), \( d \approx \frac{1}{\lambda(\epsilon = 0.1, \alpha = 0.51)} \log \frac{K}{m} = 0.9684 \) for \( K = 100 \) and \( m = 10 \), since \( \lambda(\epsilon = 0.1, \alpha = 0.21) = 1.3681 \) and \( \lambda(\epsilon = 0.1, \alpha = 0.51) = 2.3777 \). For the baseline CUSUM-based scheme, i.e., \( N_{\alpha=0}(b, d) \) with \( \alpha = 0 \), we consider two different choices of the shrinkage parameter \( d \): one designed for \( \epsilon = 0.1 \) and the other designed for \( \epsilon = 0 \). Since \( \lambda(\epsilon = 0.1, \alpha = 0) = 0.4572 \) and \( \lambda(\epsilon = 0, \alpha = 0) = 1 \), by (17), we derive two optimal \( d \) values for the baseline scheme: \( d \approx \frac{1}{\lambda(\epsilon = 0.1, \alpha = 0)} \log \frac{K}{m} = 5.0363 \). and \( d \approx \frac{1}{\lambda(\epsilon = 0, \alpha = 0)} \log \frac{K}{m} = 2.3026 \).

In summary, we will compare the following eight different schemes.

- Our proposed scheme \( N_\alpha(b, d) \) in (6) with \( \alpha_{\text{oracle}} = 0.21 \) and \( d = 1.6831 \) optimized for \( m = 10 \) and \( \epsilon = 0.1 \);
- Our proposed scheme \( N_\alpha(b, d) \) in (6) with \( \alpha_{\text{opt}} = 0.51 \) and \( d = 0.9684 \) optimized for \( m = 10 \) and \( \epsilon = 0.1 \);
- The baseline CUSUM-based scheme \( N_{\alpha=0}(b, d) \) with \( d = 2.306 \) optimized for \( m = 10 \) and \( \epsilon = 0 \);
- The baseline CUSUM-based scheme \( N_{\alpha=0}(b, d) \) with \( d = 5.0363 \) optimized for \( m = 10 \) and \( \epsilon = 0.1 \);
- The MAX scheme \( N_{\alpha=0.21,\text{max}}(b) \) in (7);
- The SUM scheme \( N_{\alpha=0.21,\text{sum}}(b) \) in (8);
- The method \( N_{XS}(b, p_0 = 0.1) \) in Xie and Siegmund (2013) based on generalized likelihood ratio:
\[
N_{XS}(b, p_0) = \inf \left\{ n \geq 1 : \max_{0 \leq i < n} \sum_{k=1}^{K} \log \left( 1 - p_0 + p_0 \exp \left[ \left( U_{k,n,i}^+ \right)^2 / 2 \right] \right) \geq b \right\},
\]
where for all \( 1 \leq k \leq K, 0 \leq i < n \),
\[
U_{k,n,i}^+ = \max \left( 0, \frac{1}{\sqrt{n-i}} \sum_{j=i+1}^{n} X_{k,j} \right).
\]
- The method \( N_{\text{Chan,1}}(b) \) in Chan (2017) under the idealized model that is an extension of the
Table 1: A comparison of the detection delays of 9 schemes with $\gamma = 5000$ under the gross error model. The smallest and largest standard errors of these 9 schemes are also reported under each post-change hypothesis based on 1000 repetitions in Monte Carlo simulations.

<table>
<thead>
<tr>
<th></th>
<th># affected local data streams</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>30</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Smallest standard error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.43</td>
<td>0.16</td>
<td>0.10</td>
<td>0.07</td>
<td>0.06</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Largest standard error</td>
<td></td>
<td>1.35</td>
<td>0.35</td>
<td>0.27</td>
<td>0.22</td>
<td>0.22</td>
<td>0.17</td>
<td>0.14</td>
<td>0.12</td>
<td>0.12</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Our proposed robust scheme

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\alpha=0.21}(b = 16.40, d = 1.6831)$</td>
<td>46.2</td>
<td>21.1</td>
<td>15.1</td>
<td>11.4</td>
<td>10.1</td>
<td>8.2</td>
<td>7.2</td>
<td>6.0</td>
<td>4.9</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>$N_{\alpha=0.51}(b = 9.26, d = 0.9684)$</td>
<td>49.3</td>
<td>22.6</td>
<td>16.2</td>
<td>12.2</td>
<td>10.9</td>
<td>8.9</td>
<td>7.8</td>
<td>6.5</td>
<td>5.3</td>
<td>4.2</td>
<td></td>
</tr>
</tbody>
</table>

Other methods for comparison

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\alpha=0}(b = 84.74, d = 2.3026)$</td>
<td>94.5</td>
<td>41.0</td>
<td>27.6</td>
<td>19.7</td>
<td>17.0</td>
<td>12.9</td>
<td>10.9</td>
<td>8.6</td>
<td>6.5</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>$N_{\alpha=0}(b = 41.51, d = 5.0363)$</td>
<td>74.7</td>
<td>35.1</td>
<td>25.1</td>
<td>19.1</td>
<td>16.9</td>
<td>13.7</td>
<td>12.0</td>
<td>10.1</td>
<td>8.3</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>$N_{\alpha=0.21,\text{max}}(b = 8.16)$</td>
<td>31.5</td>
<td>21.8</td>
<td>19.4</td>
<td>17.5</td>
<td>16.8</td>
<td>15.8</td>
<td>15.1</td>
<td>14.3</td>
<td>13.4</td>
<td>12.4</td>
<td></td>
</tr>
<tr>
<td>$N_{\alpha=0.21,\text{sum}}(b = 70.25)$</td>
<td>70.9</td>
<td>29.7</td>
<td>19.8</td>
<td>13.8</td>
<td>11.6</td>
<td>8.7</td>
<td>7.0</td>
<td>5.3</td>
<td>3.7</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>$N_{\text{Chan,1}}(b = 22.55, p_0 = 0.1)$</td>
<td>74.7</td>
<td>35.7</td>
<td>25.3</td>
<td>19.1</td>
<td>16.9</td>
<td>13.4</td>
<td>11.5</td>
<td>9.3</td>
<td>7.2</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td>$N_{\text{Chan,2}}(b = 48.7, p_0 = 0.1)$</td>
<td>407.3</td>
<td>86.4</td>
<td>55.5</td>
<td>38.4</td>
<td>32.9</td>
<td>24.2</td>
<td>19.8</td>
<td>14.9</td>
<td>10.3</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td><em>(Standard error)</em></td>
<td>12.1</td>
<td>(0.76)</td>
<td>(0.53)</td>
<td>(0.3)</td>
<td>(0.25)</td>
<td>(0.19)</td>
<td>(0.15)</td>
<td>(0.1)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td></td>
</tr>
<tr>
<td>$N_{\chi^2}(b = 130, p_0 = 0.1)$</td>
<td>290.6</td>
<td>97.5</td>
<td>58.3</td>
<td>38.4</td>
<td>32</td>
<td>22.7</td>
<td>17.6</td>
<td>12.7</td>
<td>8.1</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td><em>(Standard error)</em></td>
<td>(5.85)</td>
<td>(2.21)</td>
<td>(1.12)</td>
<td>(0.68)</td>
<td>(0.64)</td>
<td>(0.41)</td>
<td>(0.31)</td>
<td>(0.22)</td>
<td>(0.15)</td>
<td>(0.08)</td>
<td></td>
</tr>
</tbody>
</table>

SUM scheme in Mei (2010):

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

where $W_{k,n}^*$ is the CUSUM statistics in (5).

- The method $N_{\text{Chan,2}}(b, p_0 = 0.1)$ in Chan (2017) which is similar as $N_{\chi^2}(b, p_0)$:

$$N_{\text{Chan,2}}(b, p_0) = \inf \left\{ n \geq 1 : \max_{0 \leq i < n} \sum_{k=1}^{K} \log \left( 1 - p_0 + 2(\sqrt{2} - 1)p_0 \exp \left[ \left( U_{k,n,i}^+ \right)^2 / 2 \right] \right) \geq b \right\}.$$  

For each of these 9 schemes $T(b)$, we first find the appropriate values of the threshold $b$ to satisfy the false alarm constraint $\gamma \approx 5000$ under the gross error model in (1) with $\epsilon = 0.1$ (within the range of sampling error). Next, using the obtained global threshold value $b$, we simulate the detection delay 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 simulated detection delays of these nine schemes under 10 different post-change hypothesis, depending on different numbers of affected local data streams. Since our proposed scheme $N_{\alpha=0.21}(b, d = 1.6831)$ is optimized for the case when $m = 10$ out of data streams...
are affected under the gross error models, it is not surprising that it indeed has the smallest detection delays among all comparison methods when 10 data streams are affected. In particular, our proposed schemes $N_\alpha(b,d)$ have much smaller detection delay than the three CUSUM-based schemes $N_{\alpha=0}(b,d = 5.0363)$, $N_{\alpha=0}(b,d = 2.3026)$ and $N_{\text{Chan},1}(b,p_0 = 0.1)$. This illustrates that the improvement of $L_\alpha$-CUSUM statistics with $\alpha = 0.21$ is significant as compared to the baseline CUSUM statistics in the presence of outliers.

Moreover, compared with the choice of $\alpha_{\text{oracle}} = 0.21$, our proposed scheme with $\alpha_{\text{opt}} = 0.51$ yields overall larger detection delays under those 10 different post-change hypothesis. This is consistent to the previous discussion that $\alpha_{\text{oracle}}$ would be better than $\alpha_{\text{opt}}$ when the contamination ratio $\epsilon$ and contamination distribution $g$ are known. Note $\alpha_{\text{opt}} = 0.51$ does not use any information about $\epsilon$ and $g$ but still led smaller detection delays than the two baseline CUSUM-based schemes $N_{\alpha=0}(b,d = 5.0363)$ and $N_{\alpha=0}(b,d = 2.3026)$, which suggests the usefulness of $\alpha_{\text{opt}}$, especially when the contaminations are unknown.

In addition, the detection delays of the two likelihood-ratio-based methods $N_{\text{XS}}(b,p_0)$ and $N_{\text{Chan},2}(b,p_0)$ are extremely large, especially when the number of affected data stream is small. The reason is that they do not suppose that $f_{\theta_1} = N(1,1)$ is known and are designed to be efficient against $f_{\theta} = N(\theta,1)$ for all $\theta > 0$. Hence they want to detect say $f_{\theta} = N(3,1)$ quickly as well. Due to the presence of outliers, a significant proportion of the observations have values close to 3 and these two methods, $N_{\text{XS}}(b,p_0)$ and $N_{\text{Chan},2}(b,p_0)$, will take this into the consideration and detect a possible change of distribution to $f_{\theta} = N(3,1)$ having occurred. Since the detection delays of $N_{\text{XS}}(b,p_0 = 0.1)$ and $N_{\text{Chan},2}(b,p_0 = 0.1)$ are very large, we use separate rows in Table 1 to show the standard deviation of their detection delays.

It is also interesting to note that the MAX-scheme $N_{\alpha=0.21,\text{max}}(b)$ and the SUM-scheme $N_{\alpha=0.21,\text{sum}}(b)$ are designed for the case when $m = 1$ or $m = K$ features are affected, and Table 1 confirmed that their detection delays are indeed the smallest in their respective designed scenarios. However, when the number of affected features $m$ is moderate, our proposed scheme $N_{\alpha=0.21}(b,d)$ will have smaller detection delay, which implies our proposed scheme with soft-thresholding transformation could be more robust to the number of affected features.

Next, for our proposed robust scheme $N_\alpha(b,d)$ with two choices of $\alpha_{\text{oracle}}, \alpha_{\text{opt}}$, we want to investigate how much efficiency it will lose as compared to the other seven schemes under the idealized model with $\epsilon = 0$. We re-calculate the threshold $b$ for each of these schemes $T(b)$, so as to
satisfy the false alarm constraint $\gamma \approx 5000$ under the idealized model with $\epsilon = 0$.

Table 2 summarizes the results of our second simulation study on the detection delays of these 9 schemes under 10 different post-change hypotheses. Among all schemes, $N_{XS}(b, p_0)$ and $N_{Chan,2}(b, p_0)$ generally yield the competing smallest detection delay. However, we want to emphasize that both schemes are computationally expensive. Specifically, even if we use a time window of size $k$ as in Chan (2017) to speed up the implementation of $N_{XS}(b, p_0)$ and $N_{Chan,2}(b, p_0)$, at each time $n$, $O(Kk^2)$ computations are needed to get the global monitoring statistics, whereas our proposed scheme only require $O(K)$ computations to get the global monitoring statistics. For instance, for a given global threshold $b$ around 4.25, it took about 130 minutes on average to finish 1000 Monte Carlo simulation runs in our laptop. If we did not know $b \approx 4.25$ and wanted to search for 10 different values of $b$'s by bisection method based on 1000 Monte Carlo runs for each $b$, it would have taken about $10 \times 130 = 1300$ computer minutes for the case of $\gamma = 5000$. Meanwhile, due to the nice recursive formula, our proposed schemes can be implemented in real-time. For instance, it took about 15 minutes to find such threshold $b$ from a range of values for our proposed schemes based on 1000 Monte Carlo runs (the time is shorter if our initial guess range of $b$ is closer) and all of these simulations are conducted on a Windows 10 Laptop with Intel i5-6200U CPU 2.30 GHz.

In addition, under the idealized model with $\epsilon = 0$, the corresponding $\alpha_{oracle} = 0$, which suggest that the baseline CUSUM scheme $N_{\alpha=0}(b, d = 2.3026)$ should have good performance when $m = 10$ data streams are affected. Moreover, in corollary 3.1, we show the detection delay of our proposed scheme nearly achieves the optimal detection lower bound in Chan (2017), which can be validated from the numerical results in Table 2 since it compares well with the best possible method.

Another interesting observation from Table 2 is that the detection delay of our proposed robust scheme $N_{\alpha=0.21}(b, d = 1.6831)$ is comparable with that of $N_{\alpha=0}(b, d = 2.3026)$, and it just takes 6.3% more time steps to raise a correct global alarm under the idealized model when $m = 10$ data streams are affected. Recall that in Table 1, $N_{\alpha=0}(b, d = 2.3026)$ takes 68.3% more time steps than $N_{\alpha=0.21}(b, d = 1.6831)$ to raise a global alarm under the gross error model with $\epsilon = 0.1$. In other words, our proposed robust scheme $N_{\alpha=0.21}(b, d = 1.6831)$ sacrifices about 6.3% efficiency under the idealized model with $\epsilon = 0$, but can gain 68.3% efficiency under the gross error model with proportion of outliers $\epsilon = 0.1$.

In the third experiment, we want to investigate the impact of contamination rate $\epsilon$ on the false alarms, and illustrate the robustness of our proposed $L_\alpha$-CUSUM statistics with respect to
A comparison of the detection delays of 9 schemes with \( \gamma = 5000 \) under the idealized model. The smallest and largest standard errors of these 9 schemes are also reported under each post-change hypothesis based on 1000 repetitions in Monte Carlo simulations.

<table>
<thead>
<tr>
<th>Gross error model with ( \epsilon = 0 )</th>
<th># affected local data streams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smallest standard error</td>
<td>1</td>
</tr>
<tr>
<td>Smallest standard error</td>
<td>0.29</td>
</tr>
<tr>
<td>Largest standard error</td>
<td>0.58</td>
</tr>
<tr>
<td>Our proposed robust scheme</td>
<td></td>
</tr>
<tr>
<td>( N_{\alpha=0}(b = 11.69, d = 1.6831) )</td>
<td>33.5</td>
</tr>
<tr>
<td>( N_{\alpha=0.51}(b = 7.63, d = 0.9684) )</td>
<td>39.4</td>
</tr>
</tbody>
</table>

Comparison of other methods

| \( N_{\alpha=0}(b = 21.52, d = 2.3026) \) | 33.6 | 15.2 | 11.0 | 8.4 | 7.5 | 6.1 | 5.3 | 4.5 | 3.7 | 3.0 |
| \( N_{\alpha=0}(b = 7.35, d = 5.0363) \) | 22.4 | 13.8 | 11.1 | 9.3 | 8.6 | 7.6 | 7.0 | 6.3 | 5.5 | 4.8 |
| \( N_{\alpha=0.21,\text{max}}(b = 7.14) \) | 24.4 | 17.1 | 15.4 | 14.1 | 13.6 | 12.8 | 12.2 | 11.6 | 10.9 | 10.2 |
| \( N_{\alpha=0.21,\text{sum}}(b = 58.81) \) | 56.0 | 23.2 | 15.5 | 10.8 | 9.1 | 6.8 | 5.6 | 4.2 | 3.0 | 2.0 |
| \( N_{\text{chan},1}(b = 3.44, p_0 = 0.1) \) | 26.7 | 14.2 | 10.9 | 8.6 | 7.8 | 6.3 | 5.5 | 4.5 | 3.4 | 2.3 |
| \( N_{\text{chan},2}(b = 4.25, p_0 = 0.1) \) | 26.3 | 13.1 | 9.7 | 7.2 | 6.3 | 4.8 | 3.9 | 2.9 | 2.0 | 1.1 |
| \( N_{XS}(b = 19.5, p_0 = 0.1) \) | 30.9 | 13.2 | 9.2 | 7.2 | 5.7 | 4.7 | 3.5 | 2.5 | 1.8 | 1.0 |

\( \epsilon \). Since the MAX-scheme \( N_{\alpha=0.21,\text{max}}(b = 7.14) \) and the SUM-scheme \( N_{\alpha=0.21,\text{sum}}(b = 58.81) \) are based on local \( L_\alpha \)-CUSUM statistics, their robustness properties to the outliers are similar to our proposed scheme \( N_{\alpha=0.21}(b = 11.69, d = 1.6831) \) and \( N_{\alpha=0.51}(b = 7.63, d = 0.9684) \). To highlight the robustness of our proposed \( L_\alpha \)-CUSUM statistics, we compare our proposed schemes \( N_{\alpha=0.21}(b = 11.69, d = 1.6831) \) and \( N_{\alpha=0.51}(b = 7.63, d = 0.9684) \) with other four schemes: two baseline CUSUM schemes and Chan’s two methods.

Figure 7 reports the curve of \( \log E^{(\infty)}_{\theta_0}(T) \) as the contamination ratio \( \epsilon \) varies from 0.02 to 0.2 with stepsize 0.02. It is clear from the figure that 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 \( L_\alpha \)-CUSUM statistics-based method with \( \alpha_{\text{oracle}} = 0.21 \) and \( \alpha_{\text{opt}} = 0.51 \) decrease rather slowly. This suggests that our proposed scheme is more robust in the sense of keeping \( \log E^{(\infty)}_{\theta_0}(T) \) more stable with a small departure from the assumed model. Moreover, note the curve for \( \alpha_{\text{opt}} = 0.51 \) decreases slower than the curve for \( \alpha_{\text{oracle}} = 0.21 \), which implies the performance of \( \alpha_{\text{opt}} \) is better than \( \alpha_{\text{oracle}} \) in term of keeping the false alarm constraint stable to the contaminations.

In the last experiment, we focus on the sensitivity of our proposed scheme \( N_\alpha(b, d) \) with the misspecified post-change parameter \( \theta \). Specifically, we fix the number of affected features \( m = 10 \) and set the true post-change parameter \( \theta \) to be 1, 1.5, 2, 2.5, and 3. Then, we simulate the detection
delay of our proposed schemes $N_{\alpha=0.21}(b = 11.69, d = 1.6831)$, $N_{\alpha=0.51}(b = 7.63, d = 0.9684)$, the CUSUM-based scheme $N_{\alpha=0}(b = 84.74, d = 2.3026)$, $N_{\text{Chan},1}(b = 22.55, p_0 = 0.1)$, $N_{\text{Chan},2}(b = 48.7, p_0 = 0.1)$ and $N_{\mu_S}(b = 130, p_0 = 0.1)$. The results are summarized in Table 3. First, we can see although $\alpha_{\text{oracle}} = 0.21$ is designed to be optimal when the true post-change parameter $\theta = \theta_1 = 1$ with $\epsilon = 0.1$ and $g = N(0, 3^2)$, it still has the smallest detection delay among those three schemes with the true change parameter is larger than 1. Second, although the overall performance of our proposed scheme with the choice of $\alpha$ to be $\alpha_{\text{opt}} = 0.51$ is not as good as the the choice of $\alpha$ to be $\alpha_{\text{oracle}} = 0.21$, it still has a smaller detection delay than the CUSUM-based method when the true post-change post-change parameter is smaller than 2. Moreover, it does not use any knowledge of outliers $\epsilon$ and $g$. Those results demonstrate that generally our proposed scheme $N_{\alpha}(b, d)$ are not sensitive to the small misspecified post-change parameter $\theta$.

6. Case study. In this section, we conduct a case study based on a real dataset of tonnage signal collected from a progressive forming manufacturing process. The dataset includes 307 normal samples and 2 different groups of fault samples. Each group contains 69 samples which are collected under the faults due to missing part occurring in the forming station (hereafter called Fault #1) and the pre-forming station (hereafter called Fault #2). Additionally, there are $p = 2^{11} = 2048$ measurement points in each tonnage signal. We want to build efficient monitoring scheme to detect the faults due to missing part occurring in the forming station while avoid making false alarm on
which minimizes efficiency improvement for $\epsilon$ coefficients to effectively detect some small but persistent changes.

two selected standardized Haar coefficients. Clearly, we may not detect the fault 1 samples if we

on the

Specifically, discrete Haar transform is applied on each tonnage signal data and we just keep the

explicit interpretation of the changes in the profile observations, see Zhou, Sun and Shi (2006) as

nonlinear profile data (Fan, 1996; Zhou, Sun and Shi, 2006; Lee et al., 2012). In this article, Haar

the random fault #2 samples.

In literature, wavelet-based approaches have been widely used for analyzing and monitoring nonlinear profile data (Fan, 1996; Zhou, Sun and Shi, 2006; Lee et al., 2012). In this article, Haar transform is chosen as an illustration of our proposed scheme because Haar coefficients have an explicit interpretation of the changes in the profile observations, see Zhou, Sun and Shi (2006) as an example about applying Haar transform and the physical interpretation of the Haar coefficients. Specifically, discrete Haar transform is applied on each tonnage signal data and we just keep the first $p = 512$ Haar coefficients.

We use $c_{k,n}$ denotes the $k^{th}$ Haar coefficient of the $n^{th}$ tonnage signal data. Then we consider the normalized standardized Haar coefficients by

$$X_{k,n} = \frac{c_{k,n} - \hat{\mu}_k}{\hat{\sigma}_k},$$

(29)

where $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are the sample mean and variance of all in-control normal tonnage signal data on the $k^{th}$ Haar coefficient. Figure 8 shows the projection of all normal and faulty samples on two selected standardized Haar coefficients. Clearly, we may not detect the fault 1 samples if we just using the first Haar coefficient. This illustrates the necessary to monitor a large number of coefficients to effectively detect some small but persistent changes.

After standardizing those Haar coefficients, we assume those $X_{k,n}$’s are i.i.d with standard normal distribution $N(0, 1)$ for the in-control tonnage samples and have some mean shifts for those faulty tonnage samples. To apply our proposed scheme, we set $\theta_1 = 1$, i.e., the minimal magnitude of shift is 1 and the number of affected coefficients $m = 50$. We will use our proposed scheme with the choice of $\alpha = 0.51$, which maximizes the false alarm breakdown point, and the choice of $\alpha = 0.21$, which minimizes efficiency improvement for $\epsilon = 0.1$ and $g = N(0, 3^2)$. We compare the performance

<table>
<thead>
<tr>
<th>True post-change $\theta$ value</th>
<th>$\theta = 1$</th>
<th>$\theta = 1.5$</th>
<th>$\theta = 2$</th>
<th>$\theta = 2.5$</th>
<th>$\theta = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{a=0.21}(b = 16.40, d = 1.6831)$</td>
<td>10.1 ± 0.06</td>
<td>6.5 ± 0.03</td>
<td>5.2 ± 0.02</td>
<td>4.6 ± 0.02</td>
<td>4.5 ± 0.01</td>
</tr>
<tr>
<td>$N_{a=0.51}(b = 9.26, d = 0.9684)$</td>
<td>10.9 ± 0.06</td>
<td>7.4 ± 0.03</td>
<td>6.4 ± 0.02</td>
<td>6.5 ± 0.02</td>
<td>7.4 ± 0.02</td>
</tr>
<tr>
<td>CUSUM-based scheme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{a=0}(b = 84.74, d = 2.3026)$</td>
<td>17.0 ± 0.08</td>
<td>10.0 ± 0.05</td>
<td>7.2 ± 0.03</td>
<td>5.7 ± 0.02</td>
<td>4.8 ± 0.02</td>
</tr>
<tr>
<td>$N_{Chan,1}(b = 22.55, p_0 = 0.1)$</td>
<td>16.8 ± 0.10</td>
<td>9.9 ± 0.04</td>
<td>7.1 ± 0.03</td>
<td>5.6 ± 0.02</td>
<td>4.7 ± 0.02</td>
</tr>
<tr>
<td>$N_{Chan,2}(b = 48.7, p_0 = 0.1)$</td>
<td>32.8 ± 0.18</td>
<td>14.9 ± 0.08</td>
<td>8.5 ± 0.05</td>
<td>5.6 ± 0.03</td>
<td>3.9 ± 0.02</td>
</tr>
<tr>
<td>$N_{XS}(b = 130, p_0 = 0.1)$</td>
<td>32.3 ± 0.61</td>
<td>14.7 ± 0.26</td>
<td>8.4 ± 0.15</td>
<td>5.6 ± 0.08</td>
<td>3.9 ± 0.07</td>
</tr>
</tbody>
</table>
of those two choices of $\alpha$ with the baseline CUSUM-based scheme $N_{\alpha=0}(b,d)$, Xie and Siegmund method $N_{XS}(b,p_0 = 0.1)$ and Chan’s two methods $N_{Chan,1}(b,p_0 = 0.1)$ and $N_{Chan,2}(b,p_0 = 0.1)$. All of those schemes are conducted by using the normalized Haar coefficients data $X_{k,n}$ in (29).

To evaluate the detection efficiency of those methods, we first find the appropriate values of the global threshold $b$ such that the average run length of each scheme is 300 when the samples are collected by sampling from the 307 in-control tonnage samples with probability 90% and from the 69 Fault #1 tonnage samples with probability 10%. Then, using the obtained global threshold value $b$, we simulate the detection delay when the samples are sequentially collected by sampling from the 69 Fault #1 tonnage samples with probability 90% and from the Fault 2 tonnage samples with probability 10%. All Monte Carlo simulations are based on 100 repetitions. The results of detection delay and standard error are summarized in Table 4.

From Table 4, we can see our proposed schemes yield very small detection delay for detecting the smaller persistent change caused by Fault #1 compared with other methods. Thus, they are robust to the larger but transient change caused by Fault #2.

7. Appendix. In this section, we provide the detailed proofs for Theorem 3.1, Theorem 3.2, Theorem 4.1 and Corollary 3.1.

7.1. Proof of Theorem 3.1. For any $x \geq 0$, by Chebyshev’s inequality,

$$E_{\epsilon}[N_{\alpha}(b,d)] \geq xP_{\epsilon}^{(\infty)}(N_{\alpha}(b,d) \geq x)$$

$$= x \left[ 1 - P_{\epsilon}^{(\infty)}(N_{\alpha}(b,d) < x) \right]$$

$$= x \left[ 1 - P_{\epsilon}^{(\infty)}(\sum_{k=1}^{K} \max\{0, W_{\alpha,k,n} - d\} \geq b) \right]$$

for some $1 \leq n \leq x$.

$$\geq x \left[ 1 - xP_{\epsilon}^{(\infty)}(\sum_{k=1}^{K} \max\{0, W_{\alpha,k}^* - d\} \geq b) \right],$$

(30)
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 $P_{\epsilon}^{(\infty)}$. Now if we define the log-moment generating function of the $W_{\alpha,k}^*$’s

$$\psi_{\alpha}(\theta) = \log E_{\epsilon}^{(\infty)} \exp\{\theta \max(0, W_{\alpha,k}^* - d)\}$$

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

$$\exp(K\psi_{\alpha}(\theta)) = E_{\epsilon}^{(\infty)} \exp\{\theta \sum_{k=1}^{K} \max(0, W_{\alpha,k}^* - d)\} \geq e^{\theta b} P_{\epsilon}^{(\infty)} (\sum_{k=1}^{K} \max\{0, W_{\alpha,k}^* - d\} \geq b)$$

for $\theta > 0$. Combining (30) and (32) yields

$$E_{\epsilon}^{(\infty)}[N_{\alpha}(b, d)] \geq x [1 - x \exp(-\theta b + K\psi_{\alpha}(\theta))]$$

for all $x \geq 0$. Since $x(1 - xu)$ is maximized at $x = 1/(2u)$ with the maximum value $1/(4u)$. We conclude from (33) that

$$E_{\epsilon}^{(\infty)}[N_{\alpha}(b, d)] \geq \frac{1}{4} \exp (\theta b - K\psi_{\alpha}(\theta)).$$

for any $\theta > 0$ as long as $\psi_{\alpha}(\theta)$ in (31) is well-defined.

The remaining proof is to utilize the assumption of $\lambda(\epsilon, \alpha) > 0$ in (12) in Assumption (3.2) to show that the upper limiting $W_{\alpha,k}^*$ of the proposed $L_{\alpha}$-CUSUM statistics is well-defined and derive a careful analysis of $\psi_{\alpha}(\theta)$ in (31). When $\alpha = 0$, the $L_{\alpha}$-CUSUM statistics become the classical CUSUM statistics, and the corresponding analysis is well-known, see Liu, Zhang and Mei (2017).

Here our main insight is that our proposed $L_{\alpha}$-CUSUM statistics $W_{\alpha,k,n}$ for detecting a change from $h_0(x)$ to $h_1(x)$ in (1) 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 $L_{\alpha}$-CUSUM statistics can be derived through those of the classical CUSUM statistics.

By the assumption of $\lambda(\epsilon, \alpha) > 0$ in (12) in Assumption 3.2, if we define a new function

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

(35)
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 \( k \)th local data stream is defined recursively by

\[
W'_{k,n} = \max \{0, W'_{k,n-1} + \log \frac{h_2(X_{k,n})}{h_0(X_{k,n})}\} = \max \{0, W'_{n-1} + \lambda(\epsilon, \alpha) \left[ f_1(X_{k,n}) \right]^\alpha - \left[ f_0(X_{k,n}) \right]^\alpha / \alpha \}.
\]

Compared with our proposed \( L_\alpha \)-CUSUM statistics \( W_{\alpha,k,n} \), it is clear that

\[
W'_{k,n} = \lambda(\epsilon, \alpha) W_{\alpha,k,n},
\]

and thus our proposed \( L_\alpha \)-CUSUM statistics \( W_{\alpha,k,n} \)’s are equivalent to the standard CUSUM statistics \( W'_{k,n} \) up to a positive constant \( \lambda(\epsilon, \alpha) \).

By the classical results on the CUSUM, see Appendix 2 on Page 245 of Siegmund (1985), 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

\[
G(x) = P^{(\infty)}(W^*_{\alpha,k} \geq x) = P^{(\infty)}(\limsup_{n \to \infty} W'_{k,n} \geq \lambda(\epsilon, \alpha)x) \leq e^{-\lambda(\epsilon, \alpha)x}.
\]

Now we shall use (36) to derive information bound of \( \psi_\alpha(\theta) \) in (31). In order to simplify our arguments, we abuse the notation and simply denote \( \lambda(\epsilon, \alpha) \) by \( \lambda \) in the remaining proof of the theorem. By the definition of \( \psi_\alpha(\theta) \) in (31) and the tail probability \( G(x) \) in (36), for \( \theta > 0 \),

\[
\psi_\alpha(\theta) = \log [P^{(\infty)}(W^*_{\alpha,k} \leq d) - \int_{d}^{\infty} e^{\theta(x-d)}dG(x)]
\]

\[
= \log [1 + \theta \int_{d}^{\infty} e^{\theta(x-d)}G(x)dx]
\]

\[
\leq \log [1 + \theta \int_{d}^{\infty} e^{\theta(x-d)}e^{-\lambda x}dx]
\]

\[
= \log \left( 1 + \frac{\theta}{\lambda - \theta} e^{-d\lambda} \right) = \frac{\theta}{\lambda - \theta} e^{-d\lambda},
\]

where the second equation is based on the integration by parts. Clearly, relation (37) holds for any \( 0 < \theta < \lambda = \lambda(\epsilon, \alpha) \).

By (34) and (37), we have

\[
\mathbf{E}^{\infty} N_\alpha(b, d) \geq \frac{1}{4} \exp \left( \theta b - \frac{K\theta}{\lambda - \theta} e^{-d\lambda} \right)
\]

for all \( 0 < \theta < \lambda = \lambda(\epsilon, \alpha) \). When \( \lambda b > K \exp\{-\lambda d\} \), relation (13) follows at once from (38) 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 3.1.
7.2. Proof of Theorem 3.2. To prove the detection delay bound (15) in Theorem 3.2, without loss of generality, assume the first \( m \) data streams are affected. Consider a new stopping time

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

Clearly \( N_{\alpha}(b, d) \leq T'(b, d) \), and thus

\[
D_{\epsilon}(N_{\alpha}(b, d)) \leq D_{\epsilon}(T'(b, d)).
\]

Next, by the recursive definition of \( W_{\alpha,k,n} \) in (4), using the same approach in Theorem 2 of Lorden (1971) that connects the recursive CUSUM-type scheme to the random walks, we have

\[
D_{\epsilon}(T'(b, d))) \leq E_{1}T''(b, d),
\]

where \( E_{1} \) 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 \( mI_{\theta}(\epsilon, \alpha) \) exceeds the bound \( b + md \), and is defined as

\[
T''(b, d) = \inf\{n \geq 1 : \sum_{i=1}^{n} \sum_{k=1}^{m} \left\lfloor f_{1}(X_{k,i})\right\rfloor^{\alpha} - \left\lfloor f_{0}(X_{k,i})\right\rfloor^{\alpha} \geq b + md\}.
\]

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

\[
E_{1}T''(b, d) \leq \frac{1 + o(1)}{mI_{\theta}(\epsilon, \alpha)} (b + md).
\]

Relation (15) then follows at once from the above relations, which completes the proof of Theorem 3.2.

7.3. Proof of Corollary 3.1. The choice of \( b = b_{\gamma} \) in (16) follows directly from Theorem 3.1. To prove (17), we abuse the notation and use \( \lambda \) to denote \( \lambda(\epsilon, \alpha) \) for simplification. By Theorem 3.2, the optimal \( d \) is the non-negative value that minimize the function

\[
\ell(d) := \frac{b_{\gamma}}{m} + d = \frac{1}{\lambda m} (\sqrt{\log(4\gamma)} + \sqrt{Ke^{-\lambda d}}) + 1 + \log(4\gamma).
\]

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,

\[
\ell'(d) = -\frac{1}{m} (\sqrt{Ke^{-\lambda d}} + \frac{\sqrt{\log(4\gamma)}}{2}) + 1 + \frac{\log(4\gamma)}{4m}.
\]
\[ \ell''(d) = \frac{\lambda}{m}(\sqrt{K}e^{-\lambda d} + \frac{\sqrt{\log(4\gamma)}}{2})\sqrt{K}e^{-\lambda d} > 0. \]

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

\[ d_{opt} = \frac{1}{\lambda} \log \left( \frac{K}{(\sqrt{m + \frac{1}{4}\log(4\gamma)} - \frac{1}{2}\sqrt{\log(4\gamma)})^2} \right) \]

\[ = \frac{1}{\lambda} \left\{ \log \left[ \frac{m + \frac{1}{4}\log(4\gamma) + \frac{1}{2}\sqrt{\log(4\gamma)}}{m} \right]^2 + \log \frac{K}{m} \right\}, \]

which is equivalent to those in (17) under the assumption that \( m = m(K) << \min(\log \gamma, K) \).

Plugging \( d = d_{opt} \) in (40) back to (16) yields (18), and thus the corollary is proved. \( \square \)

7.4. Proof of Theorem 4.1. Before providing the detailed proof of Theorem 4.1, let us prove the following probability result that is interesting on its own.

**Lemma 7.1.** Suppose that \( Y \) is a continuous random variable that takes both positive and negative values, and assume that its moment generating function \( \varphi(\lambda) = \mathbb{E}[e^{\lambda Y}] \) is well defined over \( -\infty < \lambda < \infty \). Then there exists a constant \( \lambda^* > 0 \) satisfying \( \mathbb{E}[e^{\lambda^* Y}] = 1 \) if and only if \( \mathbb{E}(Y) < 0 \).

**Proof:** Let us first present several facts of the moment generating function \( \varphi(\lambda) = \mathbb{E}[e^{\lambda Y}] \). First, \( \varphi(\lambda) \) is a strict convex function of \( \lambda \) since \( \varphi''(\lambda) = \mathbb{E}[Y^2 e^{\lambda Y}] > 0 \), as \( Y \) is not identical 0. Second, under our assumption, \( \varphi(\lambda) \to +\infty \) as \( \lambda \to \pm \infty \). To see this, note that there exists a constant \( y_0 > 0 \), such that \( \mathbb{P}(Y \geq y_0) > 0 \). By Chebyshev’s inequality, as \( \lambda > 0 \), \( \varphi(\lambda) = \mathbb{E}[e^{\lambda Y}] \geq e^{\lambda y_0} \mathbb{P}(Y \geq y_0) \), which goes to \( \infty \) as \( \lambda \to \infty \). Similarly, we can show that \( \lim_{\lambda \to -\infty} \varphi(\lambda) = +\infty \).

To show the “if” direction, assume \( \mathbb{E}(Y) < 0 \). Since \( \varphi(0) = 1 \) and \( \varphi'(0) = \mathbb{E}(Y) \), there must exist a positive \( \lambda_0 > 0 \) such that \( \varphi(\lambda_0) < 1 \). However, \( \varphi(\lambda) \to \infty \) as \( \lambda \to \infty \). Hence, there exists a \( \lambda^* \in (\lambda_0, \infty) \) such that \( \varphi(\lambda^*) = 1 \).

For the “only if” direction, since \( \varphi(0) = \varphi(\lambda^*) = 1 \), there exists a positive value \( \lambda_1 \in (0, \lambda^*) \) such that \( \varphi'(\lambda_1) = 0 \). Since \( \varphi(\lambda) \) is convex, \( \varphi'(\lambda) \) must be decreasing. Thus \( \mathbb{E}(Y) = \varphi'(0) < \varphi'(\lambda_1) = 0 \). This completes the proof of the lemma. \( \square \)
Now we are ready to prove Theorem 4.1. Let us begin with a high-level sketch of the proof. To find the breakdown point of our proposed scheme $T_\alpha(b, d)$, we need to investigate the asymptotic properties of $E_h^{(\infty)}[N_\alpha(b, d)]$ for any $h = (1 - \epsilon)f_0 + \epsilon g$ as $b \to \infty$, where $E_h^{(\infty)}$ denotes the expectation of run length when there is no change and all data come from the density function $h$ here and the remaining of the proof. Since we assume $f_0(x) - f_1(x)$ take both positive and negative values, $Y = \frac{|f_1(X)|^\alpha - |f_0(X)|^\alpha}{\alpha}$ is a continuous random variable that takes both positive and negative values.

By Lemma 7.1, it turns out the asymptotic properties depend on whether the following expectation is positive or negative:

$$
\mu_{\epsilon, h} = E_h \left[ \frac{f_1(X)^\alpha - f_0(X)^\alpha}{\alpha} \right]
= (1 - \epsilon)E_{f_0} \left[ \frac{f_1(X)^\alpha - f_0(X)^\alpha}{\alpha} \right] + \epsilon E_g \left[ \frac{f_1(X)^\alpha - f_0(X)^\alpha}{\alpha} \right]
$$

As we will show below, $\log E_h^{(\infty)}[N_\alpha(b, d)]$ is of order $b$ if $\mu_{\epsilon, h} < 0$ but becomes of order $\log(b)$ if $\mu_{\epsilon, h} > 0$. Next, in order for $T_\alpha(b, d)$ to satisfy the false alarm constraint $\gamma$ under the idealized model with $\epsilon = 0$, we must have $b \sim \log \gamma$ as it can be shown that $\mu_{\epsilon=0, h} < 0$ for any $\alpha \geq 0$ when $f_0$ and $f_1$ are from the same location family. Hence, the false alarm breakdown point can be found by finding the smallest $\epsilon$ value such that $\mu_{\epsilon, h} > 0$.

Next, let us show that $\mu_{\epsilon, h} < 0$ is a sufficient condition that $\log E_h^{(\infty)}[N_\alpha(b, d)]$ is of order $b$. By Lemma 7.1, if $\mu_{\epsilon, h} < 0$, then there exists a positive real value $\lambda > 0$ such that

$$
E_h \exp \left\{ \lambda \left( \frac{f_1(X)^\alpha - f_0(X)^\alpha}{\alpha} \right) \right\} = 1.
$$

This is exactly Assumption 3.2 with $h_0 = h$, and thus the conclusions of Theorem 3.1 holds when $h_0$ is replaced by $h$. In particular, for fixed $d$ and $K$, as $b$ goes to $\infty$, we have

$$
\log E_h^{\infty} N_\alpha(b, d) \geq (1 + o(1))\lambda b.
$$

Meanwhile, if $\mu_{\epsilon, h} > 0$, we will show that $\log E_h^{(\infty)}[N_\alpha(b, d)]$ is of order $\log(b)$. To see this, $E_h^{(\infty)} N_\alpha(b, d)$ is the expected sample size of $N_\alpha(b, d)$ when the data are i.i.d. from $h$, which can also be regarded as the detection delay with the post-change distribution $h_1 = h$ when the change occurs at time $\nu = 1$. Indeed, $\mu_{\epsilon, h} > 0$ is actually Assumption 3.1 with $h_1 = h$, and thus the arguments on the detection delay analysis in Theorem 3.2 applies. Hence,

$$
\log E_h^{(\infty)} N_\alpha(b, d) \leq (1 + o(1))\log b.
$$
Therefore, combining the above results with the definition of breakdown point in Definition 4.1, the breakdown point of our proposed scheme $N_\alpha(b,d)$ is

\begin{equation}
\epsilon^*(N_\alpha) = \inf\{\epsilon \geq 0 : \sup_{h \in h_{0,\epsilon}} \mu_{\epsilon,h} > 0\},
\end{equation}

where $\mu_{\epsilon,h}$ is defined in (41).

The remaining proof is based on a careful analysis of $\mu_{\epsilon,h}$ in (41) for any arbitrary outlier density function $g$. For any $h(x) = (1 - \epsilon)f_0(x) + \epsilon g(x) \in h_{0,\epsilon}$, by (41), we have

\begin{equation}
\mu_{\epsilon,h} = -\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,
\end{equation}

where $d_\alpha(f_0, f_1)$ is defined in (24) and is the density power divergence between $f_0$ and $f_1$ proposed by Basu et al. (1998). 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 (23), it is clear from (45) that

\begin{equation}
\sup_{h \in h_{0,\epsilon}} \mu_{\epsilon,h} = -\frac{1 - \epsilon}{1 + \alpha}d_\alpha(f_0, f_1) + \epsilon M(\alpha).
\end{equation}

Therefore, by (44), if both $d_\alpha(f_0, f_1)$ and $M(\alpha)$ are finite, the false alarm breakdown point of $N_\alpha$ should be

\begin{equation}
\epsilon^*(N_\alpha) = \frac{d_\alpha(f_0, f_1)}{d_\alpha(f_0, f_1) + (1 + \alpha)M(\alpha)}.
\end{equation}

If $d_\alpha(f_0, f_1)$ is finite but $M(\alpha) = +\infty$, by (44) and (46), $\epsilon^*(N_\alpha) = 0$. 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 (44) and (46), we have $\epsilon^*(N_\alpha) = \frac{\rho}{\rho + (1 + \alpha)}$ no matter $\rho$ is finite or not. Therefore, for all cases, the false alarm breakdown point of $N_\alpha$ have the same expression in (47), which completes the proof of Theorem 4.1.

\begin{flushright}
\square
\end{flushright}

References.


