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.	

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],

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

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, 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 stopping time T that minimizes the detection delay $\overline{\mathbf{E}}(T)$ in (2) subject to the global false alarm constraint

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

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

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

($= \infty$ 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

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

($= \infty$ 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).