Statistica Sinica Preprint No: SS-2015-0316.R2	
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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

Statistica Sinica

SCALABLE SUM-SHRINKAGE SCHEMES FOR DISTRIBUTED MONITORING LARGE-SCALE DATA STREAMS

Kun Liu, Ruizhi Zhang and Yajun Mei

Georgia Institute of Technology

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. The usefulness of our proposed SUM-Shrinkage approach is illustrated in an example of monitoring large-scale independent normally distributed data streams when the local post-change mean shifts are unknown and can be positive or negative.

Key words and phrases: 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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 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 (1931)), moving average control charts, 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)). These procedures not only hold attractive theoretical properties, but also are computationally simple. See, for example, Lorden (1971), Pollak (1985, 1987), Moustakides (1986), Lai (1995, 2001), Kulldorff (2001). For a review, see the books such as Basseville and Nikiforov (1993), Poor and Hadjiliadis (2009), and Tartakovsky et al. (2015).

Research has been limited in the context of monitoring large-scale data streams, especially when the occurring event might affect some, but not

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

all, local data streams. Existing methods include the MAX-scheme (which uses the maximum of local CUSUM statistics as the global statistic, see Tartakovsky et al. (2006)), the SUM-scheme (which uses the sum of local CUSUM statistics as the global statistic, see Mei (2010)), the mixture-schemes proposed in Xie and Siegmund (2013), and the simultaneous-estimation-based schemes in Wang and Mei (2015). While the first two of these schemes are computationally efficient but are generally statistically inefficient unless the number of affected data streams is either very small or very large, the last two schemes enjoy nice statistical properties under general settings, but are computationally infeasible for online monitoring large-scale data streams over a long time period. Our research intends to balance the trade-off between statistical efficiency and computational efficiency when monitoring large-scale data streams.

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, introduced by Rago et al. (1996) and later by Appadwedula et al. (2005) and Tay et al. (2007). 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

final decision is made at a central location, called the fusion center. The key feature of such a network is that while taking observations at the local sensors is 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 monitor the system effectively under the networks resource constraints in the computing power, memory and communications. An 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 be 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 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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

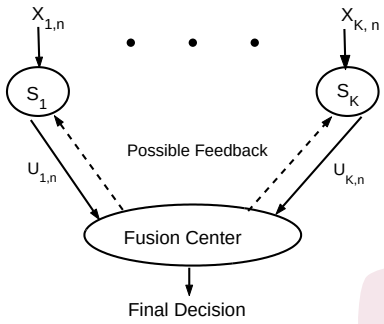

Figure 1: A configuration of censoring sensor networks.

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. 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 is interesting, a charge often made is that one loses much information at the global level by combining local detection procedures, not raw observation themselves, to make a global decision. There are two 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; these methods are known

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

to be inefficient when the number of affected data streams is moderate, see Mei (2010) and Xie and Siegmund (2013).

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 when the number of affected data streams is moderate. Our idea is to generalize the SUM scheme in Mei (2010) by introducing the shrinkage function to local detection statistics in the hope of filtering out those unchanging local data streams. We 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 4 for the comparison of our proposed schemes with those in Xie and Siegmund (2013). The parallel local monitoring approach does allow us to develop scalable schemes, and the loss of statistical efficiency is the price we pay for the computational efficiency. A common 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 (1937), Donoho and Johnstone (1994), Fan and Lin (1998), Candès (2006),

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

and the references there. Our motivation 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 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 Section 4 for the scenario of monitoring large-scale independent normally distributed data when the post-change means of local data streams are unknown.

2. Preliminaries and Background

For a general setting, assume there are K independent data streams in a system.

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

Initially, the system is "in control", but at some *unknown* time ν, an unde-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

sired 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. (2006), Mei (2010), and Xie and Siegmund (2013). Also see Lèvy-Leduc and Roueff (2009) for an application of the unstructured problem to anomaly detection in computer networks. In particular, we focus on the scenario in which 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 are 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 might think, see Xie et al. (2013), and Liu et al. (2015), who monitor the independent residuals from some spatio-temporal models instead of dependent raw data, in applications to solar flare and hot forming processes.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

For the purpose of generalization, we do not specify the kind of local changes these K data streams might 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 (2015) when the local data are dependent from hidden Markov models. The requirements for these $W_{k,n}$'s are that they not only should be able to detect local changes quickly, but also can be computed efficiently for our proposed scheme to be scalable. It can be highly non-trivial to construct such $W_{k,n}$'s in practice, see an example in Section 4.

We 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 (2.1),

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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," the Average Run Length (ARL) to false alarm. The definition of detection delay of the global monitoring scheme is 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. One definition of the detection delay of T is the "worst case" delay given in Lorden (1971),

$$\overline{\mathbf{E}}(T) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbf{E}^{(\nu)} \Big((T - \nu + 1)^+ \Big| \mathcal{F}_{\nu-1} \Big). \tag{2.2}$$

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 up to 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 (2.2) subject to the global false alarm

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

constraint

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

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

3. Our Proposed SUM-Shrinkage Methodology

Now we turn to our proposed methodology. Assume, for a moment, that the local detection statistics $W_{k,n}$'s (in the log-likelihood scale) have been constructed for each local k-th data streams at time n. We suggest using the global monitoring statistic of the general "SUM-shrinkage" form

$$G_n = \sum_{k=1}^{K} h_k(W_{k,n}), \tag{3.1}$$

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

$$N_G(a) = \inf\{n \geq 1 : G_n \geq a\}. \tag{3.2}$$

Our proposed $N_G(a)$ in (3.2) has two key components in its global monitoring statistic G_n in (3.1): the local detection statistic $W_{k,n}$'s; the shrinkage transformations $h_k(\cdot)$'s. 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 (3.1) play the role of dimension reduction by automatically filtering out non-changing local data streams and focusing on those local data streams that appear to be affected by the occurring event.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

Our proposed "SUM-shrinkage" methodology in (3.1)-(3.2) has a broad range of applications. For instance, Mei (2011) applied the idea to develop an efficient communication policy between sensors and fusion center in the context of censoring sensor networks. Depending on which kind of local models or local changes are of interest, local detection statistic $W_{k,n}$ can be defined for such dependent observations as those from the recursive schemes in Fuh and Mei (2015) for hidden Markov models, or those from the non-parametric rank-based detection schemes in Gordon and Pollak (1994). Little information seems to be lost if we do not observe those local data streams with small values of the $W_{k,n}$ since they make limited contributions in the global monitoring statistic G_n in (3.1). This motivated Liu et al. (2015) 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 can occur in a manufacturing process with K possible stages but only r sensors are available for monitoring. In such a problem, the order-thresholding transformation at (3.5) can be combined with missing data techniques not only to construct the global monitoring statistic G_n in (3.1) 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 (2015) essentially tackle the similar problem

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

of missing data, but using the hard-thresholding transformation at (3.3).

Subsection 3.1 contains three choices of shrinkage functions h_k at (3.1), and Subsection 3.2 includes some general properties of $N_G(a)$ that are related to the global false alarm constraint in (2.3). Subsection 3.3 discusses how to choose the tuning parameters in the shrinkage functions h_k in (3.1) when the local data streams are homogeneous.

3.1. Shrinkage Transformation

Evidently a suitable choice of the h_k in the SUM-shrinkage monitoring statistic G_n in (3.1) depends on the assumptions and contexts of applications. As an illustration, we list three shrinkage transformations.

- Hard-thresholding: $h(x) = x\mathbf{1}\{x \geq b\}$ for some constant b. (3.3)
- Soft-thresholding: $h(x) = \max\{x - b, 0\}$ for some constant b. (3.4)
- Order-thresholding: $h(x) = x\mathbf{1}\{x \geq w_{(r)}\}$, where $w_{(r)}$ is the r-th largest statistic of $w_1, \cdots, w_K$. (3.5)

There are many other shrinkage functions, such as $h(x) = \exp(bx)$. By semi-Bayesian arguments, the transformation $h(x) = \log(1-p_0+p_0 \exp(\max(0, x)))$ was proposed by Xie and Siegmund (2013) in a completely different context.

To better understand the shrinkage transformations in (3.3)-(3.5), we motivate them from the communication efficiency viewpoint, first presented

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

in Mei (2011) in the context of the 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 $W_{k,n}$ that are large. Specifically, at time n, the message from the sensor to the fusion center is given by

$$U_{k,n} = \begin{cases} W_{k,n}, & \text{if } W_{k,n} \geq b_k \\ \text{NULL}, & \text{if } W_{k,n} < b_k \end{cases}, \tag{3.6}$$

where $b_k \geq 0$ is the local censoring parameter at the k-th sensor (or data stream). In practice, the message "NULL" could represent that the sensor is silent.

After receiving the local sensor messages $U_{k,n}$ in (3.6), the fusion center combines them suitably to make a global decision. There are many approaches to doing so. Two schemes are based on the summation of all sensor messages $U_{k,n}$, depending on how to interpret the "NULL" values. If we treat the "NULL" values as the lower limit 0, then the fusion center raises a global alarm at time

$$\begin{aligned} N_{hard}(a) &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} U_{k,n} \geq a\Big\} \\ &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} W_{k,n}\mathbf{1}\{W_{k,n} \geq b_k\} \geq a\Big\}. \end{aligned} \tag{3.7}$$

This scheme is referred as the hard-thresholding, since it is a special case of

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

the global statistic in (3.1) when the shrinkage functions h_k are the hard-thresholding transformation in (3.3).

If we treat the "NULL" values as the upper limit b_k, then the fusion center computes 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 (3.4). We can call this a soft-thresholding scheme when it raises an alarm at time

$$N_{soft}(a) \;=\; \inf\Big\{n \geq 1 : \sum_{k=1}^{K} \max\{W_{k,n} - b_k, 0\} \geq a\Big\}. \qquad (3.8)$$

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 (3.2) with the soft-thresholding transformation in (3.4).

A third approach occurs when the fusion center has prior knowledge that (at most) r out of K data streams will be affected by the occurring event. Such 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 (3.3) and the order-thresholding transformation in (3.5), and it yields a global scheme for which the stopping

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

time is

$$N_{comb,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} U_{(k),n} \geq a\Big\}. \tag{3.9}$$

A special case of $N_{comb,r}(a)$ in (3.9) has the order-thresholding transformation in (3.5) applied directly to the local detection statistics $W_{k,n}$ themselves. Specifically, we order the K local CUSUM statistics $W_{1,n}, \ldots, W_{K,n}$ as $W_{(1),n} \geq W_{(2),n} \geq \ldots \geq W_{(K),n}$. Then the order-thresholding scheme is defined by the stopping time

$$N_{order,r}(a) = \inf\Big\{n \geq 1 : \sum_{k=1}^{r} W_{(k),n} \geq a\Big\}, \tag{3.10}$$

which corresponds to the order-thresholding transformation in (3.5).

Based on our experience, the soft-thresholding transformation, as a continuous function, often yields smaller detection delays than the hard-thresholding transformation, a discontinuous function, in finite-sample Monte Carlo simulations. The soft- and order- thresholding transformations have comparable finite-sample performances, but the soft-thresholding transformation is computationally and theoretically simpler. We use the soft-thresholding transformation in (3.4) as a concrete demonstration, when needed.

For the soft-thresholding scheme, $N_{soft}(a)$ in (3.8), statistical intuition is a little more complicated; we provide a semi-Bayesian interpretation of

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

why it works. At a given time n, let Z_k be the indicator of whether the distribution of the k-th local data stream changes for $k = 1, \ldots, K$. Suppose each local data stream has a prior probability π_k of being affected by the event, and 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, $W_{k,n}$ representing the evidence of possible change (in logarithm scale) and applicable only when $Z_k = 1$. 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 and the observed data $X_{k,n}$ 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 is

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

and the generalized log-likelihood ratio is

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

exactly the form of the soft-thresholding scheme $N_{soft}(a)$ in (3.8), with $b_k = \log((1 - \pi_k)/\pi_k)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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

Given the choices of the local detection statistics $W_{k,n}$ and the shrinkage transformation $h_k(\cdot)$, an important question is how to determine the global threshold a in (3.2) so that the proposed SUM-shrinkage scheme satisfies the global false alarm constraint on γ in (2.3). This requires one to accurately characterize the relationship between the threshold a and the ARL to the false alarm $\mathbf{E}^{(\infty)}(N_G(a))$.

As the global monitoring statistic G_n in (3.1) is the sum of K (independent) random variables, one would expect that the Central Limited Theorem (CLT) would be useful when the shrinkage transformation keeps most non-zero values, e.g., the hard-thresholding or soft-thresholding transformations in (3.3) or (3.4) when the censoring parameters b's are not large, whereas the compound Poisson process would be needed when the shrinkage transformation keeps only few non-zero values, e.g., the order-thresholding transformation in (3.5) with a not so large r value. Rigorous proofs are beyond our scope and will be investigated elsewhere.

Below we will use Chebyshev's inequalities to provide a crude relationship between the threshold a and the ARL to the false alarm $\mathbf{E}^{(\infty)}(N_G(a))$. Assume that under the pre-change hypothesis $\mathbf{P}^{(\infty)}$, for each k, the shrinkage transformation of local detection statistics, $h_k(W_{k,n})$, converge to their

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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 $h_k(W_{k,n})$, and has a well-defined log-moment generating function

$$\psi_k(\theta) = \log \mathbf{E}^{(\infty)} \exp(\theta H_k^*) \tag{3.11}$$

for some $\theta \geq 0$.

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

$$\mathbf{E}^{(\infty)}(N_G(a)) \geq \frac{1}{4} \exp\Big(\theta a - \sum_{k=1}^{K} \psi_k(\theta)\Big) \tag{3.12}$$

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

$$a = \inf_{\theta \in \Theta} \left(\frac{1}{\theta} \big(\log(4\gamma) + \sum_{k=1}^{K} \psi_k(\theta) \big) \right) \tag{3.13}$$

guarantees that $N_G(a)$ in (3.2) satisfies the global false alarm constraint γ in (2.3).

Proof: Relation (3.13) follows directly from (3.12), and it suffices to show that (3.12) holds for any $\theta \in \Theta$. By the definition of $N_G(a)$ in (3.2) and the use of Chebyshev's inequality twice, once to $N_G(a) \geq 0$ and the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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 $h_k(W_{k,n})$, and the last equation uses the assumption that these K data streams are independent across different data streams. 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 (3.12). □

The results in Theorem 1 are non-asymptotic, and hold for any K and γ. To demonstrate their usefulness, 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 (3.4). We suppress the script k and derive the log-moment generating function $\psi(\theta)$ in (3.11) for the soft-thresholding transformation

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

$h(W_n) = \max(W_n - b, 0)$ for large b. We further assume that, as $n \to \infty$, the local detection statistic W_n converges to an asymptotically exponentially distributed variable W^* under the pre-change hypothesis,

$$\mathbf{P}^{(\infty)}(W^* > x) \approx \lambda e^{-x}, \tag{3.14}$$

for some constant $\lambda > 0$. A non-asymptotic result is often true for many local detection statistic W_n such as CUSUM: for *any* $x > 0$,

$$\mathbf{P}^{(\infty)}(W^* > x) \leq e^{-x}, \tag{3.15}$$

see Appendix 2 of Siegmund (1985). Under (3.14), we have $\mathbf{P}^{(\infty)}(W^* \leq b) = 1 - \lambda e^{-b}$ for large b. Combining the definition of $\psi(\theta)$ in (3.11) with the fact that $H^* = 0$ whenever $W^* \leq b$ yields that

$$\begin{aligned} \psi(\theta) &= \log \mathbf{E}^{(\infty)} \exp(\theta H^*) = \log[\mathbf{P}^{(\infty)}(W^* \leq b) + \int_b^\infty e^{\theta(x-b)} \lambda e^{-x} dx] \\ &= \log\left(1 + \frac{\theta \lambda e^{-b}}{1-\theta}\right). \end{aligned} \tag{3.16}$$

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 (3.13) is

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

As $A/\theta + B/(1-\theta)$ has a minimum value $(\sqrt{A}+\sqrt{B})^2$ over $0 \le \theta \le 1$ for any $A, B > 0$, (3.13) in Theorem 1 gives

$$a \approx \left(\sqrt{\log(4\gamma)} + \sqrt{K\lambda e^{-b}}\right)^2. \tag{3.17}$$

In(3.17) we see the challenges of monitoring large-scale data streams: the asymptotic expression of a in (3.17) depends on the asymptotic relationship between $\log(\gamma)$ and $K\lambda e^{-b}$. When $\log(\gamma) >> K$, we have the classical result on the threshold of $a = (1 + o(1)) \log(\gamma)$, see Lorden (1971). When $K\lambda e^{-b} >> \log(\gamma)$, we have

$$a \approx K\lambda e^{-b} + 2\sqrt{K\lambda e^{-b}}\sqrt{\log \gamma}. \tag{3.18}$$

This suggests that $K\lambda e^{-b}$ plays a dominant role to determine the threshold a for $N_G(a)$ to satisfy the false alarm constraint γ in (2.3) when b is large and $Ke^{-b} >> \log(\gamma)$.

3.3. The Choice of Censoring Parameters

In this subsection, we discuss the optimal choice of the censoring parameters b_k in (3.6). For illustration and simplicity we consider the homogeneous case, $b_k \equiv b$, when local data streams are identically distributed for different k, e.g., relations (3.14), (3.15), and $\psi_k(\theta) \equiv \psi(\theta)$ in (3.16) hold for all $k = 1, \cdots, K$. We provide two optimal choices of the censoring parameter b for the soft-thresholding scheme $N_{soft}(a)$ in (3.8): one from the communi-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

cation efficiency aspect, and the other from the statistical efficiency aspect. It turns out that they are closely related.

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

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

where the second-to-last inequality follows from (3.15). Thus a choice of

$$b_{opt,1} = \log(\eta^{-1}) \tag{3.19}$$

will guarantee that on average, at most $100\eta\%$ of K sensors transmit messages at any given time when no event occurs. When η is small, one can use the refined asymptotic approximation (3.14), instead of the non-asymptotic bound (3.15), in the above arguments. Then the $b_{opt,1}$ can further improved as $b^*_{opt,1} = \log(\lambda/\eta)$ under the communication constraint.

Next, we 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, when the global threshold value a is *given*, our

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

proposed scheme $N_{soft}(a)$ in (3.8) is increasing as a function of the censoring parameter $b_k \equiv b$, and 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 (2.3), different global threshold values a are needed for these schemes with different b, and thus it is natural to find the censoring parameter b that yields the smallest detection delay $\overline{\mathbf{E}}(T)$ in (2.2).

We 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 general and holds for many local detection statistics including CUSUM, see Lorden (1971). Then the detection delay of the soft-thresholding scheme $N_{soft}(a)$ in (3.8) is bounded above by

$$(1 + o(1))\frac{1}{I}\left(b + \frac{a}{w_0}\right). \tag{3.20}$$

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 (3.20) 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 (2005) when the $W_{k,n}$ are local CUSUM statistics.

If we keep only on the first-order major term of a in (3.18), plugging it

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

into (3.20) yields that the detection delay of the soft-thresholding scheme $N_{soft}(a)$ in (3.8) (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 $K\lambda e^{-b} = w_0$, so the optimal b value (up to first-order) is given by

$$b_{opt,2} = \log \frac{\lambda K}{w_0}, \tag{3.21}$$

where $\lambda > 0$ is the constant in (3.14) that only depends on the asymptotic properties of the $W_{k,n}$.

When we have prior knowledge that w_0 local data streams are affected but we do not know which ones, 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 (3.21) when the fraction of affected data stream $w_0/K \to 0$.

A direct comparison of (3.19) and (3.21) suggests that the two optimal b values are asymptotically equivalent if we set $\eta = w_0/K$. Moreover, by (3.6) and (3.14), when there are no changes, the average fraction of transmitting

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

sensors at any time step n is

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbf{P}^{(\infty)}(U_{k,n}\neq \text{NULL}) &= \mathbf{P}^{(\infty)}(U_{k,n}\neq \text{NULL}) = \mathbf{P}^{(\infty)}(W_{k,n}\geq b_{opt,2}) \\ &= \lambda e^{-b_{opt,2}} = w_0/K. \end{aligned} \tag{3.22}$$

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 is best to design the schemes that on average allow w_0 out of K local data streams to transmit local detection statistics to the fusion center when no change event occurs (and possibly more than w_0 data streams when a change occurs). Due to this equivalence, in our simulations, the censoring parameter b is chosen based on (3.19), which is non-asymptotic and easier to compute.

4. An Example: Unknown Post-Change Normal Means

Suppose that we are monitoring K independent normally distributed data streams $X_{k,n}$ in (2.1). Initially, the data $X_{k,n}$ 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. We do not know which subset of local data streams are affected, and here another new challenge is that we do not know the values of the post-change means μ_k's when affected. We want to develop a

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

system-wise online monitoring scheme that can detect the change as soon as possible, subject to the global false alarm constraint γ in (2.3).

Xie and Siegmund (2013) investigated 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 scheme they proposed was motivated from a semi-Bayesian approach; it is defined by

$$T_{XS}(a, p_0) = \inf\left\{ n \geq 1 : \max_{0 \leq i < n} \sum_{k=1}^{K} \log(1 - p_0 + p_0 \exp\left[\left(U^+_{k,n,i}\right)^2/2\right] \geq a \right\}, \quad (4.1)$$

where $U^+_{k,n,i} = \max\left(0, \sum_{j=i+1}^{n} X_{k,j}\right)/\sqrt{n-i}$ for all $0 \leq i < n$ and $1 \leq k \leq K$. One can also simplify the memory requirement by keeping a large window of the most recent observations. Wang and Mei (2015) proposed a global Shiryaev-Robert procedure by simultaneously estimating all K unknown post-change means μ_k via shrinkage estimation. These schemes are not scalable, and not suitable in the context of censoring sensor networks in Figure 1. 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, and not feasible for large K. Here we illustrate how

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

to tackle this problem based on our proposed SUM-shrinkage statistics in (3.1). We need 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. if the local detection statistics $W_{k,n}$s are defined, we can use any shrinkage transformation to develop a global monitoring scheme.

In this section, we consider the soft-thresholding scheme $N_{soft}(a)$ in (3.8). For simplicity, we assume that all censoring parameters b_k in (3.8) are the same, $b_k \equiv b_1$ for some constant $b_1 > 0$. Our focus is how to construct the local detection statistics $W_{k,n}$'s suitably.

Subsection 4.1 provides an overview of our proposed soft-thresholding scheme in (3.8) that only uses a fixed number of $6K$ registers to store all past information and involves $O(K)$ computations at each given time step n. Simulation results are summarized in Subsection 4.2.

4.1. Our Proposed Local Detection Statistics $W_{k,n}$

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 of Lorden and Pollak (2008) from one-sided to two-sided. As detecting negative local mean shift of the $X_{k,n}$ is equivalent to detecting positive local mean shift of the $-X_{k,n}$, we propose the local detection statistic for each

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

local data stream at time n,

$$W_{k,n} = \max\,(W_{k,n}^{(1)}, W_{k,n}^{(2)}). \tag{4.2}$$

Here $W_{k,n}^{(1)}$ and $W_{k,n}^{(2)}$ are the local detection statistics of Lorden and Pollak (2008) for detecting positive and negative mean shifts, respectively:

$$\begin{aligned} W_{k,n}^{(1)} &= \max\left(W_{k,n-1}^{(1)} + \hat{\mu}_{k,n}^{(1)} X_{k,n} - \frac{1}{2}(\hat{\mu}_{k,n}^{(1)})^2, 0\right), \\ W_{k,n}^{(2)} &= \max\left(W_{k,n-1}^{(2)} + \hat{\mu}_{k,n}^{(2)} X_{k,n} - \frac{1}{2}(\hat{\mu}_{k,n}^{(2)})^2, 0\right), \end{aligned} \tag{4.3}$$

where

$$\hat{\mu}_{k,n}^{(1)} = \max\left(\rho, \frac{s + S_{k,n}^{(1)}}{t + T_{k,n}^{(1)}}\right) > 0, \qquad \hat{\mu}_{k,n}^{(2)} = \min\left(-\rho, \frac{-s + S_{k,n}^{(2)}}{t + T_{k,n}^{(2)}}\right) < 0, \tag{4.4}$$

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

$$\begin{pmatrix} S_{k,n}^{(j)} \\ T_{k,n}^{(j)} \end{pmatrix} = \begin{cases} \begin{pmatrix} S_{k,n-1}^{(j)} + X_{k,n-1} \\ T_{k,n-1}^{(j)} + 1 \end{pmatrix} & \text{if } W_{k,n-1}^{(j)} > 0 \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } W_{k,n-1}^{(j)} = 0 \end{cases} \tag{4.5}$$

Here the $\hat{\mu}_{k,n}^{(1)}$ and $\hat{\mu}_{k,n}^{(2)}$ in (4.4) are the estimates of the post-change mean when restricted to the positive and negative values, respectively, under the assumption that $|\mu| \geq \rho$. The two-sided local detection statistic $W_{k,n}$ in (4.2) is always nonnegative for any k at any time step n, and it is large

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

when there is a local mean shift no matter whether such mean shift is positive or negative.

The proposed soft-thresholding scheme $N_{soft}(a)$ in (3.8) can be easily implemented in the censoring sensor network context by parallel computing the K local detection statistics $W_{k,n}$'s recursively through (4.2)-(4.5) at the local sensor levels. We can use $6K$ 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 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. Our scheme can be implemented in censoring sensor networks 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

4.2. Simulation Results

In this subsection, we report the numerical simulation results of the soft-thresholding scheme $N_{soft}(a)$ in (3.8) when the local detection statistics $W_{k,n}$'s are defined recursively through (4.2)-(4.5), and the censoring parameters $b_k \equiv b_1$ for all k. For the purpose of comparison, we follow Xie and Siegmund Xie and Siegmund (2013) 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 considered six schemes: the Xie and Siegmund schemes $T_{XS}(a, p_0)$ in (4.1) with $p_0 = 1$ and 0.1, and four schemes employed our proposed soft-thresholding schemes $N_{soft}(a)$ in (3.8) with censoring parameters: $b_1 = 0, 0.5, \log(10), \log(100)$. 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 $W_{k,n}$ values to the global monitoring statistic G_n when there are no changes. When computing the local detection statistics $W_{k,n}$ in (4.2), we set $\rho = 0.25, t = 4$ and $s = 1$ as in Lorden and Pollak (2008).

For each of these schemes, we first numerically searched the threshold a to satisfy the global false alarm constraint γ in (2.3). Two values of γ were

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

considered: $\gamma = 5000$, so that we can compare with those results from Xie and Siegmund (2013); $\gamma = 5 \times 10^4$ to see the effect of false alarm constraint γ on the detection delays of our schemes. We are unable to numerically find the global threshold a of the Xie and Siegmund scheme for the case $\gamma = 5 \times 10^4$ in a reasonable time, and there we only report the performance of our proposed schemes. For the detection delays, we considered various post-change hypotheses and, for each post-change hypothesis, we simulated the $\mathbf{E}(T(a))$ when the event occurs at time $\nu = 1$, and used this as an estimate of the detection delay $\mathbf{D}(T(a))$. All simulated values were based on 2500 Monte Carlo runs.

Table 1 summarizes detection delays when the change is instantaneous if a local data stream is affected. For the Xie and Siegmund scheme $T_{XS}(a, p_0)$ in (4.1), our simulated detection delay results are slightly different from their reported results in their paper, possibly because our simulation was based on 2500 runs instead of the 500 runs in their paper. The Xie and Siegmund schemes $T_{XS}(a, p_0)$ in (4.1) 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 schemes. The Xie and Siegmund schemes are not scalable and cannot be implemented in the context of distributed monitoring in censoring sensor

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

Table 1: 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 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 (4.1)									
	$T_{XS}(a = 53.5, p_0 = 1)$	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 (3.8)									
	$N_{soft}(a = 127.86, b_1 = 0)$	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 (3.8)									
	$N_{soft}(a = 136.07, b_1 = 0)$	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

networks. Our proposed schemes can be easily implemented by parallel computing in a recursive manner at the local sensors level and thus are

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

scalable.

All simulations were done on a Windows 8 Laptop with Intel i7-4700MQ CPU 2.40GHz using MATLAB R2013b. For each row of Table 1, 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). For the Xie and Siegmund scheme, for the *given* global threshold a around 53.5 provided in their paper, it took about one and a half hours on average to finish one Monte Carlo simulation run. 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. Once the global threshold a is found, it is straightforward to simulate the detection delays in Table 1. When $\gamma = 5000$, our proposed schemes are at least 10 times faster than those of Xie and Siegmund. For instance, when exactly one data stream is affected, it took 4.94 seconds to simulate the detection delay of our proposed schemes, and 41.02 seconds to simulate theirs. The

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

computational advantage of our proposed schemes is evident.

Acknowledgements

This research was supported in part by the NSF grants DMS-0954704 and CMMI-1362876. The authors are grateful to two anonymous reviewers for the detailed and constructive comments that greatly improved the quality and presentation of the article.

References

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

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

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

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

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

Durrett, R. (1996). *Probability: Theory and Examples.* Second edition. Duxbury Press, Belmont,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

CA. MR1609153

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

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

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

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

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

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

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

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

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

dimensional network traffic data. *Ann. Appl. Stat.* **3**, 637–662. MR2750676

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

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

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

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

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

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

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

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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

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

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

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.

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

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

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0316

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

Tartakovsky, A. G. and Veeravalli, V. V. (2004). Change-point Detection in Multichannel and Distributed Systems. *Applied Sequential Methodologies* **17**, 339–370, Statist. Textbooks Monogr., 173, Dekker, New York. MR2159163

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

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.

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

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

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive NW, Atlanta, GA 30332-0205, U.S.A.

E-mail: {kliu80, zrz123, ymei3}@gatech.edu

(Received September 2015; accepted January 2017)