Online Change-Point Detection of Linear Regression Models

Jun Geng, *Member, IEEE*, Bingwen Zhang, Lauren M. Huie and Lifeng Lai, *Member, IEEE*

***Abstract*—In this paper, we consider the problem of quickly detecting an abrupt change in linear regression models. Specifically, an observer sequentially obtains a sequence of observations, whose underlying linear model changes at an unknown time. Moreover, the pre-change linear model is perfectly known by the observer but the post-change linear model is unknown. The observer aims to design an efficient online algorithm to detect the presence of the change via his sequential observations. Based on different assumptions on the change-point, both non-Bayesian and Bayesian problem formulations are considered. In the non-Bayesian setting, the change-point is modeled as a fixed but unknown constant. Two performance metrics, namely the worst case detection delay (WADD) and the average run length to false alarm (ARL2FA), are adopted to evaluate the performance of detection algorithms. In the Bayesian setting, the change-point is modeled as a geometrically distributed random variable. For this case, the average detection delay (ADD) and the probability of false alarm (PFA) are used as performance metrics. We propose a novel algorithm, namely the parallel-sum algorithm, for the purpose of change detection. For both setups, we show that the proposed algorithm has a low computational complexity while still offering a good performance in terms of the performance metrics of the respective setting.**

***Index Terms*—Change-point detection; linear regression model; sequential analysis.**

I. Introduction

Linear regression is a basic but important tool in statistics, signal processing and machine learning. It has wide range applications in data fitting, classification, feature or subset selection [2], beam forming [3], cognitive radio network [4], economic data analysis [5], biomedical science [6], etc. Many efforts have been devoted into the problem of estimating the coefficients in the linear regression model based on observation data [7]–[11]. The underlying assumption in these work is that all data come from a single linear model. However, in many applications, the underlying model may change over time [12]. For example, in building economic growth models, it is more appropriate to assume that various economic indicators obey different models in different time period as the economic growth pattern undergoes structural changes over the years [13]. As another example, in monitoring the health of control systems, the presence of a problem will cause the system to change from a model of normal state to another model of abnormal state [14]. In such applications, it is of interest to detect the presence of such changes in the underlying model quickly.

Motivated by above applications, we focus on *on-line change detection problem* in linear regression models in this work. In particular, an observer keeps monitoring the explanatory variables $\mathbf{x}_n$ and the dependent variable y_n of a linear model. At an unknown time t, the linear model changes to another linear model with unknown coefficients. The observer aims to design an on-line algorithm to quickly detect the presence of such change based on his sequential observations.

We formulate this problem in the framework of quickest change-point detection (QCD). Based on different assumptions on change-point t, both non-Bayesian and Bayesian setups are considered in this paper. In the non-Bayesian setup, the change time t is assumed to be a fixed but unknown number. Specifically, Lorden's setup [15] is considered. In this case, the observer aims to minimize the worst case average detection delay (WADD) while keeping the average run length to false alarm (ARL2FA), namely the expected duration between two successive false alarms, under control. WADD and ARL2FA will be precisely defined in the model section. In the Bayesian setup, the change-point is assumed to be a geometrically distributed random variable [16], [17]. Correspondingly, the observer wishes to minimize the average detection delay (ADD) (average over the prior distribution of the change-point) such that the probability of false alarm (PFA) is under control.

In the considered problem, the post-change coefficient in the linear regression model is unknown to the observer; hence the formulated problem is closely related to the QCD problem with unknown post-change parameters [18]–[21]. However, as we will discuss in the sequel, typical existing methods, including generalized likelihood ratio (GLR) based cumulative sum (CUSUM) algorithm and GLR-Shiryaev algorithm, have a high computational complexity. In this paper, we focus on designing schemes that have a low complexity yet still offer reasonable performance. In particular, we propose a low complexity parallel-sum algorithm. In this algorithm, the observer calculates the correlations between y_n and each individual component of $\mathbf{x}_n$ and then compares the sum of these calculated statistics with a pre-designed threshold. If the threshold is exceeded, which indicates that y_n strongly

The work of J. Geng was supported by the National Natural Science Foundation of China under grant 61601144 and by the Fundamental Research Funds for the Central Universities under grant AUGA5710013915. The work of B. Zhang and L. Lai was supported by the National Science Foundation under grant ECCS-1711468. This paper was presented in part at IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, Mar. 2016 [1].

J. Geng is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, 150001, China (Email: j-geng@hit.edu.cn). B. Zhang is with the Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA (Email: bzhang@wpi.edu). Lauren M. Huie is with Air Force Research Laboratory, Rome, NY, 13440, USA (Email: lauren.huie@us.af.mil). L. Lai is with the Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA (Email: lflai@ucdavis.edu).

depends on some components in $\mathbf{x}_n$, the observer raises an alarm. The performance of the proposed algorithm is analyzed for both non-Bayesian and Bayesian formulations. In the non-Bayesian formulation, to guarantee ARL2FA to be no less than a preset level γ, we show that WADD of the parallel-sum algorithm is on the order of $O(\log\gamma)$ when $p/\gamma \to 0$, in which p is the dimension of $\mathbf{x}_n$, and is on the order of $O(\log p)$ when $p/\gamma \to c$ with c being a constant. In the Bayesian formulation, to guarantee PFA to be no larger than a given threshold α, we show that ADD of the proposed algorithm is on the order of $O(|\log\alpha|)$ when $p\alpha \to 0$ and is on the order of $O(\log p)$ when $p\alpha \to c$. The proposed algorithm is neither optimal nor asymptotically optimal. However, it has a low computational complexity and its detection delay is reasonable. At time slot n, the computational complexity of the proposed algorithm is on the order of $O(np)$.

Our paper is related to several interesting papers on change detection in the linear model. For example, [22] extends Lorden's results to detecting the change in the linear regression model and proposes a family of asymptotically optimal algorithms by exploring the relationship between one-sided sequential probability ratio test (SPRT) and QCD problem. [23] proposes a first order asymptotically optimal detection algorithm in which the unknown mean of the dependent variable in the linear model is replaced by its one-step ahead estimate. [14] adopts the window-limited GLR-CUSUM for the change detection in the stochastic dynamic system. [24] decomposes the unknown post-change parameter space into several subspaces, and for each subspace the observer runs a recursive GLR test for detection purpose. [25] and [26] discuss detecting changes in the linear regression parameter under both fixed sample setting and sequential setting. One can find more discussions on this topic in a recent book [27]. However, all these aforementioned works are focus on the non-Bayesian setting, and most of existing algorithms have a high computational complexity. In this paper, we consider both non-Bayesian and Bayesian settings, for which we propose a low complexity algorithm and analyze its performance in detail.

We now briefly review other related papers. There are a series of works such as [28], [29] that consider the problem of monitoring model or structural change. However, these works focus on the probability of detecting the change-point while our work focuses on analyzing the detection delay. Some other works, such as [13], [30], consider the scenario that the structure of data in the dataset undergo several changes. These works commonly assume that the whole dataset is available to the observer, and the observer aims to design the offline algorithm to estimate the location change-point; hence the estimation error is of interest. However, in our work, observations come to the observer in a sequential manner, and the observer aims to design online change detection algorithm; hence the detection delay and the false alarm are of interest. Our problem is also related to the sequential joint detection and estimation problem considered in [31], [32], as in our model the post-change linear coefficient is unknown to the observer. However, the observations in [31], [32] are identically distributed. In this paper, the distribution of observations has an abrupt change, and we focus on quickly detecting the occurrence of the change rather than estimating the unknown linear coefficient. In addition, the proposed parallel-sum algorithm is similar to both the SUM-CUSUM algorithm [33] and to the mixture-based CUSUM algorithm [34] in a multi-sensor setting. These existing algorithms design the detection statistic by taking the sum or the weighted sum of the local CUSUM statistics, while the proposed parallel-sum algorithm taking the sum of the correlation coefficients between the dependent variable and the components in the corresponding explanatory variable. Further, SUM-CUSUM and mixture-based CUSUM are designed for the sequential change detection of multiple data streams, while the proposed parallel-sum algorithm is for the detection of change in the linear regression models.

The remainder of this paper is organized as follows. The mathematical model is described in Section II. Section III presents the proposed algorithms and the main conclusions of this paper. Numerical examples are provided in Section IV to illustrate the analytical results obtained in this work. Finally, Section V offers concluding remarks.

II. MODEL

We consider the change-point detection problem in a linear regression model. For a given sequence of $p \times 1$ explanatory vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots$, the observations $z_1, z_2, \ldots$, obey the following linear model

$$z_n = \begin{cases} \boldsymbol{\beta}_0^T \mathbf{x}_n + \epsilon_n & n = 1, 2, \ldots, t-1 \\ \boldsymbol{\beta}_1^T \mathbf{x}_n + \epsilon_n & n = t, t+1, \ldots \end{cases}, \tag{1}$$

in which $\epsilon_1, \epsilon_2, \ldots,$ are i.i.d. $\mathcal{N}(0,1)$ random variables that model the observation noise, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ model the pre-change and the post-change linear regression coefficients, respectively. In addition, $\boldsymbol{\beta}_0$ is perfectly known by the observer but $\boldsymbol{\beta}_1$ is unknown. The value of the change-point t is unknown. Hence, (1) indicates that the relationship between $\mathbf{x}_n$ and z_n abruptly changes to an unknown linear model from a known linear model at some unknown time t.

We note that (1) can be transformed to a simpler but equivalent form. Since $\boldsymbol{\beta}_0$ is perfectly known, by setting $y_n = z_n - \boldsymbol{\beta}_0^T \mathbf{x}_n$, we can obtain

$$y_n = \begin{cases} \mathbf{0}^T \mathbf{x}_n + \epsilon_n & n < t \\ \mathbf{a}^T \mathbf{x}_n + \epsilon_n & n \geq t \end{cases}, \tag{2}$$

in which $\mathbf{a} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$. In the remainder of the paper, we will focus on the simplified model (2).

Let $\mathbf{x}_n = [x_{1,n}, x_{2,n}, \ldots, x_{p,n}]^T$. We assume that $\{\mathbf{x}_n\}$ satisfy the following regularity condition:

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=k+1}^{k+m} \mathbf{x}_n \mathbf{x}_n^T = \mathbf{R} \quad \text{uniformly for all } k \geq 0, \tag{3}$$

where $\mathbf{R}$ is a $p \times p$ positive definite matrix. This uniform convergence assumption entails the fact that for any $\mathbf{a} \neq \mathbf{0}$, there exists a common upper bound for $\mathbf{a}^T \mathbf{x}_n$, $n = 1, 2, \ldots$. This fact will be used several times in the detailed proofs.

It is of interest to consider the case that the abrupt change only modifies a few components in the linear coefficient. Hence, we assume that $\mathbf{a}$ is an $s-$sparse vector, i.e., $\mathbf{a}$ only contains at most s non-zero components. s is assumed to be known to the observer.

Let $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$. We specify the feasible set of $\mathbf{a}$, denoted as $\mathcal{A}$, in an element-wise manner. Particularly, if the i^{th} element of the linear coefficient is modified by the abrupt change, then a_i falls in the set:

$$\mathcal{A}_i = \{a_i | a_i \in (-\infty, -b_i] \cup [b_i, \infty)\}, \tag{4}$$

in which $b_i > 0$. Furthermore, denote $\bar{\mathcal{A}}_i := \{a_i = 0\}$ and let

$$\mathcal{A}_k = \bigcup_{(i_1,\ldots,i_p)\in\mathcal{P}} (\mathcal{A}_{i_1} \cup \cdots \cup \mathcal{A}_{i_k} \cup \bar{\mathcal{A}}_{i_{k+1}} \cup \cdots \cup \bar{\mathcal{A}}_{i_p}), \tag{5}$$

where $\mathcal{P}$ consists of all permutations of set $\{1, 2, \ldots, p\}$. Hence $\mathcal{A}_k$ contains all vectors with exactly k-nonzero components. Then, the feasible set of $\mathbf{a}$ is expressed as $\mathcal{A} = \bigcup_{1\le k\le s} \mathcal{A}_k$. We note that $\mathbf{a} = \mathbf{0}$ is excluded from $\mathcal{A}$.

The observer aims to detect the change-point t via the observation sequence $\{y_n\}$ and the deterministic explanatory vector sequence $\{\mathbf{x}_n\}$. Let τ be the stopping time when the observer declares that a change has occurred. The goal of the observer is to, intuitively speaking, minimize the detection delay $(\tau - t)^+$ while keeping the false alarm $\{\tau < t\}$ under control. Two formal mathematical formulations, based on different assumptions on the change-point t, are considered in this paper.

In the non-Bayesian formulation, t is assumed to be a fixed but unknown number. The detection problem is formulated as

$$\begin{aligned} &\text{minimize}_\tau \quad \text{WADD}(\tau; \mathbf{a}) := \\ &\quad \sup_{t\ge 1} \operatorname{esssup} \mathbb{E}_t^{\mathbf{a}}[(\tau - t + 1)^+ | \mathcal{F}_{t-1}], \quad \text{for all } \mathbf{a} \in \mathcal{A} \\ &\text{subject to} \quad \text{ARL2FA}(\tau) := \mathbb{E}_\infty[\tau] \ge \gamma, \end{aligned} \tag{6}$$

in which $\mathbb{E}_t^{\mathbf{a}}$ is the expectation with respect to $P_t^{\mathbf{a}}$, and $P_t^{\mathbf{a}}$ is the probability measure of the observations when the change occurs at t with the post change linear coefficient being $\mathbf{a}$, $\mathbb{E}_\infty$ is the expectation under the probability measure that change never happens ($t = \infty$), and $\mathcal{F}_{t-1}$ is the sigma field generated by $\{y_n\}_{n=1}^{t-1}$. (6) is known as Lorden's formulation [15], which is a min-max setting aiming to minimize the worst case average detection delay over both change-point t and observations up to $t-1$. $\mathbb{E}_\infty[\tau]$ is termed as the average run length to false alarm. Since no change happens in the event $\{t = \infty\}$, the declaration at τ is a false alarm. Hence the constraint in (6) requires that the expected duration to a false alarm is no less than γ.

In the Bayesian formulation, t is modeled as a geometrically distributed random variable. Particularly, we assume that

$$P(t = m) = \begin{cases} \pi_0, & m = 0, \\ (1 - \pi_0)\rho(1 - \rho)^{m-1}, & m = 1, 2, \ldots, \end{cases} \tag{7}$$

in which $\pi_0, \rho \in (0, 1)$ are known parameters. Define probability measure $P_\pi^{\mathbf{a}}$ for a measurable event F as

$$\begin{aligned} P_\pi^{\mathbf{a}}(F) &:= \sum_{m=0}^{\infty} P_t^{\mathbf{a}}(F | t = m) P(t = m) \\ &= \sum_{m=0}^{\infty} P_m^{\mathbf{a}}(F) P(t = m). \end{aligned} \tag{8}$$

The problem under the Bayesian framework is then formulated as

$$\begin{aligned} &\text{minimize}_\tau \quad \text{ADD}(\tau; \mathbf{a}) := \mathbb{E}_\pi^{\mathbf{a}}[\tau - t | \tau \ge t], \text{ for all } \mathbf{a} \in \mathcal{A}. \\ &\text{subject to} \quad \text{PFA}(\tau) := P_\infty(\tau < t) \le \alpha, \end{aligned} \tag{9}$$

in which $\mathbb{E}_\pi^{\mathbf{a}}$ is the expectation with respect to $P_\pi^{\mathbf{a}}$, and P_∞ is the probability measure that the change never happens. Note that on the event $\{\tau < t\}$, all observations are generated by the pre-change distribution; hence the false alarm probability does not depend on the post-change coefficient $\mathbf{a}$. Therefore, (9) aims to minimize the average detection delay while keeping the probability of false alarm under control.

We note that both (6) and (9) are multi-objective optimization problems. Optimal solutions for these two proposed problems are generally difficult to obtain. Hence, in this paper, we aim to propose low complexity algorithms and to analyze their performances.

III. The Parallel-Sum Algorithm

A. Challenges of Existing Methods

Let $f_0(y_n)$ and $f_1(y_n; \mathbf{a})$ denote the pre-change distribution and the post-change distribution of y_n, respectively. We note that $\{y_n, n < t\}$ are i.i.d. random variables with distribution $\mathcal{N}(0, \sigma^2)$, and $\{y_n, n \ge t\}$ are independent variables with distribution $\mathcal{N}(\mathbf{a}^T\mathbf{x}_n, \sigma^2)$. Hence, the likelihood ratio (LR) is

$$L(y_n; \mathbf{a}) = \frac{f_1(y_n; \mathbf{a})}{f_0(y_n)} = \exp\left\{\mathbf{a}^T\mathbf{x}_n y_n - \frac{1}{2}\mathbf{a}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{a}\right\}. \tag{10}$$

We further denote

$$\mu_{\mathbf{a}} := \frac{1}{2}\mathbf{a}^T\mathbf{R}\mathbf{a} = \frac{1}{2}\sum_{i=1}^{p} a_i^2 r_{i,i} + \frac{1}{2}\sum_{i\ne j} a_i a_j r_{i,j}, \tag{11}$$

in which $r_{i,j}$ is the element located at the i^{th} row and the j^{th} column in $\mathbf{R}$.

For the classic quickest change-point detection problem, it is well known that the CUSUM procedure is optimal for Lorden's formulation and the Shiryaev procedure is optimal for the Bayesian formulation. Particularly, for a given $\mathbf{a}$, the CUSUM statistic can be written as

$$C_n(\mathbf{a}) = \max_{1\le m\le n} \prod_{k=m}^{n} L(y_k; \mathbf{a}), \tag{12}$$

and the Shiryaev statistic can be written as

$$R_{n,\rho}(\mathbf{a}) = \frac{\pi_0}{(1 - \pi_0)\rho}\prod_{k=1}^{n}\frac{L(y_k; \mathbf{a})}{1 - \rho} + \sum_{m=1}^{n}\prod_{k=m}^{n}\frac{L(y_k; \mathbf{a})}{1 - \rho}. \tag{13}$$

In our problem, the observer does not know the true value of $\mathbf{a}$. Hence, it is natural to consider the GLR based detection procedures. More specifically, GLR-CUSUM procedure can be written as

$$\tau_C^{(GLR)} := \min\left\{n \geq 0 : \sup_{\mathbf{a}\in\mathcal{A}} C_n(\mathbf{a}) \geq B\right\}, \tag{14}$$

and the GLR-Shiryaev procedure can be written as

$$\tau_R^{(GLR)} := \min\left\{n \geq 0 : \sup_{\mathbf{a}\in\mathcal{A}} R_{n,\rho}(\mathbf{a}) \geq B\right\}. \tag{15}$$

It is easy to see that, for both (14) and (15), the observer has to estimate $\mathbf{a}$ by solving $\sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} L(y_k;\mathbf{a})$ for each $m \in \{1,\ldots,n\}$[1], which is equivalent to solving

$$\inf_{\mathbf{a}\in\mathcal{A}} \sum_{k=m}^{n} (y_k - \mathbf{a}^T\mathbf{x}_k)^2 \quad \text{for } m = 1,\ldots,n. \tag{16}$$

The challenges of solving this problem include

- (16) is a non-convex problem because the feasible set $\mathcal{A}$ is non-convex. It is known that to find an $s-$sparse solution of an underdetermined system is NP hard.
- One may consider using the popular l_1-relaxation techniques, such as LASSO, to solve for the $s-$sparse solution. However, l_1-relaxation techniques cannot guarantee to find the optimal solution of (16) for all $m = 1,\ldots,n$. Specifically, when m is close to n, e.g. $n - m \sim o(s\log p)$, the observer does not have enough samples for a successful recovery [35]–[37].
- Even if the LASSO algorithm could work in solving (16), its computational complexity is high. For example, if LASSO is implemented using LARS algorithms, the computation complexity of recovering a $p-$dimensional $s-$sparse vector using $n-m$ measurements is $O(p^3+(n-m)p^2)$ [38]. Hence, the total computational complexity of solving (16) is $O(np^3+n^2p^2)$. Note that the computational complexity increases dramatically when the observer obtains more samples.

[22] proposes another GLR based detection procedure and shows that it is asymptotically optimal for Lorden's QCD problem. In that algorithm, the observer has to calculate the detection statistic by solving

$$\sup_{||\mathbf{a}_{m:n}||\geq\theta} \left(\mathbf{a}^T\mathbf{X}_{m:n}^T\mathbf{Y}_{m:n} - \frac{1}{2}\mathbf{a}^T\mathbf{X}_{m:n}^T\mathbf{X}_{m:n}\mathbf{a}\right) \quad \text{for } m = 1,\ldots,n, \tag{17}$$

in which $\mathbf{a}_{m:n} = (n-m+1)^{-1}(\mathbf{X}_{m:n}^T\mathbf{X}_{m:n})^{1/2}\mathbf{a}$, $\mathbf{X}_{m:n} = [\mathbf{x}_m,\ldots,\mathbf{x}_n]^T$ and $\mathbf{Y}_{m:n} = [\mathbf{y}_m,\ldots,\mathbf{y}_n]^T$. Though (17) is a convex optimization problem, considering the calculation of matrix multiplication and singular value decomposition, the computational complexity of solving (17) is at least $O(np^3 + n^2p^2)$ at time slot n.

Using mixture based detection procedures is another way to deal with the post-change uncertainty. When the post-change distribution contains unknowns, it has been shown that the mixture based Shiryaev procedure is asymptotically optimal for the exponential family under Bayesian QCD, the mixture based CUSUM procedure and the mixture based SR procedure are asymptotically optimal for the non-Bayesian QCD under some regularity conditions. One can refer to Section 7.5, Section 8.3 and Section 8.5 in [27] and references therein for more details. Let $g(\mathbf{a})$ be a prior distribution for $\mathbf{a}$. In our context, the mixture CUSUM procedure can be written as

$$\tau_C^{(mix)} := \min\left\{n \geq 0 : \int_{\mathcal{A}} C_n(\mathbf{a})g(\mathbf{a})d\mathbf{a} \geq B\right\}, \tag{18}$$

and the mixture Shiryaev procedure can be written as

$$\tau_R^{(mix)} := \min\left\{n \geq 0 : \int_{\mathcal{A}} R_{n,\rho}(\mathbf{a})g(\mathbf{a})d\mathbf{a} \geq B\right\}. \tag{19}$$

The challenges of mixture base detection procedures include: 1) they require prior distribution $g(\mathbf{a})$, which may not be available in some applications; 2) they usually involve high computational complexity since the integral is difficult to calculate especially when $\mathbf{a}$ has a complicated multivariate pdf.

Finally, we note that the asymptotic lower bound of the detection delay for non-Bayesian formulation has been established in several existing works [22], [23]. The detection delay is lower bounded by $|\log\gamma|/\mu_{\mathbf{a}}(1+o(1))$. In this paper, we are interested in finding low complexity algorithms (which may be sub-optimal) to avoid the huge calculation burden involved in both GLR and mixture type detection procedure.

B. Parallel-Sum Algorithm for the Non-Bayesian Setup

In this subsection, we propose a low complexity algorithm, termed as parallel-sum algorithm, for Lorden's formulation. The proposed detection procedure is described as follows:

$$W_i(m,n;a_i) := \kappa a_i \sum_{k=m}^{n} x_{i,k}y_k - \frac{\kappa}{2}a_i^2\sum_{k=m}^{n} x_{i,k}^2, \quad \text{for } 1 \leq i \leq p, \tag{20}$$

$$U(m,n) := \sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(m,n;a_i), \tag{21}$$

$$S_n := \sup_{1\leq m\leq n} U(m,n), \tag{22}$$

$$\tau_S := \inf\{n \geq 0 : S_n \geq \log B\}, \tag{23}$$

in which κ is a constant, and B is a properly selected threshold to control ARL2FA.

The main idea of the proposed parallel-sum algorithm is to use the correlation between y_k and $\mathbf{x}_k$ to detect the change-point. From (2), we see that y_k does not depend on $x_{i,k}$ before the change as the linear coefficients are $\mathbf{0}$. After the change, y_k depends on $x_{i,k}$ if $a_i \neq 0$. Furthermore, a_i reflects the correlation strength between y_k and $x_{i,k}$. Actually, $W_i(m,n;a_i)$ defined in (20) is a measurement of the correlation between y_k and $x_{i,k}$. In particular, on the event $\{t = m\}$, $W_i(m,n;a_i)$ is close to zero if the i^{th} component in $\mathbf{a}$ is unchanged and tends to be positive if changed. Hence, the observer wants to sum

[1]For Bayesian setting, we use the case $\pi_0 = 0$ to explain the GLR-Shiryeav procedure involves high computational burden.

up all s positive W_i's to speed up the detection procedure. This idea is reflected by $U(m,n)$ in (21). As the change-point t is unknown, the observer then searches over all time instants within $[1,n]$ in (22) and detect the change-point via a threshold rule in (23). This follows a similar idea of constructing the CUSUM procedure from the one-side SPRT procedure [15], [22].

The performance of the proposed parallel-sum algorithm is presented in the following theorem.

Theorem III.1. *By setting*

$$\log B = 2\kappa s\left[\log\left(2p + p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{\max}]\right) + \log\gamma\right], \quad (24)$$

in which N_{max} is a finite random variable, and $\mathbb{E}[N_{\max}] \le c_1 p$ with c_1 being a constant independent of p. One can guarantee that

$$\mathit{ARL2FA}[\tau_S] \ge \gamma. \quad (25)$$

Furthermore, the detection delay is bounded by

$$\mathit{WADD}(\tau_S;\mathbf{a}) \le \frac{2|\log B|}{\kappa\sum_{i=1}^p a_i^2 r_{i,i} + 2\kappa\sum_{i\neq j} a_i a_j r_{i,j}}(1+o(1)) \quad (26)$$

as $\gamma\to\infty$.

Proof: Please see Section A. ∎

Remark III.2. *1) In the asymptotic analysis when p, s are constants and $\gamma\to\infty$, i.e., roughly speaking, the observer has infinitely many (compared with dimension p) post-change observations to detect the change-point, we have $\log B = 2\kappa s\log\gamma(1+o(1))$ and*

$$\frac{\mathit{WADD}(\tau_S;\mathbf{a})}{\inf_\tau \mathit{WADD}(\tau;\mathbf{a})} \le 2s\frac{\sum_{i=1}^p a_i^2 r_{i,i} + \sum_{i\neq j} a_i a_j r_{i,j}}{\sum_{i=1}^p a_i^2 r_{i,i} + 2\sum_{i\neq j} a_i a_j r_{i,j}}.$$

Hence, when the components in $\mathbf{x}_n$ are mutually uncorrelated, i.e., $\mathbf{R}$ is a diagonal matrix, the performance loss of the proposed algorithm is no more than $2s$.
2) In the high dimension setting when $p\to\infty$, $s\to\infty$, $\gamma\to\infty$ and $\gamma/p\to c$ (c is constant that could be zero), we have $\log B \sim O(s\log p)$. Note that the denominator in (26) is on the order of $O(s)$ (since there are only s non-zero components in $\mathbf{a}$); hence the detection delay $\mathit{WADD}(\tau_S;\mathbf{a}) \sim O(\log p)$. That is, the observer only needs $O(\log p)$ post-change observations on average to detect the change-point. Recall that in the sparse recovery problem, one needs $O(s\log p)$ observations to recover an $s-$sparse vector. Here, we require fewer observations for the purpose of detection.
3) From (24) and (26), we note that the constant κ does not affect the upperbound of WADD in the non-Bayesian case. However, as will be shown in the sequel, κ plays a role in the upperbound of ADD in the Bayesian case.

C. Parallel-Sum Algorithm for the Bayesian Setup

In this subsection, we construct the the parallel-sum algorithm for the Bayesian formulation. Specifically, the proposed detection procedure is described as follows:

$$W_i(m,n;a_i) := \kappa a_i\sum_{k=m}^n x_{i,k}y_k - \frac{\kappa}{2}a_i^2\sum_{k=m}^n x_{i,k}^2, \quad \text{for } 1\le i\le p, \quad (27)$$

$$V_i(m,n;a_i) := W_i(m,n;a_i) + (n-m+1)\mu, \quad (28)$$

$$U(m,n) := \sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^p V_i(m,n;a_i) = \sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^p W_i(m,n;a_i) + p(n-m+1)\mu, \quad (29)$$

$$S_n := \sup_{1\le m\le n} U(m,n), \quad (30)$$

$$\tau_S := \inf\{n\ge 0 : S_n \ge \log B\}. \quad (31)$$

With a little abuse of notations, we still use $U(m,n)$, S_n and τ_S in the Bayesian case to denote the detection procedure. However, these notations can be clearly distinguished from the ones for the non-Bayesian setting in a given context. Similar to the non-Bayesian case, the parallel-sum algorithm for the Bayesian formulation also explores the correlated information between y_k and $\mathbf{x}_k$ for the purpose of change-point detection. However, the proposed algorithm in the Bayesian case contains one more parameter μ in (28), which can be designed by the observer to speed up the detection procedure by exploring the prior knowledge of the change-point.

The analysis of the proposed parallel-sum algorithm requires some additional mild assumptions. In particular, let

$$\xi_k := \kappa y_k\sum_{i=1}^p a_i x_{i,k} - \frac{\kappa}{2}\sum_{i=1}^p (a_i x_{i,k})^2 + p\mu.$$

$$I_\xi := \sum_{i=1}^p\left[\frac{\kappa}{2}a_i^2 r_{ii} + \kappa\sum_{i\neq j} a_i a_j r_{ij}\right] + p\mu.$$

As will be shown in the appendix, on the event $\{t=m\}$,

$$\frac{1}{n}\sum_{k=m}^{m+n-1}\xi_k \overset{a.s.}{\to} I_\xi.$$

Define

$$T_\delta := \inf\left\{n\ge 0 : \left|n^{-1}\sum_{k=m}^{m+n-1}\xi_k - I_\xi\right| > \delta\right\}, \quad (32)$$

hence $T_\delta<\infty$ almost surely. We further assume that

$$\mathbb{E}_m^{\mathbf{a}}[T_\delta] < \infty \text{ and } \mathbb{E}_\pi^{\mathbf{a}}[T_\delta] < \infty \text{ for all } \mathbf{a}\in\mathcal{A}. \quad (33)$$

With Assumption (33), we have the following result:

Theorem III.3. *Let*

$$c_2 = (1-\kappa)^{-\frac{1}{2}}\exp\left\{\frac{p}{s}\mu\right\}.$$

By setting

$$\log B = s|\log\alpha| + s\log\frac{\rho c_2}{(c_2-1)[1-(1-\rho)c_2]},$$

and choosing $\kappa < 1$*,*

$$\frac{s}{2p}\log(1-\kappa) \le \mu < \frac{s}{2p}\log(1-\kappa) + \frac{s}{p}|\log(1-\rho)|, \quad (34)$$

one can guarantee that

$$PFA[\tau_S] \le \alpha \quad (35)$$

for all $\mathbf{a} \in \mathcal{A}$*. Furthermore, the average detection delay is bounded by*

$$\begin{aligned} ADD(\tau_S;\mathbf{a}) &\le \mathbb{E}_\pi^{\mathbf{a}}[\tau_S - t|\tau_S \ge t] \\ &= \frac{2|\log B| + 2c_3 s}{\kappa\sum_{i=1}^p a_i^2 r_{ii} + 2\kappa\sum_{i\ne j} a_i a_j r_{ij} + 2p\mu}(1+o(1)) \quad (36) \end{aligned}$$

as $\alpha \to 0$*, in which* c_3 *is a constant independent of* p*.*

Proof: Please see Section B. ∎

Remark III.4. *1) In the asymptotic analysis when* p*,* s *are constants and* $\alpha \to 0$*, it is easy to see that* (34) *is satisfied if we choose*

$$\mu = \frac{s}{2p}\log(1-\kappa) + \frac{s}{p}\log\frac{1-\alpha}{1-\rho}.$$

With this selection, we have $c_2 = (1-\alpha)/(1-\rho)$ *and hence* $\log B = 2s|\log\alpha|(1+o(1))$ *as* $\alpha \to 0$*. Correspondingly,*

$$ADD(\tau_S;\mathbf{a}) \le \frac{4s|\log\alpha|}{\vartheta(\mathbf{a},\mathbf{R},\kappa)}(1+o(1)),$$

in which

$$\begin{aligned} \vartheta(\mathbf{a},\mathbf{R},\kappa) = \kappa\sum_{i=1}^p a_i^2 r_{ii} + 2\kappa\sum_{i\ne j} a_i a_j r_{ij} \\ +2s|\log(1-\rho)| + s\log(1-\kappa). \end{aligned}$$

By adjusting the value of κ*, we can obtain a family of upper bounds for the detection delay. In this case, we have* $ADD(\tau_S;\mathbf{a}) \sim O(|\log\alpha|)$ *for all* $\mathbf{a} \in \mathcal{A}$*.*

2) In the high dimension setting when $p \to \infty$*,* $s \to \infty$*,* $\alpha \to 0$ *and* $p\alpha \to c$ *(*c *is a constant that could also be infinity), it is easy to see that* (34) *is satisfied if we choose*

$$\mu = \frac{s}{2p}\log(1-\kappa) + \frac{s}{p}\log\frac{1-p^{-1}}{1-\rho},$$

we then have $\log B \sim O(s\log p)$*. Since the denominator in* (36) *is on the order of* $O(s)$*, the detection delay* $ADD(\tau_S;\mathbf{a}) \sim O(\log p)$*. Hence, similar to the conclusion obtained in the non-Bayesian case, we require less observations for the purpose of online change-point detection than that for the sparse recovery.*

D. Implementation and Discussion

The proposed parallel-sum algorithm can be easily implemented. From (21) and (29), the main calculation of the parallel-sum algorithm, for both non-Bayesian and Bayesian cases, is to solve the optimization problem

$$\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^p W_i(m,n;a_i). \quad (37)$$

By solving $\frac{\partial}{\partial a_i}W_i(m,n;a_i) = 0$, we can easily obtain that $W_i(m,n;a_i)$ achieves its maximum at

$$a_i^* = \frac{\sum_{k=m}^n x_{i,k}y_k}{\sum_{k=m}^n x_{i,k}^2}. \quad (38)$$

Let

$$\hat{a}_i := \arg\max_{a_i\in\mathcal{A}_i} W_i(m,n;a_i). \quad (39)$$

It is easy to see that $\hat{a}_i = a_i^*$ if $a_i^* \in \mathcal{A}_i$, otherwise $\hat{a}_i$ equals to one of the two candidates $\{-b_i, b_i\}$.

Let $\hat{\mathbf{a}}^* = [\hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_p^*]^T$ be the optimal solution for (37). Denote the order statistics of $\{W_i(m,n;\hat{a}_i), i = 1,\ldots,p\}$ as

$$\begin{aligned} W_{(1)}(m,n;\hat{a}_{(1)}) \ge W_{(2)}(m,n;\hat{a}_{(2)}) \ge \\ \cdots \ge W_{(p)}(m,n;\hat{a}_{(p)}). \end{aligned} \quad (40)$$

It is easy to see that the optimal estimation $\hat{\mathbf{a}}^*$ is given as

$$\hat{a}_i^* = \begin{cases} \hat{a}_i & \text{if } W_i(m,n;\hat{a}_i) \ge W_{(s)}(m,n;\hat{a}_{(s)}) \\ 0 & \text{otherwise} \end{cases}. \quad (41)$$

As a result, for the non-Bayesian case, we have

$$U(m,n) = \sum_{i=1}^s W_{(i)}(m,n;\hat{a}_{(i)}), \quad (42)$$

and for the Bayesian case

$$U(m,n) = \sum_{i=1}^s W_{(i)}(m,n;\hat{a}_{(i)}) + p(n-m+1)\mu. \quad (43)$$

We now discuss the computational complexity of the proposed algorithm. The main computation of the parallel-sum algorithm consists of four parts: 1) Calculating $W_i(m,n;\hat{a}_i)$ for $m = 1,\ldots,n$. Since $\sum_{k=m}^n x_{i,k}y_k$ and $\sum_{k=m}^n x_{i,k}^2$ can be calculated recursively for adjacent values of m, the computational complexity of calculating $\{W_i(m,n;\hat{a}_i), m = 1,\ldots,n\}$ is on the same level of calculating $W_i(1,n;\hat{a}_i)$, which is on the level of $O(n)$. As the observer has to find W_i's for $i = 1,\ldots,p$, the total amount of computation in this part is $O(np)$; 2) Finding $\{W_{(i)}(m,n;\hat{a}_{(i)}), i = 1,\ldots,s\}$ for $m = 1,\ldots,n$. The computational complexity of searching the s^{th} largest number from a group of p numbers is known as $O(p)$, hence the total computational amount in this step is also $O(np)$; 3) Calculating $U(m,n)$ for $m = 1,\ldots,n$. The amount of calculation is $O(ns)$ in this step. 4) Calculating S_n. The computational complexity of finding the largest number from n numbers is $O(n)$. As a result, the computational complex of proposed algorithm at time slot n is $O(np)$.

One may notice that the computational complexity increases as n increases; hence the amount of computation explodes when $n \to \infty$. For implementation purpose, one can truncate the proposed algorithm by a window with length l_w. Specifically, one can modify S_n defined in (23) and (31) as

$$S_n := \sup_{n-l_w+1 \le m \le n} U(m,n). \tag{44}$$

With this modification, the computational complexity will be limited to $O(l_w p)$ for each time slot. In practice, we should set l_w slightly larger than the average detection delay. For example, in the high-dimensional case, the detection delay is $O(\log p)$ as we have discussed in Remark III.2 and Remark III.4; hence we can set, for instance, $l_w = O(\log p \log\log p)$. Then, Markov's inequality indicates that $P_m^{\mathbf{a}}(\tau - m > l_w) \to 0$ as $p \to \infty$. In this paper, we do not focus on the window-based parallel-sum algorithm, hence the rigorous analysis of the window-based algorithm is left for future work.

The parallel-sum algorithm reduces the computational complexity at the expense of asymptotic optimality. Denote $L(y_n, a_i)$ as the LR of the component-wise hypothesis testing $H_0^{(i)} : y_n = \epsilon_n$ versus $H_1^{(i)} : y_n = a_i x_{i,n} + \epsilon_n$. Then, for both non-Bayesian and Bayesian setup, one can easily verify that

$$W_i(m,n;a_i) = \kappa \sum_{k=m}^{n} \log L(y_k, a_i).$$

Hence, the parallel algorithm is based on the statistic of component-wise hypothesis testing rather than on the statistic (10) of the original model $H_0 : y_n = \epsilon_n$ versus $H_1 : y_n = \mathbf{a}^T \mathbf{x}_n + \epsilon_n$. On the one hand, the proposed algorithm significantly reduces the computational complexity as the component-wise hypothesis testing only considers one unknown parameter each time. On the other hand, the proposed algorithm is not asymptotic optimal anymore because of the model mismatch.

IV. Numerical Examples

In this section, we provide numerical examples to illustrate the theoretic results obtained in our paper. In the first numerical example, we assume that $p = 15$ and $s = 3$, the post-change linear coefficient $\mathbf{a}$ is given as $a_1 = 0.8$, $a_2 = 0.65$, $a_3 = 0.5$, and $a_i = 0$ for the rest of components in $\mathbf{a}$. We set $\mathcal{A}_i = [0.4, 2.5]$ for all $i \in \{1, \ldots, p\}$. $\mathbf{R}$, the covariance matrix of $\mathbf{x}_n$, is randomly selected as $\mathbf{R} = \mathrm{diag}[1.32, 1.18, 1.04, 0.93, 0.86, 0.84, 0.71, 0.64, 0.52, 0.42, 0.39, 0.28, 0.17, 0.14, 0.03]$. In the simulation, the explanatory variable $\mathbf{x}_n$ are generated independently by an underlying distribution; hence the uniform convergence assumption (3) is satisfied. Particularly, two underlying distributions, namely the Gaussian distribution with zero mean and the Possion distribution with its expectation shifted to zero, are tested in the simulation.

Figure 1 illustrates the performance of the proposed parallel-sum algorithm under the non-Bayesian setting. In particular, the blue line with squares is the performance of the parallel-sum algorithm when $\mathbf{x}_n$'s are independent and identically Gaussian distributed, and the green line with diamonds is the performance when $\mathbf{x}_n$'s are independent and identically Possion distributed. The black dot-dash line is calculated as $|\log\gamma|/\mu_{\mathbf{a}}$, which is the lower bound of WADD presented in Theorem 1 [22]. The black dash-line is the upper bound of the parallel-sum algorithm, which is presented in Theorem III.1. Figure 1 presents the relationship between WADD and ARL2FA for the proposed parallel-sum algorithm. From the simulation, we can see that the parallel-sum algorithm is not asymptotically optimal since it diverges from the lower bound as γ increases. However, we note that the detection delay of the parallel-sum algorithm still increases only linearly with $\log\gamma$, and the computation complexity of this algorithm is low.

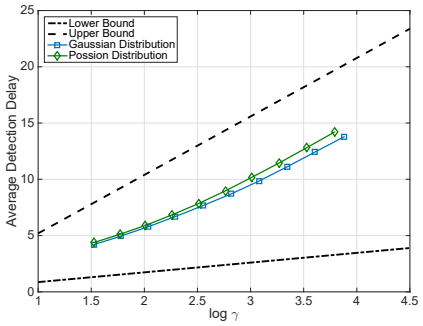


Fig. 1. WADD versus ARL2FA when $p = 15$, $s = 3$

Figure 2 illustrates the relationship between ADD and PFA for the proposed parallel-sum algorithm under the Bayesian setting. In this simulation, we set $\rho = 0.2$ and we choose $\kappa = 0.35$, $\mu = 0.0014$. The performance result is similar to the one obtained in the non-Bayesian simulation. In particular, the performances under Gaussian distribution and Possion distribution are close to each other, which verifies our theoretical results that the asymptotic performance is independent of the underlying distribution $\mathbf{x}_n$. In addition, the performance of the proposed algorithm diverges from the lower bound hence it is not asymptotically optimal, but the detection delay is still on the order of $|\log\alpha|$ as the performance is upper bounded by the result in Theorem III.3. The computational complexity of the proposed algorithm is also low.

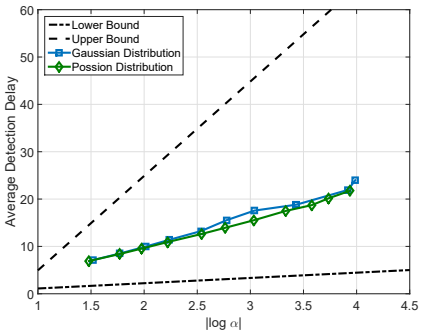


Fig. 2. ADD vs. PFA under when $p = 15$, $s = 3$

TABLE I
PERFORMANCE OF THE PARALLEL-SUM ALGORITHM UNDER DIFFERENT THRESHOLDS

$\log B$	approximated ARL2FA	change declaration time	change-point	detection delay
84.67	10	45	60	false alarm
105.39	10^2	72	60	12
126.11	10^3	81	60	21
146.84	10^4	87	60	27

Finally, we test our proposed algorithm on a real dataset, which is published on the webpage of the Center for Machine Learning and Intelligent Systems at University of California, Irvine [39]. This data set is comprised of measured EMG signals for six different kinds of hand movements of different persons. Specifically, each kind of hand movements is repeated and measured 30 times, and each time the signal is recorded by a 2-channel EMG system; hence each person has $30 \times 2 \times 6 = 360$ different measurements. In data processing, we use the last measurement as dependent variable y_n and the rest of measurements as $\mathbf{x}_n$; hence $p = 359$ in this numerical example. We then concatenate two different person's data to model the change-point. 60 samples for each person are selected; hence the real change-point is located at $t = 60$ and the total time duration is 120. Since the change-point is fixed (but unknown to the observer in the simulation), we implement the proposed algorithm for non-Bayesain formulation and select $s = 9$ in our simulation. The evolution of the detection statistic S_n over time is shown in Figure 3. As we can see, S_n tends to increase for $n > 60$. The performance under different threshold $\log B$ is listed in Table I, which shows the efficiency of the proposed algorithm.

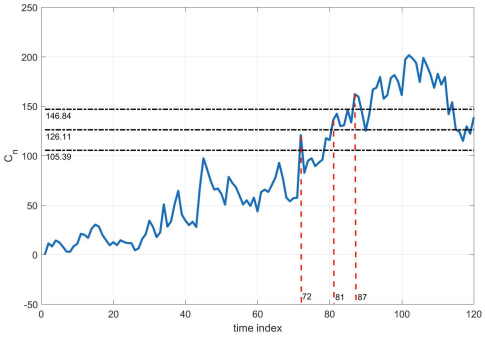


Fig. 3. The evolution of statics S_n over time slot

V. CONCLUSION

In this paper, we have considered the problem of quickly detecting an abrupt change in the linear model. Both non-Bayesian and Bayesian formulations are considered. For each case, we have proposed a low complexity online algorithm. When p and s are fixed, the average detection delay for the proposed strategy is on the order of $O(\log \gamma)$ for the non-Bayesian formulation as $\gamma \to \infty$ and is on the order of $O(|\log \alpha|)$ for the Bayesian formulation as $\alpha \to 0$. When $p \to \infty$, the average detection delay of the proposed algorithm has been shown to be upper bounded by $O(\log p)$ for both non-Bayesian and Bayesian formulations.

APPENDIX A
PROOF OF THEOREM III.1

We prove Theorem III.1 by exploring the relationship between Lorden's QCD problem and the one-sided SPRT problem.

Consider a sequential hypothesis testing problem that $\{(\mathbf{x}_n, y_n)\}_{n=1}^{\infty}$ obeys one of the following hypothesis:

$$H_0 : y_n = \mathbf{0}^T \mathbf{x}_n + \epsilon_n \text{ versus } H_1 : y_n = \mathbf{a}^T \mathbf{x}_n + \epsilon_n. \quad (45)$$

Denote $P_\infty(\cdot)$ and $P^{\mathbf{a}}(\cdot)$ as probability measures under H_0 and H_1, respectively. In the one-sided SPRT problem, the observer wants to take as many (even infinitely many) observations as possible when H_0 is true, and wants to take as few observations as possible when H_1 is true. Specifically, a testing procedure can be defined as a stopping time τ. $\{\tau = n\}$ indicates the number of observations taken by the observer when he claims H_1 to be true. $\{\tau = \infty\}$ is the event that the procedure takes infinitely many observations. For a given $\mathbf{a} \in \mathcal{A}$, the one-sided SPRT problem aims to solve

$$\begin{aligned} &\text{minimize}_\tau \ \ \mathbb{E}^{\mathbf{a}}[\tau], \\ &\text{subject to } P_\infty(\tau < \infty) \le \alpha. \end{aligned} \quad (46)$$

The relationship between one-sided SPRT and Lorden's QCD problem is firstly revealed by Lorden in [15]. [22] extends this relationship to the linear model. We rewrite this well known result in our context as the following lemma.

Lemma A.1. *(Proposition 1 in [22]) Suppose τ_1 is a stopping time for one-sided SPRT problem with respect to $\{(\mathbf{x}_n, y_n)\}_{n=1}^{\infty}$ such that*

$$P_\infty(\tau_1 < \infty) \le \alpha, \quad 0 < \alpha < 1. \quad (47)$$

For each $k = 1, 2, \ldots,$ let τ_k denote the stopping time obtained by applying τ_1 to $\{(\mathbf{x}_n, y_n)\}_{n=k}^{\infty}$ and define

$$\tau^* = \inf\{\tau_k + k - 1 | k = 1, 2, \ldots\}. \quad (48)$$

Then τ^* is also a stopping time, and for the problem formulation defined in (6) *it satisfies*

$$\mathit{ARL2FA}(\tau^*) \ge \frac{1}{\alpha} \quad (49)$$

and

$$\mathit{WADD}(\tau^*; \mathbf{a}) \le \mathbb{E}^{\mathbf{a}}[\tau]. \quad (50)$$

Using this lemma, we will study the performance of following algorithm for the one-sided SPRT problem (46):

$$\begin{aligned} &W_i(1, n; a_i) = \kappa a_i \sum_{k=1}^{n} x_{i,k} y_k - \frac{\kappa}{2} a_i^2 \sum_{k=1}^{n} x_{i,k}^2, \\ &U_n = \sup_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^{p} W_i(1, n; a_i), \\ &\tau_1 = \inf\{n \ge 0 : U_n \ge \log B\}. \end{aligned} \quad (51)$$

Note that τ_S in (23) can be equivalently written as $\tau_S = \inf\{\tau_k + k - 1 | k = 1, 2, \ldots\}$. Due to Lemma A.1, it suffices to study the performance of $\mathbb{E}^{\mathbf{a}}[\tau_1]$ and $P_\infty(\tau_1 < \infty)$ in (46).

Lemma A.2. *(Detection delay) For a given threshold B, as $B \to \infty$ we have*

$$\mathbb{E}^{\mathbf{a}}[\tau_1] \leq \frac{2|\log B|}{\kappa \sum_{i=1}^p a_i^2 r_{i,i} + 2\kappa \sum_{i \neq j} a_i a_j r_{i,j}} (1 + o(1)). \quad (52)$$

for any $\mathbf{a} \in \mathcal{A}$.

Lemma A.3. *(False alarm probability) For a given threshold B, the error probability of τ_1 is given as*

$$P_\infty(\tau_1 < \infty) \leq 2pB^{-\frac{1}{\kappa s}} + \left(p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{max}]\right)(\log B)^{-\frac{1}{4}} B^{-\frac{1}{2\kappa s}}, \quad (53)$$

in which N_{max} is a finite random variable and $\mathbb{E}[N_{max}] \leq c_1 p$, where c_1 is a constant independent of p.

With these two lemmas, by setting

$$\log B = 2\kappa s \left[\log\left(2p + p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{\max}]\right) + \log\gamma\right],$$

we have

$$\begin{aligned} &P_\infty(\tau_1 < \infty) \\ &\leq 2pB^{-\frac{1}{\kappa s}} + \left(p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{max}]\right)(\log B)^{-\frac{1}{4}} B^{-\frac{1}{2\kappa s}} \\ &\leq \left(2p + p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{max}]\right) B^{-\frac{1}{2\kappa s}} = \frac{1}{\gamma}. \end{aligned}$$

Theorem III.1 then follows by exploring the results (49) and (50). In the rest of this subsection, we will prove Lemma A.2 and Lemma A.3.

A. Proof of Lemma A.2

To support the proof of Lemma A.2, we first rewrite a supporting lemma from [22] in our context.

Lemma A.4. *(Lemma 2 in [22]) Let $\xi_1, \xi_2, \ldots,$ be a sequence of independent random variables. Let $\bar{U}_n = \sum_{k=1}^n \xi_k$. Suppose that $\{\tau_B, B \in (1, \infty)\}$ is a sequence of stopping rules with respect to $\{\xi_n\}$. Suppose further $\mathbb{E}[\tau_B] \to \infty$ as $B \to \infty$. Let $\sup_{k \geq 1} \mathbb{E}[|\xi_k|] < \infty$, and $\frac{1}{n}\sum_{k=1}^n \mathbb{E}[\xi_k] \to \nu$ as $n \to \infty$ with ν finite. Then $\mathbb{E}[\bar{U}_{\tau_B}]/\mathbb{E}[\tau_B] \to \nu$ as $B \to \infty$.*

Lemma A.4 says Wald's identity holds asymptotically when ξ_i's have different expectations but their average value converges to a finite limit.

We then prove Lemma A.2. Assume that a genie knows the true post-change linear coefficient $\mathbf{a}$, and he uses the statistic

$$\bar{U}_n = \sum_{i=1}^p W_i(1, n; a_i) = \sum_{k=1}^n \left[\kappa y_k \sum_{i=1}^p a_i x_{i,k} - \frac{\kappa}{2}\sum_{i=1}^p (a_i x_{i,k})^2\right]$$

for detection. Let $\bar{U}_0 = 0$, define $\xi_k = \bar{U}_k - \bar{U}_{k-1} = \kappa y_k \sum_{i=1}^p a_i x_{i,k} - \frac{\kappa}{2}\sum_{i=1}^p (a_i x_{i,k})^2$, $k = 1, 2, \ldots$. Then, for a deterministic explanatory variable sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$, $\{\xi_k\}$ is a sequence of independent random variables under the alternative hypothesis. In addition, we have

$$\mathbb{E}^{\mathbf{a}}[\xi_k] = \frac{\kappa}{2}\sum_{i=1}^p a_i^2 x_{i,k}^2 + \kappa \sum_{i \neq j} a_i a_j x_{i,k} x_{j,k}.$$

By the assumption of uniform convergence (3), we have

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}^{\mathbf{a}}[\xi_k] \to \frac{\kappa}{2}\sum_{i=1}^p a_i^2 r_{i,i} + \kappa \sum_{i \neq j} a_i a_j r_{i,j}. \quad (54)$$

Further

$$\begin{aligned} \mathbb{E}^{\mathbf{a}}[|\xi_k|] &\leq \kappa \mathbb{E}^{\mathbf{a}}\left[\left|y_n \sum_{i=1}^p a_i x_{i,k}\right|\right] + \frac{\kappa}{2}\mathbb{E}^{\mathbf{a}}\left[\left|\sum_{i=1}^p (a_i x_{i,k})^2\right|\right] \\ &= \kappa \mathbb{E}^{\mathbf{a}}\left[\left|(\mathbf{a}^T\mathbf{x}_k + \epsilon_k)\mathbf{a}^T\mathbf{x}_k\right|\right] + \frac{\kappa}{2}\mathbb{E}^{\mathbf{a}}\left[\left|\sum_{i=1}^p (a_i x_{i,k})^2\right|\right] \\ &\leq \kappa \mathbf{a}^T\mathbf{x}_k\mathbf{x}_k^T\mathbf{a} + \kappa \mathbf{a}^T\mathbf{x}_k \mathbb{E}^{\mathbf{a}}[|\epsilon_k|] + \frac{\kappa}{2}\left|\sum_{i=1}^p (a_i x_{i,k})^2\right| \\ &< \infty, \end{aligned} \quad (55)$$

in which the last step is due to the uniform convergence assumption (3) and $\mathbb{E}^{\mathbf{a}}[|\epsilon_k|] < \infty$. Let $\tau_B = \inf\{n \geq 1 : \bar{U}_n \geq \log B\}$. Using Lemma A.4 and ignoring the overshoot of $\bar{U}_{\tau_B}$, as $B \to \infty$, we have

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}^{\mathbf{a}}[\xi_k] = \frac{\mathbb{E}^{\mathbf{a}}[\bar{U}_{\tau_B}]}{\mathbb{E}^{\mathbf{a}}[\tau_B]} = \frac{|\log B|}{\mathbb{E}^{\mathbf{a}}[\tau_B]}, \quad (56)$$

or equivalently,

$$\mathbb{E}^{\mathbf{a}}[\tau_B] = \frac{2|\log B|}{\kappa \sum_{i=1}^p a_i^2 r_{i,i} + 2\kappa \sum_{i \neq j} a_i a_j r_{i,j}}.$$

We note that $\tau_1 \leq \tau_B$ since $U_n \geq \bar{U}_n$; hence Lemma A.2 holds.

B. Proof of Lemma A.3

In the following, we study the false alarm probability of τ_1 for the one-sided SPRT problem (46). By solving $\frac{\partial}{\partial a_i} W_i(1, n; a_i) = 0$, it is easy to see that

$$a_i^* = \frac{\sum_{k=1}^n x_{i,k}\epsilon_i}{\sum_{k=1}^n x_{i,k}^2} \quad (57)$$

maximizes the value of $W_i(1, n; a_i)$ under P_∞. Specifically,

$$W_i(1, n; a_i^*) = \frac{\kappa}{2}\frac{(\sum_{k=1}^n x_{i,k}\epsilon_k)^2}{\sum_{k=1}^n x_{i,k}^2} = \frac{\kappa}{2}\left(\sum_{k=1}^n w_k \epsilon_k\right)^2 \quad (58)$$

with $w_k = x_{i,k}/\sqrt{\sum_{k=1}^n x_{i,k}^2}$. Since $\sum_{k=1}^n w_k\epsilon_k$ is a linear combination of Gaussian random variables with weights satisfying $\sum_{k=1}^n w_k^2 = 1$, $\sum_{k=1}^n w_k\epsilon_k$ is distributed as $\mathcal{N}(0, 1)$. As a result, $\frac{2}{\kappa}W_i(1, n; a_i^*)$ is χ_1^2 distributed.

We further note that as $n \to \infty$,

$$a_i^* = \frac{\frac{1}{n}\sum_{k=1}^n x_{i,k}\epsilon_k}{\frac{1}{n}\sum_{k=1}^n x_{i,k}^2} \overset{P_\infty - a.s.}{\to} \frac{0}{r_{i,i}} = 0. \quad (59)$$

Recall that $\mathcal{A}_i = \{a_i | a_i \in (-\infty, -b_i] \cup [b_i, \infty)\}$. Therefore, (59) indicates that there exists a finite random variable N_i such that $-b_i < a_i^* < b_i$ almost surely when $n > N_i$. Let $N_{max} = \max_i N_i$, then N_{max} is a finite random variable with

$$\mathbb{E}[N_{max}] \leq \mathbb{E}\left[\sum_{i=1}^{p} N_i\right] \leq c_1 p, \tag{60}$$

where $c_1 = \max_i \mathbb{E}[N_i]$ is a constant that is independent of p. Further, let N be a large constant, we have

$$\begin{aligned}
&P_\infty(\tau_1 < \infty) = P_\infty\left[U_{\tau_1} \geq \log B\right] \\
&= P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(1, \tau_1; a_i) \geq \log B\right] \\
&\leq P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(1, \tau_1; a_i) \geq \log B, \tau_1 \leq N\right] \\
&+P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(1, \tau_1; a_i) \geq \log B, \tau_1 > N_{max}\right] \\
&+P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(1, \tau_1; a_i) \geq \log B, N_{max} \geq \tau_1 > N\right]. \quad (61)
\end{aligned}$$

We then bound these three items on the right hand side of (61) individually. To bound the first item, we have to introduce some notations. Let

$$\hat{\mathbf{a}}^* = \arg\max_{\mathbf{a}\in\mathcal{A}} U_n.$$

Follow a discussion that is similar from (39) to (42), we have

$$U_n = \sum_{i=1}^{s} W_{(i)}(1, n; \hat{a}_{(i)}), \tag{62}$$

in which $\hat{a}_i = \arg\max_{a_i \in \mathcal{A}_i} W_i(1, n; a_i)$, and $W_{(i)}(1, n; \hat{a}_{(i)})$ is the i^{th} order statistic of $\{W_k(1, n; \hat{a}_k), k = 1, \ldots, p\}$. Then, for the first item, we have

$$\begin{aligned}
&P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(1, \tau_1; a_i) \geq \log B, \tau_1 \leq N\right] \\
&\overset{(a)}{=} \sum_{n=1}^{\infty} P_\infty\left[\sum_{i=1}^{s} W_{(i)}\left(1, n; \hat{a}_{(i)}\right) \geq \log B, \tau_1 = n, \tau_1 \leq N\right] \\
&\leq \sum_{n=1}^{N} P_\infty\left[W_{(1)}\left(1, n; \hat{a}_{(1)}\right) \geq \frac{\log B}{s}\right] \\
&\leq \sum_{n=1}^{N} P_\infty\left[\exists i \in \{1, \ldots, p\} \text{ such that } W_i(1, n; \hat{a}_i) \geq \frac{\log B}{s}\right] \\
&\leq \sum_{n=1}^{N}\sum_{i=1}^{p} P_\infty\left[W_i(1, n; \hat{a}_i) \geq \frac{\log B}{s}\right] \\
&\overset{(b)}{\leq} \sum_{n=1}^{N}\sum_{i=1}^{p} P_\infty\left[\frac{2}{\kappa} W_i(1, n; a_i^*) \geq \frac{2}{\kappa}\frac{\log B}{s}\right] \\
&= NpP_\infty\left[\chi_1^2 \geq \frac{2}{\kappa}\frac{\log B}{s}\right] \\
&\overset{(c)}{\leq} Np\sqrt{\frac{\kappa s}{4\pi}}\frac{1}{B^{1/\kappa s}[\log B]^{1/2}}, \quad (63)
\end{aligned}$$

in which (a) holds because of (62), (b) is due to definitions of $\hat{a}_i$ and a_i^*, and (c) is true because of the tail bounds inequality

$$P(X > x) \leq \frac{\exp(-x^2/2)}{x\sqrt{2\pi}}$$

for a standard normal random variable X.

We then bound the second item in (61). For $x_{i,k}$ under P_∞, we generate another two probability measures $Q_b^-(y_k)$ and $Q_b^+(y_k)$. In particular, $Q_b^-(y_k)$ is generated by linear transformation $y_k = -b_i x_{i,k} + \epsilon_k$ and $Q_b^+(y_k)$ by $y_k = b_i x_{i,k} + \epsilon_k$. A direct calculation shows that the Radon-Nikodym derivatives of Q_b^-, Q_b^+ and P_∞ for $(y_1, \ldots, y_n)$ are given as

$$\begin{aligned}
\frac{dQ_b^-}{dP_\infty} &= \exp\left\{-b_i\sum_{k=1}^{n} x_{i,k}y_k - \frac{1}{2}\sum_{k=1}^{n} b_i^2 x_{i,k}^2\right\} \\
&= \exp\left\{\frac{1}{\kappa}W_i(1, n; -b_i)\right\}, \\
\frac{dQ_b^+}{dP_\infty} &= \exp\left\{b_i\sum_{k=1}^{n} x_{i,k}y_k - \frac{1}{2}\sum_{k=1}^{n} b_i^2 x_{i,k}^2\right\} \\
&= \exp\left\{\frac{1}{\kappa}W_i(1, n; b_i)\right\}.
\end{aligned}$$

Note that $\hat{a}_i$ equals to either $-b_i$ or b_i, $\forall i \in \{1, \ldots, p\}$, on the event $\{\tau_1 > N_{max}\}$. Then, for the second item on the right hand side of (61), we have

$$\begin{aligned}
&P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(1, \tau_1; a_i) \geq \log B, \tau_1 > N_{max}\right] \\
&= P_\infty\left[\sum_{i=1}^{s} W_{(i)}\left(1, \tau_1; \hat{a}_{(i)}\right) \geq \log B, \tau_1 > N_{max}\right] \\
&\leq P_\infty\left[W_{(1)}\left(1, \tau_1; \hat{a}_{(1)}\right) \geq \frac{\log B}{s}, \tau_1 > N_{max}\right] \\
&\leq \sum_{i=1}^{p} P_\infty\left[W_i\left(1, \tau_1; \hat{a}_i\right) \geq \frac{\log B}{s}, \tau_1 > N_{max}\right] \\
&= \sum_{i=1}^{p} P_\infty\left[W_i\left(1, \tau_1; \hat{a}_i\right) \geq \frac{\log B}{s},\right. \\
&\qquad \left.\{\hat{a}_i = -b_i \text{ or } \hat{a}_i = b_i\}, \tau_1 > N_{max}\right] \\
&\leq \sum_{i=1}^{p}\left[P_\infty\left[W_i\left(1, \tau_1; -b_i\right) \geq \frac{\log B}{s}\right]\right. \\
&\qquad \left.+P_\infty\left[W_i\left(1, \tau_1; b_i\right) \geq \frac{\log B}{s}\right]\right] \\
&= \sum_{i=1}^{p}\left[\int_{\left\{W_i(1,\tau_1;-b_i) \geq \frac{\log B}{s}\right\}} \frac{dP_\infty}{dQ_b^-} dQ_b^-\right. \\
&\qquad \left.+\int_{\left\{W_i(1,\tau_1;b_i) \geq \frac{\log B}{s}\right\}} \frac{dP_\infty}{dQ_b^+} dQ_b^+\right] \\
&\overset{(a)}{\leq} \sum_{i=1}^{p} \frac{1}{e^{\log B/\kappa s}}\left[Q_b^-\left[W_i\left(1, \tau_1; -b_i\right) \geq \frac{\log B}{s}\right]\right. \\
&\qquad \left.+Q_b^+\left[W_i\left(1, \tau_1; b_i\right) \geq \frac{\log B}{s}\right]\right] \\
&= \frac{2p}{B^{1/\kappa s}}, \quad (64)
\end{aligned}$$

in which (a) holds because of inequalities (65) and (66) in the following

$$\begin{aligned}
&\int_{\left\{W_i(1,\tau_1;b_i)\geq\frac{\log B}{s}\right\}}\frac{dP_\infty}{dQ_b^+}dQ_b^+\\
&=\sum_{n=1}^{\infty}\int_{\left\{W_i(1,\tau_1;b_i)\geq\frac{\log B}{s},\tau_1=n\right\}}\frac{dP_\infty}{dQ_b^+}dQ_b^+\\
&=\sum_{n=1}^{\infty}\int_{\left\{W_i(1,n;b_i)\geq\frac{\log B}{s},\tau_1=n\right\}}\exp\left\{-\frac{1}{\kappa}W_i(1,n;b_i)\right\}dQ_b^+\\
&\leq\exp\left\{-\frac{1}{\kappa}\frac{\log B}{s}\right\}\sum_{n=1}^{\infty}\int_{\left\{W_i(1,n;b_i)\geq\frac{\log B}{s},\tau_1=n\right\}}dQ_b^+\\
&=\exp\left\{-\frac{\log B}{\kappa s}\right\}Q_b^+\left[W_i\left(1,\tau_1;b_i\right)\geq\frac{\log B}{s}\right]. \qquad (65)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\int_{\left\{W_i(1,\tau_1;-b_i)>\frac{\log B}{s}\right\}}\frac{dP_\infty}{dQ_b^-}dQ_b^-\\
&\quad\leq\exp\left\{-\frac{\log B}{\kappa s}\right\}Q_b^-\left[W_i\left(1,\tau_1;-b_i\right)\geq\frac{\log B}{s}\right]. \qquad (66)
\end{aligned}$$

The third term in the right hand of (61) can be bounded by the Markov inequality:

$$\begin{aligned}
&P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau_1;a_i)\geq\log B,N_{max}\geq\tau_1>N\right]\\
&\leq P(N_{max}>N)\leq\frac{\mathbb{E}[N_{max}]}{N}. \qquad (67)
\end{aligned}$$

By setting

$$N=B^{1/2\kappa s}(\log B)^{1/4}$$

and adding three bounds together, we obtain

$$\begin{aligned}
&P_\infty(\tau_1<\infty)\\
&\leq Np\sqrt{\frac{\kappa s}{4\pi}}\frac{1}{B^{1/\kappa s}[\log B]^{1/2}}+\frac{2p}{B^{1/\kappa s}}+\frac{\mathbb{E}[N_{max}]}{N}\\
&=2pB^{-\frac{1}{\kappa s}}+\left(p\sqrt{\frac{\kappa s}{4\pi}}+\mathbb{E}[N_{max}]\right)(\log B)^{-\frac{1}{4}}B^{-\frac{1}{2\kappa s}}.
\end{aligned}$$

This ends the proof.

APPENDIX B
PROOFS FOR THE BAYESIAN SETUP

The proof of Theorem III.3 relies on the following two supporting lemmas.

Lemma B.1. *(Detection Delay) If $\kappa<1$ and $\mu\geq\frac{s}{2p}\log(1-\kappa)$, then as $B\to\infty$*

$$\begin{aligned}
&\mathbb{E}_\pi^{\mathbf{a}}[\tau_S-t|\tau_S\geq t]\\
&\quad\leq\frac{\log B+c_3 s}{\frac{\kappa}{2}\sum_{i=1}^{p}a_i^2r_{ii}+\kappa\sum_{i\neq j}a_ia_jr_{ij}+p\mu}(1+o(1)), \qquad (68)
\end{aligned}$$

in which c_3 is a constant that is independent of p.

Lemma B.2. *(False Alarm) If $\kappa<1$ and*

$$\frac{s}{2p}\log(1-\kappa)\leq\mu<\frac{s}{2p}\log(1-\kappa)+\frac{s}{p}|\log(1-\rho)|, \qquad (69)$$

then for threshold B,

$$P_\infty(\tau_S<t)\leq\frac{1}{B^{1/s}}\frac{\rho c_2}{c_2-1}\frac{1}{1-(1-\rho)c_2}, \qquad (70)$$

in which $c_2=(1-\kappa)^{-1/2}\exp\{p\mu/s\}$.

Theorem III.3 then can be proved by setting

$$\log B=s|\log\alpha|+s\log\frac{\rho c_2}{(c_2-1)[1-(1-\rho)c_2]}. \qquad (71)$$

Putting this threshold into (70), we have $P_\infty(\tau_S<t)\leq\alpha$. Putting (71) into (68), we can obtain the upperbound of the detection delay presented in Theorem III.3. In the rest of this subsection, we prove the above two supporting lemmas.

A. Proof of Lemma B.1:

In the following, we study the detection delay of τ_S defined in (31). Assume that a genie knows the true post-change linear coefficient $\mathbf{a}$, and he uses the statistic

$$\begin{aligned}
\bar{U}(m,n)&=\sum_{i=1}^{p}\left[W_i(m,n;a_i)+(n-m+1)\mu\right]\\
&=\sum_{k=m}^{n}\left[\kappa y_k\sum_{i=1}^{p}a_ix_{i,k}-\frac{\kappa}{2}\sum_{i=1}^{p}(a_ix_{i,k})^2+p\mu\right]. \qquad (72)
\end{aligned}$$

Let $\xi_k=\kappa y_k\sum_{i=1}^{p}a_ix_{i,k}-\frac{\kappa}{2}\sum_{i=1}^{p}(a_ix_{i,k})^2+p\mu$ (note that this is the same ξ_k defined in (32)). Let

$$\tilde{\tau}_S=\inf\{n\geq0:\bar{U}(1,n)\geq\log B\}. \qquad (73)$$

Note that $\bar{U}(1,n)\leq S_n$ since the definition of S_n takes supreme over $\{\mathbf{a}\in\mathcal{A}\}$ and over $\{1\leq m\leq n\}$; therefore $\tilde{\tau}_S\geq\tau_S$. It is sufficient to find an upper bound for $\tilde{\tau}_S$.

On the event $\{t=m\}$, it is easy to see that as $n\to\infty$

$$\begin{aligned}
&\frac{1}{n}\bar{U}(m,m+n-1)=\frac{1}{n}\sum_{k=m}^{m+n-1}\xi_k\\
&=\sum_{i=1}^{p}\left[\kappa a_i\frac{1}{n}\sum_{k=m}^{m+n-1}x_{i,k}y_k-\frac{\kappa}{2}a_i^2\frac{1}{n}\sum_{k=m}^{m+n-1}x_{i,k}^2\right]+p\mu\\
&\overset{a.s.}{\to}\sum_{i=1}^{p}\left[\frac{\kappa}{2}a_i^2r_{ii}+\kappa\sum_{i\neq j}a_ia_jr_{ij}\right]+p\mu. \qquad (74)
\end{aligned}$$

Denote

$$I_\xi=\sum_{i=1}^{p}\left[\frac{\kappa}{2}a_i^2r_{ii}+\kappa\sum_{i\neq j}a_ia_jr_{ij}\right]+p\mu.$$

Then, we can rewrite the T_δ in (32) as

$$T_\delta=\inf\{n\geq0:|n^{-1}\bar{U}(m,m+n-1)-I_\xi|>\delta\}. \qquad (75)$$

On the event $\{\tilde{\tau}_S>T_\delta+(m-1)\}$, we have

$$\bar{U}(m,\tilde{\tau}_S-1)>(\tilde{\tau}_S-m+1)(I_\xi-\delta)$$

or equivalently,

$$\tilde{\tau}_S-m+1<\frac{\bar{U}(m,\tilde{\tau}_S-1)}{I_\xi-\delta}<\frac{\log B-\bar{U}(1,m-1)}{I_\xi-\delta}.$$

Then we have

$$\begin{aligned}&\tilde{\tau}_S - m + 1\\ &< \frac{\log B - \bar{U}(1,m-1)}{I_\xi - \delta}\mathbf{1}_{\{\tilde{\tau}_S > T_\delta + (m-1)\}} + T_\delta \mathbf{1}_{\{\tilde{\tau}_S \le T_\delta + (m-1)\}}\\ &\le \frac{\log B - \bar{U}(1,m-1)}{I_\xi - \delta} + T_\delta. \end{aligned} \tag{76}$$

Taking the conditional expectation on both sides, we have

$$\begin{aligned}&\mathbb{E}_m^{\mathbf{a}}[\tilde{\tau}_S - m | \tilde{\tau}_S \ge m]\\ &\le \frac{\log B}{I_\xi - \delta} - \mathbb{E}_m^{\mathbf{a}}\left[\frac{\bar{U}(1,m-1)}{I_\xi - \delta}\Big|\tilde{\tau}_S \ge m\right] + \mathbb{E}_m^{\mathbf{a}}\left[T_\delta \Big| \tilde{\tau}_S \ge m\right].\end{aligned}$$

In addition, we have

$$\begin{aligned}&\mathbb{E}_\pi^{\mathbf{a}}[\tilde{\tau}_S - t | \tilde{\tau}_S \ge t]\\ &\le \frac{\log B}{I_\xi - \delta} - \mathbb{E}_\pi^{\mathbf{a}}\left[\frac{\bar{U}(1,t-1)}{I_\xi - \delta}\Big|\tilde{\tau}_S \ge t\right] + \mathbb{E}_\pi^{\mathbf{a}}\left[T_\delta \Big| \tilde{\tau}_S \ge t\right].\end{aligned} \tag{77}$$

The uniform convergence assumption (3) entails the fact that for any $\mathbf{a} \neq \mathbf{0}$, there exists a common upper bound for $\mathbf{a}^T\mathbf{x}_k, k = 1, 2, \ldots$. For a given $\mathbf{a}$, let $c_a := \max_i (a_i x_{i,k})^2$, we then have

$$\begin{aligned}&\mathbb{E}_m^{\mathbf{a}}\left[\bar{U}(1,m-1)\right] = \mathbb{E}_\infty\left[\bar{U}(1,m-1)\right]\\ &= \sum_{k=1}^{m-1}\left[p\mu - \frac{\kappa}{2}\sum_{i=1}^{p}(a_i x_{i,k})^2\right]\\ &\ge \sum_{k=1}^{m-1}\left[p\mu - \frac{\kappa}{2}s c_a\right].\end{aligned}$$

The last step is true as $\mathbf{a}$ has at most s non-zeros. For the case of $\{m = 0\}$ and $\{m = 1\}$, we use the convention

$$\mathbb{E}_m^{\mathbf{a}}[\bar{U}(1,m-1)] = \mathbb{E}_m^{\mathbf{a}}\left[\sum_{k=1}^{m-1}\xi_k\right] = 0.$$

Then

$$\begin{aligned}&\mathbb{E}_\pi^{\mathbf{a}}[\bar{U}(1,t-1)] = \sum_{m=0}^{\infty}\pi_m \mathbb{E}_m^{\mathbf{a}}[\bar{U}(1,m-1)]\\ &\ge \sum_{m=2}^{\infty}(1-\pi_0)\rho(1-\rho)^{m-1}(m-1)(p\mu - \frac{\kappa}{2}s c_a)\\ &= s\frac{(1-\pi_0)(1-\rho)}{\rho}\left[\frac{p}{s}\mu - \frac{\kappa}{2}c_a\right]\\ &\ge s c_3,\end{aligned} \tag{78}$$

in which

$$c_3 = \frac{(1-\pi_0)(1-\rho)}{2\rho}\left[\log(1-\kappa) - \kappa c_a\right].$$

The last inequality holds due to the condition $\mu \ge \frac{s}{2p}\log(1-\kappa)$. Hence, c_3 is a constant independent of s and p.

Since we have assumed that $\mathbb{E}_\pi^{\mathbf{a}}[T_\delta] < \infty$, and $\{\tilde{\tau}_S \ge t\}$ is an almost sure event as $B \to \infty$, by (77) we have

$$\begin{aligned}&\mathbb{E}_\pi^{\mathbf{a}}[\tilde{\tau}_S - t | \tilde{\tau}_S \ge t]\\ &\le \left(\frac{\log B}{I_\xi - \delta} - \frac{\mathbb{E}_\pi^{\mathbf{a}}\left[\bar{U}(1,t-1)\right]}{I_\xi - \delta}\right)(1 + o(1))\\ &\le \frac{\log B + c_3 s}{I_\xi - \delta}(1 + o(1)).\end{aligned} \tag{79}$$

Then, Lemma B.1 follows the fact that δ is arbitrarily close to zero and that $I_\xi = \frac{\kappa}{2}\sum_{i=1}^{p} a_i^2 r_{i,i} + \kappa\sum_{i\neq j} a_i a_j r_{i,j} + p\mu$.

B. Proof of Lemma B.2:

In the following, we study the false alarm probability of τ_S defined in (31). To proceed, we first recall some notations in Section III-D. Specifically, $\hat{\mathbf{a}}^* = [\hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_p^*]^T$ is the optimal estimation of $\mathbf{a}$ in (29). Note that $\hat{\mathbf{a}}^*$ is also optimal for $\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p} W_i(m,n;a_i)$. Further $\hat{a}_i = \arg\max_{a_i\in\mathcal{A}_i} W_i(m,n;a_i)$ and $W_{(i)}(m,n;\hat{a}_{(i)})$ is the i^{th} order statistic of $\{W_i(m,n;\hat{a}_i)\}_{i=1}^{p}$. With these notations, for a constant N, we have

$$\begin{aligned}&P_\infty(\tau_S \le N) = P_\infty\left(\max_{1\le n\le N}\exp\{S_n\} \ge B\right)\\ &= P_\infty\left(\max_{1\le n\le N}\exp\left\{\sup_{1\le m\le n}\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p} V_i(m,n;a_i)\right\} \ge B\right)\\ &= P_\infty\left(\max_{1\le n\le N}\exp\left\{\sup_{1\le m\le n}\sum_{i=1}^{p} V_i(m,n;\hat{a}_i^*)\right\} \ge B\right)\\ &= P_\infty\left(\max_{1\le n\le N}\sup_{1\le m\le n}\prod_{i=1}^{p} e^{V_i(m,n;\hat{a}_i^*)} \ge B\right)\\ &= P_\infty\left(\max_{1\le n\le N}\sup_{1\le m\le n} e^{p(n-m+1)\mu}\prod_{i=1}^{p} e^{W_i(m,n;\hat{a}_i^*)} \ge B\right)\\ &\overset{(a)}{=} P_\infty\left(\max_{1\le n\le N}\sup_{1\le m\le n} e^{p(n-m+1)\mu}\prod_{i=1}^{s} e^{W_{(i)}(m,n;\hat{a}_{(i)})} \ge B\right)\\ &\le P_\infty\left(\max_{1\le n\le N}\sup_{1\le m\le n} e^{\frac{p}{s}(n-m+1)\mu} e^{W_{(1)}(m,n;\hat{a}_{(1)})} \ge B^{\frac{1}{s}}\right),\end{aligned} \tag{80}$$

where (a) is true due to (41) and (43). In the following, we will construct a submartingale and apply Doob's martingale inequality to bound the false alarm probability. We have

$$\begin{aligned}&W_{(1)}(m,n;\hat{a}_{(1)}) = \max_{1\le i\le p} W_i(m,n;\hat{a}_i)\\ &= \max_{1\le i\le p}\sup_{a_i\in\mathcal{A}_i} W_i(m,n;a_i)\\ &= \max_{1\le i\le p}\sup_{a_i\in\mathcal{A}_i}\left[\kappa a_i\sum_{k=m}^{n} x_{i,k}y_k - \frac{\kappa}{2}\sum_{k=m}^{n}(a_i x_{i,k})^2\right]\\ &\le \max_{1\le i\le p}\kappa\sum_{k=m}^{n}\sup_{a_i\in\mathcal{A}_i}\left[a_i x_{i,k}y_k - \frac{1}{2}(a_i x_{i,k})^2\right]\\ &\overset{(a)}{=} \max_{1\le i\le p}\frac{\kappa}{2}\sum_{k=m}^{n} y_k^2 = \frac{\kappa}{2}\sum_{k=m}^{n} y_k^2,\end{aligned} \tag{81}$$

in which (a) is true as $a_i x_{i,k}y_k - \frac{1}{2}(a_i x_{i,k})^2$ achieves its maximum value $y_k^2/2$ when $a_i = y_k/x_{i,k}$. Putting (81) into (80), we have

$$\begin{aligned}&P_\infty(\tau_S \le N)\\ &\le P_\infty\left(\max_{1\le n\le N}\sup_{1\le m\le n} e^{\frac{p}{s}(n-m+1)\mu} e^{\frac{\kappa}{2}\sum_{k=m}^{n} y_k^2} \ge B^{\frac{1}{s}}\right)\\ &\le P_\infty\left(\max_{1\le n\le N}\sum_{m=1}^{n}\prod_{k=m}^{n} e^{\frac{\kappa}{2}y_k^2 + \frac{p}{s}\mu} \ge B^{\frac{1}{s}}\right).\end{aligned} \tag{82}$$

Let

$$M_n := \sum_{m=1}^{n}\prod_{k=m}^{n}\exp\left\{\frac{\kappa}{2}y_k^2+\frac{p}{s}\mu\right\} = (M_{n-1}+1)\exp\left\{\frac{\kappa}{2}y_n^2+\frac{p}{s}\mu\right\}. \quad (83)$$

We note that M_n could be a submartingale. Particularly, let $\mathcal{F}_n := \sigma\{y_1,\ldots,y_n\}$, we have

$$\mathbb{E}_\infty[M_n|\mathcal{F}_{n-1}] = (M_{n-1}+1)\exp\left\{\frac{p}{s}\mu\right\}\mathbb{E}_\infty\left[\exp\left\{\frac{\kappa}{2}y_n^2\right\}\right].$$

Since we have $\kappa<1$ in the condition, then $\mathbb{E}_\infty\left[\exp\left\{\frac{\kappa}{2}y_n^2\right\}\right]$ is integrable and $\mathbb{E}_\infty\left[\exp\left\{\frac{\kappa}{2}y_n^2\right\}\right] = (1-\kappa)^{-1/2}$. In addition, the condition $\frac{s}{2p}\log(1-\kappa)\le\mu$ guarantees $(1-\kappa)^{-\frac{1}{2}}\exp\left\{\frac{p}{s}\mu\right\}\ge 1$; hence we have $\mathbb{E}_\infty[M_n|\mathcal{F}_{n-1}]\ge M_{n-1}+1>M_{n-1}$, i.e., M_n is a submartingale. In addition,

$$\mathbb{E}_\infty[M_n] = \sum_{m=1}^{n}\prod_{k=m}^{n}\mathbb{E}_\infty\left[\exp\left\{\frac{\kappa}{2}y_k^2+\frac{p}{s}\mu\right\}\right] = \sum_{m=1}^{n}\prod_{k=m}^{n}\left[(1-\kappa)^{-\frac{1}{2}}\exp\left\{\frac{p}{s}\mu\right\}\right] = \frac{c_2(c_2^n-1)}{c_2-1}, \quad (84)$$

in which $c_2 := (1-\kappa)^{-\frac{1}{2}}\exp\{p\mu/s\}$. By Doob's martingale inequality

$$P_\infty\left(\max_{1\le n\le N}M_n\ge B^{1/s}\right)\le\frac{\mathbb{E}_\infty[M_N]}{B^{1/s}}. \quad (85)$$

Combining (82) and (84), we have

$$P_\infty(\tau_S\le N)\le P_\infty\left(\max_{1\le n\le N}M_n\ge B^{1/s}\right)\le\frac{1}{B^{1/s}}\frac{c_2(c_2^N-1)}{c_2-1}.$$

Further,

$$P_\pi(\tau_S<t) = \sum_{N=1}^{\infty}\pi_N P_\infty(\tau_S\le N-1) \le \sum_{N=1}^{\infty}(1-\pi_0)\rho(1-\rho)^{N-1}\frac{1}{B^{1/s}}\frac{c_2}{c_2-1}c_2^{N-1} = \frac{1}{B^{1/s}}\frac{\rho c_2}{c_2-1}\frac{1-\pi_0}{1-(1-\rho)c_2}. \quad (86)$$

in which the last step is because the condition $\mu<\frac{s}{2p}\log(1-\kappa)+\frac{s}{p}|\log(1-\rho)|$ guarantees $(1-\rho)c_2<1$.

REFERENCES

[1] J. Geng, B. Zhang, L. M. Huie, and L. Lai, "Online change detection of linear regression models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (Shanghai, China), Mar. 2016.

[2] S. Kallummil and S. Kalyani, "High SNR consisitent linear moder order selection ans subset selection," *IEEE Trans. Signal Processing*, vol. 64, pp. 4307–4322, Aug. 2016.

[3] X. Jiang, W. Zeng, H. So, A. M. Zoubir, and T. Kirubarajan, "Beamforming via noncovex linear regression," *IEEE Trans. Signal Processing*, vol. 64, pp. 1714–1728, Apr. 2016.

[4] G. Mateos, J. A. Bazeque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Processing*, vol. 58, pp. 5262–5276, Oct. 2010.

[5] P. Perron, *Dealing With Structutral Breaks Palgrave Handbook of Econometrics*. New York, NY, USA: Palgrave Macmillan, 2006.

[6] N. Zhang and D. Siegmund, "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 63, pp. 22–32, Mar. 2007.

[7] G. Papageorgiou, P. Bouboulis, and S. Theodoridis, "Distributed sparse linear regression," *IEEE Trans. Signal Processing*, vol. 63, pp. 3872–3887, Aug. 2015.

[8] T. Hu, Q. Wu, and D. Zhou, "Convergence of gradient descent for minimum error entropy principle in linear regression," *IEEE Trans. Signal Processing*, vol. 64, pp. 6571–6579, Dec. 2016.

[9] E. G. Larsson and Y. Selen, "Linear regression with a sparse parameter vector," *IEEE Trans. Signal Processing*, vol. 55, pp. 451–460, Feb. 2007.

[10] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online adaptive estimation of sparse signals: where RLS meets the l_1 norm," *IEEE Trans. Signal Processing*, vol. 58, pp. 3436–3447, July 2010.

[11] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[12] P. Mattsson, D. Zachariah, and P. Stoica, "Recursive identification method for piecewise arx models: A sparse estimation approach," *IEEE Trans. Signal Processing*, vol. 64, pp. 5082–5093, Oct. 2016.

[13] J. Bai and P. Perron, "Estimating and testing linear models with multiple structural changes," *Econometrica*, vol. 66, pp. 47–78, 1998.

[14] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society*, vol. 57, pp. 613–658, 1995.

[15] G. Lorden, "Procedures for reacting to a change in distribution," *Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[16] A. N. Shiryaev, "On optimal methods in quickest detection problems," *Theory of Probability and Its Applications*, vol. 8, pp. 22–46, 1963.

[17] A. N. Shiryaev, "The problem of the most rapid detection of a disturbance in a stationary process," *Soviet Math. Dokl.*, no. 2, pp. 795–799, 1961. (translation from Dokl. Akad. Nauk SSSR vol. 138, pp. 1039-1042, 1961).

[18] T. L. Lai, "Information bounds and quickest detection of parameter changes in stochastic systems," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2917–2929, Nov. 1998.

[19] T. Banerjee and V. V. Veeravalli, "Data-efficient minimax quickest change detection with composite post-change distribution," *IEEE Trans. Inform. Theory*, vol. 61, pp. 5172–5184, Sept. 2015.

[20] A. G. Tartakovsky and A. S. Polunchenko, "Quickest changepoint detection in distributed multisensor systems under unknown parameters," in *Proc. Intl. Conf. on Information Fusion*, (Cologne, Germany), pp. 878–885, Jul. 2008.

[21] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Trans. Inform. Theory*, vol. 61, pp. 6926–6938, Dec. 2015.

[22] Q. Yao, "Asymptotically optimal detection of a change in a linear model," vol. 12, no. 3, pp. 201–210, 1993.

[23] B. Yakir, A. Krieger, and M. Pollak, "Detecting a change in regression: first-order optimality," *Annals of Statistics*, vol. 27, no. 6, pp. 1896–1913, 1999.

[24] I. V. Nikiforov, "A simple change detection scheme," *Signal Processing*, vol. 81, pp. 149–172, 2001.

[25] H.-J. Kim and D. Siegmund, "The likelihood ratio test for a change-point in simple linear regression," vol. 76, pp. 409–423, Aug. 1989.

[26] Y. Cao, Y. Xie, and N. Gebraeel, "Multi-sensor slope change detection," *Annals of Operations Research*, pp. 1–27, April 2016.

[27] M. Basseville, I. V. Nikiforov, and A. G. Tartakovsky, *Sequential Analysis: Hypothesis Testing and ChangePoint Detection*. CRC Press, 2013.

[28] C. Chu, M. Stinchcombe, and H. White, "Change-point monitoring in linear models," *Econometrica*, vol. 64, pp. 1045–1065, Sept. 1996.

[29] A. Aue, L. Horvath, M. Huskova, and P. Kokoszka, "Change-point monitoring in linear models," *The Econometrics Journal*, vol. 9, pp. 373–403, Sept. 2006.

[30] B. Zhang, J. Geng, and L. Lai, "Multiple change-points estimation in linear regression models via sparse group lasso," *IEEE Trans. Signal Processing*, vol. 63, pp. 2209–2224, May 2015.

[31] Y. Yilmaz, G. V. Moustakides, and X. Wang, "Sequential joint detection and estimation," *Theory of Probability and Its Applications*, vol. 59, no. 3, pp. 452–465, 2015.

[32] Y. Yilmaz, S. Li, and X. Wang, "Sequential joint detection and estimation: Optimum tests and applications," *IEEE Trans. Signal Processing*, vol. 64, pp. 5311–5326, Oct. 2016.

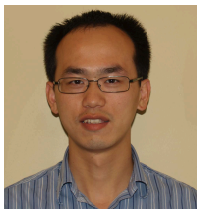
[33] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.
[34] G. Fellouris and G. Sokolov, "Second-order asymptotic optimality in multisensor sequential change detection," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3662–3675, 2013.
[35] D. L. Donoho, "Compressive sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289 – 1306, Apr. 2006.
[36] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, Feb. 2006.
[37] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, UK: Cambridge University Press, 2008.
[38] B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.
[39] "sEMG for basic hand movements data set." https://archive.ics.uci.edu/ml/datasets/sEMG+for+Basic+Hand+movements.

Jun Geng (S'13-M'15) received the B. E. and M. E. degrees from Harbin Institute of Technology, Harbin, China in 2007 and 2009 respectively, and the Ph.D. degree from Worcester Polytechnic Institute, MA, United States in 2015. Since June 2015, he has been an associate professor at Harbin Institute of Technology. Dr. Geng's research interests include stochastic signal processing, wireless communications and other related areas.

Bingwen Zhang received the B. E. degree from University of Science and Technology of China, Hefei, China in 2011, the M. S. degree from Worcester Polytechnic Institute, MA, in 2013, and the Ph.D. degree in Electrical and Computer Engineering from Worcester Polytechnic Institute, MA, in 2017. His research interests are in statistical learning and data mining.

Lauren M. Huie received the B.S. degree in Electrical Engineering from the State University of New York at Binghamton in 2005. In 2007 she received the M.S. degree in Electrical Engineering from The Pennsylvania State University. She received the Ph.D. degree in 2013 from the State University of New York at Binghamton.

She is currently with the Air Force Research Laboratory Information Directorate in Rome, NY. Her current research interests include sensor networks, estimation and detection theory, and physical layer security.



Lifeng Lai (M'07) received the B.E. and M. E. degrees from Zhejiang University, Hangzhou, China in 2001 and 2004 respectively, and the PhD degree from The Ohio State University at Columbus, OH, in 2007. He was a postdoctoral research associate at Princeton University from 2007 to 2009, an assistant professor at University of Arkansas, Little Rock from 2009 to 2012, and an assistant professor at Worcester Polytechnic Institute. Since 2016, he has been an associate professor at University of California, Davis. Dr. Lais research interests include information theory, stochastic signal processing and their applications in wireless communications, security and other related areas.

Dr. Lai was a Distinguished University Fellow of the Ohio State University from 2004 to 2007. He is a co-recipient of the Best Paper Award from IEEE Global Communications Conference (Globecom) in 2008, the Best Paper Award from IEEE Conference on Communications (ICC) in 2011 and the Best Paper Award from IEEE Smart Grid Communications (SmartGridComm) in 2012. He received the National Science Foundation CAREER Award in 2011, and Northrop Young Researcher Award in 2012. He served as a Guest Editor for IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing Techniques for Wireless Physical Layer Security from 2012 to 2013, and an Editor for IEEE Transactions on Wireless Communications from 2013 to 2018. He is currently serving as an Associate Editor for IEEE Transactions on Information Forensics and Security.