Statistica Sinica Preprint No: SS-2015-0316R2	
Title	Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams
Manuscript ID	SS-2015-0316.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0316
Complete List of Authors	Yajun Mei Kun Liu and Ruizhi Zhang
Corresponding Author	Yajun Mei
E-mail	ymei@isye.gatech.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams

Kun Liu, Ruizhi Zhang & Yajun Mei*
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30329-0205, USA

January 25, 2017

Abstract

In this article, we investigate the problem of monitoring independent large-scale data streams where an undesired event may occur at some unknown time and affect only a few unknown data streams. Motivated by parallel and distributed computing, we propose to develop scalable global monitoring schemes by parallel running local detection procedures and by using the sum of the shrinkage transformation of local detection statistics as a global statistic to make a decision. Our approach is illustrated in two concrete examples: one is censoring sensor networks when local sensors can be nonhomogeneous with completely specified local post-change distributions, and the other is the problem of monitoring normally distributed data streams when the local post-change mean shifts are unknown and can be positive or negative. Numerical simulation studies demonstrate the usefulness of the proposed SUM-Shrinkage schemes.

Keywords: change-point, CUSUM, parallel computing, quickest detection, sensor networks.

1 Introduction

In the modern information age, one often faces the need to online monitor large-scale data streams with the aim of offering the potential for early detection of a "trigger" event. Ideally, one would like to develop a global monitoring scheme that can detect the occurring event as quickly as possible while controlling the system-wise global false alarm rate. From the statistical point of view, this is a sequential change-point

*emails: {kliu80, zrz123, ymei3}@gatech.edu. This research was partially supported by the NSF grants DMS-0954704 and CMMI-1362876.

detection or quickest change detection problem, which has a variety of applications such as industrial quality control, signal detection and biosurveillance. The classical version of this problem, where one monitors independent and identically distributed (iid) *univariate* or *low-dimensional multivariate* observations from a single data stream, is a well-developed area, and many classical procedures have been developed such as the Shewhart's chart (Shewhart [32]), moving average control charts, Page's CUSUM procedure (Page [26]), Shiryaev-Roberts procedure (Shiryaev [33], Roberts [31]), window-limited procedures (Lai [14]) and scan statistics (Glaz, Naus and Wallenstein [10]). All these classical procedures not only hold attractive theoretical properties, but also are computationally simple. See, for example, Lorden [18], Pollak [27, 28], Moustakides [24], Lai [14, 15], Kulldorff [13]. For a review, see the books such as Basseville and Nikiforov [3], Poor and Hadjiliadis [29], Tartakovsky, Nikiforov, and Basseville [35].

However, research is limited in the context of monitoring large-scale data streams, especially when the occurring event might affect some, but not all, local data streams. Existing methods include MAX-scheme (which uses the maximum of local CUSUM statistics as the global statistic, see Tartakovsky et al. [36]), SUM-scheme (which uses the sum of local CUSUM statistics as the global statistic, see Mei [21]), the mixture-schemes proposed in Xie and Siegmund [40], and the simultaneous-estimation-based schemes in Wang and Mei [42]. Unfortunately, while these first two schemes are computationally efficient, they are generally statistically inefficient unless in extreme cases where the number of affected data streams is either very small or very large. Meanwhile, the last two families of schemes enjoy nice statistical properties under general settings, but they are computationally infeasible for online monitoring large-scale data streams over long time period due to the lack of recursive forms and the requirements of large local memory and computation power to store and process past information. This research intends to fill the gap to balance the tradeoff between statistical efficiency and computational efficiency when monitoring large-scale data streams and the number of affected data streams is moderate. While many classical likelihood-ratio-based quickest change detection methods can be extended from one or low dimension to high-dimension or large-scale data streams from the theoretical viewpoint, they are generally computationally infeasible in the context of large-scale data streams. As mentioned in Breiman [4], in order for the profession of statistics to remain healthy, more algorithm-based methods should be developed. This is exactly what needs to be done in the subfield of quickest change detection or sequential change-point detection. We feel that the current main bottleneck is on the algorithm or methodology aspect, and in particular, new ideas and new approaches are needed to develop efficient *scalable* global schemes in the sense of being able to be implemented for monitoring large-scale data streams over a long period of time.

In this article, we present a general and flexible approach that can provide efficient scalable global schemes when monitoring large-scale data streams. Our research is motivated by censoring sensor networks in engineering, which was introduced by Rago, Willett, and Bar-Shalom [30] and later by Appadwedula,

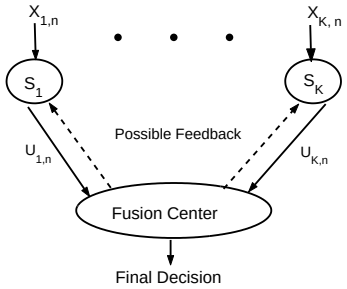

Figure 1: General setting of a widely used configuration of censoring sensor networks.

Veeravalli, and Jones [1] and Tay, Tsitsiklis, and Win [38]. Figure 1 illustrates the general setting of a widely used configuration of censoring sensor networks, in which the data streams $X_{k,n}$'s are observed at the remote, distributed sensors, but the final decision is made at a central location, called the fusion center. The key feature of such a network is that while sensing (i.e., taking observations at the local sensors) are generally cheap and affordable, communication between remote sensors and fusion center is expensive in terms of both energy and limited bandwidth. The question then becomes how the fusion center can still monitor the system effectively under the networks resource constraints. A more concrete example is the National Syndromic Surveillance Program BioSense Platform at the Centers for Disease Control and Prevention (CDC), where the computing power and memory of any centralized server would have become limited as compared to *daily* summary data from all state and local health departments as well as many hospitals, and thus the CDC's BioSense Platform is designed to be a distributed computing system that can detect a global level disease outbreak.

We propose to develop scalable schemes for monitoring large-scale data streams by taking advantage of parallel and distributed computing and the fact that many efficient and computationally simple local procedures are available to detect changes in local data streams. To be more specific, suppose we are monitoring a large number K of data streams, and for each local data stream, an efficient local detection procedure is available based upon some local detection statistics that can be computed recursively over time n, e.g., involving $O(1)$ computations and $O(1)$ memory requirements at each time. Then our proposed methodology is to run these K local detection procedures in parallel before combining them into a global monitoring scheme. Thus the computation and memory requirements of our proposed scheme do not increase over time n, and are fixed as a function of K at each time step n when new observations are taken, thereby yielding a scalable global monitoring scheme. While the parallel local monitoring approach sounds interesting, one allegation often made is that we will lose much information at the global level if we combine local detection procedures, not raw observation themselves, to make a global decision. Indeed, as mentioned earlier, there are two existing methods that combine local detection procedures together: the MAX and SUM schemes

that use the maximum or sum of local CUSUM to raise a global alarm, but both methods are known to be inefficient when the number of affected data streams is moderate, see Mei [21] and Xie and Siegmund [40].

In this article, we demonstrate that the problem is not on the parallel local monitoring approach itself, but on how to combine the local detection procedures suitably in the scenario when the number of affected data streams is moderate. Our key idea is to generalize the SUM scheme in Mei [21] by introducing the shrinkage function to local detection statistics in the hope of filtering out those unchanging local data streams. We would like to acknowledge that there might be inherent loss of statistical efficiency in the parallel local monitoring approach as compared to the (non-recursive) global monitoring approach that uses all raw observations, e.g., see Section 5 for the comparison of our proposed schemes with those in Xie and Siegmund [40]. However, the parallel local monitoring approach allows us to develop scalable schemes, and the loss of statistical efficiency is the price we pay for the computational efficiency. It is also worth pointing out that a well-known view in the standard off-line statistical inference literature is the necessity of shrinkage for high-dimensional data in order to improve power or efficiency. Thus, from the methodology point of view, our proposed methodologies are analogous to those off-line statistical methods such as (adaptive) truncation, and soft- and hard- thresholding, see Neyman [25], Donoho and Johnstone [6], Fan and Lin [8]. Also see Candès [5] and the references there. However, our motivation here is different and our application to distributed quickest change detection is new.

The remainder of this article is organized as follows. In Section 2, we present some preliminaries and background information of quickest change detection or sequential change-point detection, and also discuss two existing methodologies for parallel local monitoring. In Section 3, we propose our "SUM-shrinkage" methodology under a general setting of monitoring large-scale independent data streams, and provide general theoretical results. We exemplify our methodology in two concrete examples: Section 4 considers the censoring sensor networks when the local data streams may be nonhomogeneous but the pre-change and post-change distributions of local data streams are given, and Section 5 investigates the scenario of normally distributed data when the post-change means of local data streams are unknown. In both Sections 4 and 5, numerical Monte Carlo simulation studies are conducted to illustrate the performance of our proposed methods.

2 Preliminaries and Background

Let us present our problem under a general setting, and two specific examples will be given in later sections. Assume there are K independent data streams in a system.

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

$$\cdots \qquad \cdots$$
$$\text{Data Stream } K: \qquad X_{K,1}, X_{K,2}, \cdots .$$

Initially, the system is "in control", but at some *unknown* time ν, an undesired event may occur and affect a few unknown local data streams in the sense of changing the local distributions of the $X_{k,n}$'s.

Here we assume that the online monitoring is conducted under the *unstructured* environment in the sense that we do not make any assumptions to relate the occurring event to the local data streams, see Tartakovsky et al. [36], Mei [21], and Xie and Siegmund [40]. Also see Lévy-Leduc and Roueff [16] for an application of the unstructured problem to anomaly detection in computer networks. In particular, we focus on the scenario when the occurring event changes the local distributions of affected local data streams, and we do not aim to detect changes on the correlation between different data streams. Hence, the data $X_{k,n}$'s will be assumed to be independent across different data streams, but can be flexible otherwise. For instance, the $X_{k,n}$'s may or may not be identically distributed across different local data streams, can be dependent over time within each local data stream, and can be *univariate* or *low-dimensional multivariate*. In addition, in many practical applications, the assumption of the independence across different data streams is not as restrictive as one thought, see Xie, Huang and Willett [39], and Liu, Mei and Shi [17], to monitor the independent residuals from some spatio-temporal models, instead of dependent raw data, in two real-world applications in solar flare and hot-forming process.

For the purpose of generalization, we do not specify which kind of local changes these K data streams may have. Instead we assume that there is a local detection statistic $W_{k,n}$ (in the log-likelihood scale) for the k-th local data stream at each time step n that summarizes the evidence regarding a possible local change based on the first n local observations $(X_{k,1}, \ldots, X_{k,n})$ for each $k = 1, \ldots, K$. For instance, $W_{k,n}$ can be the well-known CUSUM or Shiryeav-Robert statistics (in the log-likelihood scale) when the local data are independent over time, or can be the recursive quasi-generalized-likelihood-ratio test in Fuh and Mei [9] when the local data are dependent from hidden Markov models. It is important to point out that the $W_{k,n}$'s not only should be able to detect local changes quickly, but also can be computed efficiently in order for our proposed scheme to be scalable. However, we should emphasize that it can be highly non-trivial to construct such $W_{k,n}$'s in practice, see a concrete example in Section 5.

Next, let us review the definition of a global monitoring scheme and the criteria to evaluate it under the minimax setting. A global monitoring scheme can be defined as a stopping time T with respect to the K-dimensional vector data $\{(X_{1,n}, \cdots, X_{K,n})\}_{n\geq 1}$. In particular, when $T = t$, one raises an alarm at time t to indicate that a change has occurred somewhere in the first t time steps. When monitoring K independent data streams in (1), it is well-known in statistics that even if each local false alarm rate is well controlled, the global false alarm rate can be significant when the number K of data streams is large. In the literature of sequential change-point detection, for a global monitoring scheme that raise an alarm at time T, its global

false alarm rate is often evaluated by $1/\mathbf{E}^{(\infty)}(T)$, where $\mathbf{E}^{(\infty)}(T)$ is the expectation of T when the system is "in control," and is often called the Average Run Length (ARL) to false alarm. Meanwhile, the definition of detection delay of the global monitoring scheme is a little more complicated. Assume that the event occurs at the unknown time ν, and the global monitoring scheme raises an alarm at time $T \geq \nu$. Then the detection delay is $T-\nu+1$, but we must take into account of the randomness of T and the uncertainty of ν. A widely used rigourous definition of the detection delay of T is the following "worst case" detection delay defined in Lorden [18],

$$\overline{\mathbf{E}}(T) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbf{E}^{(\nu)}\Big((T-\nu+1)^+ \Big| \mathcal{F}_{\nu-1}\Big). \tag{2}$$

Here "ess sup" is over all possible scenarios of global pre-change information $\mathcal{F}_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]})$, $X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1})$ is local pre-change information for the k-th data stream at time ν, and $\mathbf{P}^{(\nu)}$ and $\mathbf{E}^{(\nu)}$ denote the probability measure and expectation when the event occurs at time ν.

The standard minimax formulation is to find a global monitoring scheme with a stopping time T that minimizes the detection delay $\overline{\mathbf{E}}(T)$ in (2) subject to the global false alarm constraint

$$\mathbf{E}^{(\infty)}(T) \geq \gamma, \tag{3}$$

where $\gamma > 0$ is a pre-specified constant.

In this paper, we are interested in finding a global monitoring scheme that is scalable for large number K of local data steams when we do not know which subset of local data streams might be affected. Our approaches are based on parallel local monitoring to construct a global stopping time T from the local detection statistics $W_{k,n}$'s that can be computed efficiently. There are two existing methods for parallel local monitoring. The first one is to raise an alarm at the global level whenever any local detection procedures raises a local alarm. If we normalize the local detection statistics $W_{k,n}$'s, this can be rewritten as raising an alarm at the global level at time

$$T_{\max}(a) = \inf\{n \geq 1 : \max_{1 \leq k \leq K} W_{k,n} \geq a\}, \tag{4}$$

($= \infty$ if such n does not exist) where $a > 0$ is a pre-specified constant. Below we will call the scheme in (4) the "MAX" scheme. The second method is the "SUM" scheme developed in Mei [21] that is defined by the stopping time

$$T_{\text{sum}}(a) = \inf\{n \geq 1 : \sum_{k=1}^{K} W_{k,n} \geq a\}, \tag{5}$$

($= \infty$ if such n does not exist). As mentioned in Mei [21], the "MAX" scheme $T_{\max}(a)$ in (4) works well when one or very few data streams are affected, whereas the "SUM" scheme $T_{\text{sum}}(a)$ in (5) works well only when many data streams are affected. Here and below the threshold a of a scheme $T(a)$ is a pre-specified constant so that the scheme $T(a)$ satisfies the false alarm constraint γ in (3).

3 Our Proposed SUM-Shrinkage Methodology

Now we are ready to present our proposed methodology under a general setting. Assume, for a moment, the local detection statistics $W_{k,n}$'s (in the log-likelihood scale) have been constructed. We suggest to define the global monitoring statistic of the general "SUM-shrinkage" form

$$G_n = \sum_{k=1}^{K} h_k(W_{k,n}), \tag{6}$$

where $h_k(\cdot) \geq 0$ are some suitable shrinkage transformation functions. Then our proposed SUM-shrinkage scheme raises a global alarm at the time

$$N_G(a) = \inf\{n \geq 1 : G_n \geq a\}. \tag{7}$$

Note that our proposed SUM-shrinkage scheme $N_G(a)$ in (7) has two key components in its global monitoring statistic G_n in (6): one is the local detection statistic $W_{k,n}$'s, and the other is the shrinkage transformations $h_k(\cdot)$'s. In the next two sections, we will demonstrate how to specifically choose these two components in different contexts. Intuitively, the local detection statistics $W_{k,n}$'s should be easily computed and able to detect local changes quickly. The shrinkage functions h_k's in (6) play the role of dimension reduction by automatically filtering out those non-changing local data streams and by focusing only on those local data streams that appear to be affected by the occurring event.

Besides the iid examples in the next two sections, we should point out that our proposed "SUM-shrinkage" methodology in (6)-(7) has a broad range of other applications. For instance, depending on which kind of local models or local changes we are interested in, the local detection statistics $W_{k,n}$'s can be defined for dependent observations such as those from the recursive schemes in Fuh and Mei [9] for hidden Markov models, or those from the non-parametric rank-based detection schemes in Gordon and Pollak [11]. In addition, as an extension, the proposed SUM-shrinkage scheme $N_G(a)$ enjoys the nice properties of the SUM scheme $T_{\text{sum}}(a)$ in (5): it does not assume that all local data streams are affected by the occurring event simultaneously, and thus is applicable when different local data streams are affected by the occurring event at different time steps, see Mei [21] and Xie and Siegmund [40]. Moreover, little information seems to be lost if we do not observe those local data streams with small values of $W_{k,n}$'s since they make limited contributions in our proposed global monitoring statistic G_n in (6). This motivated Liu, Mei and Shi [17] to develop an efficient adaptive sensor relocation policy when one only has ability to observe r out of K data streams at each time step. This may occur in manufacturing process control when there are K possible stages in the process but there are only r expensive sensors available to monitor the process. In such a problem, the order-thresholding transformation in (10) below can be combined with missing data techniques to be used not only in the global monitoring statistic G_n in (6) for quickest detection, but also in a greedy manner to adaptively observe those r data streams with the largest $W_{k,n}$'s values at each time step. Banerjee and Veeravalli [2]

essentially tackle the similar problem of missing data, but using the hard-thresholding transformation in (8) below. We feel the spirit of SUM-Shrinkage can have many other applications, and hopefully our research opens new research opportunities and directions, especially on monitoring large-scale data streams.

To better understand the general properties of our proposed SUM-shrinkage scheme $N_G(a)$ in (7), the remaining of this section is divided into three subsections. Subsection 3.1 contains three explicit choices of shrinkage functions h_k's in (6), and Subsection 3.2 includes some general properties of $N_G(a)$ that is related to the global false alarm constraint in (3). Subsection 3.3 takes a further step to discusses how to choose the tuning parameters in the shrinkage functions h_k's in (6) when the local data streams are homogeneous.

3.1 Shrinkage Transformation

Evidently a suitable choice of the h_k's in the SUM-shrinkage monitoring statistic G_n in (6) will depend on the assumptions and contexts of applications. As an explicit illustration, below we will demonstrate the following three shrinkage transformations of the form

- Hard-thresholding: $h(x) = x\mathbf{1}\{x \geq b\}$ for some constant b, (8)
- Soft-thresholding: $h(x) = \max\{x - b, 0\}$ for some constant b, (9)
- Order-thresholding: $h(x) = x\mathbf{1}\{x \geq w_{(r)}\}$, where $w_{(r)}$ is the r-th largest statistic of $w_1, \cdots, w_K$. (10)

Of course, besides those in (8)-(10), there are many other kinds of the shrinkage functions such as $h(x) = \exp(bx)$. Also by semi-Bayesian arguments, the transformation $h(x) = \log[1 - p_0 + p_0 \exp(\max(0, x))]$ is proposed and used in the schemes of Xie and Siegmund [40] in a completely different manner.

Based on our experiences, the soft-thresholding transformation, as a continuous function, often yields smaller detection delays than the hard-thresholding transformation, which is a discontinuous function, in the finite-sample Monte Carlo simulations. Moreover, the soft- and order- thresholding transformations have comparable finite-sample performances, but the soft-thresholding transformation is computationally and theoretically simpler. Thus our analysis below will use the soft-thresholding transformation in (9) as a concrete demonstration when needed.

To better understand the shrinkage transformations in (8)-(10), below let us motivate them from the communication efficiency viewpoint, which were first presented in Mei [22] in the context of censoring sensor networks in Figure 1. To prolong the reliability and lifetime of the network system, it is natural for the local sensors to transmit only those local detection statistics $W_{k,n}$'s that are large. Specifically, at time n, the sensor message from the sensor to the fusion center is given by

$$U_{k,n} = \begin{cases} W_{k,n}, & \text{if } W_{k,n} \geq b_k \\ \text{NULL}, & \text{if } W_{k,n} < b_k \end{cases}, \tag{11}$$

where $b_k \geq 0$ is the local censoring parameter at the k-th sensor (or data stream). In practice, the message "NULL" could be represented by the situation when the sensor does not send any message to the fusion center, e.g., the sensor is silent.

After receiving the local sensor messages $U_{k,n}$'s in (11), the fusion center then combines them together suitably to make a global decision. There are many approaches to do so, and the first two schemes are based on the summation of all sensor messages $U_{k,n}$'s, depending on how to interpret the "NULL" values. If we treat the "NULL" values as lower limit 0, then the fusion center raises a global alarm at time

$$\begin{aligned} N_{hard}(a) &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} U_{k,n} \geq a\Big\} \\ &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} W_{k,n}\mathbf{1}\{W_{k,n} \geq b_k\} \geq a\Big\}. \end{aligned} \tag{12}$$

Below this scheme will be referred as the hard-thresholding scheme, since it is a special case of the global statistic in (6) when the shrinkage functions h_k's are the hard-thresholding transformation in (8).

Meanwhile, if we treat the "NULL" values as the upper limit b_k's, then the fusion center will compute the global monitoring statistic

$$G_n = \sum_{k=1}^{K} U_{k,n} = \sum_{k=1}^{K} \max\{W_{k,n}, b_k\} = \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} + \sum_{k=1}^{K} b_k,$$

which is closely related to the soft-thresholding transformation in (9). Hence, we can define the soft-thresholding scheme that raises an alarm at time

$$N_{soft}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} \geq a\Big\}. \tag{13}$$

Here we keep the threshold of $N_{soft}(a)$ as a instead of $a - \sum_{k=1}^{K} b_k$, so that $N_{soft}(a)$ is the special case of our proposed SUM-shrinkage scheme $N_G(a)$ in (7) with the soft-thresholding transformation in (9).

The third approach occurs when the fusion center has a prior knowledge that (at most) r out of K data streams will be affected by the occurring event. Such a prior knowledge may be defined by the network fault-tolerant design to avoid risking failure. In this case, it is reasonable for the fusion center to order all sensor messages $U_{k,n}$'s as $U_{(1),n} \geq \ldots \geq U_{(K),n}$, and raise an alarm if the sum of the r largest $U_{k,n}$'s is too large. This is a combination of the hard-thresholding transformation in (8) and the order-thresholding transformation in (10), and it yields a global scheme that is defined by the stopping time

$$N_{comb,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} U_{(k),n} \geq a\Big\}. \tag{14}$$

For simplicity, the "NULL" values of $U_{k,n}$'s in the scheme $N_{comb,r}(a)$ in (14) will be treated as the lower limit 0 in our simulation below. A special case of $N_{comb,r}(a)$ in (14) is when the order-thresholding transformation

in (10) is applied directly to the local detection statistics $W_{k,n}$'s in (33) themselves. Specifically, we order the K local CUSUM statistics $W_{1,n}, \ldots, W_{K,n}$ from largest to smallest: $W_{(1),n} \geq W_{(2),n} \geq \ldots \geq W_{(K),n}$. Then the order-thresholding scheme can be defined by the stopping time

$$N_{order,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} W_{(k),n} \geq a\Big\}. \tag{15}$$

Of course, $N_{order,r}(a)$ is a special case of $N_{comb,r}(a)$ if the local censoring parameter $b_k \equiv 0$, under the reasonable assumption that the local detection statistics $W_{k,n}$'s are non-negative.

It is useful to highlight that the hard-thresholding scheme, $N_{hard}(a)$ in (12), becomes the SUM scheme $T_{\rm sum}(a)$ in (5) when $b_k = 0$ for all k, and becomes the MAX scheme $T_{\rm max}(a)$ in (4) when $b_k = a$ for all k. Likewise, for the order-thresholding scheme, $N_{order,r}(a)$, it becomes the SUM scheme when $r = K$, but becomes the MAX scheme when $r = 1$. Hence, we should expect that the censoring parameter $b_k \in (0, a)$ or $r \in [1, K]$ would adapt our proposed schemes to different sparsity post-change level depending on how many local data streams are affected.

As for the soft-thresholding scheme, $N_{soft}(a)$ in (13), it becomes the SUM scheme $T_{\rm sum}(a)$ in (5) when $b_k = 0$ for all k, and becomes the MAX scheme $T_{\rm max}(b)$ in (4) when $a = 0$ and $b_k \equiv b$ for all k. The statistical intuition of the soft-thresholding transformation is a little more complicated, and below we provide a semi-Bayesian interpretation why the soft-thresholding scheme in (13) works. At a given time n, let Z_k be the indicator whether the distribution of the k-th local data stream changes for $k = 1, \ldots, K$. Assume that each local data stream has a prior probability π_k getting affected by the event, and assume that $Z_1, \ldots, Z_K$ are iid with probability mass function $\mathbf{P}(Z_k = 1) = \pi_k = 1 - \mathbf{P}(Z_k = 0)$. Treat Z_k's as the hidden states, and recall that $W_{k,n}$ represents the evidence of possible change (in logarithm scale) and is applicable only when $Z_k = 1$ (since $Z_k = 0$ implies that there is no change at the k-th data stream). Then when testing $H_0 : Z_1 = \ldots = Z_K = 0$ (no change), the Log-Likelihood Ratio (LLR) statistic of the hidden state Z_k's and the observed data $X_{k,n}$'s is

$$\begin{aligned} LLR(n) &= \sum_{k=1}^{K} \{Z_k(\log \pi_k + W_{k,n}) + (1 - Z_k)\log(1 - \pi_k)\} - \sum_{k=1}^{K} \log(1 - \pi_k) \\ &= \sum_{k=1}^{K} Z_k\{W_{k,n} - \log((1 - \pi_k)/\pi_k)\} \end{aligned}$$

Since the Z_k's are unobservable, it is natural to maximize $LLR(n)$ over $Z_1, \ldots, Z_K \in \{0, 1\}$. Hence, the maximum likelihood estimator of the Z_k's is that

$$\hat{Z}_k = \begin{cases} 1, & \text{if } W_{k,n} \geq \log((1 - \pi_k)/\pi_k) \\ 0, & \text{otherwise} \end{cases}, \quad \text{for } k = 1, \ldots, K,$$

and the generalized log-likelihood ratio becomes

$$\max_{Z_k's} LLR(n) = \sum_{k=1}^{K} \max\{W_{k,n} - \log((1-\pi_k)/\pi_k), 0\},$$

which is exactly the form of the soft-thresholding scheme $N_{soft}(a)$ in (13) with $b_k = \log((1-\pi_k)/\pi_k)$.

3.2 The Choice of Threshold a to Satisfy the False Alarm Constraint

Given the choices of the local detection statistics $W_{k,n}$'s and the shrinkage transformation $h_k(\cdot)$'s, an important remaining question is how to determine the global threshold a in (7) so that the proposed SUM-shrinkage scheme $N_G(a)$ in (7) satisfies the global false alarm constraint γ in (3). This is nontrivial, as it requires one to accurately characterize the relationship between the threshold a and the ARL to false alarm $\mathbf{E}^{(\infty)}(N_G(a))$.

Intuitively, the global monitoring statistic G_n in (6) is the sum of K (independent) random variables, one would expect that the Central Limited Theorem (CLT) will be useful when the shrinkage transformation keeps most non-zero values, e.g., the hard-thresholding or soft-thresholding transformations in (8) or (9) when the censoring parameters b's are not large, whereas the compound Poisson process will be needed when the shrinkage transformation only keeps very few non-zero values, e.g., the order-thresholding transformation in (10) with not so large r value. The rigorous theoretical proofs are beyond the scope of this article and will be investigated elsewhere. Below we will use Chebyshev's inequalities to provide a general non-asymptotic, conservative bound on the global threshold a of our proposed SUM-shrinkage scheme $N_G(a)$ in (7) in terms of the false alarm constraint γ in (3).

To do so, let us assume that under the pre-change hypothesis $\mathbf{P}^{(\infty)}$, for each k, the shrinkage transformation of local detection statistic, $h_k(W_{k,n})$, converges to their limit H_k^* as $n \to \infty$. We further assume that for each $k = 1, \cdots, K$, the limit H_k^* is stochastically larger than any finite-time version $h_k(W_{k,n})$'s, and has a well-defined log-moment generating function

$$\psi_k(\theta) = \log \mathbf{E}^{(\infty)} \exp(\theta H_k^*) \tag{16}$$

for some $\theta \geq 0$. With this definition, the following theorem provides a non-asymptotic lower bound on the ARL to false alarm of the proposed SUM-shrinkage scheme $N_G(a)$ in (7) and a non-asymptotic, conservative choice threshold a of $N_G(a)$ to satisfy the false alarm constraint γ in (3).

Theorem 3.1. *Assume $\psi_k(\theta)$ are well-defined for all $\theta \in \Theta$, a sub-interval of $[0, \infty)$, for all $k = 1, \cdots, K$. Then,*

$$\mathbf{E}^{(\infty)}(N_G(a)) \geq \frac{1}{4} \exp\left(\theta a - \sum_{k=1}^{K} \psi_k(\theta)\right) \tag{17}$$

for all $\theta \in \Theta$, *and thus a choice of threshold*

$$a = \inf_{\theta \in \Theta} \left(\frac{1}{\theta} \big(\log(4\gamma) + \sum_{k=1}^{K} \psi_k(\theta) \big) \right) \tag{18}$$

will guarantee that $N_G(a)$ *in (7) satisfies the global false alarm constraint* γ *in (3).*

Proof: Relation (18) follows directly from (17), and it suffices to show that (17) holds for any $\theta \in \Theta$. To see this, by the definition of $N_G(a)$ in (7) and by applying Chebyshev's inequality twice: once to $N_G(a) \geq 0$ and the second to $\sum_{k=1}^{K} H_k^*$, for any $x > 0$

$$\begin{aligned}
\mathbf{E}^{(\infty)}(N_G(a)) &\geq x\mathbf{P}^{(\infty)}\Big(N_G(a) \geq x\Big) \\
&= x\Big[1 - \mathbf{P}^{(\infty)}\Big(N_G(a) < x\Big)\Big] \\
&= x\Big[1 - \mathbf{P}^{(\infty)}\Big(\sum_{k=1}^{K} h_k(W_{k,n}) \geq a \text{ for some } 1 \leq n \leq x\Big)\Big] \\
&\geq x\Big[1 - x\mathbf{P}^{(\infty)}\Big(\sum_{k=1}^{K} H_k^* \geq a\Big)\Big] \\
&\geq x\Big[1 - xe^{-\theta a}\mathbf{E}^{(\infty)} \exp\Big(\theta \sum_{k=1}^{K} H_k^*\Big)\Big] \\
&= x\Big[1 - xe^{-\theta a} \exp\Big(\sum_{k=1}^{K} \psi_k(\theta)\Big)\Big].
\end{aligned}$$

Here the second inequality follows from the assumption that H_k^* is stochastically larger than $h_k(W_{k,n})$'s, and the last equation uses the assumption that these K data streams are independent across different data streams. Note that for any $u > 0$, the function $x(1 - xu)$ is maximized at $x = 1/(2u)$ with the maximum value $1/(4u)$. This completes the proof of (17), and thus the theorem holds. □

It is important to emphasize that the results in Theorem 3.1 are non-asymptotic, and hold for any K and γ. To demonstrate their usefulness, let us consider a concrete homogeneous case when the $W_{k,n}$'s are identically distributed over k under the pre-change hypothesis and all local data streams use the same soft-thresholding transformation (9). In this case, let us suppress the script k and first derive the log-moment generating function $\psi(\theta)$ in (16) for the soft-thresholding transformation $h(W_n) = \max(W_n - b, 0)$ for large b. For that purpose, we further assume that as $n \to \infty$, the local detection statistic W_n's converges to an asymptotically exponentially distributed variable W^* under the pre-change hypothesis. That is, we assume that as $x \to \infty$,

$$\mathbf{P}^{(\infty)}(W^* > x) \approx \lambda e^{-x}, \tag{19}$$

for some constant $\lambda > 0$. In fact, the following non-asymptotic result is often true for many local detection

statistic W_n's such as CUSUM: for *any* $x > 0$,

$$\mathbf{P}^{(\infty)}(W^* > x) \leq e^{-x}, \tag{20}$$

see Appendix 2 on Page 245 of Siegmund [34]. Under the assumption (19), we have $\mathbf{P}^{(\infty)}(W^* \leq b) = 1-\lambda e^{-b}$ for large b. Combining the definition of $\psi(\theta)$ in (16) with the fact that $H^* = 0$ whenever $W^* \leq b$ yields that

$$\psi(\theta) = \log \mathbf{E}^{(\infty)} \exp(\theta H^*) = \log[\mathbf{P}^{(\infty)}(W^* \leq b) + \int_b^\infty e^{\theta(x-b)} \lambda e^{-x} dx] = \log\left(1 + \frac{\theta\lambda e^{-b}}{1-\theta}\right). \tag{21}$$

Clearly, $\psi(\theta)$ is well-defined over $\theta \in \Theta = [0, 1)$. If we further assume that b is large, or equivalently, λe^{-b} is small, using the approximation $\log(1 + x) \approx x$ yields that $\psi(\theta) \approx \theta\lambda e^{-b}/(1 - \theta)$. Thus the term inside the infimum in relation (18) becomes

$$\frac{1}{\theta}\left(\log(4\gamma) + K\psi(\theta)\right) \approx \frac{1}{\theta}\log(4\gamma) + \frac{1}{1-\theta}(K\lambda e^{-b}).$$

Note that $A/\theta + B/(1 - \theta)$ has a minimum value $(\sqrt{A} + \sqrt{B})^2$ over $0 \leq \theta \leq 1$ for any $A, B > 0$. Hence, relation (18) in Theorem 3.1 becomes

$$a \approx \left(\sqrt{\log(4\gamma)} + \sqrt{K\lambda e^{-b}}\right)^2. \tag{22}$$

It is interesting to note that (22) demonstrates the theoretical challenges of monitoring large-scale data streams: the asymptotic expression of a in (22) depends on the asymptotic relationship between $\log(\gamma)$ and $K\lambda e^{-b}$. When $\log(\gamma) >> K$, we would have the classical result on the threshold of $a = (1 + o(1)) \log(\gamma)$, see Lorden [18]. However, when $K\lambda e^{-b} >> \log(\gamma)$, the first two major terms of (22) become

$$a \approx K\lambda e^{-b} + 2\sqrt{K\lambda e^{-b}}\sqrt{\log \gamma}. \tag{23}$$

This suggests that $K\lambda e^{-b}$ will play a dominant role to determine the threshold a for $N_G(a)$ to satisfy the false alarm constraint γ in (3) when b is large and $Ke^{-b} >> \log(\gamma)$.

It is interesting to point out that the Chebyshev's inequalities approximation of a in (23) is asymptotically equivalent to the following CLT approximation of a suggested by Professor Benjamin Yakir in a personal communication. Under the homogenous scenario of $\psi_k(\theta) = \psi(\theta)$ for all k, instead of Chebyshev's inequality, one can use the CLT approximation to $\sum_{k=1}^K H_k^*$:

$$\mathbf{P}^{(\infty)}\Big(\sum_{k=1}^K H_k^* \geq a\Big) \approx \mathbf{P}\Big(N(0,1) \geq \frac{a - K\mu_H}{\sigma_H\sqrt{K}}\Big),$$

where μ_H and σ_H^2 are the mean and variance of the limiting statistic H^*, and can be found as the derivatives of the log-moment generating function $\psi(\theta)$ in (16):

$$\mu_H = \dot{\psi}(0) \quad \text{and} \quad \sigma_H^2 = \ddot{\psi}(0), \tag{24}$$

In addition, if we approximate the distribution of $N_G(a)$ as an exponentially distribution, which holds for many sequential change-point detection schemes in the literature as well as the SUM scheme in Mei[21], then we have $\mathbf{E}^{(\infty)}(N_G) \approx x/\mathbf{P}^{(\infty)}(N_G(a) \leq x)$ for moderately large x. Combining these above two approximations yields a heuristic approximation

$$a = K\mu_H + z_{1/\gamma}\sigma_H\sqrt{K} \tag{25}$$

where $z_{1/\gamma} = z$ so that $\mathbf{P}(N(0,1) \geq z) = 1/\gamma$. Our numerical simulations below supports this heuristic approximation, but rigorous justifications turn out to be highly technical due to the complicated correlation structures of the $W_{k,n}$'s over time domain n, and will be presented elsewhere. In the special case of the soft-thresholding transformation, it is straightforward to compute from (21) and (24) that $\mu_H = \lambda e^{-b}$ and $\sigma_H^2 = \lambda e^{-b}(2 - \lambda e^{-b}) \approx 2\lambda e^{-b}$ for large b. This yields that

$$a \approx K\lambda e^{-b} + z_{1/\gamma}\sqrt{2K\lambda e^{-b}}. \tag{26}$$

By the well-known fact that $\frac{1}{z}\phi(z) \leq \mathbf{P}(N(0,1) \geq z) \leq \frac{1}{z+1/z}\phi(z)$ for any $z > 0$ where $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$ is the pdf of $N(0,1)$, it is not difficult to show that $z_{1/\gamma} \approx \sqrt{2\log\gamma}$ for large γ. Hence, when b is large and $K\lambda e^{-b} >> \log\gamma$, the CLT approximation in (26) is asymptotically equivalent to the Chebyshev's inequalities approximation in (23). This further demonstrates that the bound of a in (18) of Theorem 3.1 derived from the Chebyshev's inequalities is not bad.

3.3 The Choice of Censoring Parameters

So far we analyzed the properties of our proposed SUM-shrinkage schemes for fixed shrinkage transformations, and the censoring parameters b_k's in the shrinkage transformation were defined intuitively in (11) from the communication considerations in the context of censoring sensor networks. In this subsection, we take a further step to discuss the optimal choice of the censoring parameters b_k's in (11). For the purpose of illustration and simplicity, we consider the homogeneous case of $b_k \equiv b$ when local data streams are identically distributed for different k, e.g., relations (19), (20), and $\psi_k(\theta) \equiv \psi(\theta)$ in (21) hold for all $k = 1, \cdots, K$. Here we provide two different optimal choices of the censoring parameter b for the soft-thresholding scheme $N_{soft}(a)$ in (13): one from the communication efficiency aspect, and the other from the statistical efficiency aspect, and it turns out that they are closely related.

Let us first choose the censoring parameter b based on the communication considerations in censoring sensor networks. Assume that the average fraction of transmitting sensors at any time step is restricted to be at most $\eta \in (0,1)$ when no change occurs. In this case, when no event occurs, the average fraction of transmitting sensors at any time step n is

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{P}^{(\infty)}(W_{k,n} \geq b) \leq \frac{1}{K}\sum_{k=1}^{K}\exp(-b) \leq \exp(-b),$$

where the second-to-last inequality follows from the non-asymptotic bound (20). Thus a choice of

$$b_{opt,1} = \log(\eta^{-1}) \tag{27}$$

will guarantee that on average, at most $100\eta\%$ of K sensors will transmit messages at any given time when no event occurs. When η is small, one may use the refined asymptotic approximation (19), instead of the non-asymptotic bound (20), in the above arguments. Then the $b_{opt,1}$ in (27) can further improved as $b_{opt,1} = \log(\lambda/\eta)$ under the communication constraint.

Next, let us choose the censoring parameter b based on the statistical efficiency considerations in the scenario when w_0 out of K local data streams are affected. Intuitively, our proposed scheme $N_{soft}(a)$ in (13) is increasing as a function of the censoring parameter $b_k \equiv b$ when the global threshold value a is *given*. That is, a larger value of b implies both larger ARL to false alarm and larger detection delays. Hence, subject to the false alarm constraint γ in (3), different global threshold values a's are needed for these schemes with different b's, and thus it is natural to find the censoring parameter b that yields the smallest detection delay $\overline{\mathbf{E}}(T)$ in (2).

To do so, we further assume that those affected local streams have the same post-change statistical properties in the sense that the detection delay of a local scheme $N_k(c) = \inf\{n \geq 1 : W_{k,n} \geq c\}$ is $(1 + o(1))c/I$ for some constant $I > 0$ as $c \to \infty$. This assumption is very general and holds for many local detection statistics including CUSUM, see Lorden [18]. Then the detection delay of the soft-thresholding scheme $N_{soft}(a)$ in (13) will be bounded above by

$$(1 + o(1))\frac{1}{I}\left(b + \frac{a}{w_0}\right). \tag{28}$$

To see this, at time step n, if $w_{k,n} \geq b + a/w_0$ for all of those w_0 affected local data streams, then $N_{soft}(a) \leq n$ since $\sum_{k=1}^{K} \max(w_{k,n} - b, 0) \geq w_0(a/w_0) = a$. Relation (28) follows at once from the detection delays of $N_k(c)$ with $c = b + a/w_0$ for those w_0 affected data streams, and similar ideas have been applied in the proof of Theorem 3 in Mei [20] in a concrete scenario when the $W_{k,n}$'s are local CUSUM statistics.

Now let us keep only on the first-order major term of a in (23), plugging it into (28) yields that the detection delay of the soft-thresholding scheme $N_{soft}(a)$ in (13) (up to the first-order) is

$$\frac{1}{I}\left(b + \frac{K\lambda e^{-b}}{w_0}\right).$$

Taking derivatives with respect to b and setting it to 0, the detection delay bound is minimized when $K\lambda e^{-b} = w_0$, i.e., the optimal b value (up to first-order) is given by

$$b_{opt,2} = \log\frac{\lambda K}{w_0}, \tag{29}$$

where $\lambda > 0$ is the constant in (19) that only depends on the asymptotic properties of the $W_{k,n}$'s.

It is useful to add some remarks that help us better understand the optimal $b_{opt,2}$ value in (29). First, when we have a prior knowledge that w_0 local data streams are affected but we do not know which w_0 data streams are affected, it is reasonable to assume that each local data stream has the same probability $\pi = w_0/K$ of being affected. By the semi-Bayesian interpretation of the soft-thresholding transformation in subsection 3.1, the local censoring parameters b_k's should be chosen as $b_k = \log((1-\pi)/\pi) = \log((K-w_0)/w_0)$, which is asymptotically equivalent to (29) when the fraction of affected data stream $w_0/K \to 0$.

Meanwhile, it is also interesting to interpret the optimal $b_{opt,2}$ value in (29) back to the communication constraints in sensor networks. A direct comparison of (27) and (29) suggests that these two optimal b values are asymptotically equivalent if we set $\eta = w_0/K$. Moreover, by (11) and (19), when there are no changes, the average fraction of transmitting sensors at any time step n is

$$\frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) = \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) = \mathbf{P}^{(\infty)}(W_{k,n} \geq b_{opt,2}) = \lambda e^{-b_{opt,2}} = w_0/K. \tag{30}$$

This demonstrates a simple but useful equivalence relationship between communication efficiency and statistical efficiency: if we want to optimize the detection delay performance (up to first-order) when w_0 data streams are affected, then it will be the best to design the schemes that on average allow w_0 out of K local data streams (i.e., a fraction w_0/K local data streams) to transmit local detection statistics to the fusion center at every time when no change event occurs (and possibly more than w_0 data streams when a change occurs). Due to this equivalence, in our simulations below, the censoring parameter b will be chosen based on (27), which is non-asymptotic and easier to compute.

4 A First Example: Known Post-Change Distributions

In this section we assume that we are monitoring K data streams $X_{k,n}$'s in (1) in the context of censoring sensor networks in Figure 1 when the data streams might be nonhomogeneous. We follow the literature to make a restrictive assumption that the pre-change and post-change distributions of the $X_{k,n}$'s are completely specified. That is, we assume that for each $k = 1, \ldots, K$, the observations $X_{k,n}$'s at the k-th sensor or data stream are iid with one known density f_k before the change, and iid with another known density g_k after the change if the k-th data stream is affected, where the f_k's and g_k's are completely specified densities with respect to a suitable measure μ_k, see, for example, Tartakovsky and Veeravalli [37]. For each $1 \leq k \leq K$, we assume that the Kullback-Leibler (KL) information number

$$I(g_k, f_k) = \int \log \frac{g_k(x)}{f_k(x)} g_k(x) d\mu_k(x) \tag{31}$$

is finite and positive, and

$$\int \Big(\log \frac{g_k(x)}{f_k(x)} \Big)^2 g_k(x) d\mu_k(x) < \infty. \tag{32}$$

For easy understanding the application of our proposed SUM-shrinkage schemes, we divide this section into three subsections. The local detection statistics $W_{k,n}$'s are constructed in Subsection 4.1, and then the suitable choices of tuning parameters in the shrinkage transformations are discussed in Subsection 4.2. Numerical simulation results are reported in Subsection 4.3.

4.1 Local detection statistics

When the local pre-change and post-change distributions, f_k and g_k, are completely specified, the corresponding local change detection problem has been well-studied, and many efficient local detection statistics $W_{k,n}$'s are available in the literature. For instance, the $W_{k,n}$'s can be chosen as the well-known local CUSUM statistics (Page [26]) that are defined recursively by

$$W_{k,n} = \max\Big(W_{k,n-1} + \log\frac{g_k(X_{k,n})}{f_k(X_{k,n})},\ 0\Big), \tag{33}$$

for $n \geq 1$ and $W_{k,0} = 0$ for $k = 1, \cdots, K$. As shown in Lorden [18] and Moustakides [24], the local CUSUM statistics $W_{k,n}$'s in (33) enjoy certain optimality properties when detecting the local change.

Besides the local CUSUM statistics, another popular local detection statistic is the local Shiryaev-Roberts statistic (Shiryaev [33], Roberts [31]) which can be defined in the log-likelihood ratio scale by

$$\hat{W}_{k,n} = \log\Big(\exp(\hat{W}_{k,n-1}) + 1\Big) + \log\frac{g_k(X_{k,n})}{f_k(X_{k,n})} \tag{34}$$

for $n \geq 1$ and $\hat{W}_{k,0} = 0$. It is well-known that the local Shiryaev-Roberts statistics $\hat{W}_{k,n}$'s in (34) yield an efficient local detection procedure whose performance is similar to that of local CUSUM statistics in (33) when detecting a local change in distribution from f_k to g_k, see Pollak [27, 28].

In our numerical analysis below, the local detection statistics $W_{k,n}$'s can also be defined as $\hat{W}_{k,n}$'s in (34), the local Shiryaev-Roberts statistics in logarithm scale, or better yet, its positive part $\max\{\hat{W}_{k,n}, 0\}$. Our numerical simulation experiences suggest that the performances of global monitoring schemes based upon local Shiryaev-Roberts statistics are similar to those based upon local CUSUM statistics in (33) when monitoring K data streams. Unfortunately it is still an open question to investigate the theoretical properties of Shiryaev-Roberts-type schemes in the context of K data streams, and thus we will focus on the local CUSUM statistics $W_{k,n}$'s in (33) as the local detection statistics below.

4.2 Local shrinkage transformations

With the local detection statistics $W_{k,n}$'s in (33), local sensors can send sensor messages $U_{k,n} = W_{k,n} I\{W_{k,n} \geq b_k\}$ as in (11), and then the fusion center can use one of the four global monitoring schemes in (12)-(15). A key question is how to choose the local censoring parameters b_k's in (11) suitably to ensure statistical

efficiency. Intuitively, the b_k's should be the same when the sensors are homogeneous, but they probably should be different when the sensors are nonhomogeneous.

It turns out that a "good" choice is

$$b_k = \rho_k b_g \tag{35}$$

for $k = 1, \ldots, K$ for some common global-level constant $b_g \geq 0$, where

$$\rho_k = \frac{I(g_k, f_k)}{\sum_{k=1}^{K} I(g_k, f_k)} \tag{36}$$

and $I(g_k, f_k)$ is the KL information number defined in (31).

At the high-level, ρ_k in (36) can be thought of as the weight of the k-th data stream in the overall final decision. From the technical viewpoint, the choice of $b_k = \rho_k b_g$ in (35) allows those affected local sensors to send local messages $U_{k,n}$'s with large values to the fusion center at roughly the same time, thereby leading quick detection of the occurring event. A more rigorous justification of our choice of ρ_k in (36), or b_k in (35)-(36), is briefly summarized in [22], though we should emphasize that the choice of b_k in (35)-(36) is a sufficient but not necessarily necessary condition in order for our proposed schemes in (12)-(15) to enjoy good properties.

As for the choice of the common global-level constant $b_g > 0$ in (35), in the general non-homogeneous context, it remains an open problem to optimally choose the common global-level constant $b_g > 0$ in (35) from the statistical efficiency viewpoint as in (29). Fortunately, one can still easily determine it from the communication efficiency viewpoint when the average fraction of transmitting sensors at any time step is restricted to be at most $\eta \in (0, 1)$ when no change occurs. In this case, when no event occurs, the average fraction of transmitting sensors at any time step n is

$$\frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(W_{k,n} \geq \rho_k b_g) \leq \frac{1}{K}\sum_{k=1}^{K} \exp(-\rho_k b_g) \leq \exp(-\rho_{\min} b_g),$$

where $\rho_{\min} = \min_{1 \leq k \leq K} \rho_k$ and the second-to-last inequality follows from the well-known properties of the local CUSUM statistics that $\mathbf{P}^{(\infty)}(W_{k,n} \geq a) \leq \exp(-a)$ for all $a > 0$, see, for example, Appendix 2 on Page 245 of Siegmund [34]. Thus a choice of $b_g = (1/\rho_{\min}) \log \eta^{-1}$ will guarantee that on average, at most $100\eta\%$ of K sensors will transmit messages at any given time when no event occurs. It is not difficult to see that this general result reduces to $b_{opt,1}$ in (27) in subsection 3.3 under the homogeneous case when ρ_k in (36) are constant over k.

4.3 Numerical Simulations

In this subsection we report our numerical simulation results to illustrate the usefulness of the proposed schemes in (12)-(15). Suppose that there are $K = 100$ independent and identical sensors in a system, and

the observations at each sensor are iid with mean 0 and variance 1 before the change and with mean 1 and variance 1 after the change if affected. In our simulation study, we simply assume that the change is instantaneous if a sensor is affected, but we do not know which subset of sensors will be affected.

For the purpose of comparison, we conduct numerical simulations for six families of global monitoring schemes:

- the "MAX" scheme $T_{\max}(a)$ in (4),
- the "SUM" scheme $T_{\text{sum}}(a)$ in (5),
- the order thresholding scheme $N_{order,r}(a)$ in (15) with $r = 10$,
- the hard thresholding scheme $N_{hard}(a)$ in (12),
- the soft thresholding scheme $N_{soft}(a)$ in (13),
- the combined thresholding schemes $N_{comb,r}(a)$ in (14) with $r = 10$.

The first three schemes require all local sensors to send all local CUSUM statistics $W_{k,n}$'s values to the fusion center at each and every time step, and corresponds to the case when the local censoring parameter $b_k \equiv 0$ for all $k = 1, \cdots, K$. For order-thresholding in the families of $N_{order,r}(a)$ and $N_{comb,r}(a)$, we choose $r = 10$ to better understand the scenario when 10 out of 100 sensors are affected by the occurring event. For each of the last three schemes in the list, i.e., our three proposed schemes (12)-(14), we further consider three different values of the local censoring parameters b_k's:

(i) $b_k \equiv 1/2 \approx -\log(0.607)$ for all k,

(ii) $b_k \equiv -\log(0.1) = 2.3026$ for all k,

(iii) $b_k \equiv -\log(0.01) = 4.6052$ for all k.

The choices of these values will guarantee that when no event occurs, on average at most $\eta = 60.7\%, 10\%$, and 1% of $K = 100$ homogeneous sensors will transmit messages at any given time, respectively. Therefore, there are a total of $3 + 3 * 3 = 12$ specific schemes in our numerical simulation study.

For each of these 12 specific schemes $T(a)$, we first find the appropriate values of the global threshold a to satisfy the false alarm constraint $\mathbf{E}^{(\infty)}(T(a)) \approx \gamma = 5000$ (within the range of sampling error). Next, using the obtained global threshold value a, 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 $m = 2500$ repetitions.

Table 1 summarizes our simulated detection delays of these 12 schemes under 8 different post-change hypothesis, depending on the number of affected sensors. From Table 1, among these 12 specific schemes,

Table 1: A comparison of the detection delays of six families of schemes with $\gamma = 5000$. The smallest and largest standard errors of these 12 schemes are also reported under each post-change hypothesis based on 2500 repetitions in Monte Carlo simulations.

	# sensors affected								
	1	3	5	8	10	20	30	50	100
Smallest standard error	0.18	0.07	0.05	0.03	0.03	0.02	0.01	0.01	0.00
Largest standard error	0.35	0.12	0.07	0.06	0.05	0.04	0.03	0.03	0.03
Schemes with $b_k \equiv 0$									
$T_{\max}(a = 11.27)$	23.3	16.3	14.4	13.0	12.4	10.9	10.2	9.5	8.7
$T_{\text{sum}}(a = 88.66)$	52.1	21.8	14.7	10.3	8.7	5.2	3.9	2.9	2.0
$N_{order,r=10}(a = 44.11)$	34.1	15.5	11.2	8.5	7.5	5.5	4.8	4.1	3.4
Schemes $N_{hard}(a)$ in (12) with different positive b_k's									
$N_{hard}(a = 85.60, b_k = 0.50)$	52.9	21.9	14.9	10.3	8.7	5.2	4.0	2.9	2.0
$N_{hard}(a = 52.21, b_k = 2.3026)$	50.6	20.7	13.8	9.6	8.2	5.2	4.2	3.2	2.4
$N_{hard}(a = 26.31, b_k = 4.6052)$	39.8	16.0	11.5	8.8	7.9	5.9	5.2	4.4	3.8
Schemes $N_{soft}(a)$ in (13) with different positive b_k's									
$N_{soft}(a = 63.92, b_k = 0.50)$	48.2	20.2	13.7	9.7	8.2	5.1	4.0	3.0	2.0
$N_{soft}(a = 21.56, b_k = 2.3026)$	33.9	15.4	11.2	8.5	7.5	5.3	4.5	3.7	3.0
$N_{soft}(a = 8.29, b_k = 4.6052)$	25.2	13.8	11.1	9.2	8.4	6.7	5.9	5.2	4.4
Schemes $N_{comb,r}(a)$ in (14) with $r = 10$ and different positive b_k's									
$N_{comb,r}(a = 44.11, b_k = 0.50)$	34.1	15.5	11.2	8.5	7.5	5.5	4.8	4.1	3.4
$N_{comb,r}(a = 43.88, b_k = 2.3026)$	38.5	16.8	11.7	8.6	7.5	5.5	4.7	4.0	3.3
$N_{comb,r}(a = 26.31, b_k = 4.6052)$	39.8	16.0	11.5	8.8	7.9	5.9	5.2	4.4	3.8

when a small number ($1 \sim 3$) of 100 homogeneous sensors are affected by the event, the "MAX" scheme $T_{\max}(a)$ is the best (in the sense of smallest detection delay), the "SUM" scheme $T_{\text{sum}}(a)$ is the worst, and all other schemes are in-between. Similarly, when a large number (20 or more) of 100 homogeneous sensors are affected, the order is reserved: $T_{\text{sum}}(a)$ is the best, $T_{\max}(a)$ is the worst, and all other schemes are in-between. However, when $5 \sim 10$ sensors are affected, the schemes with order-thresholding $r = 10$ yield the smallest detection delays, since they are designed to detect the scenario when 10 sensors are affected by the event. In addition, it is clear from Table 1 that for each given scheme, the fewer affected sensors we have, the larger detection delay it will have. All these results are consistent with our intuition.

It is worth emphasizing that for the families of the hard- and soft- thresholding schemes, $N_{hard}(a)$ in (12) and $N_{soft}(a)$ in (13), a larger censoring value of b_k actually leads to a smaller detection delay when only a few sensors are affected. This suggests that a larger censoring value b_k may actually be necessary for efficient detection when the affected sensors are sparse. Indeed, our theoretical results in (27) and (29) in Section 3.3 suggest that the soft-thresholding schemes $N_{soft}(a)$ with censoring parameters $b_k \equiv 1/2 \approx -\log(0.607)$, $b_k \equiv -\log(0.1) = 2.3026$, and $b_k \equiv -\log(0.01) = 4.6052$ would have the smallest detection delay when the number of truly affected sensors are around 60.7, 10, and 1, respectively. These theoretical results are consistent with Monte Carlo simulation results in Table 1.

A surprising and possibly counter-intuitive result in Table 1 is the effect of not so large values of censoring parameters b_k's in finite sample simulations. For instance, the performances of the "SUM" scheme $T_{\text{sum}}(a)$ and the hard thresholding scheme $N_{hard}(a, b_k = 0.50)$ are similar in view of sampling errors. Likewise, the top-r thresholding scheme $N_{order,r=10}(a)$ and the combined thresholding scheme $N_{comb,r=10}(a, b_k = 0.50)$ also have identical performances. The interpretation in the censoring sensor networks context is as follows: using our proposed communication policy in (11), we only need $\exp(-b_k) = \exp(-0.5) = 60.7\%$ of 100 sensors to transmit information to the fusion center at any given time when no event occurs, but we can still be as effective as the full transmission scenario when all sensors transmit information at all time steps. In other words, much communication costs can be saved by our proposed schemes $N_{hard}(a)$ or $N_{comb,r}(a)$ with not so large values of b_k's.

It is also interesting to see the effect of the order-thresholding parameter r in finite sample simulations when the hard-thresholding parameters b_k's are large. From Table 1, when the false alarm constraint γ in (3) is only moderately large, e.g., $\gamma = 5000$, the performances of $N_{hard}(a, b_k)$ and $N_{comb,r=10}(a, b_k)$ are identical when $b_k = 4.6052$ — they not only have the same global threshold a, but also have the same detection delays. Intuitively, the stopping time $N_{comb,r}(a, b_k)$ is decreasing as a function of r, and thus we have $N_{hard}(a, b_k) = N_{comb,r=K}(a, b_k) \leq N_{comb,r=10}(a, b_k)$ when $b_k = 4.6052$. So one may wonder why our numerical simulations lead to identical results? One explanation is that with such a choice of $b_k = 4.6052$, when no event occurs, on average there is at most 1 non-zero sensor message received in the fusion center

at any given time, and thus there is little difference whether one uses the sum of the largest $r = 10$ sensor messages or uses the sum of all $K = 100$ sensor messages. Hence similar performances are observed in finite-sample simulations.

5 A Second Example: Unknown Post-Change Means

Suppose that we are monitoring K data streams $X_{k,n}$'s in (1). Initially, the data $X_{k,n}$'s are iid $N(0,1)$. At some unknown time ν, the distribution of the k-th local data stream might change to $N(\mu_k, 1)$ if affected. As in the previous section, we do not know which subset of local data streams are affected, but here we add a new challenge that we do not know the values of the post-change means μ_k's when affected. We want to develop a system-wise online monitoring scheme that can detect the change as soon as possible, subject to the global false alarm constraint γ in (3).

Xie and Siegmund [40] investigates this problem under the assumption that the post-change mean $\mu_k > 0$ for all k. By assuming that the fraction p_0 of affected data stream is known, the main scheme they proposed is motivated from a semi-Bayesian approach and is defined by

$$T_{XS}(a, p_0) = \inf\left\{ n \geq 1 : \max_{0 \leq i < n} \sum_{k=1}^{K} \log(1 - p_0 + p_0 \exp\left[\left(U^+_{k,n,i}\right)^2 / 2 \right] \geq a \right\}. \tag{37}$$

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

$$U^+_{k,n,i} = \max\left(0, \frac{1}{\sqrt{n-i}} \sum_{j=i+1}^{n} X_{k,j}\right).$$

Some simplified versions have also been proposed to reduce the memory requirement to a large window of the most recent observations. In addition, Wang and Mei [42] proposed a global Shiryaev-Robert procedure by simultaneously estimating all K unknown post-change means μ_k via shrinkage estimation. However, all these existing schemes are not scalable, and also not suitable in the context of censoring sensor networks in Figure 1: besides being computationally expensive, the implementation of their schemes requires the fusion center to have full access to all data streams at each time step.

It has been an open problem to develop a scalable global monitoring scheme that is able to detect both positive and negative local mean shifts for affected local data streams. Part of the reason is that for the K local data streams, there are 2^K potential different combinations of positive or negative local shifts, which is huge for a large K. Here we illustrate how to tackle this open problem based upon our proposed SUM-shrinkage statistics in (6). The main challenge is to choose a suitable local detection statistic $W_{k,n}$ that can be easily computed and has the ability to detect both positive and negative local mean shifts. Once such local detection statistic $W_{k,n}$'s are defined, we can use any shrinkage transformation such as hard-, soft- or order- thresholding to develop a global monitoring scheme.

In this section, we use the soft-thresholding transformation as a demonstration, and consider the soft-thresholding scheme $N_{soft}(a)$ in (13). For simplicity, we assume that all censoring parameters b_k's in (13) are the same, i.e., $b_k \equiv b_1$ for some constant $b_1 > 0$. Our focus is how we can construct the local detection statistics $W_{k,n}$'s suitably.

The remainder of this section is as follows. Subsection 5.1 reviews the recursive register approach of Lorden and Pollak [19] for monitoring a single data stream, which is adapted to monitoring positive and negative mean shifts in Subsection 5.2. Subsection 5.3 provides an overview of our proposed soft-thresholding scheme in (13) as well as an efficient numerical algorithm of our scheme that only uses fixed $6K$ registers to store all past information and involves $O(K)$ computations at each given time step n. Numerical simulation results are summarized in subsection 5.4.

5.1 The Recursive Register Approach of Lorden and Pollak [19]

To abuse notation, in this subsection we suppress the subscript k of the k-th data stream, and consider the local monitoring problem with respect to the one-dimensional data stream $\{X_1, X_2, \ldots\}$ whose distribution may change from $N(0,1)$ to $N(\mu,1)$ with unknown post-change mean μ at some unknown time ν. Lorden and Pollak [19] focuses on the case when the unknown post-change mean $\mu > 0$, and makes a technical assumption that $\mu \geq \rho$, where $\rho \geq 0$ is the smallest mean shift that is meaningful in practice, e.g. $\rho = 0.25$.

A high-level description of the recursive register approach of Lorden and Pollak [19] is as follows. Recall that the CUSUM statistics are defined in (33), and for one-dimensional normal distributed data, the CUSUM statistics have a simpler recursive form

$$W_n = \max(W_{n-1} + \mu X_n - \frac{1}{2}\mu^2, 0) \tag{38}$$

and $W_0 = 0$. When μ is unknown, we can continue to use this recursive formula to define a detection statistic if we replace the true unknown μ by its estimate from the past observed data. A key observation in Lorden and Pollak [19] is that at each given time step n, the CUSUM-type detection statistics can produce a candidate post-change time $\hat{\nu} \in \{0, 1, \cdots, n-1\}$, and thus the observations $X_{\hat{\nu}}, X_{\hat{\nu}+1}, \cdots, X_{n-1}$ can be used to estimated the post-change mean μ in (38). Specifically, at any given time step n, define $\hat{\nu}$ as the largest $0 \leq i \leq n-1$ such that $W_i = 0$, and denote by T_n and S_n the total number and the summation of observations X_i's between the candidate post-change time $\hat{\nu}$ and time step $n-1$. That is,

$$T_n = n - \hat{\nu} \qquad \text{and} \qquad S_n = \sum_{i=\hat{\nu}}^{n-1} X_i. \tag{39}$$

By the method of moments estimator or maximum likelihood estimator method, the post-change mean μ can be estimated by S_n/T_n at time step n. If we treat the pre-specified nonnegative constants s and t as a prior, then a Bayes-type estimate of μ is $\hat{\mu}_n = (s + S_n)/(t + T_n)$, which includes S_n/T_n as a special case

when $s = t = 0$. After taking into account that ρ is the smallest post-change mean we are interested in, one can estimate μ at time step n by

$$\hat{\mu}_n = \max\Big(\rho, \frac{s + S_n}{t + T_n}\Big). \tag{40}$$

From the algorithm viewpoint, the recursive register approach of Lorden and Pollak [19] can be recursively implemented as follows. Let $S_0 = T_0 = W_0 = X_0 = 0$, and $\hat{\mu}_1 = \rho$. For all $n \geq 1$,

$$W_n = \max\left(W_{n-1} + \hat{\mu}_n X_n - \frac{1}{2}(\hat{\mu}_n)^2, 0\right), \tag{41}$$

where $\hat{\mu}_n$ is defined in (40) and

$$\begin{pmatrix} S_n \\ T_n \end{pmatrix} = \begin{cases} \begin{pmatrix} S_{n-1} + X_{n-1} \\ T_{n-1} + 1 \end{pmatrix} & \text{if } W_{n-1} > 0 \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } W_{n-1} = 0. \end{cases} \tag{42}$$

In other words, the local detection statistics W_n's can be computed recursively as the part of three-dimensional vectors (S_n, T_n, W_n), or four-dimensional vectors $(S_n, T_n, \hat{\mu}_n, W_n)$. It is important to note that $(S_n, T_n, \hat{\mu}_n)$ only uses the observations up to time $n-1$ for the purpose of estimating the post-change mean μ, so that the data X_n is reserved for the local detection statistics W_n for the purpose of detecting changes. It was shown that the detection scheme based on the detection statistic W_n in (41) is asymptotically optimal whenever the true post-change mean $\mu \geq \rho > 0$, see Theorems 3.1-3.3 of Lorden and Pollak [19].

5.2 Our Proposed Local Detection Statistics $W_{k,n}$'s

Since we are interested in detecting both positive and negative local mean shifts for affected data streams, we propose to extend the detection statistic W_n in (41) of Lorden and Pollak [19] from one-sided to two-sided. Observe that detecting negative local mean shift of $X_{k,n}$'s is equivalent to detecting positive local mean shift of $-X_{k,n}$'s, we propose the following two-sided local detection statistic for each local data stream at time n :

$$W_{k,n} = \max\left(W_{k,n}^{(1)}, W_{k,n}^{(2)}\right), \tag{43}$$

where $W_{k,n}^{(1)}$ and $W_{k,n}^{(2)}$ are the local detection statistics of Lorden and Pollak [19] for detecting positive and negative mean shifts, respectively. Specifically,

$$\begin{aligned} W_{k,n}^{(1)} &= \max\left(W_{k,n-1}^{(1)} + \hat{\mu}_{k,n}^{(1)} X_{k,n} - \frac{1}{2}(\hat{\mu}_{k,n}^{(1)})^2, 0\right), \\ W_{k,n}^{(2)} &= \max\left(W_{k,n-1}^{(2)} + \hat{\mu}_{k,n}^{(2)} X_{k,n} - \frac{1}{2}(\hat{\mu}_{k,n}^{(2)})^2, 0\right), \end{aligned} \tag{44}$$

where

$$\hat{\mu}_{k,n}^{(1)} = \max\Big(\rho, \frac{s + S_{k,n}^{(1)}}{t + T_{k,n}^{(1)}}\Big) > 0, \qquad \hat{\mu}_{k,n}^{(2)} = \min\Big(-\rho, \frac{-s + S_{k,n}^{(2)}}{t + T_{k,n}^{(2)}}\Big) < 0, \tag{45}$$

and for $j = 1, 2$ and for any k, the sequences $(S_{k,n}^{(j)}, T_{k,n}^{(j)})$ are defined recursively

$$\begin{pmatrix} S_{k,n}^{(j)} \\ T_{k,n}^{(j)} \end{pmatrix} = \begin{cases} \begin{pmatrix} S_{k,n-1}^{(j)} + X_{k,n-1} \\ T_{k,n-1}^{(j)} + 1 \end{pmatrix} & \text{if } W_{k,n-1}^{(j)} > 0 \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } W_{k,n-1}^{(j)} = 0 \end{cases} \tag{46}$$

Note that $\hat{\mu}_{k,n}^{(1)}$ and $\hat{\mu}_{k,n}^{(2)}$ in (45) are the estimates of the post-change mean when restricted to the positive and negative values, respectively, under the assumption that $|\mu| \geq \rho$. Clearly, $W_{k,n}^{(1)}$ is designed to detect positive local mean shift, whereas $W_{k,n}^{(2)}$ is to detect negative local mean shifts. Also the two-sided local detection statistic $W_{k,n}$ in (43) is always nonnegative for any k at any time step n, and it will become large when there is a local mean shift no matter whether such mean shift is positive or negative.

5.3 Overview of Our Proposed Scheme

Note that the proposed soft-thresholding scheme $N_{soft}(a)$ in (13) can be easily implemented in the censoring sensor network context by parallel computing the K local detection statistics $W_{k,n}$'s recursively through (43)-(46) at the local sensor levels. To be more specific, we can use the following $6K$ registers to adaptively store all past information at each time step after observing new data: $(S_k^{(j)}, T_k^{(j)}, W_k^{(j)})$ for $j = 1, 2$ and $k = 1, 2 \cdots, K$. At any given time step n, we can first update the $4K$ registers in $(S_k^{(j)}, T_k^{(j)})$ using the past data and compute the $2K$ estimates $\hat{\mu}_k^{(j)}$ of the post-change means μ_k's. Then after we observe new observations, $(X_{1,n}, \cdots, X_{K,n})$, we only need to update the $2K$ registers $W_k^{(j)}$'s and compute the values of K local detection statistics W_k's, which allows us to easily compute the global monitoring statistic G. Including the $3K$ intermediate variables $(\hat{\mu}_k^{(j)}, W_k)$ and the global monitoring statistic G, the proposed scheme only needs $9K + 1$ registers to adaptively store all relevant information and involves $O(K)$ computations at any given time step n. Moreover, our proposed scheme can be implemented in the context of censoring sensor networks in the previous section where most computations are done at the remote sensors. Hence, our proposed scheme is scalable and can be easily implemented to online monitor large-scale data streams over a long time period.

When the local detection statistics $W_{k,n}$'s are defined in (43)-(46), an overview of our proposed soft-thresholding scheme, $N_{soft}(a)$ in (13), is illustrated in the following algorithm:

Algorithm: Implementation of N_{soft} in (13) when the $W_{k,n}$'s are defined in (43)-(46)

Initial parameters: ρ, s, t, and b_1 for $k = 1, \cdots, K$.

Set: A terminal threshold a.

Algorithm:

initialize $n = 0$, and set all initial observations $X_k = 0$ and all $8K$ initial registers $S_k^{(j)} = T_k^{(j)} = \mu_k^{(j)} = W_k^{(j)} = 0$, for $k = 1, \ldots, K$ and $j = 1, 2$.

While the scheme N_{soft} has not raised an alarm

do 1. Update $4K$ registers $(S_k^{(j)}, T_k^{(j)})$ via (46).

2. Compute the $2K$ intermediate variables $\hat{\mu}_k^{(j)}$ from (45) which are the estimates of the post-change means.

3. Input new observations from all K data streams, denoted by $(X_1, \cdots, X_K)$.

4. For $k = 1, \ldots, p$, recompute the local monitoring statistics $W_k^{(j)}$'s in (44) and W_k in (43).

5. Compute the global monitoring statistics

$$G = \sum_{k=1}^{K} \max(W_k - b_1, 0)$$

if $G \geq a$ **terminate:** Raising an alarm at time n and declaring that a change has occurred;

end the while loop

5.4 Simulation Results

In this subsection, we report the numerical simulation results of the soft-thresholding scheme $N_{soft}(a)$ in (13) when the local detection statistics $W_{k,n}$'s are defined recursively through (43)-(46), and the censoring parameters $b_k \equiv b_1$ for all k. For the purpose of comparison, we follow Xie and Siegmund [40] to assume that there are $K = 100$ independent normal data streams. For each $k = 1, \cdots, K$, the data $X_{k,n}$'s of the k-th data stream are iid $N(0, 1)$ before the change, but are iid $N(1, 1)$ after the k-th data stream is affected by the occurring event.

In our simulations, we consider six schemes: two of them are the Xie and Siegmund schemes $T_{XS}(a, p_0)$ in (37) with $p_0 = 1$ and 0.1; and the remaining four schemes are our proposed soft-thresholding schemes $N_{soft}(a)$ in (13) with four different censoring parameters: $b_1 = 0, 0.5, \log(10), \log(100)$. As in the previous sections, the three non-zero b_1 values imply that on average at most $\exp(-b_1) \approx 60.7\%, 10\%$ and 1% out of 100 local data streams produce significant $W_{k,n}$'s values to the global monitoring statistic G_n when there are no changes. When computing the local detection statistics $W_{k,n}$'s in (43), we set $\rho = 0.25, t = 4$ and $s = 1$ as in Lorden and Pollak [19].

Table 2: A comparison of detection delays when the change is instantaneous and the post-change mean $\mu_k = 1$ if affected. The smallest and largest standard errors of the schemes are also reported under each post-change hypothesis based on 2500 repetitions in Monte Carlo simulations.

γ		# local data streams affected								
		1	3	5	8	10	20	30	50	100
	Smallest standard error	0.19	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.00
	Largest standard error	0.40	0.14	0.08	0.05	0.04	0.03	0.02	0.02	0.01
5000	Xie and Siegmund's schemes $T_{XS}(a, p_0)$ in (37)									
	$T_{XS}(a = 53.5, p_0 = 1)$	52.4	18.3	11.1	7.1	5.7	2.9	2.0	1.2	1.0
	$T_{XS}(a = 19.5, p_0 = 0.1)$	31.1	13.4	9.2	6.7	5.7	3.5	2.5	1.8	1.0
	Soft-thresholding Schemes $N_{soft}(a)$ in (13)									
	$N_{soft}(a = 127.86, b_1 = 0)$	75.0	35.4	25.2	18.5	16.0	10.3	8.1	6.1	4.1
	$N_{soft}(a = 84.91, b_1 = 0.5)$	72.1	33.9	24.1	17.7	15.3	10.0	7.9	6.0	4.2
	$N_{soft}(a = 24.01, b_1 = \log(10))$	45.8	22.0	16.4	12.8	11.5	8.5	7.3	6.1	5.0
	$N_{soft}(a = 7.88, b_1 = \log(100))$	29.0	17.2	14.2	12.0	11.2	9.2	8.3	7.3	6.4
5×10^4	Soft-thresholding Schemes $N_{soft}(a)$ in (13)									
	$N_{soft}(a = 136.07, b_1 = 0)$	89.0	39.9	27.9	20.2	17.4	11.1	8.7	6.5	4.4
	$N_{soft}(a = 92.79, b_1 = 0.5)$	85.7	38.2	26.8	19.4	16.7	10.7	8.4	6.3	4.4
	$N_{soft}(a = 29.05, b_1 = \log(10))$	55.1	25.3	18.4	14.1	12.6	9.1	7.8	6.5	5.2
	$N_{soft}(a = 11.11, b_1 = \log(100))$	35.5	19.7	16.0	13.4	12.4	10.0	8.9	7.9	6.8

For each of these six schemes $T(a)$, we first numerically search the threshold a to satisfy the global false alarm constraint γ in (3). Two different values of γ are considered. One is $\gamma = 5000$, so that we can compare with those results from Xie and Siegmund [40]. The other is $\gamma = 5 \times 10^4$ to see the effect of false alarm constraint γ on the detection delays of our proposed schemes. Note that we are unable to numerically find the global threshold a of the Xie and Siegmund scheme for the case of $\gamma = 5 \times 10^4$ in a reasonable time, and thus we will only report the performance of our proposed schemes. Next, for the detection delays of $T(a)$, we consider various post-change hypotheses, and for each post-change hypothesis, we simulate the $\mathbf{E}(T(a))$ when the event occurs at time $\nu = 1$, and use this as an estimate of the detection delay $\mathbf{D}(T(a))$. All simulated values are based on 2500 Monte Carlo runs.

Table 2 summarizes the detection delays in the scenario when the change is instantaneous if a local data stream is affected. For the Xie and Siegmund scheme $T_{XS}(a, p_0)$ in (37), our simulated detection delay results are slightly different from their reported results in their paper, possibly because our simulation is

based on 2500 runs instead of 500 runs in their paper. Note that the Xie and Siegmund schemes $T_{XS}(a, p_0)$ in (37) involve expensive computations, and require the fusion center to have full access to all raw data. Thus it is not surprising that their schemes have smaller detection delays than our proposed soft-thresholding schemes. However, we want to emphasize that the Xie and Siegmund schemes are not scalable and cannot be implemented in the context of distributed monitoring in censoring sensor networks. Meanwhile, our proposed schemes can be easily implemented by parallel computing in a recursive manner at the local sensors level and thus are scalable and also suitable to the censoring sensor network context.

A more reasonable comparison is to compare the results in Table 2 with those in Table 1 which were conducted under the assumption that the post-change mean of each affected local data stream is $\mu = 1$. When at least 5 local data streams are affected, the detection delays of $N_{soft}(a = 24.01, b_1 = \log(10))$ in Table 2 are only $2 \sim 5$ larger than those of $N_{soft}(a = 21.56, b_k = 2.3026)$ in Table 1. Since the schemes in Table 2 are able to detect both positive or negative mean shifts, one may be willing to pay the price of slightly larger detection delays at the given post-change mean $\mu = 1$ so as to effectively detect other local mean shifts, especially the negative shifts. In addition, it is interesting to see from Table 2 that as the false alarm constraint γ increases from 5000 to 5×10^4, the global threshold a of our proposed soft-thresholding schemes $N_{soft}(a, b_1)$ increases moderately for any given censoring parameters b_1, but the detection delays of our proposed soft-thresholding schemes increase only marginally when at least 5 data streams are affected.

All simulations were done on a Windows 8 Laptop with Intel i7-4700MQ CPU 2.40GHz using MATLAB R2013b. For each of these schemes $T(a)$ (i.e., each row of Table 2), the most time consuming part was to search for the global threshold a so that $\mathbf{E}^{(\infty)}(T(a)) \approx \gamma$. When $\gamma = 5000$, it took about 8 minutes to find such a from *a range of values* for our proposed schemes based on 2500 Monte Carlo runs (the time is shorter if our initial guess range of a is closer). Meanwhile, for the Xie and Siegmund scheme, for a *given* global threshold a around 53.5 which was provided in their paper, it took about one and a half hour on average to finish one Monte Carlo simulation run in our laptop. If we did not know $a \approx 53.5$ and wanted to try 10 different values of a's by bisection method based on 2500 Monte Carlo runs for each a, it would have taken about $10 \times 1.5 \times 2500 = 37500$ computer hours for the case of $\gamma = 5000$. When $\gamma = 5 \times 10^4$, it took us about one hour to find the global threshold a for our proposed schemes, but we are unable to numerically implement the Xie and Siegmund schemes since their computational time will be in days for each Monte carlo run. Once the global threshold a is found, it is straightforward to simulate the detection delays in Table 2. When $\gamma = 5000$, our proposed schemes are at least 10 times faster than the Xie and Siegmund schemes. For instance, when exactly one data stream is affected, it took 4.94 seconds to simulate the detection delay of our proposed schemes, whereas it took 41.02 seconds to simulate those of the Xie and Siegmund schemes. Hence, as compared to the Xie and Siegmund schemes, the computational advantage of our proposed schemes is evident.

References

[1] Appadwedula, S., Veeravalli, V. V., and Jones, D. L. (2005). Energy-efficient detection in sensor networks. *IEEE J. Sel. Areas Commun.*, *23*, 693–702.

[2] Banerjee, T. and Veeravalli, V. V., (2015). Data-efficient quickest change detection in sensor networks. *IEEE Trans. Signal Processing*, **63**, 3727–3735. MR3359859

[3] Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Applications.* Englewood Cliffs, Prentice-Hall. MR1210954

[4] Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Sciences*, *16*, 199–231. MR1874152

[5] Candès, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numerica*, *15*, 257–325. MR2269743

[6] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455. MR1311089

[7] Durrett, R. (1996). *Probability: Theory and Examples*. Second edition. Duxbury Press, Belmont, CA. MR1609153

[8] Fan, J. and Lin, S. K. (1998). Test of significance when data are curves. *Journal of American Statistical Association*, **93**, 1007–1021. MR1649196

[9] Fuh, C.D. and Mei, Y. (2015). Quickest change detection and Kullback-Leibler divergence for two-state hidden Markov models. *IEEE Trans. Signal Processing*, **63**, 4866–4878. MR3385842

[10] Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. Springer-Verlag, New York. MR1869112

[11] Gordon, L. and Pollak, M. (1994). An efficient sequential nonparametric scheme for detecting a change of distribution. *Ann. Statist.* **22** 763–804. MR1292540

[12] Kiefer, J. and Sacks, J. (1963). Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* **34** 705–750. MR0150907

[13] Kulldorff, M. (2001). Prospective Time-Periodic Geographic Disease Surveillance Using a Scan Statistic, *J. R. Stat. Soc. Ser. A* **164** 61–72. MR1819022

[14] Lai, T. L. (1995). Sequential change-point detection in quality control and dynamical systems (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 613–658. MR1354072

[15] Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statist. Sinica* **11** 303–408. MR1844531

[16] Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *Ann. Appl. Stat.* **3** 637–662. MR2750676

[17] Liu, K., Mei, Y., and Shi, J. (2015). An adaptive sampling strategy for online high-dimensional process monitoring. *Technometrics* **57** 305–319. MR3384946

[18] Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42** 1897–1908. MR0309251

[19] Lorden, G. and Pollak, M. (2008). Sequential change-point detection procedures that are nearly optimal and computationally simple. *Sequential Analysis* **27** 476-512. MR2460209

[20] Mei, Y. (2005). Information bounds and quickest change detection in decentralized decision systems. *IEEE Trans. Inform. Theory* **51** 2669–2681. MR2246385

[21] Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **97** 419-433. MR2650748

[22] Mei, Y. (2011). Quickest detection in censoring sensor networks. In *IEEE International Symposium on Information Theory (ISIT)*, page 2148–2152, Aug. 2011.

[23] Montgomery, D. C. (1991). *Introduction to Statistical Quality Control* (2nd edition). Wiley, New York.

[24] Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.* **14** 1379–1387. MR0868306

[25] Neyman, J. (1937). Smooth test for goodness-of-fit. *Skand. Aktuarietidskr.* **20** 149–199.

[26] Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115. MR0088850

[27] Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13** 206–227. MR0773162

[28] Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist.* **15** 749–779. MR0888438

[29] Poor, H. V. and Hadjiliadis, O. (2009). *Quickest Detection*. Cambridge Univ. Press, New York, 2009. MR2482527

[30] Rago, C., Willett, P., and Bar-Shalom, Y. (1996). Censoring sensors: A low-communication-rate scheme for distributed detection. *IEEE Trans. Aerosp. Electon. Syst.*, *32*, 554–568.

[31] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics* **8** 411–430. MR0196887

[32] Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. D Van Norstrand, New York. Preprinted by ASQC Quality Press, Wisconsin, 1980.

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

[34] Siegmund, D. (1985): *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York. MR0799155

[35] Tartakovsky, A., Nikiforov, I., and Basseville, M. (2015). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Monographs on Statistics and Applied Probability, 136. CRC Press, Boca Raton, FL. MR3241619

[36] Tartakovsky, A. G., Rozovskiia, B. L., Blazeka, R. B. and Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods (with discussions). *Statistical Methodology* **3** 252–340. MR2240956

[37] Tartakovsky, A. G. and Veeravalli, V. V. (2004). Change-point Detection in Multichannel and Distributed Systems. *Applied Sequential Methodologies*, 339–370, Statist. Textbooks Monogr., 173, Dekker, New York. MR2159163

[38] Tay, W. P., Tsitsiklis, J. N. and Win, M. Z. (2007). Asymptotic performance of a censoring sensor network. *IEEE Trans. Inform. Theory* **53** 4191–4209. MR2446562

[39] Xie, Y., Huang, J., and Willett, R. (2013). Changepoint detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, **7**, 12–27.

[40] Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection. *Ann. Stat.*, **41** 670–692. MR3099117

[41] Veeravalli, V. V. (2001). Decentralized quickest change detection. *IEEE Trans. Inform. Theory* **47** 1657–1665. MR1830119

[42] Wang, Y. and Mei, Y. (2015). Large-Scale multi-stream quickest change detection via shrinkage post-change estimation. *IEEE Trans. Inform. Theory* **61** 6926–6938. MR3430730