JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 538, 794–808: Theory and Methods
https://doi.org/10.1080/01621459.2020.1819295

Large-Scale Datastreams Surveillance via Pattern-Oriented-Sampling

Haojie Ren[a,d], Changliang Zou[b], Nan Chen[c], and Runze Li[d]

[a]School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China; [b]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China; [c]Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore; [d]Department of Statistics, The Pennsylvania State University at University Park, State College, PA

ABSTRACT

Monitoring large-scale datastreams with limited resources has become increasingly important for real-time detection of abnormal activities in many applications. Despite the availability of large datasets, the challenges associated with designing an efficient change-detection when clustering or spatial pattern exists are not yet well addressed. In this article, a design-adaptive testing procedure is developed when only a limited number of streaming observations can be accessed at each time. We derive an optimal sampling strategy, the pattern-oriented-sampling, with which the proposed test possesses asymptotically and locally best power under alternatives. Then, a sequential change-detection procedure is proposed by integrating this test with generalized likelihood ratio approach. Benefiting from dynamically estimating the optimal sampling design, the proposed procedure is able to improve the sensitivity in detecting clustered changes compared with existing procedures. Its advantages are demonstrated in numerical simulations and a real data example. Ignoring the neighboring information of spatially structured data will tend to diminish the detection effectiveness of traditional detection procedures. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2019
Accepted August 2020

KEYWORDS
Design-adaptive test; Generalized likelihood ratio method; Kernel smoothing; Optimal sampling; Sequential change detection; Statistical monitoring

1. Introduction

1.1. Motivation and Model

Motivated by an empirical analysis of New York hourly traffic data in Section 4.2, we study online monitoring of large-scale datastreams. Such problems can be formulated as a change-point detection as follows. Suppose that m streams of observations $\mathbf{X}_t = (X_{t1}, \ldots, X_{tm})^\top$ are collected over time, admitting a change-point model

$$X_{tj} = \begin{cases} \mu_0(\mathbf{s}_{tj}) + e_{tj}, & \text{for } t = 1, \ldots, \tau, \\ \mu_1(\mathbf{s}_{tj}) + e_{tj}, & \text{for } t = \tau + 1, \ldots, \end{cases}, \quad j = 1, \ldots, m, \tag{1}$$

where τ is an unknown change-point, $\mathbf{e}_t = (e_{t1}, \ldots, e_{tm})^\top, t = 1, 2, \ldots$, are independent and identically distributed (iid) random vectors satisfying that $\mathbb{E}(\mathbf{e}_t) = \mathbf{0}$, and $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are the mean functions before and after the change point, respectively. It is assumed that $\mu_1(\mathbf{s})$ is not equal to $\mu_0(\mathbf{s})$ in one or more regions $\Omega \subset \Gamma$, the support of $\mathbf{s} \in \mathbb{R}^d$, and their difference, $\mu_1(\mathbf{s}) - \mu_0(\mathbf{s})$, is (at least locally) smooth. Each datastream X_{tj} is associated with a d-dimensional auxiliary covariate $\mathbf{s}_{tj} \in \mathcal{S}$, such as spatial information or other characteristic information, where we assume that there are total m_a possible covariates $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^{m_a} \subset \Gamma$. Our goal is to raise an alarm as quickly as possible after τ.

For the New York hourly traffic data, the $\mathbf{s}$ variable is the location, which can be represented as a two-dimensional vector. One would put a sensor at every intersection in a city to record full information on the city's traffic network and monitor continuously over time. For a big city like New York, this will definitely demand a huge storage space and computational power. Thus, it is of great interest to select representative intersections or locations in the city to monitor traffic network. This motivates us to study how to dynamically sample a small portion of $\mathbf{s}$'s from all possible locations (i.e., $m \ll m_a$) so that we can carry out real-time detection of abnormal activities or events.

With rapid advance of technology, this type of large-scale datastreams arise in many applications. As another example, wireless sensor networks provide us the capability to build large-scale systems for real-time monitoring of environmental disasters such as landslides. It is common that a large number of sensors are installed in different locations of a mountain in advance, each of which records a datastream measuring the landslide movement or acceleration at its location. If there is an emergency signal at some places, nearby sensors are more likely to be affected than those that are spatially distant. Consequently, only a small proportion of sensors or datastreams, spatially adjacent, tend to be active or significant (Ciampalini et al. 2015). Similarly, disease outbreaks tend to affect spatially contiguous areas (Neill 2012), either because of contagion (e.g., human-to-human transmission) or because the cases share a common source (e.g., contaminated drinking water). Other examples include earthquake detection and network flow surveillance (Liu, Liu, and Ansari 2014).

CONTACT Changliang Zou nk.chlzou@gmail.com School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, 94 Weijin Road, Nankai District, Tianjin 300071, P. R. China and Runze Li rzli@psu.edu Department of Statistics, The Pennsylvania State University, University Park, PA 16802.
All authors equally contributed to this work, and the authors are listed in seniority.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

1.2. Challenges and Connections to Existing Works

A common feature for the aforementioned applications is the *clustering* pattern among these massive datastreams; it is thus anticipated that a location and its adjacent neighbors fall in a similar type of region, either significant or insignificant. Also, the number of affected datastreams by an event is usually not large, that is, certain sparsity structure exists. Taking the monitoring of landslides as an example, X_{tj} can be the sensor observation of landslide acceleration at the location $\mathbf{s}_{tj}$ and time t. An emergency signal may result that $\mu_1(\cdot)$ differs from $\mu_0(\cdot)$ in some small regions, namely, the nonzero components of $\mu_1(\cdot) - \mu_0(\cdot)$ are clustered. Hence, our first task is to take into account clustering structures in massive data.

Another major challenge in real-time detection of abnormal activities for large-scale datastreams is the limited budgets available for online monitoring. The second task of this work is to study how to optimally use a given budget to detect changes in the process in (1). In general, there are two types of limited budgets: computational and measurement budgets. The computational budget is referred to as the computer memory and storage space for real-time analysis of the large volume datastreams. The measurement budget is referred to the measurement constrains or costs that make the collection of streaming observations of all spatial points at each time impossible. For example, researchers monitoring the landslides may be reluctant to maintain all the sensors in operating conditions all the time considering the power consumption and service life of sensors. As a result, these constraints often enable us to obtain, at each time point, only a limited number of streaming observations X_{tj}'s at a small subset of spatial locations $\mathbf{s}_{tj}$'s from a large pool of given points to identity change signals. Therefore, it is crucial to design a dynamic sampling strategy to automatically select m informative points over time when m is significantly less than the size of total possible points m_a.

Recently, there is a great deal of effort to develop new methods for detecting change-points that can accommodate high-dimensionality and dependence within components, but in a nonsequential setting (see, e.g., Wang and Samworth 2018; Enikeeva and Harchaoui 2019, and the references therein). Some authors adapt various sequential change-detection methods to large-scale surveillance, such as Veeravalli (2001), Tartakovsky et al. (2006), Zou and Qiu (2009), Mei (2010), Xie and Siegmund (2013), Zou et al. (2015), and Li (2019), but they did not consider the constraints on resources, which may greatly hamper their applicability to massive data applications. A more related work is Liu, Mei, and Shi (2015) which proposed a sampling strategy assuming that only partial observations are available. Their procedure is powerful in general but does not consider the spatial structure of the anomalies. See also Xian et al. (2019) for more discussions. Xian, Wang, and Liu (2018) and Wang et al. (2018) further proposed sampling and detection schemes with spatial information, but their test statistics are essentially in a form of extreme value and thus may not be fully efficient for the potentially clustered signals. Some works, for example, Yan, Paynabar, and Shi (2018) among many others, proposed to use dimension reduction techniques to deal with the spatiotemporal correlation structure, but their settings are completely different from ours.

1.3. Our Contribution

In this article, we propose a new procedure that can optimally detect changes in the process (1) under a given computation or measurement budget at each time point. The proposed procedure stems from a design-adaptive testing procedure, which is to select the most informative sample points from the full set of locations and then to construct a kernel-based model specification test statistic based on the selected observations. Under mild conditions, we derive the optimal sampling strategy, the *pattern-oriented-sampling* (POS), with which the proposed test possesses asymptotically and locally best power under alternatives. Then, the test is adapted to sequential change-point detection by using the generalized-likelihood-ratio-based scheme (Siegmund and Venkatraman 1995). A data-driven approach to dynamically estimate the optimal sampling is developed, giving the proposed method an edge over conventional methods in terms of the detection ability when clustering pattern exhibits.

The POS procedure addresses two key questions in a unified framework: how to construct an efficient statistic via aggregation of large-scale datastream observations, and how to arrange limited resources to improve detection sensitivity with a dynamic sampling strategy. Our simulation results clearly demonstrate the superiority of the proposed procedure over existing ones in terms of the finite-sample performance.

1.4. Organization

The remainder of this article is organized as follows. In Section 2, we present the construction of optimal designs in the context of a two-stage test and its theoretical properties. Extensions to the sequential detection problem are given in Section 3, along with detailed discussions on asymptotic optimality and practical implementation. Numerical studies and a real-data example are conducted in Section 4. Section 5 concludes the article with some remarks, and theoretical proofs are delineated in the Appendix. Some technical details and additional numerical results are provided in the supplementary materials.

Notations. For a $m \times m$ matrix $\mathbf{A}$, let $\|\mathbf{A}\|_S$ be its spectral norm (largest eigenvalue in absolute value) and $\|\mathbf{A}\|_F$ be its Frobenius norm (the square root of the sum of the squared eigenvalues), respectively. Let $A_m \sim B_m$ denote that there is a constant $C > 1$ such that $B_m/C \le A_m \le B_m C$ with probability tending to 1. The $A_m \approx B_m$ means that two quantities A_m and B_m are asymptotically equivalent, in the sense that both $A_m/B_m \xrightarrow{p} 1$. Two diagonal matrices $\mathbf{A}_m \approx \mathbf{B}_m$ are asymptotically equivalent if and only if their diagonal components are asymptotically and uniformly equivalent. We use diag($\mathbf{a}$) and diag($\mathbf{A}$) to denote the $m \times m$ diagonal matrices with diagonal entries $\mathbf{a} = (a_1, \ldots, a_m)^\top$, and the $m \times 1$ vector of diagonal components of the matrix $\mathbf{A}$, respectively.

2. Optimal Sampling Plan With Clustering Patterns

Given a measurement constraint, at time t, only m streaming observations $\{X_{tj}\}_{j=1}^m$ can be observed from a chosen subset of entire sampling space, $\{\mathbf{s}_{tj}\}_{j=1}^m \subset \mathcal{S}$, where $m \ll m_a$. To make inference on the model (1), we start with the associated testing

problem at a given time point. For simplicity, we suppress the dependence on t which should not cause any confusion.

2.1. Kernel-Based Tests

Suppose that $\mathbf{X} = (X_1, \ldots, X_m)^\top$ is a random vector observed with $\{\mathbf{s}_j\}_{j=1}^m$ by model (1). Consider the following hypothesis testing problem

$$\mathbb{H}_0 : \mu(\mathbf{s}_j) = \mu_0(\mathbf{s}_j) \text{ for all } j \longleftrightarrow \mathbb{H}_1 : \mu(\mathbf{s}_j) = \mu_1(\mathbf{s}_j), \quad (2)$$

where $\mu_1(\mathbf{s}) \neq \mu_0(\mathbf{s})$, for some $\mathbf{s}_j \in \Omega$. As a convention in the practice of monitoring problems, $\mu_0(\mathbf{s})$ can be either user specified or estimated by sufficiently large historical samples prior to the surveillance procedure (Liu, Mei, and Shi 2015; Zou et al. 2015). Hence, we assume $\mu_0(\mathbf{s})$ is known and particularly $\mu_0(\mathbf{s}) \equiv 0$ without loss of generality. Our implicit assumption is that for data with clustered signals, a location and its adjacent neighbors have similar status, either having nonzero $\mu_1(\mathbf{s})$ or not. Furthermore, the signal strength at one location is similar to those at its adjacent neighbors. To utilize information within its neighbors, assume $\mu(\mathbf{s})$ is a smooth function of $\mathbf{s}$. This assumption enables us to construct a local aggregation of standardized test statistics at points located adjacent to $\mathbf{s}_j$ instead of using the test statistics at $\mathbf{s}_j$ only. In the same spirit of nonparametric regression with fixed design, let us regard $\mu(\mathbf{s}) = \mathbb{E}(X \mid \mathbf{s})$. It is worth to clarify that our goal is to derive an optimal sampling plan on spatial point $\mathbf{s}$ rather than the nonparametric regression on $\mu(\mathbf{s})$. Suppose that $f(\mathbf{s})$ is the sampling density function of the spatial point $\mathbf{s}$. Our target is to find the optimal $f(\mathbf{s})$ so that our test for (2) has the best power under the alternative.

Intuitively, $\mathbb{E}\{\mu^2(\mathbf{s})\}$ may be used to measure the derivation of $\mu(\mathbf{s})$ from $\mu_0(\mathbf{s})$. Here, $\mathbb{E}$ is the expectation with respect to the sampling density $f(\mathbf{s})$. Note that $\mathbb{E}\{\mu^2(\mathbf{s})\} = \mathbb{E}\{X\mathbb{E}(X \mid \mathbf{s})\} \geq 0$. Our test can be based on $\mathbb{E}\{X\mathbb{E}(X \mid \mathbf{s})\}$, which is a popular quantity in the context of model specification test (Guerre and Lavergne 2005). For a given sampling density f, a nonparametric approximation to $m\mathbb{E}\{X\mathbb{E}(X \mid \mathbf{s})\}$ is

$$D_{\mathrm{f}} = \frac{1}{(m-1)} \sum_{j=1}^{m} \sum_{k \neq j}^{m} \frac{K_h(\mathbf{s}_k - \mathbf{s}_j) X_k X_j}{\sqrt{f(\mathbf{s}_k)}\sqrt{f(\mathbf{s}_j)}}, \quad (3)$$

where $h > 0$ is a bandwidth depending on the number of observations m, $K(\cdot)$ is a kernel function with $K_h(\cdot) = K(\cdot/h)/h^d$. For notation simplicity, we suppress the dependence of D_{f} on m and h, but emphasize its dependence on $f(\cdot)$.

Under some mild conditions, it can be shown by theory of U-statistics that D_{f} is asymptotically normal under the null hypothesis, say

$$T_{\mathrm{f}} := (D_{\mathrm{f}} - \mu_{\mathrm{f}}^{(0)})/\sigma_{\mathrm{f}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad (4)$$

where $\mu_{\mathrm{f}}^{(0)}$ and σ_{f}^2 are the asymptotic mean and variance of D_{f} under $\mathbb{H}_0$, respectively. This is a well-known result if $X_j, j = 1, \ldots, m$ are independent (see, e.g., Zheng 1996). More discussions can be found in Proposition A.1 in the Appendix. Under certain local alternatives, the asymptotic distribution of D_{f} is also a normal distribution but with a different mean, $\mu_{\mathrm{f}}^{(1)}$, which depends on $\mu(\mathbf{s})$ and $f(\mathbf{s})$. It evokes some insight: if an appropriate sampling distribution $f(\mathbf{s})$ can be chosen, the power of D_{f} can be maximized. Next, we will present a result that sheds lights on how to determine the optimal sampling distribution. First, the following assumptions are needed to facilitate the derivation.

Assumption 1 (Spatial sampling points). The sampling density of $\mathbf{s}$, $f(\mathbf{s})$, is bounded away from zero on the compact support Γ and has bounded derivative.

Assumption 2 (Covariance structure). The noise $\mathbf{e} = \Lambda^{1/2}\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ and ε_i's are independent variables with mean zero and variance σ^2. Λ is a correlation matrix with the components $\rho_{j_1 j_2} = \rho\left(\|\mathbf{s}_{j_1} - \mathbf{s}_{j_2}\|/\lambda\right)$ with some $\lambda > 0$, where $\rho(\cdot)$ is a continuous, nonnegative and nonincreasing function on $\mathbb{R}$ with $\rho(0) = 1$ and $\int_0^\infty \rho(x)dx < \infty$.

Assumption 3 (Moments condition). For a fixed $C < \infty$, $\mathbb{E}(|\varepsilon_i|^\gamma) \leq C < \infty$ with $\gamma \geq 4$.

Assumption 4 (Kernel function and bandwidth). $K(\cdot)$ is a Lipschitz continuous, nonnegative, symmetric and bounded kernel function from $\mathbb{R}^d$ that integrates to 1 and has bounded derivatives. The h and λ satisfy $h \to 0$ and $mh^d/(\eta_m \log m) \to \infty$ as $m \to \infty$, where $\eta_m = \max(1, m\lambda^d)$.

Remark 1. Assumption 1 implies that the density function of $\mathbf{s}$ is positive, which ensures that the denominator used in our statistic D_{f} is bounded away from 0 with high probability. The condition of bounded derivatives is not the weakest possible; we only require $f(\cdot)$ to be Lipschitz continuous if some other conditions are imposed. In the asymptotic analysis of conventional model specification tests, the errors are usually assumed to be independent, which is not always met in high-dimensional settings. Assumption 2 allows the presence of spatial correlations which widens the scope of applications of our method. Assumption 3 is necessary in establishing asymptotic normality of D_{f}. Assumption 4 is commonly used in kernel-based methods. The condition $mh^d/(\eta_m \log m) \to \infty$ is required due to the presence of spatial-correlation. When $m\lambda^d \to 0$, say the correlation is sufficiently small, this condition reduces to the conventional one $mh^d/\log m \to \infty$.

We consider a sequence of local alternatives

$$\mathbb{H}_{1m} : \mu(\mathbf{s}) = \delta_m l(\mathbf{s}) \text{ for } \mathbf{s} \in \Gamma, \quad (5)$$

where $l(\mathbf{s})$ is a twice continuously differentiable function on Γ and bounded away from zero on some $\Omega \subseteq \Gamma$, and $\delta_m \to 0$ as $m \to \infty$. Under (5), the difference between $\mu_1(\cdot)$ and $\mu_0(\cdot)$, quantified by $\delta_m l(\mathbf{s})$, goes to zero as $m \to \infty$.

Theorem 1. Suppose Assumptions 1–4 hold. Under the local alternative (5) with $\delta_m = (mh^{d/2}/\eta_m)^{-1/2}$, the test based on D_{f} reaches its asymptotic best power if the sampling density $f(\mathbf{s}) \propto \mu^2(\mathbf{s})$.

Under the local alternative (5), $\mathbb{E}(D_{\mathrm{f}}) \hat{=} \mu_{\mathrm{f}}^{(1)} = \mu_{\mathrm{f}}^{(0)} + \mathbb{E}_{\mathrm{f}}\{l^2(\mathbf{s})\}$, and its asymptotic variance σ_{f}^2 does not depend on $l(\mathbf{s})$, so an explicit power expression, depending on $\mu_{\mathrm{f}}^{(1)}$, can be obtained. Theorem 1 enlightens us to construct a locally most powerful kernel-based test by choosing the sampling distribution $f(\mathbf{s})$ as a linear function of $l^2(\mathbf{s})$, that is, $\mu^2(\mathbf{s})$.

2.2. Two-Stage Tests With Adaptive Sampling

In practice, the statistics D_f can be carried out with an estimated $f(\mathbf{s})$ if a two-stage procedure is used. The theory of two-stage combination test can be described as follows: at the first stage, compute the test statistic T_{f_1} by (4) based on the observations $\{\mathbf{s}_{1j}, X_{1j}\}_{j=1}^m$, where $\{\mathbf{s}_{1j}\}_{j=1}^m$ come from a given distribution $f_1(\mathbf{s})$, and then make an early decision by rejecting $\mathbb{H}_0$ if $T_{f_1} > z_{\alpha_1}$ or otherwise continuing to test at the second stage if $T_{f_1} \le z_{\alpha_1}$. At the second stage, compute the test statistic T_{f_2} based on new observations $\{\mathbf{s}_{2j}, X_{2j}\}_{j=1}^m$ and reject $\mathbb{H}_0$ if $T_{f_2} > z_{\alpha_2}$, where $\{\mathbf{s}_{2j}\}_{j=1}^m$ are sampled from another distribution $f_2(\mathbf{s})$. The constants α_1 and α_2 are subject to a control of the error rate at a prespecified level, α, by the test, and $z_{\alpha_i} (i = 1, 2)$ denotes the α_i upper quantile of the standard normal distribution.

Traditionally, one sets $f_2(\cdot) = f_1(\cdot)$. However, Theorem 1 motivates us to sample from $f_2(\mathbf{s}) \propto \mu^2(\mathbf{s})$ at the second stage. A consistent estimator of $\mu(\mathbf{s})$ based on the first stage samples is

$$\hat{\mu}(\mathbf{s}) = \frac{\sum_{j=1}^m K_{h_E}(\mathbf{s}_{1j} - \mathbf{s}) X_{1j}}{\sum_{j=1}^m K_{h_E}(\mathbf{s}_{1j} - \mathbf{s})}, \tag{6}$$

where h_E is a prespecified bandwidth for estimation, satisfying the condition $h_E \to 0$ and $mh_E \to \infty$ as $m \to \infty$. Then $f_2(\mathbf{s})$ can be chosen as

$$f_2(\mathbf{s}) = \max\left\{\zeta_m, \hat{\mu}^2(\mathbf{s})\right\} \Big/ \int \max\left\{\zeta_m, \hat{\mu}^2(\mathbf{s})\right\} d\mathbf{s}, \tag{7}$$

where ζ_m is fixed as $O\{\eta_m (\log m)^c / (mh_E^d)\}$ for $c > 1$. The use of ζ_m in (7) is to ensure that the estimated density is bounded away from zero, especially under the null hypothesis where $\mu(\mathbf{s}) = 0, \forall \mathbf{s} \in \Gamma$. From Lemma A.2 in the Appendix, $\sup_{\mathbf{s}} |\hat{\mu}(\mathbf{s})| = o_p(\zeta_m^{1/2})$ under $\mathbb{H}_0$, and consequently, with probability tending to one $f_2(\mathbf{s}) = 1/|\Gamma|$, that is, a uniform distribution. Hereafter, we shall refer the testing method based on this sampling procedure as POS-based test.

The next result shows that the effect of replacing $\mu(\mathbf{s})$ by appropriate estimators can be asymptotically negligible and the efficiency of the locally most powerful test can still be achieved under certain conditions.

Theorem 2. Consider the local alternative (5). Suppose Assumptions 1–4 hold. If $(h_E^2/h)^{d/2}/(\log m)^c \to \infty$, the power function of the POS test with $f_2(\mathbf{s})$ in (7) can be approximated by

$$\beta_{f_1,f_2} = \Phi(-z_1) + \Phi(z_1)\,\Phi(-z_2),$$

where $z_1 = z_{\alpha_1} - \mathbb{E}_{f_1}\{l^2(\mathbf{s})\}/\kappa(f_1, \Gamma)$, $z_2 = z_{\alpha_2} - \{\int l^4(\mathbf{s})d\mathbf{s} / \int l^2(\mathbf{s})d\mathbf{s}\}/\kappa(f_T, \Gamma)$, $f_T(\mathbf{s}) = l^2(\mathbf{s}) / \int l^2(\mathbf{s})d\mathbf{s}$, and for $\Omega \subseteq \Gamma$

$$\kappa^2(f, \Omega) = \begin{cases} 2|\Omega| \int_{\mathbb{R}^d} K^2(\mathbf{u})d\mathbf{u}, & \text{if } m\lambda^d \to 0, \\ 2\int_{\mathbb{R}^d} K^2(\mathbf{u})d\mathbf{u} & \\ \times \left(\int_{\mathbb{R}^+} \rho(u)du\right)^2 \int_\Omega f^2(\mathbf{s})d\mathbf{s}, & \text{if } m^{1-2\gamma}\lambda^d \to \infty. \end{cases}$$

In this result, we need to distinguish the two cases, $m\lambda^d \to 0$ or $m\lambda^d \to \infty$, which correspond to the situation that the spatial-correlation is negligible or nonnegligible, respectively. It is easy to verify that the power function of the two-stage test with traditionally $f_2(\mathbf{s}) = f_1(\mathbf{s})$ is

$$\beta_{f_1,f_1} = \Phi(-z_1) + \Phi(z_1)\,\Phi\left(-z_{\alpha_2} + \mathbb{E}_{f_1}\{l^2(\mathbf{s})\}/\kappa(f_1, \Gamma)\right).$$

Thus, using the estimated density $f_2(\mathbf{s})$ in (7) for sampling at the second stage yields a more powerful test. For example, when $m\lambda^d \to 0$ this is clear, because $\mathbb{E}_{f_1}\{l^2(\mathbf{s})\} \le \int l^4(\mathbf{s})d\mathbf{s} / \int l^2(\mathbf{s})d\mathbf{s}$; see the Appendix for more detailed derivation.

Furthermore, it turns out that the advantage of the POS test with the optimal sampling strategy could be much more prominent when $\mu(\mathbf{s})$ exhibits sparse and "clustered" pattern. Consider a sequence of "singular" local alternatives

$$\mathbb{H}'_{1m} : \mu(\mathbf{s}) = \delta'_m l(\mathbf{s}) \text{ for } \mathbf{s} \in \Omega_m, \tag{8}$$

where $\Omega_m \subset \Gamma$ satisfying $|\Omega_m| \approx a_m$ with $a_m \to 0$ being a real deterministic sequence and $l(\mathbf{s})$ is bounded away from zero on Ω_m almost everywhere. In other words, $l(\mathbf{s}) = 0$ for $\mathbf{s} \notin \Omega_m$ and the support of $l(\mathbf{s})$ depends on m. The main feature of these "singular" alternatives is that they have narrow spikes as m increases. Loosely speaking, the $\mathbb{H}'_{1m}$ can be thought of as representing sparse alternatives, while the $\mathbb{H}_{1m}$ in (5) as representing dense alternatives. We have the following result.

Corollary 1. Consider the local alternative (8) where $\delta'_m = (mh^{d/2}a_m^{1/2}/\eta_m)^{-1/2}$, $a_m/h^d \to \infty$ and $\inf_{\mathbf{s}\in\Omega_m} l(\mathbf{s}) \ge \underline{l} > 0$. Suppose the conditions in Theorem 2 hold. If $(h_E^2/h)^{d/2} / \{a_m^{1/2}(\log m)^c\} \to \infty$, the asymptotic power of the POS test with $f_2(\mathbf{s})$ in (7) is not smaller than $\alpha_1 + (1 - \alpha_1)\Phi\left(-z'_2\right)$, where $z'_2 = z_{\alpha_2} - \underline{l}^2/(2\int_{u\in\mathbb{R}^+} \rho(u)du \int_{\mathbb{R}^d} K^2(\mathbf{u})d\mathbf{u})^{1/2}$.

In this result, Assumption $a_m/h^d \to \infty$ implies that our method could work well as long as Ω_m's size goes to zero slower than h as the number of observations m increases regardless of number or shapes of change areas. By Theorem 2, we can see that the two-stage test with a fixed sampling density f on Γ in both stages has nontrivial power (larger than the size $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$) against a contiguous alternative of order $(mh^{d/2}a_m/\eta_m)^{-1/2}$ under $\mathbb{H}'_{1m}$. In contrast, the test with "optimal sampling" has nontrivial power provided that δ'_m is as large as $(mh^{d/2}a_m^{1/2}/\eta_m)^{-1/2}$, resulting in an improvement. This can be intuitively understood that we actually put much more sampling points within the region Ω_m at the second stage through "optimal sampling." In this situation, the POS test would also outperform the one-stage test with the size $m' = 2m$ from the asymptotic viewpoint, if the same bandwidth is used. An illustrative example can be found in the supplementary materials.

Another benefit of the proposed test under this sparse alternative is that the assumptions imposed on h_E are more relaxed in Corollary 1 compared with that in Theorem 2. Consider the following case. If $m\lambda^d \to 0$ and $\mu(\mathbf{s})$ differs from 0 only in a small region Ω_m so that $a_m = h^{d/2}$, the optimal rate of nonparametric estimation $h_E = O(m^{-1/(4+d)})$ can be allowed as long as the bandwidth h used for testing satisfies $mh^{3(4+d)/4} \to 0$.

Remark 2. To generate m samples at the second stage, one can use acceptance-rejection algorithm to sample from the

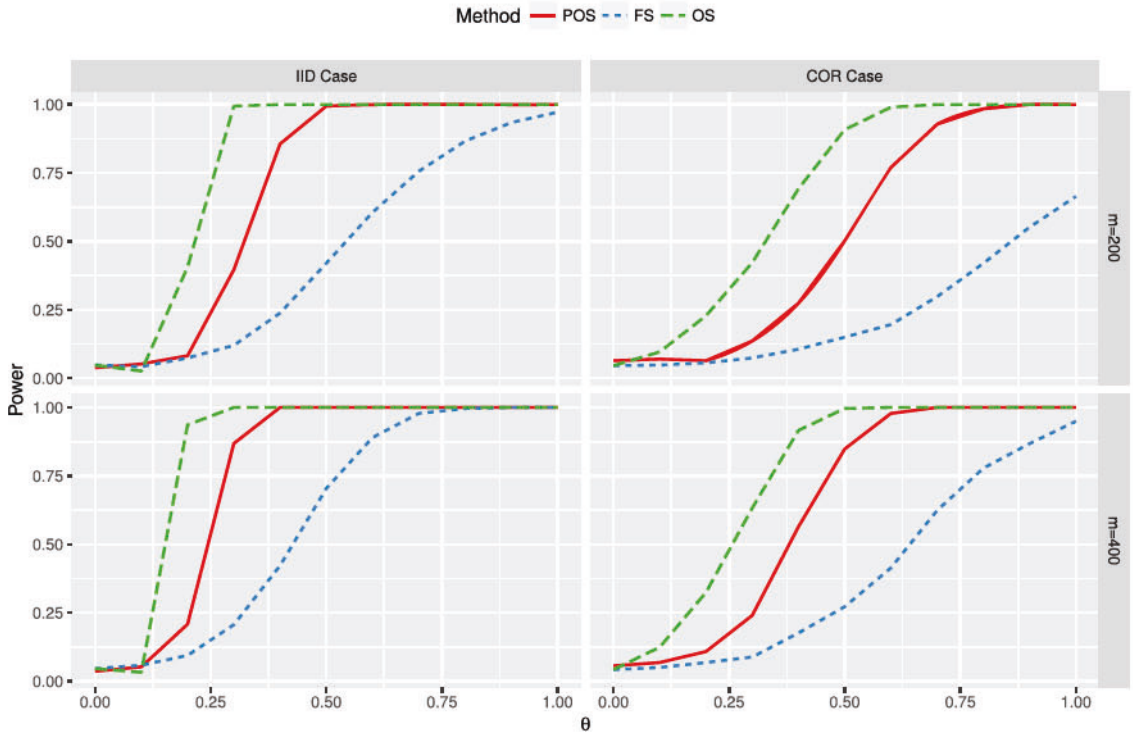

Figure 1. Power comparison under Scenario (I) when $m = 200, 400$ and errors follow $\mathcal{N}(0, 1)$ for the tests with different sampling methods at the second stage: our POS procedure (POS; solid red line); fixed sampling procedure (FS; dot blue line); oracle sampling procedure (OS; dash green line).

estimated continuous density (Gilks and Wild 1992). Here, we suggest to apply a discrete distribution to approximate the sampling density $f_2(\mathbf{s})$. In particular, we estimate the sampling probabilities by (7) for all m_a possible design points in sample space $\mathcal{S}$ and then draw m sampling points by a multinomial distribution with probabilities $\{f_2(\mathbf{s}_j)\}_{j=1}^{m_a}$, where $\mathbf{s}_j \in \mathcal{S}$. In practice, this discrete approximation is sufficient for large m_a and is fast to implement with the computational complexity $O(m_a m h_E^d)$. In this way, the computation of the two-stage test with POS is $O(m^2 h^d + m_a m h_E^d)$, where $O(m^2 h^d)$ is the complexity for computing T_{f_i}.

We use a toy example to demonstrate the performance of the proposed test. Suppose that $\mu(\mathbf{s}) = \theta \exp(2\mathbf{s})\mathbb{I}_\Omega(\mathbf{s})$, where $\Gamma = [0, 1]$, $\theta \geq 0$, $\Omega = [0.4, 0.5]$ and $\mathbb{I}(\cdot)$ is the indicator function. We consider both independent and autoregressive covariance structures; see Scenario (I) in Section 4 for details. The significant levels α_1 and α_2 for the two stages are equal and the global significant level is controlled as $\alpha = 0.05$. Figure 1 compares the power curves of our proposed method, POS, with two other two-stage tests. They differ only in their sampling distributions at the second stage. FS uses a fixed uniform sampling and OS uses the "oracle" sampling density function $\mu^2(\mathbf{s})/\int \mu^2(\mathbf{s})d\mathbf{s}$ as if $\mu(\mathbf{s})$ is known. At the first stage, all three methods use uniform sampling. The improvement of our adaptive-sampling-based test over FS is clear. The test with "oracle" sampling has superior performance as expected, but the difference between OS and POS becomes smaller as m increases, which is consistent with the assertion in Theorem 2. Another example in the supplementary materials shows the performance of the proposed test under the "sparse local alternatives in (8)."

3. Sequential Change Detection Based on Dynamic Sampling

3.1. The Procedure

When data are collected over time, we have the random vector $\mathbf{X}_t$ of which the components are obtained at spatial points $\mathcal{S}_t = \{\mathbf{s}_{tj}\}_{j=1}^m \subset \mathcal{S}$ at time t. For detecting changes under model (1), we notice that

$$T_{\mathrm{f}_t} \overset{d}{\approx} \begin{cases} \mathcal{N}(0, 1), & \text{for} \quad t = 1, \ldots, \tau, \\ \mathcal{N}(\theta_t, 1) & \text{for} \quad t = \tau + 1, \ldots, \end{cases} \tag{9}$$

where $\theta_t > 0$, T_{f_t} is the test statistic in (4) based on $\{\mathbf{X}_t, \mathcal{S}_t\}$ and the sampling density $f_t(\cdot)$ and the "$\overset{d}{\approx}$" means asymptotically distributed. This motivates us to use a standard sequential change-point detection approach based on the generalized likelihood ratio (GLR) statistic, which is defined as

$$Q_t = \max_{0 \leq k < t} \frac{1}{\sqrt{t-k}} \sum_{i=k+1}^{t} T_{\mathrm{f}_i}. \tag{10}$$

The process triggers a signal if $Q_t \geq L$, where L is chosen to achieve a specified average run length (ARL) under the null state. Generally speaking, the Q_t can adaptively detect the change-point τ with large probability as $t - \tau$ becomes large. Note that the original monitoring problem is analogous to "high-dimensional" detection, the use of T_{f_t} plays an important role of dimension reduction that exploits the clustering information and casts all current observations into a univariate value, facilitating the construction of change-detection procedure.

In fact, many methods in sequential change detection, like the popular cumulative (CUSUM) or exponentially weighted

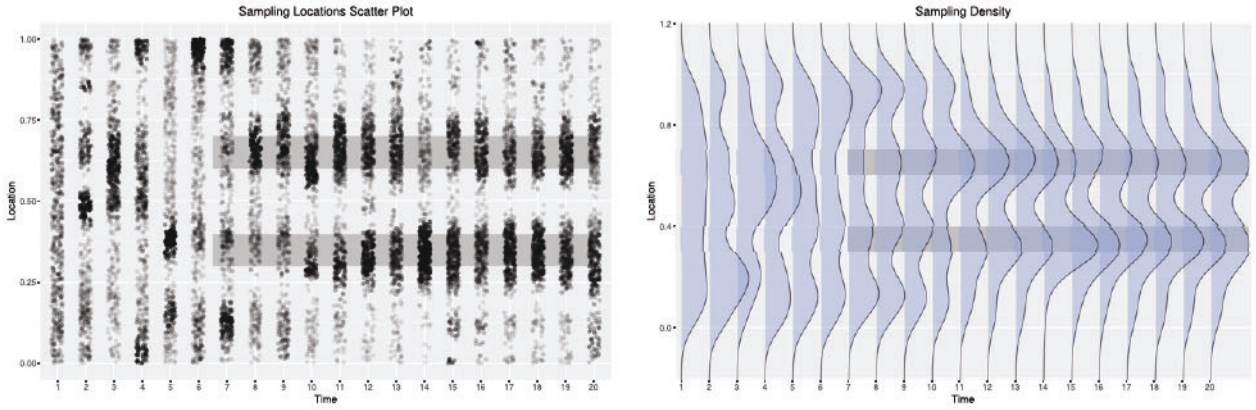

Figure 2. Left: Sampling locations at each time point under Scenario (I) with $\Omega = [0.3, 0.4] \cup [0.6, 0.7]$ and $\theta = 5$ when $\tau = 7$ and errors follow $\mathcal{N}(0, 1)$. Black dots stand for sampling locations while the gray areas represent those in which change signals occur. Right: Corresponding sampling densities at each time point.

moving average charting schemes, can be used to detect T_{f_t}'s mean change. In particular, we notice that the signal magnitude, θ_t in (9), is likely to be time-varying under the alternative condition and thus the GLR would not be optimal in terms of mean-change detection. One workable alternative is to apply some methods for detecting dynamic mean changes, such as Han and Tsung (2006) which developed a reference-free procedure that can trace and detect mean changes without knowing the change pattern. Because the GLR detection procedure has the robust performance without selecting tuning parameters and is a diagnostic aid to identify the change point, we focus on GLR method though our ideas are readily extended to other change-point detection approaches.

We also generalize the two-stage adaptive sampling strategy in Section 2.2 to a dynamic sampling strategy so that T_{f_t} would tend to automatically select the variables that have changes with a high probability. Clearly, at time point t, an ideal sampling distribution for the time point $t + 1$ would be estimated by the whole observations under the alternative, that is, $\{(\mathbf{X}_{\tau+1}, \mathcal{S}_{\tau+1}), \ldots, (\mathbf{X}_t, \mathcal{S}_t)\}$ if $t > \tau$. However, accurate estimation of τ is not feasible when the detection scheme is ongoing. To this end, we suggest to obtain the estimators $\hat{\mu}_t(\mathbf{s})$ by (6) based on the most recent ω observations up to time point t, say, $\{(\mathbf{X}_{t-\omega+1}, \mathcal{S}_{t-\omega+1}), \ldots, (\mathbf{X}_t, \mathcal{S}_t)\}$, and then combine them together by

$$\tilde{\mu}_{\omega,t}(\mathbf{s}) = \frac{1}{\omega} \sum_{i=t-\omega+1}^{t} \hat{\mu}_i(\mathbf{s}), \tag{11}$$

where ω is a pre-chosen window size. At least after the time point $t - \omega + 1$ time point, we will have "homogenous" samples to estimate $\mu(\mathbf{s})$. This choice may not be the best strategy but it is sufficient for practical use due to its simplicity. Then, the $\mathcal{S}_{t+1}$ can be generated from a new sampling distribution by (7) based on the estimator $\tilde{\mu}_{\omega,t}(\mathbf{s})$.

We use the example of Scenario (I) in Section 4 to illustrate this procedure. Figure 2 presents the sampling locations and the estimated sampling density at each time point with two change areas $\Omega = [0.3, 0.4] \cup [0.6, 0.7]$ and $\theta = 5$ when the change-point is at $\tau = 7$ and errors follow $\mathcal{N}(0, 1)$. It is evident that the sampling locations do not exhibit any structure before the change occurs, that is, $t \leq \tau$. In contrast, most locations selected after the change point are around Ω (the gray areas). Accordingly, the estimated sampling density tends to be bimodal around the two clustered areas after the change point. In particular, within the time interval $t \in [7, 12]$, more and more sampled locations gather in the areas with change signals, because $\tilde{\mu}_{\omega,t}(s)$ is more and more accurately estimated. The sampling points determined by this procedure perform reasonably well as they can dynamically capture change patterns. Two illustrative examples with $d = 2$ can be found in the supplementary materials.

Our proposed procedure for online detection using POS is outlined as follows:

Algorithm 1. (*Online detection using pattern-oriented-sampling*)

Step 1. At time point t, compute T_{f_t} by (4) based on $\{\mathbf{X}_t, \mathcal{S}_t\}$.
Step 2. Obtain Q_t by (10). If $Q_t \geq L$, stop; otherwise, go to Step 3.
Step 3. Compute $\hat{\mu}_t(\mathbf{s})$ by (6) based on $\{\mathbf{X}_t, \mathcal{S}_t\}$, and then obtain $f_{t+1}(\mathbf{s})$ by (7) with $\tilde{\mu}_{\omega,t}(\mathbf{s})$ in (11).
Step 4. Sample m covariates from $f_{t+1}(\mathbf{s})$ as $\mathcal{S}_{t+1}$ and get $\mathbf{X}_{t+1}$. Return to Step 1.

To better understand the advantage of the proposed POS procedure over the GLR method with the fixed-sampling strategy, we discuss some asymptotic properties under the alternative conditions. Suppose a threshold L_{FS} is chosen so that $\Pr(\max_{1 \leq t' \leq t} \tilde{Q}_{t'} < L_{\mathrm{FS}}) = 1 - \alpha$ for some $\alpha \in (0, 1)$, where $\tilde{Q}_t$ is the detection statistic similar to Q_t in (10) but using the FS strategy with a given density f on Γ.

Theorem 3. Suppose Assumptions 1–4 hold.

(i) Under the null state $t \leq \tau$, $\Pr(\max_{1 \leq t' \leq t} Q_{t'} < L_{\mathrm{FS}}) \to 1 - \alpha$ as $m \to \infty$.
(ii) Assume that after the change-point τ, a change specified by the alternative occurs in the process. If the conditions in Corollary 1 are all satisfied, at any time point $t > \tau$, the power of POS is not smaller than that of the FS procedure.

The first part of this theorem claims that the POS and FS procedures perform equivalently from asymptotic viewpoints when the process is under the null state. Under the alternative state, the proposed procedure would reject the null hypothesis at the time t with a larger probability, which results in faster detection than the method with a fixed-sampling density.

Besides quickly detecting a process change, a by-product of our proposed procedure is a diagnostic aid to identify when the process change occurs after a signal is triggered at the time point $t = t_s$. The suggested change-point estimator is

$$\hat{\tau}_{t_s} = \text{argmax}_{k\in[0,t_s)} \frac{1}{\sqrt{t_s - k}} \sum_{i=k+1}^{t_s} T_{\mathrm{f}_i}. \tag{12}$$

Finally, we present an asymptotic result on the consistency of $\hat{\tau}_{t_s}$, which ensures that it is asymptotically effective.

Corollary 2. Assume the conditions in Theorem 3(ii) all hold and process stops at t_s. Then, as $m\to\infty$ and $t_s - \tau\to\infty$, $|\hat{\tau}_{t_s} - \tau| = O_p(1)$.

Some numerical results in the supplementary materials demonstrate that $\hat{\tau}_{t_s}$ works reasonable well in finite-sample settings.

3.2. Practical Guidelines

This section provides guidelines on how to design and implement the proposed scheme. Several practical issues are discussed, including the choices of bandwidths and determination of the threshold L.

Like many other smoothing-based tests, the performance of the proposed procedure depends on the bandwidth h in the test statistic D_{f} and the h_{E} in the density estimator (6). By Lemma A.2 in the Appendix, the optimal h_{E} for the estimator $\hat{\mu}(\mathbf{s})$ is of order $O\{(m/\eta_m)^{-1/(4+d)}\}$. We therefore consider to use $(m/\log m)^{-1/(4+d)}$ as a rule of thumb when there is no information about the magnitude of spatial correlation.

It is widely acknowledged that the optimal h for nonparametric estimation is generally not optimal for testing (Hart 2013). The selection of h for optimal power is an open problem. Asymptotically, a range of bandwidths that satisfy Assumption 4 will maintain the consistency of the test, while a specific bandwidth will maximize the power. The amount of smoothing applied will affect the power of the test, but our simulations demonstrate that the observed significance changes only mildly over a wide range of values of h. In addition, a larger bandwidth generally leads to higher power. This can be understood from the power function given in Theorem 2. However, in practice, the condition $h \to 0$ will be violated if h is too large, and an inappropriately large h will yield an excessive false alarm rate. Based on our numerical experience, we recommend the empirical bandwidth $h = h_{\mathrm{E}}^{2+c'}$ with some $c' > 0$ so that the condition $(h_{\mathrm{E}}^2/h)^{d/2}/\{a_m^{1/2}(\log m)^c\}\to\infty$ in Theorem 3 is roughly valid. This formula works well for a wide range of models and sizes, as shown in Section 4. Selecting an optimal h warrants future attention.

It seems that determining L is nontrivial as its accurate value depends on m, $K(\cdot)$, h and h_{E} given an ARL under the null state. Simulation- or resampling-based approaches (Chatterjee and Qiu 2009) can be applied here but they usually require extensive computation in massive data. As revealed by Theorem 3(i), the distribution of Q_t is asymptotically free of those quantities and thus the corresponding L could be approximated by the control limit L_{FS} designed for the GLR scheme with a simple FS strategy. This L_{FS} can be found by either the simulation method or large-sample approximation (Han and Tsung 2006). From the simulation results in the next section, we can see that this approximation method works reasonably well in most cases.

The optimal choice of the window size ω used in $\tilde{\mu}_{\omega,t}(\mathbf{s})$ depends on the signal strength, say δ'_m (or δ_m). A smaller ω leads to a quicker reaction to larger signals, while a larger ω would be more efficient when δ'_m is small as more observations will be collected to estimate $\mu(\mathbf{s})$ after the change occurs. However, our simulation results show that the performance of POS is not affected by this value too much, and in general $\omega \in [5, 10]$ is suitable.

The GLR procedure may be computationally intensive when t is very large because its complexity is linear with t. Conventionally, a moving window of size w_Q is employed to alleviate the computational burden, that is, building the GLR statistics as $\tilde{Q}_t = \max_{t-w_Q\le k<t} 1/\sqrt{t-k}\sum_{i=k+1}^{t} T_{\mathrm{f}_i}$. Generally, increasing the value of w_Q could improve the ability of the GLR to detect small signals (Hawkins and Zamba 2005). Some simulations results in the supplementary materials appear that the proposed method leads to similar detection abilities as long as $w_Q \ge 50$. We use $\tilde{Q}_t$ and choose $w_Q = 50$ in the simulation study. In this case, the complexity of online search is bounded by $O(w_Q)$. Moreover, the POS procedure only needs to store w_Q test statistics T_{f_t} for computing $\tilde{Q}_t$ besides m observations collected at each time point. This advantage of memory cost is crucial, especially for online monitoring of large-scale datastreams.

For implementing the POS, we need to compute $\mu_{\mathrm{f}_t}^{(0)}$ and $\sigma_{\mathrm{f}_t}^2$. It can be easily seen that $\mu_{\mathrm{f}_t}^{(0)} = \mathrm{tr}(\mathbf{V}_{\mathrm{f}_t})$ and $\sigma_{\mathrm{f}_t}^2 \approx 2\mathrm{tr}(\mathbf{V}_{\mathrm{f}_t}^2)$, where $\mathbf{V}_{\mathrm{f}_t} = \mathrm{diag}\left(1/\sqrt{\mathbf{f}_t}\right)\boldsymbol{\Lambda}^{1/2}\mathbf{K}_h\boldsymbol{\Lambda}^{1/2}\mathrm{diag}\left(1/\sqrt{\mathbf{f}_t}\right)$, $\mathbf{K}_h = \left(\frac{K_h(\mathbf{s}_i-\mathbf{s}_j)}{(m-1)}\mathbb{I}(i\ne j)\right)_{m\times m}$ and $\mathbf{f}_t = \{f_t(\mathbf{s}_1),\ldots,f_t(\mathbf{s}_m)\}^\top$. The $\boldsymbol{\Lambda}$ can be estimated from a historical sample and accordingly $\mu_{\mathrm{f}_t}^{(0)}$ and $\sigma_{\mathrm{f}_t}^2$ can be updated online once $f_t(\cdot)$ is obtained.

As a side note, it is clear that better performances would be expected when m is larger. Some examples can be found in the supplementary materials. How large m is allowed depends on practitioners' resource constraints, such as the storage space, computational power, and processing time.

3.3. Extensions to the Case With Sampling Cost

In some applications, practitioners may be reluctant to relocate the sampling points frequently due to its complexity or high setup cost. For example, in environmental surveillance, changing sampling locations requires redeployment of sensors and resources. In such cases, we can simplify the detection procedure via classifying the process into two statuses with a warning threshold. That is, a fixed set of sample points is always used to save sampling cost when the statistics are less than the warning threshold, while a new set of sample points is selected adaptively by our POS procedure to improve the monitoring efficiency when the test statistic exceeds the warning threshold.

Specially, let $\mathcal{S}_0$ with $|\mathcal{S}_0| = m$ be the fixed design set, and L_w is a suitable warning threshold to determine whether a dynamic sampling is needed at the next time point. The modified POS procedure for the case with sampling cost is outlined as follows.

Algorithm 2. (*Modified POS (MPOS) procedure in the presence of sampling cost*)

Step 1. At time point t, compute T_{f_t} by (4) based on $\{\mathbf{X}_t, \mathcal{S}_t\}$.
Step 2. Obtain Q_t by (10). If $Q_t \geq L$, stop; otherwise, go to Step 3.
Step 3. If $Q_t < L_w$, let $\mathcal{S}_{t+1} = \mathcal{S}_0$; if $Q_t \in [L_w, L)$, obtain $f_{t+1}(\mathbf{s})$ by (7) with $\tilde{\mu}_{\omega,t}(\mathbf{s})$ in (11), sample m covariates from $f_{t+1}(\mathbf{s})$ as $\mathcal{S}_{t+1}$.
Step 4. Get the corresponding $\mathbf{X}_{t+1}$ at $\mathcal{S}_{t+1}$, and return to Step 1.

Determination of the warning threshold L_w is related to the sampling cost. Let v be the unit sampling cost and accordingly $m \times v$ is the cost of regenerating design at each time point. Under the null state, the average cost can be expressed by $\Pr_{\mathbb{H}_0}\{Q_t \in [L_w, L)\} \times mv \times \text{ARL}_0$, where ARL_0 is the nominal ARL under the null state. Hence, one can decide a reliable L_w according to a prespecified average cost under the null state. In general, a smaller L_w tends to yield dynamic sampling more often and consequently detect the change faster. In particular, the procedures with $L_w = 0$ and $L_w = L$ reduce to the POS and FS methods, respectively. The MPOS adaptively selects m most informative samples via altering sampling strategies when the warning signal triggered, which is in a similar spirit to traditional variable sampling schemes with sampling sizes or sampling intervals in statistical process monitoring (Li and Qiu 2014).

It is worth pointing out that this modified procedure can result in cost saving compared to the standard POS when the process changes. For fair comparison, we may consider the POS procedure with a size $m' = \lfloor m\Pr_{\mathbb{H}_0}\{Q_t \in [L_w, L)\}\rfloor$. Accordingly, the POS and its modified one, MPOS, would have the same ARL and average cost under the null state. Denote ARL_{POS} and ARL_{MPOS} as the ARLs (after the change-point τ) of the POS and MPOS under the alternative state, respectively. Then, the average costs of POS and MPOS would be $m'v \times \text{ARL}_{\text{POS}}$ and $\Pr_{\mathbb{H}_1}\{Q_t \in [L_w, L)\} \times mv \times \text{ARL}_{\text{MPOS}}$, respectively. Thus, the expected proportion of cost saving of the MPOS with respect to POS is

$$1 - \frac{\Pr_{\mathbb{H}_1}\{Q_t \in [L_w, L)\} \times \text{ARL}_{\text{MPOS}}}{\Pr_{\mathbb{H}_0}\{Q_t \in [L_w, L)\} \times \text{ARL}_{\text{POS}}}.$$

With a larger subsampling size, ARL_{MPOS} is likely to be much smaller than ARL_{POS}, resulting in considerable savings. Some numerical evidence can be found in Section 4.

4. Numerical Study

In this section, we conduct simulation study and use a real-data example to examine the performance of our proposed POS detection procedure.

4.1. Simulation Results

Consider the covariate space $\mathcal{S} \subset [0,1]^d$ with $m_a = 10{,}000$ possible sampling points generated uniformly. We follow model (1) and set $\mu_0(\mathbf{s}) = 0$ without loss of generality. The number of sampling points is $m = 200$ at each time point. The random error $\mathbf{e}_t$ is a m-dimensional variable with mean zero and covariance matrix $\mathbf{\Lambda}$. Two classes of $\mathbf{\Lambda}$ are investigated: one is the identity matrix; the other one has the components $\rho_{j_1 j_2} = \exp(-||\mathbf{s}_{j_1} - \mathbf{s}_{j_2}||/\lambda)$ and $\lambda = 1/5m^{-3/4}$. We denote these two as "IID" and "COR" cases, respectively. We consider two cases for generating errors: one is that all the errors are from $\mathcal{N}(0,1)$; and the other one is a mixed one, in which the errors are randomly from $\mathcal{N}(0,1)$, normalized chi-squared distribution with three degrees of freedom (χ^2_3) and normalized Student's t-distribution with five degrees of freedom (t_5). Assume the change point is $\tau = 50$ and Ω is the region where changes occur. Three scenarios for $\mu_1(\mathbf{s})$ are considered:

- Scenario (I): $d = 1$, $\mu_1(\mathbf{s}) = \theta \exp(2\mathbf{s})\mathbb{I}_\Omega(\mathbf{s})$ and $\Omega = [0.4, 0.5]$;
- Scenario (II): $d = 1$, $\mu_1(\mathbf{s}) = \theta(\mathbf{s} + 0.5)^2\mathbb{I}_\Omega(\mathbf{s})$ and $\Omega = [0.3, 0.4] \cup [0.6, 0.7]$;
- Scenario (III): $d = 2$, $\mu_1(\mathbf{s}) = \theta(a'\mathbf{s}+2)^2\mathbb{I}_\Omega(\mathbf{s}), a' = (1, -2)$ and $\Omega = \{\mathbf{s} : \|\mathbf{s}\|^2 \in [0.4, 0.5]\}$.

All the simulation results are based on 5000 replications.

For the sake of simplicity and consistency with the literature, the ARL (detection delay) is used to evaluate the performance of monitoring scheme and ARL_0 is fixed as 200. As discussed in Section 3.2, we consider the empirical bandwidth formula $h_{\text{E}} = 1.5m^{-1/(4+d)} \cdot \text{sd}(\mathbf{s})$ in the density estimator (6) and $h = h_{\text{E}}^{2+c'}$ in the test statistics T_{f_t} for some $c' > 0$, where $\text{sd}(\cdot)$ denotes the sample standard deviation. Table 1 reports the ARL values of POS with different bandwidths when the errors follow $\mathcal{N}(0,1)$. We observe that three different values of $c' \in [0.05, 0.2]$ present similar results and their ARLs are not significantly different. Meanwhile, the POS is not affected too much when the window size ω used in $\tilde{\mu}_{\omega,t}(s)$ is selected in $[5, 10]$. Hence, $c' = 0.1$ and $\omega = 5$ are used in the rest of simulations. Moreover, the POS procedure has satisfactory ARL performances under the null state (close to the nominal value 200 when $\theta = 0$) under all the three scenarios. In some cases, the ARL values deviate a little from the nominal level, but the deviations are generally in an acceptable range considering we are using asymptotic approximation here. Our additional results (not reported here) show that the deviation becomes less pronounced as the m increases. These results together with those provided in the supplementary materials for the other error distributions demonstrate that the proposed method for determining the threshold L is effective even with finite-sample sizes.

We next compare the proposed scheme POS with three related approaches. The first approach uses the fixed-sampling (FS) strategy with the uniform distribution at each step. The second one, named as TRAS proposed by Liu, Mei, and Shi (2015), uses the sum of top-r local statistics to monitor and uses the top-m local statistics to sample adaptively. Here, we choose $r = 5$ for the method TRAS following the recommendation made in that article. The third one is Wang et al.'s (2018) SASAM method which monitors a largest local statistic and updates the observations with random sampling in the whole space in conjunction with a directional sampling around the location with the maximum statistic. Because both POS and SASAM considered clustering information, we choose to use the same bandwidth for them to have a fair comparison.

Table 1. Average run lengths of POS with different bandwidths when the errors follow $\mathcal{N}(0,1)$.

ω	Scenario	θ \ c'	IID case 0.05	IID case 0.1	IID case 0.2	COR case 0.05	COR case 0.1	COR case 0.2
5	(I)	0	203 (222)	193 (216)	200 (218)	194 (183)	200 (183)	197 (183)
		0.2	8.9 (5.3)	9.1 (5.5)	10.1 (6.4)	46.7 (42.8)	51.3 (47.5)	53.5 (50.5)
		0.4	2.3 (1.1)	2.3 (1.1)	2.4 (1.1)	5.6 (3.0)	5.9 (3.2)	6.2 (3.5)
	(II)	0	203 (222)	193 (216)	200 (218)	194 (183)	200 (183)	197 (183)
		0.2	64.0 (61.8)	65.0 (62.3)	73.2 (70.0)	132 (133)	142 (145)	136 (134)
		0.4	6.6 (3.7)	6.7 (3.9)	7.4 (4.3)	34.9 (30.8)	38.3 (33.2)	40.2 (34.4)
	(III)	0	191 (180)	203 (189)	199 (187)	205 (190)	209 (192)	207 (190)
		0.2	26.4 (21.6)	31.2 (26.2)	40.9 (36.1)	32.7 (27.4)	38.8 (33.5)	55.5 (49.4)
		0.4	3.7 (1.6)	3.9 (1.7)	4.3 (1.9)	3.9 (1.7)	4.1 (1.8)	4.8 (2.1)
10	(I)	0	207 (192)	214 (197)	204 (188)	199 (188)	204 (189)	201 (185)
		0.2	9.6 (5.7)	10.1 (6.0)	10.7 (6.4)	48.1 (45.8)	50.4 (47.6)	54.7 (51.4)
		0.4	2.4 (1.2)	2.4 (1.2)	2.5 (1.3)	6.7 (3.6)	6.9 (3.6)	7.4 (3.9)
	(II)	0	207 (192)	214 (197)	204 (188)	199 (188)	204 (189)	201 (185)
		0.2	53.9 (49.4)	59.8 (55.9)	61.7 (58.1)	146 (141)	146 (145)	147 (144)
		0.4	7.1 (4.1)	7.5 (4.2)	8.2 (4.6)	33.6 (27.8)	34.4 (28.9)	38.6 (33.5)
	(III)	0	221 (202)	217 (199)	211 (198)	209 (195)	206 (192)	201 (186)
		0.2	27.5 (21.8)	30.5 (24.6)	38.4 (32.9)	30.0 (24.6)	34.3 (28.9)	45.5 (40.1)
		0.4	4.4 (2.1)	4.7 (2.2)	5.2 (2.5)	4.7 (2.2)	4.9 (2.4)	5.7 (2.7)

NOTE: The h is chosen as $h = h_E^{2+c'}$ with $c' = 0.05, 0.1, 0.2$. Numbers in parentheses are standard deviations of the run length.

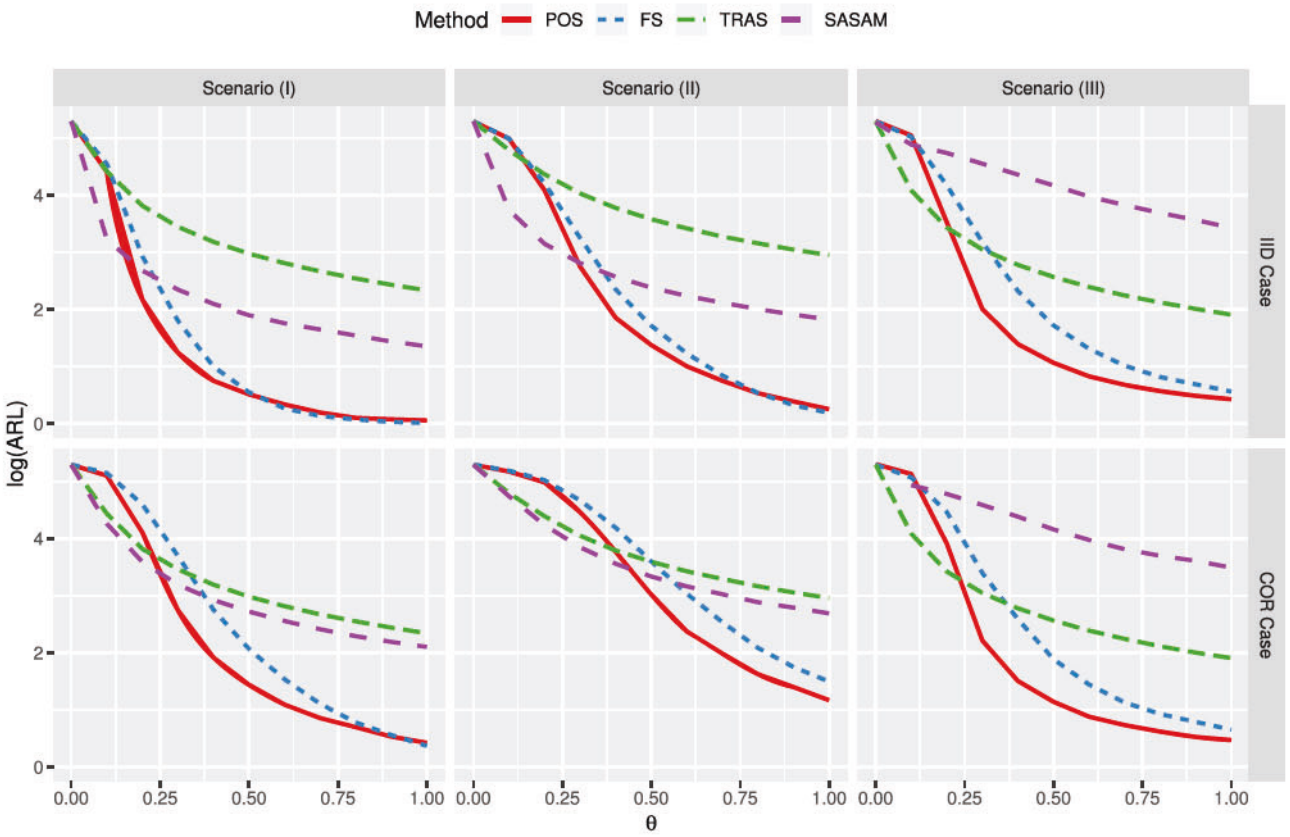

Figure 3. ARL comparisons (in log-scale) under three scenarios when errors randomly come from $\mathcal{N}(0,1)$ or normalized t_5 or χ^2_3. POS: the proposed method (solid red line); FS: the detection scheme based on fixed sampling (dot blue line); TRAS: the top-r-based adaptive sampling scheme by Liu, Mei, and Shi (2015) with $r = 5$ (dash green line); SASAM: the spatial-adaptive sampling and monitoring procedure by Wang et al. (2018) (long-dash purple line).

Figure 3 illustrates the ARL curves (in the log scale) against different signal magnitudes θ for the mixed errors. All plots indicate that the proposed POS is sensitive to the process changes. It outperforms the FS method uniformly, except for the very large θ. This is consistent with the asymptotic comparison. In most cases, our method has smaller ARLs than the TRAS; the

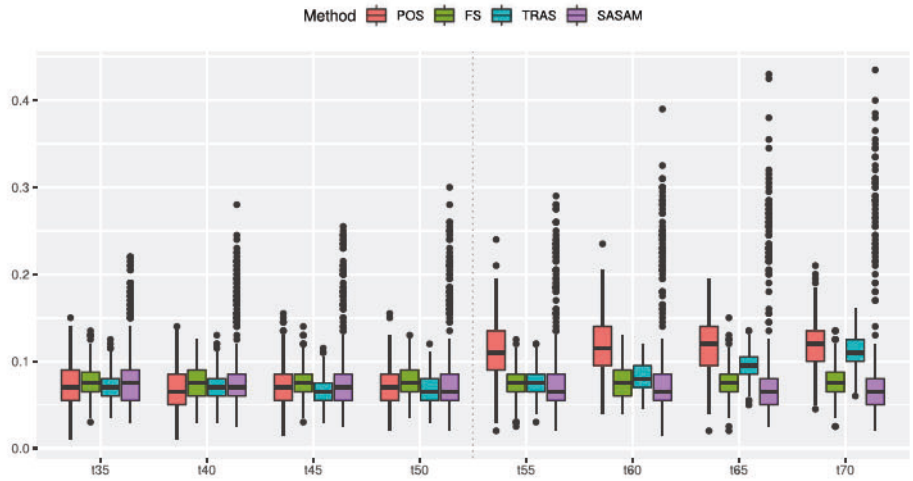

Figure 4. Boxplots of proportions of informative sampling points from the 35th time point to 70th time point. Dotted-line stands for the change point.

ARL curves of TRAS decline slowly as the abnormal signal θ increases. This is not surprising, since the clustering structure of the covariates is completely ignored in the construction of TRAS test statistics. This can be better appreciated by comparing FS and TRAS: even though without dynamic sampling, the FS could be more efficient than TRAS, especially when the abnormal signal is moderate or large, due to the use of the local aggregation statistics D_{f}. When θ is too small, the TRAS performs better than POS. This can be understood that the estimator $\tilde{\mu}_{\omega,t}(\mathbf{s})$ may fail to capture the main feature if the signal-to-noise ratio is not large, and accordingly the power of the test statistic T_{f_t} would be considerably compromised due to local aggregation. In contrast, the TRAS extracts the information on change from each individual marginally, and thus could be more efficient in such cases. Despite utilizing the spatial information to certain degree, the SASAM is not as efficient as POS in terms of change detection when θ is not too small. The "optimal" sampling in POS allows us to take better account of the clustering information than the SASAM which roughly exploits some covariates information around the one with the largest local statistics. Similar results with $\mathcal{N}(0,1)$ distributed errors are provided in the supplementary materials.

To evaluate the sampling performance of all methods, we further compare the proportions of sampling points located in the change regions at each time point. Figure 4 presents the boxplots of the proportions of informative sampling points under Scenario (III) with $\theta = 0.2$ when errors are from $\mathcal{N}(0,1)$. It is evident that POS is able to deliver a more stable performance than the competitors, implying that it tends to select more points in the change regions after the change point $\tau = 50$.

Finally, we consider the extension MPOS in Section 3.3. We consider POS with the subsampling size $m' = 100$ and MPOS with the subsampling size $m = 200$ or 500. That is $\Pr_{\mathbb{H}_0}\{Q_t \in [L_w, L)\}$ equals 0.5, 0.2 for MPOS, respectively. With this setting, POS and MPOS would have the same average cost under the null state. Figure 5 displays the comparison between POS and MPOS in terms of ARLs and cost saving proportions. Clearly, with a larger subsampling size at each time point, the MPOS system can detect the changes faster, and consequently the cost saving of MPOS relative to POS is quite substantial in most cases.

4.2. A Real-Data Application

It becomes increasingly important to monitor the real-time traffic conditions for smart traffic management. In a typical setting, sensors are installed at different segments to detect the traffic volumes and average speed. These data provide useful information in identifying congested areas, shifts in traffic patterns, and abnormal events. However, transmitting the entire city network with high sampling rate for real time analysis is expensive, requiring exhaustive computation and communication resources. Alternatively, using our proposed POS procedure can quickly detect changes in traffic patterns, and provide information for quick traffic control and warnings for road users. In this study, we demonstrate the POS procedure to monitor the real traffic in New York City. The dataset was obtained from New York City Hourly Traffic Data (Donovan et al. 2016). We aim to monitor the average traffic volume, defined as the average number of cars while passing a road junction within a given time interval. For illustration, we consider the traffic monitoring for Manhattan region, as shown in the left panel of Figure 6.

In this example, we focus on $m_a = 2646$ possible sampling locations in the year of 2010 and fix $m = 100$ for each sampling point. For the fixed sampling strategy, we choose 100 typical locations to study the traffic situation of this area, which completely ignores the clustering information of traffic congestion in practice. For the POS procedure, we set $h_{\mathrm{E}} = 0.15$, $h = 0.02$, and $\omega = 5$ in the study. As a convention, we split the dataset into two parts, the data before October as the training data and the data from October to December for monitoring. First, we remove the outliers based on the normalized training data via a rule of thumb threshold, 4. Then we use the training data to estimate the normal mean function $\mu_0(\mathbf{s})$ as well as the parameters $\mu_{\mathrm{f}_t}^{(0)}$ and σ_{f_t} under $\mathbb{H}_0$. Following the GLR procedure and dynamic sampling strategy, the test statistics of the POS and FS procedures are plotted in the right panel of Figure 6. We can observe that the two curves have similar trends, but the POS triggers alarms since the 56th point but the FS method does not produce sustained signals. Retrospective study suggested that the change point could be caused by a blizzard hitting New York. We present the sampling locations of our procedure during the

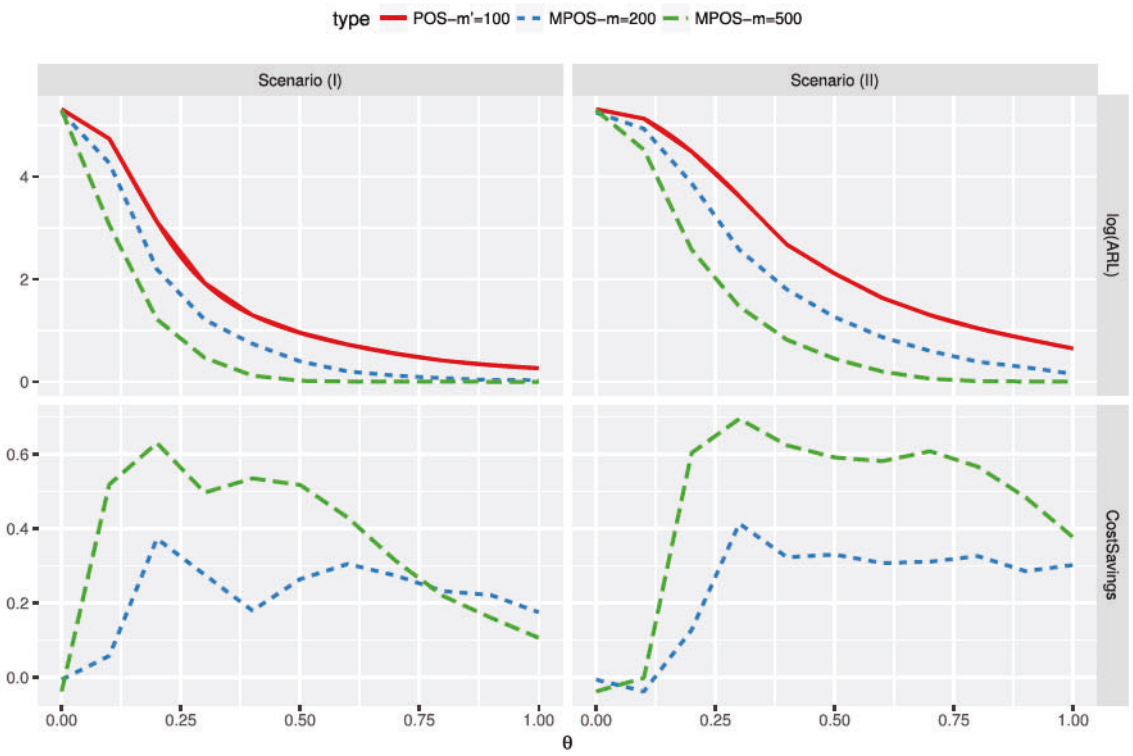

Figure 5. Comparisons between POS and MPOS in terms of ARL and cost saving proportions under Scenarios (I) and (II) when errors are independent from $\mathcal{N}(0, 1)$. The POS with $m' = 100$ (solid red line), MPOS with $m = 200$ (dot blue line) and MPOS with $m = 500$ (dash green line) are considered.

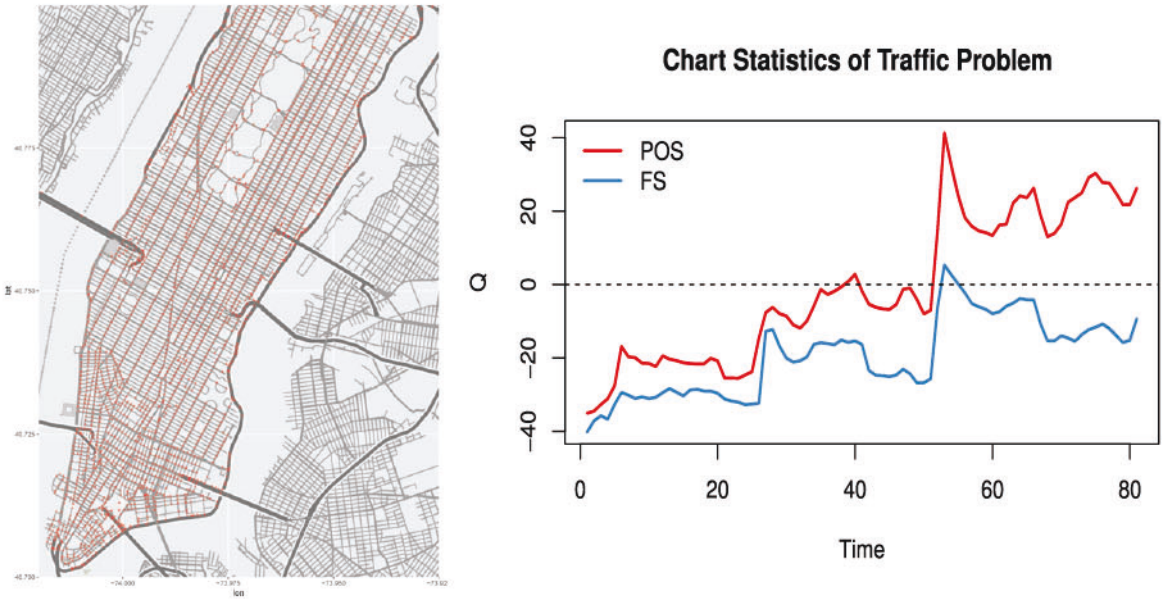

Figure 6. Left panel: The map of Manhattan. The red points stand for the road junction nodes; right panel: the test statistics of the POS and FS procedures. For ease of illustration, the values of test statistics minus the control threshold are plotted so that the "zero" line is the threshold for an alarm decision.

out of control period in Figure 7. It is clear that the locations sampled by POS exhibit certain clustering pattern, and more locations around Broadway region (top-left area) are selected. Compared with the sampling locations from FS strategy, it is not surprising to find that the FS strategy fails to detect the change because only few junction nodes around Broadway were chosen. These results indicate that the POS strategy is appealing in the sense that, even with constraints on resources in practice, it can still enable fast and accurate detection by efficiently using clustering patterns.

5. Concluding Remarks

This article is designed for online detecting the mean function change when limited resources can be used. However, the variances or the correlation structures of datastreams may change over time in many applications. It requires more research to extend our procedure to such problems, in which a new statistic and the corresponding optimal sampling density are required. Moreover, our method employs a kernel-based statistic under the assumption that the difference between the mean functions

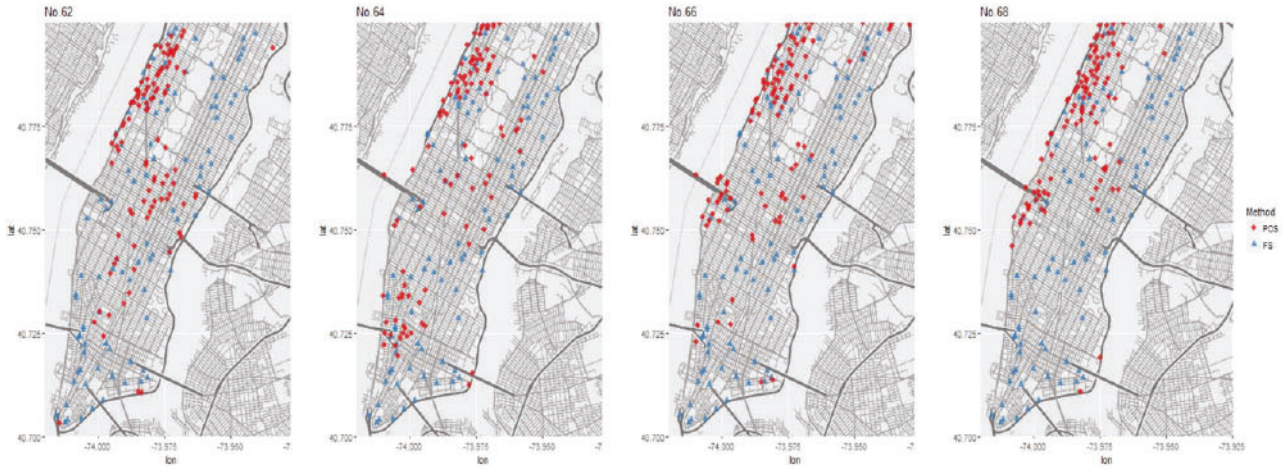

Figure 7. Sampling locations selected by the POS and FS procedures from the 62nd time point to the 68th time point. Red circles stand for sampling locations by POS, while blue triangles stand for the fixed junction nodes.

before and after the change point possess certain smoothness, which may cause difficulties in detecting the scattered or isolated changes. Hence, a robust method, which is able to automatically adapt to both clustered and scattered changes, is of interest.

Another important problem is to identify those locations whose datastream signals have changed, which amounts to conducting a large-scale hypothesis tests for each individual datastream simultaneously and sequentially. It is necessary to study this dynamic testing problem with some error rate control when the clustering or spatial pattern among datastreams exists. In addition, we study the optimality of the sampling procedure POS under the framework of two-stage test. Though our theoretical and numerical results demonstrate that this sampling procedure works well for online detection, there is still a need to further investigate its properties from the sequential viewpoints, such as whether it possess certain optimality in terms of expected delay.

Appendix: Proofs

Before we present the proofs of main theorems, we first state two lemmas whose proof can be found in the supplementary materials. A few well-known theorems we will frequently use are also presented in the supplementary materials. Throughout this discussion, unless otherwise stated, $\sup_{\mathbf{s}}$ will be taken over the entire Ω without the boundary. Let $\boldsymbol{\mu} = \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_m)\}^\top$ and $\mathbf{U}_h = \operatorname{diag}\left(1/\sqrt{\mathbf{f}}\right) \mathbf{K}_h \operatorname{diag}\left(1/\sqrt{\mathbf{f}}\right)$, where $\mathbf{K}_h = \{k_{ij}(h)\}_{m\times m}$ and $\mathbf{f} = \{f(\mathbf{s}_1), \ldots, f(\mathbf{s}_m)\}^\top$. Write

$$D_{\mathrm{f}} = \boldsymbol{\varepsilon}^\top \mathbf{V}_h \boldsymbol{\varepsilon} + 2\boldsymbol{\varepsilon}^\top \boldsymbol{\Lambda}^{1/2} \mathbf{U}_h \boldsymbol{\mu} + \boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\mu}$$
$$:= \boldsymbol{\varepsilon}^\top \mathbf{W}_h \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top \operatorname{diag}(\mathbf{V}_h) \boldsymbol{\varepsilon} + 2\boldsymbol{\varepsilon}^\top \boldsymbol{\Lambda}^{1/2} \mathbf{U}_h \boldsymbol{\mu} + \boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\mu}, \quad \text{(A.1)}$$

where $\mathbf{W}_h = \mathbf{V}_h - \operatorname{diag}(\mathbf{V}_h) = \{w_{ij}(h)\}_{m\times m}$, and $\mathbf{V}_h = \operatorname{diag}\left(1/\sqrt{\mathbf{f}}\right) \boldsymbol{\Lambda}^{1/2} \mathbf{K}_h \boldsymbol{\Lambda}^{1/2} \operatorname{diag}\left(1/\sqrt{\mathbf{f}}\right)$. Denote $v_h^2 = 2\sum_{i=1}^m \sum_{j=1}^m w_{ij}^2(h)$.

Lemma A.1. Suppose the conditions in Theorem 1 hold. We have: (i) $\boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\mu} / (\eta_m h^{-d/2}) \xrightarrow{p} \mu(f, \Gamma)$, where $\mu(f, \Gamma) = \mathbb{E}_{\mathrm{f}}\{I^2(\mathbf{s})\}$; (ii) $v_h^2 / (h^{-d} \eta_m^2) \xrightarrow{p} \kappa^2(f, \Gamma)$.

The next lemma is related to uniform convergence of Nadaraya–Watson estimator under the spatial correlation structure we are considering.

Lemma A.2. Suppose the conditions in Corollary 1 all hold. The Nadaraya-Watson estimator of $\mu(\mathbf{s})$ satisfies

$$\sup_{\mathbf{s}} |\hat{\mu}(\mathbf{s}) - \mu(\mathbf{s})| = O_p\left(h_{\mathrm{E}}^2 \delta_m' + \sqrt{\frac{a_m \eta_m \log m}{m h_{\mathrm{E}}^d}}\right).$$

To prove Theorem 1, we will establish the asymptotic distribution of D_{f} under the null and alternative hypotheses, respectively, given by the next two propositions whose proof will be often referred. Lemma S.3 which is concerned about the asymptotic normality of a quadratic form of iid random variables, will be applied.

Proposition A.1. If Assumptions 1–4 hold, then $\Pr(\frac{D_{\mathrm{f}} - \mu_{\mathrm{f}}^{(0)}}{v_h} > z_\alpha \mid \mathbf{s}_1, \ldots, \mathbf{s}_m) \xrightarrow{p} \alpha$ under $\mathbb{H}_0$.

Proof. By (A.1), $\mathbf{W}_h$ is symmetric with $w_{ii}(h) = 0$ for all i. To apply Lemma S.3 to $\boldsymbol{\varepsilon}^\top \mathbf{W}_h \boldsymbol{\varepsilon}$, we will check the condition, $\|\mathbf{W}_h\|_S^2 / \|\mathbf{W}_h\|_F^2 \xrightarrow{p} 0$. Lemma S.4 and Assumption 1 imply that $\|\operatorname{diag}\left(1/\sqrt{\mathbf{f}}\right)\|_S = O_p(1)$. Note that

$$\|\mathbf{K}_h\|_S^2 = \max_{\|\mathbf{u}\|=1} \|\mathbf{K}_h \mathbf{u}\|^2 \le \left(\max_{1\le i\le m} \sum_{j=1}^m |k_{ij}(h)|\right)^2$$
$$= \max_{1\le i\le m} f^2(\mathbf{s}_i) = O_p(1),$$

where we used the Cauchy inequality and Lemma S.4 (implied by Assumption 4). By the fact that $\|\boldsymbol{\Lambda}^{1/2} \mathbf{K}_h \boldsymbol{\Lambda}^{1/2}\|_S^2 \le \|\boldsymbol{\Lambda}\|_S^2 \|\mathbf{K}_h\|_S^2$, we can claim that

$$\|\mathbf{W}_h\|_S^2 \le \|\mathbf{V}_h\|_S^2 = O_p\left\{\left(\max_{1\le i\le m} \sum_{j=1}^m |\rho_{ij}|\right)^2\right\} = 1 + O_p(m^2 \lambda^{2d})$$
$$= O_p(\eta_m^2), \quad \text{(A.2)}$$

where the second equality is due to Lemma S.5.

Next, we analyze $\|\mathbf{W}_h\|_F^2$. Observe first that

$$\|\mathbf{V}_h\|_F^2 = \sum_{i=1}^m \sum_{j=1}^m \frac{\left\{\sum_{l=1}^m \rho_{il} k_{lj}(h)\right\}^2}{f(\mathbf{s}_i) f(\mathbf{s}_j)} + o_p(1) \sum_{i=1}^m \sum_{j=1}^m \left\{\sum_{l=1}^m \rho_{il} k_{lj}(h)\right\}^2.$$

Given $\mathbf{s}_i$ and $\mathbf{s}_j$, we have as $\lambda/h \to 0$,

$$\begin{aligned}\sum_{l\neq i} \rho_{il} k_{lj}(h) &\approx \frac{1}{h^d} \int \rho\left(\frac{\mathbf{s}_i - \mathbf{s}}{\lambda}\right) K\left(\frac{\mathbf{s}_j - \mathbf{s}}{h}\right) f(\mathbf{s}) d\mathbf{s} \\ &= \frac{\lambda^d}{h^d} \int K\left(\frac{\mathbf{s}_i - \mathbf{s}_j + \lambda \mathbf{u}}{h}\right) \rho(\|\mathbf{u}\|) f(\mathbf{s}_i - \lambda \mathbf{u}) d\mathbf{u} \\ &\approx \frac{\lambda^d}{h^d} K\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h}\right) f(\mathbf{s}_i) \int \rho(u) du.\end{aligned}$$

Accordingly, we have

$$\left\{\sum_{l=1}^m \rho_{il} k_{lj}(h)\right\}^2 \approx \frac{K_h^2(\mathbf{s}_i - \mathbf{s}_j)}{(m-1)^2} \left\{1 + m^2 \lambda^{2d} \left(f(\mathbf{s}_i) \int \rho(u) du\right)^2\right\}. \tag{A.3}$$

By $h^{-d} K^2\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h}\right) \leq C h^{-d}$, and the Bernstein inequality for U-statistics in Lemma S.2,

$$\begin{aligned}&\Pr\left(\left|\frac{1}{m(m-1)} \sum\sum_{1\leq i\neq j\leq m} h^{-d} \right.\right. \\ &\quad \left.\left. \times \left[K^2\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h}\right) - \mathbb{E}\left\{K^2\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h}\right)\right\}\right]\right| > t\right) \\ &\leq 2 \exp\left(-\frac{1}{2C} \frac{(m-1)h^d t^2}{1 + t/3}\right) \to 0,\end{aligned} \tag{A.4}$$

for any $t > 0$ since $mh^d \to \infty$. Because

$$\begin{aligned}h^{-d} \mathbb{E}\left\{K^2\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h}\right)\right\} &= \int\int K^2(\mathbf{u}) f(\mathbf{s}) f(\mathbf{s} + h\mathbf{u}) d\mathbf{s} d\mathbf{u} \\ &\to \int K^2(\mathbf{u}) d\mathbf{u} \int f^2(\mathbf{s}) d\mathbf{s}, \quad i \neq j,\end{aligned}$$

when $h \to 0$ by the Lebesgue dominated convergence theorem, the limit of $m^{-2} h^{-d} \sum\sum_{i,j}^m K^2\left(\frac{\mathbf{s}_i - \mathbf{s}_j}{h}\right)$ is given by $\int K^2(\mathbf{u}) d\mathbf{u} \int f^2(\mathbf{s}) d\mathbf{s}$. Consequently, we can conclude that

$$\|\mathbf{V}_h\|_F^2 = O_p\left\{h^{-d}\left(1 + m^2 \lambda^{2d}\right)\right\} = O_p(h^{-d} \eta_m^2). \tag{A.5}$$

Using (A.3), we have $\|\mathrm{diag}(\mathbf{V}_h)\|_F^2 = o_p(\|\mathbf{V}_h\|_F^2)$. Combining this with (A.2) and (A.5), the condition $\|\mathbf{W}_h\|_S^2 / \|\mathbf{W}_h\|_F^2 \xrightarrow{p} 0$ holds.

As a consequence, we have

$$\Pr(\boldsymbol{\varepsilon}^\top \mathbf{W}_h \boldsymbol{\varepsilon} / \nu_h > z_\alpha \mid \mathbf{s}_1, \ldots, \mathbf{s}_m) \xrightarrow{p} \alpha$$

by Lemma S.3. Finally, it can be checked that $\boldsymbol{\varepsilon}^\top \mathrm{diag}(\mathbf{V}_h) \boldsymbol{\varepsilon} - \mu_{\mathrm{f}}^{(0)} = o_p(\nu_h)$, from which the assertion of the proposition holds. □

Proposition A.2. If the conditions in Theorem 1 all hold, then

$$\Pr\left(\frac{D_{\mathrm{f}} - \mu_{\mathrm{f}}^{(0)}}{\nu_h} > z_\alpha \mid \mathbf{s}_1, \ldots, \mathbf{s}_m\right) - \Phi\left(\frac{\mu(f,\Gamma)}{\kappa(f,\Gamma)} - z_\alpha\right) \xrightarrow{p} 0.$$

Proof. Under the alternative, $D_{\mathrm{f}} = \boldsymbol{\varepsilon}^\top \mathbf{V}_h \boldsymbol{\varepsilon} + 2\boldsymbol{\varepsilon}^\top \boldsymbol{\Lambda}^{1/2} \mathbf{U}_h \boldsymbol{\mu} + \boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\mu}$. By Proposition A.1, $\Pr\{(\boldsymbol{\varepsilon}^\top \mathbf{V}_h - \mu_{\mathrm{f}}^{(0)}) \boldsymbol{\varepsilon} / \nu_h > z_\alpha \mid \mathbf{s}_1, \ldots, \mathbf{s}_m\} \xrightarrow{p} \alpha$. From Lemma A.1, $\boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\mu} / \nu_h \xrightarrow{p} \mu(f,\Gamma)/\kappa(f,\Gamma)$. Note that $\mathbb{E}\left\{\boldsymbol{\varepsilon}^\top \boldsymbol{\Lambda}^{1/2} \mathbf{U}_h \boldsymbol{\mu} \mid (\mathbf{s}_1, \ldots, \mathbf{s}_m)\right\} = 0$, and

$$\begin{aligned}\mathrm{var}\left\{\boldsymbol{\varepsilon}^\top \boldsymbol{\Lambda}^{1/2} \mathbf{U}_h \boldsymbol{\mu} \mid (\mathbf{s}_1, \ldots, \mathbf{s}_m)\right\} &= \boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\Lambda} \mathbf{U}_h \boldsymbol{\mu} \\ &\leq \|\boldsymbol{\mu}^\top \mathbf{U}_h \boldsymbol{\mu}\| \|\boldsymbol{\Lambda}\|_S \|\mathbf{U}_h\|_S \\ &= O(\eta_m h^{-d/2}) O_p(\eta_m) O_p(1).\end{aligned}$$

The last equality is based on $\|\mathbf{U}_h\|_S \leq \|\mathbf{K}_h\|_S \|\mathrm{diag}\left(1/\sqrt{\mathbf{f}}\right)\|_S$. Thus, $\boldsymbol{\varepsilon}^\top \boldsymbol{\Lambda}^{1/2} \mathbf{U}_h \boldsymbol{\mu} / \nu_h = O_p(h^{d/4})$. Simple algebra yields the conclusion. □

Proof of Theorem 1. Under the local alternative hypothesis, Proposition A.2 revealed that the power of the test based on D_{f} depends only on $\mu(f,\Gamma)/\kappa(f,\Gamma)$.

If $m\lambda^d \to 0$, we can see that $\kappa(f,\Gamma)$ is a constant not depending on $f(\mathbf{s})$ under a given kernel function $K(\cdot)$. By Cauchy–Schwarz inequality, we can find that $\mathbb{E}\{l^2(\mathbf{s})\} \leq \sqrt{\int l^4(\mathbf{s}) d\mathbf{s} \int f^2(\mathbf{s}) d\mathbf{s}}$, and the equality holds if and only if $l^2(\mathbf{s}) = c f(\mathbf{s})$, where c is some constant. Thus, if the sampling distribution of $\mathbf{s}$ is a linear function of $l^2(\mathbf{s})$, that is, $f(\mathbf{s}) = l^2(\mathbf{s}) / \int l^2(\mathbf{s}) d\mathbf{s}$, the power can be maximized. If $m\lambda^d \to \infty$, we have

$$\mu(f,\Gamma)/\kappa(f,\Gamma) \propto \frac{\int l^2(\mathbf{s}) f(\mathbf{s}) d\mathbf{s}}{\sqrt{\int f^2(\mathbf{s}) d\mathbf{s}}}.$$

Again, the locally best power is achieved by choosing $f(\mathbf{s}) = l^2(\mathbf{s}) / \int l^2(\mathbf{s}) d\mathbf{s}$. □

Proof of Theorem 2. By the definition of the two-stage test, the power conditioned on $\{\mathbf{s}_{kj}\}_{j=1}^m, k = 1, 2$ is given by

$$\begin{aligned}\beta_{\mathrm{f}_1,\mathrm{f}_2} = \Pr\left(T_{\mathrm{f}_1} > z_{\alpha_1} \mid \{\mathbf{s}_{1j}\}_{j=1}^m\right) + \Pr\left(T_{\mathrm{f}_1} < z_{\alpha_1} \mid \{\mathbf{s}_{1j}\}_{j=1}^m\right) \\ \times \Pr\left(T_{\mathrm{f}_2} > z_{\alpha_2} \mid \{\mathbf{s}_{kj}\}_{j=1}^m, k = 1, 2\right),\end{aligned}$$

where we use the fact that T_{f_2} is independent of T_{f_1} conditioned on $\{\mathbf{s}_{kj}\}_{j=1}^m, k = 1, 2$.

By Proposition A.2, we have

$$\Pr\left(T_{\mathrm{f}_1} > z_{\alpha_1} \mid \{\mathbf{s}_{1j}\}_{j=1}^m\right) - \Phi(-z_1) \xrightarrow{p} 0.$$

By the proof of Theorem 1, we know that

$$\Pr\left(T_{\mathrm{f}_2} > z_{\alpha_2} \mid \{\mathbf{s}_{kj}\}_{j=1}^m, k = 1, 2\right) - \Phi(-\tilde{z}_2) \xrightarrow{p} 0,$$

$\tilde{z}_2 = z_{\alpha_2} - \mathbb{E}_{\mathrm{f}_2}\{l^2(\mathbf{s})\} / \kappa(f_2, \Omega)$. Using the uniform convergence given in Lemma A.2, we have

$$\mathbb{E}_{\mathrm{f}_2}\{l^2(\mathbf{s})\} - \int l^4(\mathbf{s}) d\mathbf{s} / \int l^2(\mathbf{s}) d\mathbf{s} \xrightarrow{p} 0,$$

provided that the order of signal strength is larger than that of the maximum noise level, almost everywhere, say

$$(mh^{d/2}/\eta_m)^{-1/2} / \sqrt{\frac{\eta_m \log m}{mh_{\mathrm{E}}^d}} \to \infty.$$

The condition $(h_{\mathrm{E}}^2/h)^{d/2} / (\log m)^c \to \infty$ given in the theorem implies this theorem holds. □

Proof of Corollary 1. For simplicity, we assume that the number of signal region is one, say Ω_m, and the proof for the case with more than one region is similar. By Lemma A.2 and the condition $(h_E^2/h)^{d/2}/\{a_m^{1/2}(\log m)^c\}\to\infty$, we can see that the Nadaraya–Watson estimator of $\mu(\mathbf{s})$ is still a uniformly consistent one except for the boundary, in the sense that

$$\sup_{\mathbf{s}\in(\Gamma\setminus\Omega_m)\setminus\mathbb{B}_h}|\hat{\mu}(\mathbf{s})| = O_p\left(\sqrt{\frac{a_m\eta_m\log m}{mh_E^d}}\right),$$
$$\sup_{\mathbf{s}\in\Omega_m\setminus\mathbb{B}_h}|\hat{\mu}(\mathbf{s})-\mu(\mathbf{s})| = O_p\left(h_E^2\delta_m' + \sqrt{\frac{a_m\eta_m\log m}{mh_E^d}}\right), \quad \text{(A.6)}$$

where $\mathbb{B}_h$ denotes a d-dimensional ball with radius h that is around the boundary of Ω_m. Thus,

$$\int_{\Omega_m\cup\mathbb{B}_h} f_2(\mathbf{s})d\mathbf{s} = 1 + O_p\left(\sqrt{\frac{a_m\eta_m\log m}{mh_E^d}}\right),$$
$$\Pr\left\{f_2(\mathbf{s}_{2j}) < \zeta_m^{1/2}, \forall \mathbf{s}_{2j}\notin\Omega_m\cup\mathbb{B}_h\right\}\to 1,$$

where ζ_m is defined in (7).

Accordingly, using the same procedure in the proof of Proposition A.2, we can show that

$$\Pr(T_{f_2} > z_{\alpha_2} \mid \{\mathbf{s}_{kj}\}_{j=1}^m, k=1,2) - \Phi(\mu(f_2,\Omega_m)/\kappa(f_2,\Omega_m) - z_{\alpha_2}) \xrightarrow{p} 0,$$

where by (A.6)

$$\kappa^2(f_2,\Omega_m) \approx 2h^{-d}|\Omega_m|\int K^2(\mathbf{u})d\mathbf{u}, \text{ if } m\lambda^d\to 0,$$
$$\kappa^2(f_2,\Omega_m) \approx \frac{2m^2\lambda^{2d}(\int\rho(u)du)^2}{h^d(\int_{\Omega_m} l^2(\mathbf{s})d\mathbf{s})^2}\int K^2(\mathbf{u})d\mathbf{u}\int_{\Omega_m} l^4(\mathbf{s})d\mathbf{s}, \text{ if } m\lambda^d\to\infty,$$

and

$$\begin{aligned}\mu(f_2,\Omega_m) &= \int_{\Omega_m\cup\mathbb{B}_h} l^2(\mathbf{s})f_2(\mathbf{s})d\mathbf{s}\\ &\approx \int_{\Omega_m\setminus\mathbb{B}_h} l^2(\mathbf{s})g(\mathbf{s})d\mathbf{s}\\ &\approx \int_{\Omega_m\setminus\mathbb{B}_h} l^4(\mathbf{s})d\mathbf{s}/\int_{\Omega_m\setminus\mathbb{B}_h} l^2(\mathbf{s})d\mathbf{s}\\ &\approx \int_{\Omega_m} l^4(\mathbf{s})d\mathbf{s}/\int_{\Omega_m} l^2(\mathbf{s})d\mathbf{s}.\end{aligned}$$

Finally, by the assumption that $a_m\to 0$, we can see that the asymptotic power of the two-stage test under $m\lambda^d\to 0$ is 1. The proof can be completed by using the condition $\inf_{\mathbf{s}\in\Omega_m} l(\mathbf{s}) \geq \underline{l} > 0$ for the case that $m\lambda^d\to\infty$. □

Proof of Theorem 3.

(i) First of all, by Proposition A.1, we have

$$\sup_{x\in\mathbb{R}}\left|\Pr(\tilde{T}_{f_i} < x \mid \mathcal{S}_i) - \Phi(x)\right| \leq O_p(h^{d/8}),\ 1\leq i\leq t.$$

Note that $\tilde{T}_{f_1},\ldots\tilde{T}_{f_t}$ are independent. Thus, $\tilde{T}_{f_i}$'s are asymptotically equivalent to t iid $\mathcal{N}(0,1)$ random variables, provided that $th^{d/8}\to 0$.

On the other hand,

$$\begin{aligned}&\Pr\left(T_{f_1} < x,\ldots,T_{f_{\omega+1}} < x \mid \{\mathcal{S}_i\}_{i=1}^{\omega+1}\right)\\ &= \Pr\left(T_{f_1} < x,\ldots,T_{f_\omega} < x \mid \{\mathcal{S}_i\}_{i=1}^{\omega}\right)\Pr\left(T_{f_{\omega+1}} < x \mid \{\mathcal{S}_i\}_{i=1}^{\omega+1}\right),\end{aligned} \quad \text{(A.7)}$$

where we use the fact that $T_{f_{\omega+1}}$ is independent of $\{T_{f_i}\}_{i=1}^{\omega}$ conditioned on $\{\mathcal{S}_i\}_{i=1}^{\omega}$. By Lemma A.2, we know that

$$\max_{2\leq i\leq t}\sup_{\mathbf{s}}|f_i(\mathbf{s})-1| = O_p\left(\sqrt{\frac{\eta_m\log m}{mh_E^d}}\right).$$

Thus, again we have

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(T_{f_{\omega+1}} < x \mid \{\mathcal{S}_i\}_{i=1}^{\omega+1}\right) - \Phi(x)\right| \leq O_p(h^{d/8}).$$

By iteratively using (A.7), we see that the joint distribution of T_{f_i}'s are also asymptotically equivalent to that of t iid $\mathcal{N}(0,1)$ random variables. Notice that $\max_{1\leq t'\leq t}\tilde{Q}_{t'}$ and $\max_{1\leq t'\leq t}Q_{t'}$ are continuous functions of $\{\tilde{T}_{f_i}\}_{i=1}^t$ and $\{T_{f_i}\}_{i=1}^t$, respectively. Hence, the assertion holds immediately by the continuous mapping theorem.

(ii) Consider $Q_t^* = \max_{0\leq k<t}\frac{1}{\sqrt{t-k}}\sum_{i=k+1}^t Z_i$, where

$$Z_i \overset{\text{iid}}{\sim} \begin{cases}\mathcal{N}(0,1), & \text{for } i=1,\ldots,\tau,\\ \mathcal{N}(\theta_i,1), & \text{for } t=\tau+1,\ldots,t,\end{cases} \quad \text{(A.8)}$$

and $\theta_i > 0$. By induction, it can be seen that the distribution function of Q_t^* is nonincreasing in θ_i, $i=\tau+1,\ldots,t$. Similar to the proof of (i), we can show that the joint distributions of $\{T_{f_i}\}_{i=1}^t$ and $\{\tilde{T}_{f_i}\}_{i=1}^t$ are asymptotically equivalent to those of t iid $\mathcal{N}(\theta_i,1)$ and $\mathcal{N}(\tilde{\theta}_i,1)$ random variables, respectively. Thus, it suffices to show that $\theta_i\geq\tilde{\theta}_i$ for $i>\tau$.

By Corollary 1, we know that $\theta_{\tau+1} = \tilde{\theta}_{\tau+1}$. By Lemma A.2, we can see that the $\mu_i(\mathbf{s}), i\geq\tau+2$ has the following properties

$$\sup_{\mathbf{s}\in(\Gamma\setminus\Omega_m)\setminus\mathbb{B}_h}|\hat{\mu}_i(\mathbf{s})| = O_p\left(\sqrt{\frac{a_m\eta_m\log m}{mh_E^d}}\right),$$
$$\sup_{\mathbf{s}\in\Omega_m\setminus\mathbb{B}_h}\left|\hat{\mu}_i(\mathbf{s}) - \frac{i-\tau-1}{\omega}\mu(\mathbf{s})\right| = O_p\left(h_E^2\delta_m' + \sqrt{\frac{a_m\eta_m\log m}{mh_E^d}}\right),\ i=\tau+2,\ldots,t.$$

Accordingly, by the same procedure used in the proof of Theorem 2 and Corollary 1, we can show that $\theta_i\geq\tilde{\theta}_i$ for $i=\tau+2,\ldots,t$. The theorem is proved.

□

Proof of Corollary 2. Consider a neighborhood of τ, $[\tau-\Delta,\tau+\Delta]$ for some $\Delta>0$. Denote $R_k = \frac{1}{\sqrt{t_s-k}}\sum_{i=k+1}^{t_s}T_{f_i}$. It suffices to show that $\Pr(R_\tau - R_{\tau\pm\Delta} > 0)\to 1$ for sufficiently large Δ, as $m\to\infty$ and $t_s\to\infty$. We first consider $R_{\tau+\Delta}$. By Corollary 1, we know that

$$T_{f_i} = Z_i + \theta + o_p(1),\ i=\tau+2,\ldots,t_s,$$

under the alternative state, where $\mathbb{E}(Z_i)\approx 0$, $\text{var}(Z_i)\approx 1$ and $\theta>0$. Furthermore, by the definition of T_{f_i}, we have

$$\mathbb{E}(R_\tau) = \mathbb{E}\{\mathbb{E}(R_\tau \mid \{\mathcal{S}_j\}_{j=\tau+1}^{t_s})\} \geq \sqrt{t_s-\tau-1}\theta(1+o(1))$$
$$\text{var}(R_\tau) = \mathbb{E}\{\text{var}(R_\tau \mid \{\mathcal{S}_j\}_{j=\tau+1}^{t_s})\} + o(1) = \frac{1}{t_s-\tau}\sum_{i=\tau+1}^{t_s}\text{var}(T_{f_i}) + o(1) = 1+o(1).$$

Thus, we have

$$\begin{aligned}&\Pr(R_\tau - R_{\tau+\Delta} > 0)\\ \geq&\Pr\left\{\left(\sqrt{t_s-\tau-1}-\sqrt{t_s-\tau-\Delta}\right)\theta > O_p(\sqrt{\text{var}\,(R_\tau)})\right.\\ &\left.+O_p(\sqrt{\text{var}\,(R_{\tau+\Delta})})\right\}\to 1,\end{aligned}$$

as $m\to\infty$ and Δ is sufficiently large. Similarly, we have $\Pr(R_\tau - R_{\tau-\Delta} > 0)\to 1$ as well, from which we can complete the proof. □

Supplementary Materials

The supplementary materials contain some lemmas and additional simulation results.

Acknowledgments

The authors thank the editor, associate editor, and three anonymous referees for their many helpful comments that have resulted in significant improvements in the article. This work was completed when Ren was a postdoctoral researcher at Pennsylvania State University.

Funding

Ren and Li's research were supported by NSF grants DMS 1820702, DMS 1953196 and DMS 2015539. Zou was supported by NNSF of China grants 11931001, 11690015, 11925106, 11771332 and NSF of Tianjin 18JCJQJC46000. Chen was partially supported by Singapore AcRF Tier 1 grant R-266-000-123-114.

References

Chatterjee, S., and Qiu, P. (2009), "Distribution-Free Cumulative Sum Control Charts Using Bootstrap-Based Control Limits," *The Annals of Applied Statistics*, 3, 349–369. [800]

Ciampalini, A., Raspini, F., Bianchini, S., Frodella, W., Bardi, F., Lagomarsino, D., Di Traglia, F., Moretti, S., Proietti, C., and Pagliara, P. (2015), "Remote Sensing as Tool for Development of Landslide Databases: The Case of the Messina Province (Italy) Geodatabase," *Geomorphology*, 249, 103–118. [794]

Donovan, B., Mori, A., Agrawal, N., Meng, Y., Lee, J., and Work, D. D. (2016), "New York City Hourly Traffic Estimates (2010–2013)," University of Illinois at Urbana-Champaign, DOI: 10.13012/B2IDB-4900670_V1. [803]

Enikeeva, F., and Harchaoui, Z. (2019), "High-Dimensional Change-Point Detection With Sparse Alternatives," *The Annals of Statistics*, 47, 2051–2079. [795]

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society*, Series C, 41, 337–348. [798]

Guerre, E., and Lavergne, P. (2005), "Data-Driven Rate-Optimal Specification Testing in Regression Models," *The Annals of Statistics*, 33, 840–870. [796]

Han, D., and Tsung, F. (2006), "A Reference-Free Cuscore Chart for Dynamic Mean Change Detection and a Unified Framework for Charting Performance Comparison," *Journal of the American Statistical Association*, 101, 368–386. [799,800]

Hart, J. (2013), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer. [800]

Hawkins, D. M., and Zamba, K. (2005), "Statistical Process Control for Shifts in Mean or Variance Using a Changepoint Formulation," *Technometrics*, 47, 164–173. [800]

Li, J. (2019), "A Two-Stage Online Monitoring Procedure for High-Dimensional Data Streams," *Journal of Quality Technology*, 51, 392–406. [795]

Li, Z., and Qiu, P. (2014), "Statistical Process Control Using a Dynamic Sampling Scheme," *Technometrics*, 56, 325–335. [801]

Liu, J., Liu, F., and Ansari, N. (2014), "Monitoring and Analyzing Big Traffic Data of a Large-Scale Cellular Network With Hadoop," *IEEE Network*, 28, 32–39. [794]

Liu, K., Mei, Y., and Shi, J. (2015), "An Adaptive Sampling Strategy for Online High-Dimensional Process Monitoring," *Technometrics*, 57, 305–319. [795,796,801,802]

Mei, Y. (2010), "Efficient Scalable Schemes for Monitoring a Large Number of Data Streams," *Biometrika*, 97, 419–433. [795]

Neill, D. B. (2012), "Fast Subset Scan for Spatial Pattern Detection," *Journal of the Royal Statistical Society*, Series B, 74, 337–360. [794]

Siegmund, D., and Venkatraman, E. (1995), "Using the Generalized Likelihood Ratio Statistic for Sequential Detection of a Change-Point," *The Annals of Statistics*, 23, 255–271. [795]

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Kim, H. (2006), "A Novel Approach to Detection of Intrusions in Computer Networks via Adaptive Sequential and Batch-Sequential Change-Point Detection Methods," *IEEE Transactions on Signal Processing*, 54, 3372–3382. [795]

Veeravalli, V. V. (2001), "Decentralized Quickest Change Detection," *IEEE Transactions on Information Theory*, 47, 1657–1665. [795]

Wang, A., Xian, X., Tsung, F., and Liu, K. (2018), "A Spatial-Adaptive Sampling Procedure for Online Monitoring of Big Data Streams," *Journal of Quality Technology*, 50, 329–343. [795,801,802]

Wang, T., and Samworth, R. J. (2018), "High Dimensional Change Point Estimation via Sparse Projection," *Journal of the Royal Statistical Society*, Series B, 80, 57–83. [795]

Xian, X., Wang, A., and Liu, K. (2018), "A Nonparametric Adaptive Sampling Strategy for Online Monitoring of Big Data Streams," *Technometrics*, 60, 14–25. [795]

Xian, X., Zhang, C., Bonk, S., and Liu, K. (2019), "Online Monitoring of Big Data Streams: A Rank-Based Sampling Algorithm by Data Augmentation," *Journal of Quality Technology*, to appear. DOI: 10.1080/00224065.2019.1681924 [795]

Xie, Y., and Siegmund, D. (2013), "Sequential Multi-Sensor Change-Point Detection," *The Annals of Statistics*, 41, 670–692. [795]

Yan, H., Paynabar, K., and Shi, J. (2018), "Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition," *Technometrics*, 60, 181–197. [795]

Zheng, J. X. (1996), "A Consistent Test of Functional Form via Nonparametric Estimation Techniques," *Journal of Econometrics*, 75, 263–289. [796]

Zou, C., and Qiu, P. (2009), "Multivariate Statistical Process Control Using LASSO," *Journal of the American Statistical Association*, 104, 1598–1596. [795]

Zou, C., Wang, Z., Zi, X., and Jiang, W. (2015), "An Efficient Online Monitoring Method for High-Dimensional Data Streams," *Technometrics*, 57, 374–387. [795,796]