INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Thompson Sampling-Based Partially Observable Online Change Detection for Exponential Families

Jie Guo, Hao Yan, Chen Zhang

To cite this article:
Jie Guo, Hao Yan, Chen Zhang (2024) Thompson Sampling-Based Partially Observable Online Change Detection for Exponential Families. INFORMS Journal on Data Science 3(2):145-161. https://doi.org/10.1287/ijds.2022.00011

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 3, No. 2, October–December 2024, pp. 145–161
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

Thompson Sampling-Based Partially Observable Online Change Detection for Exponential Families

Jie Guo,[a] Hao Yan,[b] Chen Zhang[a,*]

[a] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China; [b] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, Arizona 85287

*Corresponding author

Contact: guojie19@mails.tsinghua.edu.cn, https://orcid.org/0000-0002-8319-524X (JG); haoyan@asu.edu, https://orcid.org/0000-0002-4322-7323 (HY); zhangchen01@tsinghua.edu.cn, https://orcid.org/0000-0002-8319-524X (CZ)

Received: April 28, 2022
Revised: April 26, 2023
Accepted: July 29, 2023
Published Online in Articles in Advance: January 9, 2024

https://doi.org/10.1287/ijds.2022.00011

Abstract. This paper proposes a holistic sequential change detection framework for partially observable high-dimensional data streams with exponential-family distributions. The framework first proposes a general composite decomposition for exponential-family distributed data by projecting its natural parameter onto normal bases and abnormal bases, which enables efficient inference for sparse changes. Then, the inference results are used for detection scheme construction, and different types of test statistics can be compacted in our framework. Last, by further designing the test statistic as the reward function in the combinatorial multi-armed bandit problem, a Thompson sampling-based sensor allocation strategy is constructed to select the most anomalous variables. Theoretical properties of the detection framework are discussed. Finally, examples of Gaussian, Poisson, and binomial distributed data streams are given in numerical studies and case studies to evaluate the performance of our proposed method.

History: Bianca Maria Colosimo served as the senior editor for this article.

Funding: C. Zhang is partially supported by the NSFC [Grants 71932006 and 72271138], the BNSF [Grant 9222014], and the ASFC [Grant 2020Z063058001]. H. Yan is partially supported by NIH [R21 AI157618] and NSF [CMMI 2316654].

Data Ethics & Reproducibility Note: No data ethics considerations are foreseen related to this paper. The code capsule is available on Code Ocean at https://codeocean.com/capsule/8794940/tree/v1 and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2022.00011).

Keywords: sequential sparse change detection • partially observable data • Bayesian framework • combinatorial multi-armed bandit • Thompson sampling • exponential family

1. Introduction

Sequential change detection has received a considerable research focus in the last two decades (Tartakovsky 2019). Generally, for a process with p-dimensional variables, $\mathbf{X} \in \mathcal{R}^p$, sequential samples are observed, and it assumes $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_\tau \overset{iid}{\sim} f_0$ and $\mathbf{X}_{\tau+1}, \mathbf{X}_{\tau+2}, \ldots \overset{iid}{\sim} f_1$, where τ is an unknown change point. f_0 is the before-change distribution, which is usually assumed known or can be well estimated from historical reference samples, whereas f_1 is the postchange distribution, which is generally unknown. In this problem, it is desired to detect the change as quickly as possible or minimize the change detection delay while maintaining a predefined false alarm rate. In recent years, as the process to be monitored becomes increasingly complicated, an increasing number of variables are of interest and to be monitored, and, consequently, the dimension of $\mathbf{X}_t$ becomes extremely higher. Compared with traditional multivariate change detection, two particular challenges specific to monitoring high-dimensional data streams have attracted extensive attention. First, high-dimensional data pose pressure for modeling their correlation structure, which is known as the curse of dimensionality. As a solution, different types of dimension reduction methods have been proposed and incorporated into the detection schemes, such as principal component analysis (Zhang et al. 2018), dictionary learning (Lu et al. 2016), matrix decomposition (Yan et al. 2014), etc. The second challenge is the sparse structure of change. The change can only be in a quite small subset of p variables—that is, f_1 differs from f_0 in only a small portion of dimensions. This increases the difficulty of change detection dramatically because the sparse change can easily be submerged by the random noise. To deal with this, sparse learning methods, such as sparse principal component analysis (Zhang et al. 2018) and smooth-sparse decomposition (Yan et al. 2017, 2018), are highly recommended to be combined into sequential change detection.

Besides the two challenges above caused by high dimensions, another fundamental obstacle to efficient high-dimensional change detection is sensing resource constraint. In many applications, for each time point,

only m variables, $m < p$, can be accessible because of limited sensing resource constraint, such as in the following three scenarios. (1) The number of sensors may be limited due to high allocating or purchasing costs. For example, in military operations, unmanned vehicles are commonly used for intrusion detection. However, the high cost of these vehicles often limits the number that can be deployed, forcing practitioners to strategically choose areas with high suspicion of insecurity to monitor. (2) Only a limited number of sensors can be set at the "ON" mode due to limited battery life. For example, in environmental monitoring, such as monitoring landslides, earthquakes, volcano eruptions, etc., the cost of replacing sensors' batteries is often high in the harsh environment, so only a limited number of sensors can be set at the "ON" mode due to the limited battery life. (3) Only partial data collected at each acquisition time can be used for real-time analysis due to the limited transmission bandwidth or the computer's memory. For example, when monitoring anomalies in streaming images or videos, the high resolution rate and high acquisition rate put pressure on the transmission bandwidth and also the computer's memory, which makes it difficult to analyze such data in real time. In a word, these resource constraints make it impossible to obtain full observations of all variables at each time point. It results in the partially observable sequential change detection (POSCD) problem, whose solution requires not only a powerful change detection scheme that can deal with partially observable data, but also an adaptive sensor allocation strategy to identify which variables should be observed for the next time point, such that the change information can be observed maximally.

For the adaptive sensor allocation strategy, we aim to sequentially select the subset of variables with the maximal change information, so as to minimize the average detection delay. This is intimately related to the combinatorial multi-armed bandit (CMAB) problem, which is extensively studied in reinforcement learning and online learning. Under the classic setting of a CMAB problem, a bandit is faced with a system of p arms (i.e., variables), and each arm's reward follows an unknown distribution with an unknown mean. At each round, the bandit needs to select a subset of m arms to play and observes their joint reward, where $m < p$. The goal is to maximize the total expected reward in a fixed number of rounds. In the POSCD problem, the reward function should relate to the system change information, given that we would like to maximize the chance of detecting the system change by selecting the variables related to the change. One of the main challenges in CMAB is to balance the trade-off between exploration and exploitation (Agrawal and Goyal 2012, Chen et al. 2013). Exploration encourages us to explore more variables by selecting those arms with large abnormal variance, which means their parameters are not learned precisely so far based on their past collected observations. Inversely, exploitation encourages us to select those arms with large abnormal mean, but little abnormal variance, which means they are expected to locally maximize the change information based on their past collected observations.

In the last five years, increasing works have been proposed for the POSCD problem. However, there are still many research gaps to be filled. First, most of the current works either assume that the variables are independent or that they can only deal with low-dimensional cases. They cannot be extended to high dimensions due to the curse of dimensionality. Even if applying them brutally, they cannot deal with the sparse change as well, leading to their poor detection power. Second, most existing works assume Gaussian distribution for $\mathbf{X}_t$, whereas there are many cases far from Gaussian distribution in practice. For example, the number of emails arriving at a mailbox per day conforms to Poisson distribution; The operation time consumed by a worker in a manufacturing workshop is often exponentially distributed; On the quality inspection line, the number of defective products in a batch generally follows binomial distribution. As such, weaker assumptions than Gaussian distribution should be more desirable to depict and monitor $\mathbf{X}_t$.

To fill the research gap, we develop a general Bayesian learning framework for high-dimensional POSCD, which is applicable to general exponential-family distributed data. In particular, a nonlinear form of composite decomposition is first developed for general high-dimensional exponential-family distributed data representation. It decomposes the data onto normal bases and abnormal bases, and the coefficients on those bases are estimated via Bayesian sparse learning. Built upon it, various test statistics can be incorporated for change detection. By further treating the test statistic as the reward function in CMAB, a Thompson sampling strategy can be implemented for adaptive sensor allocation with the balance of exploration and exploitation. Theoretical properties of the proposed framework are discussed accordingly. Three specific data distributions in exponential family—that is, Gaussian distribution, Poisson distribution, and binomial distribution—are discussed in numerical experiments.

The remainder of the article is organized as follows. In Section 2, we review the literature on some related topics to the POSCD problem. Section 3 develops our general framework for change detection with partial observations and variable selection strategy. Section 4 introduces its implementation procedures more specifically, including the Bayesian online variational inference for abnormal parameters, the construction of test statistics, the sensor selection strategy based on Thompson sampling, and some theoretical properties. Taking Gaussian, Poisson, and binomial distributions as examples, Section 5 and Section 6 present simulation studies on synthetic data and three real-world case studies, respectively, to further illustrate the efficacy and applicability of the proposed methodology. Some conclusive remarks are given in Section 7.

2. Literature Review

2.1. Partially Observable Sequential Change Detection

POSCD is an emerging topic that has not been fully addressed. Previous research has focused mainly on assuming the data's distribution to be the most common Gaussian distribution. The most pioneering work, Liu et al. (2015), proposed a local likelihood ratio test for each observed variable and also added a compensation parameter for each unobserved variable to represent their changed probabilities. It treats different dimensions as independent variables and selects variables independently according to their local statistics. However, not considering the correlation between variables leads to certain information loss because although a variable is not observed, its information can still be inferred based on its correlation with other observed variables. So, later, Wang et al. (2018) constructed kernel functions to represent the spatial correlation of variables according to their locations. Therefore, this correlation is constrained to spatial distribution represented by kernel function, which cannot be applied to more extensive instances. Considering the general high-dimensional variables with spatio-temporal correlation, Gómez et al. (2022) addressed the POSCD by regarding the data with unobserved variables in a time window as a streaming tensor and decomposing it into a low-rank component and a sparse component. This decomposition enables the values of the unobserved variables to be approximated by estimating the low-rank component. Meanwhile, the abnormal information contained in the sparse component can be learned. Then, a β-greedy adaptive sampling strategy is designed. That is, for β percentage of time points, the observable variables are selected randomly. For the other $1-\beta$ percentage, the observable variables with the largest abnormal information are selected. However, this cannot give a theoretical guarantee of optimal sampling because β is designated greedily. There are some works exploiting the relationship between POSCD and CMAB and trying to establish the theoretical properties for POSCD algorithms with the help of the CMAB framework. For example, Zhang and Hoi (2019) related POSCD to CMAB problems, proposed an adaptive sampling strategy based on the upper confidence bound (UCB) algorithm, and gave the regret bound between the optimal set and the selected set following their algorithm. However, because the construction of UCB needs to compute the inverse of covariance matrix of variables first, its computation complexity is too high to apply to high-dimensional data. Later, Zhang and Mei (2020) applied the Shiryaev-Robert-Pollak procedure to monitor the change of the process of independently and identically distributed variables and introduced Thompson sampling to sample those changed variables. However, it applies the Thompson sampling strategy not by sampling the parameter from its posterior distribution, but from a distribution that converges to the posterior in the limit condition when the change probability in the geometric prior converges to zero. It also imposes independence assumption on variables, which may lead to certain information loss, as we mentioned earlier. Furthermore, it assumes that the changed magnitude is known in advance to construct its test statistic, which is unrealistic. As we stated before, all these works assumed that the observations were from Gaussian distribution. However, in reality, there are many data streams following non-Gaussian distributions, such as Poisson distribution, binomial distribution, etc. For non-Gaussian distributions, Xian et al. (2018) extended the work of Liu et al. (2015) by constructing an antirank detection statistic. But it also assumes that variables are independent from each other, which is not efficient for change detection in correlated variables, as we stated before. To tackle this problem, Xian et al. (2021) revised the rank-based statistic of Xian et al. (2018) by automatically augmenting information for unobservable variables based on the observed ones and, accordingly, allocating the monitoring resources to the most suspicious variables. However, these two methods both need to tune the compensation parameters for those unobserved variables, and their sampling strategies cannot guarantee optimal sampling. Furthermore, except Gómez et al. (2022), all the current existing works do not address the challenge of detecting sparse changes and, hence, have limited detection power for high-dimensional cases.

2.2. Sparse Change Detection for High-Dimensional Data Streams

In the traditional fully observable scenario, for high-dimensional change detection, representation-learning-based methods, such as dimension reduction or low-rank approximation, have been used for better modeling high-dimensional data (Yan et al. 2014, Lu et al. 2016). To further address sparse changes, many works also incorporated the idea of sparse regularization or variable selection into the detection statistic (Wang and Jiang 2009, Zou and Qiu 2009, Mei 2010). However, they assumed that the correlations between variables were known in advance or could be estimated from historical data. Or, more simply, they imposed independence assumption on variables and just used identity matrix as the correlation matrix. As a result, these methods lose their efficiency for complexly structured data, such as profile data, image data, spatio-temporal data, etc. More recently, for modeling and monitoring multichannel profile data with sparse changes, Zhang et al. (2018) proposed a sparse functional principal component analysis (PCA) to extract the prominent features in these high-dimensional profiles, and, at the same time, the sparse PCA scores were used to construct a detection statistic for online monitoring. For image data with sparse changes, Yan

et al. (2017) decomposed image data on smooth bases and sparse bases, which were designed for normal regions and abnormal regions, respectively. Then, the coefficients on the sparse abnormal bases were learned to indicate the occurrence of anomaly, based on which the LASSO-based detection statistic was constructed. However, they are all for Gaussian-distributed high-dimensional data and cannot be applied to more general distributions. To release this assumption, for continuously distributed non-Gaussian high-dimensional variables, Zhang et al. (2020) first projected data onto several subspaces using random projection for dimension reduction and then constructed a local nonparametric control chart based on the spatial rank test for each subspace. Finally, these local charts were fused together for decision making. Mukherjee and Marozzi (2021) proposed nonparametric monitoring schemes based on specific distance metrics. For discretely distributed non-Gaussian data, Li et al. (2018b) proposed a change detection method for high-dimensional categorical data by relating the high-dimensional data to arrays of low-dimensional features and then classifying these features into multiple groups to discover various anomalous patterns. However, all these methods can only be applied to fully observable scenarios, and their extensions to partially observable cases are still of no trial.

2.3. Combinatorial Multi-armed Bandit

Another related research field is sequential variable selection for adaptive sampling. As we stated in Section 1, this is intimately related to the CMAB problem. So far, a number of studies have been done on sampling strategies for the CMAB problem, or the multi-armed bandit problem, which is a special case of CMAB, assuming that only one arm is chosen at each time point, instead of a set of m arms, and the rewards of different arms are independent. One commonly used sampling strategy is the upper confidence bound (Chen et al. 2013). Its basic idea is to compute the mean and variance for each arm and, consequently, get its confidence interval. Then, the arms whose confidence intervals have the largest upper bounds are selected. In this way, UCB can balance exploration and exploitation efficiently. Another strategy is Thompson sampling, proposed by Thompson (1933). It is a randomized Bayesian algorithm. Priors on the parameters of each arm are first imposed, and their posteriors are updated based on the reward of the played subset of arms. Then, a random sample is drawn from these posteriors and is treated as the true parameter for each arm. Based on it, the arms able to achieve the maximal expected joint reward are selected to play for the next time point. Its empirical performance and theoretical properties have been carefully exploited in Kaufmann et al. (2012) and Agrawal and Goyal (2012). Built upon these two classic strategies, some works also studied how to identify the best top-K arms (Bubeck et al. 2013, Jun et al. 2016) or the outlier arms (Zhuang et al. 2017, Ban and He 2020). However, these methods assume that the system is static—that is, the true reward distributions of arms do not change sequentially—and they target the identification of a single arm or top-K arms. Yet, for our POSCD problem, the reward distributions of arms can change at an unknown time point, and the goal is identifying the system change. We aim to detect this change at the system level, but do not intend to identify which arm is different from others at the arm level. Recently, there are some works that propose the idea of a nonstationary multi-armed bandit by assuming that the reward of each arm is not stationary and can change at unknown time points. Adaptive sampling policies as extensions of the UCB (Cao et al. 2019, Zhou et al. 2020) or Thompson sampling (Ghatak 2020, Cavenaghi et al. 2021) are proposed to identify the time-varying best arms to minimize the total expected regret. However, their goal is still best arm identification instead of system-level change detection. As such, their sampling policies cannot be applied to our case.

3. General POSCD Framework

This section gives a general description of the proposed framework for the POSCD problem. In particular, we first introduce our model formulation with composite decomposition for high-dimensional exponential-family distributed data in Section 3.1. Then, we construct the connection between variable selection and CMAB in Section 3.2. Built upon the connection, we introduce how to apply Thompson sampling for adaptive sensor allocation in Section 3.3.

In this paper, for notation simplification, for a vector $\boldsymbol{\alpha}$, we denote $\boldsymbol{\alpha}_Z$ as the subvector selected by Z, which is the index set of the selected dimensions. Similarly, we denote $\mathbf{A}_Z$ as the extracted submatrix of $\mathbf{A}$ composed by the extracted rows by Z. For example, if $\mathbf{A} \in \mathcal{R}^{p\times a}$, then $\mathbf{A}_Z \in \mathcal{R}^{m\times a}$. α_i denotes the ith component of a vector $\boldsymbol{\alpha}$. A_{ij} denotes the (ith, jth) component of a matrix $\mathbf{A}$. $\mathbf{A}_{i:}$ denotes the ith row of a matrix $\mathbf{A}$. $\mathbf{A}_{:j}$ denotes the jth column of a matrix $\mathbf{A}$. Let $\mathbf{1}_m$ denote a column vector with all m components equal to one. The notation $\oslash$ represents the element-wise division, and $\circ$ represents the element-wise product.

3.1. High-Dimensional Data Composite Decomposition

Consider a system consisting of p variables, whose observations at each time point t are denoted as $\mathbf{X}_t = (X_{1t}, \ldots, X_{pt})' \in \mathcal{R}^{p\times 1}$. We assume X_{it} has the same distribution type for $i = 1, \ldots, p$, which belongs to the exponential family—that is,

$$p(X_{it}\,|\,\eta_{it}) = \exp\left(\frac{X_{it}\eta_{it} - b(\eta_{it})}{a(\phi_i)} + c(X_{it}, \phi_i)\right), \quad i = 1, \ldots, p. \tag{1}$$

Here, η_{it} is the natural parameter (canonical parameter), and ϕ_i is the scale parameter, which is assumed not to

change over time. $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions for every specific distribution in the exponential family. There is a link function $g(\cdot)$ relating the natural parameter η_{it} to $\pi_{it} = \mathbb{E}[X_{it}]$, such that $g(\pi_{it}) = \eta_{it}$ or $\pi_{it} = g^{-1}(\eta_{it})$. Denote $\boldsymbol{\eta}_t = (\eta_{1t}, \ldots, \eta_{pt})' \in \mathcal{R}^{p\times 1}$, such that the joint distribution can be represented as $p(\mathbf{X}_t|\boldsymbol{\eta}_t) = \prod_{i=1}^{p} p(X_{it}|\eta_{it})$. We further introduce the correlation of $\mathbf{X}_t$ by the correlation of $\boldsymbol{\eta}_t$, following the general decomposition formulation for the exponential family (Chiquet et al. 2018, Li et al. 2018a). In particular, for high-dimensional representation, we assume the natural parameter $\boldsymbol{\eta}_t$ is linearly expanded on a before-change feature space with k_b bases $\mathbf{B}_b = (\mathbf{b}_{b1}, \ldots, \mathbf{b}_{bk_b})' \in \mathcal{R}^{p\times k_b}$—that is,

$$\boldsymbol{\eta}_t = \mathbf{B}_b\boldsymbol{\theta}_t, \quad \forall t \le \tau, \tag{2}$$

where $\boldsymbol{\theta}_t \in \mathcal{R}^{k_b\times 1}$ are the coefficients. In this paper, we further assume $\boldsymbol{\theta}_t \overset{iid}{\sim} N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$, $\forall t$. Here, $\mathbf{B}_b$ can either be learned by historical reference samples via matrix decomposition algorithms or be set as notable spaces, such as spline space (Meier et al. 2009), Fourier space, kernel space (Qiu et al. 2010), etc. $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ can also be estimated from historical before-change samples, and the estimation procedure is added to Online Appendix E.

When $\mathbf{X}_t$ occurs sparse changes after unknown time τ, $\boldsymbol{\eta}_t$ would change. Consider that many sparse change patterns may occur in the system, and the chance that each change pattern happens is pretty small. We may further define an anomaly dictionary with a set of abnormal bases $\mathbf{B}_a = (\mathbf{b}_{a1}, \ldots, \mathbf{b}_{ak_a})' \in \mathcal{R}^{p\times k_a}$. $\mathbf{B}_a$ can either be set by domain knowledge from practitioners if certain specific change patterns are of interest or be learned from collected historical anomaly data via the dictionary learning approach. Then, a composite decomposition approach can be formulated to describe the postchange distribution with sparse change patterns (Zhang et al. 2016, Yan et al. 2017)—that is,

$$\boldsymbol{\eta}_t = \mathbf{B}_b\boldsymbol{\theta}_t + \mathbf{B}_a\boldsymbol{\theta}_a, \quad \forall t > \tau, \tag{3}$$

where $\boldsymbol{\theta}_a \in \mathcal{R}^{k_a\times 1}$ are the postchange coefficients. Combining (2) and (3), we can define the change point model as

$$\begin{cases} H_0 : \boldsymbol{\theta}_a = \mathbf{0}, & t = 1, \ldots, \tau, \\ H_1 : \boldsymbol{\theta}_a = \boldsymbol{\xi}, & t = \tau+1, \ldots, \end{cases} \tag{4}$$

where $\boldsymbol{\xi}$ is the unknown change magnitude with nonzero values. In a real online change detection scenario, because τ is unknown, our goal is to construct a sequential change detection scheme for (4) based on the partially observed subset of $\mathbf{X}_t, t = 1, 2, \ldots$. Here, we introduce a sensing indicator variable $\mathbf{z}_t = (z_{1t}, \ldots, z_{pt})'$, such that $z_{it} = 1$ if and only if X_{it} is observed at time point t, and the sensing constraint can be expressed as $\sum_{i=1}^{p} z_{it} = m$, $\forall t$. Denote $Z(t)$ to be the vector of indices corresponding to the observed dimensions for $\mathbf{X}_t$. $\mathbf{X}_{Z(t)} \in \mathcal{R}^{m\times 1}$ represents the observed data for time point t. We would like to construct a detection scheme for (4).

In this paper, we restrict our attention to the scenario with only one change point. Once the change occurs after τ, it would keep. This restriction is very common in literature, especially in statistical quality control (Montgomery 2007). The detection scheme is decided by a stopping time T and a test statistic Λ_t. The stopping time is defined as the time point that Λ_t first exceeds a predefined detection threshold h—that is, $T = \inf_t\{\Lambda_t > h\}$. Thus, $T = n$ means that there exists a change among the first n time points. Two fundamental metrics are used to evaluate the performance of such a detection scheme for online change point detection: Average Run Length (ARL) before a false alarm occurs under normal condition—that is, $ARL = \mathbb{E}[T|\tau = \infty]$—and Average Detection Delay (ADD) after a change occurs under abnormal condition—that is, $ADD = \mathbb{E}[T - \tau|T > \tau, \tau < \infty]$. In practice, conditional on a fixed ARL, which controls the false alarm rate, a desired change detection scheme manages to achieve as small of an ADD as possible.

3.2. CMAB Formulation for Change Detection

We formulate our change detection procedure with adaptive sampling under CMAB framework. In our problem, there are p variables (i.e., arms) in the system. According to our formulation, the ith arm has mean $g^{-1}(\mathbf{B}_{a,i:}\boldsymbol{\theta}_a + \mathbf{B}_{b,i:}\boldsymbol{\theta}_t)$. Each time the bandit pulls a set of m arms and gets a reward, which is the joint function of the observations of the set of arms. The goal is looking for a sampling policy that can select arms to maximize the expectation of the total reward. Because our goal is to select the most possibly changed dimensions, a good reward function should include the information of change—that is, $\boldsymbol{\theta}_a$. Denote $R_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(t)})$ as the reward function of selecting an arm subset $Z(t)$ under the parameter $\boldsymbol{\theta}_a$. The reward function $R_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(t)})$ is desired to satisfy the following two requirements. First, under a normal condition ($\boldsymbol{\theta}_a = \mathbf{0}$), for any subset $Z(t)$, this reward should be small. Second, under an abnormal condition ($\boldsymbol{\theta}_a \neq \mathbf{0}$), we desire that when the selected set $Z(t)$ approximates the true anomalous set more, the reward should be larger. This is consistent with the formulation of the hypothesis test statistic. That is, for any subset $Z(t)$, if there is no change, the test statistic would be small; When there is a change, the test statistic would be large and become larger as the included changed dimensions or change magnitudes increase. Evidently, we can select proper hypothesis test statistics as the reward functions in CMAB, such as the commonly used test statistics, including the likelihood ratio test, Bayes factor, etc.

3.3. Adaptive Sampling via Thompson Sampling

Built upon the reward function, at each time point, we aim to select a subset $Z(t)$ to maximize the expected reward—that is, the chance of detecting the change. Because $\boldsymbol{\theta}_a$ is unknown, one way is to select arms that can maximize the reward function $R_{\hat{\boldsymbol{\theta}}_a}(\mathbf{X}_{Z(t)})$, based on

the estimated $\hat{\boldsymbol{\theta}}_a$ so far, which is exploitation. Meanwhile, the estimation uncertainty in $\hat{\boldsymbol{\theta}}_a$ should also be considered, which is exploration.

Thompson sampling is the most commonly used method to balance exploitation and exploration for CMAB problems, which has superior empirical performance and abundant theoretical properties in the literature (Agrawal and Goyal 2012, Kaufmann et al. 2012, Durand and Gagné 2014). The idea of Thompson sampling is selecting a set $\hat{Z}$ with probability

$$\int \mathbb{I}\left(\hat{Z} = \arg\max_{Z} \mathbb{E}[R_{\boldsymbol{\theta}_a}(\mathbf{X}_Z)]\right)\tilde{p}(\boldsymbol{\theta}_a)d\boldsymbol{\theta}_a, \tag{5}$$

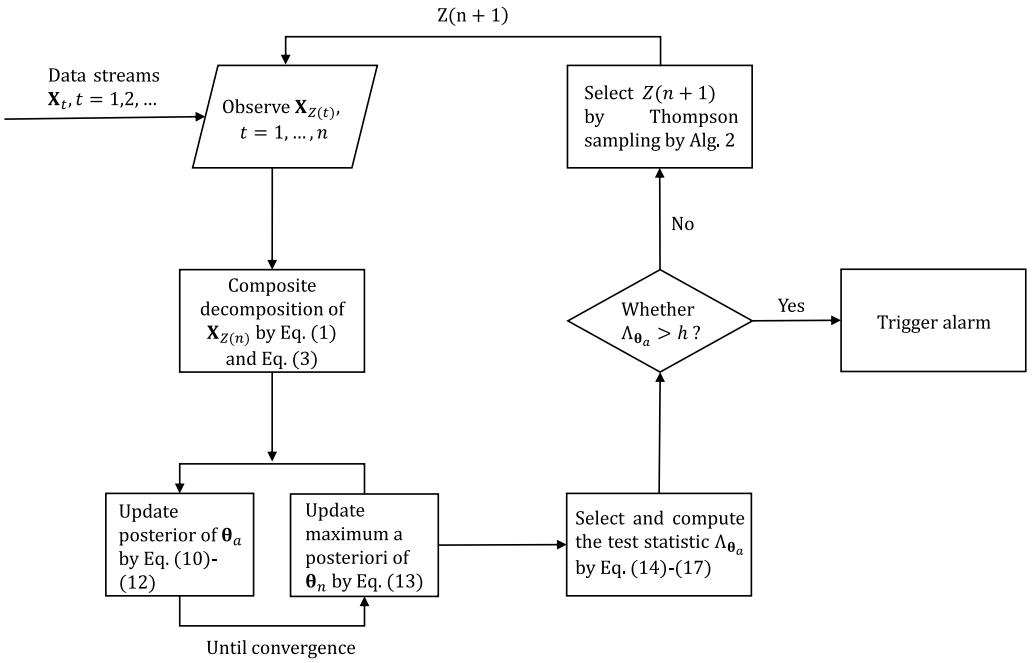
where $\tilde{p}(\boldsymbol{\theta}_a)$ is the posterior distribution of $\boldsymbol{\theta}_a$. Intuitively, the indicator function $\mathbb{I}(\cdot)$ in (5) focuses on exploiting a best $\hat{Z}$, and the integral over all $\tilde{p}(\boldsymbol{\theta}_a)$ explores all possible $\boldsymbol{\theta}_a$.

In summary, the change detection procedure with adaptive sampling under CMAB framework can be generally formulated as follows:

1. For the current time n, construct the composite decomposition of $\mathbf{X}_{Z(n)}$. Based on the historical observations—that is, $\mathbf{X}_{Z(t)}, t = 1, \ldots, n$—estimate the posteriors of $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_a$ iteratively until convergence;
2. Select and compute the test statistic $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)})$ based on the estimation of parameters; If $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) > h$, where h is the change threshold, a change alarm is triggered;
3. Otherwise, set the test statistic as the reward function—that is, $R_{\boldsymbol{\theta}_a}(\mathbf{X}_Z) \equiv \Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z)$—and decide $Z(n+1)$ using the Thompson sampling strategy.

We also draw a flowchart in Figure 1 to show this process more clearly.

Figure 1. Flowchart for the Proposed Algorithm at Time n



4. Specific Implementation Procedure

In this section, we will discuss how to implement the above three steps in detail. First, we talk about the online inference of $\boldsymbol{\theta}_a$ in the Bayesian framework in Section 4.1 and Section 4.2. Then, we introduce some popular test statistics for change detection in Section 4.3. Treating the test statistics as the reward function, we introduce the detailed Thompson sampling strategy in Section 4.4. Some theoretical properties are last discussed in Section 4.5.

4.1. Bayesian Estimation for $\boldsymbol{\theta}_a$ with Spike-Slab Prior

Because Thompson sampling relies on the inference of posterior distribution of $\boldsymbol{\theta}_a$—that is, $\tilde{p}(\boldsymbol{\theta}_a)$—we propose an online Bayesian estimation framework for $\tilde{p}(\boldsymbol{\theta}_a)$ based on sequential samples. First, we need to decide its prior distribution. Considering the changed dimension is sparse, $\boldsymbol{\theta}_a$ would be a sparse vector. Hence, we propose to impose a Bayesian sparse prior on $\boldsymbol{\theta}_a$ by assuming that each component of $\boldsymbol{\theta}_a$ follows a spike-slab distribution independently. This prior has been commonly used as a Bayesian prior for sparse vector estimation and variable selection (Ishwaran and Rao 2005, Ročková 2018). In particular, the prior distribution of $\boldsymbol{\theta}_a$ has the form of

$$\begin{aligned} &p_0(\theta_{a,j}|r_j) \sim N(0, r_j\sigma_j^2 + (1-r_j)v\sigma_j^2), \\ &p_0(r_j) \sim Bernoulli(w_j), \quad j = 1, \ldots, k_a. \end{aligned} \tag{6}$$

Here, binary variables $\mathbf{r} = (r_1, \ldots, r_{k_a})'$ are introduced to indicate whether the corresponding components in $\boldsymbol{\theta}_a$ are nonzero. Each $r_j, j = 1, \ldots, k_a$, is a Bernoulli random trial governed by common success rate $p(r_j = 1) = w_j$. If

$r_j = 0$, the prior $p_0(\theta_{a,j})$ follows the Gaussian distribution with zero mean and variance $v\sigma_j^2$, with $v \ll 1$—that is, the "spike"—which demonstrates that the probability $p_0(\theta_{a,j} = 0)$ almost equals one. Otherwise, $p_0(\theta_{a,j})$ follows the Gaussian distribution with zero mean and variance σ_j^2—that is, the "slab"—which demonstrates that the probability $p_0(\theta_{a,j} \neq 0)$ is large.

For $\tilde{p}(\boldsymbol{\theta}_a)$, to better estimate the changes with small magnitude in $\boldsymbol{\theta}_a$, supposing the current time point is n, we borrow the idea of the exponentially weighted moving average in Montgomery (2007) and calculate $\tilde{p}(\boldsymbol{\theta}_a)$ with the previous n samples $\mathbf{X}_{Z(t)}$, $t = 1, \ldots, n$. Consider that samples in recent time points are more likely to represent the current system state and can help better detect the changes than samples in the past time points. Thus, we would like to impose more weights on the later time points in the estimation. As such, we enforce time-decayed weights $\boldsymbol{\lambda}_n = (\lambda_{n1}, \ldots, \lambda_{nn})'$ on the n samples, in the sense that $\lambda_{n1} < \cdots < \lambda_{nn}$, and get the weighted posterior distribution of $\boldsymbol{\theta}_a$ as

$$p(\boldsymbol{\theta}_a, \mathbf{r} | \mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)}) \propto p_0(\boldsymbol{\theta}_a, \mathbf{r}) \prod_{t=1}^{n} p(\mathbf{X}_{Z(t)} | \boldsymbol{\theta}_a, \mathbf{r})^{\lambda_{nt}}. \quad (7)$$

In this paper, we use the exponentially decayed weights—that is, $\lambda_{nt} = (1-\lambda)^{n-t}$ with a small positive value $\lambda \in [0.05, 0.25]$.

4.2. Online Variational Inference

However, consider that (7) does not have a closed-form solution due to the complex hierarchical structure of the spike-slab prior distribution of $\boldsymbol{\theta}_a$. Hence, we propose an efficient variational Bayesian (VB) inference for $\boldsymbol{\theta}_a$. The core idea of the variational Bayesian approach is to approximate (7) via another simpler distribution family, defined as $\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r}) = \prod_{j=1}^{k_a} \tilde{p}_j(\theta_{a,j}, r_j)$, by minimizing their Kullback-Leibler divergence—that is, $KL[\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r}) || p(\boldsymbol{\theta}_a, \mathbf{r} | \mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)})]$. Here, $\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})$ is fully factorized as k_a independent distributions, which originates from the famous mean field assumption in VB to make the derivation of the Kullback-Leibler divergence tractable. Although $\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})$ cannot capture the posterior dependence across $\boldsymbol{\theta}_{a,j}$, $j = 1, \ldots, k_a$, it can approximate each of their marginals well. VB has a theoretical guarantee that $\tilde{p}(\boldsymbol{\theta}_{a,j})$ can converge to the Gaussian distribution, which is centered at the true value of $\boldsymbol{\theta}_{a,j}$ in total variation distance, $j = 1, \ldots, k_a$, which is the frequentist consistency of VB methods (Wang and Blei 2019). In this paper, to capture the sparsity of $\boldsymbol{\theta}_a$ well, we restrict $\tilde{p}_j(\theta_{a,j}, r_j)$ to be a spike-slab distribution as well—that is,

$$\begin{aligned} \tilde{p}_j(\theta_{a,j} | r_j) &\sim N(r_j \mu_j, r_j s_j^2 + (1 - r_j) v s_j^2), \\ \tilde{p}_j(r_j) &\sim Bernoulli(\alpha_j). \end{aligned} \quad (8)$$

Solving the best approximated posterior distribution $\tilde{p}_j(\theta_{a,j}, r_j) = \tilde{p}_j(\theta_{a,j} | r_j) \tilde{p}_j(r_j)$ indicates to find $\{\mu_j, s_j, \alpha_j\}$ that minimize the Kullback-Leibler divergence. This minimization cannot be solved directly because the Kullback-Leibler divergence is hard to derive directly. To simplify it, the evidence lower bound (ELBO) of $\ln p(\mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)})$—that is, $J(\tilde{p})$—is derived from the Kullback-Leibler divergence:

$$\begin{aligned} &J(\tilde{p}) \\ &= \ln p(\mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)}) - KL[\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r}) || p(\boldsymbol{\theta}_a, \mathbf{r} | \mathbf{X}_{Z(1)}, \ldots, \mathbf{X}_{Z(n)})] \\ &= \mathbb{E}_{\tilde{p}} \left[\sum_{t=1}^{n} \lambda_{nt} \ln p(\mathbf{X}_{Z(t)} | \boldsymbol{\theta}_a) + \ln p(\boldsymbol{\theta}_a, \mathbf{r}) - \ln \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r}) \right] \\ &= \sum_{t=1}^{n} \lambda_{nt} ((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' (\mathbf{B}_{a,Z(t)}(\boldsymbol{\alpha} \circ \boldsymbol{\mu}) + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t)) \\ &\quad + \sum_{j=1}^{k_a} \left(\frac{1}{2} - \frac{s_j^2}{2\sigma_j^2} + \frac{1}{2} \ln \frac{s_j^2}{\sigma_j^2} + \left(\ln w_j - \ln \alpha_j - \frac{\mu_j^2}{2\sigma_j^2} \right) \alpha_j \right. \\ &\quad \left. + (\ln(1 - w_j) - \ln(1 - \alpha_j))(1 - \alpha_j) \right) \\ &\quad - \sum_{t=1}^{n} \lambda_{nt} (\mathbf{1}_m' \mathbb{E}_{\tilde{p}} [b(\mathbf{B}_{a,Z(t)} \boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)} \boldsymbol{\theta}_t) \oslash a(\boldsymbol{\phi}_{Z(t)})] \\ &\quad + \mathbf{1}_m' c(\mathbf{X}_{Z(t)}, \boldsymbol{\phi}_{Z(t)})). \end{aligned} \quad (9)$$

Denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k_a})'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{k_a})'$, $\mathbf{s} = (s_1, \ldots, s_{k_a})'$. $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ impose on every dimension. $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. Then, minimizing the Kullback-Leibler divergence can be converted to maximizing $J(\tilde{p})$. By taking the partial derivatives of $J(\tilde{p})$, it yields coordinate gradient updates:

$$\frac{\partial J(\tilde{p})}{\partial \mu_j} = \sum_{t=1}^{n} \lambda_{nt} ((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' \mathbf{B}_{a,Z(t)j} \alpha_j) - \frac{\alpha_j \mu_j}{\sigma_j^2} - \frac{\partial d}{\partial \mu_j}, \quad (10)$$

$$\frac{\partial J(\tilde{p})}{\partial \alpha_j} = \sum_{t=1}^{n} \lambda_{nt} ((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))' \mathbf{B}_{a,Z(t)j} \mu_j) + \ln w_j - \frac{\mu_j^2}{2\sigma_j^2} - \ln \alpha_j - \ln(1 - w_j) + \ln(1 - \alpha_j) - \frac{\partial d}{\partial \alpha_j}, \quad (11)$$

$$\frac{\partial J(\tilde{p})}{\partial s_j} = -\frac{s_j}{\sigma_j^2} + \frac{1}{s_j} - \frac{\partial d}{\partial s_j}. \quad (12)$$

In the above equations, for notation simplification, we define $d = \sum_{t=1}^{n} \lambda_{nt} \mathbf{1}_m' \mathbb{E}_{\tilde{p}} [b(\mathbf{B}_{a,Z(t)} \boldsymbol{\theta}_a + \mathbf{B}_{b,Z(t)} \tilde{\boldsymbol{\theta}}_t) \oslash a(\boldsymbol{\phi}_{Z(t)})]$. Then, update $\{\boldsymbol{\mu}, \mathbf{s}, \boldsymbol{\alpha}\}$ using the gradient ascent method.

Based on $\tilde{p}(\boldsymbol{\theta}_a | \mathbf{r})$ and $\tilde{p}(\mathbf{r})$, we can further update the posterior estimate of $\boldsymbol{\theta}_n$ using Bayes' rule. Because we assume $\boldsymbol{\theta}_n$ is identically and independently distributed for different n, the likelihood function $p(\mathbf{X}_{Z(n)} | \boldsymbol{\theta}_n)$ is only related to the current observation $\mathbf{X}_{Z(n)}$. Here, for simple implementation, we adopt the maximum a posteriori (MAP) estimation method. Although we assume $p_0(\boldsymbol{\theta}_n) \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$ as a universal prior for the exponential family (Petterson et al. 2009), conjugate priors for specific distributions are also compatible. Then, the posterior of $\boldsymbol{\theta}_n$

can be obtained as $\tilde{p}(\boldsymbol{\theta}_n) \propto p(\mathbf{X}_{Z(n)}|\boldsymbol{\theta}_n, \boldsymbol{\theta}_a)p_0(\boldsymbol{\theta}_n)$. Then, the MAP estimation—that is, $\tilde{\boldsymbol{\theta}}_n$—can be achieved via that gradient ascent method as well with

$$\frac{\partial \ln \tilde{p}(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n} = -\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) + \mathbf{B}'_{b,Z(n)}((\mathbf{X}_{Z(n)} - b'(\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a + \mathbf{B}_{b,Z(n)}\boldsymbol{\theta}_n)) \oslash a(\boldsymbol{\phi}_{Z(t)})), \tag{13}$$

where $b'(\cdot)$ is the first derivative of $b(\cdot)$. Taking Gaussian, Poisson, and binomial distributions as examples, the specific forms of (10)–(13) are derived in Online Appendix A. The architecture of the full estimation procedure is shown in Algorithm 1. It is worth mentioning that other prior distributions, such as the spike-slab Laplace prior (Ročková 2018), and other corresponding Bayesian estimation procedures, such as Markov chain Monte Carlo (MCMC) (George and McCulloch 1997), would also be used for the inference of $\boldsymbol{\theta}_a$ and be compatible with the whole detection framework, based on the practical demand.

Algorithm 1 (Variational Bayesian Inference for $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_n$)

Input: Data $\mathbf{X}_{Z(t)}, t = 1, \ldots, n$, $\tilde{\boldsymbol{\theta}}_t, t = 1, \ldots, n-1$
Initialize $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\mathbf{s}$, $\tilde{\boldsymbol{\theta}}_n$ according to prior knowledge.
repeat
  for $j = 1, \ldots, k_a$ **do**
    Update gradients $\frac{\partial J(\tilde{p})}{\partial \mu_j}$, $\frac{\partial J(\tilde{p})}{\partial \alpha_j}$, $\frac{\partial J(\tilde{p})}{\partial s_j}$ via (10), (11), and (12).
    Update μ_j, s_j and α_j using gradient ascent method.
  Update gradient $\frac{\partial \ln \tilde{p}(\boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n}$ via (13). Update $\tilde{\boldsymbol{\theta}}_n$ using gradient ascent method.
until *Converge*;
return $\boldsymbol{\mu}, \mathbf{s}, \boldsymbol{\alpha}, \tilde{\boldsymbol{\theta}}_n$

4.3. Choices of Test Statistics

As mentioned in Section 3.2, as long as a test statistic can guarantee that for any subset $Z(t)$, if there is no change, the test statistic would be small, whereas if there is a change, the test statistic would be large and become larger as the included changed dimensions or change magnitudes increase, this test statistic can be compatible in our framework. In this paper, we consider the four most commonly used test statistics as follows: simple likelihood ratio test, general likelihood ratio test, Bayesian factor, and posterior Bayesian factor. However, other test statistics, such as the Lagrange multiplier, the Wald test, the sequential probability ratio test, and so on, can also be good candidates.

The simple likelihood ratio test (SLRT) can be derived by formulating the hypothesis test as $H_0 : \boldsymbol{\theta}_a = \mathbf{0}$, $H_1 : \boldsymbol{\theta}_a = \boldsymbol{\mu}_a$, where $\boldsymbol{\mu}_a = \boldsymbol{\mu} \circ \boldsymbol{\alpha}$ is the MAP estimation:

$$\begin{aligned}\Lambda^{SLRT}_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) &= 2\ln\frac{p(\mathbf{X}_{Z(n)}|H_1)}{p(\mathbf{X}_{Z(n)}|H_0)} \\ &= (\mathbf{X}_{Z(n)} \oslash a(\boldsymbol{\phi}_{Z(n)}))'\mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a - \mathbf{1}'_m \\ &\quad (b(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n + \mathbf{B}_{a,Z(n)}\boldsymbol{\mu}_a) \oslash a(\boldsymbol{\phi}_{Z(n)})) \\ &\quad + \mathbf{1}'_m(b(\mathbf{B}_{b,Z(n)}\tilde{\boldsymbol{\theta}}_n) \oslash a(\boldsymbol{\phi}_{Z(n)})).\end{aligned} \tag{14}$$

The generalized likelihood ratio test (GLRT) with exponentially weighted moving average can be derived by formulating the hypothesis test as $H_0 : \boldsymbol{\theta}_a = \mathbf{0}, H_1 : \boldsymbol{\theta}_a \neq \mathbf{0}$:

$$\begin{aligned}\Lambda^{GLRT}_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) &= 2\max_{\boldsymbol{\theta}_a}\sum_{t=1}^{n}\lambda_{nt}\ln\frac{p(\mathbf{X}_{Z(t)}|H_1)}{p(\mathbf{X}_{Z(t)}|H_0)} \\ &= \sum_{t=1}^{n}\lambda_{nt}((\mathbf{X}_{Z(t)} \oslash a(\boldsymbol{\phi}_{Z(t)}))'\mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a \\ &\quad - \mathbf{1}'_m(b(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t + \mathbf{B}_{a,Z(t)}\hat{\boldsymbol{\theta}}_a) \oslash a(\boldsymbol{\phi}_{Z(t)})) \\ &\quad + \mathbf{1}'_m(b(\mathbf{B}_{b,Z(t)}\tilde{\boldsymbol{\theta}}_t) \oslash a(\boldsymbol{\phi}_{Z(t)}))),\end{aligned} \tag{15}$$

where $\hat{\boldsymbol{\theta}}_a$ maximizes the weighted log-likelihood ratio—that is, $\sum_{t=1}^{n}\lambda_{nt}\ln\frac{p(\mathbf{X}_{Z(t)}|H_1)}{p(\mathbf{X}_{Z(t)}|H_0)}$—which can be estimated using Newton's method in Hardin and Hilbe (2012).

The Bayes factor (BF) is a test statistic in Bayesian framework. It can derived by formulating the hypothesis test as $H_0 : \boldsymbol{\theta}_a = \mathbf{0}, H_1 : \boldsymbol{\theta}_a \sim p_0(\boldsymbol{\theta}_a)$, where p_0 is the prior of $\boldsymbol{\theta}_a$ (in our case, it is the spike-slab distribution in (6)):

$$\begin{aligned}\Lambda^{BF}_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) &= \frac{p(\mathbf{X}_{Z(n)}|H_1)}{p(\mathbf{X}_{Z(n)}|H_0)} \\ &= \frac{\sum_{\mathbf{r}\in\mathbb{R}} p_0(\mathbf{r})\int p(\mathbf{X}_{Z(n)}|\boldsymbol{\theta}_a, \tilde{\boldsymbol{\theta}}_n)p_0(\boldsymbol{\theta}_a|\mathbf{r})d\boldsymbol{\theta}_a}{p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n)},\end{aligned} \tag{16}$$

where $\mathbb{R}$ is the value set of $\mathbf{r}$.

The posterior Bayes factor (PBF) is similar to the BF, but considers using the estimated posterior of $\boldsymbol{\theta}_a$ as $H_1 : \boldsymbol{\theta}_a \sim \tilde{p}(\boldsymbol{\theta}_a)$, where $\tilde{p}(\boldsymbol{\theta}_a)$ is the posterior estimated via Algorithm 1:

$$\begin{aligned}\Lambda^{PBF}_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) &= \frac{p(\mathbf{X}_{Z(n)}|H_1)}{p(\mathbf{X}_{Z(n)}|H_0)} \\ &= \frac{\sum_{\mathbf{r}\in\mathbb{R}} \tilde{p}(\mathbf{r})\int p(\mathbf{X}_{Z(n)}|\boldsymbol{\theta}_a, \tilde{\boldsymbol{\theta}}_n)\tilde{p}(\boldsymbol{\theta}_a|\mathbf{r})d\boldsymbol{\theta}_a}{p(\mathbf{X}_{Z(n)}|\tilde{\boldsymbol{\theta}}_n)}.\end{aligned} \tag{17}$$

Frequentist tests, such as SLRT and GLRT, are easy to compute, applicable to a wide range of models and data types, and provide a rigorous significance test for

comparing models. But they have limitations for not considering the uncertainty of parameter estimations and, thus, cannot provide reliable results with small sample size. To make those up, Bayesian tests, such as BF and PBF, combine prior or posterior information, which may come from the prior knowledge of experts or historical data, to consider the uncertainty of parameters and, thus, have more robust performance, especially when $\boldsymbol{\theta}_a$ may even change over time. However, computing the marginal likelihoods in BF or PBF is always complex, and their performance is quite sensitive to the prior or posterior estimation. Improper designation of the prior or inaccurate estimation of the posterior may ruin the efficiency of Bayesian tests. Taking Gaussian, Poisson, and binomial distributions as examples, the specific forms of the four test statistics are shown in Online Appendix B. For each of these test statistics, we can set a detection threshold h according to a predefined confidence level (false alarm rate) and define that if $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_{Z(n)}) > h$, the test statistic triggers a change alarm. Otherwise, decide next $Z(n+1)$ and wait for $\mathbf{X}_{n+1}$.

4.4. Thompson Sampling for Variable Selection

Now, we discuss how to select the sampling index set $Z(n+1)$ for the next time point. As discussed in Section 3.2, the test statistic can be regarded as the reward function in CMAB. Then, based on the estimated $\tilde{p}(\boldsymbol{\theta}_a)$, Thompson sampling in (5) can be applied to choose $Z(n+1)$. In particular, instead of computing (5) analytically, Thompson sampling actually implements a practical sampling procedure. First, it samples a $\boldsymbol{\theta}_a^\star$ from $\tilde{p}(\boldsymbol{\theta}_a)$ for exploration and then selects $Z(n+1)$ by maximizing the expected reward for one time step for exploitation—that is,

$$Z(n+1) = \arg\max_{Z} \mathbb{E}[\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z) \mid \boldsymbol{\theta}_a^\star]. \tag{18}$$

Balancing exploration and exploitation well, Thompson sampling achieves a good property that the average per period regret between the optimal Z and $Z(n+1)$ converges to zero as n goes on (Russo et al. 2018). Note that for some exponential-family distributions, if complex test statistics are used, their $\mathbb{E}[\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z) \mid \boldsymbol{\theta}_a^\star]$ can be very complex and have nonlinear items of $\mathbf{X}$, which makes it very hard to derive the expectation of $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z)$. In this case, we can further approximate the expectation using a stochastic way by first drawing a $\boldsymbol{\theta}_a^\star \sim \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})$. Then, draw $\mathbf{X}^\star$ from its probability distribution $\mathbf{X}^\star \sim p(\mathbf{X} \mid \boldsymbol{\eta} = \mathbf{B}_a\boldsymbol{\theta}_a^\star + \mathbf{B}_b\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the prior mean of $\boldsymbol{\theta}_t$. Then, use $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z^\star)$ as an approximation of $\mathbb{E}[\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z) \mid \boldsymbol{\theta}_a^\star]$.

Remark 1. Note that for some exponential-family distributions with complex test statistics, such as PBF for Poisson and binomial distributions, $\Lambda_{\boldsymbol{\theta}_a}(\mathbf{X}_Z^\star)$ would be a nonlinear function of Z. Then, all the $\binom{p}{m}$ subsets need to be evaluated for selecting the best Z in (18). This is a common problem for many CMAB strategies (Chen et al. 2016), where they usually assume that an oracle computer center can evaluate all the combinations and do not need to worry about the computation cost. However, in practice, enumerating all possible combinations of the arms may still be undesirable, especially when the number of arms is large. Fortunately, $\Lambda_{\boldsymbol{\theta}_a}^{SLRT}(\mathbf{X}_Z^\star)$ or $\Lambda_{\boldsymbol{\theta}_a}^{GLRT}(\mathbf{X}_Z^\star)$ can always be decomposed as the linear sum of $X_i^\star$, $i = 1, \ldots, p$, for exponential-family distributions and take SLRT as an example, which is similar for GLRT:

$$\begin{aligned}
&\Lambda_{\boldsymbol{\theta}_a}^{SLRT}(\mathbf{X}_Z^\star) \\
&= (\mathbf{X}_Z^\star \oslash a(\boldsymbol{\phi}_Z))' \mathbf{B}_{a,Z}\boldsymbol{\mu}_a - \mathbf{1}_m'(b(\mathbf{B}_{b,Z}\boldsymbol{\theta}_0 + \mathbf{B}_{a,Z}\boldsymbol{\mu}_a) \oslash a(\boldsymbol{\phi}_Z)) \\
&\quad + \mathbf{1}_m'(b(\mathbf{B}_{b,Z}\boldsymbol{\theta}_0) \oslash a(\boldsymbol{\phi}_Z)) \\
&= \sum_{i \in Z} \frac{1}{a(\phi_i)} (X_i^\star \mathbf{B}_{a,i:}\boldsymbol{\mu}_a - b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0 + \mathbf{B}_{a,i:}\boldsymbol{\mu}_a) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0)).
\end{aligned}$$

Consequently, we can select each component of Z in a sequential way by ranking

$$\Lambda_i^{SLRT} = \frac{1}{a(\phi_i)} (X_i^\star \mathbf{B}_{a,i:}\boldsymbol{\mu}_a - b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0 + \mathbf{B}_{a,i:}\boldsymbol{\mu}_a) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0)), \quad i = 1, \ldots, p,$$

from the largest to the smallest and select the top m items and avoid enumerating all possible sets of Z. Consequently, the complexity of the sampling process reduces from $O(p^m)$ to $O(p(k_a + k_b) + p\log p))$ and makes it desirable to apply to extremely high-dimensional data. As such, we highlight the sampling procedure based on SLRT in Algorithm 2.

Algorithm 2 (Thompson Sampling Procedure Based on SLRT)

Input: $\tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})$, $a(\boldsymbol{\phi})$, $\mathbf{B}_a$, $\mathbf{B}_b$, $\boldsymbol{\theta}_0$.
Output: $Z(n+1)$
Sample $\boldsymbol{\theta}_a^\star \sim \tilde{p}(\boldsymbol{\theta}_a, \mathbf{r})$. Draw $\mathbf{X}^\star$ from its probability distribution—that is,
$\mathbf{X}^\star \sim p(\mathbf{X} \mid \boldsymbol{\eta} = \mathbf{B}_a\boldsymbol{\theta}_a^\star + \mathbf{B}_b\boldsymbol{\theta}_0)$.
Compute $\Lambda_i^{SLRT} = \frac{1}{a(\phi_i)}(X_i^\star \mathbf{B}_{a,i:}\boldsymbol{\mu}_a - b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0 + \mathbf{B}_{a,i:}\boldsymbol{\mu}_a) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0))$, $i = 1, \ldots, p$.
Rank the Λ_i^{SLRT}, for $i = 1, \ldots, p$ from the largest to the smallest, and select the top m indexes as $Z(n+1)$.

Hereafter, we name our proposed Exponential Family-Bayesian Spike-Slab Composite Decomposition change detection scheme based on SLRT as EF-BSSCD(S), based on GLRT as EF-BSSCD(G), based on PBF as EF-BSSCD(P), and based on BF as ORACLE (because in the following numerical studies of this paper, we use the true distribution of $\boldsymbol{\theta}_a$ as the prior distribution in BF, which actually is not feasible in reality. So, it is predictable that BF will perform the best—that is, oracle—detection performance).

4.5. Theoretical Properties

Considering the computational merit of SLRT as discussed above, in this paper, we discuss the theoretical properties of EF-BSSCD(S) under asymptotic conditions for general exponential-family distributions. However, note that other test statistics, such as GLRT and PBF, also have similar theoretical results following the similar proof procedure, though their detailed derivations are not included in this paper due to space limitation.

Theorem 1. *When there is no change in the system, as* $n \to \infty$, $\lambda \to 0$, *and* $m \to \infty$ (*forcing* $p \to \infty$), *we have* $\tilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_n$ *and* $\tilde{p}_j(\theta_{a,j}) \xrightarrow{d} \delta_0$, *for* $j = 1, \ldots, k_a$, *where* δ_0 *is a point mass at zero. That means* $\mu_j \to 0$ *and* $s_j \to 0$. *Consequently, we have* $\Lambda_i^{SLRT} \to 0$ *in the same rate of* $O(1/\sqrt{n})$, *for* $i = 1, \ldots, p$.

As stated in Algorithm 2, we select $Z(n+1)$ by ranking Λ_i^{SLRT}, for $i = 1, \ldots, p$, and selecting the top m items. Theorem 1 indicates that when the system has no change, under limit conditions, $Z(n+1)$ are selected from all variables randomly.

Theorem 2. *When the system has a change, assume the change relates to a certain base set* $\mathcal{A} \subset \{1, \ldots, k_a\}$. *And the change magnitude vector can be represented as* $\boldsymbol{\xi} = (0, \ldots, \xi_l, \ldots, 0)'$, *where* ξ_l *is the change magnitude happening on the* l*th basis, and* $l \in \mathcal{A}$. *As* $n \to \infty$ *and* $\lambda \to 0$ *and* $m \to \infty$ (*forcing* $p \to \infty$), *we have* $\tilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_n$ *and* $\tilde{p}(\theta_{a,l}) \xrightarrow{d} \delta_{\xi_l}$, $\forall l \in \mathcal{A}$, *where* δ_{ξ_l} *is a point mass at* ξ_l, *and* $\tilde{p}(\theta_{a,j}) \xrightarrow{d} \delta_0$, $\forall j \notin \mathcal{A}$. *That means* $\mu_l \to \xi_l$, $\alpha_l \to 1$, $s_l \to 0$, $\forall l \in \mathcal{A}$, *and* $\mu_j \to 0$, $s_j \to 0$, $\forall j \notin \mathcal{A}$. *Consequently,* $\mathbb{E}[\Lambda_i^{SLRT}] \to \frac{1}{a(\phi_i)}(b(\mathbf{B}_{a,i:}\boldsymbol{\xi} + \mathbf{B}_{b,i:}\boldsymbol{\theta}_0)\mathbf{B}_{a,i:}\boldsymbol{\xi} - b(\mathbf{B}_{a,i:}\boldsymbol{\xi} + \mathbf{B}_{b,i:}\boldsymbol{\theta}_0) + b(\mathbf{B}_{b,i:}\boldsymbol{\theta}_0))$, *for* $i = 1, \ldots, p$, *is monotonically increasing with the absolute true changed value* $|\xi_l|$ *and dimension* i*'s relation to the* l*th base* (*i.e.*, $\mathbf{B}_{a,il}$), $\forall l \in \mathcal{A}$.

Theorem 2 indicates that in the abnormal condition, we prefer to choose those dimensions most influenced by the most significant abnormal patterns. The detailed verifications of Theorems 1 and 2 are shown in Online Appendix C.

5. Simulation

In this section, we conduct extensive experiments on synthetic data to evaluate the performance of our proposed EF-BSSCD schemes. We also compare them with the following baselines for the POSCD problem:

TRAS. A change detection algorithm based on top-r local cumulative sum (CUSUM) statistics and the corresponding adaptive sampling in Liu et al. (2015);

CMAB(s). A simplified adaptive sampling strategy using UCB in Zhang and Hoi (2019);

SASAM. An adaptive sampling policy considering spatial correlation of variables in Wang et al. (2018);

NAS. A nonparametric adaptive sampling based on anti-rank statistics in Xian et al. (2018);

R-SADA. A rank-based nonparametric adaptive sampling algorithm combined with data augmentation in Xian et al. (2021).

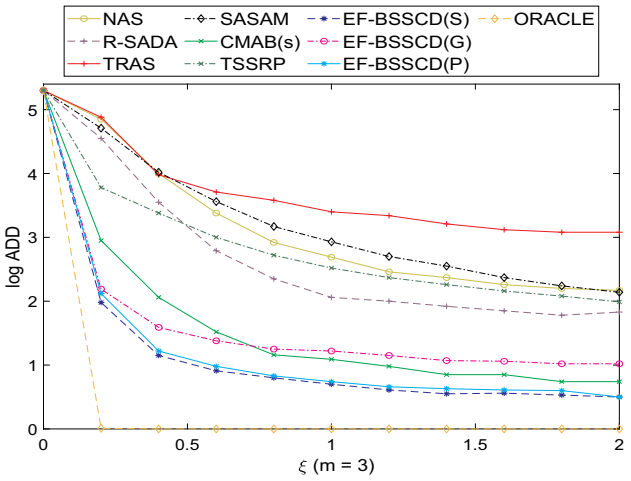
TSSRP. An adaptive sampling algorithm incorporated with Thompson sampling in Zhang and Mei (2020).

5.1. Experiments for Gaussian Data

Consider $\mathbf{X}_t \in \mathcal{R}^{10\times 1}$. We set $\mathbf{B}_b \in \mathcal{R}^{10\times 2}$ as the two lowest-frequency Fourier bases and $\mathbf{B}_a \in \mathcal{R}^{10\times 7}$ as seven four-order B-spline bases with 11 equally spaced knots. We set up the experiment with two phases. In phase I, we first generate $\boldsymbol{\theta}_t$ from $N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\theta}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \sigma_0^2\mathbf{I}$ and $\sigma_0^2 = 0.1$. Then, we generate N_0 normal samples of $\mathbf{X}_t$ from $N(\mathbf{B}_b\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_e)$ as the historical reference samples, where $\boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}$ with $\sigma_e = 0.05$. After that, we use the these samples to estimate $\boldsymbol{\theta}_0$, σ_0, and σ_e—that is, $\hat{\boldsymbol{\theta}}_0$, $\hat{\sigma}_0$, and $\hat{\sigma}_e$—using the procedure shown in Online Appendix E. In phase II, for each replication, we first generate $\boldsymbol{\theta}_t$ from $N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$ and then generate background signals from $N(\mathbf{B}_b\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_e)$ for every t before the change point $\tau = 50$. For later time points $\tau < t \le T = 1{,}000$, we generate $\boldsymbol{\theta}_t$ from $N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$ and also randomly draw j from one to k_a and set $\theta_{a,j} = \xi$. Then, we generate abnormal signals from $N(\mathbf{B}_b\boldsymbol{\theta}_t + \mathbf{B}_a\boldsymbol{\theta}_a, \boldsymbol{\Sigma}_e)$. After that, we run the monitoring scheme and get the detection delay. Replicate the experiment 1,000 times to calculate ADD for a specific change magnitude ξ as the performance evaluation criterion.

For EF-BSSCD-based methods, we set $\lambda = 0.1$, $w_j = 0.1$, $\sigma_j = 2$, $v = 10^{-6}$. The parameter-setting guidelines of EF-BSSCD methods are added in Online Appendix D. As to other baseline methods, although they are used in different circumstances, such as different distributions, with or without data correlation, they can still be applied to our simulated data and compared with our method. We set all their parameters according to the recommendations in their papers, which are shown in Online Appendix D. For all the methods, we tune their detection thresholds to ensure that their average run length before the first false alarm—that is, ARL is exactly 200 (controlling the false alarm rate)—such that their detection performance under abnormal cases can be fairly compared. The ADD values in logarithmic scale for ξ ranging from zero to two with change granularity equaling 0.2 when $m = 3$ are shown in Figure 2. The detailed ADDs together with the standard deviations are shown in Online Appendix D. Clearly, except ORACLE, which needs true prior information of $\boldsymbol{\theta}_a$ and yet may not be available in practice, EF-BSSCD(S) has the smallest ADD generally, followed by EF-BSSCD(P). EF-BSSCD(S) and EF-BSSCD(P) both use the spike-slab estimation of $\boldsymbol{\theta}_a$, which contains sparse change information, in their test statistics.

Figure 2. (Color online) Logarithm of ADDs for Gaussian Distribution with $p = 10$ and $m = 3$



But EF-BSSCD(G) uses the maximum likelihood estimation (MLE) instead and, consequently, cannot track the sparse change very accurately. As to other baselines, CMAB(s) is the most powerful and performs comparably to EF-BSSCD(G) because it also considers the correlation between variables when constructing test statistics and can balance exploration and exploitation using the UCB method. Under nonparametric CUSUM modeling, R-SADA performs better than NAS. Under univariate CUSUM modeling, SASAM performs better than TRAS. That is because R-SADA and SASAM both consider the correlation between variables and help obtain the information for those unobserved variables. TSSRP performs better than NAS, SASAM, and TRAS because it uses Thompson sampling to balance exploration and exploitation. In general, EF-BSSCD-based methods have much superior performance than other baselines.

5.2. Experiments for Poisson Data

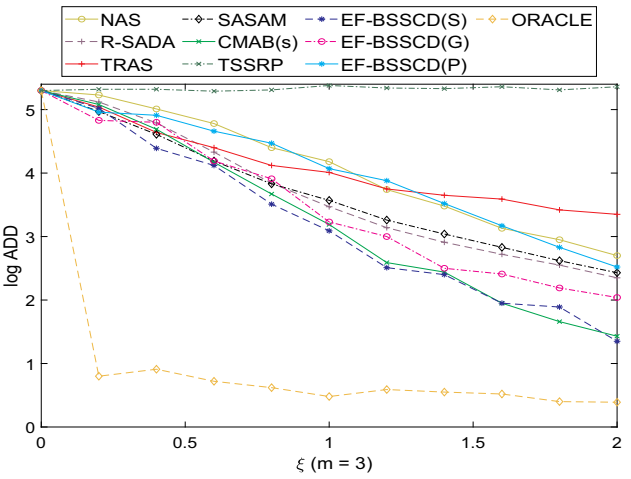
We consider $\mathbf{X}_t \in \mathcal{R}^{10\times 1}$. $\mathbf{B}_b$, $\mathbf{B}_a$, $\boldsymbol{\theta}_t$, and $\boldsymbol{\theta}_a$ are set in the same way as the Gaussian distribution. In phase I, we first generate N_0 normal samples of $\mathbf{X}_t$ from Poisson distribution with mean $\exp(\mathbf{B}_b\boldsymbol{\theta}_t)$ as the historical reference samples. Then, we use the N_0 normal samples to estimate $\boldsymbol{\theta}_0$ and σ_0—that is, $\hat{\boldsymbol{\theta}}_0$ and $\hat{\sigma}_0$—using the procedure in Online Appendix E. In phase II, for each replication, we first generate samples of $\mathbf{X}_t$ from Poisson distribution with mean $\exp(\mathbf{B}_b\boldsymbol{\theta}_t)$ for every t before the change point $\tau = 50$. Then, we randomly draw j from one to k_a, set $\theta_{a,j} = \xi$ for later time points $\tau < t \le T = 1{,}000$, and generate abnormal samples of $\mathbf{X}_t$ from Poisson distribution with mean $\exp(\mathbf{B}_b\boldsymbol{\theta}_t + \mathbf{B}_a\boldsymbol{\theta}_a)$. Also, we run the monitoring scheme and get the detection delay. We replicate the experiment 1,000 times to calculate the ADD for a specific change magnitude ξ.

For EF-BSSCD methods, we set $\lambda = 0.1$, $w_j = 0.5$, $\sigma_j = 2$, $v = 10^{-6}$. The ADD values in logarithmic scale for ξ ranging from zero to two with change granularity equaling 0.2 when $m = 3$ are shown in Figure 3. Except ORACLE, EF-BSSCD(S) performs best at the most magnitudes. CMAB(s) is a little inferior to EF-BSSCD(S) because CMAB(s) is derived from the Gaussian form. Note that, although the test statistic of CMAB(s) is derived from the Gaussian form, it performs best among all baselines because it uses the mean of observations in the test statistic, which converges to a Gaussian distribution when the number of observations approaches infinity, according to central limit theorem. So, its test statistic is not much deviated for non-Gaussian distributions. Conversely, EF-BSSCD(G) only uses the MLE of $\boldsymbol{\theta}_a$, which cannot estimate the parameter accurately, especially when it is a sparse vector. Therefore, its performance is not as good as EF-BSSCD(S) and also CMAB(s). EF-BSSCD(P) performs not so well compared with EF-BSSCD(S) or EF-BSSCD(G) because its performance relies on the accurate estimation of the variance of $\boldsymbol{\theta}_a$. However, because the variance of the observations is large given Poisson distributions, the variance of $\boldsymbol{\theta}_a$ is estimated to be large. Therefore, the performance of PBF is not as good as SLRT, which only uses the expectation of $\boldsymbol{\theta}_a$. R-SADA and SASAM perform closely, better than TRAS and NAS at most change magnitudes, because they both use the correlation information between variables, the same reason as in Section 5.1. As to TSSRP, it does not detect the change because the form of Shiryaev-Roberts-Pollak statistic is sensitive to data distributions.

5.3. Experiments for Binomial Data

We consider $\mathbf{X}_t \in \mathcal{R}^{10\times 1}$. $\mathbf{B}_b$, $\mathbf{B}_a$, $\boldsymbol{\theta}_t$, and $\boldsymbol{\theta}_a$ are set in the same way as Gaussian distribution. In phase I, we first generate N_0 normal samples of $\mathbf{X}_t$ from binomial distribution with mean $\mathbf{N}_p \circ \exp(\mathbf{B}_b\boldsymbol{\theta}_t) \oslash (\mathbf{1}_p + \exp(\mathbf{B}_b\boldsymbol{\theta}_t))$ as

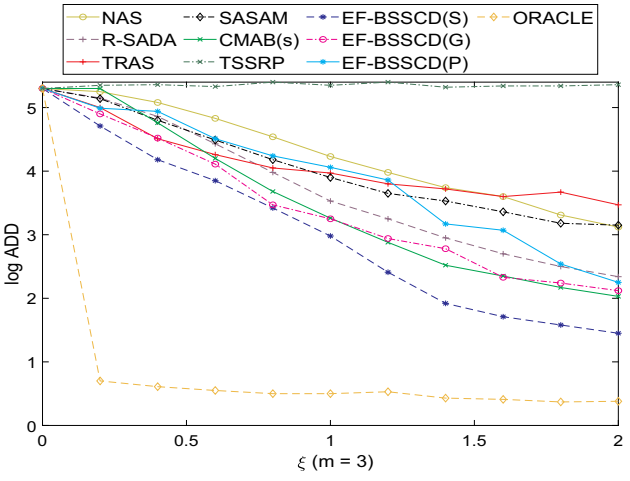
Figure 3. (Color online) Logarithm of ADDs for Poisson Distribution with $p = 10$ and $m = 3$



the historical reference samples, where $\mathbf{N}_p$ denotes a column vector with all p components equal to N and $N = 10$ is the number of trials for binomial distribution. Then, we use the N_0 normal samples to estimate $\boldsymbol{\theta}_0$ and σ_0—that is, $\hat{\boldsymbol{\theta}}_0$ and $\hat{\sigma}_0$—via the method shown in Online Appendix E. In phase II, for each replication, we first generate normal samples of $\mathbf{X}_t$ from the binomial distribution with mean $\mathbf{N}_p \circ \exp(\mathbf{B}_b\boldsymbol{\theta}_t) \oslash (\mathbf{1}_p + \exp(\mathbf{B}_b\boldsymbol{\theta}_t))$ for every t before the change point $\tau = 50$. Then, for every time point $\tau < t \le T = 1{,}000$, we randomly draw j from one to k_a, set $\theta_{a,j} = \xi$, and generate abnormal samples of $\mathbf{X}_t$ from the binomial distribution with mean $\mathbf{N}_p \circ \exp(\mathbf{B}_b\boldsymbol{\theta}_t + \mathbf{B}_a\boldsymbol{\theta}_a) \oslash (\mathbf{1}_p + \exp(\mathbf{B}_b\boldsymbol{\theta}_t + \mathbf{B}_a\boldsymbol{\theta}_a))$. Also, we run the monitoring scheme and get the detection delay. We replicate the experiment for 1,000 times to calculate ADD for a specific change magnitude ξ.

For EF-BSSCD methods, we set $\lambda = 0.1$, $w_j = 0.5$, $\sigma_j = 2$, $v = 10^{-6}$. The ADD values in logarithmic scale for ξ ranging from zero to two with change granularity equaling 0.2 when $m = 3$ are shown in Figure 4. Except ORACLE, EF-BSSCD(S) performs best. CMAB(s) and EF-BSSCD(G) are a little inferior to EF-BSSCD(S), due to the same reason as in Section 5.2. Also, EF-BSSCD(P) performs not so well compared with EF-BSSCD(S) or EF-BSSCD(G), due to the same reason as that in Section 5.2. Under nonparametric CUSUM modeling, R-SADA performs better than NAS because R-SADA considers the correlation information between variables. Under univariate CUSUM modeling, TRAS performs better than SASAM at small magnitudes, but worse at large magnitudes. That is because SASAM only uses maximum local CUSUM statistics, but TRAS uses a top-r local detection scheme based on the sum of the largest-r local CUSUM statistics, which is effective when multiple variables simultaneously shift by a small magnitude. TSSRP also does not detect the change because the form of Shiryaev-Roberts-Pollak statistic is sensitive to data distributions.

Figure 4 (Color online) Logarithm of ADDs for Binomial Distribution with $p = 10$ and $m = 3$



In conclusion, EF-BSSCD(S) is the most robust scheme among various distributions in the exponential family and can achieve the most superior performance compared with other baselines. EF-BSSCD(G) performs almost comparably with CMAB(s), but is a little inferior to EF-BSSCD(S) because it does not combine the sparse change information in its statistic formulation. However, EF-BSSCD(P) is good only for Gaussian distribution, but not satisfactory for Poisson and binomial distributions due to the large variance in the generated data and, thus, the large variance in the estimated parameter. Note that because some test statistics have large computational costs, such as PBF for the Poisson and binomial distributions, as stated in Section 4.4, they are not suitable to handle high-dimensional data. Except for them, for other computational-efficient test statistics, to further show their efficacy for high-dimensional data, we also run experiments with $p = 100$ and $m = 30$ for Gaussian, Poisson, and binomial distributions, respectively, and the detection results are shown in Online Appendix D. The conclusions for high-dimensional cases are almost the same as those in Sections 5.1, 5.2, and 5.3. Also, the advantage of EF-BSSCD methods is more evident compared with other baselines because the change is sparser under the high-dimensional cases.

6. Case Study

In the practical online monitoring process, there are rich circumstances where the sampling resources are in shortage. It urges us to use adaptive sampling methods to select the most informative variables to observe for quick change detection. Here, we present three real-world data sets, which need to detect the change in constrained resource circumstances. The first is solar flare detection, whose data follow a Gaussian distribution. The second is COVID-19 outbreak detection, in which the tested positive number in a district every day approximately follows a Poisson distribution. The third is counted defect change detection in a roll-to-roll manufacturing system, where the number of defects in every batch follows a binomial distribution. We apply EF-BSSCD(S) in these three problems, to demonstrate its applicability to real cases.

6.1. Solar Flare Detection with Gaussian Distributed Data

This is a commonly used data set for POSCD problems in literature (Liu et al. 2015, Wang et al. 2018, Xian et al. 2021). The occurrence of solar flare is harmful to the earth's radio communications. Hence, it's necessary to monitor it in real time and trigger alarms as soon as possible. Satellites with cameras in space take hundreds of pictures of the sun's surface in one minute for the monitoring tasks. However, those abundant pictures cannot be sent back to the earth due to the limited transmission rate in space. That can be solved by our proposed

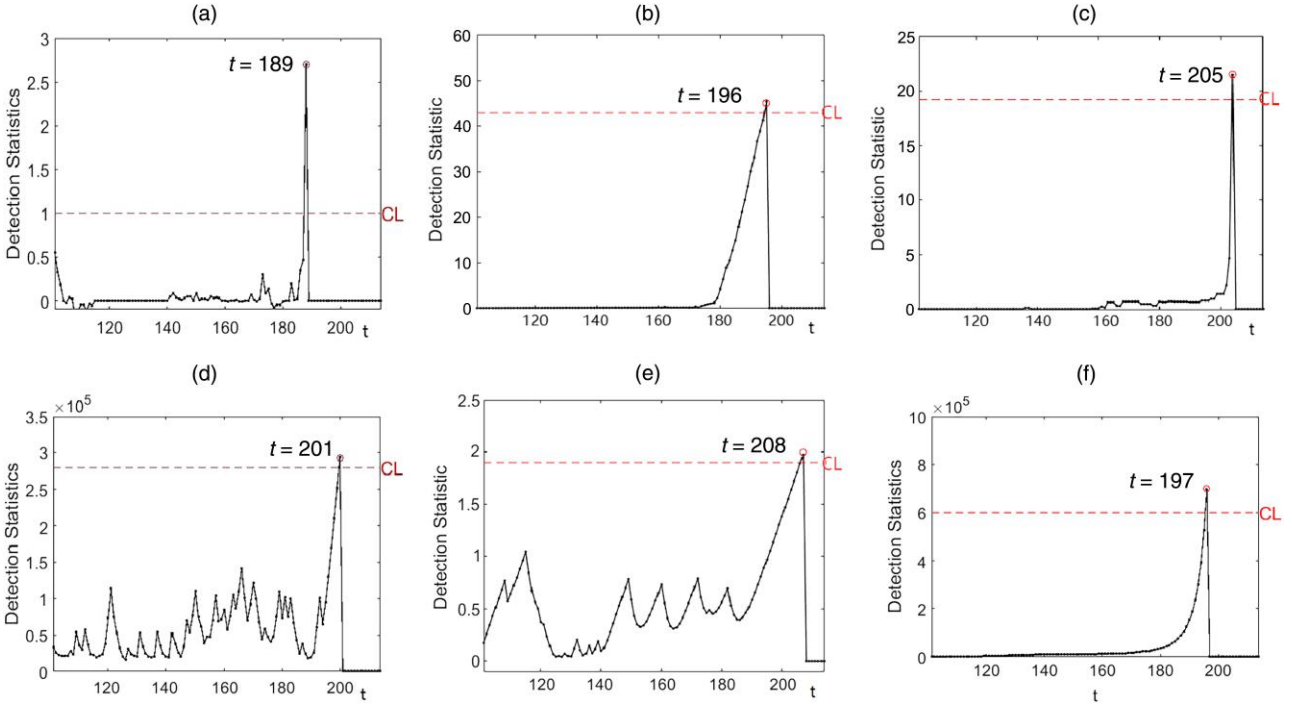
EF-BSSCD algorithms. This data set is of video format and contains 300 frames of images, and each image has 67,744 pixels distributed on a 232×292 grid. One solar flare appears at time $t = 187 \sim 202$. After removing the background information of data using normal frames, the residuals can be regarded as approximately following Gaussian distribution (Xie et al. 2012). We conduct PCA for the before-change frames and extract the first 20 features as the normal bases $\mathbf{B}_b \in \mathcal{R}^{67,744\times 20}$. The extracted PCA scores representing $\hat{\boldsymbol{\theta}}_t$ can be used to further compute the prior mean and covariance of $\boldsymbol{\theta}_t$—that is, $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\Sigma}}_0$. Correspondingly, $\hat{\sigma}_e$ can be computed from the reconstruction error of PCA. As for anomaly bases, it is desirable to construct $\mathbf{B}_a$ according to the size and shape of possible abnormal patterns, which can be obtained from historical images of solar flares. In particular, consider that the pattern of solar flares approximates small circle piles and forms many free shapes by these circle piles. According to this prior information, we generate $\mathbf{B}_a \in \mathcal{R}^{67,744\times 256}$ from three-order B-spline bases with 19 equally spaced knots.

To show the detection efficacy of EF-BSSCD(S), we compare it with the other baselines. Here, we do not compare with CMAB(s) because it needs to construct the covariance matrix of $\mathbf{X}_t$, whose size is $67,744 \times 67,744$. This requires more than 32 GB of computer memory , which is too time-consuming to implement and inefficient for online learning. For EF-BSSCD(S), we set $\sigma_j = 10$, $w_j = 0.1$ for $j = 1, 2, \ldots, k_a$, $v = 10^{-6}$, and $\lambda = 0.1$. Because of the limited transmission bandwidth, only $m = 400$ out of 67,744 pixels are available, whereas Liu et al. (2015) assumes $m = 2{,}000$, and Wang et al. (2018) and Xian et al. (2021) assume $m = 500$. To make different methods comparable, for all the methods, we generate 3,000 normal time points using the first 100 normal samples using bootstrap techniques and decide the detection threshold h such that $ARL = 1{,}100$, according to Wang et al. (2018). We start the online monitoring for data since $t = 101$. The detection statistics of all methods are shown in Figure 5. As we can see, the detection delays (DDs) of the six methods are $DD_{EF-BSSCD(S)} = 3$, $DD_{TRAS} = 10$, $DD_{SASAM} = 19$, $DD_{R-SADA} = 15$, $DD_{NAS} = 22$, and $DD_{TSSRP} = 11$, respectively. EF-BSSCD(S) has the smallest $DD = 3$, outperforming other methods and achieving efficient online anomaly detection.

6.2. COVID-19 Outbreak Detection with Poisson Distributed Data

Pandemics can cause panic among people and damage public health, economics, social communication, etc. Therefore, monitoring the pandemic is very crucial to keep the whole society stable. Medical testing resources, such as testing machines, testing kits, etc., are always in shortage, especially at the beginning of a new pandemic. This leads to the difficulty of data collection for efficient pandemic monitoring. For instance, the outbreak detection of COVID-19 has been a hotspot in recent years. To better understand the COVID-19 status, different types of testing are typically distributed. The Centers for Disease Control and Prevention (CDC) has classified the

Figure 5. (Color online) Detection Statistics for Solar Flare Case with (a) $DD_{EF-BSSCD(S)} = 3$, (b) $DD_{TRAS} = 10$, (c) $DD_{SASAM} = 19$, (d) $DD_{R-SADA} = 15$, (e) $DD_{NAS} = 22$, and (f) $DD_{TSSRP} = 11$



testing for COVID-19 into the following two categories: (1) *Diagnostic testing* is intended to identify current infection in individuals and is performed when a person has symptoms consistent with COVID-19 or has recently been known or suspected to have had exposure to COVID-19. (2) *Screening testing* is recommended for unvaccinated people to identify those who are asymptomatic and do not have been known, suspected, or reported exposure to COVID-19. Screening helps to identify unknown cases so that measures can be taken to prevent further transmission. At the early stage of COVID-19, screening test kits were so limited that we could not test the whole population in an area, but only a subset (sample) of them, and report the change of the infectious rate based on these sampled partial observations. That made it difficult to monitor the change of the infectious rate in an area. Therefore, we apply EF-BSSCD schemes to solve this problem.

In our experiment, we mimic this monitoring scenario by using a data set containing 80 days' tested positive numbers from August 19, 2020, to November 6, 2020, in 18 districts in King County in Washington state, which are proved following Poisson distribution. The data are publicly available at https://kingcounty.gov/depts/health/covid-19/data/summary-dashboard.aspx. Furthermore, we make the following two assumptions for simplicity: (1) To understand the behavior of the COVID-19 infection, only screening tests are used. (2) We need to decide to which districts to distribute the testing kits. If a district is selected, a fixed number of testing kits will be distributed to this region to obtain the number of infected patients as observed tested positive numbers for each time point. For King County, there is an infectious rate change occurring around October 7, 2020 (the 50th day) at the district with zip code 98101, as shown in Figure 6(a), which was later confirmed as the start point of the second wave of COVID-19 by experts. Conversely, most of the other districts, such as 98065 and 98121, do not show a significant change in this time range, as shown in Figure 6, (b) and (c). This means the changed districts are sparse.

Like the case study of solar flare detection, we set the first 12 principal components extracted from the first 50 normal days as $\mathbf{B}_b$—that is, $\mathbf{B}_b \in \mathcal{R}^{18\times 12}$—using PCA methods for exponential family (Collins et al. 2001). Because these 18 districts are dispersed and few of them are adjacent, we set $\mathbf{B}_a = \mathbf{I}$. In our experiment, we assume that every day, due to the shortage of test kits, only 10 out of 18 districts can be tested and the positive numbers of them can be obtained. For EF-BSSCD(S), we set the parameters $\sigma_j = 1$, $w_j = 0.5$ for $j = 1, 2, \ldots, k_a$, $v = 10^{-6}$, and $\lambda = 0.1$. Similar to Section 5.1, for each detection scheme, we generate 1,000 normal time points using the first 50 normal samples by bootstrap techniques and set the detection threshold as $ARL = 550$. After applying those baselines to this data set, the obtained DDs are $DD_{EF-BSSCD(S)} = 1$, $DD_{TRAS} = 2$, $DD_{SASAM} = 9$, $DD_{R-SADA} = 30$, $DD_{NAS} = 30$, $DD_{CMAB(s)} = 8$, and $DD_{TSSRP} = 8$. Their detection statistics for a particular replication are shown in Figure 7. EF-BSSCD(S) has the smallest $DD = 1$, outperforming other methods and achieving efficient online anomaly detection for Poisson distribution. The performance of TRAS is also good in this COVID-19 case. SASAM, CMAB(s), and TSSRP almost trigger alarm at the same time. However, NAS and R-SADA are not able to detect the change in this Poisson case because the change magnitude in this case is not so large for them to trigger an alarm.

6.3. Manufacturing Defect Change Detection with Binomial Distributed Data

Defect detection and quality control are crucial to keep manufacturing systems in a good situation. Generally, the number of defects in a production batch follows binomial distribution. So, developing online statistical process control methods for binomial data is quite important and has been studied in a lot of literature. In this experiment, we consider a roll-to-roll manufacturing system, where the product is produced through a roll-to-roll process by the leveling roller. During this manufacturing process, online manual inspections for different functional categories for each sequential batch

Figure 6. (Color online) Tested Positive Numbers of Three Districts, with Zip Codes (a) 98101, (b) 98065, and (c) 98121, from August 19, 2020, to November 6, 2020

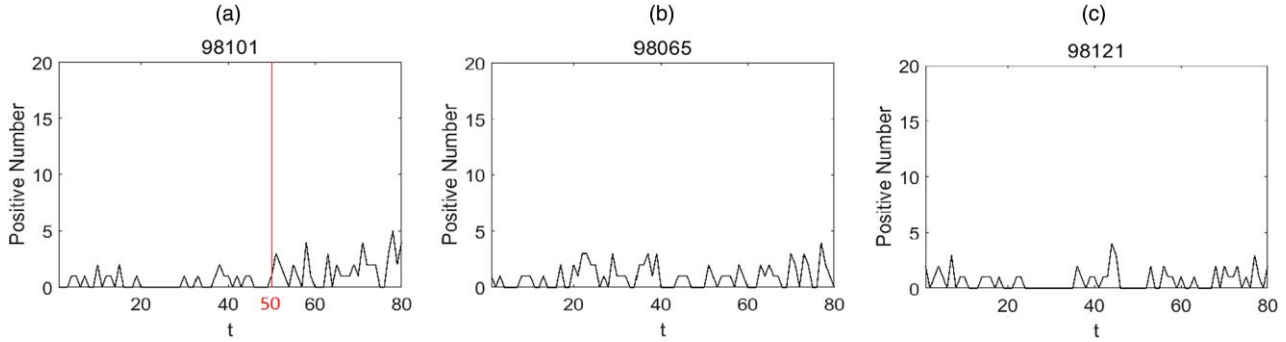
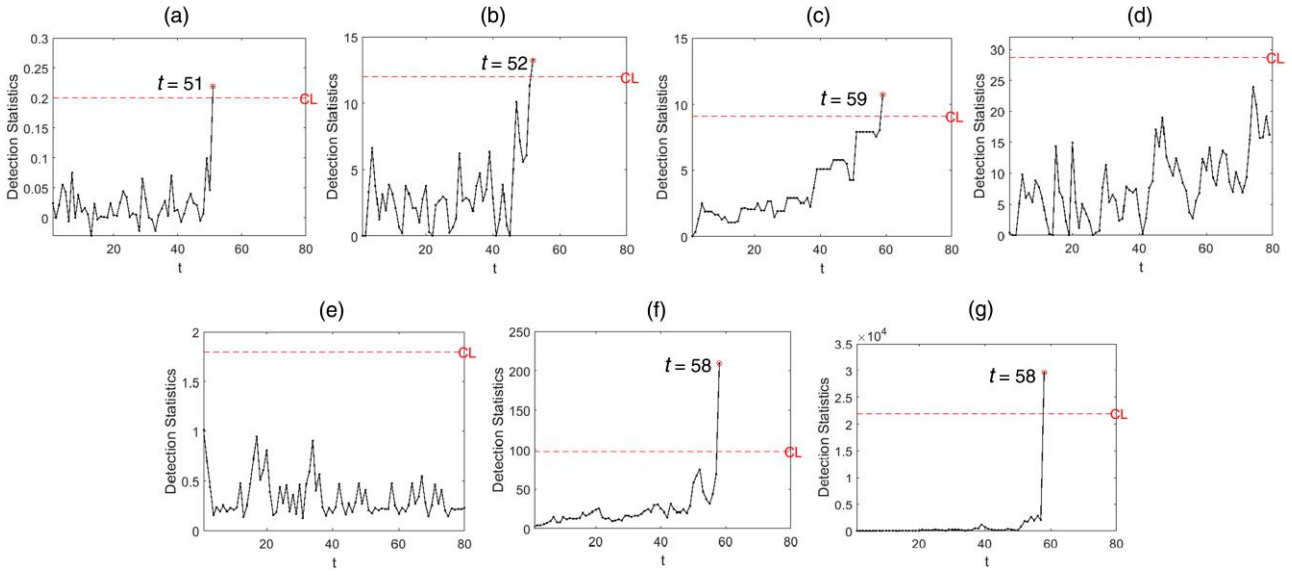


Figure 7. (Color online) Detection Statistics for COVID-19 Outbreak with (a) $DD_{EF-BSSCD(S)} = 1$, (b) $DD_{TRAS} = 2$, (c) $DD_{SASAM} = 9$, (d) $DD_{R-SADA} = 30$, (e) $DD_{NAS} = 30$, (f) $DD_{CMAB(s)} = 8$, and (g) $DD_{TSSRP} = 8$



are performed to detect defective batches. When the manufacturing system is in control (IC), the number of defects of each category in a batch follows binomial distribution. When the manufacturing system goes out-of-control (OC), it will lead to certain categories having larger numbers of defects. However, it is hard and time-consuming to manually inspect all functional categories for each batch due to the high cost of sophisticated manual inspection. One solution is to decide a subset of categories to inspect for each batch, such that, sequentially, we can switch to the influenced OC categories and, consequently, detect the change in the manufacturing process as quickly as possible. For this experiment scenario, the data set has numbers of defects for total 22 categories of 80 sequential batches. Each batch contains $N = 20$ products. The anomaly appears after the 50th batch with only one category influenced, as shown in Figure 8(a). Yet, there is no change for other categories, as shown in Figure 8, (b) and (c) as two examples. This indicates that the change is quite sparse.

Like the former two cases, we set the first 12 principle components extracted from the first 50 normal batches as $\mathbf{B}_b$—that is, $\mathbf{B}_b \in \mathcal{R}^{22\times 12}$—using PCA methods for exponential family (Collins et al. 2001). According to practitioners' domain knowledge, the anomaly happening to any category has little correlation with other categories, so we set $\mathbf{B}_a = \mathbf{I}$. We assume, because of the limitation of the inspection cost, that only 15 out of 22 categories' defects can be inspected and obtained for each batch. For EF-BSSCD(S), we set the parameters $\sigma_j = 3$, $w_j = 0.5$ for $j = 1, 2, \ldots, k_a$, $v = 10^{-6}$, and $\lambda = 0.1$. To compare with other baselines, we generate 1,000 normal sequential batches using the first 50 normal samples through bootstrap techniques and set $ARL = 280$ for all the baselines. The DDs are $DD_{EF-BSSCD(S)} = 3$, $DD_{TRAS} = 6$, $DD_{SASAM} = 14$, $DD_{R-SADA} = 13$, $DD_{NAS} = 30$, $DD_{CMAB(s)} = 30$, and $DD_{TSSRP} = 13$. Their detection statistics for a particular

Figure 8. (Color online) Number of Defects of Three Functional Categories

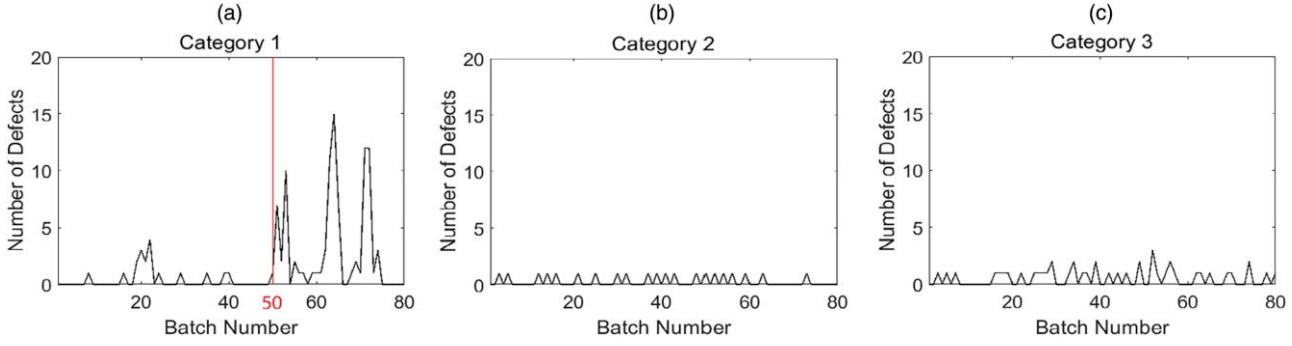
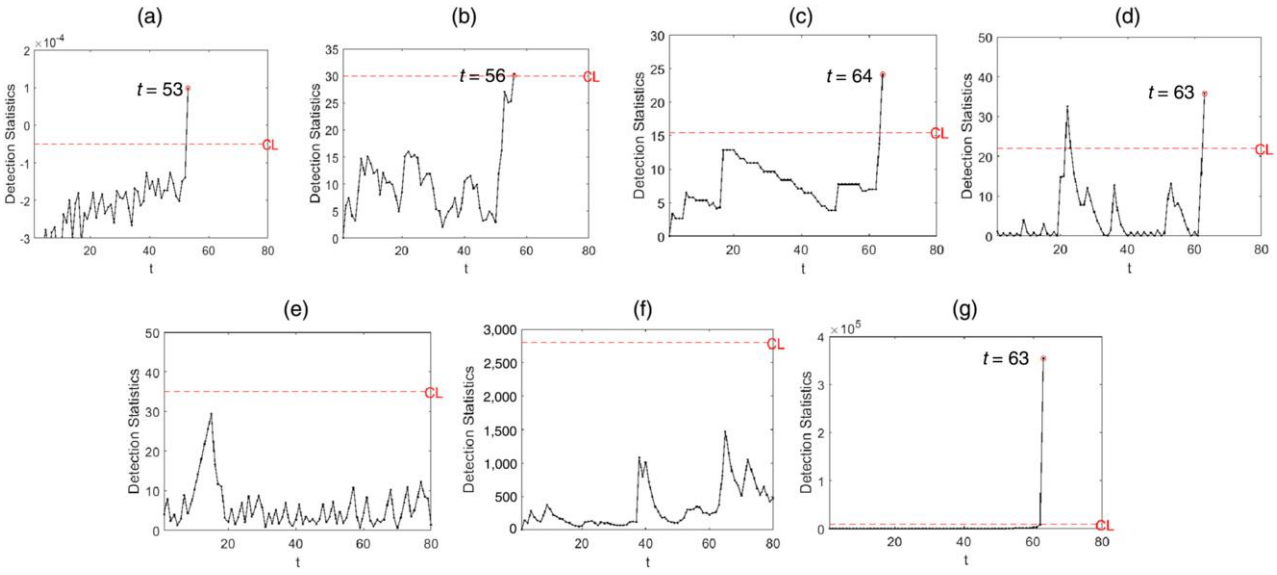


Figure 9. (Color online) Detection Statistics for Counted Defect Change with (a) $DD_{EF-BSSCD(S)} = 3$, (b) $DD_{TRAS} = 6$, (c) $DD_{SASAM} = 14$, (d) $DD_{R-SADA} = 13$ (One False Alarm Exists), (e) $DD_{NAS} = 30$, (f) $DD_{CMAB(s)} = 30$, and (g) $DD_{TSSRP} = 13$



replication are shown in Figure 9. EF-BSSCD(S) has the smallest $DD = 3$. Like the COVID-19 case, the DD of TRAS is a little longer than EF-BSSCD(S), but also fine to apply to this case, which demonstrates that it is also robust to various real data. SASAM, R-SADA, and TSSRP detect the change around at the same time. However, CMAB(s) and NAS both fail to detect the change in this case, due to their lower efficiency for sparse change.

7. Conclusion

This paper proposes a holistic framework for high-dimensional POSCD for exponential-family distributed data. We first use a composite decomposition for feature extraction for high-dimensional data, whose decomposition coefficients on abnormal bases can be regarded as indicators of sparse changes and used for test statistic construction. Multiple test statistics, such as SLRT, GLRT, PBF, etc., can be incorporated. By further treating the test statistic as the reward function in a CMAB problem, an adaptive sampling strategy based on Thompson sampling with the balance of exploration and exploitation is conducted. Taking the SLRT-based detection procedure as an example, theoretical properties of the proposed algorithm under asymptotic conditions are justified. Furthermore, extensive experiments for Gaussian, Poisson, and binomial distributions also demonstrate the superiority of the proposed framework.

References

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *25th Annu. Conf. Learn. Theory* (PMLR, New York), 39.1–39.26.

Ban Y, He J (2020) Generic outlier detection in multi-armed bandit. *Proc. 26th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 913–923.

Bubeck S, Wang T, Viswanathan N (2013) Multiple identifications in multi-armed bandits. *30th Internat. Conf. Machine Learning* (JMLR, Norfolk, MA), 258–265.

Cao Y, Wen Z, Kveton B, Xie Y (2019) Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. *22nd Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 418–427.

Cavenaghi E, Sottocornola G, Stella F, Zanker M (2021) Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm. *Entropy* 23(3):380.

Chen W, Wang Y, Yuan Y (2013) Combinatorial multi-armed bandit: General framework and applications. *Internat. Conf. Machine Learn.* (JMLR, Norfolk, MA), 151–159.

Chen W, Wang Y, Yuan Y, Wang Q (2016) Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *J. Machine Learn. Res.* 17(1):1746–1778.

Chiquet J, Mariadassou M, Robin S (2018) Variational inference for probabilistic Poisson PCA. *Ann. Appl. Statist.* 12(4):2674–2698.

Collins M, Dasgupta S, Schapire RE (2001) A generalization of principal components analysis to the exponential family. *Adv. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 617–624.

Durand A, Gagné C (2014) Thompson sampling for combinatorial bandits and its application to online feature selection. *Workshops Twenty-Eighth AAAI Conf. Artificial Intelligence* (AAAI Press, Palo Alto, CA).

George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. *Statist. Sinica* 7(2):339–373.

Ghatak G (2020) A change-detection-based Thompson sampling framework for non-stationary bandits. *IEEE Trans. Comput.* 70(10):1670–1676.

Gómez AME, Li D, Paynabar K (2022) An adaptive sampling strategy for online monitoring and diagnosis of high-dimensional streaming data. *Technometrics* 64(2):253–269.

Hardin JW, Hilbe JW (2012) *Generalized Linear Models and Extensions* (Stata Press Books, College Station, TX).

Ishwaran H, Rao JS (2005) Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* 33(2):730–773.

Jun KS, Jamieson K, Nowak R, Zhu X (2016) Top arm identification in multi-armed bandits with batch arm pulls. *Proc. 19th Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 139–148.

Kaufmann E, Korda N, Munos R (2012) Thompson sampling: An asymptotically optimal finite-time analysis. Bshouty NH, Stoltz G, Vayatis N, Zeugmann T, eds. *Internat. Conf. Algorithmic Learn. Theory* (Springer, Berlin, Heidelberg), 199–213.

Li G, Huang JZ, Shen H (2018a) Exponential family functional data analysis via a low-rank model. *Biometrics* 74(4):1301–1310.

Li J, Zhang J, Pang N, Qin X (2018b) Weighted outlier detection of high-dimensional categorical data using feature grouping. *IEEE Trans. Systems Man Cybernetics Systems* 50(11):4295–4308.

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

Lu X, Yuan Y, Zheng X (2016) Joint dictionary learning for multispectral change detection. *IEEE Trans. Cybernetics* 47(4):884–897.

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

Meier L, Van de Geer S, Bühlmann P (2009) High-dimensional additive modeling. *Ann. Statist.* 37(6B):3779–3821.

Montgomery DC (2007) *Introduction to Statistical Quality Control* (John Wiley & Sons, Hoboken, NJ).

Mukherjee A, Marozzi M (2021) Nonparametric phase-II control charts for monitoring high-dimensional processes with unknown parameters. *J. Quality Tech.* 54(1):44–64.

Petterson J, Yu J, McAuley J, Caetano T (2009) Exponential family graph matching and ranking. *Adv. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 1455–1463.

Qiu P, Zou C, Wang Z (2010) Nonparametric profile monitoring by mixed effects modeling. *Technometrics* 52(3):265–277.

Ročková V (2018) Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* 46(1):401–437.

Russo DJ, Van Roy B, Kazerouni A, Osband I, Wen Z (2018) A tutorial on Thompson sampling. *Foundations Trends® Machine Learn.* 11(1):1–96.

Tartakovsky A (2019) *Sequential Change Detection and Hypothesis Testing: General Non-iid Stochastic Models and Asymptotically Optimal Rules* (CRC Press, Boca Raton, FL).

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3–4):285–294.

Wang Y, Blei DM (2019) Frequentist consistency of variational bayes. *J. Amer. Statist. Assoc.* 114(527):1147–1161.

Wang K, Jiang W (2009) High-dimensional process monitoring and fault isolation via variable selection. *J. Quality Tech.* 41(3):247–258.

Wang A, Xian X, Tsung F, Liu K (2018) A spatial-adaptive sampling procedure for online monitoring of big data streams. *J. Quality Tech.* 50(4):329–343.

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

Xian X, Zhang C, Bonk S, Liu K (2021) Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation. *J. Quality Tech.* 53(2):135–153.

Xie Y, Huang J, Willett R (2012) Change-point detection for high-dimensional time series with missing data. *IEEE J. Selected Topics Signal Processing* 7(1):12–27.

Yan H, Paynabar K, Shi J (2014) Image-based process monitoring using low-rank tensor decomposition. *IEEE Trans. Automation Sci. Engrg.* 12(1):216–227.

Yan H, Paynabar K, Shi J (2017) Anomaly detection in images with smooth background via smooth-sparse decomposition. *Technometrics* 59(1):102–114.

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

Zhang C, Hoi SC (2019) Partially observable multi-sensor sequential change detection: A combinatorial multi-armed bandit approach. *Proc. Conf. AAAI Artificial Intelligence*, vol. 33, 5733–5740.

Zhang W, Mei Y (2020) Bandit change-point detection for real-time monitoring high-dimensional data under sampling control. Preprint, submitted September 24, https://arxiv.org/abs/2009.11891.

Zhang C, Chen N, Wu J (2020) Spatial rank-based high-dimensional monitoring through random projection. *J. Quality Tech.* 52(2): 111–127.

Zhang L, Wang K, Chen N (2016) Monitoring wafers' geometric quality using an additive Gaussian process model. *IIE Trans.* 48(1):1–15.

Zhang C, Yan H, Lee S, Shi J (2018) Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis. *IISE Trans.* 50(10):878–891.

Zhou H, Wang L, Varshney L, Lim EP (2020) A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. *Proc. Conf. AAAI Artificial Intelligence*, vol. 34, 6933–6940.

Zhuang H, Wang C, Wang Y (2017) Identifying outlier arms in multi-armed bandit. *Adv. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 5204–5213.

Zou C, Qiu P (2009) Multivariate statistical process control using lasso. *J. Amer. Statist. Assoc.* 104(488):1586–1596.