Statistica Sinica Preprint No: SS-2023-0195	
Title	Dynamic Statistical Learning in Massive Datastreams
Manuscript ID	SS-2023-0195
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0195
Complete List of Authors	Jingshen Wang, Lilun Du, Changliang Zou and Zhenke Wu
Corresponding Authors	Lilun Du
E-mails	lilundu@cityu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

Statistica Sinica

Dynamic Statistical Learning in Massive Datastreams

Jingshen Wang, Lilun Du[1], Changliang Zou, and Zhenke Wu

University of California, Berkeley, City University of Hong Kong, Nankai University, and University of Michigan

Abstract: Technological advances have necessitated statistical methodologies for analyzing large-scale datastreams comprising multiple indefinitely time series. This manuscript proposes a dynamic tracking and screening (DTS) framework for online learning and model updating. Utilizing the sequential nature of datastreams, a robust estimation approach is developed under a linear varying coefficient model framework. This accommodates unequally-spaced design points and updates coefficient estimates without storing historical data. A data-driven choice of an optimal smoothing parameter is proposed, alongside a new multiple testing procedure for the streaming environment. Statistical guarantees of the procedure are provided, along with simulation studies on its finite-sample performance. The methods are demonstrated through a mobile health example estimating when subjects' sleep and physical activities unusually influence their mood.

Key words and phrases: Consistency, Kernel smoothing, Multiple testing, Varying coefficient.

[1]Corresponding author

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

1. Introduction

1.1 Background and motivation

Highly developed information and sensor technologies constantly generate and store massive longitudinal data sets that become available sequentially at a high frequency. Ranging from telecommunications (Black and Hickey, 2003), environmental monitoring (Guerriero et al., 2009), retail banking (Tsung et al., 2007), health care (Spiegelhalter et al., 2012), and network monitoring (Vaughan et al., 2013), such a type of data collection is pervasive and is referred to as streaming data throughout this manuscript. Other than the high-frequency feature, as massive datastreams are often collected from distinct classes of subjects often in highly dynamic real-life environments, it is commonly believed that they may contain a growing number of irregular patterns (Gama, 2010).

In this context, a statistical methodology that is relevant to streaming data analysis often pertains to algorithms that enable us to (1) dynamically revise statistical models and update the statistical inferential results by incorporating local dynamic changes, (2) efficiently store summary statistics from past history without the need to store an ever-increasing data history (Aggarwal, 2007), and (3) identify individual datastreams whose behavioral patterns deviate significantly from that of most individuals. In this manuscript, we propose a dynamic statistical learning procedure–so called *dynamic tracking and screening* (DTS)–which is able to temporally adapt to time-varying structures, to incor-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

porate time-varying covariates and to identify irregular datastreams as soon and accurately as possible.

Our motivating example comes from the Intern Health Study (IHS)–an ongoing multi-site cohort study enrolled more than 3,000 medical interns–which aims to assess behavioral phenotypes that precipitate stress episodes and mood changes during the first year of residency training (Kalmbach et al., 2018; Kious et al., 2019). Here, the datastreams represent daily ecological momentary assessments (EMA) via a mobile App and temporal behavioral patterns collected from wristbands that are preassigned to the medical interns. As a medical internship–the first phase of professional medical training in the United States–is a stressful period in the career of physicians, the residents are faced with difficult decisions, long work hours and sleep deprivation. A timely identification of individuals with sleep- or activity-sensitive emotional states informs the policy maker right interventions, e.g, the mobile App can send out sleep or activity message in the hope of promoting healthy outcomes. In this context, as new data batch arrives every day (such as hours of sleep, daily step counts and daily mood scores), our DTS framework quickly revises the statistical model and update the underlying parameter estimation, hence it enables an efficient detection of medical interns potentially at high stress level in a timely manner.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

1.2 Model setup and our contribution

Driven by the aforementioned examples, we formulate the dynamic tracking and screening problem as follows. Suppose that we have p datastreams for the units indexed by $j = 1, \ldots, p$. Suppose the study begins at the time point t_1, and we are at the current time point t_m. At each point t_i, $i = 1, \ldots, m$, we observe the response y_{ij} and the covariates $\mathbf{X}_{ij} \in \mathbb{R}^d$. We consider the linear varying coefficient model in the form of

$$y_{ij} = \begin{cases} \mathbf{X}_{ij}^\top \boldsymbol{\beta}(t_i) + \sigma(t_i)\varepsilon_{ij}, & \text{for} \quad t_i \in (0, \tau_j], \\ \mathbf{X}_{ij}^\top \{\boldsymbol{\beta}(t_i) + \boldsymbol{\delta}_j(t_i)\} + \sigma(t_i)\varepsilon_{ij}, & \text{for} \quad t_i > \tau_j, \end{cases} \tag{1.1}$$

for $j = 1, \ldots, p$, $i = 1, 2, \ldots$, where ε_{ij} is the random noise that satisfies $\mathbb{E}(\varepsilon_{ij} \mid \mathbf{X}_{ij}) = 0$ and $\text{var}(\varepsilon_{ij} \mid \mathbf{X}_{ij}) = 1$ for theoretical treatments, $\sigma^2(\cdot)$ is the variance function, and τ_j is an unknown change-point in the jth stream. The coefficient $\boldsymbol{\beta}(\cdot)$, the drift $\boldsymbol{\delta}_j(\cdot)$, and the variance function $\sigma^2(\cdot)$ are assumed to be smooth functions. In particular, this implies that $\boldsymbol{\delta}_j(t) = 0$ for $t \leq \tau_j$. In model (1.1), different streams are assumed to share the same coefficient function $\boldsymbol{\beta}(\cdot)$, while the change-points τ_j's and the drift functions $\boldsymbol{\delta}_j(\cdot)$'s are allowed to vary among different streams.

To better understand our model in (1.1), take IHS for example, the response y_{ij} is the mood score (self-reported through the mobile App), and the covariates $\mathbf{X}_{ij}$ include traits of individual which may affect the mood score, such as hours of sleep, step counts, or average resting heart rate. As the relationship between the mood score and the covariates is potentially affected by factors such as

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

temperature or daylight hours that change with time, it is more suitable to treat the regression coefficients as dynamic functions of the time. In addition, the time-varying coefficient $\beta_r(\cdot)$ (i.e., the rth coordinate of $\boldsymbol{\beta}(\cdot)$) captures the mean change in the mood score if, for example, the individual sleeps one hour less while holding other predictors in the model constant, and therefore is shared across different streams. Lastly, since the medical interns work in high-stress environments, some of them may experience episodes of mood change after an initial period during which they adapt to daily routines. Such irregular patterns would often last for a period, which leads to (1.1).

Given Model (1.1), at the current time point t_m, our goal is two-folds. First, we want to provide accurate estimates of $\boldsymbol{\beta}(t_m)$ and $\sigma^2(t_m)$ without having to store the entire historical trajectory for each subject (Section 2.1). We refer to this parameter estimation step as the dynamic tracking step. Second, we aim to sequentially detect the occurrence of the changes as soon as possible *for each stream* (Section 2.2). We define, formally, $\mathcal{O}_{t_m} = \{j : \|\boldsymbol{\delta}_j(t_m)\| \neq 0,\ j = 1, \ldots, p\} \subset \{1, \ldots, p\}$ as the subset that contains the indices of the irregular datastreams at the current time t_m, where $\|\cdot\|$ represents the vector L_2 norm. We refer to this second change-point detection step as the dynamic screening step.

Although there have been some recent articles expressing concern about the online updating method for analysis of datastreams (see Schifano et al. (2016); Luo and Song (2020) and the references therein), the issues of developing effec-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

tive methodologies and theories for statistical modeling and learning of massive datastreams still remain. As most of the existing procedures and formulae were mainly developed based on the assumption that the observations come from the same model across time and sources. The primary goal of this manuscript is to provide a dynamic statistical learning–dynamic tracking and screening (DTS)–procedure that fully explores the dynamical features of datastreams. We summarize our contribution from two perspectives.

From a statistical methodology standpoint, our dynamic tracking and screening (DTS) procedure efficiently adapts local dynamic structures in the streaming data environment and detect the irregular patterns as soon as they occur. Specifically, we demonstrate that incorporating exponentially weighted loss functions into our DTS procedure allows the estimates and the test statistics to be updated sequentially in a timely manner (Eqs. (2.2) and (2.10)). As a result, DTS can quickly revise the underlying statistical model as the new data arrive, without the need to store an ever-increasing data history (see Section S1.1 for a discussion on computational complexity).

From a theoretical perspective, our theoretical investigations show that the proposed estimators in DTS are uniformly consistent under some regularity conditions on between-and-within streams dependence even when the proportion of irregular streams does not vanish to zero as p goes to infinity, and the optimal convergence rates of the proposed estimators are presented as separate results (Theorem 2). It is also worth noting that, different from the classical nonpara-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

metric estimation literature where the smoothing parameter is often chosen to minimize the global mean squared error, our DTS chooses the smoothing parameter adaptively so that the averaged prediction error is minimized (Theorem 3). With the help of efficient tracking, a new multiple testing procedure tailed to the streaming environment is developed for screening purposes, and we show that the false discovery rate (FDR) with the data-driven threshold can be controlled at the nominal level uniformly at all time points (Theorem 4).

1.3 Connections to existing work

Model (1.1) is built upon a linear varying coefficient (VC) model that incorporates potential structural changes raised by irregular patterns. The VC models have been extensively studied in the past two decades, especially in the field of longitudinal data analysis, and are known to be very powerful tools for analyzing the relationship between a response and a group of covariates due to its efficiency and flexibility; see Fan and Zhang (2008) for a comprehensive review. The VC models are particularly useful to explore the dynamic pattern in our problem discussed above. Nevertheless, our framework differs substantially from existing literature in various aspects. First, traditional methods usually assume that all of the longitudinal observations have the same model structure, which is not appropriate in the present problem. Second, as the data are collected sequentially and our aim is prospective learning rather than retrospective analysis, only partial information is available rather than the whole functional

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

curves as in the standard VC models. We need to make decisions based on the available observations up to the current time, and with the implicit assumption that more recent observations are more useful for making decisions. Third, as the data collection process runs with high speed, a sequential procedure which is capable of updating parameters with minimal storage requirements is highly desirable. These differences play an important role in the setup of our approaches for parameter estimation and hypothesis testing.

There are some efforts to adapt various sequential change-detection methods to large-scale datastreams surveillance, such as Tartakovsky et al. (2006), Mei (2010), Xie and Siegmund (2013), Zou et al. (2015), Chan (2017) and Ren et al. (2022). Different from our goal in developing dynamic statistical learning, their settings are completely from ours because they aim to minimize the overall expectation delay while controlling the average run length under the null hypothesis that none of the datastreams experience changes. Recent works on sequential testing based on the sequential probability ratio test (SPRT) rules such as Bartroff (2018) and Song and Fellouris (2019) are computationally intensive, making it infeasible for large-scale studies such as those arising from IHS where millions of tests are conducted simultaneously at each time.

More closely related works are Marshall et al. (2004), Grigg et al. (2009), Spiegelhalter et al. (2012), and Gandy and Lau (2012), which considered various applications of Benjamini and Hochberg's FDR control procedure for statistical surveillance. Those methods usually assume that the coefficient function $\boldsymbol{\beta}(\cdot)$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

is either a constant or fully known, which is not appropriate to assume in the present problem since we need to exploit the time-varying structure of (1.1) in a data-driven manner. Recently, some authors considered dynamic testing systems to solve the curve monitoring problem, e.g., see Qiu and Xiang (2014) and Qiu and Xiang (2015), among others. Those methods are simple and effective, however, they are designed under the assumption that the regular pattern is known in advance of detection and only focus on the irregular behavior of local streams, without taking the global false discoveries into account. It should be also emphasized that to the authors' best knowledge, a common feature of most of the literature on sequential detection consists in the fact that either the datastreams are independent (Gandy and Lau, 2012; Xie and Siegmund, 2013) or the streaming observations are independent in the time domain (Ren et al., 2022). However, such assumption is often violated in practice, especially for large-scale datastreams with high frequency observations, which may in turn hamper their applicability to massive data applications. To this end, this article suggests a systematic DTS procedure by making connections to the VC models and some sequential detection problems. We address two key challenges in a unified framework: constructing efficient estimators and multiple testing procedures under model (1.1), and investigating theoretical properties under lenient requirements on the datastreams.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

1.4 Organization

Our paper is structured as follows. In Section 2, we propose the dynamic tracking and screening procedure, followed by investigating its theoretical properties in Section 3. In Section 4, we conduct simulation experiments to show the finite-sample performance of the proposed method in comparison with some others, and then apply our approach on the Intern Health Study for further illustration in Section 5. Section 6 offers a summarizing discussion. We provide some practical guidance about our DTS procedure and additional simulation results on dynamic estimation in Supplementary Material. The proofs and some technical details are delineated in the Supplementary Material.

2. Methodology

At the current time point t_m, we have access to the observations up to time point t_m, i.e., $\{(y_{1j}, \mathbf{X}_{1j}), \ldots, (y_{mj}, \mathbf{X}_{mj})\}$. In addition, because our over-aching goal is to dynamically identify individuals with irregular behaviors and to estimate $\boldsymbol{\beta}(t_m)$ that captures the shared time-varying coefficient across all individuals, we work under the setting that individuals with the irregular pattern $\boldsymbol{\delta}_j(t_m)$ are minority in the study cohort. In other words, the cardinality of $\mathcal{O}_{t_m}$ is small compared to p.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

2.1 Dynamic Tracking

We start with describing our approach for dynamic tracking on estimating $\boldsymbol{\beta}(t_m)$. Because different datastreams have different time-varying coefficients (some have $\boldsymbol{\beta}(t_m)$ and some have $\boldsymbol{\beta}(t_m)+\boldsymbol{\delta}_j(t_m)$), unless the irregular set $\mathcal{O}_{t_m}$ is known as a prior, we cannot naively combine all observed data $\{(y_{1j}, \mathbf{X}_{1j}), \ldots, (y_{mj}, \mathbf{X}_{mj})\}_{j=1}^{p}$ together to construct an accurate estimator of $\boldsymbol{\beta}(t_m)$. Instead, at current time t_m, our dynamic tracking strategy for estimating $\boldsymbol{\beta}(t_m)$ is to first estimate each individual stream coefficients and then perform a robust quantile-based approach to combine those estimates after screening out the irregular ones.

For each datastream j, at the current time t_m, based on the observed data $\{(y_{1j}, \mathbf{X}_{1j}), \ldots, (y_{mj}, \mathbf{X}_{mj})\}$ up to the time point t_m, we consider the following loss function to better incorporate local dynamics:

$$Q_{mj,\lambda}(\mathbf{b}) \equiv \sum_{i=1}^{m} \left(y_{ij} - \mathbf{X}_{ij}^{\top}\mathbf{b}\right)^2 \lambda^{t_m - t_i} \triangleq \sum_{i=1}^{m} \left(y_{ij} - \mathbf{X}_{ij}^{\top}\mathbf{b}\right)^2 w_i(t_m), \tag{2.1}$$

where $\lambda \in (0,1)$ is a smoothing parameter, and $w_i(t_m)$ assigns different weights to different individuals.

The exponential weighting function $w_i(t_m)$ invests at least two merits into our framework. First, it respects the local dynamic nature of our streaming data environment. Because the weighting function assigns smaller weights to individuals farther away from the current time point t_m, and all observations up to t_m are incorporated in $Q_{mj,\lambda}(\mathbf{b})$ to improve statistical estimation efficiency. We note that $Q_{mj,\lambda}(\mathbf{b})$ combines the ideas of local smoothing and exponential

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

weighting schemes used in the exponentially weighted moving average (EWMA) procedures through the term $\lambda^{t_m - t_i}$, which can be regarded as weights defined by a special kernel function (Runger and Prabhu, 1996). Second, the exponential weighting function admits recursive expressions in both tracking and screening steps, whereas some commonly used kernels, such as the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$, may not result in recursive formulae. In other words, though this weighting scheme is just one choice among a broad class of weighting functions, it serves the purpose of our DTS procedure well. In addition, as opposed to the traditional VC model, making one-sided kernel functions is necessary in our problem, since only the observations on one side of t_m are available (Wu and Chu, 1993).

Benefited from the exponential weighting function $w_i(t_m)$, we can quickly obtain an estimate of $\boldsymbol{\beta}(t_m)$ that not only minimizes the lost function $Q_{mj,\lambda}(\mathbf{b})$ but also can be updated efficiently based on the previous time point estimate $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_{m-1})$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m) &= \mathbf{A}_{mj}^{-1}\left\{w_{m-1}(t_m)\mathbf{A}_{m-1,j}\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_{m-1}) + \mathbf{X}_{mj}y_{mj}\right\}, \\ \mathbf{A}_{mj} &= w_{m-1}(t_m)\mathbf{A}_{m-1,j} + \mathbf{X}_{mj}\mathbf{X}_{mj}^{\top}. \end{aligned} \tag{2.2}$$

Then, by defining $e_{ij} = y_{ij} - \mathbf{X}_{ij}^{\top}\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_i)$, the variance function $\sigma^2(\cdot)$ at t_m can be estimated using the observations on the jth stream by

$$\widehat{\sigma}^2_{j,\lambda}(t_m) = \varphi_m^{-1}\sum_{i=1}^{m} w_i(t_m)e_{ij}^2, \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

which again can be recursively updated:

$$\widehat{\sigma}^2_{j,\lambda}(t_m) = \varphi_m^{-1}\left\{w_{m-1}(t_m)\varphi_{m-1}\widehat{\sigma}^2_{j,\lambda}(t_{m-1}) + e^2_{mj}\right\}, \tag{2.4}$$

with $\varphi_m = \sum_{i=1}^m w_i(t_m)$. In Section 3, we prove in Theorem 1 that $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m)$ and $\widehat{\sigma}^2_{j,\lambda}(t_m)$ are appealing estimators given a properly chosen λ.

In the second step, because the majority of $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m)$'s are correctly centered at $\boldsymbol{\beta}(\cdot)$ and the others substantively deviate from $\boldsymbol{\beta}(\cdot)$ pushed by the irregular pattern $\boldsymbol{\delta}_j(t_m)$. This motivates us to adopt a robust quantile-based method to better estimate $\boldsymbol{\beta}(\cdot)$ and $\sigma^2(\cdot)$. We estimate $\boldsymbol{\beta}(t_m)$ component-wise. For the rth component $\beta_r(t_m)$, the signal set $\mathcal{O}_{r,t_m} = \{j:\ \delta_{jr}(t_m) \neq 0\} \subset \{1,\ldots,p\}$ can be divided into two subsets

$$\mathcal{O}^+_{r,t_m} \cup \mathcal{O}^-_{r,t_m} := \{j:\ \delta_{jr}(t_m) > 0\} \cup \{j:\ \delta_{jr}(t_m) < 0\},$$

where $\mathcal{O}^+_{r,t_m}$ contains the subjects with positive biases and $\mathcal{O}^-_{r,t_m}$ includes the rest. By definition, $\mathcal{O}_{t_m} = \cup_{r=1}^d \mathcal{O}_{r,t_m}$. Accordingly, we define

$$\pi^+_{r,t_m} = \frac{\mathrm{Card}(\mathcal{O}^+_{r,t_m})}{\mathrm{Card}(\mathcal{O}_{r,t_m})}, \quad \pi^-_{r,t_m} = \frac{\mathrm{Card}(\mathcal{O}^-_{r,t_m})}{\mathrm{Card}(\mathcal{O}_{r,t_m})}, \quad \pi_{r,t_m} = \frac{1}{2} - \frac{\pi^+_{r,t_m} - \pi^-_{r,t_m}}{2},$$

with $\mathrm{Card}(\cdot)$ being the cardinality of any set. Let $\widehat{\beta}_{jr,\lambda}(t)$ be the rth component of $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$, for $r = 1,\ldots,d$. We estimate $\beta_r(t_m)$ through the π_{r,t_m}-th quantile of $\widehat{\beta}_{1r,\lambda}(t_m), \cdots, \widehat{\beta}_{pr,\lambda}(t_m)$, i.e.,

$$\widetilde{\beta}_{r,\lambda}(t_m) = \inf\left\{\beta_r:\ \widehat{F}_p(\beta_r, t_m) \geq \pi_{r,t_m}\right\}, \tag{2.5}$$

where $\widehat{F}_p(\beta_r, t_m) = p^{-1}\sum_{j=1}^p \mathbb{I}(\widehat{\beta}_{jr,\lambda}(t_m) \leq \beta_r)$ and $\mathbb{I}(\cdot)$ is an indicator function.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

We note that π_{r,t_m} is unknown but it can be estimated consistently, and the estimation details shall be provided in Appendix S1.1.

Constructing an estimate of $\sigma^2(t_m)$ using the information of all datastreams is much simpler because the consistency of $\widehat{\sigma}^2_{j,\lambda}(t_m)$ in (3.2) can always be obtained, regardless of whether j is an outlying stream. Naturally, we estimate $\sigma^2(t_m)$ by the pooled average from all datastreams, which is $\widetilde{\sigma}^2_{\lambda}(t_m) = p^{-1}\sum_{j=1}^{p}\widehat{\sigma}^2_{j,\lambda}(t_m)$.

Like many other smoothing-based procedures, setting the value of the smoothing parameter λ is a critical and non-trivial task. A larger λ may gain on the variance side, but loses on the bias side. The optimal choice of λ often depends on the "smoothness" of $\boldsymbol{\beta}(t_m)$, a quantity that is only estimable under strict, often unrealistic, assumptions. Cross-validation has been frequently adopted in selecting the bandwidth in the VC models or longitudinal data analysis literature (Hoover et al., 1998). The selection of an optimal λ is particularly challenging in the present problem as we need to determine it dynamically for the current time, say $\lambda(t_m)$. This is because the population coefficient function typically varies with time and it does not make sense to assume that $\boldsymbol{\beta}(t_m)$ and $\boldsymbol{\beta}(t_{m'})$ have the same degree of smoothness if $|t_m - t_{m'}|$ is large. This implies that the optimality of estimation across all the time points simultaneously cannot be achieved with a single choice of λ (Zhang and Lee, 2000).

Note that at the current time point t_m, we are only concerned about the estimation of $\boldsymbol{\beta}(t_m)$ rather than the entire regression coefficients; that is, in our

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

dynamic tracking procedure we do not update the estimation of $\boldsymbol{\beta}(t)$ for $t < t_m$ after receiving the streaming observation at t_m. As a result, it is reasonable to define the averaged predictive squared error (APSE) of $\widetilde{\boldsymbol{\beta}}_\lambda(t_m)$ by

$$\mathrm{APSE}_\lambda(t_m) = \mathrm{Card}(\mathcal{I}_{t_m})^{-1} \sum_{j \in \mathcal{I}_{t_m}} \{y^*_{mj} - \mathbf{X}_{mj}^\top \widetilde{\boldsymbol{\beta}}_\lambda(t_m)\}^2,$$

where $\mathcal{I}_{t_m} \subset \{1, \ldots, p\}$ is a subset that contains most of the noise streams, y^*_{mj} is a new observation at $(\mathbf{X}_{mj}, t_m)$, say $y^*_{mj} = \mathbf{X}_{mj}^\top \boldsymbol{\beta}(t_m) + \varepsilon^*_{mj}$, where ε^*_{mj} is a new realization of ε_{mj}. Note that $\mathcal{I}_{t_m}$ is allowed to contaminate with some outlying streams provided that its size is small relative to p. We shall provide a heuristic algorithm to select $\mathcal{I}_{t_m}$ in Appendix S1.1. By the smoothness assumption of $\boldsymbol{\beta}(\cdot)$, we propose a one-step APSE criterion to choose λ dynamically from

$$\widehat{\lambda}(t_m) = \arg\inf_\lambda \widehat{\mathrm{APSE}}_\lambda(t_m), \tag{2.6}$$

where $\widehat{\mathrm{APSE}}_\lambda(t_m) = \mathrm{Card}(\mathcal{I}_{t_m})^{-1} \sum_{j \in \mathcal{I}_{t_m}} \{y_{mj} - \mathbf{X}_{mj}^\top \widetilde{\boldsymbol{\beta}}_\lambda(t_{m-1})\}^2$.

2.2 Dynamic screening

Now, let us turn to construct an effective screening procedure. By defining the normalizing transformation of the response y_{ij} as $z_{ij} = \{y_{ij} - \mathbf{X}_{ij}^\top \boldsymbol{\beta}(t_i)\}/\sigma(t_i)$, we have under model (1.1)

$$z_{ij} = \begin{cases} \varepsilon_{ij}, & \text{for} \quad t_i \in (0, \tau_j], \\ \gamma_j(t_i) + \varepsilon_{ij}, & \text{for} \quad t_i > \tau_j, \end{cases} \tag{2.7}$$

where $\gamma_j(t_i) = \mathbf{X}_{ij}^\top \boldsymbol{\delta}_j(t_i)/\sigma(t_i)$, and $\boldsymbol{\delta}_j(\cdot)$ is a smooth function defined on (τ_j, ∞), for $j = 1, \cdots, p$, $i = 1, 2, \cdots$. In this context, the goal of change points detection

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

is to sequentially check if z_{ij} has zero mean at each time point. More formally, at current time $t = t_m$, we test the null hypotheses

$$H_{mj}^0 : \ \mathbb{E}(z_{mj} \mid \mathbf{X}_{mj}) = 0, \quad j = 1, \ldots, p. \tag{2.8}$$

As $\boldsymbol{\beta}(t_i)$ and $\sigma^2(t_i)$ can be well estimated by $\widetilde{\boldsymbol{\beta}}_\lambda(t_i)$ and $\widetilde{\sigma}_\lambda^2(t_i)$, respectively, $\widetilde{z}_{ij} = \{y_{ij} - \mathbf{X}_{ij}^\top \widetilde{\boldsymbol{\beta}}_\lambda(t_i)\}/\widetilde{\sigma}_\lambda(t_i)$ yields a natural estimate of z_{ij}. Similar to the spirit of (2.2), we estimate $\gamma_j(t_m)$ through

$$\widehat{\gamma}_{j,\lambda}(t_m) := \varphi_m^{-1} \sum_{i=1}^{m} w_i(t_m) \widetilde{z}_{ij}, \tag{2.9}$$

which serves as a proper quantity at current time point t_m for checking H_{mj}^0. The dynamic screening procedure rejects the null hypothesis whenever the streaming pattern of jth datastream deviates from that of the majority, i.e., if $|\widehat{\gamma}_{j,\lambda}(t_m)|$ exceeds a pre-specified threshold at t_m.

Again, benefiting from the exponential weight function $w_i(t_k)$, we have the following recursive form to quickly update $\widehat{\gamma}_{j,\lambda}(t_m)$ as the new data arrive

$$\widehat{\gamma}_{j,\lambda}(t_m) = \varphi_m^{-1} \left\{ w_{m-1}(t_m) \varphi_{m-1} \widehat{\gamma}_{j,\lambda}(t_{m-1}) + \widetilde{z}_{mj} \right\}, \tag{2.10}$$

where $\varphi_m = \sum_{i=1}^m w_i(t_m)$. Under the null hypotheses and certain regularity conditions, we prove that $\widehat{\gamma}_{j,\lambda}(t_m)$ is asymptotically normal. Nevertheless, because reliable estimates of $\widehat{\gamma}_{j,\lambda}(t_m)$'s long-run variance are extremely challenging to obtain in the presence of those irregular patterns, multiple testing procedures that directly adjust for p-values are not suitable to achieve our goal.

We consider an alternative way to apply the well known Benjamini and Hochberg (1995)'s FDR procedure (BH) with statistical guarantees (Theorem

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

4). Concretely, we reject H_{mj}^0 if $|\widehat{\gamma}_{j,\lambda}(t_m)| > L$, where L is a data-driven threshold (Storey et al., 2004) given by

$$L = \inf\left[u : \frac{\#\{j : |\widehat{\gamma}_{j,\lambda}(t_*)| \geq u\}}{\#\{j : |\widehat{\gamma}_{j,\lambda}(t_m)| \geq u\} \vee 1} \leq \alpha\right] \tag{2.11}$$

for a desired FDR level α and some small $t_* > 0$. The screening set is denoted by $\widehat{\mathcal{O}}_{t_m}$. In (2.11), we use a "warm-up" sample to construct a series of null test statistics, denoted as $\widehat{\gamma}_{j,\lambda}(t_*), j = 1, \ldots, p$. Such a warm-up sample is generally available since a conventional assumption in the practice of change-detection is that the changes would be unlikely to occur at the beginning of monitoring. In our motivating example, this is also a reasonable assumption in the context of regular medical residence training where work and life routines are being established as baseline from which deviations may be detected. Intuitively, if most of $|\widehat{\gamma}_{j,\lambda}(t_*)|$'s are from null states, $\#\{j : |\widehat{\gamma}_{j,\lambda}(t_*)| \geq u\}$ would be a reasonable approximation to $\#\{j : |\widehat{\gamma}_{j,\lambda}(t_m)| \geq u, H_{mj}^0\}$. The advantage of this empirical formula is that we do not need to estimate other nuisance parameters and the temporal correlation structures are allowed to be different across datastreams. In Section 3, we shall prove that, under mild conditions, the false discovery proportion (FDP), defined as

$$\mathrm{FDP}(t_m) := \frac{\#\{j : \widehat{\gamma}_{j,\lambda}(t_m) \geq L, j \notin \mathcal{O}_{t_m}\}}{\#\{j : \widehat{\gamma}_{j,\lambda}(t_m) \geq L\} \vee 1} \tag{2.12}$$

with the threshold L is asymptotically controlled at the level α, and such control is valid uniformly at t_m.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

2.3 Dynamic Tracking and Screening (DTS) Procedure

Our proposed DTS procedure is summarized as follows. Some practical guidances about our DTS procedures are given in the Appendix S1.1.

1. (Initiation) Set $\Lambda = \{\lambda_k, k = 1, \ldots, q\}$ and an FDR level α.

2. (Choice of the smoothing parameter) Given the observations $\{(y_{mj}, \mathbf{X}_{mj})\}_{j=1}^{p}$ at the time point t_m, find the optimal λ by $\widehat{\lambda}(t_m) = \arg\inf_{\lambda\in\Lambda} \widehat{\mathrm{APSE}}_\lambda(t_m)$.

3. (Dynamic tracking) Update the estimators $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t_m)$ and $\widehat{\sigma}^2_{j,\lambda}(t_m)$ by (2.2) and (2.3) for each $\lambda \in \Lambda$. Obtain $\widetilde{\boldsymbol{\beta}}_{\widehat{\lambda}(t_m)}(t_m)$ by (2.5) and $\widetilde{\sigma}^2_{\widehat{\lambda}(t_m)}(t_m)$.

4. (Dynamic screening) Compute $\widetilde{z}_{mj}$ and calculate the test statistics $\widehat{\gamma}_{j,\lambda(t_m)}(t_m)$ for $j = 1, \ldots, p$ using (2.10). Search for the threshold L by (2.11); Display the discoveries with the level α, i.e., $\widehat{\mathcal{O}}_{t_m}$.

In step 2, since $\widehat{\lambda}(t_m)$ can only be approximately identified within a compact set of the parameter space and there might exist more than one local minimum, in practice we recommend to find $\widehat{\lambda}(t_m)$ from a pre-specified set with some admissible values, say $\Lambda = \{\lambda_k, k = 1, \ldots, q\}$.

3. Theoretical investigations

In this section, we derive the asymptotic properties of the proposed estimators. We first discuss the assumptions that are needed for the analysis and then

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

summarize the main theorems. The proofs are provided in the Supplemental Material.

3.1 Assumptions

As we are considering the problem of dynamic tracking in an unending sequence, we assume that t_i is deterministic and the incremental time $t_i - t_{i-1} = n_i$ is lower and upper bounded, with $t_0 = 0$. The bandwidth in the exponential weight (i.e., λ^{t-t_i}) is $h = 1/\{-m\log(\lambda)\}$. We make the following assumptions to establish the theoretical foundation of our DTS procedure.

Assumption 1. *The time series processes* $\mathcal{D}_j = \{\mathbf{X}_{ij}, \varepsilon_{ij}, i = 1, \ldots, m\}$ *are strictly stationary and strongly* ρ*-mixing for each* j*. Let* $\rho_j(l)$ *for* $l = 1, 2, \ldots$ *be the mixing coefficients corresponding to the* j*-th time series* $\mathcal{D}_j$*. It holds that* $\rho_j(l) \leq \rho(l)$ *for all* $1 \leq j \leq p$*, where the coefficients* $\rho(l)$ *satisfy that* $\sum_l \rho(l) < \infty$*. Moreover, assume that the eigenvalues of* $\mathbf{\Gamma}_j := \mathbb{E}(\mathbf{X}_{ij}\mathbf{X}_{ij}^\top)$ *are uniformly bounded by zero and infinity.*

Assumption 2. *Suppose* $\{\mathcal{D}_j, j = 1, \cdots, p\}$ *satisfy the block dependence structure in the sense that there exists a partition of the data streams* $\{\mathcal{D}_{j,k}, j = 1, \cdots, J, k = 1, \cdots, n_j\}$ *such that* $\mathcal{D}_{j_1,k_1}$ *are independent with* $\mathcal{D}_{j_2,k_2}$ *for* $j_1 \neq j_2$*, and the maximal block size in each partition is of the order* $O(p/J)$*. In addition, we assume that* $J = p^\zeta$*, for some* $\zeta > 0$*.*

Assumption 3. *There is a real number* $\theta > 20/3$ *such that* $\sup_{i,j} \mathbb{E}(|\varepsilon_{ij}|^\theta) < \infty$*.*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

Assumption 4. *For $t \in [t_*, t_m]$, the varying coefficients $\beta_r(t)$ and $\delta_{jr}(t)$ for $r = 1, \ldots, k$ and the variance function $\sigma^2(t)$ satisfy the Lipschitz conitunity with the dilation constant $A \to 0$.*

Assumption 5. *The number of data streams p diverges to infinity that $p = O(m)$. The tuning parameter $h \to 0$ has the property that $\log(m)/(mh) \to 0$, $A(mh)^{3/2}/\sqrt{\log(m)} \to 0$, and $h \geq Cm^{-2/5}$ for some positive constant $C > 0$. Moreover, assume that $p^{\zeta}/(mh) \to \infty$.*

Assumption 6. *Assume $|\delta_{jr}(t)|/\sqrt{\log(m)/(mhp^{\zeta})} \to \infty$, for $j \in \mathcal{O}_{r,t_m}$ and $t > \tau_j$.*

The condition on the mixing rate $\rho(l)$ in Assumption 1 is not stringent and it can be satisfied if $\rho(l)$ decays to zeros by sufficiently high polynomial rates. Assumption 2 implies that any data stream can be correlated with at most other $O(p^{1-\xi})$ data streams. The moment conditions in Assumptions 3 are used to derive the uniform consistency of the regression coefficients; see Hansen (2008) for similar assumptions. Assumption 4 is the Lipschitz condition with dilation constant shrinking to zero, which is suitable for analyzing unending sequences; a common rate of $A = O(N^{-1})$, where N represents the ending of the sequence, which could be much larger than m. The regression coefficient function $\boldsymbol{\beta}(t)$ in (4.1) of the simulation study satisfies this condition. Assumption 5 imposes restriction on the relative growth of p and m, and the smoothing parameter λ is chosen at a rate faster than the optimal rate in the standard nonparametric

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

regression problems so that the bias term will be negligible. The effective sample size mh is required at least as large as $Cm^{0.6}$ so that the block size used to prove Lemma 4 tends to infinity. Moreover, the lower bound rate $-2/5$ is closely related to the moment condition $\theta > 20/3$ in Assumption 3; see Vogt and Linton (2017) for similar discussions. Assumption 6 guarantees that the signals of the alternative streams dominate the noise so that the quantile-based estimator will be consistent. This condition can be removed if we assume that the signals in the alternative are sparse.

3.2 Main results

We first discuss the asymptotic properties of $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$ and $\widehat{\sigma}^2_{j,\lambda}(t)$, for $t \in [t_*, t_m]$, where $[0, t_*]$ serves as a warm-up period. Throughout this paper, we assume $\mathrm{Card}(\mathcal{O}_{t_m}) \leq cp$ for some $c < 1$, which includes the sparse setting $\mathrm{Card}(\mathcal{O}_{t_m}) = o(p)$. Theorem 1 establishes the uniform consistency of $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$ and $\widehat{\sigma}^2_{j,\lambda}(t)$ for all $t \in [t_*, t_m]$.

Theorem 1. *Under Assumptions 1 and 3-5, $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$ and $\widehat{\sigma}^2_{j,\lambda}(t)$ satisfy*

$$\sup_{t\in[t_*,t_m]} \left\|\widehat{\boldsymbol{\beta}}_{j,\lambda}(t) - \{\boldsymbol{\beta}(t) + \boldsymbol{\delta}_j(t)\}\right\| = O_p(Amh + \sqrt{\log(m)/(mh)}), \tag{3.1}$$

$$\sup_{t\in[t_*,t_m]} |\widehat{\sigma}^2_{j,\lambda}(t) - \sigma^2(t)| = O_p(Amh + \sqrt{\log(m)/(mh)}), \tag{3.2}$$

where $mh = 1/\{-\log(\lambda)\}$ does not depend on m.

In fact, $\log(m)$ appeared in Theorem 1 can be replaced by $\log(mh)$, implying that the uniform convergence rate only depends on λ, not m. To the best of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

our knowledge, the uniform convergence rate of one-sided kernel smoother with correlated errors over a possibly unbounded support has not been thoroughly investigated in the literature. In Theorem 1, we establish the uniform consistency results of the proposed estimators given the time series are ρ-mixing. In the right-hand side of (3.1), $O(Amh)$ is a bound for bias while $O_p(\sqrt{\log(m)/(mh)})$ is a bound for the maximum level of variation. Hence, to make the estimators uniformly consistent, h (equivalently λ), that is similar to the bandwidth in classical nonparametric regression, is required to satisfy $Amh \to 0$ and $mh/\{\log(m)\} \to \infty$. Compared to the results of local polynomial smoothers in nonparametric regression, the vanishing rate of the bias in our estimator is h rather than h^2, which is due to the use of one-sided kernel.

Next theorem provides the uniform convergence rates of $\widetilde{\boldsymbol{\beta}}_\lambda(t)$ and $\widetilde{\sigma}^2_\lambda(t)$.

Theorem 2. *Under Assumptions 1-6, we have*

$$\sup_{t\in[t_*,t_m]} \left\|\widetilde{\boldsymbol{\beta}}_\lambda(t) - \boldsymbol{\beta}(t)\right\| = O_p\left(\sqrt{\log(m)/(mhp^\zeta)}\right), \tag{3.3}$$

$$\sup_{t\in[t_*,t_m]} |\widetilde{\sigma}^2_\lambda(t) - \sigma^2(t)| = O_p\left(\sqrt{\log(m)/(mhp^\zeta)}\right), \tag{3.4}$$

where ζ satisfies that $p^\zeta/(mh) \to \infty$ and ζ appeared in Assumption 2.

The incremental rate $p^{-\zeta}$ due to information fusion is determined by the number of blocks p^ζ, which should diverge to infinity at a rate faster than mh to erase the normal approximation error, where $1/\sqrt{mh}$ is the Berry-Esseen bound of the $\sqrt{mh}\big[\widehat{\boldsymbol{\beta}}_{j,\lambda}(t) - \{\boldsymbol{\beta}(t) + \boldsymbol{\delta}_j(t)\}\big]$ under ρ-mixing assumption as indicated in Lemma 4 of the Supplement. As shown in Theorem 1, as long as

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

$A(mh)^{3/2}/\log(m) \to 0$ so that the bias in $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$ is negligible and $p^{\zeta}/(mh) \to \infty$, we observe that $\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$ has a faster rate of convergence than does $\widehat{\boldsymbol{\beta}}_{j,\lambda}(t)$, which suggests a significant efficiency again due to the fusion of information across streams. This uniform convergence result is particularly helpful for justifying the role of $\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$ and $\widetilde{\sigma}^2_{\lambda}(t)$ in the dynamic screening procedure described in Section 2.2. Though uniform convergence results for kernel estimation with dependent data were discussed in the literature, such as Hansen (2008) and Vogt and Linton (2017), technical arguments for this theorem are highly non-trivial and may be interesting in their own rights because our quantile-based estimator $\widetilde{\boldsymbol{\beta}}_{\lambda}(t)$ is not a linear statistic.

Next result shows that $\widehat{\lambda}(t_m)$ is asymptotically optimal in the sense that it minimizes the averaged predictive squared error.

Theorem 3. *Under Assumptions 1-6, provided that the number of outlying datastreams $\mathcal{I}_{t_m}$ is negligible, as $m \to \infty$, we have*

$$\frac{\mathrm{APSE}_{\widehat{\lambda}(t_m)}(t_m)}{\inf_{\lambda\in(0,1)} \mathrm{APSE}_{\lambda}(t_m)} \to 1$$

in probability, where $\widehat{\lambda}(t_m)$ is obtained by restricting $h = 1/\{-m\log(\lambda)\} \in [C(m)^{-1/2+\delta}, \infty)$ in (2.6) for some constants $C > 0$ and $\delta > 0$.

Theorem 3 is derived under the assumption that $\mathcal{I}_{t_m}$ is not contaminated by many alternative datastreams. The choice of $\mathcal{I}_{t_m}$ proposed in Section S1.1 ensures that the proposed method is able to deliver satisfactory performance in both simulations and the real-data example.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

Our main theoretical result on the asymptotic validity of the DTS method for both FDP and FDR control is given by the next theorem. We need an additional condition on the change magnitude.

Assumption 7. *As* $m \to \infty$, $\psi_m \to \infty$, *where* $\psi_m = |\mathcal{C}_\mu(t_m)|$, $\mathcal{C}_\mu(t_m) = \{j \in \mathcal{O}_{t_m} : |\gamma_j(t_m)|/\nu_m \to \infty, (m-\tau_j)h \to \infty\}$ *and* $\nu_m = \sqrt{\log(p)/(mh)}$.

Remark 1 Assumption 7 is a technical condition for establishing the FDP control of DTS. The ν_m represents the convergence rate of $\widetilde{\boldsymbol{\beta}}_\lambda(t)$ and $\widetilde{\sigma}^2_\lambda(t)$ as discussed in Theorem 2, and the implication of this assumption is that the number of outlying streams with identifiable signal strengths is not too small as $m \to \infty$ and $p \to \infty$ (but ψ_m may still be small relative to p, i.e., $\psi_m/p \to 0$). This seems to be a necessary condition for FDP control under the sparse scenario, say $\text{Card}(\mathcal{O}_{t_m}) = o(p)$. For example, in the context of multiple testing, Liu and Shao (2014) showed that even with the true p-values, no method is able to control FDP with a high probability if the number of true alternatives is fixed as the number of hypothesis tests goes to infinity. To see this clearer, notice that the key step is to show the validity of (2.12) in which the convergence of empirical sum such like $\sum_{j \notin \mathcal{O}_{t_m}} \text{I}(\widehat{\gamma}_{j,\lambda}(t_m) \geq u)$ is needed. When u is extremely large, the number of nonzero terms in the summation would be finite and consequently the convergence would fail. The condition that $\psi_m \to \infty$ helps to rule out such pathologic cases.

Theorem 4. *Suppose Assumptions 1-7 hold. For any* $\alpha \in (0,1)$, *the* FDP *of*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

the DTS *method satisfies* $\mathrm{FDP}(t_m) \leq \alpha + o_p(1)$ *uniformly at* t_m. *It follows that* $\limsup_{(m,p)\to\infty} \mathrm{FDR}(t_m) \leq \alpha$ *uniformly at* t_m.

This theorem shows that the DTS procedure can control the FDR level uniformly at t_m. Numerical study shows that our data-driven FDR control approach works well in finite-sample cases and thus greatly facilitates our screening procedure.

4. Simulation

4.1 Simulation setup

To demonstrate the finite sample performance of the proposed method, we consider model (1.1) with time dependent $\mathbf{X}_{ij} = (1, X_{ij})^\top$ covariates, where X_{ij} is generated from the mean zero Gaussian process with $\mathrm{cov}\{X_{j_1}(t_{i_1}), X_{j_2}(t_{i_2})\} = 0.8^{|t_{i_1} - t_{i_2}|}$. The possible ending point is set as N which could be much larger than m. To mimic the real world scenario, our data generating process incorporates not only between-stream dependence but also temporal correlation introduced by the noise variable. In doing so, for each stream j, we first generate $\varepsilon_j := (\varepsilon_{1j}, \ldots, \varepsilon_{Nj})^\top$ from a mean zero Gaussian process with autocorrelation $\rho_{\mathrm{Tempo}} \in \{0, 0.5\}$. Then, we multiply the stacked noise matrix $(\varepsilon_1, \ldots, \varepsilon_p)^\top$ with a diagonal block structured correlation matrix Σ_{Block} of block size $n_{\mathrm{Block}} = 200$, and the within block correlation is set to be $\rho_{\mathrm{Block}} \in \{0, 0.5\}$. The number of streams $p = 800$. The noise level $\sigma^2(t) \in \{1, 8\}$. Let $t_i = i$ for $i = 1, \cdots, N$,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

and we use the first 300 time points as warm-up period. Finally, we consider two cases in which the time points $N \in \{2400, 4800\}$.

Next, for any $t \in \{t_1, \ldots, t_N\}$, we generate the coefficient $\boldsymbol{\beta}(t) = (1, \beta(t))^\top$ through

$$\beta(t) = \begin{cases} \frac{\sin\{(14s)^{1.5} - 14s\}\exp(7s)}{20} + 3, & \text{if } s = \frac{t}{N} \in (0, \frac{1}{2}), \\ \frac{\sin\{(7-14s)^{1.5} - 7 + 14s\}\exp(7-7s)}{20} + 3, & \text{if } s = \frac{t}{N} \in [\frac{1}{2}, 1]. \end{cases} \tag{4.1}$$

A pictorial illustration of $\beta(t)$ is given in Figure S1. For $t \leq \frac{N}{2}$, we generate the drift $\boldsymbol{\delta}_j(t) = (\delta_j(t), 0)^\top$ with the following signal lengths and patterns:

$$\delta_j(t) = \begin{cases} 10, & \text{if } j \in \{1, \ldots, p/10\},\ t \in [\frac{N}{6} + 1, \frac{N}{4}] \cup [\frac{N}{3} + 1, \frac{11N}{24}], \\ 1, & \text{if } j \in \{p/10 + 1, \ldots, p/5\},\ t \in [\frac{N}{6} + 1, \frac{N}{4}] \cup [\frac{N}{3} + 1, \frac{11N}{24}], \\ 0, & \text{otherwise.} \end{cases}$$

We refer to this period with rather stable change points as "fixed signal period." When $t \in [\frac{N}{2} + 1, N]$, we consider a more realistic scenario: once a signal occurs, both the signal strength and length change over time. For $j = 1, \ldots, p$, assume that the jth stream has $\widetilde{T}_j$ change points $\tau_{j1}, \cdots, \tau_{j,\widetilde{T}_j}$, where $\widetilde{T}_j = T_j \mathbf{1}_{T_j \leq 5}$, and T_j follows Poisson distribution with mean 3. Under this data generating process, overall about 1/5 datastreams contain signals. Then, to well separate the signals between two adjacent changes points, we generate random change points under the constraint that $|\tau_{jk} - \tau_{j,k+1}| > 200$, for $j = 1, \ldots, p$, $k = 1, \ldots, \widetilde{T}_j - 1$. The length of the signals is randomly sampled from Uniform$[30, 80]$. The size of the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

signal is a nonlinear function of t

$$\delta_j(t) = \frac{1}{3}\sin\left(\frac{9t}{2N}\pi\right) + \omega_j, \text{ if } j \in \mathcal{O}_t,$$

where ω_j is either 2 or 7 based on a random draw. The smoothing parameter is adaptively chosen from $\lambda_l = \exp\left(-C_l N^{-0.3}\right), C_l = 0.10 + l/10, l = 1, \ldots, 10$. The numerical results presented in this section are evaluated through 200 Monte Carlo replications.

4.2 Simulation results for dynamic testing

We present simulation results for dynamic testing in this Section, while the simulation results on dynamic estimation are provided in Appendix S2.

We compare the DTS testing procedure with three competitors. The first one is the moving-window-based nonparametric test (MWNT) proposed by Zheng (1996). The second approach we compare with is based on estimating the long-run covariance matrix via Andrews (1991), where the author proposes heteroskedasticity and autocorrelation consistent (HAC) estimation of covariance matrices. Lastly, we compare the performance of the DTS testing procedure based on the naive pooled estimator $\widehat{\beta}_{\lambda,\text{pool}}(t)$ in (S2.1). The details of these three methods can be found in Appendix S1.2.

In the following sections, DTS refers to the proposed testing procedure, DTS (Pooled) refers to the DTS procedure with $\widehat{\beta}_{\lambda,\text{pool}}(t_m)$, HAC-LFDR referes to the testing procedure based on (S1.3) (Andrews, 1991) with the local false discovery rate adjustment procedure (Efron, 2004), HAS-BC refers to the procedure based

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

on (S1.3) with the Benjamini-Hochberg procedure to adjust for the multiple comparison effect, and MWNT refers to the testing procedure built on Zheng (1996). Since it is difficult to detect if the dependence between data streams exists in reality, we implement the described decorrelation strategies even if the simulated data streams are independent.

4.2.1 Computational efficiency comparison

Computational efficiency is a vital concern to screen out the irregular individuals in the massive datastreams, as of which the primary goal is to find out the signals as soon as they occur. To illustrate the benefits of our proposal, we report the average runtime for DTS, HAC- (with LFDR) and MWNT- (with $b_n = 0.03N$) based procedures in Table 1. We also note that the simulations are paralleled on 50 nodes (each node is equipped with 2.5 GHz Intel Xeon 10-core Ivy Bridge processors) via the packages "doSNOW" (for parallelization) and "rlecuyer" (for correct parallelization of random numbers).

Table 1: Computation time (unit: second) of the three testing procedures for dependent datastreams when $(N, p) = (4800, 800)$, $\rho_{\text{Block}} = \rho_{\text{Tempo}} = 0.5$ and $\sigma^2(t) = 1$.

	DTS	HAC	MWNT
$N = 2400$	35.46	6931.77	1893.55
$N = 3600$	50.12	8691.21	2817.22
$N = 4800$	68.49	12,724.01	3644.08

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

From the results in Table 1, due to the usage of updating formulae in Section 2.1, our DTS procedures have clear advantages over MWNT and the HAC procedures. This advantage also suggests the need of careful designs when using a standard model specification test in a dynamic streaming environment.

4.2.2 FDR, TPR and the length of delay comparison

In this section, we evaluate the performance of DTS in terms of FDR controls, true positive rate (TPR), and the length of delay. The latter one is defined as the minimum number of steps that a procedure takes to detect a signal within each alternative period. In the literature of sequential change detection (e.g., see Zou et al. (2015)), this detection delay corresponds to the well-known *run-length*. If there is no such signal of detection, the detection delay is simply set as the length of that specific signal period. We record the medians of detection delays and TPRs amongst all shift periods in each replication. The nominal FDR level is set to be 0.1. To avoid redundancy, in this part, we report the simulation results in two extreme cases: (i) independent datastreams without temporal correlation, $(N, p) = (4800, 800)$ and noise level $\sigma^2(t) = 1$, and (ii) dependent datastreams with temporal correlation, $(N, p) = (2400, 800)$ and noise level $\sigma^2(t) = 8$. Figure 1 and 3 show results for independent datastreams, and Figure 2 and 4 show results for dependent datastreams.

For the false discovery rate control, from the results in Figures 1 and 2, we observe that both DTS and, to a lesser extent, MWNT- and HAC- based

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

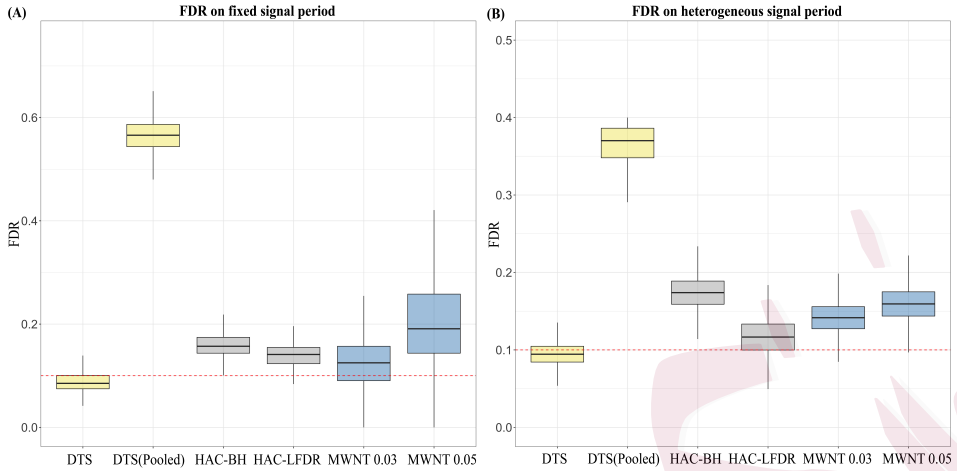

Figure 1: For independent streams without temporal correlation, $(N, p) = (4800, 800)$, $\sigma^2(t) = 1$: Figure (A) is the boxplot of the empirical FDR in the fixed signal period; Figure (B) is the boxplot of the empirical FDR in the heterogeneous signal period.

procedures control FDR at desired levels within acceptable ranges when the datastreams are independent. The Pooled-DTS procedure, built on the inconsistent estimate of $\boldsymbol{\beta}(t)$, fails to control the FDR at the nominal level. When datastreams are correlated, the decorrelation-based testing procedures (HAC-BH, HAC-LFDR and MWNT) tend to have higher FDR than the independent case. In a sharp contrast, the new DTS procedure that utilizes the empirical distribution to approximate the number of false positives is able to deliver very accurate FDR control irrespective of the correlation structures and signal patterns.

In terms of signal detection comparison displayed in Figure 3-4, not surpris-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

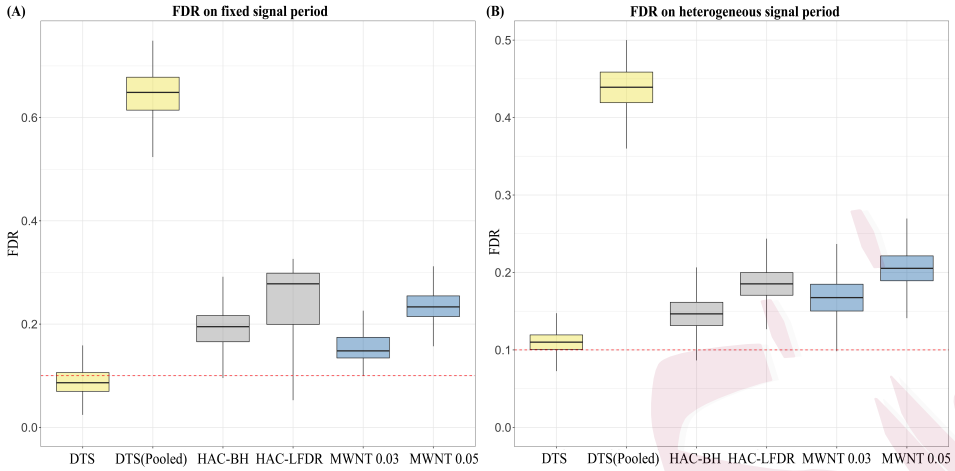

Figure 2: For dependent streams $\rho_{\text{Block}} = 0.5$ with temporal correlation $\rho_{\text{Tempo}} = 0.5$, $(N, p) = (2400, 800)$, $\sigma^2(t) = 8$: Figure (a) is the boxplot of the empirical FDR in the fixed signal period; Figure (b) is the boxplot of the empirical FDR in the heterogeneous signal period.

ingly, since the pooled-DTS does not take the outliers into account and $\boldsymbol{\beta}(t)$ is poorly estimated, it yields low detection power. Our DTS method outperforms the other competitors by a significant margin in both scenarios from the viewpoint of detection delay. It also delivers satisfactory performance in terms of TPR. Compared to DTS, MWNT- and HAC-based procedures show comparative detection power due to the signal accumulation, but the long detection delay seems to be unavoidable. Based on the same reasoning, MWNT shows higher detection power with $b_n = 0.05N$, although longer bandwidth yields longer detection delay. In sum, the DTS addresses the need to dynamically detect the streaming pattern and provides more robust performance than the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

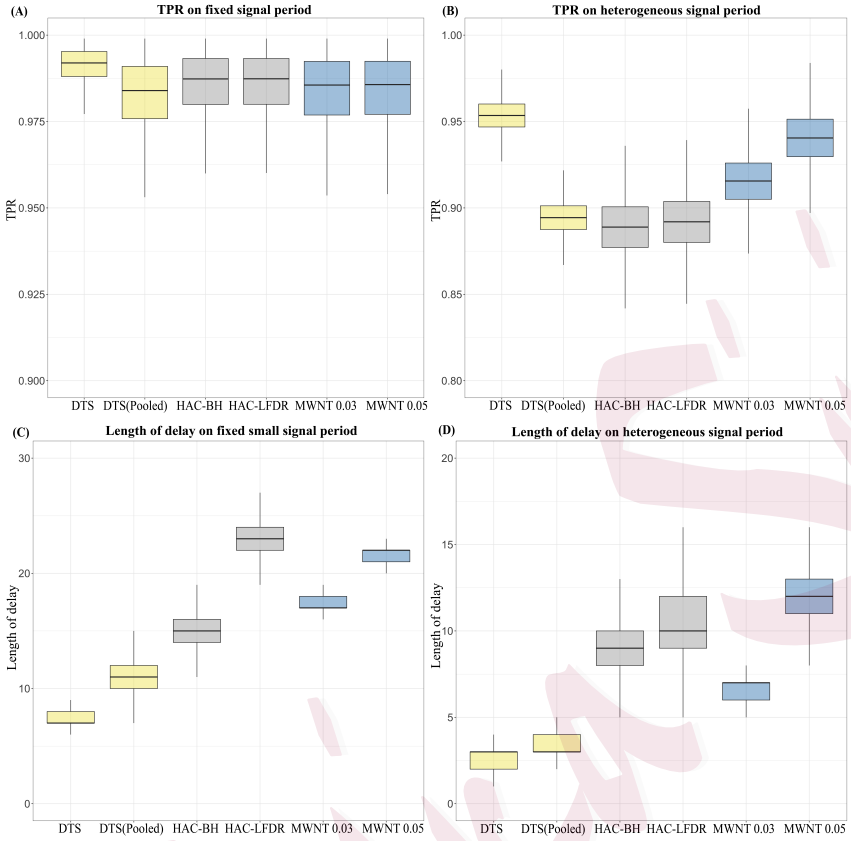

Figure 3: For independent streams without temporal correlation, $(N, p) = (4800, 800)$, $\sigma^2(t) = 1$: the boxplots of the medians of TPRs and detection delays in the fixed signal and the heterogeneous periods.

existing methods from the viewpoints of detection delay and power.

5. Analysis of Intern Health Study Data

The Intern Health Study (IHS) is an on-going mobile health cohort study that enrolls more than 3,000 interns annually. The long-term goal of the IHS is to elucidate the pathophysiological architecture underlying depression to facilitate

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

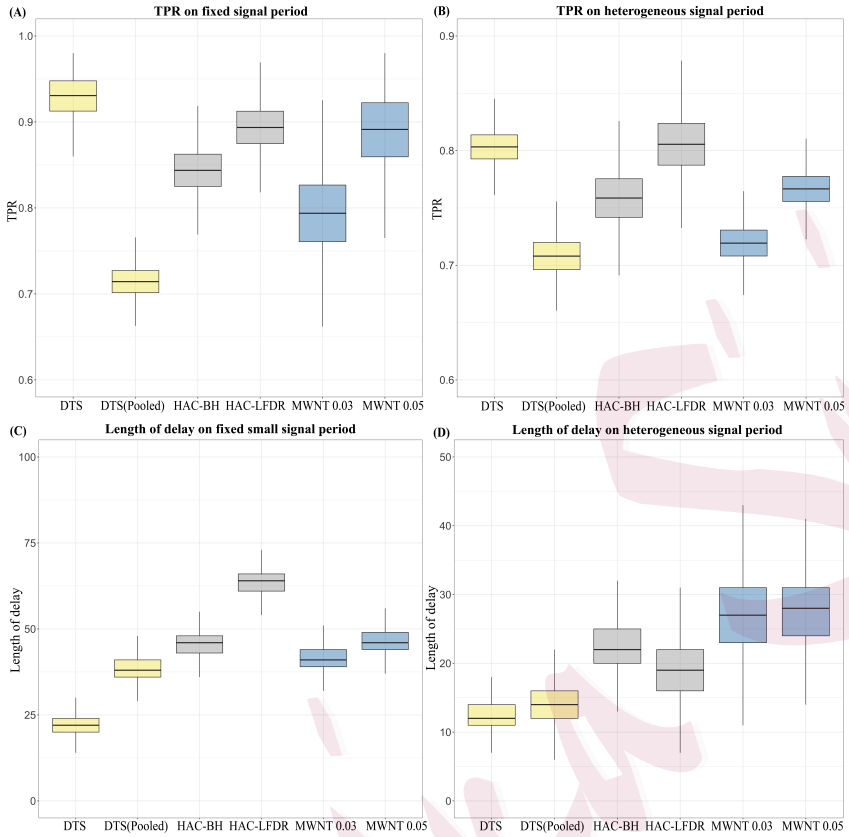

Figure 4: The boxplots of the medians of TPRs and detection delays in the fixed signal and the heterogeneous periods for dependent streams $\rho_{\text{Block}} = 0.5$ with temporal correlation $\rho_{\text{Tempo}} = 0.5$, $(N, p) = (2400, 800)$, $\sigma^2(t) = 8$.

the development of improved treatments. One important goal of IHS is to identify subjects with short-term risks for mood changes whenever new data from the mobile app "MyDataHelps" are updated on a daily basis.

In this case study, for illustration purposes, we have restricted our attention to the 2018 IHS cohort with $1,565$ subjects enrolled from July to December. During the study period, the study tracks medical interns using phones and wear-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

ables. The outcomes are daily self-reported mood valence (measured through a one-question survey; one of two cardinal symptoms of depression, Löwe et al. (2005)). Participants are prompted to enter their daily mood rated from 1-10 every day at a user-specified time between 5 pm and 10 pm. The time-varying covariates are daily steps prior to the survey (as a proxy for activity) and daily sleep duration that ended in the same day. Both covariates are important potential predictors of mood (Kalmbach et al., 2018). Our data set thus consists of data from $p = 1,565$ subjects over $m = 182$ days. We treat the first 40 days as the warm-up phase where data are used to initialize the varying coefficient $\boldsymbol{\beta}(\cdot)$ estimate. This warm-up phase can be viewed as study baseline, because during this period work and life routines are being established and subjects are usually not suffered from sustained stress.

The upper panels of Figure 5 show the *online* time-varying effect estimates of daily step counts (cubic root transformed; left) and daily sleep hours (square root transformed; right) upon the mood score along with 95% pointwise confidence bands. In other words, in each time point, we only use the data point collected prior to this time point. Although the actual magnitudes of the two estimated effect curves depend on the scale of the predictors, we obtain prominent positive effect over time for both the step and sleep predictors indicating their dynamic and positive effects upon daily mood uniformly over time. The proposed DTS method also detects periods when an individual's mood trajectory as a function of time, sleep hours and step counts cannot be described by a

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

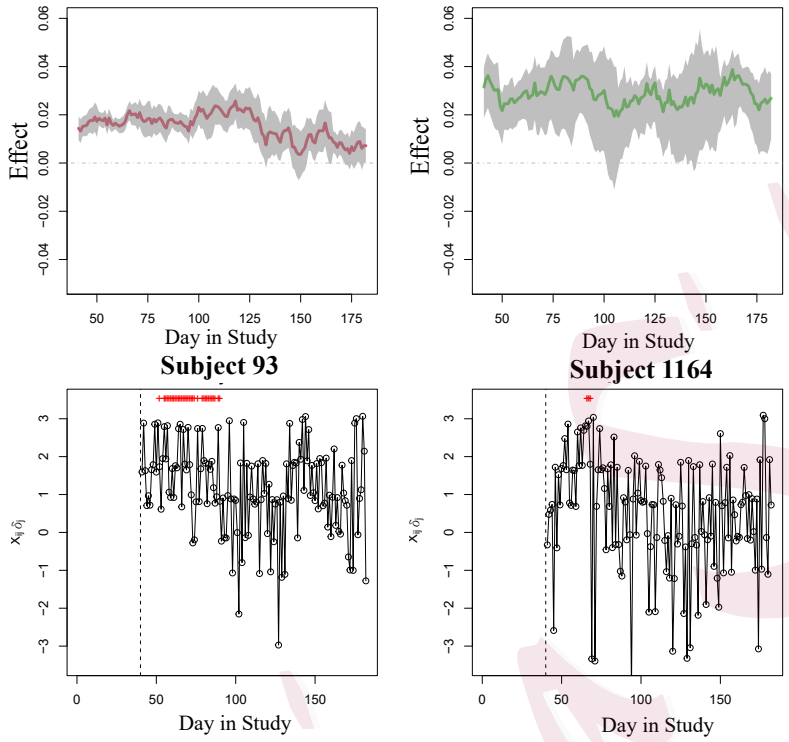

Figure 5: Upper panels: the estimated main time-varying effects for (left) step counts (cubic root scale) and (right) sleep hours (square root scale; Lower panels: the estimated $\mathbf{X}_{ij}^{\top}\boldsymbol{\delta}_j(t_i)$ for two random subjects. The red horizontal pluses indicate the days with detected deviation from the population model; both are online estimates/decisions.

null population model. In the lower panels of Figure 5, we illustrate for two randomly chosen subjects the time points when we detected such deviations from a population dynamic model (red pluses at the top). Relative to the subject on the right with mood scores oscillating around values predicted by a population mean model, the subject on the left has mood scores that are too high to be well characterized by the population model. The estimated coefficients

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

indicate longer periods of extra effects of sleep and activity. Our results show that individuals have distinct timings and duration when the joint effect of sleep hours and step counts changes the mood to a different degree than others in the population, highlighting the timings to intervene upon sleep and activities, for example, through push notifications via their mobile phones.

6. Concluding remarks

We conclude this article with several remarks. Firstly, although the local linear kernel estimator has certain advantages over the local constant kernel estimator, as shown in the literature (Fan and Zhang, 1999), our simulation results show that these two estimators yield similar performance in the dynamic screening. Thus, the local constant kernel procedure (2.1) is chosen for simplicity. Systematic study of local polynomial smoothers in the present problem warrants future research.

Secondly, one important issue with varying coefficient models is how to incorporate the within-subject correlation structure into the estimation or testing procedure. This issue has been investigated, and the methodology has been well established, especially for longitudinal data analysis; e.g., see Sun et al. (2007). For estimation, Theorem 2 shows that the consistency of the proposed estimators is valid under quite general correlation assumptions. Though it has been shown that an estimator can be improved by incorporating the within-subject correlation into the estimation procedure (Fan et al., 2007), such an improved

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

procedure would generally require iterative steps (at least two steps) and the corresponding estimators do not permit recursive calculation. Although it may be computationally feasible to perform a complicated estimation for fixed longitudinal data, fast implementation is likely to be our first priority in massive streaming cases. Certainly, it is of interest to see how the correlation structure can be accommodated into our dynamic tracking procedure.

Finally, the variance function $\sigma^2(t)$ may experience change after some time point for some datastreams and we could use the similar quantile-based estimator to estimate the common variance function across streams. However, the difference of the proportions of positive and negative shift from the common variance function may not be well estimated due to the fact that the sampling distribution of the variance-type estimator is chi-squared distributed. We leave this interesting question for future research.

Supplementary Materials

The supplementary file contains practical implementations and three competing methods used in the simulation studies, additional simulation results, several key lemmas, and the proofs of Theorems 1–4.

Acknowledgements

The authors thank the Associate Editor and two anonymous referees for their exceptional comments that lead to improvement of the paper. Wang acknowl-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

edges the support of NSF DMS-2220537. Du's research is supported by Hong Kong RGC-GRF-16302620 and CityU Start-up Grant (Grant No: 7200774). Zou was supported by the National Key R&D Program of China (Grant Nos. 2022YFA1003703, 2022YFA1003800) and the National Natural Science Foundation of China (Grant Nos. 11925106, 12231011, 11931001,12226007,12326325). This work was partially supported by grants from the National Institutes of Health (R01 MH101459 to ZW), and an investigator award from Precision Health Initiative at the University of Michigan to ZW. We thank Dr. Srijan Sen for generous support in the IHS data access.

References

Aggarwal, C. C. (2007). *Data streams: models and algorithms*, Volume 31. Springer Science & Business Media.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society 59*(3), 817–858.

Bartroff, J. (2018). Multiple hypothesis tests controlling generalized error rates for sequential data. *Statistica Sinica 28*(1), 363–398.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Black, M. and R. Hickey (2003). Learning classification rules for telecom customer call data under concept drift. *Soft Computing 8*(2), 102–108.

Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics 45*(6), 2736–2763.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association 99*(465), 96–104.

Fan, J., T. Huang, and R. Li (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association 102*(478), 632–641.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and its Interface 1*(1), 179–195.

Gama, J. (2010). *Knowledge discovery from data streams*. Chapman and Hall/CRC.

Gandy, A. and F. D.-H. Lau (2012). Non-restarting cumulative sum charts and control of the false discovery rate. *Biometrika 100*(1), 261–268.

Grigg, O., D. Spiegelhalter, and H. Jones (2009). Local and marginal control charts applied to methicillin resistant staphylococcus aureus bacteraemia reports in uk acute national health service trusts. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 49–66.

Guerriero, M., P. Willett, and J. Glaz (2009). Distributed target detection in sensor networks using scan statistics. *IEEE Transactions on Signal Processing 57*(7), 2629–2639.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*(3), 726–748.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Kalmbach, D. A., Y. Fang, J. T. Arnedt, A. L. Cochran, P. J. Deldin, A. I. Kaplin, and S. Sen (2018). Effects of sleep, physical activity, and shift work on daily mood: a prospective mobile monitoring study of medical interns. *Journal of General Internal Medicine 33*(6), 914–920.

Kious, B. M., A. Bakian, J. Zhao, B. Mickey, C. Guille, P. Renshaw, and S. Sen (2019). Altitude and risk of depression and anxiety: findings from the intern health study. *International Review of Psychiatry 31*(7-8), 637–645.

Liu, W. and Q.-M. Shao (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics 42*(5), 2003–2025.

Löwe, B., K. Kroenke, and K. Gräfe (2005). Detecting and monitoring depression with a two-item questionnaire (phq-2). *Journal of Psychosomatic Research 58*(2), 163–171.

Luo, L. and P. X.-K. Song (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 69–97.

Marshall, C., N. Best, A. Bottle, and P. Aylin (2004). Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 167*(3), 541–559.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

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

Qiu, P. and D. Xiang (2014). Univariate dynamic screening system: an approach for identifying individuals with irregular longitudinal behavior. *Technometrics 56*(2), 248–260.

Qiu, P. and D. Xiang (2015). Surveillance of cardiovascular diseases using a multivariate dynamic screening system. *Statistics in Medicine 34*(14), 2204–2221.

Ren, H., C. Zou, N. Chen, and R. Li (2022). Large-scale datastreams surveillance via pattern-oriented-sampling. *Journal of the American Statistical Association 117*(538), 794–808.

Runger, G. C. and S. S. Prabhu (1996). A markov chain model for the multivariate exponentially weighted moving averages control chart. *Journal of the American Statistical Association 91*(436), 1701–1706.

Schifano, E. D., J. Wu, C. Wang, J. Yan, and M.-H. Chen (2016). Online updating of statistical inference in the big data setting. *Technometrics 58*(3), 393–403.

Song, Y. and G. Fellouris (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. *The Annals of Statistics 47*(3), 1776–1803.

Spiegelhalter, D., C. Sherlaw-Johnson, M. Bardsley, I. Blunt, C. Wood, and O. Grigg (2012). Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(1), 1–47.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Sun, Y., W. Zhang, and H. Tong (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics 35*(6), 2795–2814.

Tartakovsky, A. G., B. L. Rozovskii, R. B. Blazek, and H. Kim (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*(9), 3372–3382.

Tsung, F., Z. Zhou, and W. Jiang (2007). Applying manufacturing batch techniques to fraud detection with incomplete customer information. *IIE transactions 39*(6), 671–680.

Vaughan, J., S. Stoev, and G. Michailidis (2013). Network-wide statistical modeling, prediction, and monitoring of computer traffic. *Technometrics 55*(1), 79–93.

Vogt, M. and O. Linton (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1),

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0195

5–27.

Wu, J. and C. Chu (1993). Kernel-type estimators of jump points and values of a regression function. *The Annals of Statistics 21*(3), 1545–1566.

Xie, Y. and D. Siegmund (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics 41*(2), 670–692.

Zhang, W. and S.-Y. Lee (2000). Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis 74*(1), 116–134.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics 75*(2), 263–289.

Zou, C., Z. Wang, X. Zi, and W. Jiang (2015). An efficient online monitoring method for high-dimensional data streams. *Technometrics 57*(3), 374–387.

Department of Biostatistics, University of California, Berkeley; E-mail: jingshenwang@berkeley.edu

Department of Management Sciences, City University of Hong Kong; E-mail: lilundu@cityu.edu.hk

School of Statistics and Data Science, LPMC, KLMDASR and LEBPS, Nankai University; E-mail: nk.chlzou@gmail.com

Department of Biostatistics, University of Michigan; E-mail: zhenkewu@umich.edu