Journal of Quality Technology

A Quarterly Journal of Methods, Applications and Related Topics

ISSN: 0022-4065 (Print) 2575-6230 (Online) Journal homepage: https://www.tandfonline.com/loi/ujqt20

Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation

Xiaochen Xian, Chen Zhang, Scott Bonk & Kaibo Liu

To cite this article: Xiaochen Xian, Chen Zhang, Scott Bonk & Kaibo Liu (2019): Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation, Journal of Quality Technology, DOI: 10.1080/00224065.2019.1681924

To link to this article: https://doi.org/10.1080/00224065.2019.1681924

Published online: 18 Nov 2019.

JOURNAL OF QUALITY TECHNOLOGY
https://doi.org/10.1080/00224065.2019.1681924

Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation

Xiaochen Xian[a], Chen Zhang[b], Scott Bonk[a], and Kaibo Liu[a]

[a]Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI; [b]Department of Industrial Engineering, Tsinghua University, Beijing, China

ABSTRACT
In many applications of modern quality control, process monitoring involves a large number of process variables and quality characteristics. Practitioners are desired to attain complete information about the process in order to assure quick detection of shifts that may possibly occur at any variable. However, full information is not always available during online monitoring of big data streams due to limitations of monitoring resources in practice. In this paper, a rank-based monitoring and sampling algorithm based on data augmentation is proposed to quickly detect the mean shifts in a process when only a limited portion of observations are available online. Specifically, at each observation time, the proposed method will automatically augment information for unobservable variables based on the online observations, and then intelligently allocate the monitoring resources to the most suspicious data streams. Comparing to the existing literature, this method is able to accurately infer the status of all variables in a process based on a small number of observable variables and effectively construct a global monitoring statistic with the proposed augmented vector, which leads to a quick detection of the out-of-control status even if limited shifted variables are observed in real time. Simulation studies as well as a real case study on real-time solar flare detection are conducted to demonstrate the efficacy and applicability of the proposed method.

KEYWORDS
cumulative sum (CUSUM); data augmentation; partial observations; rank-based monitoring; statistical process control (SPC)

1. Introduction

Statistical Process Control (SPC) has been developed for process monitoring to ensure stable and satisfactory performance in various systems, especially those involving multiple variables. The main objective is to detect any assignable causes as soon as they occur while maintaining a certain global false alarm rate. In the literature, SPC techniques have been widely applied in industrial, clinical, and biological environments (Wang and Jiang 2009; Zou et al. 2015; Zou and Qiu 2009).

As sensor technology advances, big data streams have become commonly available in many modern data-rich applications. Here, big data streams refer to multiple series of *real-time, continuous and sequentially ordered* observations. While receiving increasing attention (Wang and Mei 2015; Zou et al. 2015), effective online monitoring of big data streams for quick change detection is still a challenging task. The challenges are rooted in that the abnormal events are naturally complicated and unknown in advance, and the high volume and high dimensionality of big data streams place an essential demand on online data communication, storage space, memory, computational power and processing time. As a result, we often have to make a decision based on partial observations during online monitoring (Gama and Gaber 2007; Limongelli 2003; Tan et al. 2012). However, the existing SPC literature heavily relies on the assumption that the observations of all variables are fully accessible in real time.

Generally speaking, the inaccessibility to full observations during online monitoring is commonly seen in the following three scenarios (Liu, Mei, and Shi 2015; Xian, Wang, and Liu 2018): (1) due to the limitation of the number of sensors, measurements of only a subset of the interested variables can be obtained at each data acquisition time; (2) due to the limitation of the battery life of sensors, only a subset of sensors can be turned on to collect data in real time; and (3) due to the limitation of communication bandwidth, only a subset of

CONTACT Kaibo Liu ✉ kliu8@wisc.edu Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, 53706, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujqt.

measurements can be transmitted back to the data center for real-time processing.

As a specific example, environmental monitoring usually involves collecting measurements from multiple sensors at various locations to avoid hazardous events such as earthquakes, wildfires, tornadoes, and landslides. To quickly detect such calamities and reduce environmental and economic losses, practitioners are desired to access the full observations of each location at each data acquisition time. However, due to the limited lifetime of the sensor batteries and the high cost of battery replacement, in practice only a subset of the sensors are set to the "ON" mode for surveillance (Sasidhar, Sreeresmi, and Rekha 2014; Xian, Wang, and Liu 2018). In this case, although a large number of sensors are deployed, only partial observations can be actually accessed by practitioners in real time. In many other applications, the number of sensors and data communication bandwidth may also impede the acquisition or access of full observations in real time. For example, when applied in search and rescue, a swarm of unmanned aerial vehicles (UAVs) can only observe a small portion of the interested zones at a time given that the number of UAVs is limited in practice (Waharte and Trigoni 2010). In another example, when monitoring the occurrence of solar flares, only partial image information captured by the satellite can be transmitted back to earth for real-time analysis because of the communication bandwidth constraint, though the full image information can be recorded and available for analysis offline (Liu, Mei, and Shi 2015).

In all of the above examples, only a portion of the data streams are observable at a time due to resource constraints, while assignable causes may possibly happen at any variable from the whole data streams. Below we summarize our interested detection problem mathematically. Suppose we are monitoring p independently and identically distributed (i.i.d.) variables in the set $\mathcal{P} = \{1, 2, \cdots, p\}$, and the measurements of each variable, denoted as $X_j(t)$ $(j = 1, ..., p)$, form a data stream over the observation time $t = 1, 2, \cdots$. The i.i.d. assumption of the data streams, as pointed out by Wang and Mei (2013), is often satisfied when $X_j(t)$'s are selected to be the residuals of some spatial-temporal models. Thus, this assumption has been commonly used in the setting of high-dimensional monitoring literature (Wang and Mei 2015; Zou et al. 2015). $\boldsymbol{X}(t) = (X_1(t), X_2(t), ..., X_p(t))'$ denotes the measurement values of the p variables at time t. Due to resource constraints, only q $(q<p)$ out of the p variables are "observable" at each time t, where q is limited by the availability of monitoring resources in practice and therefore predetermined by the application context. The set of observable variables at time t is denoted as $\mathcal{O}(t)$, and the observed data is denoted as $\boldsymbol{X}^{\mathcal{O}}(t)$. When the process is in control (IC), $\boldsymbol{X}(t)$ is assumed to be i.i.d. across different time points, and each data stream has a general cumulative probability function (CDF) $F(x)$ and a probability density function (PDF) $f(x)$. The distribution functions $F(x)$ and $f(x)$ can be acquired empirically based on the historical IC data in Phase I analysis and this paper focuses on Phase II monitoring. Without loss of generality, $\boldsymbol{X}(t)$ is assumed to have an IC mean $\boldsymbol{\mu} = 0 = (0, 0, \cdots, 0)'$, and each variable has an IC standard deviation of 1. This standardization can be done by pre-centering and scaling the raw data. The process becomes out of control (OC) if at least one of the variables has a mean shift at an unknown change point τ, such that the mean vector of $\boldsymbol{X}(t)$ changes from $\boldsymbol{\mu} = 0$ to $\boldsymbol{\mu}' \neq 0$. To highlight the main idea of the proposed method, we focus on detecting upward mean shifts hereafter. Here it is noteworthy that the change point τ, the post-change mean $\boldsymbol{\mu}'$ and the number of affected variables are unknown. Based on the formulation, the key question is how to intelligently and sequentially decide the most informative variables to observe at each time subject to given resource constraints, in order to quickly detect the process mean shift while maintaining a pre-specified IC average run length (ARL) requirement.

In monitoring big data streams with partial observations, the status of the system depends on both the observed and unobserved variables. Thus, in theory, the detection capability for a process shift can be significantly improved if the status of the unobserved variables is appropriately inferred based on the measurements of observable variables. Specifically, in this paper, we propose to leverage data augmentation to analytically augment the statistics of the unobserved variables, which facilitates the monitoring and sampling of all variables involved in the process. Data augmentation refers to a type of methods via adding information or latent variables to the original "unobservable" or "missing" data, such that the problem becomes tractable (Van Dyk and Meng 2001). While this idea is conceptually sound, how to effectively augment unobservable variables in the context of big data streams during online monitoring is a very challenging problem and has not been explored in the past. First, it is not intuitive to find out the relationship between the observable and unobservable variables, such that the augmentation methods can be effectively integrated with the monitoring and sampling scheme to enhance the performance of SPC. Second, since we only have limited resources whereas the change point as well as the number of shifted variables are unknown, it is possible that only few OC variables are observed simultaneously after the shift occurs in the system. Thus, the augmented values should facilitate the detection of the process shift even when a very small number of OC variables are observed. Third, big data streams may not follow some well-known distributions. As a result, the augmentation method is desired to be not restricted to a certain distribution.

To address the aforementioned challenges, we propose a rank-based CUSUM monitoring and sampling method, called Rank-based Sampling Algorithm by Data Augmentation (R-SADA), to accelerate the detection of process shifts in the context of partial observations by augmenting the unobservable data with the measurements of the observed ones. Comparing to other rank-based methods (Qiu and Hawkins 2001; Xian, Wang, and Liu 2018), the proposed monitoring and sampling method is parametric since it incorporates the PDF and CDF of the process distribution into the dynamic data augmentation step. Such dynamic augmentation using specific domain knowledge makes new methodological contribution and opens new research directions in SPC. In particular, our method still inherits the merits of rank-based methods, i.e., having a robust detection power for general distributions. The main reason we use rank-based statistic is that the rank information of the data streams is naturally dependent, which allows us to dynamically augment useful statistics concerning about the system status based on observed values. Besides, since the rank information of one variable is directly associated with all other variables, the proposed method is sensitive to quickly trigger an alarm even if only a small number of OC variables are observed. Moreover, it will be proved that the proposed method can be effectively integrated with the monitoring and sampling scheme while ensuring two nice properties: (1) the IC property: no variables will be left unobserved for a long time when the process is IC; (2) the OC property: when the process becomes OC, the method tends to keep observing the suspected OC variable, which leads to quick detection of assignable causes.

The remainder of the paper is organized as follows. In Section 2, the literature about data augmentation and related SPC approaches are reviewed, which lay the groundwork for our proposed method. We then introduce the proposed R-SADA method in detail followed by the theoretical investigation of its properties in Section 3. An illustrative example is also provided to demonstrate the proposed method in this section. Numerical simulations and performance comparisons are presented in Section 4. Section 5 contains a real solar flare detection example to illustrate the application of the proposed method. Finally, Section 6 concludes the paper.

2. Data augmentation and related SPC approaches

In this section, we review the literature concerning rank-based and multivariate SPC approaches, and SPC with partial observations, which are closely related to our proposed method.

2.1. Rank-based SPC procedures and multivariate control charts

In this subsection, we focus on the rank-based methods, which are almost exclusively used in nonparametric SPC. The main idea of the rank-based methods is to utilize the rank among the observations instead of the observations themselves, and thus no parametric model is needed. An overview of this topic can be found in Qiu (2013). The literature of rank-based SPC can be classified into two categories according to the number of variables involved. For the univariate cases, see Bakir (2004), Chakraborti, Laan, and Wiel (2004), Chakraborti and Eryilmaz (2007), Chakraborti, Eryilmaz, and Human (2009), Li, Tang, and Ng (2010), and Liu, Tsung, and Zhang (2014) for details.

There are also a few extensions of the rank-based control charts to the multivariate cases. For example, Zou, Wang, and Tsung (2012) employed the multivariate spatial ranks to multivariate exponentially weighted moving average (EWMA) charts. Zi, Zou, and Tsung (2012) proposed a multivariate sign EWMA control scheme using a rank-based regression approach. Qiu and Hawkins (2001, 2003) proposed a multivariate nonparametric CUSUM control chart based on monitoring the anti-rank of variables. In particular, the anti-rank vector of $\boldsymbol{X}(t) = \left(X_1(t),\ X_2(t), \ldots, X_p(t)\right)$, denoted as $\boldsymbol{B}(t) = \left(B_1(t),\ B_2(t), \ldots, B_p(t)\right)$, is a permutation of $(1, 2,\ \cdots, p)'$, such that $X_{B_1(t)}(t) \leq X_{B_2(t)}(t) \leq \cdots \leq X_{B_p(t)}(t)$. Then a statistic $\boldsymbol{\xi}(t) = \left(\xi_1(t), \xi_2(t), \ldots, \xi_p(t)\right)$ can be constructed based on the last anti-rank $B_p(t)$, where

$$\xi_j(t) = \mathbb{I}\left(B_p(t) = j\right). \tag{1}$$

The authors proved that detecting the changes in the distributions of $\boldsymbol{X}(t)$ (with null hypothesis

$H_0 : \mu_1 = \cdots = \mu_p$) is equivalent to detecting the changes in the IC expectation of $\boldsymbol{\xi}(t)$, i.e., $\mathbb{E}(\boldsymbol{\xi}(t)) = (g_1, g_2, ..., g_p) = \boldsymbol{g}$, where g_i is the probability that the i th variable takes the largest value among all the measurements under H_0. Then, Qiu and Hawkins (2001, 2003) suggested a monitoring statistic that depicts the state of the process based on the difference between the observed anti-rank $\boldsymbol{\xi}(t)$ and its IC expectation.

To monitor multiple variables simultaneously and exploit global information, many studies have been developed to construct monitoring statistics from individual local statistics. One most straightforward and popular way is to first calculate a local statistic for each variable by some effective univariate control charts (e.g., CUSUM) and then combine the statistics together for global monitoring. For example, Woodall and Ncube (1985) and Tartakovsky et al. (2006) suggested using the largest local CUSUM statistics as a global monitoring statistic, whereas Mei (2010) proposed to use the summation of all the local statistics. These two schemes, denoted by the $T_{\max}$ and T_{sum}, show advantages in different OC scenarios. Note that the number of variables affected by assignable causes is generally unknown in practice. Thus, the $T_{\max}$ scheme is more sensitive to sparse shifts as its underlying assumption is that only one variable is affected. Nevertheless, the T_{sum} scheme can trigger an alarm much faster when shifts occur at a large number of variables. Furthermore, Mei (2011) proposed to sum the r largest local statistics (named as Top-r) as a tradeoff between $T_{\max}$ and T_{sum}. However, the performance of these monitoring statistics highly depends on the number of OC variables. In other words, these methods may lead to detection delay in our problem of interest, because it is not guaranteed that all OC variables can be observed when monitoring big data streams with partial observations. As a result, a better monitoring scheme is needed to quickly trigger an alarm, which is insensitive to the number of OC variables observed at a time.

2.2. SPC with partial observations

In the literature, there are a few works concerning SPC with partial observations available. These studies can be categorized into two types. The first category focuses on sampling partial observations over the temporal domain, i.e., adaptively adjusting the sampling intervals based on the observed values. This line of research is known as the variable sampling interval (VSI) control chart (Arnold and Reynolds, 2001; Li and Qiu, 2014; Reynolds, Amin, and Arnold 1990). Unfortunately, this strategy falls short in our interested problem here as it still requires full observations of all data streams at each sampling time. To address this issue, the second type of study considers SPC with spatial sampling strategies, i.e., observing only a portion of variables and adaptively determining which variables to observe at each time. In this category, there are two different approaches to deal with unobservable data streams. The first approach is to utilize only the observable variables to construct monitoring statistics. For example, Liu, Tsung, and Zhang (2014) applied this idea to develop an adaptive monitoring strategy in a Bayesian network. However, the proposed method is not scalable to big data streams due to its high computational burden and the prerequisite of the Bayesian network structure. The second approach dealing with unobservable data streams is by data augmentation. Data augmentation, as its name suggested, generally involves "adding more data" especially when there are unobservable values making a problem hard to solve. It has been applied widely in statistical analysis and machine learning applications such as parameter estimation (Dempster, Laird, and Rubin 1977), design of experiments (Little and Rubin 2014), regression (Allen 1974), Monte Carlo simulations (Wei and Tanner 1990) and training data constructions (Lemley, Bazrafkan, and Corcoran 2017).

Concerning data augmentation in the SPC literature, Liu, Mei, and Shi (2015) and Xian, Wang, and Liu (2018) proposed the Top-r Adaptive Sampling (TRAS) and Nonparametric Anti-rank based Sampling (NAS) strategies, respectively. The TRAS algorithm focuses on monitoring big data streams that follow normal distributions. Based on a Top-r CUSUM approach, the TRAS algorithm introduced a constant imputation parameter to the local statistics of the unobservable variables that compensates for the untaken observations. The NAS algorithm then generalized this idea to monitor arbitrarily distributed data streams by incorporating the imputation parameter with a rank-based method. Recently, Wang et al. (2018) proposed a dynamic spatial sampling method that incorporates the spatial information of the shifts when only partial observations are available. This method can also be regarded as an augmentation method as it heuristically augments the likelihood of a variable being OC according to its adjacent variables. From the perspective of data augmentation, all the aforementioned algorithms are equivalent to using an imputation method, in which a heuristic substitution is employed for the missing data (Little and Rubin

2014). More specifically, all the methods considered the imputation parameter to be pre-specified and a constant number. As a result, the estimation of unobserved data is non-informative and non-adaptive to the online measurements, which may introduce biases or uncertainties into the monitoring scheme and eventually degrade the detection performance.

3. Methodology development

In this section, we describe the details of the proposed method. The challenging question is how to utilize the observed information to effectively augment the statistics of the unobservable data and facilitate the detection of OC variables. Our innovative idea is to augment the rank information for the full data streams. The proposed method is inspired by the nonparametric anti-rank procedure in Qiu and Hawkins (2001) and Xian, Wang, and Liu (2018); however, our method is also fundamentally different from the existing ones as we dynamically and analytically augment the information for unobservable variables based on the online observations. While the proposed method can be adjusted to accommodate different distributions, it is a parametric method that requires knowing the underlying IC distribution, as the proposed data augmentation algorithm explicitly uses this underlying distributional information. However, it should be noted that the proposed method does not require the data streams to follow well-known distributions. In practice, practitioners can utilize recorded historical IC data and offline learn the empirical distributions. Since the rank information among the variables is dependent for all data streams, the proposed method can exploit the relationship among the observable and unobservable variables based on the real-time observations. Besides, given that the rank information may change dramatically for all variables even if only one variable has a shift, the proposed method is sensitive to a wide range of process changes, especially when only a small number of shifted variables are observed. Section 3.1 will discuss the details of constructing the augmented vector. Section 3.2 will propose the R-SADA method, and its properties will be investigated in Section 3.3. Then Section 3.4 provides an illustrative example of this proposed method.

3.1. Constructing a dynamic augmented vector based on anti-rank

Recall that $\boldsymbol{X}(t)$ is the vector of the measurement values of p variables at time t, and $\boldsymbol{X}^{\mathcal{O}}(t)$ is the observable subset of $\boldsymbol{X}(t)$. To facilitate monitoring big data streams in the scenario of partial observations, we are interested in constructing a dynamic augmented vector based on the observable data $\boldsymbol{X}^{\mathcal{O}}(t)$. Specifically, our approach is inspired by the anti-rank procedure in Qiu and Hawkins (2001), but we generalize the last anti-rank indicator $\boldsymbol{\xi}(t)$ (see Equation (1)) to the scenario of partial observations. While the generalized anti-rank approach has also been considered in Xian, Wang, and Liu (2018), the difference in our proposed method is that we aim to dynamically and analytically construct the augmented vector rather than heuristically using a constant value for augmentation as in Xian, Wang, and Liu (2018). To begin with, we define the augmented vector as:

$$\boldsymbol{\eta}(t) = \mathbb{E}\left(\boldsymbol{\xi}(t)|\boldsymbol{X}^{\mathcal{O}}(t)\right). \tag{2}$$

This generalization is crucial when there are unobservable variables. In addition, this formulation will allow us to incorporate distributional information of $\boldsymbol{X}(t)$, which will be explained later. Then for each variable $j \in \mathcal{P}$,

$$\eta_j(t) = \mathbb{E}\left(\xi_j(t)|\boldsymbol{X}^{\mathcal{O}}(t)\right) = \mathbb{P}\left(\xi_j(t) = 1|\boldsymbol{X}^{\mathcal{O}}(t)\right). \tag{3}$$

Intuitively, the augmented vector $\boldsymbol{\eta}(t)$ represents the probability that a variable is the largest among all p variables. Then testing the consistency of $\boldsymbol{\eta}(t)$ and its IC mean

$$\mathbb{E}(\boldsymbol{\eta}(t)) = \mathbb{E}\left[\mathbb{E}\left(\boldsymbol{\xi}(t)|\boldsymbol{X}^{\mathcal{O}}(t)\right)\right] = \mathbb{E}(\boldsymbol{\xi}(t)) = \boldsymbol{g} \tag{4}$$

is equivalent to testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = 0. \tag{5}$$

The calculation and augmentation of $\xi_j(t)$ based on Equation (3) highly rely on the true mean vector of $\boldsymbol{X}(t)$, i.e., $\boldsymbol{\mu}(t)$. Without loss of generality, to develop our method, we suppose that one variable has a mean shift in the OC scenario. However, our method is not limited to this assumption as we will show later in the simulation and case studies. Equivalently, the alternative hypothesis can be mathematically written as

$$\begin{aligned} H_1 : \mu_{j_{OC}} = \mu_{OC} > \mu_1 = \cdots = \mu_{j_{OC}-1} \\ = \mu_{j_{OC}+1} = \cdots = \mu_p = 0, \end{aligned} \tag{6}$$

where $j_{OC} \in \mathcal{P}$ is the only OC variable. In practice, the probability that a large number of variables simultaneously go OC is quite low. Therefore, this assumption is reasonable as the shifted variables are usually sparse in big data streams. Similar considerations have also been applied in the SPC literature, such as monitoring high-dimensional process (Wang and Jiang 2009; Zou, Jiang, and Tsung 2011) and sensor allocation (Liu, Zhang, and Shi 2014; Liu and Shi 2013). In practice, the OC shift magnitude μ_{OC} is unknown beforehand. We use a parameter $\mu_{\min} > 0$ to represent the interested-smallest magnitude of mean shifts to be detected (Liu, Mei, and Shi 2015). The effect of this parameter will be further discussed in Section 4.

Suppose $i(t) = \operatorname{argmax}_{j \in \mathcal{O}(t)} X_j(t)$ is the index of the largest observable variable at time t. To construct a dynamic augmented vector that timely reacts to suspicious shifts, the probability $\eta_j(t)$ is updated depending on three different scenarios: (1) $j = i(t)$, (2) $j \in \mathcal{O}(t)$ and $j \neq i(t)$, and (3) $j \notin \mathcal{O}(t)$. When an observed variable j is not among the largest in $\mathcal{O}(t)$ (scenario (2)), $\eta_j(t)$ will be set to zero since it cannot be the largest among all p variables. For $i(t)$ and the unobserved variables, their augmented values are collectively calculated according to the observed values to recognize the system status. Based on the Bayes rule and the sparse shift assumption, the augmented vector $\boldsymbol{\eta}(t)$ in Equation (3) can be calculated as follows:

1. If $j = i(t)$,

$$\eta_j(t) = \frac{F\left(X_{i(t)}(t)\right)^{p-q} \sum_{l \in O(t)} f\left(X_l(t) - \mu_{\min}\right) / f(X_l(t))}{\sum_{l \in O(t)} f\left(X_l(t) - \mu_{\min}\right) / f(X_l(t)) + (p-q)} + \frac{F\left(X_{i(t)}(t)\right)^{p-q-1} F\left(X_{i(t)}(t) - \mu_{\min}\right)(p-q)}{\sum_{l \in O(t)} f\left(X_l(t) - \mu_{\min}\right) / f(X_l(t)) + (p-q)}. \tag{7}$$

2. If $j \in \mathcal{O}(t)$ and $j \neq i(t)$,

$$\eta_j(t) = 0. \tag{8}$$

3. If $j \notin \mathcal{O}(t)$,

$$\eta_j(t) = \frac{\left(1 - F\left(X_{i(t)}(t)\right)^{p-q}\right) \sum_{l \in \mathcal{O}(t)} f\left(X_l(t) - \mu_{\min}\right) / f(X_l(t))}{(p-q)\left(\sum_{l \in \mathcal{O}(t)} f\left(X_l(t) - \mu_{\min}\right) / f(X_l(t)) + (p-q)\right)} + \frac{1 - F\left(X_{i(t)}(t)\right)^{p-q-1} F\left(X_{i(t)}(t) - \mu_{\min}\right)}{\sum_{l \in \mathcal{O}(t)} f\left(X_l(t) - \mu_{\min}\right) / f(X_l(t)) + (p-q)}. \tag{9}$$

The derivation is shown in Appendix A. Please note that $\sum_{j=1}^{p} \eta_j(t) = 1$. Clearly, unlike all the existing literature, the augmented value $\eta_j(t)$ dynamically changes as a function of the observed values $\boldsymbol{X}^{\mathcal{O}}(t)$ instead of being a constant; thus, this dynamic augmented vector is expected to better recognize the real-time status of the system and timely reacts to suspicious shifts. For example, from Equation (9) we can see that when $X_{i(t)}(t)$ is large, the value of $\eta_j(t)$ ($j \notin \mathcal{O}(t)$) gets smaller. This means that the probability of an unobserved variable being the largest gets smaller, which agrees with our expectation. Besides, the augmented value $\eta_j(t)$ relies on all observations $\boldsymbol{X}^{\mathcal{O}}(t)$ to exploit the available information to the maximum extent possible. This procedure is thus fundamentally different from the conventional monitoring schemes like $T_{\max}$ and T_{sum}, in which the individual statistics are "local" statistics in the sense that each of them relies solely on the corresponding data stream.

Thus, the advantage of our proposed scheme is twofold: (1) it allows to incorporate parametric models into a rank-based framework, and inherits its flexibility to deal with various types of distributions; and (2) given that the dynamic augmented vector depends on all observable data streams, the shift occurs at any observable variables will have an influence on the distribution of the entire augmented vector, which will be very beneficial for quick change detection. Specifically, when an OC variable is observed, $\eta_{i(t)}(t)$ tends to increase and $\eta_j(t)$ ($j \notin \mathcal{O}(t)$) decreases simultaneously. Since all the elements in $\boldsymbol{\eta}(t)$ will be changed when the process is OC, it avoids the challenge of choosing an effective combination of local statistics for global monitoring and balancing between $T_{\max}$ and T_{sum}, which is another main advantage over the existing literature. Inspired by this finding, we will further propose the monitoring procedures of the R-SADA control chart and investigate the associated properties in Sections 3.2 and 3.3, respectively. The superiority of the augmented vector $\boldsymbol{\eta}(t)$ will be further demonstrated in Section 3.4, Sections 4 and 5.

It should be noted that we used the last anti-rank indicator $\xi_j(t) = \mathbb{I}\left(B_p(t) = j\right)$ as an example in deriving $\boldsymbol{\eta}(t)$ in this subsection, which naturally leads to a one-sided control chart that is sensitive to the upward mean shift. According to Qiu and Hawkins (2001), the rank-based monitoring procedure is most effective when using the first and last anti-ranks (for detecting downward and upward shifts, respectively). Therefore, in this manuscript, we adopted this conclusion and used the last anti-rank to construct the statistics as a demonstration. However, other types of ranks can be easily constructed in our framework as well with slight modifications. For example, if downward shifts are of interest, we can use $\xi_j(t) = \mathbb{I}\left(B_1(t) = j\right)$ instead. Since the main idea is similar, for simplicity, we choose to only study the last anti-rank indicator $\xi_j(t) = \mathbb{I}\left(B_p(t) = j\right)$ hereafter.

3.2. Rank-based sampling algorithm by data augmentation (R-SADA) control chart

Based on the discussion in Section 3.1, we will detail the monitoring and sampling strategy in this subsection.

At each acquisition time, the dynamic augmented vector $\boldsymbol{\eta}(\boldsymbol{t})$ is constructed based on the partial observations available, which then contributes to a global monitoring statistic. If the monitoring statistic exceeds a control limit subject to a pre-specified IC ARL, the R-SADA control chart triggers an alarm and stops the process. Otherwise, the sampling layout is updated in the next acquisition time. In the followings, we will discuss two major components of the R-SADA method: the monitoring statistic and stopping time, and the sampling strategy.

3.2.1. Monitoring statistic and stopping time

As mentioned in Section 3.1, we need to test the expectation of the augmented vector $\mathbb{E}\boldsymbol{\eta} = \boldsymbol{g}$ regarding the null hypothesis in Equation (5). To that end, we adopt the monitoring statistics of the CUSUM approach in Qiu and Hawkins (2001). First, $\boldsymbol{S}_t^{(1)}$ and $\boldsymbol{S}_t^{(2)}$ are the CUSUM statistics for $\boldsymbol{\eta}(t)$ and $\boldsymbol{g}$, respectively, which are defined as follows:

$$\begin{cases} \boldsymbol{S}_t^{(1)} = g, \ \boldsymbol{S}_t^{(2)} = g & \text{if } C_t \le k, \\ \boldsymbol{S}_t^{(1)} = \left(\boldsymbol{S}_{t-1}^{(1)} + \boldsymbol{\eta}(t)\right)(C_t - k)/C_t & \\ \boldsymbol{S}_t^{(2)} = \left(\boldsymbol{S}_{t-1}^{(2)} + \boldsymbol{g}\right)(C_t - k)/C_t & \text{if } C_t > k. \end{cases} \tag{10}$$

$$C_t = \left(\boldsymbol{S}_{t-1}^{(1)} - \boldsymbol{S}_{t-1}^{(2)} + \boldsymbol{\eta}(t) - \boldsymbol{g}\right)' \cdot diag \left(\left(S_{1,t-1}^{(2)} + g_1\right)^{-1}, \ldots, \left(S_{p,t-1}^{(2)} + g_p\right)^{-1}\right) \cdot \left(\boldsymbol{S}_{t-1}^{(1)} - \boldsymbol{S}_{t-1}^{(2)} + \boldsymbol{\eta}(t) - \boldsymbol{g}\right), \tag{11}$$

where $\boldsymbol{S}_0^{(1)} = \boldsymbol{S}_0^{(2)} = 0$, and k is a constant. As a special case, when $k = 0$, $\boldsymbol{S}_t^{(1)} = \sum_{i=1}^{t} \boldsymbol{\eta}(i)$ is the sum of the augmented vectors, whereas $\boldsymbol{S}_t^{(2)} = t\boldsymbol{g}$ is the expected value of $\boldsymbol{S}_t^{(1)}$ under H_0. C_t is a scalar representing the "distance" between $\boldsymbol{S}_t^{(1)}$ and $\boldsymbol{S}_t^{(2)}$. The quantity k is the allowance of the CUSUM such that $\boldsymbol{S}_t^{(1)}$ and $\boldsymbol{S}_t^{(2)}$ are both reset to the IC expectation $\boldsymbol{g}$ if their distance is less than k. Since all the components of $\boldsymbol{\eta}(t)$ are changed when the system is OC, the monitoring statistic is supposed to consider all of its elements to maximize the detection capability. The monitoring statistic y_t is then defined as

$$y_t = \left(\boldsymbol{S}_t^{(1)} - \boldsymbol{S}_t^{(2)}\right)' diag \left(\frac{1}{S_{1,t}^{(2)}}, \ldots \frac{1}{S_{p,t}^{(2)}}\right) \left(\boldsymbol{S}_t^{(1)} - \boldsymbol{S}_t^{(2)}\right), \tag{12}$$

which is a classic Pearson's χ^2 statistic that measures the statistical difference between $\boldsymbol{S}_t^{(1)}$ and $\boldsymbol{S}_t^{(2)}$ when $k = 0$. Therefore, $y_t > h$ indicates a large difference between $\boldsymbol{S}_t^{(1)}$ and $\boldsymbol{S}_t^{(2)}$ and that the system is OC, where h is a constant threshold of this control chart. The choice of h depends on the pre-scribed IC ARL of the control chart, which can be obtained based on a large quantity of IC data, or using simulation and bootstrap techniques that sample from historical IC data (for details, please refer to Appendix B).

3.2.2. Sampling strategy

Another important factor to ensure the efficiency of the R-SADA method is that the sampling strategy should be able to effectively locate the most suspected variables. On one hand, if an OC variable with an upward shift is observed, it has a larger probability to be the maximum among all variables and thus its augmented value will be larger. On the other hand, if no OC variables are observable, i.e., $X_{i(t)}(t)$ in Equation (20) should not be significantly large, the augmented values $\eta_j(t)$ of the unobservable variables will increase as well. In other words, the augmented values for OC variables are expected to increase no matter if they are observed or not. Therefore, given that $\boldsymbol{S}_t^{(1)}$ is the CUSUM statistics for the augmented vector $\boldsymbol{\eta}(t)$, the variables associated with the large elements of $\boldsymbol{S}_t^{(1)}$ should be paid more attention to. Thus, we update the sampling layout and observe the variables associated with the largest elements of $\boldsymbol{S}_t^{(1)}$ at the next data acquisition time. Recall that the elements of $\boldsymbol{S}_t^{(1)}$ are denoted as $\left(S_{1,\ t}^{(1)}, \ \cdots, S_{p,\ t}^{(1)}\right)'$. Let $j_{(l),t}$ be the variable index of the decreasing order statistic of $\left(S_{1,\ t}^{(1)}, \ \cdots, S_{p,\ t}^{(1)}\right)$, i.e., $S_{j_{(1),t},t}^{(1)} \ge S_{j_{(2),t},t}^{(1)} \ge \cdots \ge S_{j_{(p),t},t}^{(1)}$. Therefore, at time $t+1$, we observe the variables whose indices are in

$$\mathcal{O}(t+1) = \left\{j_{(1),t}, \ldots, j_{(q),t}\right\}. \tag{13}$$

3.3. Properties of the R-SADA control chart

In this section, we will investigate two important properties of the R-SADA control chart under the conditions of the system being IC and OC, respectively. These two properties, namely the IC and OC properties, reveal the sampling and detection capabilities of the R-SADA control chart. Specifically, the first property below studies the sampling layout of the R-SADA control chart when the process is IC.

The IC property: Let U denote the set of variables $i \in \mathcal{P}$ that can never be observed after some finite time t_0, i.e., there exists a time t_0 such that $U = \cap_{t=t_0}^{+\infty} \mathcal{P} \backslash \mathcal{O}(t)$, where $\mathcal{P} \backslash \mathcal{O}(t)$ represents the complement of set $\mathcal{O}(t)$ with respect to $\mathcal{P}$. As $h \to \infty$, $\mathbb{P}(U = \emptyset) \to 1$, where $\emptyset$ represents the empty set. (Proof can be found in Appendix C).

The IC property indicates that the R-SADA control chart keeps all variables in surveillance when the process is IC. In other words, no matter which variable becomes OC, it will not be neglected by the proposed scheme although only a limited number of variables can be observed at each time. Specifically, the R-SADA method suspects the unobservable variables when there is no strong evidence that the observed ones are OC. In other words, it tends to sample the variables that have not been monitored for a while. Next, the OC property below further studies the sampling layout of the R-SADA method when the process is OC.

The OC property: Suppose that $k = 0$ and after time t_0, there is a shift such that $\mathbb{E}(\boldsymbol{\eta}(t)) \neq \boldsymbol{g}$. Let j^* be the OC variable with the largest mean shift such that $\mathbb{E}\left(F\left(X_{j^*}(t)\right)^{p-q-1}F\left(X_{j^*}(t)-\mu_{\min}\right)\right) > \frac{q}{p}$. Then once the variable j^* is observed at time t, there is a nonzero probability that this variable j^* will be kept observed forever, i.e., $j^* \in \mathcal{O}(\tau)$ for any $\tau \geq t$. (Proof can be found in Appendix D).

From the OC property, we proved that at least one of the OC variables with a large shift will always be observed with a nonzero probability once it is observed. The combination of the two properties indicates that when the process is OC, the sampling strategy of the R-SADA control chart will first search for suspicious variables among all variables, and then automatically stick to monitoring the suspicious OC variable once it is found to be highly likely OC. We should mention that from the technical point of view, Liu, Mei, and Shi (2015) and Xian, Wang, and Liu (2018) have also investigated the sampling layout when the process is IC and OC. However, the proof of the two properties in this study is different and much more challenging, as the augment values are dynamically changing with the observations.

3.4. An illustrative example of the R-SADA method

To illustrate the idea of the proposed method step by step, we apply the R-SADA method for a small dataset with 4 variables collected over 10 time points. These variables are randomly generated from the standard normal distribution, except that the variable $X_3(t)$ has a mean shift of magnitude 3 at times $t = 6 \sim 10$. The algorithm is applied with $q = 2$ (two variables can be observed at each time), $k = 0.3$, and $\mu_{\min} = 1.5$. Table 1 shows the values of these variables at each acquisition time (the $X_i(t)$ columns), as well as the evolution of the dynamic augmented vector and monitoring statistic (the $\eta_i(t)$ and $y(t)$ columns). In particular, Table 1 also demonstrates how the sampling layout evolves over time based on the online measurements. The variables being observed at time t are shaded in the $X_i(t)$ columns. At times $t = 1 \sim 5$ when the system is IC, the sampling strategy observes all the variables with similar frequencies. However, after the shift occurs in $X_3(t)$, the R-SADA method observes a very unusual observation $X_3(7) = 3.397$, which indicates that $X_3(t)$ is highly likely the largest variable among all four variables. And thus, all elements in the augmented vector change in time: $\eta_3(t)$ becomes very large while all other elements are almost 0. This change in the augmented vector further leads to two results: (1) the monitoring statistic $y(t)$ has a significant boost as $\boldsymbol{\eta}(t)$ is far from its expectation $\boldsymbol{g}$; (2) The increase in $\eta_3(t)$ leads to an increase in $S_{3,\ t}^{(1)}$, and thus the algorithm tends to keep observing $X_3(t)$ at the following times.

Meanwhile, Table 2 shows the monitoring and sampling procedures of the NAS algorithm proposed in Xian, Wang, and Liu (2018) on the same data. Recall that the NAS algorithm uses fixed imputation parameters to construct the generalized anti-rank indicator. In addition, it utilizes an artificial variable 0 as a reference point. Comparing to the R-SADA method, the NAS algorithm constructs the generalized anti-rank indicator based on only the comparisons among observable variables; however, it does not utilize the actual observed values, and thus may lose some information. For example, $X_2(t)$ is observed more than needed in the IC state as the fixed OC penalty accumulates there, due to the reason that the algorithm only considers the relative rank among observable

Table 1. Illustration data and evolution of statistics by implementing the proposed R-SADA method.

t	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	$\eta_1(t)$	$\eta_2(t)$	$\eta_3(t)$	$\eta_4(t)$	$y(t)$
1	0.015	0.627	0.075	0.352	0.000	0.360	0.320	0.320	0.287
2	−0.697	0.528	0.059	1.797	0.346	0.308	0.000	0.346	0.160
3	0.264	0.872	−1.446	−0.701	0.253	0.494	0.253	0.000	0.238
4	−0.120	−0.639	0.577	−0.360	0.078	0.000	0.461	0.461	0.057
5	−0.136	−1.349	−1.270	0.267	0.404	0.404	0.000	0.192	0.062
6	−0.045	−0.799	2.235	0.862	0.096	0.000	0.452	0.452	0.101
7	−0.056	0.514	3.397	0.756	0.000	0.000	1.000	0.000	1.339
8	0.401	−1.341	3.375	1.125	0.000	0.000	1.000	0.000	3.526
9	0.729	−2.378	2.726	−0.323	0.004	0.004	0.992	0.000	6.008
10	0.318	−0.511	2.998	1.607	0.002	0.002	0.996	0.000	8.667

Table 2. Illustration data and evolution of statistics by implementing the NAS method.

t	$X_1(t)$	$X_2(t)$	$X_3(t)$	$X_4(t)$	0	$\xi_1(t)$	$\xi_2(t)$	$\xi_3(t)$	$\xi_4(t)$	$\xi_5(t)$	$y(t)$
1	0.015	0.627	0.075	0.352	0	0.000	0.450	0.300	0.300	0.000	0.241
2	−0.697	0.528	0.059	1.797	0	0.300	0.450	0.000	0.300	0.000	0.390
3	0.264	0.872	−1.446	−0.701	0	0.300	0.450	0.300	0.000	0.000	0.524
4	−0.120	−0.639	0.577	−0.360	0	0.000	0.000	0.300	0.300	0.250	0.133
5	−0.136	−1.349	−1.270	0.267	0	0.300	0.000	0.000	0.300	0.250	0.217
6	−0.045	−0.799	2.235	0.862	0	0.300	0.000	0.300	0.450	0.000	0.136
7	−0.056	0.514	3.397	0.756	0	0.000	0.300	0.300	0.450	0.000	0.203
8	0.401	−1.341	3.375	1.125	0	0.300	0.300	0.450	0.000	0.000	0.025
9	0.729	−2.378	2.726	−0.323	0	0.300	0.300	0.450	0.000	0.000	0.209
10	0.318	−0.511	2.998	1.607	0	0.300	0.000	0.450	0.300	0.000	0.390

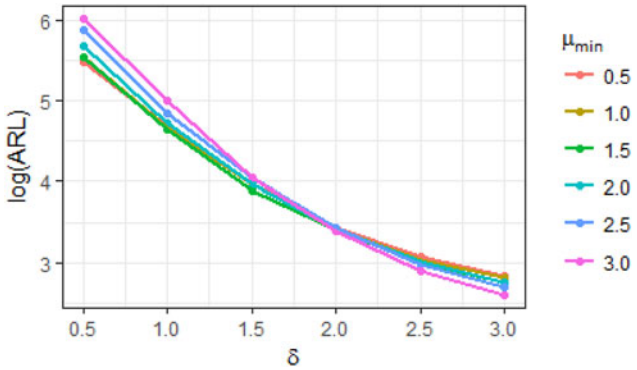

Figure 1. Comparison of OC ARLs for different values of the parameter $\mu_{\min}$.

variables although the actual observations of $X_2(t)$ are not too suspicious on an absolute scale of the distribution. For the same reason, the true OC variable $X_3(t)$ is observed at a later time point than the R-SADA method. Even after $X_3(t)$ is observed and found to be suspicious at and after time 8, the increase in the monitoring statistic $y(t)$ is less noticeable since it uses fixed imputation parameters in the generalized anti-rank indicator. As a result, this example clearly illustrates the superiority of the R-SADA strategy for mean shift detection.

4. Simulation results

In the simulation study, we evaluate the performance of the proposed R-SADA method and compare it with some representative baseline methods. Throughout this section, we consider in total $p = 100$ i.i.d. variables. For the OC cases, n variables are randomly selected to have an upward mean shift of magnitude δ. Throughout this section, the IC ARL is set to be 370, and all the results are based on 5000 simulation runs.

4.1. Parameter $\mu_{\min}$ of the R-SADA method

First of all, we observe how the parameter $\mu_{\min}$ affects the performance of the proposed method. Recall that $\mu_{\min}$ is a unique parameter in the proposed method representing the interested-smallest mean shift, as a replacement for the unknown true mean shift μ_{OC}. In this simulation, each variable independently follows the standard normal distribution $N(0, 1)$. There are $q = 10$ observable variables at each time, and only one variable has a mean shift in the OC scenario $(n = 1)$. Figure 1 shows the comparison between different values of $\mu_{\min}$ under different mean shifts δ. In particular, the x-axis of this figure represents the magnitude of the mean shifts, and the y-axis represents the log transformation of the OC ARL (ARL_1) for the proposed R-SADA method under different OC scenarios. As shown in Figure 1, ARL_1 decreases as the magnitude of mean shifts increases. This agrees with our expectation as larger mean shifts lead to more significant changes in the monitoring statistics and thus result in a quicker detection of the OC status. There are 6 curves in Figure 1, representing the log of ARL_1 of the R-SADA method under the selections of $\mu_{\min} = 0.5, 1.0, \cdots, 3.0$, respectively. It can be observed that a smaller $\mu_{\min}$ works better for smaller mean shifts, and a larger $\mu_{\min}$ helps detect larger mean shifts more quickly. For example, using $\mu_{\min} = 3.0$ leads to the quickest detection for the mean shift $\delta = 3.0$ among all the considered values of $\mu_{\min}$. However, it also results in the largest ARL for small mean shifts such as $\delta = 0.5$. When there is a large shift, the OC variable with a large mean shift can be quickly noticed once it is observed by the limited resources. Thus, a large value of $\mu_{\min}$ helps find these OC variables more quickly since it enables reallocating the resources more frequently to the unobservable variables. For small shifts, it requires the OC variables to be observed for a longer time before raising an alarm. Thus, a smaller value of $\mu_{\min}$ allows the resources to be reallocated less frequently and concentrates more on the suspicious variables. Figure 1 shows that practitioners should appropriately choose the value of $\mu_{\min}$ based on the prior knowledge of the shifts and the application context. If such information is unknown before monitoring the process, we recommend using a

Table 3. ARL_1 and corresponding standard errors (values in the parenthesis) of the R-SADA, TRAS, top-r CUSUM, and NAS algorithms under different combinations of q, n, and δ for normal distribution.

		$n = 5$				$n = 10$			
		R-SADA	TRAS	top-r CUSUM	NAS	R-SADA	TRAS	top-r CUSUM	NAS
$q = 10$									
	$\delta = 1.0$	36.1 (.46)	20.0 (.11)	9.08 (.05)	40.33 (.50)	21.8 (.27)	14.0 (.07)	6.26 (.02)	26.2 (.46)
	$\delta = 2.0$	7.09 (.08)	8.66 (.05)	3.32 (.01)	9.07 (.07)	4.63 (.05)	6.48 (.02)	2.58 (0.01)	7.49 (.06)
	$\delta = 3.0$	3.75 (.04)	6.71 (.03)	2.09 (.00)	7.66 (.05)	2.51 (.02)	5.00 (.02)	1.97 (0.00)	5.63 (.03)
$q = 20$									
	$\delta = 1.0$	12.1 (.20)	12.2 (.08)	9.08 (.05)	14.65 (.29)	7.85 (.12)	8.08 (.04)	6.26 (.02)	12.2 (.25)
	$\delta = 2.0$	3.23 (.04)	5.39 (.02)	3.32 (.01)	6.16 (.09)	2.20 (.03)	4.07 (.01)	2.58 (0.01)	5.32 (.07)
	$\delta = 3.0$	1.88 (.02)	4.24 (.02)	2.09 (.00)	4.58 (.06)	1.46 (.01)	3.25 (.01)	1.97 (0.00)	4.23 (.06)
$q = 30$									
	$\delta = 1.0$	10.2 (.16)	10.6 (.06)	9.08 (.05)	13.6 (0.23)	6.46 (.10)	6.96 (.03)	6.26 (.02)	11.5 (.19)
	$\delta = 2.0$	2.83 (.03)	4.68 (.02)	3.32 (.01)	5.45 (.05)	2.08 (.02)	3.61 (.01)	2.58 (.01)	4.25 (.05)
	$\delta = 3.0$	1.71 (.02)	3.68 (.01)	2.09 (.00)	3.74 (.04)	1.42 (.01)	2.90 (.01)	1.97 (.00)	3.06 (.03)

Table 4. ARL_1 and corresponding standard errors (values in the parenthesis) of the R-SADA, TRAS, top-r CUSUM, and NAS algorithms under different combinations of q, n, and δ for t(3) distribution.

		$n = 5$				$n = 10$			
		R-SADA	TRAS	top-r CUSUM	NAS	R-SADA	TRAS	top-r CUSUM	NAS
$q = 10$									
	$\delta = 1.0$	39.5 (.62)	31.6 (.19)	13.3 (.09)	38.6 (.47)	29.4 (.53)	21.9 (.11)	8.14 (.04)	32.4 (.46)
	$\delta = 2.0$	9.90 (.09)	12.4 (.05)	5.20 (.02)	15.2 (.14)	7.58 (.07)	9.19 (.03)	3.84 (.01)	13.3 (.12)
	$\delta = 3.0$	4.47 (.05)	9.78 (.04)	4.22 (.01)	10.5 (.09)	3.33 (.03)	7.14 (.03)	3.24 (.01)	9.66 (.08)
$q = 20$									
	$\delta = 1.0$	28.5 (.55)	21.3 (.13)	13.3 (.09)	32.5 (.54)	19.1 (.34)	14.4 (.08)	8.14 (.04)	21.9 (.44)
	$\delta = 2.0$	5.11 (.06)	8.38 (.03)	5.20 (.02)	10.2 (.09)	3.93 (.05)	6.24 (.02)	3.84 (.01)	8.72 (.06)
	$\delta = 3.0$	2.52 (.03)	6.56 (.02)	4.22 (.01)	8.29 (.06)	2.16 (.02)	4.96 (.02)	3.24 (.01)	6.00 (.04)
$q = 30$									
	$\delta = 1.0$	22.3 (.46)	18.8 (.11)	13.3 (.09)	24.2 (.45)	16.8 (.30)	12.6 (.06)	8.14 (.04)	18.9 (.46)
	$\delta = 2.0$	5.01 (.06)	7.34 (.03)	5.20 (.02)	8.63 (.07)	4.02 (.05)	5.54 (.02)	3.84 (.01)	7.00 (.06)
	$\delta = 3.0$	2.46 (.03)	5.72 (.02)	4.22 (.01)	6.49 (.05)	2.05 (.02)	4.42 (.01)	3.24 (.01)	5.60 (.03)

moderate value, e.g., $\mu_{\min} = 1.5$ as it has a relatively good performance for different magnitudes of mean shifts. As a result, $\mu_{\min}$ is selected to be 1.5 in the simulations below.

4.2. Monitoring normal data

In this simulation, we compare the R-SADA method to several popular baseline methods to thoroughly understand its performance. To be specific, the following methodologies are considered: (1) the proposed R-SADA method, (2) the TRAS algorithm which is based on the conventional univariate CUSUM statistics and a constant imputation parameter to deal with partial observations, (3) the top-r CUSUM procedure proposed in Mei (2011) which assumes all variables are observable, and (4) the NAS algorithm which is nonparametric and also uses a constant imputation parameter for tackling partial observations. Please note that only the top-r CUSUM procedure assumes full observations among the four competing methods. Besides, the random sampling method is not considered here as an additional benchmark since it has been demonstrated to be inferior to the TRAS and NAS algorithms. We compare the OC ARLs (ARL_1) under different combinations of the number of observable variables q, the number of shifted variables n, and the magnitude of mean shift δ. The following value combinations are considered in this simulation: $q = 10,\ 20,\ 30,\ \ n = 5,\ 10,$ and $\delta = 1.0,\ 2.0,\ 3.0$. For the R-SADA and top-r CUSUM algorithms, we set $\mu_{\min} = 1.5$. For the TRAS procedure, we set $\mu_{\min} = 1.5$, and the constant imputation parameter $\Delta = 0.10$, which are recommended in Liu, Mei, and Shi (2015). Moreover, the number of summands r in the top-r monitoring statistic is chosen to be the same as the number of OC variables n in the TRAS and top-r CUSUM procedures for all OC scenarios, which means $r = n$ for all cases shown in Tables 3–5. This parameter setting, as mentioned in Mei (2011), leads to the best performance in these benchmark methods. However, this comparison is a little unfair for our proposed method, as in practice we commonly do not know the number of OC variables and thus how to decide the appropriate number of summands in the benchmark methods can be quite

Table 5. ARL_1 and corresponding standard errors (values in the parenthesis) of the R-SADA, TRAS, top-r CUSUM, and NAS algorithms under different combinations of q, n, and δ for Poisson distribution.

		$n=5$				$n=10$			
		R-SADA	TRAS	Top-r CUSUM	NAS	R-SADA	TRAS	Top-r CUSUM	NAS
$q=10$									
	$\delta=1.0$	22.9 (.48)	18.8 (.12)	10.99 (.05)	27.7 (.61)	13.4 (.24)	13.0 (.07)	7.52 (.03)	25.8 (.59)
	$\delta=2.0$	5.47 (.07)	9.23 (.05)	5.10 (.01)	11.1 (.10)	3.54 (.04)	6.74 (.03)	3.90 (.01)	9.71 (.09)
	$\delta=3.0$	2.74 (.03)	6.93 (.04)	3.48 (.01)	7.88 (.05)	1.81 (.02)	5.00 (.02)	2.94 (.00)	6.92 (.04)
$q=20$									
	$\delta=1.0$	12.1 (.22)	14.7 (.08)	10.99 (.05)	19.4 (.49)	7.39 (.12)	9.99 (.05)	7.52 (.03)	12.3 (.31)
	$\delta=2.0$	3.30 (.04)	7.05 (.03)	5.10 (.01)	7.37 (.11)	2.22 (.02)	5.22 (.02)	3.90 (.01)	6.49 (.09)
	$\delta=3.0$	1.68 (.01)	5.24 (.02)	3.48 (.01)	5.76 (.07)	1.25 (.01)	3.92 (.01)	2.94 (.00)	4.38 (.06)
$q=30$									
	$\delta=1.0$	11.8 (.20)	13.0 (.07)	10.99 (.05)	16.8 (.34)	7.35 (.11)	8.96 (.04)	7.52 (.03)	10.3 (.18)
	$\delta=2.0$	3.12 (.03)	6.17 (.02)	5.10 (.01)	6.59 (.07)	2.08 (.02)	4.72 (.01)	3.90 (.01)	5.43 (.07)
	$\delta=3.0$	1.51 (.01)	4.54 (.01)	3.48 (.01)	5.16 (.04)	1.17 (.01)	3.53 (.01)	2.94 (.00)	4.27 (.03)

challenging. For the NAS method, the parameters are chosen as: $\Delta = 0.0105$ for $q = 10$, $\Delta = 0.0120$ for $q = 20$, and $\Delta = 0.0135$ for $q = 30$. The results of this simulation are summarized in Table 3.

From Table 3, it can be seen that larger n and larger δ both lead to shorter ARL_1 for all the four methods. This is because n and δ reflect the severity level of OC scenarios, and a higher severity level of OC scenarios is more noticeable by monitoring schemes. Furthermore, the number of observable variables q has a significant impact on the R-SADA, TRAS, and NAS algorithms as they all assume only q variables are available at each time. The larger q is, the more information we can obtain from the process, which thus leads to quicker detection. On the contrary, since the top-r CUSUM procedure requires that all the variables are observable, the ARL_1's of the top-r CUSUM procedure are invariant of the values of q.

We first compare the differences between the R-SADA and the TRAS algorithms. For small q ($q = 10$) and small mean shifts ($\delta = 1.0$), the TRAS algorithm shows better performance. This is because the TRAS algorithm is based on the conventional CUSUM scheme, which has an advantage in detecting small mean shifts. However, the R-SADA method is based on the order statistics, and thus the change in the augmented vector $\boldsymbol{\eta}(t)$ is less sensitive when the mean shift is small. In particular, when the monitoring resources are limited, an OC variable with a small mean shift can hardly be observed for a long time, and the increase in $\boldsymbol{\eta}(t)$ for the observed OC variables may not be significant to raise an alarm. When the monitoring resources are relatively adequate ($q = 20$ and 30), it allows the OC variables with small mean shifts ($\delta = 1.0$) to be observed for a longer time, and thus the R-SADA and TRAS algorithms have very similar performances in these scenarios. When the OC variables have moderate to large mean shifts ($\delta = 2.0$ and 3.0), the proposed R-SADA method shows significant advantages compared to TRAS algorithm. This can be explained by the framework of the R-SADA method that once an OC variable is observed, the distributions of all elements of $\boldsymbol{\eta}(t)$ are changed according to the online observations. Therefore, the R-SADA method can quickly trigger an alarm even though only few OC variables with moderate to large shift are observed.

The comparison between the R-SADA method and the top-r CUSUM algorithm is also very interesting. When there are only $q = 10$ observable variables, the top-r CUSUM algorithm significantly outperforms the R-SADA method since it assumes full observations of the process, which is consistent with our expectation. When more monitoring resources are available ($q = 20$ and 30), the performance of the R-SADA method is a little worse than the top-r CUSUM algorithm for small mean shifts ($\delta = 1.0$); nonetheless, for moderate to large mean shifts ($\delta = 2.0$ and 3.0), the R-SADA method performs very similar to, or even slightly better than the top-r CUSUM algorithm. This is surprising since the R-SADA method requires only 20% ($q = 20$) or 30% ($q = 30$) of the monitoring resources that the top-r CUSUM algorithm utilizes, but it achieves better performance. Recall that the two algorithms are based on two different schemes when constructing the monitoring statistics. As the R-SADA method depends on the rank information, the augmented vector $\boldsymbol{\eta}(t)$ is naturally more sensitive to large mean shifts, i.e., the distribution of all elements in $\boldsymbol{\eta}(t)$ quickly changes as the mean shifts occur, which thus triggers an alarm more quickly. In contrast, the disadvantage of the top-r CUSUM procedure can be explained in the following two aspects: (1) from the perspective of local statistics, it is well-known that the top-r CUSUM scheme is more efficient to detect small shifts compared to large shifts; (2) from the perspective of constructing global monitoring statistics, the top-r scheme takes only top r summands that may be related to the underlying shifts in the monitoring statistics (while the R-SADA method takes all p summands), and thus cannot react to large shifts as quickly as the R-SADA method.

Finally, the NAS algorithm has the worst overall performance because it is a nonparametric method and thus does not utilize the distributional information. Though it also considers the rank-based approach, the NAS algorithm introduces a constant imputation parameter for unobservable variables. In contrast, the R-SADA method dynamically determines the augmented vector and thus achieves a better monitoring performance.

4.3. Monitoring t-distributed data

We then study the performance of these methods when the variables follow non-normal distributions. In this study, we consider t distributions with the degree of freedom 3. Accordingly, the augmented vectors or local statistics of the three parametric methods (the R-SADA, TRAS and top-r CUSUM algorithms) are designed based on the likelihood of $t(3)$ distribution. The parameters for the R-SADA and top-r CUSUM procedures are selected as $\mu_{\min} = 1.5$. The parameters for the TRAS algorithm are selected as $\mu_{\min} = 1.5$, and $\Delta = 0.10$. For the NAS method, the parameters are chosen as: $\Delta = 0.0105$ for $q = 10$, $\Delta = 0.0120$ for $q = 20$, and $\Delta = 0.0135$ for $q = 30$. The results of this simulation are summarized in Table 4.

The results in Table 4 are similar to those in Table 3 where variables follow normal distributions. Though the performance for small shifts is compromised as there are more outliers for t distributions, the R-SADA method still demonstrates a significant advantage for detecting moderate to large mean shifts.

4.4. Monitoring poisson data

In this subsection, we consider the Poisson distributions with parameter $\lambda = 20$. Accordingly, the augmented vectors or local statistics of the three parametric methods (the R-SADA, TRAS and top-r CUSUM algorithms) are designed based on Poisson likelihood. The parameters for the R-SADA and top-r CUSUM procedures are selected as $\mu_{\min} = 1.5$. The parameters for the TRAS algorithm are selected as $\mu_{\min} = 1.5$, and $\Delta = 0.10$. For the NAS method, the parameters are chosen as: $\Delta = 0.0105$ for $q = 10$, $\Delta = 0.0120$ for $q = 20$, and $\Delta = 0.0135$ for $q = 30$. The results of this simulation are summarized in Table 5.

The results in Table 5 are also very similar to those in Table 3 where variables follow normal distributions. In other words, the R-SADA method still demonstrates a significant advantage for detecting moderate to large mean shifts, or when more monitoring resources are available ($q = 20$ and 30).

5. Case study

In this section, we will present a case study of the proposed R-SADA method using an example of solar flare detection. Solar flares are sudden releases of energy at the surface of the sun. When the solar flare occurs, the sun ejects electromagnetic waves to the vicinity of the earth, and the particles involved can be hazardous to spacecraft, satellites, astronauts in space, as well as terrestrial facilities like electrical grids. To avoid these harmful effects, it is significantly important to monitor the process and trigger an alarm as soon as the solar flare occurs. To monitor solar flares in practice, satellites with cameras in space take consecutive optical observations, which can produce a large number of solar images (observations) every second, resulting in about 1.5 TB of data every day. While such big data are recorded and available for analysis offline, the satellites are only able to send partial observations back to earth for real-time analysis due to the limited transmission rate (Aschwanden et al. 2013). In other words, conventional monitoring procedures which assume full observations of the data streams cannot be applied here to online detect the occurrence of solar flares. Specifically, this study considers that only q variables (pixels of each image) are observable at each data acquisition time.

In this example, the longitudinal data of each pixel in the captured images can be regarded as one data stream. The dataset includes in total $232 \times 146 = 33872$ pixels and $T = 250$ frames. Two apparent solar flares are observed at time $t = 137 \sim 152$ and $t = 166 \sim 218$. The background information of these data has been removed by pre-centering and pre-scaling using the mean and standard deviation of the data in the time $t = 1 \sim 50$. The residuals can be regarded as approximately normally distributed (Xie, Huang, and Willett 2013). To apply online monitoring algorithm using partial observations, we assume that there are $q = 500$ and $q = 1000$ monitoring resources available, respectively. We then apply the R-SADA and TRAS algorithms to detect solar flare occurrences and compare their performance. For both two algorithms,

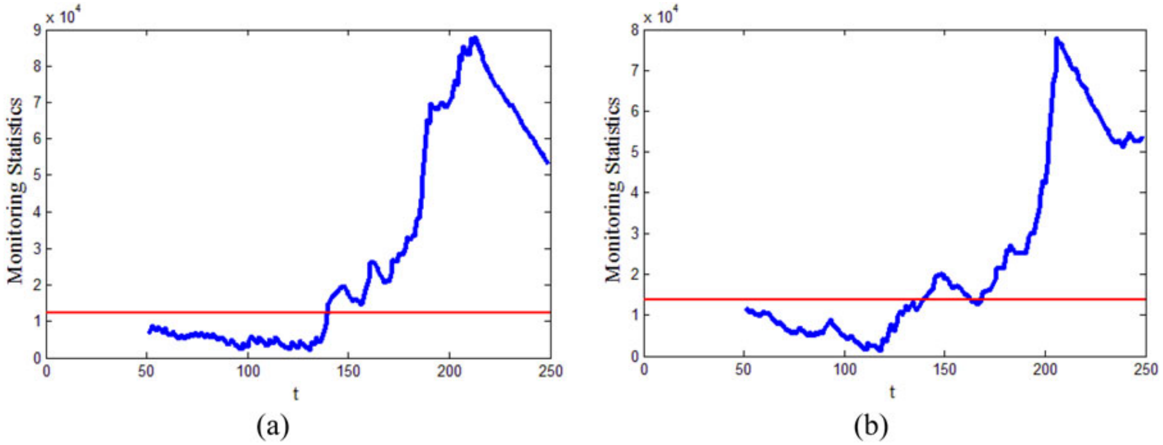

Figure 2. Monitoring statistics for the R-SADA method, when (a) $q = 500,$ and (b) $q = 1000.$

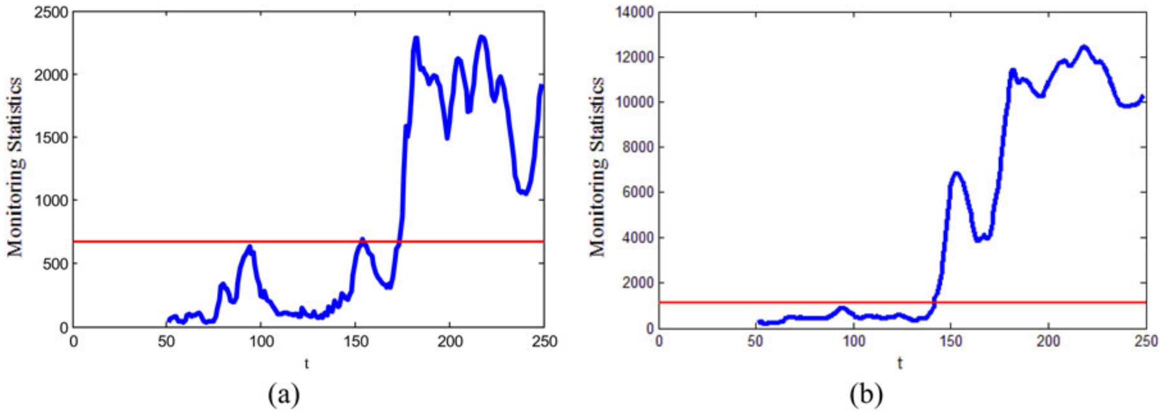

Figure 3. Monitoring statistics for the TRAS algorithm, when (a) $q = 500,$ and (b) $q = 1000.$

the parameter $\mu_{\min}$ is set as 4 since the occurrence of the solar flare incurs large mean shifts. Besides, the imputation parameter Δ for the TRAS algorithm is set to be 0.05 for $q = 500$ and 0.10 for $q = 1000,$ as they lead to the optimal performance compared to other choices. To set control limits for the two control charts, the data for time frames $t = 51 \sim 100$ are bootstrapped to determine the control limits h corresponding to an IC ARL $ARL_0 = 1000.$

The monitoring statistics for the solar flare detection of the R-SADA and TRAS algorithms are shown in Figures 2 and 3, respectively. The plots (a) and (b) in these two figures correspond to the scenarios of $q = 500$ and $q = 1000.$ When $q = 1000,$ the two algorithms have similar performances: the R-SADA method detects the occurrence at $t = 140,$ and the TRAS algorithm detects the occurrence at $t = 142.$ However, when $q = 500,$ the performance of the TRAS algorithm deteriorates fast and detects the shift at $t = 154$; On the contrary, the R-SADA method detects the shift at $t = 142.$ Since the solar flares incur very large mean shifts, the proposed R-SADA method has an obvious advantage in detecting such mean shifts though the number of observable variables is only 500. This result is consistent with the ones in the simulations.

To provide more insights on the advantage of the R-SADA method, Figure 4 shows a list of heat maps of the original images, and the augmented vector and local statistics of the R-SADA and TRAS algorithms. In these heat maps, the data values are represented in different colors, where smaller values are closer to red and larger values are brighter and closer to yellow. Figures (a) – (c) correspond to the process when $t = 120$ and the process is IC, whereas Figures (d) – (f) show the process when $t = 150$ and the solar flare occurs. In particular, Figures (a) and (d) are the heat maps of the original solar data, where an apparent solar flare can be observed in Figure (d). Figures (b) and (e) show the heat maps of the augmented vector in the R-SADA method at $t = 120$ and $t = 150.$ It should be noted that these two heat maps are plotted in the same color range. At time $t = 120,$ there is no obvious indication of the process being OC as the elements of the augmented vector are not far from their

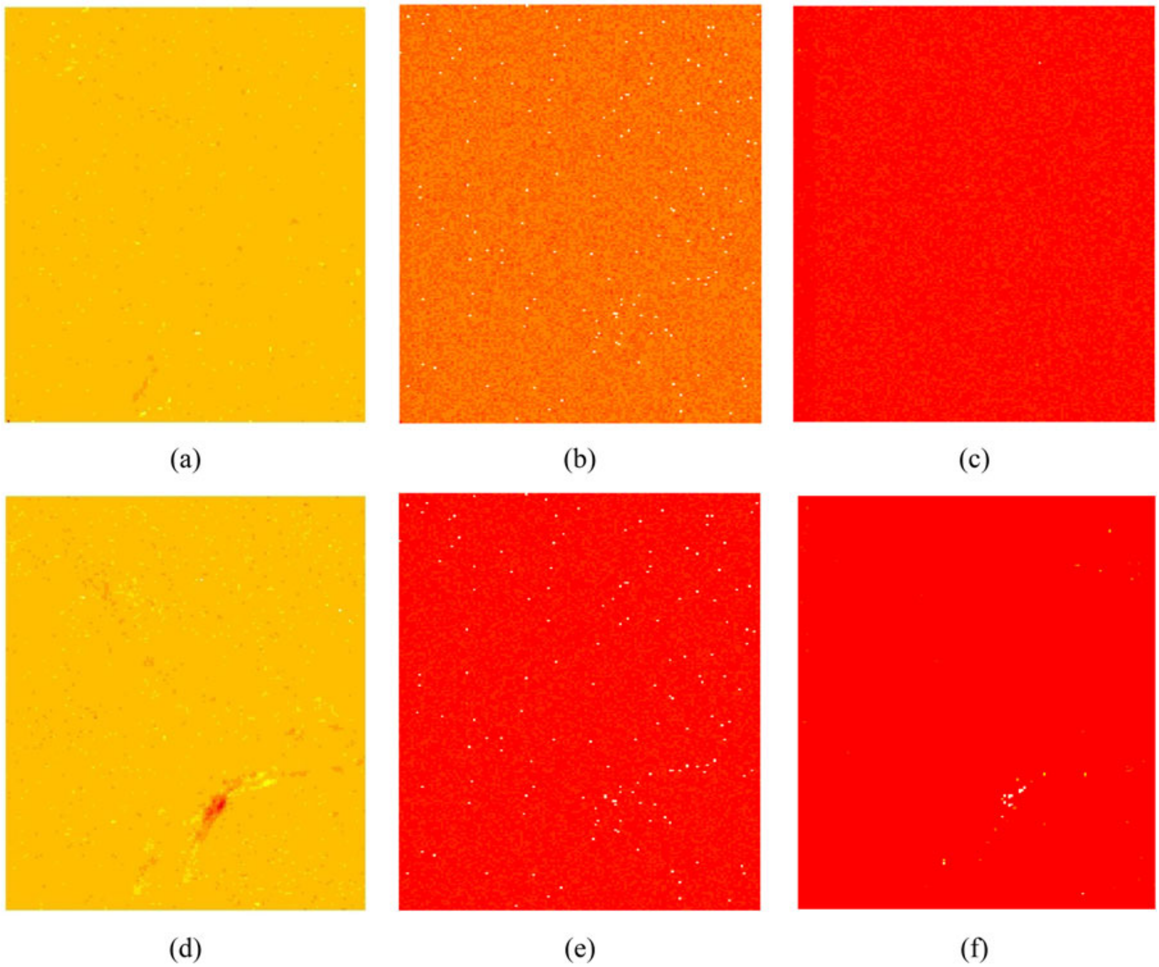

Figure 4. Heat map of the solar data, the augmented vector and local statistics for the R-SADA and TRAS algorithms when $q = 1000$. Figures (a) – (c) shows the original solar data, the augmented vector for the R-SADA method, and the local statistics for the TRAS algorithm at time $t = 120,$ respectively. Figures (d) – (f) shows the original solar flare data, the augmented vector for the R-SADA method, and the local statistics for the TRAS algorithm at time $t = 150,$ respectively.

IC means. However, when $t = 150,$ all the elements of the augmented vector decrease dramatically for the unobservable variables, as we can see that most pixels in Figure (e) get closer to red. In addition, at $t = 150,$ the maximum value of the variables that the R-SADA method observes is 8.013. The observation of such an OC variable leads to the significant changes in the entire augmented vector, which contributes to a quicker OC alarm. In contrast, Figures (c) and (f) show the heat maps of the local statistics for the TRAS algorithm when $t = 120$ and $t = 150$. It can be observed that only the statistics of the observed OC variables have significant changes when the solar flare occurs. As a result, Figure 4 clearly highlights the main idea and the advantage of the proposed R-SADA method.

6. Conclusions

In recent years, resource constraints have been well recognized as an essential challenge during online monitoring of big data streams by practitioners. Without access to full information about the process, it is critically important to effectively allocate monitoring resources and determine the status of the process, since the OC scenarios can be very complicated and hard to be noticed. In this paper, a rank-based sampling algorithm based on data augmentation is proposed to quickly detect the mean shifts in a process when only a limited portion of observations are available at each acquisition time. Specifically, our novel idea is to dynamically and analytically augment the unobservable data based on the rank information of the full observation, which facilitates the allocation of monitoring resources to OC variables and quick detection of process shift. With the dynamic augmentation, the proposed R-SADA method can simultaneously change all the elements of the augmented vector even when only limited OC variables are observed at a time. This nice characteristic allows us to leverage all the elements of the augmented vector to construct an effective global monitoring statistic, and thus quickly detect the OC status. Two theoretical properties are also investigated for the proposed R-SADA method, which guarantees its sampling efficiency in both the IC and OC scenarios. Simulation studies and a real case study on real-time solar flare detection are conducted to demonstrate the advantage of the proposed method.

There are some topics to be further studied related to the proposed R-SADA method. For example, the proposed method is less sensitive to small mean shifts. A better strategy of augmenting the unobservable variables for small mean shifts detection is desired to be explored. Besides, how to generalize the augmentation method, e.g., by relaxing the i.i.d. assumption of the variables, needs further investigation. Last but not least, it will be very interesting to theoretically investigate the detection delay of the proposed monitoring and adaptive sampling strategy under various OC scenarios.

About the authors

Dr. Xiaochen Xian is an assistant professor in the Department of Industrial and Systems Engineering, University of Florida. Her email address is xxian@ufl.edu.

Dr. Chen Zhang is currently an assistant professor in the Department of Industrial Engineering, Tsinghua University. She is a member of ASQ. Her email address is zhangchen01@tsinghua.edu.cn.

Mr. Scott Bonk is a product manager at Belvedere Trading, LLC. His email address is sbonkers33@gmail.com.

Dr. Kaibo Liu is an associate professor in the Department of Industrial and Systems Engineering, University of Wisconsin–Madison. He is a member of ASQ. His email address is kliu8@wisc.edu.

Acknowledgments

The authors thank two anonymous reviewers who provided helpful comments on the manuscript that led to significant improvement of the article.

Funding

The authors gratefully acknowledge the support provided in part by the National Science Foundation under grant NSF CMMI-1362529, the 3M, and the Air Force Office of Scientific Research under award number FA9550-18-1-0145.

References

Allen, D. M. 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16 (1):125–7. doi: 10.1080/00401706.1974.10489157.

Arnold, J. C., and M. R. Reynolds. 2001. CUSUM control charts with variable sample sizes and sampling intervals. *Journal of Quality Technology* 33 (1):66–81. doi: 10.1080/00224065.2001.11980048.

Aschwanden, M. J., P. Boerner, C. J. Schrijver, and A. Malanushenko. 2013. Automated temperature and emission measure analysis of coronal loops and active regions observed with the atmospheric imaging assembly on the solar dynamics observatory (SDO/AIA). *Solar Physics* 283 (1):5–30. doi: 10.1007/s11207-011-9876-5.

Bakir, S. T. 2004. A distribution-free Shewhart quality control chart based on signed-ranks. *Quality Engineering* 16 (4):613–23. doi: 10.1081/QEN-120038022.

Chakraborti, S., and S. Eryilmaz. 2007. A nonparametric shewhart-type signed-rank control chart based on runs. *Communications in Statistics - Simulation and Computation* 36 (2):335–56. doi: 10.1080/03610910601158427.

Chakraborti, S., S. Eryilmaz, and S. W. Human. 2009. A phase II nonparametric control chart based on precedence statistics with runs-type signaling rules. *Computational Statistics & Data Analysis* 53 (4):1054–65. doi: 10.1016/j.csda.2008.09.025.

Chakraborti, S., P. Laan, and M. A. Wiel. 2004. A class of distribution-free control charts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 53 (3): 443–62. doi: 10.1111/j.1467-9876.2004.0d489.x.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1):1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Gama, J., and M. M. Gaber. 2007. *Learning from data streams: Processing techniques in sensor networks*. New York: Springer.

Lemley, J., S. Bazrafkan, and P. Corcoran. 2017. Smart augmentation-learning an optimal data augmentation strategy. *IEEE Access* 5:5858. doi: 10.1109/ACCESS.2017.2696121.

Li, Z., and P. Qiu. 2014. Statistical process control using a dynamic sampling scheme. *Technometrics* 56 (3):325–35. doi: 10.1080/00401706.2013.844731.

Li, S.Y., L.C. Tang, and S.H. Ng. 2010. Nonparametric CUSUM and EWMA control charts for detecting mean shifts. *Journal of Quality Technology* 42 (2):209–26. doi: 10.1080/00224065.2010.11917817.

Limongelli, M. P. 2003. Optimal location of sensors for reconstruction of seismic response through spline function interpolation. *Earthquake Engineering & Structural Dynamics* 32 (7):1055–74. doi: 10.1002/eqe.262.

Little, R. J., and D. B. Rubin. 2014. *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Liu, K., Y. Mei, and J. Shi. 2015. An adaptive sampling strategy for online high-dimensional process monitoring. *Technometrics* 57 (3):305–19. doi: 10.1080/00401706.2014.947005.

Liu, K., and J. Shi. 2013. Objective-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a Bayesian network. *IIE Transactions* 45 (6):630–43. doi: 10.1080/0740817X.2012.725505.

Liu, L., F. Tsung, and J. Zhang. 2014. Adaptive nonparametric CUSUM scheme for detecting unknown shifts in location. *International Journal of Production Research* 52 (6): 1592–606. doi: 10.1080/00207543.2013.812260.

Liu, K., X. Zhang, and J. Shi. 2014. Adaptive sensor allocation strategy for process monitoring and diagnosis in a bayesian network. *IEEE Transactions on Automation Science and Engineering* 11 (2):452–62. doi: 10.1109/TASE.2013.2287101.

Mei, Y. 2010. Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* 97 (2):419–33. doi: 10.1093/biomet/asq010.

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

Qiu, P. 2013. *Introduction to statistical process control*. Boca Raton, FL: CRC Press.

Qiu, P., and D. Hawkins. 2001. A rank based multivariate CUSUM procedure. *Technometrics* 43 (2):120–32. doi: 10.1198/004017001750386242.

Qiu, P., and D. Hawkins. 2003. A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. *Journal of the Royal Statistical Society: Series D (The Statistician)* 52 (2):151–64. doi: 10.1111/1467-9884.00348.

Reynolds, M. R., R. W. Amin, and J. C. Arnold. 1990. CUSUM charts with variable sampling intervals. *Technometrics* 32 (4):371–84. doi: 10.1080/00401706.1990.10484721.

Sasidhar, K., R. Sreeresmi, and P. Rekha. 2014. A WSN lifetime improvement algorithm reaping benefits of data aggregation and state transitions. In *2014 IEEE Global Humanitarian Technology Conference - South Asia Satellite*, 201–205.

Tan, R., G. Xing, J. Chen, W. Song, and R. Huang. 2012. Fusion-based volcanic earthquake detection and timing in wireless sensor networks. *ACM Transactions on Sensor Networks* 9 (2):1–25. doi: 10.1145/2422966.2422974.

Tartakovsky, A. G., B. L. Rozovskii, R. B. Blažek, and H. Kim. 2006. Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology* 3 (3):252–93. doi: 10.1016/j.stamet.2005.05.003.

Van Dyk, D. A., and X. L. Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10 (1):1–50. doi: 10.1198/10618600152418584.

Waharte, S., and N. Trigoni. 2010. Supporting search and rescue operations with UAVs. In *2010 International Conference on Emerging Security Technologies (EST)*, IEEE, 142–147.

Wang, K., and W. Jiang. 2009. High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology* 41 (3):247–58. doi: 10.1080/00224065.2009.11917780.

Wang, Y., and Y. Mei. 2013. Monitoring multiple data streams via shrinkage post-change estimation. *IEEE Transactions on Information Theory* 61:6926–38.

Wang, Y., and Y. Mei. 2015. Large-scale multi-stream quickest change detection via shrinkage post-change estimation. *IEEE Transactions on Information Theory* 61 (12):6926–38. doi: 10.1109/TIT.2015.2495361.

Wang, A., X. Xian, F. Tsung, and K. Liu. 2018. A spatial-adaptive sampling procedure for online monitoring of big data streams. *Journal of Quality Technology* 50 (4): 329–43. doi: 10.1080/00224065.2018.1507560.

Wei, G. C., and M. A. Tanner. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85 (411):699–704. doi: 10.1080/01621459.1990.10474930.

Woodall, W. H., and M. M. Ncube. 1985. Multivariate CUSUM quality-control procedures. *Technometrics* 27 (3):285–92. doi: 10.1080/00401706.1985.10488053.

Xian, X.,. A. Wang, and K. Liu. 2018. A nonparametric adaptive sampling strategy for online monitoring of big data streams. *Technometrics* 60 (1):14–25. doi: 10.1080/00401706.2017.1317291.

Xie, Y., J. Huang, and R. Willett. 2013. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing* 7 (1): 12–27. doi: 10.1109/JSTSP.2012.2234082.

Zi, X., C. Zou, and F. Tsung. 2012. A distribution-free robust method for monitoring linear profiles using rank-based regression. *IIE Transactions* 44 (11):949–63. doi: 10.1080/0740817X.2011.649386.

Zou, C., W. Jiang, and F. Tsung. 2011. A LASSO-based SPC diagnostic framework for multivariate statistical process control. *Technometrics* 53 (3):297–309. doi: 10.1198/TECH.2011.10034.

Zou, C., W. Jiang, Z. Wang, and X. Zi. 2015. An efficient on-line monitoring method for high-dimensional data streams. *Technometrics* 57 (3):374–87. doi: 10.1080/00401706.2014.940089.

Zou, C., and P. Qiu. 2009. Multivariate statistical process control using LASSO. *Journal of the American Statistical Association* 104 (488):1586–96. doi: 10.1198/jasa.2009.tm08128.

Zou, C., Z. Wang, and F. Tsung. 2012. A spatial rank-based multivariate EWMA control chart. *Naval Research Logistics (NRL)* 59 (2):91–110. doi: 10.1002/nav.21475.

Appendix A: Derivation of Equations (7) – (9)

In this appendix, we derive the Equations (7) – (9) on the expression of the augmented vector.

As a result of the sparse shift assumption, Equation (3) can be further calculated based on Equation (6) that

$$\eta_j(t) = \mathbb{P}\Big(\xi_j(t) = 1|\boldsymbol{X}^{\mathcal{O}}(t)\Big) = \sum_{l=1}^{p} \mathbb{P}\Big(\xi_j(t) = 1,\ j_{OC} = l|\boldsymbol{X}^{\mathcal{O}}(t)\Big). \tag{14}$$

Recall that $i(t) = \text{argmax}_{j\in\mathcal{O}(t)} X_j(t)$ is the index of the largest observable variable at time t. To solve Equation (14), we consider the following three cases:

(1) If $j \in \mathcal{O}(t)$ and $j \neq i(t)$, then naturally

$$\mathbb{P}\Big(\xi_j(t) = 1,\ j_{OC} = l|\boldsymbol{X}^{\mathcal{O}}(t)\Big) = 0 \tag{15}$$

for any l since $\mathbb{P}\Big(\xi_j(t) = 1|\boldsymbol{X}^{\mathcal{O}}(t)\Big) = 0$. Therefore $\eta_j(t) = 0$ in this case.

(2) If $j = i(t)$, then

$$\begin{aligned}\mathbb{P}\Big(\xi_j(t) = 1,\ j_{OC} = l|\boldsymbol{X}^{\mathcal{O}}(t)\Big) &= \mathbb{P}\Big(\xi_j(t) = 1|j_{OC} = l,\ \boldsymbol{X}^{\mathcal{O}}(t)\Big) \\ &\quad \mathbb{P}\big(j_{OC} = l|\boldsymbol{X}^{\mathcal{O}}(t)\big) = \mathbb{P}\big(X_j(t){>}X_m(t),\ \forall m \notin \mathcal{O}(t)|j_{OC} = l\big) \\ &\quad \mathbb{P}\big(j_{OC} = l|\boldsymbol{X}^{\mathcal{O}}(t)\big).\end{aligned} \tag{16}$$

(3) If $j \notin \mathcal{O}(t)$, then

$$\mathbb{P}\left(\xi_j(t) = 1,\ j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right) = \mathbb{P}\left(\xi_j(t) = 1|j_{OC} = l,\ \mathbf{X}^{\mathcal{O}}(t)\right)$$
$$\mathbb{P}\left(j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right) = \mathbb{P}(X_j(t) > X_{i(t)},\ X_j(t) > X_m(t),$$
$$\forall m \notin \mathcal{O}(t) \text{ and } m \neq jj_{OC} = l)\mathbb{P}\left(j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right). \tag{17}$$

In the above equations,

$$\mathbb{P}\left(j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right) = \frac{p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)p\left(j_{OC} = l\right)}{\sum_{l=1}^{p} p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)p\left(j_{OC} = l\right)}$$
$$= \frac{p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)g_l}{\sum_{l=1}^{p} p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)g_l}. \tag{18}$$

It is noteworthy that for some discrete distributions, $i(t) = \text{argmax}_{j\in\mathcal{O}(t)} X_j(t)$ may not be unique. In this case, we could evenly distribute the probability to those variables with equal values. For example, if $\{j_1,\ j_2\} = \text{argmax}_{j\in\mathcal{O}(t)} X_j(t)$, let

$$\mathbb{P}\left(\xi_{j_1}(t) = 1,\ j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right) = \mathbb{P}\left(\xi_{j_2}(t) = 1,\ j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right)$$
$$= \frac{1}{2}\mathbb{P}(X_{j_1}(t)>X_m(t),\ \forall m \notin \mathcal{O}(t)|j_{OC} = l)\mathbb{P}\left(j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right) \tag{19}$$

Recall that it is assumed that all the variables are i.i.d. (with CDF F and PDF f). We can provide more detailed formulas to calculate the value of $\eta_j(t)$ accordingly. Since all the variables have the same distribution when the process is IC, by symmetry we have

$$g_j = \frac{1}{p} \text{ for } \forall j \in \mathcal{P} \text{ and } \mathbb{P}\left(X_j(t) = \max_{l\notin\mathcal{O}(t)} X_l(t)\right)$$
$$= \frac{1}{p-q} \text{ for } \forall j \notin \mathcal{O}(t). \tag{20}$$

Then we can calculate based on Equations (14)-(17) as follows. First of all,

$$\mathbb{P}(X_{i(t)}(t)>X_m(t),\ \forall m \notin \mathcal{O}(t)|j_{OC} = l)$$
$$= \begin{cases} F\left(X_{i(t)}(t)\right)^{p-q}, & \text{if } l \in O(t), \\ F\left(X_{i(t)}(t)\right)^{p-q-1}F\left(X_{i(t)}(t)-\mu_{\min}\right), & \text{if } l \notin O(t). \end{cases} \tag{21}$$

Accordingly, for $j \notin \mathcal{O}(t)$,

$$\mathbb{P}(X_j(t)>X_{i(t)},\ X_j(t)>X_m(t),\ \forall m \notin \mathcal{O}(t) \text{ and } m \neq j|j_{OC} = l)$$
$$= (1-\mathbb{P}(X_{i(t)}(t)>X_m(t),\ \forall m \notin \mathcal{O}(t)|j_{OC} = l)) \cdot \frac{1}{p-q}$$
$$= \begin{cases} \left(1-F\left(X_{i(t)}(t)\right)^{p-q}\right) \cdot \frac{1}{p-q}, & \text{if } l \in O(t), \\ \left(1-F\left(X_{i(t)}(t)\right)^{p-q-1}F\left(X_{i(t)}(t)-\mu_{\min}\right)\right) \cdot \frac{1}{p-q}, & \text{if } l \notin O(t). \end{cases} \tag{22}$$

Besides, given that

$$p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)$$
$$= \begin{cases} f\left(X_l(t)-\mu_{\min}\right)\prod_{j\in\mathcal{O}(t),\ j\neq l} f\left(X_j(t)\right), & \text{if } l \in O(t), \\ \prod_{j\in\mathcal{O}(t)} f\left(X_j(t)\right), & \text{if } l \notin O(t), \end{cases} \tag{23}$$

we can derive based on Bayes rule that

$$\mathbb{P}\left(j_{OC} = l|\mathbf{X}^{\mathcal{O}}(t)\right) = \frac{p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)g_l}{\sum_{l=1}^{p} p\left(\mathbf{X}^{\mathcal{O}}(t)|j_{OC} = l\right)g_l}$$
$$= \begin{cases} \dfrac{f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right)}{\sum_{l\in\mathcal{O}(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right) + (p-q)}, & \text{if } l \in O(t), \\ \dfrac{1}{\sum_{l\in\mathcal{O}(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right) + (p-q)}, & \text{if } l \notin O(t). \end{cases} \tag{24}$$

In this way, we finish the proof that the augmented vector $\eta_j(t)$ can be calculated as follows:

(1) If $j = i(t)$,

$$\eta_j(t) = \frac{F\left(X_{i(t)}(t)\right)^{p-q}\sum_{l\in O(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right)}{\sum_{l\in O(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right) + (p-q)}$$
$$+ \frac{F\left(X_{i(t)}(t)\right)^{p-q-1}F\left(X_{i(t)}(t)-\mu_{\min}\right)(p-q)}{\sum_{l\in O(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right) + (p-q)}.$$

(2) If $j \in \mathcal{O}(t)$ and $j \neq i(t)$,

$$\eta_j(t) = 0.$$

(3) If $j \notin \mathcal{O}(t)$,

$$\eta_j(t) = \frac{\left(1-F\left(X_{i(t)}(t)\right)^{p-q}\right)\sum_{l\in\mathcal{O}(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right)}{(p-q)\left(\sum_{l\in\mathcal{O}(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right) + (p-q)\right)}$$
$$+ \frac{1-F\left(X_{i(t)}(t)\right)^{p-q-1}F\left(X_{i(t)}(t)-\mu_{\min}\right)}{\sum_{l\in\mathcal{O}(t)} f\left(X_l(t)-\mu_{\min}\right)/f\left(X_l(t)\right) + (p-q)}.$$

Appendix B: Algorithm to find the threshold value *h*

This appendix describes the detailed algorithm to estimate the threshold value h for a given pre-scribed IC ARL, ARL_0, using collected historical IC data. This appendix is inspired by Appendix D of Liu, Mei, and Shi (2015).

Initialization:

Set $h_{\min}$ and $h_{\max}$ as the initial lower and upper bounds of h, respectively. Let $h = \frac{h_{\min}+h_{\max}}{2}$, $\overline{RL} = \infty$.

Repeat the following steps until the difference $|\overline{RL}-ARL_0$ is smaller than a small threshold:

1. Repeat the following steps for 5000 time:
   a. Generate a bootstrap dataset with 5000 samples by randomly drawing with replacement from the historical data.

b. Implement the R-SADA algorithm with threshold h and record the index of the first OC sample, RL. Update $\overline{RL}$ with the average of RL.

2. If $\overline{RL} > ARL_0$, let $h_{\max} = h$; otherwise, let $h_{\min} = h$. Then let $h = \frac{h_{\min}+h_{\max}}{2}$.

Appendix C: Proof of the IC property

In this Appendix, we will prove the IC property of the R-SADA control chart in Section 3.3. First, we consider the following two lemmas.

Lemma 1: If the variables are i.i.d., then $\mathbb{E}\eta_j(t) < \frac{1}{p}$ for $j \in \mathcal{O}(t)$ and $\mathbb{E}\eta_j(t) > \frac{1}{p}$ for $j \notin \mathcal{O}(t)$ when the process is IC.

From Equation (7) and the fact that $F\big(X_{i(t)}(t) - \mu_{\min}\big) < F\big(X_{i(t)}(t)\big)$ since $\mu_{\min} > 0$, we can see that

$$\eta_{i(t)}(t) < F\big(X_{i(t)}(t)\big)^{p-q}. \tag{25}$$

And when $j \in \mathcal{O}(t)$, $\eta_j(t)$ has probability $\frac{1}{q}$ to be nonzero and probability $\frac{q-1}{q}$ to be zero. Denote the cumulative distribution function of $X_{i(t)}(t)$ to be $F_{X_{i(t)}(t)}(x)$, then

$$F_{X_{i(t)}(t)}(x) = \mathbb{P}\big(X_{i(t)}(t) \le x\big) = \mathbb{P}\big(X_j(t) \le x,\ \forall j \in \mathcal{O}(t)\big) \\ = \mathbb{P}\big(X_j(t) \le x\big)^q = F(x)^q. \tag{26}$$

Therefore, we can calculate that

$$\mathbb{E}\eta_j(t) < \frac{1}{q}\mathbb{E}\Big[F\big(X_{i(t)}(t)\big)^{p-q}\Big] = \frac{1}{q}\int F\big(X_{i(t)}(t)\big)^{p-q} dF_{X_{i(t)}(t)}(x) \\ = \frac{1}{q}\int F(x)^{p-q} dF(x)^q = \int F(x)^{p-1} dF(x) = \int_0^1 y^{p-1} dy = \frac{1}{p}. \tag{27}$$

As $\eta_j(t) = \frac{1}{p-q}\big(1 - \eta_{i(t)}(t)\big)$ for $j \notin \mathcal{O}(t)$, we can derive similarly that $\mathbb{E}\eta_j(t) > \frac{1}{p}$ for $j \notin \mathcal{O}(t)$.

Lemma 2: There exists a time t_0 such that $S^{(1)}_{i,\ t} \le S^{(1)}_{j,\ t}$, for all $t \ge t_0$, $j \in \mathcal{P}\ U$ and $i \in U$.

We prove Lemma 2 via contradiction. Assume that there exists some $t \ge t_0$, $j \in \mathcal{P}\ U$ and $i \in U$ such that $S^{(1)}_{i,\ t} > S^{(1)}_{j,\ t}$. Then at time $t+1$, either i or j cannot be observed since $i \in U$ and $S^{(1)}_{i,\ t} > S^{(1)}_{j,\ t}$. This means that $i, j \notin \mathcal{O}(t+1)$, and $\eta_i(t+1) = \eta_j(t+1)$, then

$$S^{(1)}_{i,t+1} = \Big(S^{(1)}_{i,t} + \eta_i(t+1)\Big)(C_{t+1} - k)/C_{t+1} > \\ \Big(S^{(1)}_{j,t} + \eta_j(t+1)\Big)(C_{t+1} - k)/C_{t+1} = S^{(1)}_{j,t+1}. \tag{28}$$

Then, by induction, we can see that the variable j cannot be observed before variable i is observed. It means that $j \in U$, which contradicts the assumption. Therefore, we have proved Lemma 2.

Suppose that there exists a variable $i \in U$. From Lemma 2, we have proved that there exists a time t_0 such that $S^{(1)}_{i,\ t} \le S^{(1)}_{j,\ t}$ for any $j \in \mathcal{P}\ U$ and $t \ge t_0$. Suppose that a variable $j \in \mathcal{P}\ U$ is observable at time $t_1 < t_2 < \ldots < t_l \cdots$ and $t_1 > t_0$. Let $\omega_t = \frac{C_t - k}{C_t}$ $(0 \le \omega_t < 1,\ t \ge t_0)$. Then, we have that

$$S^{(1)}_{j,t_l} = \omega_{t_l}\Big(S^{(1)}_{j,t_l-1} + \eta_j(t_l)\Big) = S^{(1)}_{j,t_l-1} + \eta_j(t_l) - (1-\omega_{t_l}) \\ \Big(S^{(1)}_{j,t_l-1} + \eta_j(t_l)\Big). \tag{29}$$

Applying this derivation recursively on $S^{(1)}_{j,t_l-i}$ for $i = 1,\ 2,\ \cdots, t_l - t_0 - 1$ and we can get

$$S^{(1)}_{j,t_l} = \omega_{t_l}\Big(S^{(1)}_{j,t_l-1} + \eta_j(t_l)\Big) = S^{(1)}_{j,t_0} + \sum_{\tau=t_0+1}^{t_l} \eta_j(\tau) \\ - \sum_{\tau=t_0+1}^{t_l} (1-\omega_\tau)\Big(S^{(1)}_{j,\tau-1} + \eta_j(\tau)\Big). \tag{30}$$

Similarly, for $i \in U$, we have that

$$S^{(1)}_{i,t_l} = S^{(1)}_{i,t_0} + \sum_{\tau=t_0+1}^{t_l} \eta_i(\tau) - \sum_{\tau=t_0+1}^{t_l} (1-\omega_\tau)\Big(S^{(1)}_{i,\tau-1} + \eta_i(\tau)\Big). \tag{31}$$

Note that $S^{(1)}_{j,\ \tau-1} + \eta_j(\tau) = \frac{1}{\omega_\tau} S^{(1)}_{j,\ \tau} \ge \frac{1}{\omega_\tau} S^{(1)}_{i,\ \tau} = S^{(1)}_{i,\ \tau-1} + \eta_i(\tau)$, according to Lemma 2. Given that $\eta_j(t) = \eta_i(t)$ if $j \notin \mathcal{O}(t)$, it can be derived that

$$S^{(1)}_{j,t_l} - S^{(1)}_{i,t_l} = S^{(1)}_{j,t_0} - S^{(1)}_{i,t_0} + \sum_{m=1}^{l}\big(\eta_j(t_m) - \eta_i(t_m)\big) \\ - \sum_{\tau=t_0+1}^{t_l} (1-\omega_\tau)\Big[\Big(S^{(1)}_{j,\tau-1} + \eta_j(\tau)\Big) - \Big(S^{(1)}_{i,\tau-1} + \eta_i(\tau)\Big)\Big] \\ \le S^{(1)}_{j,t_0} - S^{(1)}_{i,t_0} + \sum_{m=1}^{l}\big(\eta_j(t_m) - \eta_i(t_m)\big). \tag{32}$$

From Lemma 1, $\mathbb{E}\eta_j(t_m) < \frac{1}{p} < \mathbb{E}\eta_i(t_m)$ for $m = 1, \cdots, l$. Therefore, $\sum_{m=1}^{l}\big(\eta_j(t_m) - \eta_i(t_m)\big)$ is a general random walk with mean $\mathbb{E}\eta_j(t_m) - \mathbb{E}\eta_i(t_m) < 0$. Denote L^* be the minimal l such that $\sum_{m=1}^{l}\big(\eta_j(t_m) - \eta_i(t_m)\big) < S^{(1)}_{i,t_0} - S^{(1)}_{j,t_0} < 0$, then $\mathbb{P}(L^* < \infty) = 1$ based on property of general random walk (Ross (1996)). In other words, the probability that there exists a finite l such that $S^{(1)}_{i,t_l} > S^{(1)}_{j,t_l}$ is 1, according to Equation (32). Consequently, this means that with probability 1, the above argument contradicts with Lemma 2. As a result, we have proved that there exists no such variable $i \in U$, and thus U must an empty set with probability 1.

Appendix D: Proof of the OC property

In this Appendix, we will prove the OC property of the R-SADA method in Section 3.3. Without loss of generality, here we focus on the case when $k = 0$. Consider that at time $t > t_0$, the OC variable X_{j^*} is observable (i.e.,

$j^* \in \mathcal{O}(t)$). Then $S^{(1)}_{j^*,t} > S^{(1)}_{i,\ t}$ for any variable $i \notin \mathcal{O}(t)$. Recall that $j_{(1)}$, ..., $j_{(p)}$ are the variable indices such that $S^{(1)}_{j_{(1)},t} \geq S^{(1)}_{j_{(2)},t} \geq \cdots \geq S^{(1)}_{j_{(p)},t}$. Define the difference between the increments of augmented values on variable j^* and $j_{(q+1)}$ at time $t+n$ to be

$$Z_{j^*,n} = \eta_{j^*}(t+n) - \eta_{j_{(q+1)}}(t+n). \tag{33}$$

Let $H_{j^*,N} = \sum_{n=1}^{N} Z_{j^*,n}$, $G_{j^*,N} = \sum_{n=1}^{N} \eta_{j^*}(t+n)$, $\tau_H = \inf_N \{H_{j^*,N} \leq 0\}$ and $\tau_G = \inf_N \{G_{j^*,N} \leq 0\}$. Under the condition that $N \leq \tau_G$, we claim that $\{G_{j^*,N}\}$ is a general random walk with a positive drift. From the definition of τ_H, we can see that $H_{j^*,N} > 0$ for any $N < \tau_H \leq \tau_G$. This further indicates $S^{(1)}_{j^*,t+N} > S^{(1)}_{(q+1),t+N}$ for any $N < \tau_H \leq \tau_G$. In other words, j^* is always observed at and before time τ_H. As a result, $\eta_{j^*}(t+1) - \eta_{j_{(q+1)}}(t+1)$, ..., $\eta_{j^*}(t+\tau_H-1) - \eta_{j_{(q+1)}}(t+\tau_H-1)$ are i.i.d. with mean $\mathbb{E}[\eta_{j^*}(t+N) - \eta_{j_{(q+1)}}(t+N)]$. Denote $a(t) = \frac{\sum_{l \in O(t)} f\left(X_l(t) - \mu_{\min}\right)/f(X_l(t))}{\sum_{l \in O(t)} f\left(X_l(t) - \mu_{\min}\right)/f(X_l(t)) + (p-q)}$, $b(t) = \frac{(p-q)}{\sum_{l \in O(t)} f\left(X_l(t) - \mu_{\min}\right)/f(X_l(t)) + (p-q)}$, $F_1(t) = F\left(X_{i(t)}(t)\right)^{p-q}$ and $F_2(t) = F\left(X_{i(t)}(t)\right)^{p-q-1} F\left(X_{i(t)}(t) - \mu_{\min}\right)$. Then

$$\begin{aligned}
\mathbb{E}\left[\eta_{j^*}(t+N) - \eta_{j_{(q+1)}}(t+N)\right] &\geq \mathbb{E}\left[\frac{1}{q}(a(t)F_1(t) + b(t)F_2(t)) \right. \\
&\quad \left. - \left(\frac{a(t)(1-F_1(t))}{p-q} + \frac{b(t)(1-F_2(t))}{p-q}\right)\right] \\
&= \ \mathbb{E}\left[a(t)\frac{pF_1(t)-q}{q(p-q)} + b(t)\frac{pF_2(t)-q}{q(p-q)}\right] = \frac{1}{q(p-q)} \\
\mathbb{E}\left[a(t)\left(pF_1(t)-q\right) + b(t)\left(pF_2(t)-q\right)\right] &\geq \frac{1}{q(p-q)} \\
\mathbb{E}[a(t)+b(t)]\mathbb{E}\left[\left(pF_2(t)-q\right)\right] &= \frac{1}{q(p-q)}\left(p\mathbb{E}[F_2(t)]-q\right) > 0
\end{aligned} \tag{34}$$

Therefore, $\mathbb{P}(\tau_G = \infty) > 0$, and $\mathbb{P}(\tau_H = \infty) > 0$. This leads to a conclusion that

$$\mathbb{P}\left(S^{(1)}_{j^*,t+N} - S^{(1)}_{(q+1),t+N} > 0 \text{ for any } N > 0\right) > 0, \tag{35}$$

i.e., $\mathbb{P}\left(j^* \in \mathcal{O}(t') \text{ for any } t' > t\right) > 0$. Thus, we have proved the OC property.