Efficient global monitoring statistics for high-dimensional data

Jun Li

Department of Statistics, University of California, Riverside, California

Correspondence
Jun Li, Department of Statistics, University of California, Riverside, Riverside, CA 92521, U.S.A.
Email: jun.li@ucr.edu

Abstract
Global monitoring statistics play an important role in developing efficient monitoring schemes for high-dimensional data. A number of global monitoring statistics have been proposed in the literature. However, most of them only work for certain types of abnormal scenarios under specific model assumptions. How to develop global monitoring statistics that are powerful for any abnormal scenarios under flexible model assumptions is a long-standing problem in the statistical process monitoring field. To provide a potential solution to this problem, we propose a novel class of global monitoring statistics. Our proposed global monitoring statistics are easy to calculate and can work under flexible model assumptions since they can be built on any local monitoring statistic that is suitable for monitoring a single data stream. Our simulation studies show that the proposed global monitoring statistics perform well across a broad range of settings and compare favorably with existing methods.

Keywords
CUSUM, process change detection, quantile vs quantile, statistical process control

1 INTRODUCTION

Advanced manufacturing and data acquisition technologies have made the gathering of high-dimensional data possible in many fields. The demand for efficient online monitoring tools for such data has never been greater. Depending on the purpose, two types of monitoring schemes are needed.

The first type is for applications in which changes in any of data streams indicate the same abnormality in the whole system and require a single, uniform corrective action. As a result, those applications do not require the identification of abnormal data streams. For example, in the environmental monitoring application, hundreds or thousands of sensors are usually deployed to monitor certain environmental factors. The abnormality of data from any of those sensors will indicate a general abnormality in those environmental factors. In many cases, it is not of interest to identify which sensors have caused the alarm. Since it is not required to identify abnormal data streams for this type of monitoring scheme, a popular approach for developing monitoring schemes of this type is to use a single global monitoring statistic to track all data streams jointly. As a result, a global monitoring statistic that is powerful for detecting any abnormal scenario is the key for developing any efficient monitoring scheme of this type.

The second type of monitoring scheme is for applications in which changes in different data streams indicate different problems in the system, requiring unique corrective actions. As a result, this type of monitoring scheme needs to identify which data stream is experiencing abnormal activities. For example, in the network traffic surveillance application, the network traffic data from different data streams are associated with different IP addresses. When some abnormality in the data occurs, identifying which data stream or IP address has caused the problem is very important, as doing so helps...
pinpoint the cause and guides future corrective actions. As pointed out in Li,1 the following two-stage strategy can be very effective to develop this type of monitoring scheme. In the first stage, a global monitoring statistic is used to decide whether there is any abnormal data stream. If this is the case, the second stage is activated, and a local monitoring statistic is used to decide which data streams are abnormal. As shown in Li,1 this two-stage strategy has better performance than the one-stage strategy. On the basis of this two-stage strategy, it is evident that what global monitoring statistic to use in the first stage becomes critical, since it will ultimately affect the effectiveness of the resulting monitoring scheme in terms of how quickly it will raise an alarm when some of the data streams start experiencing abnormal activities.

From the above discussions, it is clear that an efficient global monitoring statistic is important for developing efficient monitoring schemes of both types. How to construct efficient global monitoring statistics for high-dimensional data streams has been an active research topic in the statistical process monitoring field. A number of global monitoring statistics have been proposed. However, most of them only work for certain types of abnormal scenarios under specific model assumptions. For example, under the assumption that each data stream follows a normal distribution and using the cumulative sum (CUSUM) statistic as the local monitoring statistic for each data stream, Tartakovsky et al2 and Mei3 proposed using the maximum and sum of those CUSUM statistics, respectively, as the global monitoring statistic. It has been shown that the sum of those CUSUM statistics is more effective than the maximum when a moderate or large number of data streams are abnormal and vice versa when only a few data streams are abnormal. In practice, it is usually unknown in advance how many data streams will be abnormal. Therefore, neither the maximum or the sum of those CUSUM statistics as the global monitoring statistic guarantees robust performance. Xie and Siegmund4 recognized this limitation and proposed a global monitoring statistic derived from the likelihood function of a normal-mixture model. Besides being computationally intensive, their approach also needs to pre-specify the percentage of abnormal data streams. If the percentage is misspecified, their approach will sacrifice some power. Zou et al5 proposed some alternative way to combine the CUSUM statistics from all data streams to produce a single global monitoring statistic. This approach does not require prior knowledge about how many data streams are abnormal and performs well compared with the aforementioned methods across different abnormal scenarios. However, it is not clear how to extend their approach to develop some computationally efficient global monitoring statistics for situations where the local monitoring statistics are not CUSUM statistics. Recently, Liu et al6 proposed several global monitoring statistics via the so-called SUM-shrinkage technique. The SUM-shrinkage technique they use is essentially a thresholding method. Instead of taking the sum of all the local monitoring statistics as proposed in Mei,3 their new global monitoring statistics take the sum of only those that exceed some pre-specified threshold. As expected, the detection power of their monitoring schemes depends on the pre-specified threshold. In their paper, they studied several choices of threshold. However, each one works well only for certain types of abnormal scenarios. Similar to other thresholding-based methods, it is impossible to find a single threshold in their proposed global monitoring statistics that will work well for different types of abnormal scenarios.

In addition to the limitations described above, most of the existing approaches were developed under the normality assumption. When this normality assumption does not hold, none of the above approaches will perform as expected. So all this indicates the need to develop flexible global monitoring statistics that can work with any type of data and be effective for different types of abnormal scenarios. To meet this need, we propose a novel class of global monitoring statistics by making use of the order statistics of the local monitoring statistics. The unique use of the order statistics makes our proposed monitoring statistics efficient for different abnormal scenarios, as shown in our simulation studies. Our proposed global monitoring statistics are easy to calculate and can work under flexible model assumptions since they can be built on any local monitoring statistic that is suitable for monitoring a single data stream. There is vast literature on how to monitor a single data stream. Therefore, the proposed class of global monitoring statistics can easily benefit from this rich literature.

The rest of the paper is organized as follows. In Section 2, we propose a general class of global monitoring statistics and show that the global monitoring statistic studied in Zou et al5 can be considered as a special case of our proposed global monitoring statistic. In Section 3, we give three examples of our proposed global monitoring statistic from the general class and evaluate their performance through simulation studies. Finally, we provide some concluding remarks in Section 4.

2 | METHODOLOGY

2.1 | Notation

The setup for our high-dimensional data monitoring problem is the following. There are \( m \) data streams in the system. We denote the observation from the \( i \)th data stream at time \( t \) by \( X_{i,t} \), \( i = 1, \ldots, m \), \( t = 1, 2, \ldots \). Since a time series model can
be used to decorrelate the temporal correlation within each data stream and a spatial model can be used to decorrelate the spatial correlation between data streams before applying monitoring schemes, following most papers on the topic, we assume that the \( X_{i,t} \) are independent both within and between data streams. When the system is in control (IC), the underlying distribution of \( \{ X_{i,1}, X_{i,2}, \ldots \} (i = 1, \ldots, m) \) is called the IC distribution, denoted by \( F_{0,t} \). Following this setup, at a given time \( t \), we observe \( X_{1,t}, X_{2,t}, \ldots, X_{i,t}, i = 1, \ldots, m \). The task of our online monitoring scheme at time \( t \) is to determine if the distribution of \( X_{1,t}, X_{2,t}, \ldots, X_{i,t} \) is the same as \( F_{0,t} \) for all \( i = 1, 2, \ldots, m \).

The above task can be carried out by tracking a global monitoring statistic \( G_t \), which contains information collected from all data streams up to time \( t \). If \( G_t \) is within the preset control limit, we will declare that all data streams are IC and continue monitoring. If \( G_t \) exceeds the control limit, we will raise an alarm suggesting that some of the data streams are out of control (OC). In the following, we propose a novel class of global monitoring statistics that can work with any type of data and be effective for different types of OC scenarios.

### 2.2 Proposed global monitoring statistics

Because the change point can happen at different times for different data streams, a popular approach in the literature for developing the global monitoring statistic \( G_t \) is to first choose an appropriate local monitoring statistic for tracking each data stream and then combine those local monitoring statistics in a way that produces a single global monitoring statistic. We will follow this approach. More specifically, let \( W_{i,t} \) be the local monitoring statistic for the \( i \)th data stream at time \( t \) that summarizes the evidence regarding a possible local change based on the observations, \( X_{i,1}, \ldots, X_{i,t} \). Without loss of generality, we assume that a larger \( W_{i,t} \) indicates a higher probability of the \( i \)th data stream being OC. Although our proposed global monitoring statistic \( G_t \) can work with any choice of \( W_{i,t} \), in order for \( G_t \) to be efficient for detecting changes in any data stream, the \( W_{i,t} \) should be chosen to be efficient for detecting local changes.

Note that at any time \( t \), we have calculated \( W_{1,t}, \ldots, W_{m,t} \). Without loss of generality, we assume that the \( W_{i,t} \) are independent and identically distributed when the system is IC. As mentioned in Section 1, Liu et al recently proposed a SUM-shrinkage approach to construct the global monitoring statistic based on \( W_{1,t}, \ldots, W_{m,t} \). In their approach, \( W_{1,t}, \ldots, W_{m,t} \) are compared with some pre-specified threshold, and only those that exceed the threshold are used to construct the global test statistic. However, similar to all the other thresholding methods, it is impossible to choose a threshold in advance that works well for all OC scenarios.

Instead of comparing \( W_{1,t}, \ldots, W_{m,t} \) with some pre-specified threshold, we propose to compare their order statistics with their respective expected values when the system is IC. More specifically, let \( W_{(1),t} \leq W_{(2),t} \leq \ldots \leq W_{(m),t} \) be the order statistics of \( W_{1,t}, \ldots, W_{m,t} \). Note that \( W_{(i),t} \) can be also considered as the observed \( (i - 3/4)/(m - 1/2) \) quantile of the underlying distribution of \( W_{i,t} \). Here, \( (i - 3/4)/(m - 1/2) \) is the common continuity correction of \( i/m \). Let \( q_{(i),t} \) denote the expected \( (i - 3/4)/(m - 1/2) \) quantile of the IC distribution of \( W_{i,t} \). Then, \( q_{(i),t} \) can be considered as the expected value of \( W_{(i),t} \). A natural statistic that summarizes the differences between \( W_{(i),t} \) and their respective expected values \( q_{(i),t} \) is simply \( \sum_{i=1}^{m} (W_{(i),t} - q_{(i),t})^2 \). Since a larger \( W_{(i),t} \) indicates a higher probability of the \( i \)th data stream being OC, only when \( W_{(i),t} \) is larger than its expected value \( q_{(i),t} \) may indicate abnormality in the system. Therefore, we only include the difference when \( W_{(i),t} \) is larger than its expected value \( q_{(i),t} \), in our global monitoring statistic, and the new class of global monitoring statistics we propose is

\[
G_t = \sum_{i=1}^{m} (W_{(i),t} - q_{(i),t})^2 I_{[w_{(i),t}>q_{(i),t}]},
\]

where \( I_{[A]} \) is the indicator function and takes 1 if \( A \) is true and 0 otherwise. Then, our proposed monitoring scheme is to plot \( G_t \) over the time \( t \), and it raises an alarm if \( G_t > h \), where \( h \) is the control limit predetermined by the desired IC average run length (denoted by \( ARL_0 \)).

At first glance, our proposed global monitoring statistic \( G_t \) is the sum-type statistic, so it is expected to be effective when a moderate or large number of data streams are abnormal. As shown in our simulation studies in Section 3, our global monitoring statistic \( G_t \) is efficient not only for a large number of abnormal data streams but also for a few abnormal data streams. The reason why \( G_t \) can be efficient when a few data streams are abnormal is the following. In general, the extreme order statistics \( W_{(i),t} \) with small \( i \) or large \( i \) have larger variabilities than the order statistics in the middle. Therefore,
although the expression in (1) is in the form of an unweighted sum of squares, if we take into account the variabilities of different order statistics $W_{0,t}$, $G_t$ is actually a weighted sum of squares, with the more extreme order statistics receiving larger weight. This weighting scheme makes $G_t$ also sensitive for a few abnormal data streams, since those abnormal data streams will most likely drive up the extreme order statistics first. Therefore, despite its simple form, our proposed global monitoring statistic $G_t$ has a build-in adaptive mechanism that is capable of adapting to different types of abnormal scenarios.

As seen above, the proposed new class of global monitoring statistics is very general and can work with any local monitoring statistic other than the above CUSUM statistics $(1)$. To see this, note that $G_t$ can be modified using local monitoring statistics other than the above CUSUM statistics. Assume that the OC means $q_{0,t}$, $(i = 1, \ldots, m)$ of the IC distribution of the $W_{t,i}$ are needed. When the IC distribution of the $W_{t,i}$ is from some well-known distribution family, the $q_{0,t}$ can be easily obtained from that distribution family. When the IC distribution of the $W_{t,i}$ is not from any well-known distribution family, which is most often the case, we can first simulate a random sample from the IC distribution of the $W_{t,i}$ and then use the sample quantiles of this random sample to approximate the corresponding $q_{0,t}$. We will give several examples on how to obtain those approximations in the next section. It should be noted that obtaining approximations of the $q_{0,t}$ will be carried out offline before the online monitoring starts and the values will be stored beforehand. As a result, the total online computational effort in calculating $G_t$ is the same as calculating $\sum_{i=1}^{m}(W_{t,i} - a_i)^2I_{(W_{t,i} > a_i)}$ with all the $a_i$ given. Therefore, it is computationally simple to implement the proposed method online for monitoring high-dimensional data streams.

### 2.3 A special case: The global monitoring statistic proposed by Zou et al.

Assume that all the IC distributions $F_{0,t}$ are the normal distribution with mean 0 and variance 1 (denoted by $N(0, 1)$) and the OC distribution of the $i$th data stream $(i = 1, \ldots, m)$ is also some normal distribution with mean $\mu_i$ and variance 1 (denoted by $N(\mu_i, 1)$). Under those assumptions, an optimal local monitoring statistic for each data stream is the CUSUM statistic. To detect a positive mean shift, the CUSUM statistic for the $i$th data stream is defined as

$$G_t^Z = \sum_{i=1}^{m} \left\{ \log \left[ \frac{U_{0,i,t}^{-1} - 1}{(m - 1/2)/(i - 3/4) - 1} \right] \right\}^2 I_{(U_{0,i,t} > (i - 3/4)/(m - 1/2))}. \tag{3}$$

In the following, we show that $G_t^Z$ can be considered as a special case of our proposed global monitoring statistic $G_t$ in (1). To see this, note that $Q(p) = -\log(p - 1)$ is the quantile function of the standard logistic distribution. If we assume that the OC means $\mu_i$, $i = 1, \ldots, m$, are all equal, then $-\log(U_{0,i,t}^{-1} - 1)$ can be considered as the observed $(i - 3/4)/(m - 1/2)$ quantile of the underlying distribution of $S_{t,i}^+$ on the scale of the standard logistic distribution. Similarly, $-\log((i - 3/4)/(m - 1/2))^{-1}$ is the expected $(i - 3/4)/(m - 1/2)$ quantile of the underlying distribution of $S_{t,i}^+$ on the scale of the standard logistic distribution. As a result, if we choose $W_{t,i}$ to be $-\log(U_{0,i,t}^{-1} - 1)$ with $W_{0,i,t} = -\log(U_{0,i,t}^{-1} - 1)$ and $q_{0,t} = -\log((i - 3/4)/(m - 1/2))^{-1}$, our proposed global monitoring statistic $G_t$ in (1) reduces to $G_t^Z$ in (3).

The above shows that $G_t^Z$ is a special case of our proposed global monitoring statistic $G_t$. Therefore, $G_t^Z$ can be also considered as the sum of the squared differences between the observed quantiles and expected quantiles. Theoretically, $G_t^Z$ can be modified using local monitoring statistics other than the above CUSUM statistics $S_{t,i}^+$. However, the quantiles used in $G_t^Z$ are on the scale of the standard logistic distribution. To obtain those quantiles, the local monitoring statistics have to be transformed to $U_{0,i,t}$ based on their underlying IC distribution. For many commonly used local monitoring statistics, there is no analytical form available for this transformation, and it has to be approximated through the Markov chain
method or Monte Carlo simulation, which can be time-consuming, especially for high-dimensional data monitoring. This greatly restricts the applicability of $G_t^Z$ to other settings. Therefore, in Zou et al.\textsuperscript{5} all the analysis was limited to using $S^+_{L,t}$ in (2) as the local monitoring statistic, since a closed-form formula to approximate $U_{t,t}$ is available in this setting thanks to Grigg and Spiegelhalter.\textsuperscript{9} In contrast, the quantiles used in our proposed $G_t$ can be directly determined by any local monitoring statistics $W_{t,t}$, which makes our $G_t$ more versatile and more computationally efficient than $G_t^Z$ for monitoring high-dimensional data.

3 | EXAMPLES

In Section 2, we propose a general class of global monitoring statistics, which can be built on any local monitoring statistic. In this section, we provide three examples of our proposed global monitoring statistic from this general class and compare their performance with that of other existing global monitoring statistics.

3.1 | Known prechange and postchange distributions

In our first example, we assume that the distributions before and after the change are $N(0,1)$ and $N(\mu,1)$, respectively, for all the data streams, where $\mu$ is the postchange mean and is completely specified. Under this setting, we can use the CUSUM statistic $S^+_{L,t}$ defined in (2) with $\mu_t = \mu$ as the local monitoring statistic.

On the basis of this local monitoring statistic, the global monitoring statistic $G_t^Z$ defined in (3) can be used to monitor the $m$ data streams jointly. As mentioned earlier, to implement $G_t^Z$, it is important that $U_{t,t}$ can be calculated quickly. Grigg and Spiegelhalter\textsuperscript{9} developed an empirical approximation to the IC steady-state distribution of the CUSUM statistic $G_t^Z$. Their result can be used to obtain a closed-form formula to calculate $U_{t,t}$. Since this formula only works when $S^+_{L,t}$ reaches its steady state, to make use of this formula, we modify the definition of the CUSUM statistic a little. Instead of starting the CUSUM statistic at 0, ie, $S^+_{L,0} = 0$ as in (2), we start the CUSUM statistic at some value randomly drawn from the IC steady-state distribution of $S^+_{L,t}$. More specifically, we first generate $10^5$ independent sequences of $\{X_{k,i}: \ldots, X_{k,2000}\}$ ($k = 1, \ldots, 10^5$), each of which is independently drawn from $N(0,1)$, and calculate $S^+_{k,2000}$ as in (2). Then, $\{S^+_{k,2000}\}_{k=1}^{10^5}$ can serve as a random sample from the IC steady-state distribution of $S^+_{L,t}$. Our modified CUSUM statistic is then defined as follows. For $i = 1, \ldots, m$,

\[
\begin{cases}
S^{++}_{0,i} = V_i \\
S^{++}_{t,i} = \max(0, S^{++}_{t-1,i} + \mu(X_{L,i} - \frac{1}{2}\mu)), \text{ for } t \geq 1
\end{cases}
\]

where $V_i$ is randomly drawn with replacement from $\{S^+_{k,2000}\}_{k=1}^{10^5}$. The global monitoring statistic $G_t^Z$ in (3) is then calculated using $U_{t,t} = H^*(S^{++}_{t,i})$, where $H^*(\cdot)$ is the IC distribution of $S^{++}_{t,i}$. Since $S^{++}_{t,i}$ starts from the steady state, $H^*(\cdot)$ at any time $t$ follows the IC steady-state distribution. As a result, we can utilize the closed-form formula provided in Grigg and Spiegelhalter\textsuperscript{9} to calculate the above $U_{t,t}$ quickly.

Similarly, we also use the above-modified CUSUM statistic $S^{++}_{L,t}$ as $W_{t,t}$ in our proposed global monitoring statistic $G_t$ in (1). Because of this modification, the underlying distribution of $W_{t,t}$ for any time $t$ is the same as that of $\{S^+_{k,2000}\}_{k=1}^{10^5}$ obtained above. Then, its expected quantiles $q^+_{t,i}$ also remain the same for any time $t$ and can be well approximated by the corresponding sample quantiles of $\{S^+_{k,2000}\}_{k=1}^{10^5}$, which we denote by $\hat{q}^+_{t,i}$. Therefore, our proposed global monitoring statistic $G_t$ in this particular setting is

\[
G_t = \sum_{i=1}^{m} \left( S^{++}_{0,i,t} - \hat{q}^+_{t,i} \right)^2 I_{(S^{++}_{0,i,t} > \hat{q}^+_{t,i})},
\]

where $S^{++}_{1,t} \leq \ldots \leq S^{++}_{m,t}$ are the order statistics of $S^{++}_{1,t}, \ldots, S^{++}_{m,t}$.

In Liu et al.\textsuperscript{6} several global monitoring statistics based on hard thresholding, soft thresholding and order thresholding were proposed. From their simulations, the soft-thresholding method seems to work the best. Therefore, we only include their soft-thresholding-based global monitoring statistic in the following simulation study for performance comparison. To be consistent with the above $G_t^Z$ and $G_t$, we also calculate their global monitoring statistic based on the above-modified CUSUM statistic $S^{++}_{L,t}$, which is defined as

\[
G_t^L = \sum_{i=1}^{m} \max\{S^{++}_{L,i,t} - b, 0\},
\]

where $b$ is the thresholding constant. Following Liu et al.\textsuperscript{6} three choices of $b$ are considered: (a) $b_1 = 1/2$; (b) $b_2 = \log(10) = 2.3026$; and (c) $b_3 = 4.6052$.
• Simulation study

In the following, we report a simulation study to compare the performance of \( G_t, G_t^Z \), and \( G_t^L \). The general simulation settings are the following. Among the \( m \) data streams, \( m_0 \) data streams are from the IC distribution \( N(0, 1) \), and the remaining \( m_1 = m - m_0 \) data streams are from the OC distribution \( N(0.5, 1) \). We consider two choices of \( m \): \( m = 100 \) and \( 1000 \), and Table 2 lists the corresponding choices of \( m_1 \) for these two choices of \( m \).

In our simulation study, we construct the monitoring scheme by tracking \( G_t, G_t^Z \), and \( G_t^L \), respectively. If \( G_t, G_t^Z \), or \( G_t^L \) exceed its respective control limit \( h \), the corresponding monitoring scheme will stop the monitoring and raise an alarm. The control limit \( h \) for \( G_t, G_t^Z \), and \( G_t^L \) can be obtained through Monte Carlo simulation to satisfy the ARL0 requirement. The desired ARL0 for all the monitoring schemes is set at 1000. The control limits \( h \) for \( G_t, G_t^Z \), and \( G_t^L \) for different values of \( m \) are listed in Table 1.

Using those control limits, the monitoring schemes based on \( G_t, G_t^Z \), and \( G_t^L \) are then used to monitor the above \( m \) data streams with \( m_1 \) of them being OC. Since those OC data streams have changed from their IC distributions from the very beginning, the detection power of the monitoring schemes based on \( G_t, G_t^Z \), and \( G_t^L \) can be compared with the average time for the monitoring scheme to raise an alarm, ie, the average run length (denoted by ARL1). Table 2 reports the ARL1 of the monitoring schemes based on \( G_t, G_t^Z \), and \( G_t^L \) for different settings from 2500 simulations. The standard deviations of the run lengths from the 2500 simulations are also included in parentheses, and the standard errors of the ARL1 are simply those standard deviations divided by 50.

### Table 1

The control limits of the monitoring schemes based on \( G_t, G_t^Z \), and \( G_t^L \) when ARL0 = 1000

| \( m \) | \( m_1 \) | \( G_t \) \( G_t^Z \) \( G_t^L \) | \( b_1 \) | \( b_2 \) | \( b_3 \) |
|---|---|---|---|---|---|
| 100 | 80 | 64.44 (32.88) 68.04 (33.15) | 110.26 (55.02) 81.22 (40.72) | \( 62.71 \) (31.84) |
| 100 | 50 | 36.20 (14.95) 38.17 (15.01) | 48.44 (21.88) 38.23 (15.92) | \( 35.74 \) (13.62) |
| 100 | 20 | 27.08 (10.53) 28.75 (10.96) | 32.55 (14.35) 27.26 (10.53) | 28.65 (9.89) |
| 100 | 10 | 20.33 (7.26) 20.77 (7.76) | 21.19 (8.94) 19.67 (7.06) | 23.04 (7.53) |
| 100 | 5 | 17.42 (6.11) 17.63 (6.64) | 17.33 (7.21) 17.01 (5.93) | 21.08 (6.52) |
| 100 | 4 | 10.64 (3.67) 10.04 (3.84) | 9.43 (3.61) 9.15 (3.52) | 16.12 (4.56) |
| 100 | 3 | 4.88 (1.57) 3.91 (1.43) | 4.15 (1.40) 4.62 (1.97) | 11.23 (3.33) |
| 100 | 2 | 3.25 (0.97) 2.37 (0.82) | 2.81 (0.93) 4.99 (1.54) | 9.48 (2.81) |
| 100 | 1 | 2.72 (0.77) 1.89 (0.59) | 2.37 (0.74) 4.32 (1.34) | 8.56 (2.63) |

### Table 2

The ARL1 of the monitoring schemes based on \( G_t, G_t^Z \), and \( G_t^L \) from 2500 simulations

| \( m \) | \( m_1 \) | \( G_t \) \( G_t^Z \) \( G_t^L \) | \( b_1 \) | \( b_2 \) | \( b_3 \) |
|---|---|---|---|---|---|
| 100 | 80 | 64.44 (32.88) 68.04 (33.15) | 110.26 (55.02) 81.22 (40.72) | \( 62.71 \) (31.84) |
| 100 | 50 | 36.20 (14.95) 38.17 (15.01) | 48.44 (21.88) 38.23 (15.92) | \( 35.74 \) (13.62) |
| 100 | 20 | 27.08 (10.53) 28.75 (10.96) | 32.55 (14.35) 27.26 (10.53) | 28.65 (9.89) |
| 100 | 10 | 20.33 (7.26) 20.77 (7.76) | 21.19 (8.94) 19.67 (7.06) | 23.04 (7.53) |
| 100 | 5 | 17.42 (6.11) 17.63 (6.64) | 17.33 (7.21) 17.01 (5.93) | 21.08 (6.52) |
| 100 | 4 | 10.64 (3.67) 10.04 (3.84) | 9.43 (3.61) 9.15 (3.52) | 16.12 (4.56) |
| 100 | 3 | 4.88 (1.57) 3.91 (1.43) | 4.15 (1.40) 4.62 (1.97) | 11.23 (3.33) |
| 100 | 2 | 3.25 (0.97) 2.37 (0.82) | 2.81 (0.93) 4.99 (1.54) | 9.48 (2.81) |
| 100 | 1 | 2.72 (0.77) 1.89 (0.59) | 2.37 (0.74) 4.32 (1.34) | 8.56 (2.63) |

Note. The standard deviations of the run lengths from the 2500 simulations are reported in parentheses.
For the monitoring schemes based on $G^c_t$, the bold number in each row of Table 2 represents the smallest ARL1 among the three choices of $b$ for that particular OC scenario. As the table shows, the detection power of $G^c_t$ depends on the choice of $b$. If $b$ is too small, $G^c_t$ is not powerful when only a few data streams are OC, since many IC data streams may exceed $b$ and the signal in $G^c_t$ will be diluted by including those IC data streams. Similarly, if $b$ is too large, $G^c_t$ is not powerful when many data streams are OC, since many of them may not exceed $b$ and hence do not contribute to $G^c_t$. This is consistent with what is known for any thresholding-based method. However, in practice, it is rarely known in advance how many data streams will be OC, which makes it extremely difficult to come up with an appropriate $b$ for $G^c_t$ in real applications. In contrast, both $G_t$ and $G^d_t$ do not depend on any sort of tuning parameter, and they perform well across different OC scenarios with detection delays being always close to those of $G^c_t$ with the best choice of $b$.

When further comparing $G_t$ with $G^d_t$, we notice that our $G_t$ is better when a small number of data streams are OC, while $G^d_t$ is better when a large number of data streams are OC. This can be explained by the following. As described in Section 2, both of the global monitoring statistics can be viewed as the sum of the squared differences between the observed quantiles and expected quantiles. $G^d_t$ is based on the quantiles from the standard logistic distribution, while our $G_t$ is based on the quantiles from the distribution of the CUSUM statistic $S^{+s}_t$ in (4). As shown in Grigg and Spiegelhalter, the tail of the IC distribution of $S^{+s}_t$ resembles that of an exponential distribution, which implies that the IC distribution of $S^{+s}_t$ has a heavier tail than the standard logistic distribution. As a result, the extreme quantiles contribute more in $G_t$ than in $G^d_t$. This explains why $G_t$ performs better than $G^d_t$ when a small number of data streams are OC and vice versa when a large number of data streams are OC.

### 3.2 Known prechange distribution but unknown postchange distribution

In the previous example, in order to use the CUSUM statistic as the local monitoring statistic, the distribution after the change needs to be completely specified. In some real-world applications, prior knowledge of the postchange distribution may not be available. In our second example, we consider the setting where the postchange distribution is unknown. More specifically, we assume that the OC distribution of the $i$th data stream is $N(\mu_i, 1)$ with $\mu_i$ unknown and the IC distributions of all data streams are still $N(0, 1)$. To obtain the specific form of our proposed global monitoring statistic $G_t$, the key is to find the appropriate local monitoring statistic $W_{i,t}$ in this setting. There exist a few options for such a statistic in the statistical process monitoring literature. For example, Sparks proposed an adaptive CUSUM statistic, and Han and Tsung developed a reference-free cumulative score statistic. In both of the two methods, instead of using the specified $\mu_i$ in the CUSUM statistic defined in (2), an estimate of $\mu_i$ is plugged in. In Sparks, an exponentially weighted moving average (EWMA) of all the past observations is used to estimate $\mu_i$, while, in Han and Tsung, the absolute value of the current observation, $|X_{i,t}|$, is used as the estimate of $\mu_i$. Following the same idea, Lorden and Pollak proposed another estimate of $\mu_i$ to replace $\mu_i$ in the CUSUM statistic in (2) and proved the asymptotic optimality of the resulting monitoring statistic. Since the CUSUM statistic in (2) is only for detecting positive mean shifts, the monitoring statistic developed in Lorden and Pollak is also only for positive mean shifts. Recently, Liu et al extended Lorden and Pollak's monitoring statistic to detect both positive and negative mean shifts. In the following, we use this two-sided monitoring statistic in Liu et al as our local monitoring statistic $W_{i,t}$.

More specifically, define, for $t \geq 1$,

\[
C^{(1)}_{i,t} = \max \left( 0, C^{(1)}_{i,t-1} + \hat{\mu}^{(1)}_{i,t} \left( X_{i,t} - \frac{1}{2} \hat{\mu}^{(1)}_{i,t} \right) \right),
\]

\[
C^{(2)}_{i,t} = \max \left( 0, C^{(2)}_{i,t-1} + \hat{\mu}^{(2)}_{i,t} \left( X_{i,t} - \frac{1}{2} \hat{\mu}^{(2)}_{i,t} \right) \right),
\]

where $\hat{\mu}^{(1)}_{i,t}$ and $\hat{\mu}^{(2)}_{i,t}$ are the estimates of $\mu_i$ for the positive mean shift and negative mean shift, respectively, and they are given by

\[
\hat{\mu}^{(1)}_{i,t} = \max \left( \rho, \frac{s + S^{(1)}_{i,t}}{t + T^{(1)}_{i,t}} \right) > 0, \quad \hat{\mu}^{(2)}_{i,t} = \min \left( -\rho, \frac{-s + S^{(2)}_{i,t}}{t + T^{(2)}_{i,t}} \right) < 0.
\]

In the above estimates, $\rho$ is the pre-specified smallest mean shift that is meaningful, and $s$ and $t$ are also pre-specified nonnegative constants and can be considered as a prior so that the above estimates can be treated as the Bayes-type estimates. In our simulation studies, we choose $\rho = 0.25$, $s = 1$, and $t = 4$ as in Liu et al. For $j = 1, 2$, the sequences
In Liu et al,\(^6\) the above monitoring statistic \(G_t\) is calculated recursively.

\[
S_{l,t}^{(j)} = \begin{cases} 
S_{l,t-1}^{(j)} + X_{l,t-1}, & \text{if } C_{l,t-1}^{(j)} > 0, \\
0, & \text{if } C_{l,t-1}^{(j)} = 0.
\end{cases}
\]

\[
T_{l,t}^{(j)} = \begin{cases} 
T_{l,t-1}^{(j)} + 1, & \text{if } C_{l,t-1}^{(j)} > 0, \\
0, & \text{if } C_{l,t-1}^{(j)} = 0.
\end{cases}
\]

Finally, our local monitoring statistic \(C_{l,t}\) is simply

\[C_{l,t} = \max(C_{l,t}^{(1)}, C_{l,t}^{(2)}).\]

In Liu et al,\(^6\) the above monitoring statistic \(C_{l,t}\) starts from the following initial values:

\[S_{l,0}^{(1)} = S_{l,0}^{(2)} = T_{l,0}^{(1)} = T_{l,0}^{(2)} = C_{l,0}^{(1)} = C_{l,0}^{(2)} = X_{l,0} = 0.\]

Using those initial values, the IC distribution of the \(C_{l,t}\) will change over the time before it reaches its steady state. Recall that, to implement our proposed global monitoring statistic \(G_t\), the expected quantiles \(\bar{q}_{(i,t)}\) of the IC distribution of the \(C_{l,t}\) are needed. If the IC distribution of the \(C_{l,t}\) changes over the time, then we need to calculate and store \(\{q_{(i,t)}\}^m_{i=1}\) for each \(t\). To simplify our procedure, similarly to how we modified the original CUSUM statistic in the previous section, we propose to set the initial values, \(\{S_{l,0}^{(1)}, S_{l,0}^{(2)}, T_{l,0}^{(1)}, T_{l,0}^{(2)}, C_{l,0}^{(1)}, C_{l,0}^{(2)}, X_{l,0}\}\) at some value randomly drawn from the IC steady-state distribution of \((S_{l,0}^{(1)}, S_{l,0}^{(2)}, T_{l,0}^{(1)}, T_{l,0}^{(2)}, C_{l,0}^{(1)}, C_{l,0}^{(2)}, X_{l,0})\). To obtain such initial values, we generate \(10^5\) independent sequences of \(\{X_{k,1}, \ldots, X_{k,2000}\}\) \((k = 1, \ldots, 10^5)\), each of which is independently drawn from \(N(0, 1)\), and calculate \((S^{(1)}_{k,2000}, S^{(2)}_{k,2000}, T^{(1)}_{k,2000}, T^{(2)}_{k,2000}, C^{(1)}_{k,2000}, C^{(2)}_{k,2000}, X_{k,2000})\). Using the initial value \(0\), then, \((S^{(1)}_{k,2000}, S^{(2)}_{k,2000}, T^{(1)}_{k,2000}, T^{(2)}_{k,2000}, C^{(1)}_{k,2000}, C^{(2)}_{k,2000}, X_{k,2000})\) \(k=1\) can be used to approximate the IC steady-state distribution of \((S_{l,t}, S_{l,t}^{(j)}, T_{l,t}^{(j)}, C_{l,t}^{(1)}, C_{l,t}^{(2)}, X_{l,t})\).

The initial values to calculate our modified \(C_{l,t}^*\) are then defined as

\[
(S_{l,0}^{(1)}, S_{l,0}^{(2)}, T_{l,0}^{(1)}, T_{l,0}^{(2)}, C_{l,0}^{(1)}, C_{l,0}^{(2)}, X_{l,0}) = V_i,
\]

where \(V_i\) is randomly drawn from \((S^{(1)}_{k,2000}, S^{(2)}_{k,2000}, T^{(1)}_{k,2000}, T^{(2)}_{k,2000}, C^{(1)}_{k,2000}, C^{(2)}_{k,2000}, X_{k,2000})\) \(k=1\) with replacement. Since \((S_{l,t}^{(1)}, S_{l,t}^{(2)}, T_{l,t}^{(1)}, T_{l,t}^{(2)}, C_{l,t}^{(1)}, C_{l,t}^{(2)}, X_{l,t})\) starts from the steady state, \(C_{l,t}^*\) at any time \(t\) follows the IC steady-state distribution when the system is IC, and its expected quantiles \(\bar{q}_{(i,t)}\) also remain the same for any time \(t\). Then, \(\{\max(C_{k,2000}^{(1)}, C_{k,2000}^{(2)})\}_{k=1}^{10^5}\) can be used to approximate the IC steady-state distribution of \(C_{l,t}^*\), and the expected quantiles \(\bar{q}_{(i,t)}\) of \(C_{l,t}^*\) can be approximated by the corresponding sample quantiles of \(\{\max(C_{k,2000}^{(1)}, C_{k,2000}^{(2)})\}_{k=1}^{10^5}\), which we denote by \(\bar{q}_{(i,t)}^*\). Therefore, our proposed global monitoring statistic \(G_t\) in this particular setting is

\[
G_t = \sum_{i=1}^{m} \left( C_{(i,t)}^* - \bar{q}_{(i,t)}^* \right)^2 \mathbf{1}_{(C_{(i,t)}^* > \bar{q}_{(i,t)}^*)},
\]

where \(C_{(1,t)}^* \leq \ldots \leq C_{(m,t)}^*\) are the order statistics of \(C_{1,t}^*, \ldots, C_{m,t}^*\).

- **Simulation study**

Using the above modified local monitoring statistics \(C_{l,t}^*\) theoretically, it is possible to define the global monitoring statistic \(G_t^Z\) proposed by Zou et al\(^5\) in this setting accordingly. To implement this \(G_t^Z\), it is important to have a closed-form formula for the cumulative distribution function of \(C_{l,t}^*\). However, it is not easy to develop such a formula for the above \(C_{l,t}^*\). Because of this computational difficulty of \(G_t^Z\), in our simulation study, we only compare our global monitoring statistic \(G_t\) defined above with the one proposed in Liu et al.\(^6\) Their global monitoring statistic based on soft thresholding in this particular setting is defined as

\[
G_t^Z = \sum_{i=1}^{m} \max\{C_{(i,t)}^* - b, 0\}.
\]
Again, following Liu et al., three choices of \( b \) are considered: (a) \( b_1 = 1/2 \); (b) \( b_2 = \log(10) = 2.3026 \); and (c) \( b_3 = \log(100) = 4.6052 \).

The specific simulation settings are similar to those in Section 3.1. Among the \( m \) data streams, \( m_0 \) data streams are from the IC distribution \( N(0, 1) \), and the remaining \( m_1 = m - m_0 \) data streams are from the OC distribution \( N(\mu_i, 1) \), \( i = 1, \ldots, m_1 \), where \( \mu_i \) is randomly drawn from \((-0.5, 0.5)\). We consider two choices of \( m \): \( m = 100 \) and 1000, and Table 4 lists the corresponding choices of \( m_1 \) for these two choices of \( m \).

Similar to the first simulation study reported in Section 3.1, the performance of \( G_t \) and \( G_t^L \) is compared based on the ARL0 of their corresponding monitoring schemes. The desired ARL0 for the \( G_t \) and \( G_t^L \)-based monitoring schemes is set at 1000. The control limits \( h \) for those monitoring schemes, which are obtained through Monte Carlo simulation, are listed in Table 3.

On the basis of those control limits, the ARL1 of the \( G_t \) and \( G_t^L \)-based monitoring schemes are obtained from 2500 simulations, which are reported in Table 4. The standard deviations of the run lengths from the 2500 simulations are also included in parentheses. Again, for the \( G_t^L \)-based monitoring schemes, the bold number in each row represents the smallest ARL1 among the three choices of \( b \) for that particular OC scenario. As we can see from the table, the ARL1 of \( G_t^L \) depends on the choice of \( b \); \( G_t^L \) with a small \( b \) does not perform well when a small number of data streams are OC, while \( G_t^L \) with a large \( b \) does not perform well when a large number of data streams are OC. The explanation is similar to that we give in Section 3.1. Since it is rarely known in advance how many data streams will be OC in practice, it is extremely difficult to come up with an appropriate \( b \) for \( G_t^L \) in real applications. In contrast, despite the fact that there is no tuning parameter involved, our \( G_t \) performs well across different OC scenarios, and its detection delays are always close to those of \( G_t^L \) with the best choice of \( b \). This makes our \( G_t \) particularly appealing in many real-world applications.

### 3.3 Unknown prechange and postchange distributions

In the previous two examples, the prechange distributions for all data streams are assumed to be completely known and are specified by some particular parametric distribution. In some applications, it might not be easy to identify the appropriate distributions for all data streams. Therefore, in this example, we consider the setting where both the prechange and postchange distributions are unknown. Again, to obtain the specific form of our proposed global monitoring statistic \( G_t \) in this setting, we need to find the appropriate local monitoring statistic \( W_{lt} \). To deal with the unknown prechange distribution, a nonparametric monitoring statistic should be used. To deal with the unknown postchange distribution, we need a nonparametric monitoring statistic that can detect any arbitrary distributional changes. In the literature, to deal with the unknown prechange distribution, many nonparametric monitoring statistics assume that a large amount of IC reference data generated by the prechange distribution is available so that certain characteristics of the prechange distribution can be well estimated. However, in order for the effect of using estimates instead of the true values on the ARL0 to be negligible, it usually requires a substantial amount of IC reference data. In many real-world applications, it can be very challenging to have such data. Therefore, to find a good candidate for our \( W_{lt} \), we only focus on the nonparametric monitoring statistics that have the self-starting feature.

There are a few nonparametric self-starting monitoring statistics that can detect any arbitrary distributional changes in the literature. For example, Zou and Tsung proposed an EWMA statistic based on a powerful goodness-of-fit test. However, according to the simulation studies conducted in Ross and Adams, this EWMA statistic is only sensitive in detecting scale increases and is not as powerful as its competitors in detecting other types of distributional changes including location shifts. Ross and Adams further proposed two monitoring statistics based on the change-point detection (CPD) framework. Their proposed CPD statistics are shown to have better overall performance than Zou and Tsung’s EWMA statistics for detecting different distributional changes. However, like most CPD statistics, the computation of their proposed statistics is very intensive, which makes them very challenging to implement for monitoring high-dimensional data. Recently, Li proposed a nonparametric self-starting CUSUM statistic that can detect any arbitrary distributional changes. On the basis of the simulation studies in Li, the proposed monitoring statistic not only is computationally more efficient than Ross and Adams’s CPD statistics but also has better overall detection power than those CPD statistics. Therefore, in the following, we use the CUSUM statistic proposed in Li as our local monitoring statistic \( W_{lt} \).
TABLE 4 The ARL₁ comparison of the monitoring schemes based on $G_t$ and $G_t^2$.

| $m$ | $m_1$ | $G_t^2$ | $b_1$ | $b_2$ | $b_3$ |
|-----|-------|---------|-------|-------|-------|
| 1   | 71.05 (37.20) | 120.08 (62.31) | 88.54 (45.55) | **70.94** (36.97) |
| 3   | 41.05 (17.58) | 55.64 (25.85) | 44.01 (19.25) | **40.89** (17.21) |
| 5   | 31.37 (12.67) | 37.93 (16.37) | **31.89** (13.03) | 32.13 (12.57) |
| 8   | 24.51 (9.27) | 26.89 (10.98) | **24.25** (9.15) | 26.24 (9.69) |
| 100 | 21.54 (8.08) | 22.86 (9.24) | **21.10** (7.84) | 23.86 (8.73) |
| 20  | 14.43 (4.77) | **13.49** (4.85) | 14.25 (4.62) | 18.02 (5.83) |
| 50  | 8.17 (2.26) | **6.98** (2.13) | 8.59 (2.44) | 12.41 (3.60) |
| 80  | 6.17 (1.58) | **5.20** (1.40) | 6.85 (1.81) | 10.39 (2.87) |
| 100 | 5.40 (1.26) | **4.51** (1.14) | 6.16 (1.53) | 9.52 (2.53) |

| $m$ | $m_1$ | $G_t^2$ | $b_1$ | $b_2$ | $b_3$ |
|-----|-------|---------|-------|-------|-------|
| 1   | 91.72 (43.42) | 229.39 (126.38) | 161.70 (83.56) | **106.88** (51.26) |
| 3   | 56.75 (22.74) | 114.71 (56.63) | 80.57 (35.88) | **59.57** (23.93) |
| 5   | 46.40 (17.00) | 82.56 (37.59) | 59.86 (24.22) | **47.45** (18.65) |
| 8   | 37.48 (13.18) | 59.71 (25.92) | 44.29 (16.67) | **37.83** (12.81) |
| 10  | 34.05 (11.68) | 50.09 (20.84) | 38.23 (13.94) | **33.98** (11.23) |
| 20  | 24.57 (7.74) | 30.05 (11.50) | 25.26 (8.34) | **25.19** (7.62) |
| 50  | 15.36 (4.15) | 15.04 (4.95) | **14.81** (4.08) | 17.24 (4.50) |
| 80  | 11.77 (3.06) | **10.55** (3.30) | 11.36 (2.96) | 14.35 (3.55) |
| 100 | 10.42 (2.67) | **9.15** (2.78) | 10.15 (2.65) | 13.17 (3.30) |
| 1000| 150 | 8.15 (1.96) | **6.86** (1.94) | 8.19 (1.98) | 11.26 (2.61) |
| 200 | 6.94 (1.60) | **5.74** (1.53) | 7.12 (1.68) | 10.14 (2.31) |
| 300 | 5.49 (1.18) | **4.48** (1.08) | 5.90 (1.33) | 8.74 (2.00) |
| 400 | 4.70 (0.97) | **3.83** (0.89) | 5.20 (1.12) | 7.91 (1.75) |
| 500 | 4.19 (0.84) | **3.43** (0.75) | 4.72 (1.00) | 7.28 (1.59) |
| 600 | 3.78 (0.76) | **3.12** (0.69) | 4.33 (0.91) | 6.77 (1.48) |
| 700 | 3.50 (0.68) | **2.88** (0.60) | 4.09 (0.84) | 6.45 (1.38) |
| 800 | 3.28 (0.66) | **2.71** (0.59) | 3.85 (0.84) | 6.13 (1.34) |
| 900 | 3.10 (0.58) | **2.61** (0.55) | 3.70 (0.75) | 5.90 (1.25) |
| 1000| 2.95 (0.52) | **2.47** (0.54) | 3.54 (0.74) | 5.67 (1.25) |

Note. The standard deviations of the run lengths from the 2500 simulations are reported in parentheses.

Assume that, for each data stream, there are $n$ IC reference data, denoted by $X_{i-n+1}, \ldots, X_{i0}, i = 1, \ldots, m$. At time $t \geq 1$, for the $i$th data stream, we partition the real line into the following $d$ left-to-right regions:

\[
\hat{A}_{l,t,1}^{(1)} = (-\infty, \hat{q}_{l,t,1}^{(1)}], \quad \hat{A}_{l,t,2}^{(1)} = (\hat{q}_{l,t,1}^{(1)}, \hat{q}_{l,t,2}^{(1)}], \ldots, \hat{A}_{l,t,d}^{(1)} = (\hat{q}_{l,t,d-1}^{(1)}, \infty),
\]

and the following $d$ center-outward regions:

\[
\hat{A}_{l,t,1}^{(2)} = (\hat{q}_{l,t,d-1}^{(2)} \hat{q}_{l,t,d+1}^{(2)}], \quad \hat{A}_{l,t,2}^{(2)} = (\hat{q}_{l,t,d-2}^{(2)} \hat{q}_{l,t,d}^{(2)}] \cup (\hat{q}_{l,t,d+1}^{(2)}, \hat{q}_{l,t,d+2}^{(2)}], \ldots \quad \hat{A}_{l,t,d}^{(2)} = (-\infty, \hat{q}_{l,t,d-1}^{(2)}] \cup (\hat{q}_{l,t,2d-1}^{(2)}, \infty),
\]

where $\hat{q}_{l,t,j}^{(1)}$ ($j = 1, \ldots, d-1$) and $\hat{q}_{l,t,k}^{(2)}$ ($k = 1, \ldots, 2d-1$) are the $(j/d)$th and $(k/(2d))$th sample quantiles, respectively, from $X_{i-n+1}, \ldots, X_{i0}, X_{i1}, \ldots, X_{i,t-1}$. For $j = 1, \ldots, d$, define

\[
\hat{Y}_{l,t,j}^{(1)} = I(X_{i,t} \in \hat{A}_{l,t,j}^{(1)}), \quad \hat{Y}_{l,t,j}^{(2)} = I(X_{i,t} \in \hat{A}_{l,t,j}^{(2)}),
\]

and

\[
\hat{Z}_{l,t,j}^{(1)} = \sum_{i=1}^{j} \hat{Y}_{l,t,j}^{(1)}, \quad \hat{Z}_{l,t,j}^{(2)} = \sum_{i=1}^{j} \hat{Y}_{l,t,j}^{(2)}.
\]
For \( k_1, k_2 = 1, 2 \), we calculate

\[
\hat{S}_{ld}(k_1, k_2) = \max \left\{ 0, \frac{\hat{d}(k_1, k_2)}{\hat{S}_{ld}^{(k_1, k_2)}} + \frac{1}{\sqrt{d - 1}} \sum_{j=1}^{d-1} \frac{d^2}{j(d - j)} \sum_{l=1}^{\infty} b_{l,l} \log \left( \frac{\sum_{j=1}^{l} \hat{P}_{l,l}^{(k_1, k_2)}}{j/d} \right) \right\} + (1 - \hat{d}(k_1, k_2)) \log \left( \frac{1 - \sum_{l=1}^{\infty} \hat{P}_{l,l}^{(k_1, k_2)}}{1 - j/d} \right),
\]

where \( \hat{P}_{l,l}^{(k_1, k_2)} \) is defined by

\[
\hat{P}_{l,l}^{(k_1, k_2)} = \frac{a_l^{(k_1)} + N_{l,l}^{(k_1, k_2)}}{d}.
\]

and both \( N_{l,l}^{(k_1, k_2)} \) and \( N_{l,l}^{(k_2, k_1)} \) are calculated recursively by

\[
N_{l,l}^{(k_1, k_2)} = \begin{cases} 
N_{l-1,l}^{(k_1, k_2)} + 1, & \text{if } \hat{S}_{l-1}^{(k_1, k_2)} > 0, \\
0, & \text{if } \hat{S}_{l-1}^{(k_1, k_2)} = 0.
\end{cases}
\]

\[
N_{l,l}^{(k_2, k_1)} = \begin{cases} 
N_{l-1,l}^{(k_2, k_1)} + \hat{y}_{l-1,l}, & \text{if } \hat{S}_{l-1}^{(k_1, k_2)} > 0, \\
0, & \text{if } \hat{S}_{l-1}^{(k_1, k_2)} = 0.
\end{cases}
\]

The constants \( \{a_1^{(k_2)}, \ldots, a_d^{(k_2)}\} \) \((k_2 = 1, 2)\) serve as the parameters of a prior distribution and are chosen as suggested in Li.\(^1\) In particular, when using \( a_1^{(1)} \) in \( \hat{S}_{l,l}^{(1,1)} \), the prior indicates a positive location shift; therefore, \( \hat{S}_{l,l}^{(1,1)} \) is more powerful for detecting positive location shifts. When using \( a_1^{(2)} \) in \( \hat{S}_{l,l}^{(1,2)} \), the prior indicates a negative location shift, so \( \hat{S}_{l,l}^{(1,2)} \) is more powerful for detecting negative location shifts. Similarly, when using \( a_1^{(1)} \) in \( \hat{S}_{l,l}^{(2,1)} \), the prior indicates a scale increase, so \( \hat{S}_{l,l}^{(2,1)} \) is more powerful for detecting scale increases. When using \( a_1^{(2)} \) in \( \hat{S}_{l,l}^{(2,2)} \), the prior indicates a scale decrease, so \( \hat{S}_{l,l}^{(2,2)} \) is more powerful for detecting scale decreases. If we do not have any prior information about what type of changes the process might encounter, our local monitoring statistic is simply

\[
\hat{S}_{l,l} = \max(\hat{S}_{l,l}^{(1,1)}, \hat{S}_{l,l}^{(1,2)}, \hat{S}_{l,l}^{(2,1)}, \hat{S}_{l,l}^{(2,2)}),
\]

which is efficient to detect any type of distributional changes. Li\(^1\) shows that the above monitoring statistic is asymptotic distribution free. Following the suggestion in Li\(^1\) we choose \( d = 20 \) and \( n = 40 \).

In Li\(^1\), the initial values \( (\hat{S}_{l,0}^{(k_1, k_2)}, N_{l,0}^{(k_1, k_2)}, \bar{N}_{l,0}^{(k_1, k_2)}, \hat{y}_{l,0}^{(k_1)}) \), \( k_1, k_2 = 1, 2, l = 1, \ldots, d, \) and \( i = 1, \ldots, m \), for the above monitoring statistic \( \hat{S}_{l,l} \) are all set at 0. To simplify the calculation of our proposed global monitoring statistic \( \hat{G}_t \), similarly to how we modified the local monitoring statistics in the previous two examples, we propose to set the initial values \( (\hat{S}_{l,0}^{(k_1, k_2)}, N_{l,0}^{(k_1, k_2)}, \bar{N}_{l,0}^{(k_1, k_2)}, \hat{y}_{l,0}^{(k_1)}) \) at some values randomly drawn from their IC steady-state distributions. More specifically, using the distribution-free property of \( \hat{S}_{l,l} \), we generate \( 10^5 \) independent sequences of \( \{X_{k,-39}, \ldots, X_{k,0}, X_{k,1}, \ldots, X_{k,2000}\} \) \((k = 1, \ldots, 10^5)\), each of which is independently drawn from \( N(0, 1) \), and calculate

\[
\left\{ \left( \hat{S}_{k,2000}^{(k_1, k_2)}, N_{k,2000}^{(k_1, k_2)}, \bar{N}_{k,2000}^{(k_1, k_2)}, \hat{y}_{k,2000}^{(k_1)} \right) \right\}_{k=1}^{10^5}.
\]
using the initial value 0. Then, \( \{S_{k,2000}^{G\hat{m}}(k, k), S_{k,2000}^{G\hat{m}}(k, l), S_{k,2000}^{G\hat{m}}(l, k), S_{k,2000}^{G\hat{m}}(l, l)\}_{k=1}^{10} \) can be used to approximate the IC steady-state distribution of \( \{S_{lt,1}^{G\hat{m}}, N_{lt,1}^{G\hat{m}}, N_{lt,l}^{G\hat{m}}, \hat{V}_{lt,l}^{G\hat{m}}\} \). The initial values to calculate our modified \( S_{lt,1}^{G\hat{m}} \) are then defined as

\[
(S_{0,0}^{G\hat{m}}, N_{0,0}^{G\hat{m}}, N_{0,l}^{G\hat{m}}, \hat{V}_{0,l}^{G\hat{m}}) = V_l,
\]

where \( V_l \) is randomly drawn from \( \{S_{k,2000}^{G\hat{m}}(k, k), S_{k,2000}^{G\hat{m}}(k, l), S_{k,2000}^{G\hat{m}}(l, k), S_{k,2000}^{G\hat{m}}(l, l)\}_{k=1}^{10} \) with replacement. The expected quantiles \( q_{(0,l)} \) of \( S_{lt,1}^{G\hat{m}} \) can then be well approximated by the corresponding sample quantiles of \( \{\max(S_{k,2000}^{G\hat{m}}, S_{k,2000}^{G\hat{m}}, S_{k,2000}^{G\hat{m}}, S_{k,2000}^{G\hat{m}})\}_{k=1}^{10} \) which we denote by \( \hat{q}_{(0,l)}^{S} \). Therefore, our proposed global monitoring statistic \( G_t \) is

\[
G_t = \sum_{i=1}^{m} \left( \hat{S}_{(0,l)}^{i} - \hat{q}_{(0,l)}^{S} \right)^2 I_{[\hat{S}_{(0,l)}^{i} > \hat{q}_{(0,l)}^{S}]},
\]

where \( \hat{S}_{(1,1)}^{i} \leq \ldots \leq \hat{S}_{(m,1)}^{i} \) are the order statistics of \( \hat{S}_{lt,1}^{i} \).

• Simulation study

Using the above modified local monitoring statistics \( \hat{S}_{lt,1}^{i} \), again, it is difficult to implement the global monitoring statistic \( G_t^G \) proposed by Zou et al. \(^5\) since no closed-form formula for the cumulative distribution function of \( \hat{S}_{lt,1}^{i} \) is available. We can use the thresholding method proposed in Liu et al. \(^6\) to come up with some alternative global monitoring statistics. However, it is not clear how to choose a sensible threshold. Therefore, in our simulation study, we only compare our global monitoring statistic \( G_t \) defined above with two other natural competitors \( G_t^{\text{max}} = \max_{i=1, \ldots, m} \hat{S}_{lt,1}^{i} \) and \( G_t^{\text{sum}} = \sum_{i=1}^{m} \hat{S}_{lt,1}^{i} \).

Again, in our simulation study, we consider monitoring \( m \) data streams. Since \( \hat{S}_{lt,1}^{i} \) is distribution free, among the \( m \) data streams, we randomly select half of the data streams to have \( N(0, 1) \) as their IC distributions, one fifth of the data streams to have the IC distribution with 2.5 degrees of freedom as their IC distributions, and the remaining data streams to have the lognormal distribution with parameters \( \mu = 1 \) and \( \sigma = 0.5 \) as their IC distributions. For the data generated from the \( t \) or lognormal distribution, we also standardize the data so that their IC distributions have mean 0 and standard deviation 1. Among the \( m \) data streams, the first \( m_0 \) data streams follow their IC distributions all the time, and the remaining \( m_1 = m - m_0 \) data streams will experience certain distributional changes from their IC distributions at the change-point \( t = 100 \). Since \( \hat{S}_{lt,1}^{i} \) is capable of detecting any type of distributional changes, starting from the change-point \( t = 100 \), for the \( m_1 \) data streams that will experience distributional changes, we add 0.5 to the observations from the first \( [m_1/2] \) data streams to introduce the location change and multiply 1.5 to the observations from the remaining \( m_1 - [m_1/2] \) data streams to introduce the scale change. Here, \( [b] \) is the smallest integer not less than \( b \). Similar to the previous two examples, we consider two choices of \( m: m = 100 \) and 1000, and Table 6 lists the corresponding choices of \( m_1 \) for these two choices of \( m \). The desired ARL \( 0 \) for the \( G_t^{\text{max}} \)- and \( G_t^{\text{sum}} \)-based monitoring schemes is set at 1000. The control limits \( h \) for those monitoring schemes, which are obtained through Monte Carlo simulation, are listed in Table 5.

On the basis of those control limits, the ARL \( 1 \) (after the change point) of the \( G_t^{\text{max}} \)-, \( G_t^{\text{sum}} \)-, and \( G_t \)-based monitoring schemes from 2500 simulations is reported in Table 6. The standard deviations of the run lengths from the 2500 simulations are also included in parentheses. Again, the bold number in each row represents the smaller ARL \( 1 \) between \( G_t^{\text{max}} \) and \( G_t^{\text{sum}} \) for that particular OC scenario. As we can see from the table, \( G_t^{\text{max}} \) works best when only a few data streams are OC but does not perform well when a large number of data streams are OC. On the other hand, \( G_t^{\text{sum}} \) has the best performance when a large number of data streams are OC but has the worst performance when only a few data streams are OC. In contrast, our \( G_t \) performs well across different OC scenarios, and if its detection delay is not the best among all the three monitoring statistics, it is always very close to the best. This provides another example of the robust performance of our proposed global monitoring statistic \( G_t \) for detecting different OC scenarios.

| \( m = 100 \) | \( m = 1000 \) |
| --- | --- |
| \( G_t \) | \( G_t^{\text{max}} \) | \( G_t^{\text{sum}} \) | \( G_t \) | \( G_t^{\text{max}} \) | \( G_t^{\text{sum}} \) |
| \( h \) | 144.016 | 26.908 | 524.492 | 171.697 | 33.492 | 4720.355 |
TABLE 6  The ARL<sub>1</sub> comparison of the monitoring schemes based on $G_t$, $G_t^{\max}$, and $G_t^{\sum}$

| $m$ | $G_t$ | $G_t^{\max}$ | $G_t^{\sum}$ |
|-----|-------|--------------|--------------|
| 1   | 85.89 (52.00) | **80.51** (51.40) | 181.85 (136.38) |
| 3   | 53.43 (25.14) | **57.68** (27.62) | 86.00 (49.69) |
| 5   | 43.92 (16.91) | **51.91** (19.98) | 59.84 (29.17) |
| 8   | 27.70 (9.55) | **35.90** (13.36) | 32.82 (14.25) |
| 10  | 12.39 (118.56) | **111.90** (98.57) | 452.76 (430.52) |
| 20  | 17.24 (4.99) | **28.52** (9.47) | 28.33 (11.58) |
| 50  | 7.73 (1.87) | **57.68** (13.36) | 86.00 (49.69) |
| 100 | 6.63 (1.54) | **19.31** (5.63) | 5.35 (1.45) |

Note. The standard deviations of the run lengths from the 2500 simulations are reported in parentheses.

TABLE 7  The computational times (in minutes) of the monitoring schemes based on $G_t$, $G_t^{\max}$, and $G_t^{\sum}$ in calculating their ARL<sub>1</sub>s for different choices of $m_1$ in the simulation study from Section 3.1

| $m$ | $G_t$ | $G_t^{\max}$ | $b_1$ | $b_2$ | $b_3$ |
|-----|-------|--------------|-------|-------|-------|
| 100 | 0.24  | 2.34         | 0.21  | 0.20  | 0.19  |
| 1000| 2.51  | 33.27        | 2.71  | 2.11  | 1.89  |

4 | CONCLUDING REMARKS

In this paper, we introduce a general class of global monitoring statistics for high-dimensional data streams. Our proposed global monitoring statistics are easy to calculate, which makes them suitable for monitoring high-dimensional data streams. To show the computational efficiency of our proposed global monitoring statistics, we report in Table 7 the computational times of the five monitoring schemes in calculating their ARL<sub>1</sub>s for different choices of $m_1$ in the simulation study from Section 3.1. As we can see from Table 7, the monitoring schemes from the thresholding-based $G_t^{\sum}$ are the most efficient in terms of computation, since $G_t^{\sum}$ only requires the comparison of the local monitoring statistics $W_{i,t}$ with some pre-specified threshold. For our proposed monitoring scheme, after we obtain the estimates of the $q_{(i),t}$ offline beforehand, the total online computational effort in calculating our proposed $G_t$ is the same as calculating $\sum_{i=1}^{m}(W_{i,t} - a_i)^2I_{W_{i,t} > a_i}$ with all the $a_i$ given. Therefore, our $G_t$ only needs to order the $W_{i,t}$, which requires $O(m \log m)$ computations. Although this is more time-consuming than simply comparing $W_{i,t}$ to a threshold, from Table 7, we can see that the computational times of our $G_t$-based monitoring scheme are comparable with those from the $G_t^{\sum}$-based monitoring schemes. In contrast, the computational times of the monitoring scheme based on $G_t^{\max}$ are almost 10 times the computational times of our proposed monitoring scheme. This is due to the extra computation needed to convert $W_{i,t}$ to $U_{i,t}$ in calculating $G_t^{\max}$.

In our simulation studies reported in Section 3, we use the sample quantiles from a random sample of size 100 000 from the IC distribution of the $W_{i,t}$ to approximate the $q_{(i),t}$. In the following, we report new results for the simulation study.
from Section 3.1 when we use random samples of smaller sizes to estimate the $q_{(i)t}$. The size of the random IC $W_{t}$ sample used to estimate the $q_{(i)t}$ is represented by $B$ in Table 8. As mentioned earlier, our proposed global monitoring statistic is equivalent to $\sum_{i=1}^{m}(W_{(i)t} - a_{i})^{2}I_{W_{(i)t} > a_{i}}$ with $a_{i}$ being given by the estimate of $q_{(i)t}$ which we obtain offline beforehand. The control limit $h$ of our proposed monitoring scheme is then determined based on those given $a_{i}$ values. With different IC $W_{t}$ sample sizes, the values of the $a_{i}$ specified in the above monitoring statistic will be different, so is the control limit $h$ of our proposed monitoring scheme. The $h$ value corresponding to a specific $B$ value in Table 8 is the control limit we obtain for $ARL_{0} = 1000$ when the $a_{i}$ is given by the estimates of the $q_{(i)t}$ from a random IC $W_{t}$ sample of size $B$. In Table 8, the ARL when $m_{1} = 0$ indicates the simulated ARL$_{0}$. Therefore, as shown in the table, the simulated ARLs of our proposed monitoring schemes with different IC $W_{t}$ sample sizes are all close to the nominal level $ARL_{0} = 1000$. As we can also see from the table, for $m_{1} \neq 0$, the reported ARLs for our proposed monitoring schemes with different IC $W_{t}$ sample sizes are also very similar. From those results, we can see that the size of the random IC $W_{t}$ sample used to estimate the $q_{(i)t}$ has no significant impact on the performance of our proposed monitoring scheme.

As mentioned in Section 1, there are two types of monitoring schemes. For the second type of monitoring scheme, after $G_{t}$ triggers an alarm, we also need to identify which data stream is experiencing abnormal activities. For this purpose, we can use the local monitoring statistics $W_{t}$ to determine which data streams are OC as follows: If $W_{t}$ is larger than some control limit, say $c_{h}$, we conclude that the $i$th data stream is OC; otherwise, we conclude that the $i$th data stream is IC. We refer to Li1 for more details on how to choose the control limit $c_{h}$ in this situation.

Although we only consider three types of local monitoring statistics as examples in the paper, our proposed global monitoring statistic can work with any local monitoring statistic that is efficient for monitoring a single data stream. This

| $m$ | $m_{1}$ | $B = 100000$ | $B = 10000$ | $B = 5000$ | $B = 1000$ |
|-----|---------|-------------|-------------|-------------|-------------|
| 0   | 1008.94 (1044.88) | 994.45 (1021.60) | 1007.48 (1010.02) | 1003.80 (993.95) |
| 1   | 64.44 (32.88) | 65.28 (32.54) | 66.57 (33.63) | 62.58 (31.44) |
| 3   | 36.20 (14.95) | 36.43 (14.64) | 36.65 (14.49) | 36.20 (14.64) |
| 5   | 27.08 (10.53) | 27.16 (10.24) | 26.72 (10.14) | 27.26 (10.28) |
| 8   | 20.33 (7.26) | 20.29 (7.16) | 20.04 (7.45) | 21.01 (7.22) |
| 10  | 17.42 (6.11) | 17.14 (6.30) | 17.38 (6.29) | 18.54 (6.36) |
| 20  | 10.64 (3.67) | 10.46 (3.58) | 10.24 (3.64) | 11.54 (3.72) |
| 50  | 4.88 (1.57) | 4.82 (1.60) | 4.72 (1.49) | 5.54 (1.73) |
| 80  | 3.25 (0.97) | 3.23 (0.96) | 3.16 (0.96) | 3.70 (1.09) |
| 100 | 2.72 (0.77) | 2.65 (0.76) | 2.62 (0.77) | 3.07 (0.86) |

| $m$ | $m_{1}$ | $B = 100000$ | $B = 50000$ | $B = 10000$ |
|-----|---------|-------------|-------------|-------------|
| 0   | 1006.42 (1044.24) | 1004.02 (1008.25) | 994.39 (1024.27) | 996.05 (1005.91) |
| 1   | 81.56 (37.40) | 86.50 (39.92) | 81.69 (37.52) | 79.94 (36.80) |
| 3   | 51.83 (18.91) | 53.07 (19.28) | 51.59 (18.65) | 51.60 (18.31) |
| 5   | 41.89 (13.58) | 42.15 (13.88) | 42.27 (13.91) | 41.59 (13.95) |
| 8   | 33.90 (10.30) | 33.40 (10.55) | 33.75 (10.26) | 34.72 (10.47) |
| 10  | 30.19 (9.86) | 30.18 (9.44) | 30.25 (9.45) | 30.88 (9.51) |
| 20  | 21.04 (6.24) | 21.08 (6.32) | 20.97 (6.10) | 22.04 (6.37) |
| 50  | 11.95 (3.67) | 11.57 (3.47) | 11.88 (3.53) | 12.25 (3.61) |
| 80  | 8.47 (2.54) | 8.19 (2.59) | 8.37 (2.47) | 8.61 (2.58) |
| 100 | 7.05 (2.13) | 6.93 (2.06) | 7.00 (2.10) | 7.17 (2.13) |

| $m$ | $m_{1}$ | $B = 100000$ | $B = 10000$ | $B = 5000$ |
|-----|---------|-------------|-------------|-------------|
| 0   | 5.04 (1.43) | 4.97 (1.47) | 5.02 (1.48) | 5.14 (1.54) |
| 1   | 3.96 (1.13) | 3.84 (1.13) | 3.92 (1.10) | 3.97 (1.12) |
| 3   | 2.78 (0.80) | 2.75 (0.78) | 2.76 (0.77) | 2.79 (0.76) |
| 5   | 2.21 (0.60) | 2.15 (0.58) | 2.18 (0.59) | 2.21 (0.60) |
| 8   | 1.88 (0.46) | 1.82 (0.48) | 1.83 (0.46) | 1.86 (0.46) |
| 10  | 1.67 (0.48) | 1.61 (0.49) | 1.64 (0.49) | 1.67 (0.48) |
| 20  | 1.46 (0.50) | 1.41 (0.49) | 1.43 (0.50) | 1.47 (0.50) |
| 50  | 1.26 (0.44) | 1.20 (0.40) | 1.22 (0.42) | 1.26 (0.44) |
| 80  | 1.09 (0.29) | 1.09 (0.29) | 1.07 (0.26) | 1.11 (0.31) |
| 100 | 1.03 (0.17) | 1.02 (0.14) | 1.02 (0.14) | 1.03 (0.16) |
flexibility makes our proposed global monitoring statistics suitable for many different real-world applications. The simulation studies in the three examples we consider further show that our proposed global monitoring statistic performs well under a variety of OC scenarios and has the best overall detection power comparing with other existing global monitoring statistics.

ACKNOWLEDGEMENTS
The author thanks the editor and two anonymous referees for their constructive comments and suggestions, which greatly improved the quality of the paper.

ORCID
Jun Li https://orcid.org/0000-0002-0323-8255

REFERENCES
1. Li J. A two-stage online monitoring procedure for high-dimensional data streams. J Qual Technol. 2018. https://doi.org/10.1080/00224065.2018.1507562.
2. Tartakovsky AG, Rozovskia BL, Blazeka RB, Kim H. Detection of intrusions in information systems by sequential change-point methods (with discussion). Stat Method. 2006;3:252-340.
3. Mei Y. Efficient scalable schemes for monitoring a large number of data streams. Biometrika. 2010;97:419-433.
4. Xie Y, Siegmund D. Sequential multi-sensor change-point detection. Ann Stat. 2013;41:670-692.
5. Zou C, Jiang W, Wang Z, Zi X. An efficient on-line monitoring method for high-dimensional data streams. Technometrics. 2015;57:374-387.
6. Liu K, Zhang R, Mei Y. Scalable SUM-shrinkage schemes for distributed monitoring large-scale data streams. Stat Sin. 2019;29:1-22.
7. Qiu P. Introduction to Statistical Process Control. Boca Raton: FL: Chapman & Hall/CRC; 2014.
8. Zhang J. Powerful goodness-of-fit tests based on the likelihood ratio. J R Stat Soc Ser B. 2002;64:281-294.
9. Grigg OA, Spiegelhalter DJ. An empirical approximation to the null unbounded steady-state distribution of the cumulative sum statistic. Technometrics. 2008;50:501-511.
10. Sparks RS. CUSUM charts for signalling varying location shifts. J Qual Technol. 2000;32:157-171.
11. Han D, Tsung F. A reference-free cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison. J Am Stat Assoc. 2006;101:368-386.
12. Lorden G, Pollak M. Sequential change-point detection procedures that are nearly optimal and computationally simple. Seq Anal. 2008;27:476-512.
13. Zou C, Tsung F. Likelihood ratio-based distribution-free EWMA control charts. J Qual Technol. 2010;42:174-196.
14. Ross GJ, Adams NM. Two nonparametric control charts for detecting arbitrary distribution changes. J Qual Technol. 2012;44:102-116.
15. Li J. Nonparametric adaptive CUSUM chart for detecting arbitrary distributional changes. Submitted. 2017. (arXiv:1712.05072).

AUTHOR BIOGRAPHY
Jun Li received her PhD in Statistics from the Department of Statistics and Biostatistics at Rutgers University in 2006. Since then, she has been with the Department of Statistics at the University of California, Riverside, as an assistant professor (2006-2012) and an associate professor (2012 to present). She is a lifetime member of the American Statistical Association and the Institute of Mathematical Statistics. Her current research interests include statistical process control and nonparametric multivariate analysis.

How to cite this article: Li J. Efficient global monitoring statistics for high-dimensional data. Qual Reliab Engng Int. 2020;36: 18–32. https://doi.org/10.1002/qre.2557