Online Change-Point Detection of Linear Regression Models

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

***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 time, both non-Bayesian and Bayesian problem formulations are considered in this paper. 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. We proposed a low complexity algorithm, namely the parallel-sum algorithm, for change-point detection. 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 to evaluate the performance of detection algorithms. A modified version of the parallel-sum algorithm is proposed for the Bayesian formulation. For both setups, we analyze the performance of the proposed algorithms and show that they offer good performance while requiring low computational complexity.**

***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], image interpolation [4], cognitive radio network [5], economic data analysis [6], biomedical science [7], etc. Many efforts have been devoted into the problem of estimating the coefficients in the linear regression model based on a group of observing data [8]–[12]. The underlying assumption in such estimation problem is that all data come from a single linear model. However, in many other applications, the underlying model changes over time [13]. For example, in building economic growth models, it is more appropriate to assume that the available various economic indicators obey different models in different time period as the economic growth pattern undergoes structural changes over the years [14]. 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 [15]. In such applications, it is of interest to detect the presence of such changes in the underlying model quickly.

In this paper, we focus on *on-line detection* of such changes in linear regression models. In particular, an observer keeps monitoring the explanatory variables $\mathbf{x}_n$ and the dependent variable y_n. Here, y_n and $\mathbf{x}_n$ are assumed to obey a linear model at each time slot n. At the very beginning, the relationship between y_n and $\mathbf{x}_n$ is assumed to be known. However, some of the linear coefficients change at an unknown time t, and the observer does not know the post-change linear coefficients. Based on his sequential observations, the observer aims to design an on-line detection algorithm to quickly and accurately detect such change in the linear model.

We formulate this problem in the framework of quickest change-point detection. 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 [16] 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 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 [17], [18]. Correspondingly, the observer wants to minimize the average detection delay (average over the prior distribution of the change-point) subject to a false alarm probability constraint.

The optimal solutions for classic quickest detection problems, for which the post-change distribution is perfectly known by the observer, are well known. Specifically, the cumulative sum (CUSUM) procedure is optimal for Lorden's formulation [16], [19], [20] and the Shiryaev-Robert (SR) procedure is optimal for the Bayesian formulation [17], [21]. In our problem, however, the post-change coefficient in the linear regression model is unknown to the observer. It is natural to consider the generalized likelihood ratio (GLR) based algorithms for

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 CNS-1660128. 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).

problems with unknown models. However, as we will discuss in the sequel, in our setup, the GLR-CUSUM procedure and the GLR-SR procedure suffer huge computational burden. In this paper, we focus on designing schemes that have low complexity yet still offer reasonable performance. In particular, we propose a low complexity algorithm named parallel-sum algorithm. In the proposed 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 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. Specifically, in the non-Bayesian formulation, to guarantee ARL2FA to be no less than a preset threshold 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 $p\alpha \to c$. We note that the proposed algorithm is neither optimal nor asymptotically optimal; however, the proposed algorithm has very low computational complexity and its detection delay is reasonable. At time slot n, the computational complexity of the proposed parallel-sum algorithm is on the order of $O(np)$.

The problem considered in this paper is related to recent works on the quickest change-point detection problem with unknown post-change parameters. In particular, [20] shows GLR-CUSUM is asymptotic optimal for the non-Bayesian quickest detection problem when the post-change distribution contains unknown parameters. [15] adopts the window-limited GLR-CUSUM for the change detection in the stochastic dynamic system. [22] proposes the SUM algorithm, which based on the sum of local CUSUMs, to quickly detect the abrupt change in multiple independent data streams. The authors in [23] also consider the change detection problem for linear model. Particularly, the unknown post-change parameter space is decomposed into several subspaces, and for each subspace the observer runs a recursive GLR test for detection purpose. To the author's best knowledge, there are few works that considered the Bayesian quickest detection problem with unknown post-change parameters. Different from these works, we point out that the commonly analyzed GLR-CUSUM procedure suffer a huge computational burden in our problem, and we propose a low complexity detection algorithm to deal with our proposed problem. In addition, the proposed algorithm also works under Bayeisan setting and corresponding performance is analyzed.

We also briefly review other related papers. There are a series of works such as [24], [25] that consider the problem of monitoring model or structural change. However, these works focus on the probability of detection of the change-point while our work focuses on analyzing the detection delay. Some other works, such as [14], [26], also consider that the structure of data in the data set 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 is of interest.

This paper extends our previous conference publication [1] in several ways. Specifically, [1] focuses on the non-Bayesian formulation and analyzes the performance of the parallel-sum algorithm under $p/\gamma \to 0$. However, besides the contributions made in [1], this paper also considers the Bayesian formulation and analyzes the performance of the proposed algorithms under $p/\gamma \to c$. In addition, this paper provides detailed technic proofs, and also discusses the computational complexity of the proposed parallel-sum algorithm.

The remainder of this paper is organized as follows. The mathematical model is given in Section II. Section III presents the proposed algorithms and the main conclusions of this paper. In Section IV, we provide the technic proofs of the main conclusions. Numerical examples are given in Section V to illustrate the results obtained in this work. Finally, Section VI offers concluding remarks.

II. MODEL

We consider the change-point detection problem in a linear regression model. Let $\{(\mathbf{x}_n, z_n)\}_{n=1}^{\infty}$ be a sequence of observations whose underlying model changes at an unknown change-point t. For each time instant n, the scaler dependent variable z_n and the explanatory variable $\mathbf{x}_n$ obey the following model

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

in which $\epsilon_n \sim \mathcal{N}(0,1)$ models the normalized Gaussian 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. 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.

In this paper, we assume that $\mathbf{x}_n = [x_{1,n}, x_{2,n}, \ldots, x_{p,n}]^T \in \mathbb{R}^p$; hence $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are also p-dimensional real vectors. Furthermore, we assume that $\mathbf{x}_n$ has an underlying probability distribution with probability density function (pdf) $f(\mathbf{x})$. However, the observer does not have any information on $f(\mathbf{x})$ except that $\mathbf{x}$ has zero mean and that the elements of its covariance matrix $\mathbf{R}$ are finite. $\{\mathbf{x}_n\}$ and $\{\epsilon_n\}$ are independent and identically distributed (i.i.d.) over time slots n.

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 rest of this paper, we assume that the observer has conducted above transformation on his observation sequence, and we will focus on the simplified model (2) in the sequel.

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 the post-change linear coefficient $\mathbf{a}$ only contains s non-zero components. Furthermore, s is assumed to be known to the observer.

Let $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$ and let $\mathcal{A}$ be the domain of $\mathbf{a}$. Particularly, $\mathcal{A}$ is specified in the following manner: if the i^{th} component in $\mathbf{a}$ is modified by the abrupt change, then a_i falls in the set:

$$\mathcal{A}_i = \{a_i | a_i \in (-b_{i,2}, -b_{i,1}] \cup [d_{i,1}, d_{i,2})\}, \tag{3}$$

in which $b_{i,2} > b_{i,1} > 0$ and $d_{i,2} > d_{i,1} > 0$. Furthermore, we define $\bar{\mathcal{A}}_i := \{a_i = 0\}$, then $\mathcal{A}$ can be expressed as

$$\mathcal{A} = \bigcup_{(i_1,\ldots,i_p)\in\mathcal{P}} (\mathcal{A}_{i_1} \cup \cdots \cup \mathcal{A}_{i_s} \cup \bar{\mathcal{A}}_{i_{s+1}} \cup \cdots \cup \bar{\mathcal{A}}_{i_p}), \tag{4}$$

where $\mathcal{P}$ consists of all permutations of set $\{1, 2, \ldots, p\}$. We note that $\mathbf{a} = \mathbf{0}$ is excluded from $\mathcal{A}$ and we assume that $\mathcal{A}$ is also known to the observer.

The observer aims to detect the change-point t via his sequential observations $\{(\mathbf{x}_n, y_n), n = 1, 2, \ldots\}$. Let τ be the stopping time when the observer declares that a change has occurred. The goal of the observer is to, loosely speaking, minimize the detection delay $(\tau - t)^+$ while keeping the false alarm $\{\tau < t\}$ under control. Two formal mathematic formulations, based on different assumptions on the change-point t, are considered in this paper.

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

$$\begin{aligned} &\text{minimize}_\tau \quad \text{WADD}(\tau; \mathbf{a}) := \\ &\quad \sup_{t\geq 1} \operatorname{esssup} \mathbb{E}_t^{\mathbf{a}}[(\tau - t + 1)^+ | \mathcal{F}_{t-1}], \quad \text{for all } \mathbf{a} \in \mathcal{A} \\ &\text{subject to} \quad \text{ARL2FA}(\tau) := \mathbb{E}_\infty[\tau] \geq \gamma, \end{aligned} \tag{5}$$

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 $\{(\mathbf{x}_n, y_n)\}_{n=1}^{t-1}$. (5) is known as Lorden's formulation [16], which is a min-max setting that aims to minimize the worst case average detection delay (WADD) over both change-point t and observations up to $t-1$. $\mathbb{E}_\infty[\tau]$ is termed as average run length to false alarm (ARL2FA). Since no change happens in the event $\{t = \infty\}$, the declaration at τ is a false alarm; hence the constraint in (5) requires that the expected duration to a false alarm is no less than γ.

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

$$P(t = m) = \rho(1 - \rho)^{m-1}, \quad m = 1, 2, \ldots, \tag{6}$$

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

$$\begin{aligned} P_\pi^{\mathbf{a}}(F) &:= \sum_{m=1}^{\infty} P_t^{\mathbf{a}}(F | t = m) P(t = m) \\ &= \sum_{m=1}^{\infty} P_m^{\mathbf{a}}(F) P(t = m). \end{aligned} \tag{7}$$

The problem under the Bayesian framework is then formulated as

$$\begin{aligned} &\text{minimize}_\tau \quad \text{ADD}(\tau; \mathbf{a}) := \mathbb{E}_\pi^{\mathbf{a}}[\tau - t | \tau \geq t], \text{ for all } \mathbf{a} \in \mathcal{A}. \\ &\text{subject to} \quad \text{PFA}(\tau) := \sup_{\mathbf{a}\in\mathcal{A}} P_\pi^{\mathbf{a}}(\tau < t) \leq \alpha, \end{aligned} \tag{8}$$

in which $\mathbb{E}_\pi^{\mathbf{a}}$ is the expectation with respect to $P_\pi^{\mathbf{a}}$. Hence, (8) aims to minimize the average detection delay (ADD) while keeping the probability of false alarm (PFA) under control.

We note that both (5) and (8) are multi-objective optimization problems. Optimal solutions for these two proposed problems are in general difficult to obtain. Hence, in this paper, we aim to propose low complexity sub-optimal algorithms and to analyze their performances.

III. The Parallel-Sum Algorithm

A. Challenges for Existing Methods

Let $f_0(\mathbf{x}_n, y_n)$ be the joint probability density function (pdf) of $(\mathbf{x}_n, y_n)$ when $n < t$. Let $f_1(\mathbf{x}_n, y_n; \mathbf{a})$ be the joint pdf of $(\mathbf{x}_n, y_n)$ when $n > t$ and the post-change linear coefficient is $\mathbf{a}$. Even though both f_0 and f_1 are assumed to be unknown in our paper, for any given $\mathbf{a}$, the likelihood ratio (LR) can be calculated as

$$\begin{aligned} L(\mathbf{x}_n, y_n; \mathbf{a}) &:= \frac{f_1(\mathbf{x}_n, y_n; \mathbf{a})}{f_0(\mathbf{x}_n, y_n)} = \frac{f_1(y_n | \mathbf{x}_n; \mathbf{a}) f(\mathbf{x}_n)}{f_0(y_n) f(\mathbf{x}_n)} \\ &= \frac{\exp\{-\frac{1}{2}(y_n - \mathbf{a}^T \mathbf{x}_n)^2\}}{\exp\{-\frac{1}{2} y_n^2\}} \\ &= \exp\left\{\mathbf{a}^T \mathbf{x}_n y_n - \frac{1}{2} \mathbf{a}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{a}\right\}. \end{aligned} \tag{9}$$

Further, the Kullback-Leibler (KL) divergence between $f_0(\mathbf{x}_n, y_n)$ and $f_1(\mathbf{x}_n, y_n; \mathbf{a})$ is

$$\begin{aligned} D(f_1, f_0; \mathbf{a}) &= \mathbb{E}^{\mathbf{a}}[\log L(\mathbf{x}_n, y_n; \mathbf{a})] = \frac{1}{2} \mathbf{a}^T \mathbf{R} \mathbf{a} \\ &= \frac{1}{2} \sum_{i=1}^{p} a_i^2 r_{i,i} + \frac{1}{2} \sum_{i\neq j} a_i a_j r_{i,j}, \end{aligned} \tag{10}$$

in which $r_{i,j}$ is the element located at the i^{th} row and the j^{th} column in $\mathbf{R}$. Recall that $\mathbf{R}$ is the covariance matrix of $\mathbf{x}_n$. The lower bounds of the detection delay for the proposed formulations are presented in the following theorem, which restates the well known results in [20] and [21] in our context.

Theorem III.1. (Theorem 1 in [20] and Theorem 1 in [21]) *For the non-Bayesian formulation, as $\gamma \to \infty$,*

$$\begin{aligned} &\inf\{\mathit{WADD}(\tau, \mathbf{a}) : \mathit{ARL2FA}(\tau) \geq \gamma\} \\ &\qquad \geq \frac{|\log \gamma|}{D(f_1, f_0; \mathbf{a})}(1 + o(1)) \end{aligned} \tag{11}$$

for any $\mathbf{a} \in \mathcal{A}$. *For the Bayesian formulation, as* $\alpha \to 0$,

$$\inf\{ADD(\tau, \mathbf{a}) : PFA(\tau) \le \alpha\} \ge \frac{|\log \alpha|}{D(f_1, f_0; \mathbf{a}) + |\log(1-\rho)|}(1 + o(1)) \quad (12)$$

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

If the post-change linear coefficient $\mathbf{a}$ is perfectly known by the observer, it is well known that the cumulative sum (CUSUM) procedure is the optimal detection procedure for Lorden's formulation and the Shiryaev-Robert (SR) procedure is optimal for the Bayesian formulation. In our paper, the true post-change linear coefficient exhibits uncertainty to the observer since $\mathbf{a}$ could be any value in $\mathcal{A}$. Hence, it is natural to replace the likelihood ratio used in the CUSUM procedure and the SR procedure by the generalized likelihood ratio (GLR). In particular, the GLR-CUSUM procedure can be written as

$$T_n := \max_{1\le m\le n} \frac{\sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} f_1(\mathbf{x}_k, y_k; \mathbf{a})}{\prod_{k=m}^{n} f_0(\mathbf{x}_k, y_k)} = \max_{1\le m\le n} \sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} L(\mathbf{x}_k, y_k; \mathbf{a}), \quad (13)$$

$$\tau_{GLR-CUSUM} := \min\{n \ge 0 : T_n \ge B\}, \quad (14)$$

and the GRL-SR procedure can be written as

$$R_n := \sum_{m=1}^{n} \left(\frac{1}{1-\rho}\right)^{n-m+1} \frac{\sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} f_1(\mathbf{x}_k, y_k; \mathbf{a})}{\prod_{k=m}^{n} f_0(\mathbf{x}_k, y_k)} = \sum_{m=1}^{n} \left(\frac{1}{1-\rho}\right)^{n-m+1} \sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} L(\mathbf{x}_k, y_k; \mathbf{a}), \quad (15)$$

$$\tau_{GLR-SR} := \min\{n \ge 0 : R_n \ge B\}. \quad (16)$$

Note that the threshold B in (14) should be designed according to the ARL2FA constraint (5), and threshold B in (16) should be designed to satisfy the PFA constraint (8).

The GLR-CUSUM procedure has been shown to be asymptotically optimal for Lorden's formulation when the post-change parameter is unknown [20]. However, to authors' best knowledge, the optimality of the GLR-SR procedure for the Bayesian formulation is still an open problem. Though these two GLR based algorithms are natural and attractive, the huge computational burden prevents them from practical applications. In particular, we note that for both GLR based algorithms the observer has to estimate $\mathbf{a}$ by solving

$$\sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} L(\mathbf{x}_k, y_k; \mathbf{a}) = \sup_{\mathbf{a}\in\mathcal{A}} \prod_{k=m}^{n} \frac{\exp\{-\frac{1}{2}(y_k - \mathbf{a}^T\mathbf{x}_k)^2\}}{\exp\{-\frac{1}{2}y_k^2\}}$$

for each $m \in \{1, \ldots, n\}$, which is equivalent to solve

$$\inf_{\mathbf{a}\in\mathcal{A}} \sum_{k=m}^{n} (y_k - \mathbf{a}^T\mathbf{x}_k)^2 \quad \text{for } m = 1, \ldots, n. \quad (17)$$

The challenges of solving this problem include

- (17) is a non-convex problem as 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 to use 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 (17) since 1) $\mathcal{A}$ is not the whole $s-$sparse space but possesses some special structure (3) and 2) 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.
- Even if the LASSO algorithm could work in solving (17), its computational complexity is high.

Because of above reasons, we are interested in finding algorithms with low computational complexity.

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. Specifically, the proposed detection procedure is described as follows:

$$W_i(m, n; a_i) := \kappa a_i \sum_{k=m}^{n} x_{i,k} y_k - \frac{\kappa}{2} a_i^2 \sum_{k=m}^{n} x_{i,k}^2, \quad \text{for } 1 \le i \le p, \quad (18)$$

$$U(m, n) := \sup_{\mathbf{a}\in\mathcal{A}} \sum_{i=1}^{p} W_i(m, n; a_i), \quad (19)$$

$$C_n := \sup_{1\le m\le n} U(m, n), \quad (20)$$

$$\tau_c := \inf\{n \ge 0 : C_n \ge \log B\}, \quad (21)$$

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

The motivation of the above algorithm is to use the correlation between y_k and $\mathbf{x}_k$ for the change-point detection. From (2), we see that in the transformed model, 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 $a_i \ne 0$, and a_i reflects the correlation strength between y_k and $x_{i,k}$. Actually, $W_i(m, n; a_i)$ defined in (18) is a measurement of the correlation between y_k and $x_{i,k}$. If the components in $\mathbf{x}_k$ are mutually independent, we notice that

$$\mathbb{E}^{\mathbf{a}}[W_i(m, n; a_i)] = (n-m+1)\frac{\kappa}{2} a_i^2 r_{i,i} = -\mathbb{E}_\infty[W_i(m, n; a_i)].$$

That is, the conditional expectation (conditioned on $\mathbf{a}$) of $W_i(m, n; a_i)$ after the change-point is opposite to its expectation before the change-point.

When change occurs at $\{t = m\}$, $W_i(m, n; a_i)$ 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 up all s positive W_i's to speed up the detection procedure. This idea is reflected by $U(m, n)$ in (19). As the change-point t is unknown, the observer then searches over all time instants within $[1, n]$ in (20) and detect the change-point via a threshold rule in (21). This follows a similar idea of constructing the CUSUM procedure from the one-side SPRT procedure [16].

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

Theorem III.2. *By setting*

$$\log B = 2\kappa s\left[\log\left(2p + p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{\max}]\right) + \log\gamma\right], \quad (22)$$

in which N_{max} is a finite random variable whose distribution relies on $f(\mathbf{x})$, and $\mathbb{E}[N_{\max}] \le c_1 p$ with c_1 being a constant independent of p. One can guarantee that

$$ARL2FA[\tau_c] \ge \gamma. \quad (23)$$

Furthermore, the detection delay is bounded by

$$WADD(\tau_c;\mathbf{a}) \le \frac{2|\log B|}{\kappa\sum_{i=1}^p a_i^2 r_{i,i} + 2\kappa\sum_{i\ne j} a_i a_j r_{i,j}}(1+o(1)) \quad (24)$$

as $\gamma \to \infty$.

Proof: Please see Section IV-A. ■

Remark III.3. *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*

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

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 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 (24) is on the order of $O(s)$ since there are only s non-zero components in $\mathbf{a}$; hence the detection delay $WADD(\tau_c;\mathbf{a}) \sim O(\log p)$. That is, the observer only needs $O(\log p)$ post-change observations on average to detect the change-point. Recall that in sparse recovery problem, one needs $O(s\log p)$ observations to recover an $s-$sparse vector. However, we require less observations for the purpose of detection.
3) From (22) and (24), 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:

$$W_i(m,n;a_i) := \kappa a_i\sum_{k=m}^n x_{i,k}y_k - \frac{\kappa}{2}a_i^2\sum_{k=m}^n x_{i,k}^2, \quad \text{for } 1\le i\le p, \quad (25)$$

$$V_i(m,n;a_i) := W_i(m,n;a_i) + (n-m+1)\mu, \quad (26)$$

$$U(m,n) := \sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^p V_i(m,n;a_i) = \sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^p W_i(m,n;a_i) + p(n-m+1)\mu, \quad (27)$$

$$C_n := \sup_{1\le m\le n} U(m,n), \quad (28)$$

$$\tau_c := \inf\{n\ge 0 : C_n \ge \log B\}. \quad (29)$$

With a little abuse of notations, we still use $U(m,n)$, C_n and τ_c in the Bayesian case to denote the detection procedure. However, these notations can be clearly distinguished from the ones for the non-Bayesian formulation 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 designed parameter μ in (26). Specifically, μ is a factor adopted 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

$$Y_k := \kappa y_k\sum_{i=1}^p a_i x_{i,k} - \frac{\kappa}{2}\sum_{i=1}^p (a_i x_{i,k})^2 + p\mu. \quad (30)$$

Since $\{\mathbf{x}_n\}$ is i.i.d. over n, on the event $\{t=m\}$, we have

$$\frac{1}{n}\sum_{k=m}^{m+n-1} Y_k \xrightarrow{a.s.} \mathbb{E}_1^{\mathbf{a}}[Y_k]$$

by the strong law of large numbers. Define

$$T_\epsilon := \inf\left\{n\ge 0 : \left|n^{-1}\sum_{k=m}^{m+n-1} Y_k - \mathbb{E}_1^{\mathbf{a}}[Y_k]\right| > \epsilon\right\}, \quad (31)$$

hence $T_\epsilon < \infty$ almost surely. We make further assumption that

$$\mathbb{E}_m^{\mathbf{a}}[T_\epsilon] < \infty \text{ and } \mathbb{E}_\pi^{\mathbf{a}}[T_\epsilon] < \infty \text{ for all } \mathbf{a}\in\mathcal{A}. \quad (32)$$

With Assumption (32), we have the following result:

Theorem III.4. *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 $\kappa < 1$,

$$\frac{s}{2p}\log(1-\kappa) \le \mu < \frac{s}{2p}\log(1-\kappa) + \frac{s}{p}|\log(1-\rho)|, \quad (33)$$

one can guarantee that

$$PFA[\tau_c] \le \alpha \quad (34)$$

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

$$ADD(\tau_c; \mathbf{a}) \leq \mathbb{E}_\pi^{\mathbf{a}}[\tau_c - t | \tau_c \geq t] = \frac{2|\log B| + 2c_3 s}{\kappa \sum_{i=1}^p a_i^2 r_{ii} + 2\kappa \sum_{i \neq j} a_i a_j r_{ij} + 2p\mu}(1 + o(1)) \quad (35)$$

as $\alpha \to 0$*, in which*

$$c_3 := \frac{1-\rho}{\rho}\left(\frac{1}{2}|\log(1-\kappa)| + \frac{\kappa}{2}\max_{1 \leq i \leq p} a_i^2 r_{i,i}\right) \quad (36)$$

is a constant that is independent of p*.*

Proof: Please see Section IV-B. ■

Remark III.5. *1) In the asymptotic analysis when* p*,* s *are constants and* $\alpha \to 0$*, it is easy to see that* (33) *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, the lower bound of detection delay*

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

in which

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

By adjusting the value of κ*, we can obtain a family of upper bounds for the detection delay. In this case, we have* $ADD(\tau_c; \mathbf{a}) \sim O(|\log\alpha|)$ *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* (33) *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* (35) *is on the order of* $O(s)$*, the detection delay* $ADD(\tau_c; \mathbf{a}) \sim O(\log p)$*. 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 of the Parallel-Sum Algorithms

The proposed parallel-sum algorithm can be easily computed. From (19) and (27), 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 show that $W_i(m, n; a_i)$ achieves its maximum at

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

if a_i has no constraint. Hence,

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

can be easily found. In particular, $\hat{a}_i = a_i^*$ if $a_i^* \in \mathcal{A}_i$ and $\hat{a}_i$ equals to one of the four candidates $\{-b_{i,2}, -b_{i,1}, d_{i,1}, d_{i,2}\}$ otherwise.

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

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

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

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

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 then discuss the computation 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 C_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 purposes, one can truncate the proposed algorithm by a window with length l_w. Specifically, one can modify C_n defined in (21) and (29) as

$$C_n := \sup_{n - l_w + 1 \leq m \leq n} U(m, n).$$

With this modification, the computational complexity will be limited to $O(l_w p)$ for each time slot. This kind of window based algorithms was first introduced in [27] and then is analyzed in detail in [20]. For our algorithm, we can choose l_w on the order of detection delay. For example, as pointed out in Remark III.3 and Remark III.5, the detection delay is $O(\log p)$ for high dimensional settings, which indicates that the detection procedure requires $O(\log p)$ post-change observations on average to detect the change-point. Hence, roughly speaking, to set l_w on the order of $O(\log p)$ can provide enough post-change observations for the detection.

IV. Proofs

A. Proof of Theorem III.2

In this subsection, we prove Theorem III.2 by exploring the relationship between Lorden's quickest detection problem and the one-sided SPRT problem.

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

$$H_0 : y_n = \mathbf{0}^T \mathbf{x}_n + \epsilon_n \text{ versus } H_1 : y_n = \mathbf{a}^T \mathbf{x}_n + \epsilon_n. \quad (44)$$

Denote $P_\infty(\cdot)$ and $P^{\mathbf{a}}(\cdot)$ as probability measures under H_0 and H_1, respectively. Note that (44) is a sequential hypothesis testing problem rather than a change-point detection problem. 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

$$\begin{aligned} &\text{minimize}_\tau \ \mathbb{E}^{\mathbf{a}}[\tau], \\ &\text{subject to } P_\infty(\tau < \infty) \le \alpha. \end{aligned} \quad (45)$$

The relationship between one-sided SPRT and Lorden's quickest detection formulation is revealed in [16]. We rewrite the corresponding result in our context as the following lemma.

Lemma IV.1. *(Lemma 1 in [16]) Suppose τ is a stopping time for one-sided SPRT problem with respect to $\{(\mathbf{x}_n, y_n)\}_{n=1}^{\infty}$ such that*

$$P_\infty(\tau < \infty) \le \alpha, \quad 0 < \alpha < 1. \quad (46)$$

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

$$\tau^* = \inf\{\tau_k + k - 1 | k = 1, 2, \ldots\}. \quad (47)$$

Then τ^* is also a stopping time, and for the problem formulation defined in (5) *it satisfies*

$$\mathit{ARL2FA}(\tau^*) \ge \frac{1}{\alpha} \quad (48)$$

and

$$\mathit{WADD}(\tau^*) \le \mathbb{E}^{\mathbf{a}}[\tau]. \quad (49)$$

Using this lemma, we will study the performance of following algorithm for the one-sided SPRT problem (45). Consider the detection procedure

$$\begin{aligned} W_i(1, n; a_i) &= \kappa a_i \sum_{k=1}^{n} x_{i,k} y_k - \frac{\kappa}{2} a_i^2 \sum_{k=1}^{n} x_{i,k}^2, \\ U_n &= \sup_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^{p} W_i(1, n; a_i), \\ \tau_1 &= \inf\{n \ge 0 : U_n \ge \log B\}. \end{aligned} \quad (50)$$

Let τ_k be the stopping time that applies τ_1 to $\{(\mathbf{x}_n, y_n)\}_{n=k}^{\infty}$. We note that τ_c defined in (21) can be equivalently written as $\tau_c = \inf\{\tau_k + k - 1 | k = 1, 2, \ldots\}$. As a result, due to Lemma IV.1, it is sufficient to study the performance of $\mathbb{E}^{\mathbf{a}}[\tau_1]$ and $P_\infty(\tau_1 < \infty)$ in (45).

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

$$\mathbb{E}^{\mathbf{a}}[\tau_1] \le \frac{2|\log B|}{\kappa \sum_{i=1}^{p} a_i^2 r_{i,i} + 2\kappa \sum_{i \ne j} a_i a_j r_{i,j}} (1 + o(1)). \quad (51)$$

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

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

$$\begin{aligned} &P_\infty(\tau_1 < \infty) \le \\ &\quad 2pB^{-\frac{1}{\kappa s}} + \left(p\sqrt{\frac{\kappa s}{4\pi}} + \mathbb{E}[N_{max}] \right) (\log B)^{-\frac{1}{4}} B^{-\frac{1}{2\kappa s}}, \end{aligned} \quad (52)$$

in which N_{max} is a finite random variable whose distribution relies on $f(\mathbf{x})$, and $\mathbb{E}[N_{max}] \le c_1 p$, where c_1 is a constant independent of p.

With above 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

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

Then, Theorem III.2 follows by exploring the result (48) and (49).

In the rest of this subsection, we provide proofs for Lemma IV.2 and Lemma IV.3.

Proof of Lemma IV.2:

In the following, we study the detection delay of test procedure τ_1 for the one-sided SPRT problem (45). Assume

that a genie knows the true post-change linear coefficient $\mathbf{a}$, and he uses the statistic

$$\begin{aligned}\tilde{U}_n &= \sum_{i=1}^{p}\left[\kappa a_i\sum_{k=1}^{n}x_{i,k}y_k-\frac{\kappa}{2}a_i^2\sum_{k=1}^{n}x_{i,k}^2\right]\\ &= \sum_{k=1}^{n}\left[\kappa y_k\sum_{i=1}^{p}a_ix_{i,k}-\frac{\kappa}{2}\sum_{i=1}^{p}(a_ix_{i,k})^2\right]\end{aligned}$$

for detection. Let $Y_k := \kappa y_k\sum_{i=1}^{p}a_ix_{i,k}-\frac{\kappa}{2}\sum_{i=1}^{p}(a_ix_{i,k})^2$, it is easy to see $\{Y_k\}$ is a sequence of i.i.d. random variables under the alternative hypothesis, and hence $\tilde{U}_n$ is a random walk. Moreover, it is easy to verify that

$$\mathbb{E}^{\mathbf{a}}[Y_k]=\frac{\kappa}{2}\sum_{i=1}^{p}a_i^2r_{i,i}+\kappa\sum_{i\neq j}a_ia_jr_{i,j}.$$

Let $\tilde{\tau}_1=\inf\{n\geq 1:\tilde{U}_n\geq\log B\}$, using Wald's identity and ignoring the overshoot, we can obtain

$$\mathbb{E}^{\mathbf{a}}[\tilde{\tau}_1]=\frac{2|\log B|}{\kappa\sum_{i=1}^{p}a_i^2r_{i,i}+2\kappa\sum_{i\neq j}a_ia_jr_{i,j}}$$

as $B\to\infty$. We note that $U_n\geq\tilde{U}_n$ since U_n takes supreme value over $\mathbf{a}\in\mathcal{A}$. As a result, we have $\tau_1<\tilde{\tau}_1$; hence Lemma IV.2 holds.

Proof of Lemma IV.3

In the following, we study the false alarm probability of τ_1 for the one-sided SPRT problem (45). Under P_∞, by solving $\frac{\partial}{\partial a_i}W_i(1,n;a_i)=0$, we can easily obtain that

$$a_i^*=\frac{\sum_{k=1}^{n}x_{i,k}\epsilon_i}{\sum_{k=1}^{n}x_{i,k}^2}. \tag{53}$$

By the strong law of large number, as $n\to\infty$, we have

$$a_i^*=\frac{\sum_{k=1}^{n}x_{i,k}\epsilon_k}{\sum_{k=1}^{n}x_{i,k}^2}=\frac{\frac{1}{n}\sum_{k=1}^{n}x_{i,k}\epsilon_k}{\frac{1}{n}\sum_{k=1}^{n}x_{i,k}^2}\to\frac{\mathbb{E}[x_{i,k}\epsilon_k]}{\mathbb{E}[x_{i,k}^2]}=0,\quad P_\infty-\text{ almost surely }. \tag{54}$$

Recall that $\mathcal{A}_i=\{a_i|a_i\in(-b_{i,2},-b_{i,1}]\cup[d_{i,1},d_{i,2})\}$. Therefore, (54) indicates that there exists a finite random variable N_i such that $-b_{i,1}<a_i^*<d_{i,1}$ almost surely when $n>N_i$. The distribution of N_i depends on the convergence rate of $\frac{1}{n}\sum_{k=1}^{n}x_{i,k}\epsilon_k$ and $\frac{1}{n}\sum_{k=1}^{n}x_{i,k}^2$, which further depends on the marginal distribution of $\mathbf{x}$. Furthermore, we have

$$W_k(1,n;a_k^*)=\frac{\kappa}{2}\frac{(\sum_{k=1}^{n}x_{i,k}\epsilon_k)^2}{\sum_{k=1}^{n}x_{i,k}^2}=\frac{\kappa}{2}\left(\sum_{k=1}^{n}w_k\epsilon_k\right)^2 \tag{55}$$

with

$$w_k=\frac{x_{i,k}}{\sqrt{\sum_{k=1}^{n}x_{i,k}^2}}. \tag{56}$$

Denote $H_n=\sum_{k=1}^{n}w_k\epsilon_k$. H_n can be viewed as a linear combination of Gaussian random variables with random weights satisfying $\sum_{k=1}^{n}w_k^2=1$; hence for any given realization of $\{w_1,\ldots,w_n\}$, H_n is distributed as $\mathcal{N}(0,1)$. Let $\mathbf{w}=[w_1,w_2,\ldots,w_n]$ and let ϕ denote the pdf of standard Gaussian distribution, the pdf of H_n can be calculated as

$$\begin{aligned}f(h_n) &= \int f(h_n,\mathbf{w})d\mathbf{w}=\int f(h_n|\mathbf{w})f(\mathbf{w})d\mathbf{w}\\ &= \int\phi(h_n)f(\mathbf{w})d\mathbf{w}=\phi(h_n).\end{aligned} \tag{57}$$

Hence, H_n is distributed as $\mathcal{N}(0,1)$ for any n. Therefore, $\frac{2}{\kappa}W_i(1,n;a_i^*)$ is χ_1^2 distributed.

Let $N_{max}:=\max_i N_i$; hence N_{max} is an almost sure finite random variable, and

$$\mathbb{E}[N_{max}]\leq\mathbb{E}\left[\sum_{i=1}^{p}N_i\right]\leq c_1p, \tag{58}$$

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

$$\begin{aligned}&P_\infty(\tau<\infty)\\ &=P_\infty\left[U_\tau>\log B\right]\\ &=P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B\right]\\ &\leq P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B,\tau\leq N\right]\\ &\quad+P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B,\tau>N_{max}\right]\\ &\quad+P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B,N_{max}\geq\tau>N\right]\end{aligned} \tag{59}$$

We then bound these three items on the right hand side of (59) individually. To bound the first item, we have to introduce some notations. Specifically, let

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

Follow a discussion that is similar from (39) to (42), one can easily obtain

$$U_\tau=\sum_{i=1}^{s}W_{(i)}(1,\tau;\hat{a}_{(i)}), \tag{60}$$

in which

$$\hat{a}_i=\arg\max_{a_i\in\mathcal{A}_i}W_i(1,n;a_i) \tag{61}$$

and

$$W_{(1)}(1,n;\hat{a}_{(1)})\geq W_{(2)}(1,n;\hat{a}_{(2)})\geq\cdots\geq W_{(p)}(1,n;\hat{a}_{(p)})$$

is the order statistic of $W_i(1,n;\hat{a}_i)$. Then, for the first item,

we have

$$
\begin{aligned}
&P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B,\tau\le N\right]\\
&\overset{(a)}{=}P_\infty\left[\sum_{i=1}^{s}W_{(i)}\left(1,\tau;\hat{a}_{(i)}\right)>\log B,\tau\le N\right]\\
&\le P_\infty\left[W_{(1)}\left(1,\tau;\hat{a}_{(1)}\right)>\frac{\log B}{s},\tau\le N\right]\\
&=\sum_{n=1}^{\infty}P_\infty\left[W_{(1)}\left(1,n;\hat{a}_{(1)}\right)>\frac{\log B}{s},\tau=n,\tau\le N\right]\\
&\le\sum_{n=1}^{N}P_\infty\left[W_{(1)}\left(1,n;\hat{a}_{(1)}\right)>\frac{\log B}{s}\right]\\
&\le\sum_{n=1}^{N}P_\infty\left[\exists i\in\{1,\dots,p\}\text{ such that }W_i(1,n;\hat{a}_i)>\frac{\log B}{s}\right]\\
&\le\sum_{n=1}^{N}\sum_{i=1}^{p}P_\infty\left[W_i(1,n;\hat{a}_i)>\frac{\log B}{s}\right]\\
&\overset{(b)}{\le}\sum_{n=1}^{N}\sum_{i=1}^{p}P_\infty\left[\frac{2}{\kappa}W_i(1,n;a_i^*)>\frac{2}{\kappa}\frac{\log B}{s}\right]\\
&=NpP_\infty\left[\chi_1^2>\frac{2}{\kappa}\frac{\log B}{s}\right]\\
&\overset{(c)}{\le}Np\sqrt{\frac{\kappa s}{4\pi}}\frac{1}{B^{1/\kappa s}[\log B]^{1/2}},
\end{aligned}
\tag{62}
$$

in which (a) is because of (60), (b) is because of definitions of $\hat{a}_i$ and a_i^*, and (c) is because of the tail bounds inequality

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

for a standard normal random variable X.

We then bound the second item in (59). For $x_{i,k}$ under P_∞, we generate another two probability measures $Q_b(x_{i,k},y_k)$ and $Q_d(x_{i,k},y_k)$. In particular, $Q_b(x_{i,k},y_k)$ is generated by linear transformation $y_k=-b_{i,1}x_{i,k}+\epsilon_k$ and $Q_d(x_{i,k},y_k)$ by $y_k=d_{i,1}x_{i,k}+\epsilon_k$. A direct calculation shows that the Radon-Nikodym derivatives of Q_b, Q_d and P_∞ for $(x_{i,1},\dots,x_{i,n},y_1,\dots,y_n)$ are given as

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

We note that when $\tau>N_{max}$, $\hat{a}_i$ defined in (61) equals to either $-b_{i,1}$ or $d_{i,1}$ because of (54) for all $i\in\{1,\dots,p\}$. Then, for the second item on the right hand side of (59), we have

$$
\begin{aligned}
&P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B,\tau>N_{max}\right]\\
&=P_\infty\left[\sum_{i=1}^{s}W_{(i)}\left(1,\tau;\hat{a}_{(i)}\right)>\log B,\tau>N_{max}\right]\\
&\le P_\infty\left[W_{(1)}\left(1,\tau;\hat{a}_{(1)}\right)>\frac{\log B}{s},\tau>N_{max}\right]\\
&\le\sum_{i=1}^{p}P_\infty\left[W_i\left(1,\tau;\hat{a}_i\right)>\frac{\log B}{s},\tau>N_{max}\right]\\
&=\sum_{i=1}^{p}P_\infty\left[W_i\left(1,\tau;\hat{a}_i\right)>\frac{\log B}{s},\right.\\
&\qquad\left.\{\hat{a}_i=-b_{i,1}\text{ or }\hat{a}_i=d_{i,1}\},\tau>N_{max}\right]\\
&\le\sum_{i=1}^{p}\left[P_\infty\left[W_i\left(1,\tau;-b_{i,1}\right)>\frac{\log B}{s}\right]\right.\\
&\qquad\left.+P_\infty\left[W_i\left(1,\tau;d_{i,1}\right)>\frac{\log B}{s}\right]\right]\\
&=\sum_{i=1}^{p}\left[\int_{\left\{W_i(1,\tau;-b_{i,1})>\frac{\log B}{s}\right\}}\frac{dP_\infty}{dQ_b}dQ_b\right.\\
&\qquad\left.+\int_{\left\{W_i(1,\tau;d_{i,1})>\frac{\log B}{s}\right\}}\frac{dP_\infty}{dQ_d}dQ_d\right]\\
&\overset{(a)}{\le}\sum_{i=1}^{p}\frac{1}{e^{\log B/\kappa s}}\left[Q_b\left[W_i\left(1,\tau;-b_{i,1}\right)>\frac{\log B}{s}\right]\right.\\
&\qquad\left.+Q_d\left[W_i\left(1,\tau;d_{i,1}\right)>\frac{\log B}{s}\right]\right]\\
&=\frac{2p}{B^{1/\kappa s}},
\end{aligned}
\tag{63}
$$

in which (a) holds because of inequalities (64) and (65) in the following

$$
\begin{aligned}
&\int_{\left\{W_i(1,\tau;d_{i,1})>\frac{\log B}{s}\right\}}\frac{dP_\infty}{dQ_d}dQ_d\\
&=\sum_{n=1}^{\infty}\int_{\left\{W_i(1,\tau;d_{i,1})>\frac{\log B}{s},\tau=n\right\}}\frac{dP_\infty}{dQ_d}dQ_d\\
&=\sum_{n=1}^{\infty}\int_{\left\{W_i(1,n;d_{i,1})>\frac{\log B}{s},\tau=n\right\}}\exp\left\{-\frac{1}{\kappa}W_i(1,n;d_{i,1})\right\}dQ_d\\
&\le\exp\left\{-\frac{1}{\kappa}\frac{\log B}{s}\right\}\sum_{n=1}^{\infty}\int_{\left\{W_i(1,n;d_{i,1})>\frac{\log B}{s},\tau=n\right\}}dQ_d\\
&=\exp\left\{-\frac{\log B}{\kappa s}\right\}Q_d\left[W_i\left(1,\tau;d_{i,1}\right)>\frac{\log B}{s}\right].
\end{aligned}
\tag{64}
$$

Similarly, we have

$$
\begin{aligned}
&\int_{\left\{W_i(1,\tau;-b_{i,1})>\frac{\log B}{s}\right\}}\frac{dP_\infty}{dQ_b}dQ_b\\
&\qquad\le\exp\left\{-\frac{\log B}{\kappa s}\right\}Q_b\left[W_i\left(1,\tau;-b_{i,1}\right)>\frac{\log B}{s}\right].
\end{aligned}
\tag{65}
$$

The third term in the right hand of (59) can be bounded by Markov inequality. Particularly

$$P_\infty\left[\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p}W_i(1,\tau;a_i)>\log B, N_{max}\geq\tau>N\right]$$
$$\leq P(N_{max}>N)\leq\frac{\mathbb{E}[N_{max}]}{N}. \quad (66)$$

By setting

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

and adding three bounds together, we obtain that

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

This ends the proof.

B. Proofs for the Bayesian setup

In this subsection, we prove Theorem III.4. In particular, Theorem III.4 can be obtained directly from following two supporting lemmas:

Lemma IV.4. *(Detection Delay) If $\kappa<1$ and $\mu\geq\frac{s}{2p}\log(1-\kappa)$, then as $B\to\infty$*

$$\mathbb{E}_\pi^{\mathbf{a}}[\tau_c-t|\tau_c\geq t]$$
$$\leq\frac{\log B+c_3 s}{\frac{\kappa}{2}\sum_{i=1}^{p}a_i^2r_{ii}+\kappa\sum_{i\neq j}a_ia_jr_{ij}+p\mu}(1+o(1)), \quad (67)$$

in which

$$c_3:=\frac{1-\rho}{\rho}\left(\frac{1}{2}|\log(1-\kappa)|+\frac{\kappa}{2}\max_{1\leq i\leq p}a_i^2r_{i,i}\right)$$

is a constant that is independent of p.

Lemma IV.5. *(False Alarm) If $\kappa<1$ and*

$$\frac{s}{2p}\log(1-\kappa)\leq\mu<\frac{s}{2p}\log(1-\kappa)+\frac{s}{p}|\log(1-\rho)|, \quad (68)$$

then for threshold B,

$$P_\pi^{\mathbf{a}}(\tau_c<t)\leq\frac{1}{B^{1/s}}\frac{\rho c_2}{c_2-1}\frac{1}{1-(1-\rho)c_2}, \quad \forall\mathbf{a}\in\mathcal{A}, \quad (69)$$

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

Theorem III.4 then can be proved by setting

$$\log B=s|\log\alpha|+s\log\frac{\rho c_2}{(c_2-1)[1-(1-\rho)c_2]}. \quad (70)$$

Putting this threshold into (69), we have $P_\pi^{\mathbf{a}}(\tau_c<t)\leq\alpha$ for all $\mathbf{a}\in\mathcal{A}$; hence the false alarm constraint $\sup_{\mathbf{a}\in\mathcal{A}}P_\pi^{\mathbf{a}}(\tau_c<t)\leq\alpha$ is satisfied. Putting (70) into (67), we will obtain the upperbound of the detection delay presented in Theorem III.4. In the rest of this subsection, we will prove above two supporting lemmas.

Proof of Lemma IV.4:

In the following, we study the detection delay of τ_c defined in (29) for the Bayesian formulation. Assume that a genie knows the true post-change linear coefficient $\mathbf{a}$, and he uses the statistic

$$\tilde{U}(m,n)$$
$$=\sum_{i=1}^{p}\left[\kappa a_i\sum_{k=m}^{n}x_{i,k}y_k-\frac{\kappa}{2}a_i^2\sum_{k=m}^{n}x_{i,k}^2+(n-m+1)\mu\right]$$
$$=\sum_{k=m}^{n}\left[\kappa y_k\sum_{i=1}^{p}a_ix_{i,k}-\frac{\kappa}{2}\sum_{i=1}^{p}(a_ix_{i,k})^2+p\mu\right]. \quad (71)$$

Let $Y_k:=\kappa y_k\sum_{i=1}^{p}a_ix_{i,k}-\frac{\kappa}{2}\sum_{i=1}^{p}(a_ix_{i,k})^2+p\mu$ (note that this is the same Y_k defined in (30)), it is easy to see $\{Y_k,k=m,m+1,m+2,\ldots\}$ is a sequence of i.i.d. random variable on the event $\{t=m\}$ and hence $\tilde{U}(m,n)$ is a random walk. Let

$$\tilde{\tau}_c=\inf\{n\geq0:\tilde{U}(1,n)\geq\log B\}. \quad (72)$$

Note that $\tilde{U}(1,n)\leq C_n$ since the definition of C_n takes supreme over $\{\mathbf{a}\in\mathcal{A}\}$ and over $\{1\leq m\leq n\}$; hence we have $\tilde{\tau}_c>\tau_c$. Then it is sufficient to find an upper bound for $\tilde{\tau}_c$.

On the event $\{t=m\}$, by the strong law of large numbers, we have

$$\frac{1}{n}\tilde{U}(m,m+n-1)$$
$$\overset{a.s.}{\to}\mathbb{E}_1^{\mathbf{a}}[Y_k]=\frac{\kappa}{2}\sum_{i=1}^{p}a_i^2r_{i,i}+\kappa\sum_{i\neq j}a_ia_jr_{i,j}+p\mu.$$

Rewrite the T_ϵ defined in (31) as

$$T_\epsilon=\inf\{n\geq0:|n^{-1}\tilde{U}(m,m+n-1)-\mathbb{E}_1^{\mathbf{a}}[Y_k]|>\epsilon\}. \quad (73)$$

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

$$\tilde{U}(m,\tilde{\tau}_c-1)>(\tilde{\tau}_c-m+1)(\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon)$$

or equivalently,

$$\tilde{\tau}_c-m+1<\frac{\tilde{U}(m,\tilde{\tau}_c-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}<\frac{\log B-\tilde{U}(1,m-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}.$$

Then we have

$$\tilde{\tau}_c-m+1$$
$$<\frac{\log B-\tilde{U}(1,m-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}\mathbf{1}_{\{\tilde{\tau}_c>T_\epsilon+(m-1)\}}+T_\epsilon\mathbf{1}_{\{\tilde{\tau}_c\leq T_\epsilon+(m-1)\}}$$
$$\leq\frac{\log B-\tilde{U}(1,m-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}+T_\epsilon. \quad (74)$$

Taking the conditional expectation on both sides, we have

$$\mathbb{E}_m^{\mathbf{a}}[\tilde{\tau}_c-m|\tilde{\tau}_c\geq m]$$
$$\leq\mathbb{E}_m^{\mathbf{a}}\left[\frac{\log B-\tilde{U}(1,m-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}+T_\epsilon\middle|\tilde{\tau}_c\geq m\right]$$
$$=\frac{\log B}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}-\mathbb{E}_m^{\mathbf{a}}\left[\frac{\tilde{U}(1,m-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k]-\epsilon}\middle|\tilde{\tau}_c\geq m\right]+\mathbb{E}_m^{\mathbf{a}}\left[T_\epsilon\middle|\tilde{\tau}_c\geq m\right].$$

As a result, we have

$$\mathbb{E}_\pi^{\mathbf{a}}[\tilde{\tau}_c - t|\tilde{\tau}_c \geq t] \leq \frac{\log B}{\mathbb{E}_1^{\mathbf{a}}[Y_k] - \epsilon} - \mathbb{E}_\pi^{\mathbf{a}}\left[\frac{\tilde{U}(1,t-1)}{\mathbb{E}_1^{\mathbf{a}}[Y_k] - \epsilon}\Big|\tilde{\tau}_c \geq t\right] + \mathbb{E}_\pi^{\mathbf{a}}\left[T_\epsilon\Big|\tilde{\tau}_c \geq t\right].$$

Since

$$\begin{aligned}\mathbb{E}_m^{\mathbf{a}}\left[\tilde{U}(1,m-1)\right] &= \mathbb{E}_\infty\left[\tilde{U}(1,m-1)\right] = \mathbb{E}_\infty\left[\sum_{k=1}^{m-1} Y_k\right] \\ &= \sum_{k=1}^{m-1}\mathbb{E}_\infty\left[\kappa y_k \sum_{i=1}^{p} a_i x_{i,k} - \frac{\kappa}{2}\sum_{i=1}^{p}(a_i x_{i,k})^2 + p\mu\right] \\ &= (m-1)\left[p\mu - \frac{\kappa}{2}\sum_{i=1}^{p} a_i^2 r_{i,i}\right] \end{aligned} \quad (75)$$

is bounded for any given p. As a result, we have

$$\begin{aligned}\mathbb{E}_\pi^{\mathbf{a}}\left[\tilde{U}(1,t-1)\right] &= \sum_{m=1}^{\infty}\pi_m \mathbb{E}_m^{\mathbf{a}}\left[\tilde{U}(1,m-1)\right] \\ &= \frac{1-\rho}{\rho}\left(p\mu - \frac{\kappa}{2}\sum_{i=1}^{p} a_i^2 r_{i,i}\right) \\ &\geq \frac{1-\rho}{\rho}\left(\frac{1}{2}\log(1-\kappa) - \frac{\kappa}{2}\max_{1\leq i\leq p} a_i^2 r_{i,i}\right)s, \end{aligned} \quad (76)$$

in which the last inequality is because of the condition $\mu \geq \frac{s}{2p}\log(1-\kappa)$. Recall ρ is the parameter in the geometric distribution. Let

$$c_3 := \frac{1-\rho}{\rho}\left(\frac{1}{2}|\log(1-\kappa)| + \frac{\kappa}{2}\max_{1\leq i\leq p} a_i^2 r_{i,i}\right)$$

be a constant that is independent of p. Since we have assumed that $\mathbb{E}_\pi^{\mathbf{a}}[T_\epsilon] < \infty$, and $\{\tilde{\tau}_c \geq t\}$ is an almost sure event as $B \to \infty$, by (76) we have

$$\begin{aligned}&\mathbb{E}_\pi^{\mathbf{a}}[\tilde{\tau}_c - t|\tilde{\tau}_c \geq t] \\ &\leq \left(\frac{\log B}{\mathbb{E}_1^{\mathbf{a}}[Y_k] - \epsilon} - \frac{\mathbb{E}_\pi^{\mathbf{a}}\left[\tilde{U}(1,t-1)\right]}{\mathbb{E}_1^{\mathbf{a}}[Y_k] - \epsilon}\right)(1+o(1)) \\ &\leq \frac{\log B + c_3 s}{\mathbb{E}_1^{\mathbf{a}}[Y_k] - \epsilon}(1+o(1)). \end{aligned} \quad (77)$$

Then, Lemma IV.4 follows the fact that ϵ is arbitrarily close to zero and that $\mathbb{E}_1^{\mathbf{a}}[Y_k] = \frac{\kappa}{2}\sum_{i=1}^{p} a_i^2 r_{i,i} + \kappa\sum_{i\neq j} a_i a_j r_{i,j} + p\mu$.

Proof of Lemma IV.5:

In the following, we study the false alarm probability of τ_c defined in (29) for the Bayesian formulation. 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 (27). 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

$$\begin{aligned}P_\infty(\tau_c \leq N) &= P_\infty\left(\max_{1\leq n\leq N}\exp\{C_n\} \geq B\right) \\ &= P_\infty\left(\max_{1\leq n\leq N}\exp\left\{\sup_{1\leq m\leq n}\sup_{\mathbf{a}\in\mathcal{A}}\sum_{i=1}^{p} V_i(m,n;a_i)\right\} \geq B\right) \\ &= P_\infty\left(\max_{1\leq n\leq N}\exp\left\{\sup_{1\leq m\leq n}\sum_{i=1}^{p} V_i(m,n;\hat{a}_i^*)\right\} \geq B\right) \\ &= P_\infty\left(\max_{1\leq n\leq N}\sup_{1\leq m\leq n}\prod_{i=1}^{p} e^{V_i(m,n;\hat{a}_i^*)} \geq B\right) \\ &= P_\infty\left(\max_{1\leq n\leq N}\sup_{1\leq m\leq n} e^{p(n-m+1)\mu}\prod_{i=1}^{p} e^{W_i(m,n;\hat{a}_i^*)} \geq B\right) \\ &\overset{(a)}{=} P_\infty\left(\max_{1\leq n\leq N}\sup_{1\leq m\leq n} e^{p(n-m+1)\mu}\prod_{i=1}^{s} e^{W_{(i)}(m,n;\hat{a}_{(i)})} \geq B\right) \\ &\leq P_\infty\left(\max_{1\leq n\leq N}\sup_{1\leq m\leq n} e^{\frac{p}{s}(n-m+1)\mu} e^{W_{(1)}(m,n;\hat{a}_{(1)})} \geq B^{\frac{1}{s}}\right), \end{aligned} \quad (78)$$

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*. Specifically, we have

$$\begin{aligned}W_{(1)}(m,n;\hat{a}_{(1)}) &= \max_{1\leq i\leq p} W_i(m,n;\hat{a}_i) \\ &= \max_{1\leq i\leq p}\sup_{a_i\in\mathcal{A}_i} W_i(m,n;a_i) \\ &= \max_{1\leq i\leq p}\sup_{a_i\in\mathcal{A}_i}\left[\kappa a_i\sum_{k=m}^{n} x_{i,k}y_k - \frac{\kappa}{2}\sum_{k=m}^{n}(a_i x_{i,k})^2\right] \\ &\leq \max_{1\leq i\leq p}\kappa\sum_{k=m}^{n}\sup_{a_i\in\mathcal{A}_i}\left[a_i x_{i,k}y_k - \frac{1}{2}(a_i x_{i,k})^2\right] \\ &\overset{(a)}{=} \max_{1\leq i\leq p}\frac{\kappa}{2}\sum_{k=m}^{n} y_k^2 = \frac{\kappa}{2}\sum_{k=m}^{n} y_k^2, \end{aligned} \quad (79)$$

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 (79) into (78), we have

$$\begin{aligned}&P_\infty(\tau_c \leq N) \\ &\leq P_\infty\left(\max_{1\leq n\leq N}\sup_{1\leq m\leq n} e^{\frac{p}{s}(n-m+1)\mu} e^{\frac{\kappa}{2}\sum_{k=m}^{n} y_k^2} \geq B^{\frac{1}{s}}\right) \\ &\leq P_\infty\left(\max_{1\leq n\leq N}\sum_{m=1}^{n}\prod_{k=m}^{n} e^{\frac{\kappa}{2}y_k^2 + \frac{p}{s}\mu} \geq B^{\frac{1}{s}}\right). \end{aligned} \quad (80)$$

Let

$$\begin{aligned}S_n &:= \sum_{m=1}^{n}\prod_{k=m}^{n}\exp\left\{\frac{\kappa}{2}y_k^2 + \frac{p}{s}\mu\right\} \\ &= (S_{n-1}+1)\exp\left\{\frac{\kappa}{2}y_n^2 + \frac{p}{s}\mu\right\}. \end{aligned} \quad (81)$$

We note that S_n could be a submartingale. Particularly, let $\mathcal{F}_n := \sigma\{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n\}$, we have

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

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 $\frac{s}{2p}\log(1-\kappa) \leq \mu$ guarantees $(1-\kappa)^{-\frac{1}{2}}\exp\left\{\frac{p}{s}\mu\right\} \geq 1$; hence we have $\mathbb{E}_\infty[S_n|\mathcal{F}_{n-1}] \geq S_{n-1}+1 > S_{n-1}$, i.e., S_n is a submartingale. In addition,

$$\mathbb{E}_\infty[S_n] = \sum_{m=1}^{n}\prod_{k=m}^{n}\mathbb{E}_\infty\left[\exp\left\{\frac{\kappa}{2}y_k^2 + \frac{p}{s}\mu\right\}\right]$$
$$= \sum_{m=1}^{n}\prod_{k=m}^{n}\left[(1-\kappa)^{-\frac{1}{2}}\exp\left\{\frac{p}{s}\mu\right\}\right] = \frac{c_2(c_2^n-1)}{c_2-1}, \quad (82)$$

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

$$P_\infty\left(\max_{1\leq n\leq N} S_n \geq B^{1/s}\right) \leq \frac{\mathbb{E}_\infty[S_N]}{B^{1/s}}. \quad (83)$$

Combining (80) and (82), we have

$$P_\infty(\tau_c \leq N) \leq P_\infty\left(\max_{1\leq n\leq N} S_n \geq B^{1/s}\right) \leq \frac{1}{B^{1/s}}\frac{c_2(c_2^N-1)}{c_2-1}.$$

Further,

$$\begin{aligned} P_\pi(\tau_c < t) &= \sum_{N=1}^{\infty}\pi_N P_\infty(\tau_c \leq N-1) \\ &\leq \sum_{N=1}^{\infty}\rho(1-\rho)^{N-1}\frac{1}{B^{1/s}}\frac{c_2}{c_2-1}c_2^{N-1} \\ &= \frac{1}{B^{1/s}}\frac{\rho c_2}{c_2-1}\frac{1}{1-(1-\rho)c_2}. \end{aligned} \quad (84)$$

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

V. 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 $\mathbf{R} = \mathrm{diag}[1.32, 1.18, 1.04, 0.93, 0.86, 0.84, 0.71, 0.64, 0.52, 0.42, 0.39, 0.28, 0.17, 0.14, 0.03]$. The theoretic results obtained in Section III do not rely on the distribution of $\mathbf{x}_n$. In the simulation, we test our proposed algorithm under two distributions: Gaussian distribution with zero mean and Possion distribution with its expectation shifted to zero.

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 $\mathbf{x}_n$ is Gaussian distributed, and the green line with diamonds is the performance when $\mathbf{x}_n$ is Possion distributed. The black dot-dash line is the lower bound of WADD for all detection algorithms, which is presented in Theorem III.1. The black dash-line is the upper bound of the parallel-sum algorithm, which is presented in Theorem III.2. 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 linearly with $\log\gamma$, and the computation complexity of this algorithm is low.

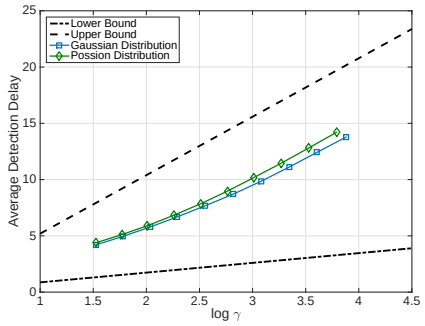

Fig. 1. WADD versus ARL2FA when $p = 15$, $s = 3$

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 irreverent to 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.4. The computational complexity of the proposed algorithm is low.

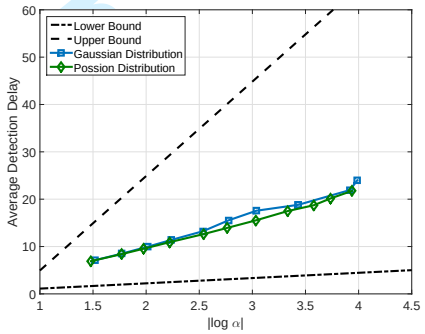

Fig. 2. ADD vs. PFA under when $p = 15$, $s = 3$

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 [28]. 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

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

by a 2-channel EMG system; hence each person totally 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 C_n over time is shown in Figure 3. As we can see, C_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.

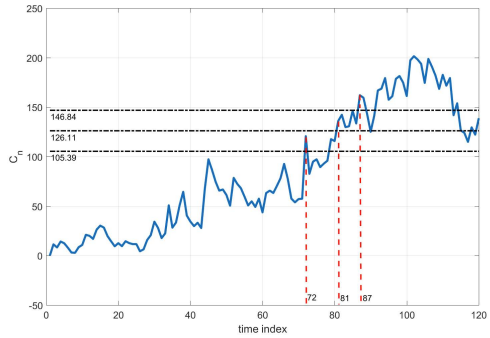

Fig. 3. The evolution of statics C_n over time slot

VI. 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.

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] X. Liu, D. Zhao, and R. Xiong et al., "Image interpolation via regularized local linear regression," *IEEE Trans. Image Processing*, vol. 20, pp. 3455–3469, Dec. 2011.

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

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

[7] 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.

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

[9] 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.

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

[11] 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.

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

[13] 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.

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

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

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

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

[18] 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).

[19] G. V. Moustakides, "Optimal stopping times for detecting changes in distribution," *Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[20] T. L. Lai, "Information bounds and quickest detection of parameter changes in stochastic systems," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[21] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic Bayesian theory of quickest change detection," *Theory of Probability and Its Applications*, vol. 49, no. 3, pp. 458–497, 2005.

[22] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

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

[24] C. Chu, M. Stinchcombe, and H. White, "Changecpoint monitoring in linear models," *Econometrica*, vol. 64, pp. 1045–1065, September 1996.

[25] A. Aue, L. Horvath, M. Huskova, and P. Kokoszka, "Changecpoint monitoring in linear models," *The Econometrics Journal*, vol. 9, pp. 373–403, September 2006.

[26] 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.

[27] A. S. Willsky and H. L. Jones, "A generalized likelihood ratio approach to detection and estimation of jumps in linear systems," *IEEE Trans. Automatic Control*, vol. AC-21, pp. 108–112, Feb. 1976.

[28] "sEMG for basic hand movements data set." https://archive.ics.uci.edu/ml/datasets/sEMG+for+Basic+Hand+movements.