# CORRELATION-BASED DYNAMIC SAMPLING FOR ONLINE HIGH DIMENSIONAL PROCESS MONITORING

Mohammad Nabhan
King Fahd University of Petroleum and Minerals, Dhahran, 31261, Kingdom of Saudi Arabia

Yajun Mei
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

Jianjun Shi
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

## ABSTRACT

Effective process monitoring of high-dimensional data streams with embedded spatial structures has been an arising challenge for environments with limited resources. Utilizing the spatial structure is key to improve monitoring performance. This article proposes a correlation-based dynamic sampling technique for change detection. Our method borrows the idea of Upper Confidence Bound algorithm and uses the correlation structure not only to calculate a global statistic, but also to infer unobserved sensors from partial observations. Simulation studies and two case studies on solar flare detection and carbon nanotubes (CNTs) buckypaper process monitoring are used to validate the effectiveness of our method.

***Keywords*****:** Adaptive sampling; partial observations; limited resources; data fusion; order thresholding; change detection.

# 1. INTRODUCTION

The ubiquitous use of sensing systems in manufacturing, healthcare, biosurveillance, network security, and service processes has created data rich environments that have presented challenges for real-time monitoring and analysis. This is especially true in the environments with limited resources, whether at the data acquisition level or processing level. For instance, when low-cost wireless sensor networks are employed for monitoring volcano activity (Pereira et al. 2014), one may want to prolong the lifetime of such networks by turning on only a limited number of battery-power sensors unless the volcano is active. When using touch-probe coordinate measuring machines (CMM) to monitor wafer manufacturing processes (Jin et al. 2012), the current profile measurement schemes are time-consuming. Therefore, it is essential to reduce the number of samples measured in wafers while still adequately monitoring process quality. Besides physical devices, the term "sensor" can also be used to denote any sources that generate relevant information. Moreover, in many real-world data rich environments, we often face resource constraints in the capacity of acquisition, transmission, analysis, or fusion of data. In biosurveillance and epidemiology, the Center for Disease Control and Preventions (CDC) has a limited capacity for drug resistance tests that monitor the resistance status for certain infectious diseases. Thus, it is crucial to decide how to effectively allocate the resources and choose which affected patients, sub-populations, or regions to monitor. Hence, in the general context of real-time or online monitoring high-dimensional data streams in resource constrained environments, it is important to dynamically sample those informative local data streams while making adequate online anomaly detection. In particular, in this paper, we investigate two such constrained environments in the form of case studies: one is to detect solar flares via satellite imaging processing with limited transmission capacities, and the other is carbon nanotubes (CNTs) buckypaper process monitoring with limitations immerging from trade-offs between scanning durations and signal to noise ratios.

There are several recent articles that tackle this problem by introducing an adaptive sampling scheme that only uses a fraction of the full observation spectrum to make real-time inferences on the state of a system. Liu et al. (2015) proposed an adaptive sampling strategy with resource limitations in which data streams are assumed to be normally distributed. Furthermore, a nonparametric adaptive sampling procedure under limited resources has been proposed by Xian et

al. (2018b). These methods assume that the data streams are spatially independent, which means that observations collected from different sensors at any given time are independent. Wang et al. (2018) proposed an adaptive sampling strategy that take a random grid and then uses finer grids when a hint of a signal occurs based on the weighted average of the likelihood ratio statistics of the data from itself as well as from its neighborhood, where the spatial information is used to define the weights. However, it is unclear how to extend to more complicated data models.

In this article, we apply the ideas of the celebrated Upper Confidence Bound (UCB) algorithm proposed by Lai (1987), Lai and Robbins (1985) in the Multi-Armed Bandit (MAB) problems to Statistical Process Control (SPC), and develop effective process monitoring of high-dimensional data streams with embedded spatial structure for environments with limited resources. In many real-world applications of SPC, the anomalies are often clustered and sparse, and thus we need to balance the tradeoff between randomly searching for possible anomalous local data streams or local regions (exploration) and performing focused sampling on local data streams or local regions near the anomalous regions for quick detection (exploitation). Now the exploration-exploitation tradeoff has been well-studied in MAB problems, and the key idea of the celebrated UCB algorithm is to use the upper confidence bound of the parameter estimation for adaptive sampling. These inspire us to explore the embedded spatial structures of local data streams/sensors to use the upper confidence bound of the local stream post-change parameter estimator to develop efficient dynamic sampling methods for online monitoring and SPC. It turns out that the existing method in Liu et al. (2015) is a special case of our proposed methods for independent data, and thus is a UCB-type algorithm for SPC. We feel that our combination of MAB and SPC is novel, and this opens a new research direction in SPC for dynamic sampling of incomplete high-dimensional data monitoring under resource constrained environments.

We should acknowledge that dynamic sampling strategies in SPC literature usually revolve around the temporal domain where the objective is mainly to inspect the quality of the product or service (Montgomery 2009). In such scenarios, the limitation is in the frequency of acquisition times, which is usually associated with the cost of data acquisition. A common example of the cost of acquisition is when the quality inspection procedure calls for a destructive test on the parts being produced. Meanwhile, our sampling strategies are over the spatial domain, and the issue lies in the capacity of deploying, observing, transmitting, or fusing all the available sensors that are

monitoring the process at any given time. The key concern that we address is how to utilize the information embedded in the spatial structure of the data streams to improve the effectiveness of the monitoring procedure. This allows for a more informative and intuitive framework when dynamically sampling the partition of streams to be observed at any given acquisition period.

A dynamic sampling strategy based on the correlation structure of data streams is characterized by how it accomplishes the following tasks at every data acquisition time $t$: (1) determining the fraction of sensors to be deployed; (2) providing an educated compensation for unobserved readings of undeployed sensors based on their correlation with measured variables from deployed sensors; (3) computing local statistics for deployed sensors based on the observed measurements while using the correlation based compensations for the undeployed ones; (4) fusing these local statistics into a single global statistic for global-level decision making.

The novelty of our proposed dynamic sampling method lies in exploiting the spatial correlation structure to provide an upper confidence bound of post-change parameter estimation and is therefore named Correlation-based Dynamic Sampling (CDS). The procedure is dynamic in the sampling process of the variables to be observed at each acquisition period, as well as in providing compensation for the unobserved variables. The dynamic behavior is achieved by combining the correlation structure with the information obtained from the observed partition of the data streams. The dynamic compensation we propose is constructed from the upper confidence bound of the marginal conditional distribution of the unobserved variables given the observed variables. When a well-structured framework such as multivariate normal distribution is assumed, the marginal conditional distribution is very well defined to be another Gaussian distribution. The marginal distribution is tractable even in high dimension when the spatial structure is readily available. This sensor assignment procedure allows for a pseudo-random sampling strategy when the process is in-control, as well as fast localization of faulty variables when the process is out-of-control. We use the term "pseudo-random" here because, although the sampling procedure tends to select a cluster of variables to be observed at any given time based on the spatial structure, the clusters themselves are randomly constructed. Furthermore, these clusters are formed from variables that are correlated. This feature of cluster formation will be illustrated further in the simulation and case studies.

The remainder of this paper is organized as follows: In Section 2, we provide a brief literature review relevant to the issue of limited resources, followed by a more detailed review of adaptive sampling methods in the literature. Next, in Section 3, we discuss in detail our proposed adaptive sampling strategy for online high-dimensional process monitoring, and also present two properties pertaining to its sampling behavior depending on the state of the system. Section 4 assesses the performance of our proposed sampling strategy on virtually simulated scenarios, while Section 5 tests the performance using two case studies involving solar flare detection and in-line Raman spectroscopy. We conclude our paper with a brief discussion of the key findings of our proposed monitoring scheme.

# 2. LITERATURE REVIEW

The following section is split into two further sub-sections. The first (Section 2.1) provides a brief review of relevant topics that address different aspects of resource limitations from our problem, whereas the second (Section 2.2) gives a detailed review of closely related procedures discussed in the literature as well as the renowned UCB algorithm in the classical MAB problem. This will lay the proper foundation for our subsequent discussions.

## 2.1 Relevant Topics for Limited Resources

There are two main problems explored in the literature that share some resemblance to our limited resources process monitoring setting from an application perspective: (i) the optimal design of sensors in a DSN system and (ii) the theory of searching and tracking targets. Regarding the first, the objective of a DSN is to find a fixed sensor layout optimized for process monitoring. However, due to the fixed layout, shifts that occur outside the predefined layout will reduce detection power, as well as diagnostic capability, as discussed in (Li and Jin 2010, Liu and Shi 2013). Studies in (Mandroli et al. 2006, Ding et al. 2006) provide inclusive reviews of the state-of-the-art advances in DSNs for enhancement in quality and productivity.

In the second example, the objective of searching and tracking target studies is to obtain an effective employment of the limited resources available to locate a target object of interest that is within an unknown location (Frost and Stone 2001, Lim et al. 2006, Zoghi and Kahaei 2010,

Ben-Gal and Kagan 2013). The main assumption of these studies is that there exists a singular object in the searching space but at some unknown locations.

This article differs from the aforementioned applications in that our objective is to develop a *dynamic* monitoring strategy in which the data streams are correlated and are flowing continuously with the uncertainty that a failure, target, or event may or may not occur to the system. Furthermore, in its core, our proposed methodology does not assume prior information on the failure characteristics. Nonetheless, it is also capable of incorporating such information seamlessly as will be demonstrated in the case studies in Section 5.

### 2.2 Review of Adaptive Sampling Methodologies

There are several ways of approaching the issue of monitoring a process with limited resources. The two most forward approaches are (i) random sampling and (ii) choosing a fixed set of variables to monitor. While both approaches can be effective in certain situations, they both suffer from not utilizing any information gained during the monitoring procedure. For example, setting fixed sensors can only detect changes in the sensors selected, but it is rare in practice to have perfect knowledge about where the fault may occur. On the other hand, while random sampling might eventually detect a change in a subset of sensors, its detection delay can be large if the magnitude of the change is not large enough to set an immediate alarm, as the process switches to monitor a different set of sensors in the next acquisition period.

One of the most relevant and recent research efforts was completed by Liu et al. (2015) who proposed an adaptive sampling strategy that is effective for the online monitoring of high-dimensional data streams. Their proposed method was based on a procedure called Top-r cumulative sum (CUSUM), which was first introduced in (Mei 2010). Although their proposed Top-r adaptive sampling strategy (TRAS) was shown to be effective for online monitoring of high dimensional data streams, it is limited to applications where there is no significant embedded correlation structure in the streams and independence across different data streams can be assumed. Furthermore, a similar adaptive sampling procedure under limited resources was proposed by Xian

et al. (2018b). Their method is a nonparametric approach that addresses a similar problem under the independent assumption except when the underlying distribution of data streams is unknown.

The aforementioned proposed algorithms in the literature monitor individual sensors or local data streams by computing local statistics based on the commonly used cumulative sum (CUSUM) procedures in statistical process control, and then take advantage of the independence assumptions across different sensors to construct the global monitoring statistic based on the sum of a few larger local CUSUM statistics. These methods address the limitation of resources by assigning a uniform non-informative constant compensation value to all the undeployed sensors.

Wang et al. (2017) proposed an adaptive sampling strategy under the assumption that data streams are spatially independent, and the occurring faults affect a local cluster of sensors within a grid. The method requires setting the cluster size, which typically would require former knowledge of fault patterns. Another study on climate simulation (Xian et al. 2018a) attempts to address the challenge of dynamically sampling data and deciding which to archive due to memory limitations. However, this problem is different than ours in the sense that the limitation is not in acquiring the data, but rather in choosing what is worth keeping.

Next, we provide a brief review on the classical multi-armed bandit problem (MAB), which includes many useful adaptive/dynamic sampling methodologies. In the simplest case of classical MAB, one assumes that there are $p$ sensors or arms, and each sensor provides a random reward from an unknown probability distribution specific to that sensor. The objective is to maximize the sum of rewards earned through allocating resources to the choice of sensors over time. Mathematically, the $i$-th sensor generates i.i.d. observations over time, say, $\{X_{i,1}, \cdots, X_{i,t}, \cdots\}$. Suppose that sensor observations $X_{i,t}$ have the common variance $\sigma^2$, but might have different means for different sensors. At each time step $t$, one can take observations only from one sensor, say, the $i^*(t)$-th sensor, and receives the reward $r_t = X_{i^*(t),t}$. Then one wants to decide which sensor to take observation at every time step to maximize the expected overall rewards, $E(\sum_{t=1}^{T} r_t)$, where $T$ is the pre-specified total number of time steps.

The crucial tradeoff one faces at the MAB is between "exploitation" of the sensor that has the highest mean and "exploration" to get more information about the mean of the other sensors.

An intuitive and appealing policy is to estimate each unknown mean $\mu_k$ by the corresponding sample mean of each sensor, and then take observations from the sensor that has the largest sample mean. Unfortunately, such a myopic policy performs poorly, due to the poor exploration of other sensors. One of the asymptotically optimal MAB policies is the notable Upper Confidence Bound (UCB) algorithm proposed by Lai (1987), Lai and Robbins (1985). The main idea of the UCB algorithm is to take observations from the arm having the highest upper confidence bound index, constructed from the Kullback-Leibler divergence between the estimated distributions of the sensors. In the setting when the sensor observations are normally distributed, the UCB-Lai algorithm chooses the arm maximizing $\bar{X}_{k,t_k} + \sigma\sqrt{\frac{2\log(T/t)}{t}}$ at each time step $t$, where $\bar{X}_{k,t_k}$ is the current observed sample mean from the $k$-th sensor, $t_k$ is the current number of observations from the $k$-th sensor, and $\sigma$ is the standard deviation of sensor observations. There are many modifications or extensions of the UCB-Lai algorithm, e.g., Auer et al. (2002) proposed a modified UCB algorithm that does not involve the total sample size $T$ in the setting of normal distributions with unequal and unknown variances.

Here we extend the classical MAB with two twists: one is the changing environments with a different reward function that is non-additive over the time domain for the purpose of quick detection, and the other is to increase the number of sampled sensors from 1 to $m \geq 1$ over the spatial domain. We propose to apply the idea of the UCB to SPC, which leads to a dynamic compensation value to the unobserved streams or sensors based on the spatial correlation structure of the data and the information obtained from the observed streams or sensors. In order to highlight our main ideas, we make a simplified assumption that the correlation structure remains unchanged and thereby stationary throughout the whole monitoring period, and we focus on detecting the sparse mean shifts of high-dimensional data with embedded spatial structures in the environments with limited resources.

## 3. CORRELATION-BASED DYNAMIC SAMPLING (CDS) STRATEGY

In this section we develop a method for effective monitoring of correlated high-dimensional data streams under the constraint of resource limitations. In our proposed strategy, we

first construct efficient local statistics for each individual data stream and consequently combine these local statistics into a single global statistic while utilizing the information embedded in the correlation structure of the streams. There are two novel ideas in the proposed strategy: (1) following the MAB algorithm to explore the spatial correlation structure and introduce a dynamic compensation value for the unobserved variables based on the confidence limit of their parameter estimates, and (2) deploying sensors to those variables efficiently to collect as much global change information after adjusting the spatial correlation.

The following subsections will elaborate on the steps of our proposed correlation-based dynamic sampling (CDS) strategy. Section 3.1 provides a detailed overview of our algorithm. Next, a detailed discussion of parameter settings is provided in Section 3.1.5. Finally, Section 3.3 discusses options for estimating or imposing the embedded spatial structure of the data streams.

### 3.1 CDS Methodology Development

In preparation to our discussion, we will first introduce the notations for the variables that will be used throughout the course of this paper. Suppose that the system to be monitored consists of $p$ variables $\mathcal{P} = \{1, \dots, p\}$ that are observable at any time $t$. The vector of observed variables at time $t$ is given by $\boldsymbol{X}_t = (X_{1,t}, \dots, X_{p,t})'$. Due to limitations in the resources available for monitoring, only a fraction of this vector is measured in real time. Let $m$ be the maximum number of variables/sensors that can be measured/deployed at any acquisition time. From the problem statement, $m$ is a process parameter dictated by the system monitoring capability. This could translate to the number of sensors available for deployment at each acquisition time, the transmission capacity, or the computational power at the data fusion center. To facilitate referencing measured variables at each time $t$, we introduce two sets $\omega_t \subset \Omega$ and $\psi_t \subset \Psi$. Here, $\Omega$ and $\Psi$ are all possible partitions of the data streams into observed and unobserved sets, respectively. Thereby, variable $X_{i,t} \in \omega_t$ if and only if it is measured at time $t$, otherwise it is assigned to set $\psi_t$. Hence, the cardinalities of $\omega_t$ and $\psi_t$ are respectively $|\omega_t| = m$ and $|\psi_t| = p - m$.

We assume that $\boldsymbol{X}_t$ comes from a multivariate normal distribution, in which the mean vector is $\mu_t$ and the covariance structure $\Sigma$. The covariance structure plays an important role in our proposed dynamic sampling procedure. Particularly, the covariance between the unobserved sensors and the observed ones (denoted by $\Sigma_{\Psi\Omega}$) is the base of inferences to be made on unobserved sensors. The in-control mean and covariance are also assumed to be known. While these parameters are not generally known, they can be estimated from an adequate amount of historical data. They can also be set to target values defined by the engineering design of the process. Without loss of generality, we assume that the data has been preprocessed to have mean 0 and standardized to have a covariance matrix equal to that of the correlation matrix. After some point in time $\tau$ during the operation of the monitored system, a change in the mean vector occurs, where a subset $\Theta$ of the variables $\boldsymbol{X}_t$ will have a non-zero mean. Moreover, we assume that the correlation structure remains unchanged during this change and thereby stationary throughout the whole monitoring period. Our objective then becomes to first detect this change with minimum delay from the onset at $\tau$. Secondly, we need to identify the subset $\Theta$ with the shifted mean, when the variables at each given time are partially observable.

There are four components to our proposed method. First, we construct the local statistics for the deployed sensors based on the observed measurements. Second, we utilize the correlation between undeployed sensors and deployed ones to determine the local statistics of the unobserved variables. Third, we select the fraction of sensors to be deployed at the next acquisition time. Finally, we fuse the local statistics into a multivariate global statistic that is used to test whether the process remains in-control. In the following subsections, we will demonstrate how we can construct each one of the components and then conclude with an overview of the proposed monitoring scheme.

### *3.1.1 Determining Local Statistics*

Our objective in this paper is to detect any change to the mean of the monitored variables. Since this shift can be either positive or negative, it is appropriate to deploy a two-sided CUSUM monitoring statistic for each variable $k$ at time $t$ defined as

$$C_{k,t} = max\ (C_{k,t}^{+}, C_{k,t}^{-}), \tag{1}$$

where the notations $C_{k,t}^{+}$ and $C_{k,t}^{-}$ represent, respectively, the positive and negative local statistics for variable $k$ at time $t$.

At any given time, we are limited by the available resources, computation power, or transmission capabilities to calculate these local statistics using partial observations. Statistics pertaining to an observed sensor $X_{k,t} \in \omega_t$ at time $t$ can be defined as CUSUM statistics (Lorden (1971)) as follows:

$$C_{k,t}^{+} = max\left(0, C_{k,t-1}^{+} + \delta X_{k,t} - \frac{\delta^2}{2}\right) \text{ and } \quad C_{k,t}^{-} = max\left(0, C_{k,t-1}^{-} - \delta X_{k,t} - \frac{\delta^2}{2}\right), \quad (2)$$

where $C_{k,0}^{+} = C_{k,0}^{-} = 0$. Here, $\delta$ is the smallest mean shift magnitude that is of interest to detect (see the guidelines in subsection 3.1.5 on how to determine the value of $\delta$).

The main difficulty is how to define the local CUSUM statistics in (2) for those unobserved variables $\left(X_{k,t} \in \psi_t\right)$. Inspired by the UCB algorithm of Lai (1987), Lai and Robbins (1985) for MAB, here we propose to salvage (2) by utilizing the spatial correlation structure to obtain the estimated upper and lower bounds, say, $U_{k,t}$ and $L_{k,t}$, on the true unobserved variable $X_{k,t}$ at time $t$ (the estimates of $U_{k,t}$ and $L_{k,t}$ will be discussed in a little bit). Then we dynamically construct the local statistic as follows:

$$C_{k,t}^{+} = \max\left(0, C_{k,t-1}^{+} + \delta U_{k,t} - \frac{\delta^2}{2}\right), \quad (3)$$

$$C_{k,t}^{-} = \max\left(0, C_{k,t-1}^{-} - \delta\, L_{k,t} - \frac{\delta^2}{2}\right). \quad (4)$$

It remains to discuss how to obtain the estimates, $U_{k,t}$ and $L_{k,t}$, for unobserved sensors $\left(X_{k,t} \in \psi_t\right)$. Since the data streams are assumed to come from a standardized multivariate normal distribution, the marginal conditional distribution of an unobserved variable $X_k \subset \Psi$ over the remaining set of observed variables $\Omega$ is also normal with mean $\mu'_k$ and variance $\sigma'_k$ given by:

$$\mu'_k = \Sigma_{k,\Omega}\Sigma_{\Omega\Omega}^{-1}X_{\Omega}, \quad (5)$$

$$\sigma'_k = 1 - \Gamma_{kk}, \quad (6)$$

where, $\Sigma_{k,\Omega}$ is the covariance between $X_k \in \Psi$ and the observed variables in $\Omega$. Moreover, $\Gamma_{kk}$ denotes the $k^{th}$ diagonal entry of $\Gamma = \Sigma_{\Psi\Omega}\Sigma_{\Omega\Omega}^{-1}\Sigma_{\Omega\Psi}$.

Using the marginal conditional distribution of an unobserved variable $X_k \in \Psi$, we can construct an $(1-\alpha)100\%$ two-sided confidence interval as follows:

$$CI_{k,t} = \left[L_{k,t}, U_{k,t}\right]$$
$$\text{where,} \quad L_{k,t} = \mu'_{k,t} - \Phi^{-1}(1-\alpha/2)\sigma'_k \ \text{and}\ U_{k,t} = \mu'_{k,t} + \Phi^{-1}(1-\alpha/2)\sigma'_k. \tag{7}$$

Here, $\Phi^{-1}(.)$ is the inverse of the cumulative standard normal distribution. Hence, the bounds of the confidence interval $U_{k,t}$ and $L_{k,t}$ will be the base of our correlation based dynamic compensation procedure given in equations (3) and (4).

It is informative to compare our proposed dynamic compensations in (3) and (4) with the static uninformative compensation in (Liu et al. 2015). In that study, the local statistics for an unobserved variable are based on a static compensation $\Delta \geq 0$, and are defined as follows:

$$C_{k,t}^{+} = C_{k,t-1}^{+} + \Delta \text{ and } C_{k,t}^{-} = C_{k,t-1}^{-} + \Delta. \tag{8}$$

However, Liu et al. (2015) did not provide any statistical justification why one needs to add a static compensation $\Delta$ for unobserved variable.

The following proposition shows that the method in (Liu et al. 2015) is a special case of our approach for independent data streams, and thus the compensation defined in equation (8) is essentially an upper bound confidence (UCB)-type algorithm in the SPC context.

**Proposition:** Our proposed dynamic compensation procedure is a generalization of the constant compensation, and is consequently equivalent to it when all data streams are spatially independent. In that case, $\Delta = \delta\Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}$.

**Proof**: For spatially independent data, and for any partition of the data into observed and unobserved sets $\Omega$ and $\Psi$, the covariance between the two sets $\Sigma_{\Psi\Omega} = 0$. Consequently:

$$\mu'_k = 0, \sigma'_k = 1, CI_k^{\pm} = \pm\Phi^{-1}(1-\alpha/2) \text{ for all } \{k: X_k \in \psi\},$$

$$C_{k,t}^{+} = \max\left(0, C_{k,t-1}^{+} + \delta\,\Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}\right),$$

$$C_{k,t}^{-} = \max\left(0, C_{k,t-1}^{-} + \delta\,\Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}\right).$$

Let $\Delta = \delta\Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}$ and choose $(\delta, \alpha)$ such that $\Delta \geq 0$. Then, the update reduces to the format in equation (8). ■

The main reason that the confidence limits, $U_{k,t}$ and $L_{k,t}$, are chosen to represent unobserved instances rather than the middle of the confidence interval is to promote exploration during the in-control phase of the process by favoring those unobserved variables that have been sampled less. It can be noted that a compensation based on the middle of the interval would require the significance level $\alpha = 1$, and yields a myopic policy that only uses the estimated means for decision making. Moreover, when an unobserved variable is independent of all observed variables, the previous proposition suggests that the compensation $\Delta = \frac{\delta}{2} - \frac{\delta^2}{2}$, which might be negative. A negative compensation will result in a monotonic decrease in the local CUSUM statistics until they hit 0, which will in turn diminish the likelihood of those variables to ever be explored at future acquisition times. Further discussions of appropriate parameter settings and their role in promoting the in-control variable exploration behavior are available in subsection 3.1.4 and 3.1.5.

### *3.1.2 Global Statistics and Out-of-Control Criteria*

Since the data streams are spatially correlated, we propose to use a multivariate CUSUM (MCUSUM) statistic as the global statistic. To be more concrete, the local CUSUM statistics $C_{i,t}$ calculated for those observed sensors at time $t$ are fused into a global CUSUM statistic as follows:

$$GC_t = ||C_{k,t}|| = \sqrt{C_{k,t}\Sigma_{\omega\omega}^{-1}C_{k,t}} \qquad k = \{n: X_n \in \omega\}. \tag{9}$$

The process is then deemed to be out-of-control at time $t$ if $GC_t > UCL$, where $UCL$ is a predefined upper control limit. Guidelines to choosing the value of the $UCL$ is discussed in subsection 3.1.5.

Recall that there are two main ways of constructing the MCUSUM statistic as described in (Pignatiello and Runger 1990). The two methods differ in the order in which the accumulation and

the quadratic transformation is performed. The first method performs the accumulation first by calculating the individual local CUSUM statistics and then combining them into a single quadratic form. On the other hand, the second method calculates local Hoteling T-square statistic (quadratic form) and then performs the accumulation using a univariate CUSUM on the result. Here we adopt the first approach of MCUSUM in equation (9) when constructing the global statistic as it fits well with the framework described in the previous section.

### *3.1.3 Sensor Reassignment*

Sensor reassignment is simply reassigning the sensors to the sets of observing sensors $\omega \subset \Omega$ and non-observing sensors $\psi \subset \Psi$ at each time step. We propose to choose the set of observing sensors that maximizes the global statistic in (9) to have the potential to detect the true change quickly. Mathematically, at each time step before taking any observations, our proposed sensor reassignment method is to choose the set of observed sensors $\omega \subset \Omega$ that is the solution to the following optimization problem:

$$\arg\max_{\omega \subset \Omega} \left\{ C_{k,t} \Sigma_{\omega\omega}^{-1} C_{k,t} \right\}, k = \{n: X_n \in \omega\} \ and \ \omega \subset \Omega \quad (1), \tag{10}$$

where $C_{i,t}$ is the local CUSUM statistic of sensor $i$ at time $t$ and $\Omega$ is the set of all possible sensor subsets of size $q$.

While the above optimization problem in (10) is well-defined from the mathematical viewpoint, it becomes very challenging to solve from the computational viewpoint, especially in high dimension situations, as the set of candidate solutions $\Omega$ becomes too large. Therefore, we propose the following greedy forward selection heuristic method to solve the combinatorial optimization problem in (10).

We start with $\omega = \Phi$; the empty set. The first variable to enter the set $\omega$ will be the variable that maximizes equation (9) when the cardinality of the set is one. The solution is the variable with the maximum local CUSUM statistic $\{X_i: C_{i,t} \geq C_{j,t} \ for \ all \ j\}$. If we partition the covariance matrix of the standardized data into the following block form,

$$\Sigma = \begin{bmatrix} \Sigma_{\omega\omega} & \Sigma_{\omega\psi} \\ \Sigma_{\psi\omega} & \Sigma_{\psi\psi} \end{bmatrix}.$$

then the inverse can be written as:

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{\omega\omega}^{-1} + \frac{1}{b}FF' & -\frac{1}{b}F \\ -\frac{1}{b}F & \frac{1}{b} \end{bmatrix},$$

where $F = \Sigma_{\omega\omega}^{-1}\Sigma_{\omega\psi}$ , and $b = 1 - \Sigma_{\psi\omega}\Sigma_{\omega\omega}^{-1}\Sigma_{\omega\psi}$. Hence if we let $G^{\omega} = C_{\omega,t}\Sigma_{\omega\omega}^{-1}C_{\omega,t}$, the global statistic with respect to the set ω, then the global statistic with respect to the joint set $\{\omega \cup \psi\}$ is

$$G^{\omega\cup\psi} = G^{\omega} + \frac{1}{b}\{C_{\psi,t}^2(1-F) - C_{\psi,t}F + FF'\}, \tag{11}$$

which means that the gain in the global statistic after adding variables in set ψ to set ω can be represented by the following:

$$G^{\omega\cup\psi/\omega} = G^{\omega\cup\psi} - G^{\omega} = \frac{1}{b}\{C_{\psi,t}^2(1-F) - C_{\psi,t}F + FF'\}. \tag{12}$$

The following variable to enter the set ω will be the variable that maximizes (9) when the cardinality of the set is two given that the first chosen variable is $X_i$. This translates to the variable $X_j$ that maximizes the gain given by (12) when the set $\omega = \{X_i\}$ and the set $\psi = \{X_j\}$. Consequently, at any step, the next variable to enter set ω given its current cardinality is the variable that maximizes the gain. The steps at each iteration of this heuristic is illustrated in algorithm 1.

| Algorithm 1: Greedy Forward Sensor Selection to Solve Equation (10) | |
|---|---|
| | Input: Empirical covariance matrix Σ, scalar $r, C_{k,t}$ for all k |
| | Forward sensor selection strategy:<br>While $(\lvert\omega\rvert < r)$, |
| 1 | Calculate the gain $G^{\omega\cup j/j}$ for all variables $X_j \notin \omega$ according to eq.(12) |
| 2 | Augment the set of ω by including $\{X_i: G^{\omega\cup i/i} \geq G^{\omega\cup j/j}\ for\ all\ j\}$ |
| 3 | Update the global statistic $G^{\omega}$ |
| | End |

The initial assignment of sensors in the sets ω and $\psi$ has no significant impact to the monitoring procedure (Liu et al. 2015). This is due to the adaptive nature of the sampling strategy that reassigns the sensors at each observation time.

### *3.1.4 Properties of CDS*

This subsection illustrates two behavioral properties of the proposed CDS procedure. These two properties address the desire to disperse sensor deployment when the system is running smoothly under the in-control state, while also quickly localizing at a fault location whenever a true fault occurs. Proofs of the proposed properties can be found in Appendices A.1 and A.2.

Recall that a variable $x_{i,t} \in \omega_t$ if and only if it is observed at time $t$. Thus, at a given time $t_0$, the set of sampled variables $x_{k,t_0}$ is given by $\omega_{t_0}$. The following property shows that when the process is in-control or when those variables in $\omega_{t_0}$ involve insignificant mean shifts, our proposed sensor deployment procedure will eventually choose variable $x_{k',t_0}$ that does not belong to a neighborhood of $\omega_{t_0}$. This implies the random behavior of our dynamic sampling method under the in-control phase, where sensors will be sampled infinitely many times as the $UCL \to \infty$. This essentially guarantees that the sensor deployment procedure will not permanently localize at any specific location.

**Property 1:** For a fix time $t_0$, we assume that $|E[x_k]| \leq \Phi(1-\alpha/2)$ for any $x_k \in \omega_{t_0}$. Consider another variable $x_{k'} \notin \omega_{t_0}$ satisfying $corr(x_{k'}, x_k) = 0$ for all $x_k \in \omega_{t_0}$. Let $UCL \to \infty$, and denote $T_{t,k'} = inf\{t \geq t_0 : x_{k'} \in \omega_t\}$, then $P(T_{t,k'} < \infty) = 1$.

Next, we will show that when a significant mean shift occurs, our proposed sensor deployment procedure has a greedy property that eventually sticks to the fault area, or to its neighborhood when we do not have enough sensors to cover the whole fault area.

**Property 2:** Denote the fault area as $\mathcal{O} = \{x_k : |E[x_k]| > \Phi(1-\alpha/2)\}$. Let $UCL \to \infty$, there exists $\mathcal{O}_0 \subseteq \mathcal{O}$ such that $P_1(O_0 \subset \omega_t \; for \; all \; t \geq t_0) = 1$ for some $t_0$.

In the event that the process is out-of-control, the second property suggests that sensors localized at the fault area will remain deployed within its neighborhood. When a fault is detected in an area, it is desired to check that area as well as its surroundings, because the main issue may be in the neighborhood rather than the initially detected location. Therefore, we are only interested in showing that a remote location, relevant to the fault area, will not be a point of interest for future

sampling. This level of flexibility allows the sampling procedure to better localize around the faulty area rather than simply sticking to an initial suspect area.

### *3.1.5 Overview of the CDS Algorithm*

Algorithm 2 illustrates the steps of the proposed CDS procedure. Compared to other procedures that assume spatial independence, our approach uses the correlation structure and the information obtained from observed sensors to dynamically compensate unobserved ones. If an unobserved sensor is positively correlated with an out-of-control sensor, then the sensor will be compensated more than the one that is independent. This makes it more likely to choose that sensor in the next step. This property will be demonstrated in the case studies in Section 5.

| Algorithm 2: Online monitoring by our proposed Correlation based Dynamic Sampling (CDS) algorithm | |
|---|---|
| | Input: Empirical covariance matrix Σ, scalar $r, \delta, \mathrm{UCL}, \alpha \in (0,1)$<br>$C_{k,0} = C_{k,0}^{+} = C_{k,0}^{-} = 0$ for all sensors |
| | For each time step t=1,2,… |
| 1 | Take observations from sensors based on the current assignments to ω and $\psi$ based on the top-r sensors at time $t-1$ |
| 2 | For sensor $k \in \omega$, recursively compute the local statistics $C_{k,t}$, $C_{k,t}^{+}$, and $C_{k,0}^{-}$ according to equation (2) |
| 3 | For sensor $k' \in \psi$, recursively compute the local statistics $C_{k',t}$, $C_{k',t}^{+}$, and $C_{k',0}^{-}$ according to equations (3) and (4) |
| 4 | Reassign sensors to the sets ω and $\psi$ according to Algorithm 1 and take observations from the updated set ω |
| 5 | Obtain the global statistic $GC_t$ based on the updated set ω from step 4 |
| 6 | If $(GC_t \geq UCL)$, raise a global alarm that a change has occurred. Otherwise go to next time step. |

Table 1 provides a miniature example with three variables to illustrate the steps of algorithm 2. In this example, we assume all three variables have mean 0 and unit variance, $\mathrm{cov}(x_1, x_2) = 0.5$ and that $x_3$ is independent of the others. Initially we only observe $x_1$; i.e., $\omega = \{x_1\}$ and $\psi = \{x_2, x_3\}$ and we set the parameters to be $\delta = 1$ and $\Phi(1-\alpha/2) = 1.04$ $(\alpha = 0.3)$.

Table 1 Demonstration of the compensation and assignment steps of CDS

| | | In-control | | | Out-of-control | | |
|---|---|---|---|---|---|---|---|
| | | $x_1$ | $x_2$ | $x_3$ | $x_1$ | $x_2$ | $x_3$ |
| $t=1$ | **Observation** | -1.58 | - | - | 1.76 | - | - |
| | $\boldsymbol{L_{k,t}}$ | - | -1.37 | 1.04 | - | -0.53 | 1.04 |
| | $\boldsymbol{U_{k,t}}$ | - | 0.57 | 1.04 | - | 1.41 | 1.04 |
| | $\boldsymbol{C_{k,t}}$ | *1.10* | 0.87 | 0.54 | *1.26* | 0.91 | 0.54 |
| $t=2$ | **Observation** | 2.20 | - | - | 0.89 | - | - |
| | $\boldsymbol{L_{k,t}}$ | - | -0.42 | 1.04 | - | -0.74 | 1.04 |
| | $\boldsymbol{U_{k,t}}$ | - | 1.52 | 1.04 | - | 1.19 | 1.04 |
| | $\boldsymbol{C_{k,t}}$ | *1.70* | 1.09 | 1.07 | *1.65* | 1.60 | 1.07 |
| $t=3$ | **Observation** | -0.03 | - | - | 0.69 | - | - |
| | $\boldsymbol{L_{k,t}}$ | - | -0.98 | 1.04 | - | -0.80 | 1.04 |
| | $\boldsymbol{U_{k,t}}$ | - | 0.96 | 1.04 | - | 1.14 | 1.04 |
| | $\boldsymbol{C_{k,t}}$ | 1.17 | 1.58 | *1.61* | 1.84 | *2.25* | 1.61 |

During the in-control phase, we notice that the compensation for the neighborhood of the observed variable $x_1$, in this case $x_2$, receives lower compensation than $x_3$. This essentially influences the algorithm to move out of this neighborhood towards the unexplored variable $x_3$. On the other hand, when $x_1$ is out-of-control ($\text{mean shift} = 1.5$), the compensation allocated to its neighborhood ($x_2$) exceeds that of the independent variable $x_3$. This promotes the exploitation of the neighborhood of $x_1$, which is consistent with the second property discussed in subsection 3.1.4.

### 3.2 Setting Input Parameters

Algorithm 2 of CDS is essentially a phase 2 procedure that can be used for high-dimensional datasets where the multivariate normal assumption is appropriate. A phase 1 analysis is required to validate the assumed underlying distribution as well as determining the required input parameters $r, \alpha, \delta$ and the UCL of algorithm 2 if they are not readily available for direct implementation. This section will provide guidelines for determining the values of these input parameters as discussed below.

- *Setting $\delta$:* From the literature review presented in Section 3.1.1 on the calculation of the local CUSUM statistic, $\delta$ represents the smallest mean change magnitude that we are

interested in detecting. In practice, the choice of $\delta$ can be a target value set using engineering knowledge in the application domain.

- *Setting $r$:* The choice of $r$ directly affects the detection power of the monitoring procedure. Setting $r$ to be too large will dilute the contribution of the out-of-control sensors to the global monitoring statistic, thereby causing an undesired delay in the detection of the mean shift. Moreover, $r \leq |\omega|$, where $|*|$ denotes the cardinality of a set. The ideal choice for $r$ would be the total number of variables associated with the faults that are of interest for detection, also referred to as the root causes. However, this is usually unknown unless it can be provided from engineering knowledge. In the case that it is unknown, choosing a small value of $r$ has been shown to be robust to various fault types (Mei 2010).
- *Setting UCL:* The $UCL$ is the threshold that determines when to stop the monitoring procedure and alert the detection of a change. The value of $UCL$ is related to the pre-scribed in-control ARL of the monitoring scheme. The practitioner can determine the optimal $UCL$ value from sufficiently large in-control measurements or via Monte Carlo simulation and bootstrap techniques (Efron and Tibshirani 1994, Chatterjee and Qiu 2009).
- *Setting $\alpha$:* The tuning parameter $\alpha \in (0,1)$ is a very crucial parameter that essentially determines the trade-off between how sporadic the behavior of the algorithm is when the process is in-control and how fast it converges to the faulty sensors when the process is out-of-control. To illustrate this further, as $\alpha$ approaches 0, the local statistic compensation provided to variable $k \in \psi$ will exceed that of variable $k \in \omega$. While this is not an issue when the process is in-control, the algorithm will not be able to converge to a unique set $\omega$ when the process goes out-of-control, as there will always be a variable in $\psi$ with a larger local statistic. On the other extreme, if $\alpha$ approaches 1, variables belonging to set $\psi$ will receive almost no compensation causing the sensor assignment of the algorithm to be static which is clearly undesired.

  To narrow down the choice of α we can initially try to find tighter bounds. From the proof of properties 1 and 2 of our CDS algorithm, the compensation requires $\delta < \Phi^{-1}(1-\alpha/2) < |\delta^*|$. Here, $\delta^*$ is the true mean when the process goes out of control. Generally speaking, $\delta^*$ is unknown and this makes it challenging to get an upper bound.

In order to obtain an appropriate value for α, we simulate the monitoring procedure iteratively with a binary search over the range of α. The criteria for terminating the search is when the percent decrease in standard deviation (denoted by $\upsilon$) of the number of times (denoted by $\eta$) that each variable is assigned to set $\omega$ is less than some predefined value $\zeta$. The details of this procedure are outlined in algorithm 3. The intuition is to determine a choice of α that provides minimum deviation between sensor sampling frequencies while maintaining the pre-specified ARL.

| Algorithm 3: Choosing the value of α | |
|---|---|
| | for $t = 0$, set $\alpha_t = \begin{cases} 2[1-\Phi(\lvert\delta^*\rvert)], & \text{if } \delta^* \text{ is known} \\ 2[1-\Phi(2\delta)], & \text{o.w} \end{cases}$<br>$\upsilon_t = \text{M}$, where M is sufficiently large |
| | for $(i = 1{:}\, I)$, ($I$: Maximum number of iterations) |
| 1 | Generate $N$ instances of $n$ in-control observation for all sensors |
| 2 | Run algorithm (1) for each instance $j$ calculating $\eta_{t,j}$ and $\upsilon_{t,j} = Var[\eta_{t,j}]$ |
| 3 | Let $\upsilon_t = E[\upsilon_{t,j}]$ |
| 4 | If $\lvert\upsilon_t - \upsilon_{t-1}\rvert/\upsilon_{t-1} < \zeta$ ; break loop |
| 5 | set $\alpha_t = \begin{cases} \alpha_t/2, & \upsilon_t < \upsilon_{t-1} \\ 3\alpha_t/2, & o.w \end{cases}$ |
| | End |

### 3.3 Estimating the Precision Matrix

The acquisition of the inverse covariance matrix, also referred to as the precision matrix (Hsieh et al. 2011), is essential for effectively implementing our proposed CDS algorithm. While we assume in our method that the precision matrix is readily available, that is rarely the case in practice. This section highlights two approaches in which we can reasonably obtain it. In practice, the precision matrix can be either learned from historical training data or alternatively imposed by the domain knowledge. The first subsection 3.3.1 provides a brief literature review of methods to empirically obtain an estimate of the precision matrix. While the second subsection 3.3.2 discusses the latter approach of imposing the matrix by leveraging domain knowledge.

#### *3.3.1 Learning the covariance structure from historical data*

The estimation of the precision matrix in high dimension is a nontrivial task that has been an area of interest for many researchers in the past years. In the literature, it is especially used to provide information on the interrelations between variables in graphical models (Scheinberg et al. 2010).

A sparse representation of the inverse covariance matrix is desirable in high dimensional settings, due to the advantages that sparsity offers. When the number of observations is limited, as is the case in many modern high-dimensional statistical problems, sparsity promotes robustness to the estimation process, which translates well to the future observations (Duchi et al. 2012). Moreover, inducing sparsity functions regularize and enhance interpretability and counter overfitting (Scheinberg et al. 2010).

Methods to estimate the precision matrix look into solving the following optimization problem, its dual or some variation of it:

$$\log\det\theta - tr(S\theta) - \rho|\theta|_1, \tag{13}$$

where $\theta = \Sigma^{-1}$, and $S$ is the empirical covariance matrix.

The objective function in (13) is a convex problem that can be solved with interior point methods in $O\left(p^6 \log\left(1/\varepsilon\right)\right)$, however this becomes infeasible for even moderate $p$. Banerjee et al. (2008) used block coordinate decent with a cost of $O(p^4)$ with their proposed algorithm COVSEL. By solving iterative LASSO problems, the graphical LASSO algorithm proposed by Friedman et al. (2008) manages to reduce the computation complexity to $O(p^3)$. The greedy gradient ascent method and alternating linearization methods (Scheinberg and Rish 2009, Scheinberg et al. 2010), as well as the projected subgradient method developed by Duchi et al. (2012) all claim to reduce the complexity to $O(p^2)$. The second order algorithm QUIC proposed by Hsieh et al. (2011) solves iterative quadratic approximations that has a reduced cost of $O(p)$ to find a Newton direction.

### *3.3.2 Imposing the covariance structure*

There are several domains in which prior knowledge of the system being monitored can be used to extract some process characteristics that can help bypass the estimation of the inverse

covariance matrix and alternatively imposing one. A prominent example of such an application is when the data is acquired in the form of images.

If we regard each pixel of an image to be a variable for monitoring purposes, then it is reasonable to assume that the value of any pixel is independent of other pixels given its neighborhood. This assumption can be translated to a precision matrix in which the entries corresponding to two pixels that are not within a certain pre-specified proximity is set to zero. This level of proximity represents the closeness of the values of nearby pixels. Naturally, different areas of an image can have a different level that is suited to the correlation of the pixels in set area. This type of structure imposition will be demonstrated in the solar flare case study discussed in subsection 5.1. Furthermore, the case study explores the sensitivity of the methodology to the choice of imposition since it may be subjective.

We note that the same estimated or imposed covariance matrix will be used in both the in-control and out-of-control phases. This assumption is suitable for systems in which the occurring faults are not expected to change the way sensors relate to each other. For example, in a manufacturing process in which both temperature and pressure are monitored, an occurring fault is not likely to change the fact that pressure will increase as the temperature rises. Moreover, in the case of images, an occurring fault is also not likely to alter the dependency between pixels within the same proximity. However, for systems in which the faults are expected to change how the observable variables relate, this assumption is restrictive and may significantly affect the performance of the method.

# 4. SIMULATIONS

This section serves as an evaluation for the performance of our proposed CDS algorithm. We compare the performance to two state-of-the-art algorithms, TRAS (Liu et al. 2015) and Top-r (Mei 2010). It is very important to note that the Top-r method assumes no limitations in the number of variables that can be observed and thereby has full access to all raw sensors or data streams. We include it in the comparison to illustrate how competitive our proposed method is, even when compared to those without sampling limitations.

### 4.1 Data Generation

In our simulations, the data is generated using the following generative model:

$$X_t = AZ_t + \varepsilon_t, \tag{14}$$

where the observed variables at time $t$ are $X_t \in R^p$, latent variables $Z_t \in R^q$ following a multivariate normal distribution $MN(0, I)$, and white noise $\varepsilon_t \in R^p$ following $N(0, \sigma_\varepsilon I)$. Matrix $A \in R^{p \times q}$ that maps the latent variables into the domain of the observed variables. Hence, the observed variables follow a multivariate normal distribution as well with $MN(0, AA^T + \sigma_\varepsilon I)$.

In the generative model described above, the transformation matrix $A$ controls the sparsity in the covariance of the observed variables $X_t$. If the matrix $A$ is block diagonal, such that each block is of size $p_i \times q_i$ with $\sum_i p_i = p$ and $\sum_i q_i = q$, then the covariance matrix of the observed variables $X_t$ will also be block diagonal with blocks of sizes $p_i \times p_i$. Therefore, as we decrease the block size in the transformation matrix $A$, we induce a higher level of sparsity in the observed variables $X_t$. In our simulations, we chose $p = 1500$ and $q = 150$. The blocks in the transformation matrix are of size $p_i \times q_i = 100 \times 10$ for all $i$, and each block is a random matrix whose entries are i.i.d. Uniform(-1,1) random variables.

### 4.2 Simulation Experiments and Results

We set the parameters of the experiment as follows. The control limits were chosen to achieve an in-control ARL of 200. The mean shifts ($\delta = 0.25,0.5,1,2,4$) were introduced in a single block of latent variables. Only 150 variables from the full observations $X_t$ could be obtained at any given time (i.e $m = 150$). Out of the 150 available observations, the test statistics were constructed using $r = 15$ variables. The confidence size for compensations was set to $\alpha = 0.27$.

We conducted the simulated experiments with the varying shift magnitude that were replicated 1000 times. The experiments were run on MATLAB R2019 on a Windows 10 operating system with AMD Ryzen 7 1700 eight-core 3.0 GHz processors and 16 GB RAM. The average computational time for each iteration of algorithm 2 was 0.036 seconds, which is adequate for many real production systems.

Table 2 Performance evaluations of the CDS algorithm under different shift magnitudes compared to benchmark techniques (Top-r and TRAS)

| | **In-control ARL (standard deviation)** | | | **Out-of-control ARL (standard deviation)** | | |
|---|---|---|---|---|---|---|
| **Shift size** | **Top-r** | **TRAS** | **CDS** | **Top-r** | **TRAS** | **CDS** |
| $\boldsymbol{\delta = 0.25}$ | 214(210) | 222(186) | 220(182) | 56(23) | 74(42) | *51(31)* |
| $\boldsymbol{\delta = 0.5}$ | 212(189) | 226(172) | 223(180) | *24(14)* | 50(31) | 34(21) |
| $\boldsymbol{\delta = 1}$ | 201(175) | 205(181) | 210(172) | *11(5)* | 17(8) | 12(7) |
| $\boldsymbol{\delta = 2}$ | 207(182) | 210(188) | 197(210) | 8(2) | 8(3) | *5(2)* |
| $\boldsymbol{\delta = 4}$ | 221(193) | 220(213) | 210(185) | 1 | 1 | 1 |

Tables 2 demonstrates that the CDS algorithm consistently outperforms the TRAS algorithm by an average 42% reduction in detection delay. Moreover, it is interesting to compare our proposed method to the Top-r procedure, which assumes no limitations on data acquisition. Although it may be expected that it would be better than our proposed CDS procedure due to the full visibility, the detection delay of CDS within 3.5% from the Top-r and can even surpasses it. This can be attributed to the global monitoring statistic of our proposed CDS method, which considers the correlation of the data streams rather than the independence assumption of the other two competing methods.

# 5. CASE STUDY

This section presents a study on two real datasets to showcase the capability of our adaptive monitoring procedure in practical scenarios. The first subsection 5.1 illustrates how the correlation based adaptive method can achieve high performance under limited transmission capacity by leveraging partial images obtained from video recording of solar flare occurrences. The second subsection 5.2 demonstrates how adaptive sampling can be utilized to monitor in-line Raman spectroscopy for CNTs manufacturing.

## 5.1 Solar flare detection

The detection of solar flares via satellite imaging is an example of a monitoring process that generates high dimensional data in which the occurrence of solar flares is regarded as the change (defect). The solar flare phenomenon, a result of various dynamical processes in the solar atmosphere, is a sudden brightening that can last from 1-15 minutes (Parker 1963). The energy released from this phenomenon can interfere with radio communications by disturbing Earth's ionosphere (Augusto et al. 2011). This serves as motivation to detect these flares upon onset with minimal delay.

Solar flare images are captured and generated in high volume at each second during the satellite's recording. Modern solar flare imaging instruments can acquire images at a frame rate up to 25 frames per second with a memory capacity of 16TB. When continuous observation of the solar disk is viable and recording at a frame rate of 25 $s^{-1}$, around 8TB of data could be acquired in a day. Due to the enormous amount of data and relatively limited memory of the imaging instruments, only one data set every 1 to 5 seconds can be archived and the rest is eliminated (Ishii et al. 2013). While it is beneficial to use all the captured frames to rapidly detect the transient (1-15 min) solar flares, transmission of the enormous amount of data is a challenge. Moreover, processing the large data with methodologies for image change detection such as (Yan et al. 2018) may not be suitable for detecting solar flares in real time. Such methods analyze the full data streams, which likely exceed the transmission and processing capabilities during online monitoring. An adaptive sampling technique that can selectively transmit partial frames not only reduces the amount of data but may also preserve information relevant to flare detection by not eliminating whole frames intermittently.

The solar flare dataset used in this study is publicly accessible in video format at http://nislab.ee.duke.edu/MOUSSE/index.html. The data is collected from satellite images that are taken at very high frequencies. Each video contains 300 frames, and at each frame is of size $232 \times 292$ pixels resulting in a total dimension of 67744 pixels. This is a very high dimensional dataset to process, especially when the number of available observations is relatively small. There are two clear occurrences of solar flares that are visible at frames t=187~202 and t=216~268, respectively.

Pre-processing of the raw data before implementing our proposed methodology is required due to the varied intensities of the pixels of original images. The time dependency of the raw pixels results in a background that is not suitable for directly using CUSUM charts for change detection. Thus the pre-processing is needed to address (i) the autocorrelation between successive frames by removing the background and (ii) the normalization of pixels to remove the effect of varying intensities. We accomplish these tasks by differencing the data with a moving average window of size 4 in a manner similar to that used by Liu et al. (2015). The processed data $X'_t$ is calculated through the relation:

$$X'_t = X_t - \frac{1}{4}\sum_{i=1}^{4} X_{t-i}\,, \quad \text{for } t \geq 5$$

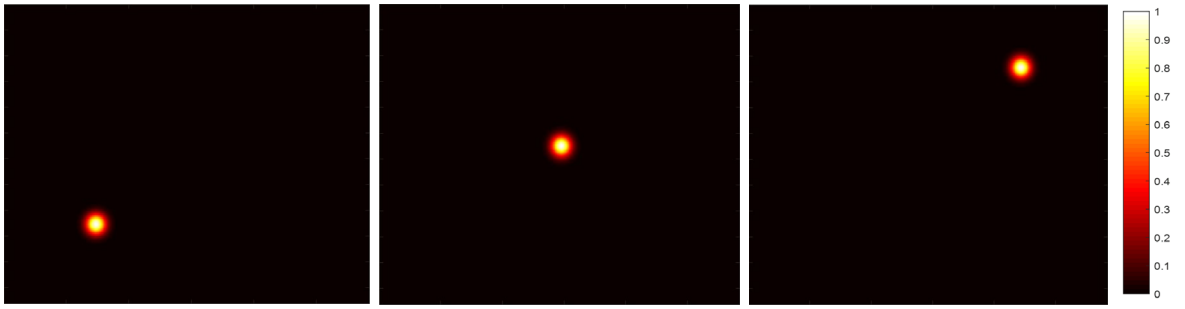
$$X'_t = X_t - \frac{1}{t-1}\sum_{i=1}^{t} X_{t-i}, \text{ for } 2 \leq t \leq 4 \text{ , and } X_1 = 0.$$

The remaining data after removing the background was found to be approximately normal, as was the case in the study in (Xie et al. 2013).

Next, we conduct the phase 1 analysis in which we consider the first 100 frames to be historical data. This will provide the input parameters of algorithm 2 for the phase 2 analysis. We set $\delta = 1$ and $\alpha = 0.27$ via algorithm 3, which corresponds to $\Delta = 0.1$. It should be noted that several manipulations of the previous parameters also yield similar results to the ones chosen here. We further assume that the number of pixels that can be transmitted for analysis at any acquisition time to be 1000 out of the available 67744 pixels in a full frame (image) of the video. In other words, the parameter $m$ is equal to 1000, while we set $r = 40$. Moreover, we set the UCL such that the false alarm rate is 0.0004 as was proposed by Liu et al. (2015). This was achieved via bootstrapping the first 100 frames with replacement.

It remains to determine the spatial covariance structure in which we opt to choose the imposition approach discussed in subsection 3.3.2 to be exponentially decaying with radius of 20 pixels. This is typically appropriate for images in general and particularly for images of solar flares, which often occur in a local area. We later discuss the sensitivity of the performance to this imposition. This specific covariance imposition is demonstrated by Figure 1, in which three plots illustrate the imposed covariance structure over the three different pixels (17107, 34214, 51321). For example, figure 1(b) is an image that is obtained when the $34{,}214^{\text{th}}$ row/column vector of the

67744×67744 pixels covariance matrix is reshaped into a 232×292 matrix, which corresponds to the dimensions of a video frame. This serves to illustrate that any given pixel is only correlated with other pixels in its proximity.




(a) (b) (c)

Figure 1 (a), (b) and (c) illustrate the imposed covariance structure over the three different pixels.

Given the determined input parameters, we can now start the phase 2 analysis for online monitoring the remaining 200 frames sequentially by implementing our proposed CDS procedure in algorithm 2 to the pre-processed $X'_t$ that remove the background based on a moving average window of size 4. In Figures 2 and 3, the images in (a) show the original frames from the raw video with the solar flare slightly noticeable on a small curvature located in the upper left corner, (b) demonstrate the observable variables, that are sampled using the benchmark TRAS method, by representing them with white pixels, and (c) illustrate the dynamic sampling results from our proposed CDS method.

Frame 186 is approximately the frame that precedes the first solar flare occurrence. Figure 2 shows that both methods behave in a random fashion, which is desirable since the process is essentially still in control (i.e. a flare has yet to occur). This can also be seen from the images at frame 215, before the second flare, which also serves to demonstrate the capability of our CDS algorithm to return to the random behavior after the end of the first flare.

Frame 198 represents the moment when the solar flare is the brightest. Figure 3 illustrates the sampled pixels at this frame as well as frame 230, when the second flare is brightest. Our

proposed CDS algorithm covers the flare area completely in both occasions. On the other hand, they are only partially covered when using methods that do not consider the correlation structure.

To demonstrate the capability to localize faster than the competing TRAS algorithm, we provide a side-by-side comparison of the sampled variables (white pixels) using the two methods before and after the detection of a flare. Figure 4 (a, b) shows the sampled pixels right before and after the detection of the flare using the benchmark TRAS algorithm at frames 194 and 195, respectively. Figure 4 (c, d) shows the sampled pixels right before and after detection by the CDS algorithm at frames 190 and 191, respectively. The ability of the CDS algorithm to outperform the TRAS algorithm, with regards to detection delay, can be attributed to the significantly faster localization. This can be clearly observed from the instantaneous localization within a single frame.

With only 1.5% pixels available from the 67744 pixels per frame, our proposed algorithm can detect the flare at frame 191; only 4 frames after its onset at frame 187. Liu et al. (2015) reported the detection of the change at frame 190 when 2000 pixels were observed at any time. While as shown in the figure, this performance deteriorates to frame 195 when the amount of pixels is cut to 1000. Our proposed CDS algorithm with only half of the resources can thus still compete with that performance due to the superior localization strategies.

Figure 5 plots the global monitoring statistic of the proposed CDS algorithm from frame 100 to the end of the captured video at frame 300. For comparison, Figure 5 (b) illustrates the monitoring statistic obtained from the competing TRAS algorithm. The first 100 frames were considered a training sample and were used to obtain the control limits using a bootstrap procedure. The control limits for both CDS and TRAS algorithms were set to a pre-specified in-control ARL of 2500 were determined to be 970 and 950, respectively. The occurrence of the second flare was very close to the first and therefore Figure 5 only shows the monitoring statistic crossing the threshold once.

This is because the 14 frame difference between the end of the first flare and the beginning of the second is insufficient to reset the declining statistic. In such scenarios, the statistic can be simply reset upon resolving the preceding out of control occurrence. In this study, the monitoring statistic was reset at frame 203 after the end of the first flare.

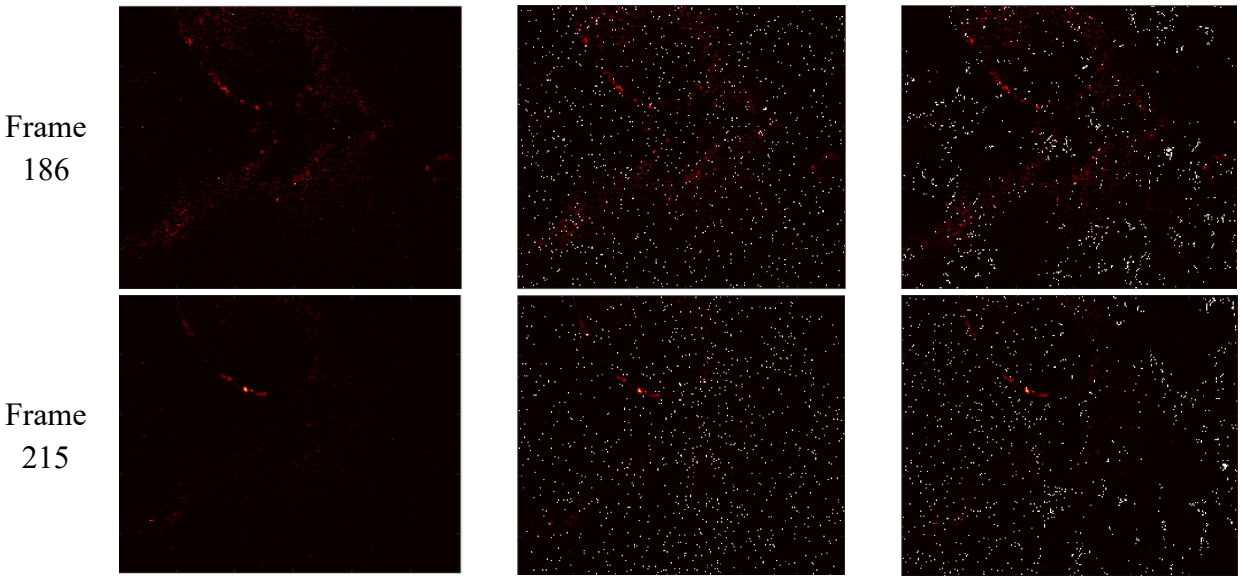



Figure 2 Monitoring frames before the two flares: (a) frame capture from video; (b) sampled pixels from the TRAS algorithm; (c) sampled pixels from the proposed CDS algorithm

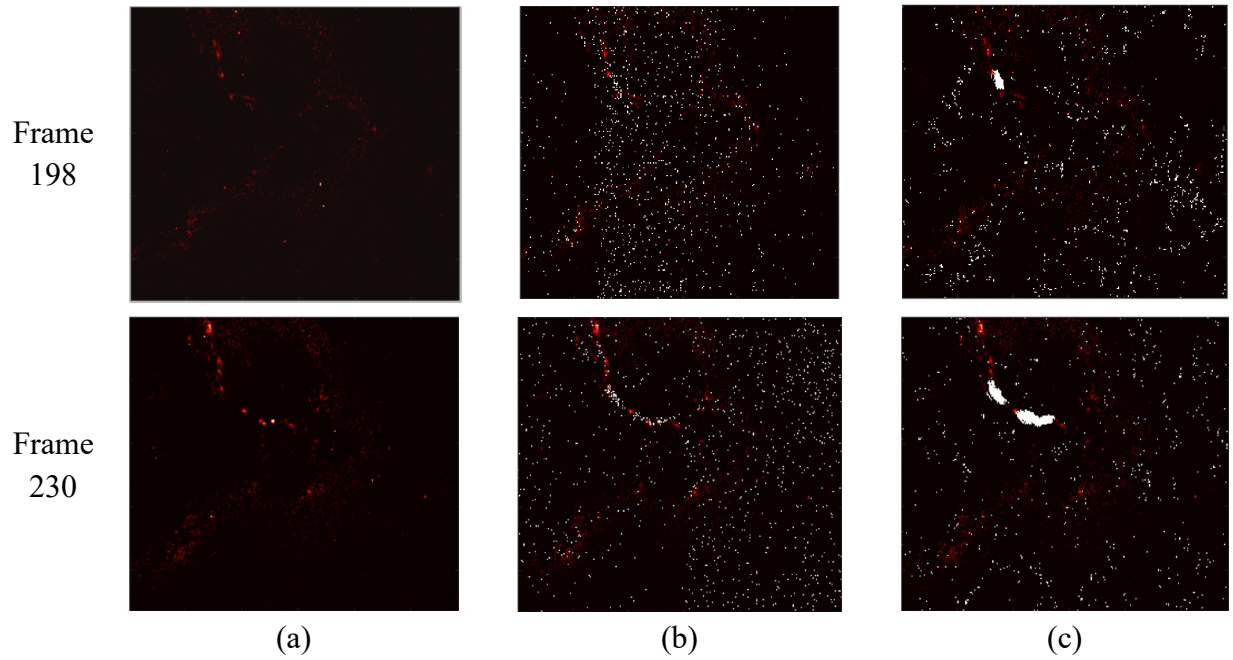
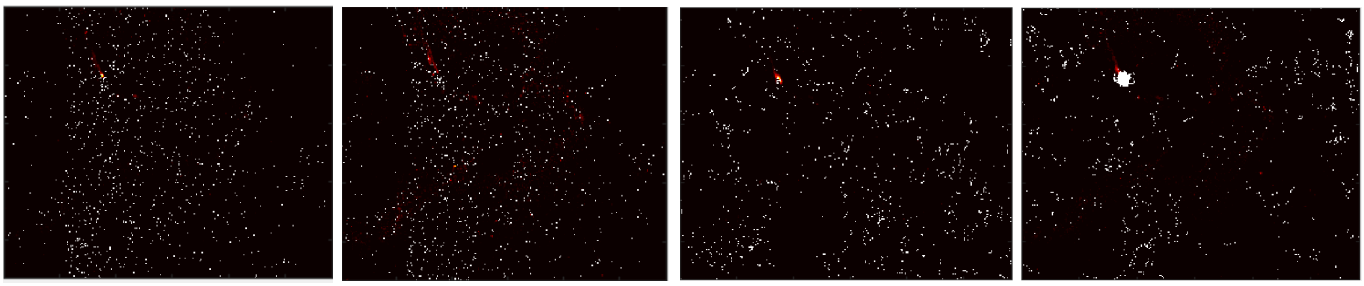
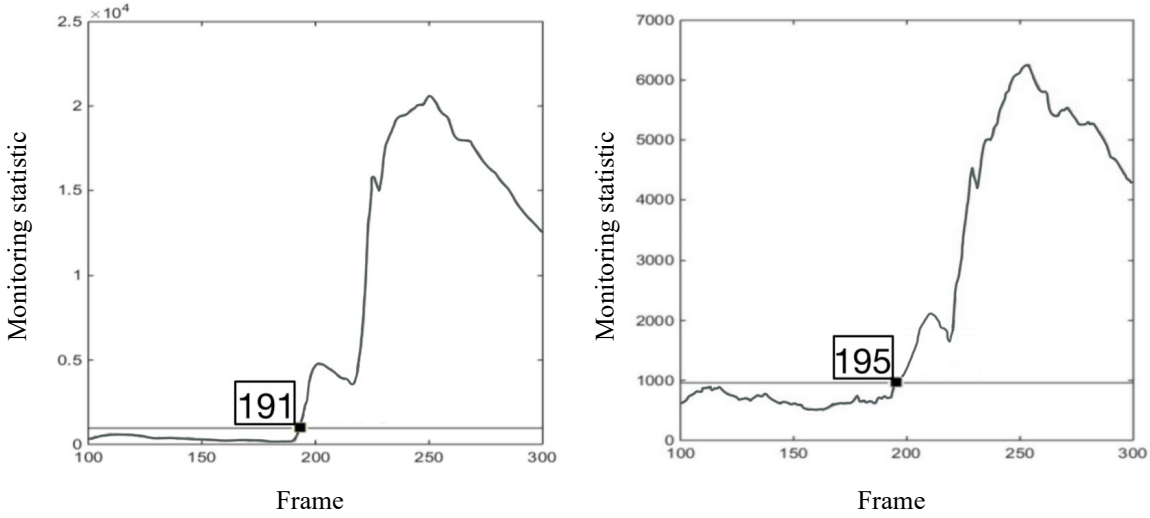



Figure 3 Monitoring at the solar flares peak: (a) frame capture from video when the flare is the brightest; (b) sampling from TRAS; (c) sampling from CDS.



(a) (b) (c) (d)

Figure 4 Detection of the first flare: (a), (b) sampling from TRAS right before and after detection; (c), (d) sampling from CDS right before and after detection




(a) Proposed CDS algorithm performance (b) TRAS algorithm performance

Figure 5 Monitoring statistics by respectively implementing the CDS/TRAS algorithms with the detection frames illustrated by the data cursors

Similarly, the CDS algorithm is capable of detecting the second flare at frame 219, only 3 frames upon onset. Meanwhile, the competing TRAS algorithm lags by 7 frames, and detects the flare at frame 223. The detection in 3 frames not only beats the TRAS algorithm under the same limitations, but also outperforms the reported detection time of frame 221 reported in (Liu et al. 2015), which had double the visibility.

Moreover, let us discuss the sensitivity of the performance of the CDS method to the imposed precision matrix. Initially, we impose the precision matrix with an exponential decay with

a radius size of 20 pixels. We aim to evaluate the change in performance as the radius size varies from 1 pixel to 100 pixels. It should be noted that a radius of 1 pixel results in the imposition of an identity matrix, which essentially means that we assume that the pixels are independent. So, we can expect that the performance of the CDS method will converge to the benchmark TRAS method as the radius goes to 1. Table 3 reports the resulting average detection delays and standard deviations from an analysis with 100 iterations.

Table 3 Average detection delays (standard deviations) of the CDS method under imposed precision matrices with different radius sizes with TRAS as the benchmark

| | | **Radius size** | | | | |
|---|---|---|---|---|---|---|
| | | **1** | **10** | **25** | **50** | **100** |
| **First Flare** | **CDS** | 7.88(1.47) | 5.72(0.46) | 4.17(0.13) | 5.85(0.83) | 7.24(1.26) |
| | **TRAS** | 7.93(1.52) | - | - | - | - |
| **Second Flare** | **CDS** | 6.75(1.19) | 4.66(0.37) | 3.58(0.09) | 4.90(0.99) | 5.86(1.10) |
| | **TRAS** | 6.81(1.15) | - | - | - | - |

The results from the sensitivity analysis in Table 3 indicate that a small deviation from the initial choice of a 20-pixel radius yield similar results. Moreover, when the radius is reduced to 1, the performance of CDS is statistically equivalent to that of the benchmark TRAS method, which is to be expected from the proposition in subsection 3.1.1. It is important to note that the standard deviation of the detection delay first decreases as the radius size increases from 1 to 25, but then increases as the radius size increases from 25 to 100. This can be attributed to the fact that the radius size controls the tradeoff between exploration and exploitation, as a too small or too large radius size results in a slower localization around the fault area and yields an increased variation. In particular, our proposed CDS method will heavily favor exploration over exploitation when the radius is size is very small, whereas a very large radius does the opposite.

Finally, we need to point out the computational challenges of online monitoring in Phase 2 analysis of solar flare when using our proposed CDS algorithm. The main challenge is to find the suitable choice of the control limit, $UCL$. It took our personal laptop (Windows 10 Laptop with Intel i7-4700MQ CPU 2.40 GHz and MATLAB R2018b) about 2 hours to find the control limit that satisfies the ARL to false alarm constraint of 2500 by bisection search method based on 2500

Monte Carlo runs. If one wants to increase the ARL to false alarm constraint, then it will become more time-consuming. The good news is that after the value of $UCL$ is determined, then it is straightforward to implement our proposed CDS algorithm when online monitoring solar flare frames. It took our personal laptop 12.6 seconds to online monitor the 200 testing frames and generate the monitoring statistics in Figure 5. This is 20 times faster than the existing method of archiving and analyzing one full data every 1 to 5 seconds (Ishii et al. 2013), and thus our proposed CDS algorithm is efficient from both computational and statistical viewpoints.

### 5.2 Fault detection of in-line Raman spectroscopy

In this subsection we evaluate the performance of our methodology in addressing the challenges of monitoring the production process of continuous carbon nanotubes (CNTs) buckypaper using inline Raman spectroscopy. The monitoring of the manufacturing process of CNTs buckypaper manufacturing in real time using in-line Raman spectroscopy has gained much interest recently (Yue et al. 2018). The ability to monitor this process in real time is critical to scale up while meeting high quality standards. However, it is challenging to detect changes in the data collected from this procedure since there are several sources for variation in Raman spectrums.

One source of variation is related to the scanning duration when obtaining the signals. Characterization of an in-line Raman spectrum may take multiple scans with a duration of ten seconds to several minutes. The longer the scanning duration the higher signal to noise (S/N) ratio, due to the rapidly moving samples. Figure 6 illustrates acquired Raman spectrums from two operating conditions (red, blue), in which the blue signals are obtained from shorter scanning durations and lower intensity at the peaks.

The higher S/N ratio, such as that of the red signals, is desirable for process monitoring. However, it comes at a cost of longer scanning durations which may delay fault detection. On the other hand, the rapidly obtained signals from shorter scanning durations may mask faults with the excessive noise. In this case, it may be beneficial to save scanning time while retaining a higher S/N ratio by acquiring partial signals. Profile monitoring methods in the literature typically approach this problem from a denoising perspective (Yue et al. 2018). Here we want to demonstrate that our proposed CDS method provides a useful alternative approach to this problem from the adaptive sampling perspective.

In a similar way to the first study, we begin the phase 1 analysis which allows us to determine the input parameters of algorithm 2 for the phase 2 analysis. The data set consists of 200 in-control profiles and 50 out-of-control instances, where the dimension of each profile is $p = 512$. The first 200 in-control profiles are used as historical data, which is used for normalization and setting the input parameters. In order to obtain signals with a high S/N ratio, approximately 10% of the Raman spectra $(m = 50)$ will be measured at any given time. For each method, a threshold that satisfies an in-control average run length (ARL) of 500 is determined by bootstrapping the 200 in-control samples. The remaining parameters are set to the following: $r = 25, \delta = 1$ and the compensation significance level was found to be $\alpha = 0.23$ $(\Delta = 0.21)$ via algorithm 3. Figure 6 illustrates the mean of out of control signals, where it can be noted that the shift approximately within the index interval [95,115] of the Raman spectrum. Finally, the covariance matrix is estimated from the first 100 in-control data using the method QUIC (Hsieh et al. 2011), which is the precision matrix estimation technique discussed in subsection 3.3.1.

This study compares our CDS procedure against the same benchmark methods in the other studies; TRAS and Top-r. The Top-r method requires full observations and therefore will be applied to data with low S/N ratios. While the two adaptive monitoring schemes (CDS, TRAS) will be implemented on partial data with high S/N ratio as illustrated by the red profiles of Figure 6 (left).

The results from implementing the different monitoring schemes are presented in Figure 7. Our CDS procedure outperforms the other benchmark methods and signals an alarm at time 209, which is 9 epochs upon failure onset. Since TRAS does not take into account the correlation structure between variables, it is unable to quickly localize at the fault region. This results in a detection delay of 18 epochs. Finally, the Top-r procedure achieves the lowest performance among the three with a detection delay of 21 epochs, even though it was implemented on complete data. This is because complete data collected at a high frequency comes at the cost of low S/N ratios as we have discussed in the introduction of this study. Hence, we can conclude from this study that there are practical scenarios where it may be beneficial to sacrifice sensor visibility in exchange for better quality data. This emphasizes the importance of making educated decisions on which sensors to acquire in real time.

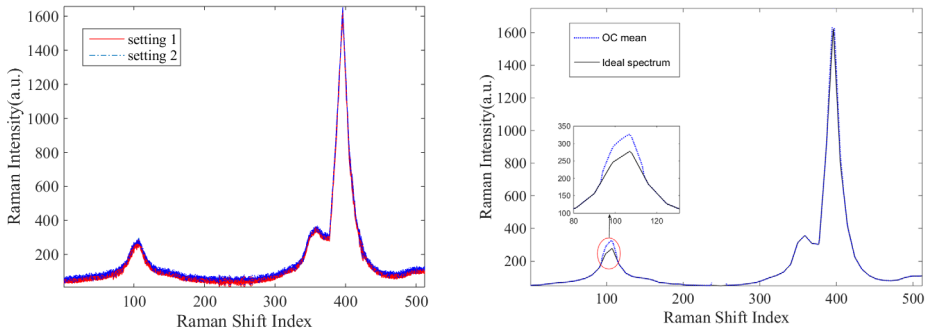



Figure 6. Left: illustration of the Raman spectra data. Right: illustration of out-of-control Raman spectrum mean shift

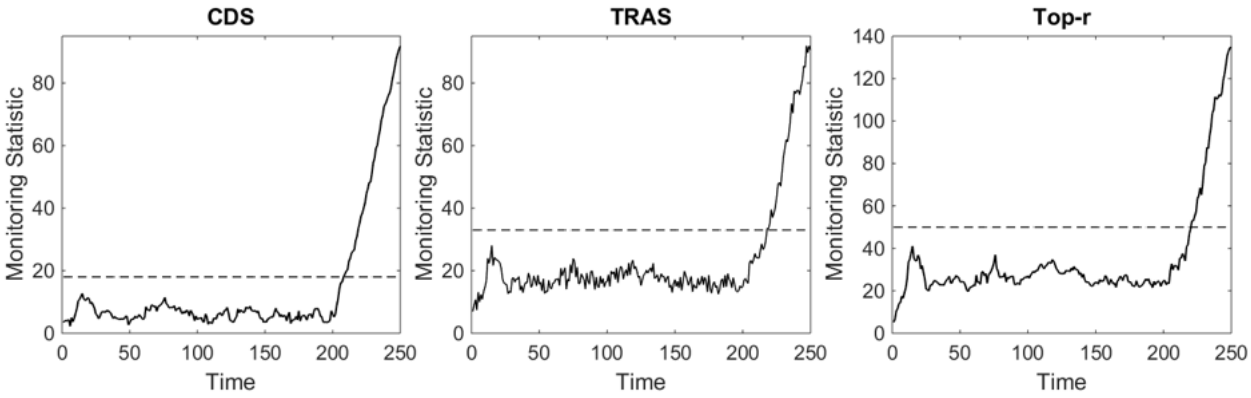



Figure 7 Monitoring statistics for in-line Raman spectra

## 6. CONCLUSION

The development of sensing technologies that generate high dimensional data has offered unprecedented process monitoring capabilities. However, with this advancement rose new challenges that require novel monitoring schemes in limited resources due to sensor availability for deployment, transmission capacity and computational power. Hence, the application of multi-armed bandit algorithms to the SPC context is useful to tackle the issue of efficient monitoring under the limited resources environments.

This paper proposes a novel correlation-based dynamic sampling strategy that constructs a dynamic compensation factor to unobserved data streams. This is performed by using the idea of

celebrated upper confidence bound (UAB) algorithm from the multi-armed bandit (MAB) problem, as well as by utilizing the correlation structure between the observed and unobserved streams. A novel integration of the Top-r procedure with multivariate CUSUM is developed to construct the global monitoring statistic used for decision making related to the state of the process. This results in a strategy that is effective in monitoring high dimensional data streams with partial observations, which consequently reduces the computational cost at the data fusion center. Moreover, utilizing the correlation structure embedded in the data streams allows for faster localization at the fault source while maintaining a random sampling behavior when the process is in-control, which was illustrated by the two properties of the dynamic sampling behavior. This allows this method to be suited for a wide area of applications, such as network processes and images as was demonstrated in the solar flare case study. Additionally, it can be implemented in advanced industrial manufacturing operations as showcased by the in-line Raman spectroscopy case study.

Note that our proposed CDS algorithm is designed to detect sparse mean shifts of high-dimensional data in the resource limited environments under the assumption that the occurrence of faults did not influence the relation between system variables. In many real-world applications, faults may alter the dependencies between system variables, and it will be interesting to develop algorithms that are able to effectively detect the changes on the spatial correlation structure in the resource limited environments. This remains an open problem, as it is unclear how to effectively estimate the post-change spatial correlation structure of high-dimensional data using partially observable data. Hopefully this paper can stimulate further research on SPC for high-dimensional data in the resource limited environments.

## Acknowledgements

The authors thank the Editor and two anonymous referees for their thoughtful and constructive comments that greatly improved the quality and presentation of this article. This research was supported in part by National Science Foundation (NSF) grants CMMI-1362876 and DMS1830344, through Georgia Institute of Technology.

REFERENCES


Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. "Finite-time analysis of the multiarmed bandit problem." *Machine Learning* 47 (2-3):235-256.

Augusto, C. R. A., A. C. Fauth, C. E. Navia, H. Shigeouka, and K. H. Tsui. 2011. "Connection among spacecrafts and ground level observations of small solar transient events." *Experimental Astronomy* 31 (2-3):177.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont. 2008. "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data." *Journal of Machine Learning Research* 9 (Mar):485-516.

Ben-Gal, I., and E. Kagan. 2013. *Probabilistic search for tracking targets: Theory and modern applications*: John Wiley & Sons.

Chatterjee, S., and P. Qiu. 2009. "Distribution-free cumulative sum control charts using bootstrap-based control limits." *The Annals of Applied Statistics*:349-369.

Ding, Y., E. A. Elsayed, S. Kumara, J.-C. Lu, F. Niu, and J. Shi. 2006. "Distributed sensing for quality and productivity improvements." *IEEE Transactions on Automation Science and Engineering* 3 (4):344-359.

Duchi, J., S. Gould, and D. Koller. 2012. "Projected subgradient methods for learning sparse gaussians." *arXiv preprint arXiv:1206.3249*.

Efron, B., and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*: CRC press.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9 (3):432-441.

Frost, J., and L. D. Stone. 2001. Review of search theory: advances and applications to search and rescue decision support. *Soza and Company LTD Fairfax VA*, Report No. CG-D-15-01.

Gut, A. 1988. *Stopped Random Walks - Limit Theorems and Applications*. Applied Probability. A Series of the Applied Probability Trust, 5. Springer-Verlag, New York.

Hsieh, C.-J., I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik. 2011. "Sparse inverse covariance matrix estimation using quadratic approximation." *Advances in Neural Information Processing Systems*.

Ishii, T. T., T. Kawate, Y. Nakatani, S. Morita, K. Ichimoto, and S. Masuda. 2013. "High-Speed Imaging System for Solar-Flare Research at Hida Observatory." *Publications of the Astronomical Society of Japan* 65 (2). doi: 10.1093/pasj/65.2.39.

Jin, R., C.-J. Chang, and J. Shi. 2012. "Sequential measurement strategy for wafer geometric profile estimation." *IIE Transactions* 44 (1):1-12.

Lai, T. L. 1987. "Adaptive treatment allocation and the multi-armed bandit problem." *The Annals of Statistics* 15 (3):1091-1114.

Lai, T. L., and H. Robbins. 1985. "Asymptotically efficient adaptive allocation rules." *Advances in Applied Mathematics* 6 (1):4-22.

Li, J., and J. Jin. 2010. "Optimal sensor allocation by integrating causal models and set-covering algorithms." *IIE Transactions* 42 (8):564-576.

Lim, H. B., M. C. Foo, and Y. Zeng. 2006. "An adaptive distributed resource allocation scheme for sensor networks." *International Conference on Mobile Ad-Hoc and Sensor Networks.*

Liu, K., Y. Mei, and J. Shi. 2015. "An adaptive sampling strategy for online high-dimensional process monitoring." *Technometrics* 57 (3):305-319.

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):630-643.

Mandroli, S. S., A. K. Shrivastava, and Y. Ding. 2006. "A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes." *IIE Transactions* 38 (4):309-328.

Mei, Y. 2010. "Efficient scalable schemes for monitoring a large number of data streams." *Biometrika* 97 (2):419-433.

Montgomery, D. C. 2009. *Introduction to Statistical Quality Control*: John Wiley & Sons (New York).

Parker, E. N. 1963. "The Solar-Flare Phenomenon and the Theory of Reconnection and Annihiliation of Magnetic Fields." *The Astrophysical Journal Supplement Series* 8:177.

Pereira, R. L., J. Trindade, F. Gonçalves, L. Suresh, D. Barbosa, and T. Vazão. 2014. "A wireless sensor network for monitoring volcano-seismic signals." *Natural Hazards and Earth System Sciences* 14 (12):3123.

Pignatiello, J. J., and G. C. Runger. 1990. "Comparisons of multivariate CUSUM charts." *Journal of Quality Technology* 22 (3):173-186.

Scheinberg, K., S. Ma, and D. Goldfarb. 2010. "Sparse inverse covariance selection via alternating linearization methods." *Advances in Neural Information Processing Systems*.

Scheinberg, K., and I. Rish. 2009. "SINCO-a greedy coordinate ascent method for sparse inverse covariance selection problem." *IBM Research Report* RC24837.

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):329-343.

Xian, X., R. Archibald, B. Mayer, K. Liu, and J. Li. 2018a. "An effective online data monitoring and saving strategy for large-scale climate simulations." *Quality Technology & Quantitative Management*:1-17.

Xian, X., A. Wang, and K. Liu. 2018b. "A nonparametric adaptive sampling strategy for online monitoring of big data streams." *Technometrics* 60 (1):14-25.

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.

Yan, H., K. Paynabar, and J. Shi. 2018. "Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition." *Technometrics* 60 (2):181-197.

Yue, X., H. Yan, J. G. Park, Z. Liang, and J. Shi. 2018. "A Wavelet-Based Penalized Mixed-Effects Decomposition for Multichannel Profile Detection of In-Line Raman Spectroscopy." *IEEE Transactions on Automation Science and Engineering, 15(3), pp.1258-1271.*

Zoghi, M., and M. Kahaei. 2010. "Adaptive sensor selection in wireless sensor networks for target tracking." *IET Signal Processing* 4 (5):530-536.

# APPENDIX

In this Appendix, A.1 and A.2 provide the proofs for properties 1 and 2 of the proposed CDS algorithm, which were discussed in subsection 3.1.4. The following Lemma 1, which essentially follows from the weak law of large numbers, will be used in the proofs in A.1 and A.2.

**Lemma 1:** For an independent and identically sequence of a bivariate normal random variables $x_t \; and \; y_t$, such that $E[x] = \mu_x > \mu_y = E[y]$:

$$\lim_{T\to\infty} P\left(\sum_{t=t_0}^{T} x_t > \sum_{t=t_0}^{T} y_t\right) \to 1$$

Proof of lemma: Define the random variable $z_t = \sum_{t=t_0}^{T} x_t - \sum_{t=t_0}^{T} y_t$, then $z_t$ is a Gaussian random walk. And we have:

$$\lim_{T\to\infty} P\left(\sum_{t=t_0}^{T} x_t > \sum_{t=t_0}^{T} y_t\right) = \lim_{T\to\infty} P(z_T > 0)$$

By assumption of $\mu_x > \mu_y$, we conclude that $z_t$ is a random walk with a positive drift $E(x_t - y_t) = \mu_x - \mu_y > 0$, then it follows that:

$$\lim_{T\to\infty}(z_T) = \infty$$

and hence,

$$\lim_{T\to\infty} P(z_T > 0) = 1 \blacksquare$$

### A.1 Proof of Property 1

At the high-level, the proof of property 1 involves two subcases. When there are no changes, the local statistics at sensors with infinitely many time observations will go back to 0, whereas the local statistics at those local sensors without any observations and not correlated to observed sensors will be linearly increasing. Hence, we will sample from those non-observed sensors eventually. The second case is for when there is an insignificant change, where the linear increase of the unobserved sensors will still outrun the increase of the observed counterparts.

Since our sensor sampling procedure (algorithm 1) starts with picking elements of $\omega_t$ according to the $\max_i C_{i,t}$, it suffices to show that for any unobserved variable $x_{k'} \notin \omega_{t_0}$ there exists a time $t$ such that $C_{k',t} > \max_k C_{k,t}$. If we take any unobserved variable $x_{k'} \notin \omega_{t_0}$ that is also not in the neighborhood of $\omega_{t_0}$ $\left(\text{i. e. } corr[x_{k'}, x_k] = 0 \text{ for all } x_k \in \omega_{t_0}\right)$, the increments of the positive and negative CUSUM will depend on $U_{k',t} = L_{k',t} = \Phi(1-\alpha/2)$. Then without loss of generality we can only consider the positive CUSUM $\left(C_{k'}^{+}\right)$. Hence, property 1 can be proven by comparing the increments of the CUSUM statistics from elements in $C_k$ to those of $C_{k'}$, and showing that there exists a time $t$ such that $C_{k',t} > \max_k C_{k,t}$. It suffices to show that there exists $T$ such that when $\delta > 0$:

$$\sum_{t=t_0}^{T}\left(\delta U_{k',t} - \frac{\delta^2}{2}\right) > \sum_{t=t_0}^{T}\left(\delta x_{k,t} - \frac{\delta^2}{2}\right),$$

or equivalently,

$$z_T = \sum_{t=t_0}^{T} \left(U_{k',t} - x_{k,t}\right) > 0.$$

The assumption on $|E[x_k]|$ from property 1 can be broken down into two cases. First, we consider the case when $|E[x_k]| < \Phi(1-\alpha/2)$. Since $x_{k',t}$ is not in a neighborhood of $x_{k,t} \in \omega_{t_0}$, $E[U_{k'}] = \Phi(1-\alpha/2)$. Hence, $z_T$ is a random walk with a positive drift and by Lemma 1:

$$P(z_T > 0) \to 1$$

The second case is when $E[x_k] = \Phi(1-\alpha/2) = E[U_{k'}]$. In this case, $z_t$ becomes a Gaussian random walk with no drift. Let $H = inf\{z_t : t \geq 1\}$, then $H \xrightarrow{as} -\infty$ as $t \to \infty$ (Gut 1988). Hence, for any two variables $x_{k',t}$ and $x_{k,t}$ there exists a time $t$ such that $C_{k',t} > C_{k,t}$. ■

### A.2 Proof of Property 2

It suffices to show that increments of significantly out-of-control samples will be greater than the compensation given to the unobserved variables outside its neighborhood. Specifically, if we define $z'_t = \sum_{t=t_0}^{T}\left(x_{k,t} - U_{k',t}\right)$, and $|E[x_k]| > \Phi(1-\alpha/2)$ by the assumption in property 2, then $z'_t$ is a random walk positive drift $(E[x_k] - E[U_{k'}])$. As $t \to \infty$ then $z'_t \to \infty$, this implies that there exists time $t_0$ such that $\forall t \geq t_0\ z'_t \geq 0$ and $C_{k',t} < C_{k,t}$. ■

It should be noted that the speed of the localization here depends on the drift $\left(E[x_k] \quad - \Phi(1-\alpha/2)\right)$: the higher the post mean shift $(E[x_k])$ is, the faster it will diverge to ∞, which translates to quicker localization. Moreover, this shows that the sampling method will not favor a variable outside of the neighborhood. However, that does not mean that it will not explore the neighborhood even after it detects a faulty area. This essentially means that our method will not necessarily stick to the initial faulty area, but may still explore the surroundings to find an even bigger fault.