SEQUENTIAL ANALYSIS
2020, VOL. 39, NO. 3, 307–335
https://doi.org/10.1080/07474946.2020.1823191

Sequential subspace change point detection

Liyan Xie[a], Yao Xie[a], and George V. Moustakides[b,c]

[a]School of Industrial and Systems Engineering (ISyE), Georgia Institute of Technology, Atlanta, Georgia, USA; [b]Department of Computer Science (CS), Rutgers University, New Brunswick, New Jersey, USA; [c]Department of Electrical and Computer Engineering (ECE), University of Patras, Patras, Greece

ABSTRACT

We consider the online monitoring of multivariate streaming data for changes that are characterized by an unknown subspace structure manifested in the covariance matrix. In particular, we consider the covariance structure changes from an identity matrix to an unknown spiked covariance model. We assume the postchange distribution is unknown and propose two detection procedures: the largest-eigenvalue Shewhart chart and the subspace-cumulative sum (CUSUM) detection procedure. We present theoretical approximations to the average run length (ARL) and the expected detection delay (EDD) for the largest-eigenvalue Shewhart chart and provide analysis for tuning parameters of the subspace-CUSUM procedure. The performance of the proposed methods is illustrated using simulation and real data for human gesture detection and seismic event detection.

ARTICLE HISTORY
Received 11 November 2019
Revised 6 January 2020, 30 January 2020
Accepted 28 February 2020

KEYWORDS
Change point detection; false alarm control; low rank; online algorithm; spiked covariance model

SUBJECT CLASSIFICATIONS
Primary 62L10; Secondary 62G10; 62G32

1. INTRODUCTION

Detecting the change from high-dimensional streaming data is a fundamental problem in various applications such as video surveillance (Sultani, Chen, and Shah 2018), sensor networks (Xie and Siegmund 2013), wearable sensors (Sprint, Cook, and Schmitter-Edgecombe 2016), and seismic events detection (Z. Li et al. 2018). In many scenarios, the change happens to the covariance structure and can be represented as a low-rank subspace to capture the similarity of signal waveforms observed at multiple sensors. We consider the fundamental problem of detecting such a change in the covariance matrix that shifts from an identity matrix to a spiked covariance model (Johnstone 2001). Different from the off-line hypothesis test considered in Berthet and Rigollet (2013), we assume a sequential setting, where the goal is to detect such a change as quickly as possible after it occurs.

A formal description of the problem is as follows. Assume a sequence of multivariate observations $x_1, x_2, \ldots, x_t, \ldots$, where $x_t \in \mathbb{R}^k$ and k is the data dimension. At a certain time τ, the distribution of the observation changes. In particular, we are interested in structural changes that happen to the covariance matrix, which we describe below: (1) the *emerging subspace*, meaning that the change is a subspace emerging from a noisy

CONTACT Yao Xie yao.c.xie@gmail.com School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.
Recommended by Tae Yang

background and thus the covariance matrix changes from an identity matrix to a spiked covariance matrix; (2) the *switching subspace*, meaning that the signals are along with different subspaces before and after the change, resulting in the covariance matrix changing from one spiked covariance matrix to another. The emerging subspace problem can arise, for instance, from weak signal detection for seismic sensor arrays (Sprint, Cook, and Schmitter-Edgecombe 2016), and the switching subspace detection can arise from monitoring principal component analysis for streaming data (Balzano, Chi, and Lu 2018). The switching subspace problem, as we will show, can be reduced to the emerging subspace problem; therefore, we focus on the emerging subspace problem.

The main contribution of this article is twofold. From the methodology perspective, we propose two sequential detection procedures: the largest-eigenvalue Shewhart chart and the subspace-cumulative sum (CUSUM) procedure. The largest-eigenvalue Shewhart chart computes the largest eigenvalue of the sample covariance matrix over a sliding window and detects a change when the statistic exceeds the threshold. The subspace-CUSUM is derived based on the likelihood ratio following the approach of classical CUSUM (Page 1954), but instead of assuming that the parameters are fully specified, we estimate the parameters and plug-in, which is analogous to the generalized likelihood ratio (GLR) statistic (Lai 1995). From the theoretical perspective, we provide a theoretical analysis of the proposed procedures, which facilitates efficient calibration of the parameters. We consider two commonly used metrics: the *average run length* (ARL) and the *expected detection delay* (EDD). Theoretical approximations can help us determine the threshold in the detection procedure efficiently. Moreover, building on Anderson's results for the distribution of eigenvectors (Anderson 1963), we provide theoretical guidelines on how to choose the parameters involved in the subspace-CUSUM procedure.

The proposed detection procedures are computationally efficient because they only require computing the leading eigenvalue and eigenvector of the sample covariance matrix, respectively. They are widely applicable to real data whenever there is a low-rank subspace change. For example, we have demonstrated its use in human activity detection from wearable sensors data and seismic event detection.

1.1. Related Work

In change point detection and industrial quality control, commonly used methods can be categorized into Shewhart chart, CUSUM, and GLR types of detection procedures.

Shewhart charts can be viewed as scan statistics over time. A change is detected when the process is out of control; that is, the detection statistic falls out of the control limit. A commonly used Shewhart chart for multivariate observations is the Hotelling T^2 control chart (Hotelling 1947), which can detect both mean and covariance shifts and the control limits are set through chi-square distributions. Modified T^2 charts based on principal component analysis are considered in Jackson (1959) and Jackson and Mudholkar (1979). The U^2 multivariate control chart in Runger (1996) considers detecting the mean shift in a known subspace. Those works do not consider the largest eigenvalue as a detection statistic.

Whereas Shewhart charts use the current subgroup samples to compute the detection statistic, the CUSUM procedure utilizes all past samples and updates the detection statistic recursively based on the log-likelihood ratio (Page 1954). A multivariate CUSUM procedure for detecting mean shift was developed in Pignatiello and Runger (1990), and a more recent work (Bodnar and Schmid 2005) presents CUSUM based on projected data. In classic CUSUM, the prechange and postchange distributions are specified completely. The subspace-CUSUM procedure here is not a typical CUSUM because we estimate the unknown subspace after the change.

Usually, the postchange distributions or their parameters are unknown and hard to prespecify. One solution is to set the postchange parameter to represent the "smallest possible change" of interest. However, when there is a parameter mismatch, the CUSUM procedure suffers from a performance loss. The GLR procedure is introduced to handle unknown postchange distributions (Lai 1995). The subspace-CUSUM procedure here is different from the GLR procedure because we do not estimate the full log-likelihood function; instead, we only estimate the subspace and introduce an additional parameter to control the performance.

Covariance shift detection has been considered in the past literature using various detection statistics. A multivariate CUSUM based on likelihood functions of multivariate Gaussian was studied in Healy (1987) considering a specific setting where the covariance changes from Σ to $c\Sigma$ for a constant c. The determinant of the sample covariance matrix was used in Alt (1985) and Alt and Smith (1988) to detect the covariance change. Chan and Zhang (2001) considered a CUSUM chart for monitoring covariance shift using the projection pursuit (Huber 1985) and likelihood ratio, with simulation studies on the performance of the proposed methods. Offline change detection of covariance change was studied in Chen and Gupta (2004) using the Schwarz information criterion (Schwarz 1978), where the change point location must satisfy certain regularity conditions to ensure the existence of the maximum likelihood estimator. Recently, Wang, Yu, and Rinaldo (2017) used wide binary segmentation through independent projection to recover the change points for the covariance matrix in the off-line setting. Avanesov and Buzun (2018) used the distance between empirical precision matrices to detect abrupt changes in covariance for the off-line case. Classical approaches usually consider the general setting, and here we are interested in detecting the structural change; that is, spiked covariance matrix.

Recent work has also considered other types of structured covariance changes. The detection of a shift in an off-diagonal submatrix of the covariance matrix was studied in Arias-Castro, Bubeck, and Lugosi (2012) using likelihood ratios. The detection of switching subspaces was studied in Jiao, Chen, and Gu (2018) based on a CUSUM-type procedure, but they only estimated the prechange subspace using historical data and assumed that the postchange subspace was known; this is different from our work because we also estimate the postchange subspace. Zhang et al. (2018) developed an off-line modeling framework for multivariate functional data based on sparse subspace clustering.

The largest-eigenvalue Shewhart chart is related to Berthet and Rigollet (2013), which studied the sparse principal component test based on sparse eigenvalue statistics. The largest eigenvalue statistic was shown to be asymptotically minimax optimal in Berthet and Rigollet (2013) for detecting whether there exists a sparse and low-rank component.

A natural sequential version of this idea is to use a sliding window and estimate the largest eigenvalue of the corresponding sample covariance matrix. However, this sequential version does not enjoy any form of (asymptotic) optimality.

1.2. Organization

The rest of the article is organized as follows. Section 2 presents the formulation of the emerging and switching subspace problems and shows a unified framework. Section 3 presents the proposed two sequential change detection procedures: the largest-eigenvalue Shewhart chart and subspace-CUSUM procedure. Section 4 presents theoretical approximations and bounds for the average run length and the expected detection delay of the largest-eigenvalue Shewhart chart, as well as theoretical calibration and parameter choice for subspace-CUSUM procedure. Section 5 contains simulation studies to demonstrate the performance of the proposed algorithms in different settings. Section 6 shows two real data examples using human gesture detection and seismic event detection. Finally, Section 7 contains our concluding remarks.

2. PROBLEM SETUP

We first introduce the spiked covariance model considered in Johnstone (2001), which assumes that a small number of directions explain most of the variance. For simplicity, we consider the spiked covariance model of rank 1 in this article. The results can be generalized to the case where rank is greater than one using similar ideas. In particular, the rank 1 spiked covariance matrix is given by

$$\Sigma = \sigma^2 I_k + \theta u u^\top,$$

where I_k denotes an identity matrix of size k; $\theta > 0$ is the signal strength; $u \in \mathbb{R}^k$ represents a basis for the subspace with unit norm $||u|| = 1$; and σ^2 is the noise variance, which will be considered known because it can be estimated from training data. It is possible to consider σ^2 unknown as well and provide estimates of this parameter along with the necessary estimates of u. However, to simplify our presentation, we decide to consider σ^2 known. The signal-to-noise ratio (SNR) is defined as $\rho = \theta/\sigma^2$.

Formally, the *emerging* subspace problem can be cast as follows

$$\begin{aligned} &x_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k), && t = 1, 2, \dots, \tau, \\ &x_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k + \theta u u^\top), && t = \tau + 1, \tau + 2, \dots \end{aligned} \tag{2.1}$$

where τ is the unknown change point that we would like to detect from data that are acquired sequentially. Similarly, the *switching* subspace problem can be formulated as follows:

$$\begin{aligned} &x_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2 I_k + \theta u_1 u_1^\top\right), && t = 1, 2, \dots, \tau, \\ &x_t \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2 I_k + \theta u_2 u_2^\top\right), && t = \tau + 1, \tau + 2, \dots \end{aligned} \tag{2.2}$$

where $u_1, u_2 \in \mathbb{R}^k$ represent bases for the subspaces before and after the change, with $||u_1||_2 = ||u_2||_2 = 1$ and u_1 is considered known. In both settings, our goal is to detect the change as quickly as possible, subject to the constraint that false detections occurring before the true change point are very rare.

The switching subspace problem (2.2) can be reduced into the emerging subspace problem (2.1) by a simple data projection. In detail, we can select any orthonormal matrix $Q \in \mathbb{R}^{(k-1)\times k}$ such that

$$Qu_1 = 0, \quad QQ^\top = I_{k-1},$$

which means that all rows of Q are orthogonal to u_1, and they are orthogonal to each other and have unit norm. Then, we project each observation x_t using the projection matrix Q onto a $k-1$ dimensional space and obtain a new sequence

$$y_t = Qx_t \in \mathbb{R}^{k-1}, t = 1, 2, \ldots.$$

Then y_t is a zero-mean random vector with covariance matrix $\sigma^2 I_{k-1}$ before the change and $\sigma^2 I_{k-1} + \theta Q u_2 u_2^\top Q^\top$ after the change. Let $u = Qu_2/||Qu_2||$, and

$$\tilde{\theta} = \theta||Qu_2||^2 = \theta\left[1 - \left(u_1^\top u_2\right)^2\right].$$

Thus, problem (2.2) can be reduced to the following:

$$\begin{aligned} y_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_{k-1}), & t &= 1, 2, \ldots, \tau, \\ y_t &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2 I_{k-1} + \tilde{\theta} u u^\top\right), & t &= \tau+1, \tau+2, \ldots \end{aligned} \tag{2.3}$$

Note that this way the switching subspace problem is reduced into the emerging subspace problem, where the new signal power $\tilde{\theta}$ depends on the angle between u_1 and u_2, which is consistent with our intuition.

We would like to emphasize that by projecting the observations onto a lower-dimensional space, we lose information, suggesting that the two versions of the problem are not exactly equivalent. Indeed, the optimum detector for the transformed data in (2.3) and the one for the original data in (2.2) do not coincide. This can be easily verified by computing the corresponding CUSUM tests and their optimum performance. Despite this difference, it is clear that with the proposed approach we put both problems under the same framework, offering computationally simple methods to solve the original problem in (2.2). Consequently, in the following analysis, we focus solely on problem (2.1).

3. DETECTION PROCEDURES

We propose two online methods: the largest-eigenvalue Shewhart chart and the subspace-CUSUM procedure. Below, denote by $\mathbb{P}_\tau$ and $\mathbb{E}_\tau$ the probability and expectation induced when there is a change point at the *deterministic* time τ. Under this definition, $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$ are the probability and the expectation under the nominal regime (a change never happens) and $\mathbb{P}_0$ and $\mathbb{E}_0$ are the probability and expectation under the alternative regime (a change happens before we take any data).

3.1. Largest-Eigenvalue Shewhart Chart

Motivated by the test statistic in Berthet and Rigollet (2013), we consider a Shewhart chart by computing the largest eigenvalue of the sample covariance matrix repeatedly over a sliding window. Assume that the window length is w. For each time $t > 0$, the *unnormalized* sample covariance matrix using the available samples is given by

$$\hat{\Sigma}_{t,\min\{t,w\}} = \begin{cases} x_1 x_1^\top + \cdots + x_t x_t^\top, & \text{for } t < w \\ x_{t-w+1} x_{t-w+1}^\top + \cdots + x_t x_t^\top, & \text{for } t \geq w. \end{cases} \tag{3.1}$$

We note that for $t = 1$ the matrix contains a single outer product and as time progresses the number of outer products increases linearly until it reaches w. After this point, namely, for $t \geq w$, the number of outer products remains equal to w.

Let $\lambda_{\max}(X)$ denote the largest eigenvalue of a symmetric matrix X. We define the largest-eigenvalue Shewhart chart as the one that stops according to the following rule:

$$T_{\mathrm{E}} = \inf\{t > 0 : \lambda_{\max}\left(\hat{\Sigma}_{t,\min\{t,w\}}\right) \geq b\}, \tag{3.2}$$

where $b > 0$ is a constant threshold selected to meet a suitable false alarm constraint. We need to emphasize that we do *not* divide by $\min\{t, w\}$ when forming the unnormalized sample covariance matrix. As we explain in Section 4.1, it is better for T_{E} to always divide by w instead of $\min\{t, w\}$. Consequently, we can omit the normalization constant w from our detection statistics by absorbing it into the threshold.

3.2. Subspace-CUSUM

The CUSUM procedure (Page 1954; Siegmund 1985) is the most popular sequential test for change detection. When the observations are independent and identically distributed before and after the change, CUSUM is known to be exactly optimum (Moustakides 1986) in the sense that it solves a very well-defined constrained optimization problem introduced in Lorden (1971). However, the CUSUM procedure can only be applied when we have exact knowledge of the pre- and postchange distributions. For our problem, this requires complete specification of all parameters, namely, the subspace u, noise power σ^2, and SNR ρ. In this section, we first introduce the exact CUSUM formulation and then present our subspace-CUSUM procedure.

3.2.1. Exact CUSUM

To derive the CUSUM procedure, let $f_\infty(\cdot), f_0(\cdot)$ denote the pre- and postchange probability density function of the observations. Then the CUSUM statistics are defined by maximizing the log-likelihood ratio statistic over all possible change point locations

$$S_t = \max_{1 \leq j \leq t} \sum_{i=j}^{t} \log \frac{f_0(x_i)}{f_\infty(x_i)},$$

which has the recursive implementation

$$S_t = (S_{t-1})^+ + \log \frac{f_0(x_t)}{f_\infty(x_t)}, \tag{3.3}$$

which enables its efficient calculation (Moustakides 1986), where $x^+ = \max\{0, x\}$. The CUSUM stopping time in turn is defined as

$$T_C = \inf\{t > 0 : S_t \geq b\}, \tag{3.4}$$

where b is a threshold selected to meet a suitable false alarm constraint. For our problem of interest (2.1) we can derive that

$$\begin{aligned}
\log \frac{f_0(x_t)}{f_\infty(x_t)} &= \log\left[\frac{\left[(2\pi)^k|\sigma^2 I_k + \theta u u^\top|\right]^{-1/2}}{\left[(2\pi)^k \sigma^{2k}\right]^{-1/2}} \times \frac{\exp\{-x_t^\top(\sigma^2 I_k + \theta u u^\top)^{-1} x_t/2\}}{\exp\{-x_t^\top x_t/(2\sigma^2)\}}\right] \\
&= \log\left[\left|I_k + \rho u u^\top\right|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}\frac{\theta}{\theta + \sigma^2}\frac{(u^\top x_t)^2}{\sigma^2}\right\}\right] \\
&= \frac{\rho}{2\sigma^2(1+\rho)}\left\{\left(u^\top x_t\right)^2 - \sigma^2\left(1 + \frac{1}{\rho}\right)\log(1+\rho)\right\}.
\end{aligned}$$

The second equality is due to the matrix inversion lemma (Woodbury 1950) that allows us to write

$$\left(\sigma^2 I_k + \theta u u^\top\right)^{-1} = \frac{1}{\sigma^2} I_k - \frac{\theta}{\theta + \sigma^2}\frac{u u^\top}{\sigma^2},$$

which, after substitution into the equation, yields the desired result. Note that the multiplicative factor $\rho/\left[2\sigma^2(1+\rho)\right]$ is positive, so we can omit it from the log-likelihood ratio when forming the CUSUM statistic in (3.3). This leads to

$$S_t = (S_{t-1})^+ + \left(u^\top x_t\right)^2 - \sigma^2\left(1 + \frac{1}{\rho}\right)\log(1+\rho). \tag{3.5}$$

Remark 3.1. We can show that the increment term in (3.5); that is,

$$\left(u^\top x_t\right)^2 - \sigma^2\left(1 + \frac{1}{\rho}\right)\log(1+\rho),$$

has the following property: its expected value is negative under the prechange probability measure and positive under the postchange probability measure. The proof relies on a simple argument based on Jensen's inequality (Rudin 2006). Due to this property, before the change, the CUSUM statistics S_t will oscillate near 0 and it will exhibit, on average, a positive drift after the occurrence of the change, forcing it, eventually, to hit or exceed the threshold.

3.2.2. Subspace-CUSUM Procedure

Usually the subspace u and SNR ρ are unknown. In this case it is impossible to form the exact CUSUM statistic depicted in (3.5). One option is to estimate the unknown parameters and substitute them back into the likelihood function. Here we propose to estimate only u and introduce a new *drift* parameter d that plays the same role as $\sigma^2(1 + 1/\rho)\log(1+\rho)$; this leads to the following subspace-CUSUM update:

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ + \left(\hat{u}_{t+w}^\top x_t\right)^2 - d. \tag{3.6}$$

To apply (3.6), we need to specify d and of course provide the estimate $\hat{u}_{t+w}$. Regarding the latter, we simply use the *unit norm* eigenvector corresponding to the largest eigenvalue of $\hat{\Sigma}_{t+w,w}$ depicted in (3.1). We denote the estimator of u as $\hat{u}_{t+w}$ because at time t the estimate will rely on the data $x_{t+1},...,x_{t+w}$ that are in the "future" of t. Practically, this is always possible by properly delaying our data by w samples. Stopping occurs similar to CUSUM; that is,

$$T_{\mathrm{SC}} = \inf\{t > 0 : \mathcal{S}_t \geq b\}. \tag{3.7}$$

Of course, in order to be fair, at the time of stopping we must make the appropriate correction; namely, if $\mathcal{S}_t$ exceeds the threshold at t for the first time, then the actual stopping takes place at $t+w$. The reason we use estimates based on "future" data is to make x_t and $\hat{u}_{t+w}$ *independent*, which in turn will help us decide what is the appropriate choice for the drift constant d in Section 4.3.

Remark 3.2. An alternative possibility is to use the GLR statistic, where both ρ and u are estimated for each possible change location κ. The GLR statistic is

$$\max_{\kappa<t}\left\{-\frac{t-\kappa}{2}\log\left(1+\hat{\rho}_{\kappa,t}\right)+\frac{1}{2\sigma^2}\frac{\hat{\rho}_{\kappa,t}}{1+\hat{\rho}_{\kappa,t}}\sum_{i=\kappa+1}^{t}\left(\hat{u}_{\kappa,t}^{\top}x_i\right)^2\right\},$$

where $\hat{\rho}_{\kappa,t}, \hat{u}_{\kappa,t}$ are estimated from samples $\{x_i\}_{i=\kappa+1}^{t}$. However, this computation is more intensive because there is no recursive implementation for the GLR statistic. Furthermore, it requires growing memory. There are finite memory versions (Lai and Shan 1999); however, they are equally complicated in their implementation. Therefore, we do not consider the GLR statistic in this article.

4. THEORETICAL ANALYSIS

To fairly compare the detection procedures discussed in the previous section, we need to calibrate them properly. The calibration process must be consistent with the performance measure we are interested in. For a given stopping time T we measure false alarms through the ARL expressed with $\mathbb{E}_\infty[T]$. For the detection capability of T we use the *worst-case* EDD defined in Lorden (1971),

$$\sup_{\tau\geq 0}\operatorname{ess\,sup}\mathbb{E}_\tau\left[(T-\tau)^+|T>\tau,x_1,...,x_\tau\right], \tag{4.1}$$

which considers the worst possible data before the change (expressed through the ess sup) and the worst possible change time τ.

In this section, we first discuss the scenarios that lead to the worst-case detection delay for the proposed procedures. Then we characterize the ARL and EDD of the largest-eigenvalue Shewhart chart. In doing so, we will also introduce some of the mathematical tools that can be used for the analysis of subspace-CUSUM. The theoretical characterization of ARL is very important because it can serve as a guideline on how to choose the threshold b used in the detection procedure. Without theoretical analysis, people usually use Monte Carlo simulation to set the threshold, which can be time consuming when the problem structure is complicated. Therefore, a theoretical way to choose the threshold can be beneficial, especially for online change point detection where computational efficiency is of great importance.

4.1. Worst-Case EDD

We now consider scenarios that lead to the worst-case detection delay. For the largest-eigenvalue Shewhart chart, assume that $1 \le t - w + 1 \le \tau < t$. Because for the detection we use $\lambda_{\max}(\hat{\Sigma}_{t,w})$ and compare it to a threshold, it is clear that the worst-case data before τ are the ones that will make $\lambda_{\max}$ as small as possible. We observe that

$$\begin{aligned}\lambda_{\max}(\hat{\Sigma}_{t,w}) &= \lambda_{\max}\left(x_{t-w+1}x_{t-w+1}^\top + \cdots + x_\tau x_\tau^\top + \cdots + x_t x_t^\top\right)\\ &\ge \lambda_{\max}\left(x_{\tau+1}x_{\tau+1}^\top + \cdots + x_t x_t^\top\right) = \lambda_{\max}(\hat{\Sigma}_{t,t-\tau}),\end{aligned}$$

which corresponds to the data $x_{t-w+1}, \ldots, x_\tau$, before the change, being all equal to zero. In fact, the worst-case scenario at any time instant τ is equivalent to forgetting all data before and including τ and restarting the procedure from $\tau + 1$ using up to w outer products in the unnormalized sample covariance matrix, exactly as we do when we start at time 0. Due to stationarity, this suggests that we can limit ourselves to the case $\tau = 0$ and compute $\mathbb{E}_0[T_{\mathrm{E}}]$ and this will constitute the worst-case EDD. Furthermore, the fact that in the beginning we do not normalize with the number of outer products is beneficial for T_{E} because it improves its ARL.

We should emphasize that if we do not force the data before the change to become 0 and use simulations to evaluate the detector with a change occurring at some time different from 0, then it is possible to arrive at misleading conclusions. Indeed, it is not uncommon for this test to appear to outperform the exact CUSUM test for low ARL values. Of course, this is impossible because the exact CUSUM is optimum for *any* ARL in the sense that it minimizes the worst-case EDD depicted in (4.1).

Let us now consider the worst-case scenario for subspace-CUSUM. We observe that

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ + \left(\hat{u}_{t+w}^\top x_t\right)^2 - d \ge 0 + \left(\hat{u}_{t+w}^\top x_t\right)^2 - d,$$

suggesting that when $\mathcal{S}_t$ restarts this is the worst that can happen for the detection delay. Therefore, the well-known property of the worst-case scenario in the exact CUSUM carries over to subspace-CUSUM. Again, because of stationarity, this allows us to fix the change point time at $\tau = 0$. Of course, as mentioned before, because $\hat{u}_{t+w}$ uses data coming from the future of t, if our detector stops at some time t (namely, when we experience $\mathcal{S}_t \ge b$ for the first time), then the *actual* time of stopping must be corrected to $t + w$. A similar correction is not necessary for CUSUM because this test has the exact information for all parameters.

Threshold b is chosen so that the ARL meets a prespecified value. In practice, b is usually determined by simulation. More specifically, by simulating multiple streams of data from prechange distribution, we can obtain the average run length for different thresholds. Therefore, the threshold can be determined by the simulation results.

A very convenient tool in accelerating the estimation of ARL (which is usually large) is the usage of the following formula that connects the ARL of CUSUM to the average of the sequential probability ratio test (SPRT) stopping time (Siegmund 1985)

$$\mathbb{E}_\infty[T_C] = \frac{\mathbb{E}_\infty[T_{\text{SPRT}}]}{\mathbb{P}_\infty(S_{T_{\text{SPRT}}} \geq b)}, \tag{4.2}$$

where the SPRT stopping time is defined as

$$T_{\text{SPRT}} = \inf\{t > 0 : S_t \notin [0, b]\}.$$

The validity of (4.2) relies on the CUSUM property that, after each restart, S_t is independent of the data before the time of the restart. Unfortunately, this key characteristic is no longer valid in the proposed subspace-CUSUM scheme because $\hat{u}_{t+w}$ uses data from the future of t. We could, however, argue that this dependence is weak. Indeed, as we will see in Lemma 4.1, each $\hat{u}_t$ is equal to u plus some small random perturbation (estimation error with the power of the order of $1/w$), with these perturbations being practically independent in time. As we observed with numerous simulations, estimating the ARL directly and through (4.2) (with S_t replaced by $\mathcal{S}_t$) results in almost indistinguishable values even for moderate window sizes w. This suggests that we can use (4.2) to estimate the ARL of the subspace-CUSUM as well. As we mentioned, in the final result, we need to add w to account for the future data used by the estimate $\hat{u}_{t+w}$.

4.2. Analysis of Largest-Eigenvalue Shewhart Chart

4.2.1. Approximate ARL of Largest-Eigenvalue Shewhart Chart

In this section, we first introduce some connection with random matrix theory, which provides the building blocks for the theoretical derivation. Then we provide the approximation to ARL as a function of threshold b after taking into account the *temporal correlation* between detection statistics. The comparison with simulation results shows the high accuracy of our results.

The study of ARL requires the understanding of the property of the largest eigenvalue under the null hypothesis; that is, the samples are independent and identically distributed Gaussian random vectors with zero mean and identity covariance matrix. To characterize the distribution of the largest eigenvalue, Johnstone (2001) used the Tracy-Widom law (Tracy and Widom 1996). Define the center constant $\mu_{w,k}$ and scaling constant $\sigma_{w,k}$:

$$\begin{aligned} \mu_{w,k} &= \left(\sqrt{w-1} + \sqrt{k}\right)^2, \\ \sigma_{w,k} &= \left(\sqrt{w-1} + \sqrt{k}\right)\left(\frac{1}{\sqrt{w-1}} + \frac{1}{\sqrt{k}}\right)^{1/3}. \end{aligned} \tag{4.3}$$

If $k/w \to \gamma < 1$, then the centered and scaled largest eigenvalue converges in distribution to a random variable W_1 with the so-called Tracy-Widom law of order one (Johnstone 2001):

$$\frac{\lambda_{\max}(\hat{\Sigma}_w) - \mu_{w,k}}{\sigma_{w,k}} \to W_1. \tag{4.4}$$

The Tracy-Widom law can be described in terms of a partial differential equation and the Airy function, and its tail can be computed numerically (using, for example, the R package RMTstat).

Remark 4.1 (Connection with random matrix theory). There has been an extensive literature on the distribution of the largest eigenvalue of the sample covariance matrix; see, for example, Johnstone (2001), Yin, Bai, and Krishnaiah (1988), Baik and Silverstein (2006), and Jiang, Leder, and Xu (2017). The so-called bulk (Edelman and Wang 2013) results are typically used for eigenvalue distributions. It treats a continuum of eigenvalues and the *extremes*, which are the (first few) largest and smallest eigenvalues. Assume that there are w samples that are k-dimensional Gaussian random vectors with zero mean and identity covariance matrix. Let $\hat{\Sigma}_w = \sum_{i=1}^{w} x_i x_i^\top$ denote the unnormalized sample covariance matrix. If $k/w \to \gamma > 0$, the largest eigenvalue of the sample covariance matrix converges to $w\big(1+\sqrt{\gamma}\big)^2$ almost surely (Geman 1980). Here we use the Tracy-Widom law to characterize its limiting distribution and tail probabilities.

If we ignore the temporal correlation of the largest eigenvalues produced by the sliding window, we can obtain a simple approximation for the ARL. If we call $p = \mathbb{P}_\infty\big(\lambda_{\max}\big(\hat{\Sigma}_{t,w}\big) > b\big)$ for $t \geq w$ then the probability of stopping at t is geometric and it is easy to see that the ARL can be expressed as $1/p$. We note that to obtain this result, we must assume that $\mathbb{P}_\infty\big(\lambda_{\max}\big(\hat{\Sigma}_{t,w}\big) > b\big) = p$ for $t < w$ as well, which is clearly not true. Because for $t < w$ the unnormalized sample covariance has less than w terms, the corresponding probability is smaller than p. This suggests that $1/p$ is a lower bound to the ARL and $w + 1/p$ is an upper bound. If $w \ll 1/p$, $\ll$ then approximating the ARL with $1/p$ is quite acceptable. We can use the Tracy-Widom law to obtain an asymptotic expression relating the ARL with the threshold b. The desired formula is depicted in the following theorem.

Theorem 4.1 (Approximation of ARL by ignoring temporal correlation). *For any* $0 < p \ll 1$ *we have* $\mathbb{E}_\infty[T_{\mathrm{E}}] \approx 1/p$, *if we select*

$$b = \sigma_{w,k} b_p + \mu_{w,k}, \tag{4.5}$$

where b_p *denotes the* p *upper percentage point of* W_1, *namely,* $\mathbb{P}\big(W_1 \geq b_p\big) = p$.

Now we aim to capture the temporal correlation between detection statistics due to overlapping time windows. We leverage a proof technique developed in Siegmund, Yakir, and Zhang (2010), which can obtain satisfactory approximation for the tail probability of the maximum of a random field.

Figure 1 illustrates the overlap of two sample covariance matrices and provides necessary notation. For each $\hat{\Sigma}_{t,w}$, define $Z_t = \lambda_{\max}\big(\hat{\Sigma}_{t,w}\big)$. We note that for any given $M > 0$,

$$\mathbb{P}_\infty(T \leq M) = \mathbb{P}_\infty\Big(\max_{1 \leq t \leq M} Z_t \geq b\Big),$$

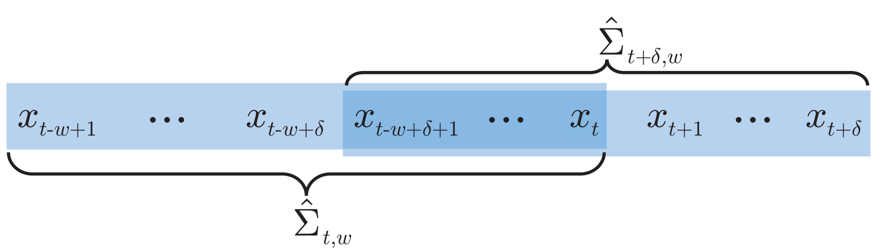

Figure 1. Illustration of the temporal correlation between largest eigenvalues, here $\delta \in \mathbb{Z}^+$.

Table 1. Comparison of the threshold b obtained from simulations and using the approximations ignoring the correlation in (4.5) and considering the correlation in (4.7). Window length $w = 200$, dimension $k = 10$. The numbers shown are b/w. Approximations that are closer to simulation values are indicated in boldface.

Target ARL	5K	10K	20K	30K	40K	50K
Simulation	1.633	1.661	1.688	1.702	1.713	1.722
Approx (4.5)	1.738	1.763	1.787	1.800	1.809	1.816
Approx (4.7)	**1.699**	**1.713**	**1.727**	**1.735**	**1.740**	**1.744**

which is the max over a set of correlated variables $\{Z_t\}_{t=1}^{M}$. Capturing the temporal dependence of $\{Z_t\}$ is challenging. Below, we assume that the dimension k and the window size w are fixed and consider the local covariance structure of the detection statistic when they only non-overlap at a small shift δ relative to the window size; that is, δ/w is small. By leveraging the properties of the local approximation, we can obtain an asymptotic approximation using the localization theorem (Siegmund, Yakir, and Zhang 2010). Define a special function $\nu(\cdot)$ that is closely related to the Laplace transform of the overshoot over the boundary of a random walk (Siegmund and Yakir 2007)

$$\nu(x) \approx \frac{\frac{2}{x}\left[\Phi\left(\frac{x}{2}\right) - 0.5\right]}{\frac{x}{2}\Phi\left(\frac{x}{2}\right) + \phi\left(\frac{x}{2}\right)}, \tag{4.6}$$

where $\phi(x)$ and $\Phi(x)$ are the probability density function and cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$. Then we have the following results.

Theorem 4.2 (ARL of largest-eigenvalue Shewhart chart). *For large values of b we can write*

$$\mathbb{E}_\infty[T_\mathrm{E}] = \left[b'\phi(b')\beta_{k,w}\nu\left(b'\sqrt{2\beta_{k,w}/w}\right)/w\right]^{-1}(1+o(1)), \tag{4.7}$$

where

$$\beta_{k,w} = 1 + \frac{\left(1 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)\left(2 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)}{c_2^2 k^{-\frac{1}{3}}/w}, \quad b' = \frac{b - (\mu_{w,k} + \sigma_{w,k}c_1)}{\sigma_{w,k}c_2},$$

with $c_1 = \mathbb{E}[W_1] = -1.21$ and $c_2 = \sqrt{\mathrm{Var}(W_1)} = 1.27$.

We perform simulations to verify the accuracy of the threshold values obtained without and with considering the temporal correlation (Theorem 4.1 and Theorem 4.2, respectively). The results are shown in Table 1. Compared with the thresholds obtained from Monte Carlo simulation, we find that the threshold, when temporal correlation (4.7) is taken into account, is more accurate than its counterpart obtained by using the Tracy-Widom law (4.5).

4.2.2. Lower Bound to EDD of Largest-Eigenvalue Shewhart Chart

We now focus on the detection performance and present a tight lower bound for the EDD of the largest-eigenvalue Shewhart chart. The results are based on a known result

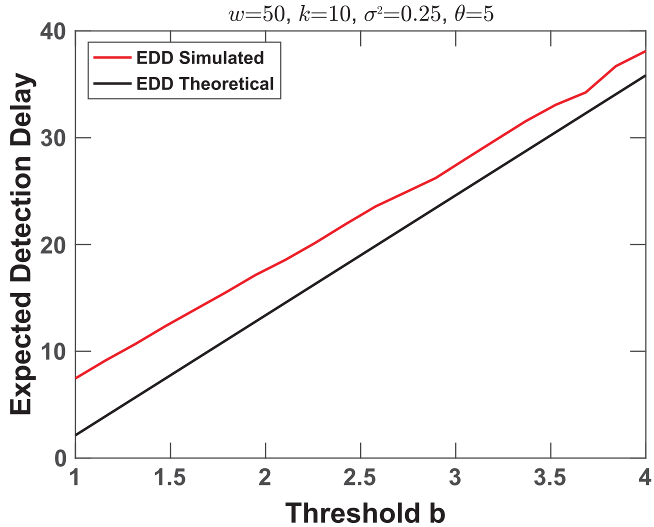

Figure 2. Simulated EDD and lower bound as a function of the threshold b.

for CUSUM (Siegmund 1985) and requires the derivation of the Kullback-Leibler divergence for our problem.

Theorem 4.3. *For large values of b we have*

$$\mathbb{E}_0[T_{\mathrm{E}}] \geq 2\frac{b' + e^{-b'} - 1}{-\log(1+\rho) + \rho}(1 + o(1)), \tag{4.8}$$

where

$$b' = \frac{1}{2\sigma^2(1+\rho)}\left[b\rho - (1+\rho)\sigma^2 \log(1+\rho)\right].$$

Consistent with intuition, in Theorem 4.3, the right-hand side of (4.8) is indeed a decreasing function of the SNR ρ. Comparing the lower bound in Theorem 4.3 with simulated average delay, as shown in Figure 2, we can show that in the regime of small detection delay (which is the main regime of interest), the lower bound serves as a reasonably good approximation.

4.3. Analysis of Subspace CUSUM

In this section, we focus on how to set the drift parameter d for the subspace-CUSUM procedure. This is an important parameter for the subspace-CUSUM to achieve desired properties of change point detection algorithms. For the drift parameter d we need the following double inequality to be true:

$$\mathbb{E}_\infty\left[\left(\hat{u}_{t+w}^\top x_t\right)^2\right] < d < \mathbb{E}_0\left[\left(\hat{u}_{t+w}^\top x_t\right)^2\right]. \tag{4.9}$$

With (4.9) we can guarantee that $\mathcal{S}_t$ mimics the behavior of the exact CUSUM statistic S_t mentioned in Remark 3.1, namely, it exhibits a negative drift before and a positive after the change. As we mentioned, the main advantage of using $\hat{\Sigma}_{t+w,w}$ is that it provides estimates $\hat{u}_{t+w}$ that are *independent* of x_t. This independence property allows for the straightforward computation of the two expectations in (4.9) and contributes toward the proper selection of d. Note that under the prechange distribution we can write

$$\mathbb{E}_\infty\left[\left(\hat{u}_{t+w}^\top x_t\right)^2\right] = \mathbb{E}_\infty\left[\hat{u}_{t+w}^\top \mathbb{E}_\infty\left[x_t x_t^\top\right]\hat{u}_{t+w}\right] = \sigma^2\mathbb{E}_\infty\left[\hat{u}_{t+w}^\top \hat{u}_{t+w}\right] = \sigma^2, \tag{4.10}$$

where the first equation is due to the independence of x_t and $\hat{u}_{t+w}$, the next one due to x_t having covariance $\sigma^2 I_k$, and the last equality is due to $\hat{u}_{t+w}$ being of unit norm.

Under the postchange regime, we need to specify the statistical behavior of $\hat{u}_{t+w}$ for the computation of $\mathbb{E}_0[(\hat{u}_{t+w}^\top x_t)^2]$. We will assume that the window size w is sufficiently large so that central limit theorem approximations (Anderson 1963; Paul 2007) are possible for $\hat{u}_{t+w}$. The required result appears in the next lemma.

Lemma 4.1. *Suppose vectors $x_1, ..., x_w$ are of dimension k and follow the distribution $\mathcal{N}(0, \sigma^2 I_k + \theta uu^\top)$. Let $\hat{\varphi}_w$ be the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix $(1/w)\left(x_1 x_1^\top + \cdots + x_w x_w^\top\right)$, then, as $w \to \infty$, we have the following central limit theorem version for $\hat{\varphi}_w$:*

$$\sqrt{w}(\hat{\varphi}_w - u) \to \mathcal{N}\left(0, \frac{1+\rho}{\rho^2}\left(I_k - uu^\top\right)\right).$$

Lemma 4.1 provides an asymptotic statistical description of the *unnormalized* estimate of u. More precisely, it characterizes the estimation error $v_w = \hat{\varphi}_w - u$. In our case we estimate the eigenvector from the matrix $\hat{\Sigma}_{t+w,w}$ but, as mentioned before, we adopt a *normalized* (unit norm) version $\hat{u}_t$. Therefore, if we fix w at a sufficiently large value and v_t denotes the estimation error of the unnormalized estimate at time t, then from Lemma 4.1 we can deduce

$$\hat{u}_{t+w} = \frac{\hat{\varphi}_{t+w}}{\|\hat{\varphi}_{t+w}\|} = \frac{u + v_{t+w}}{\|u + v_{t+w}\|}, \quad v_{t+w} \sim \mathcal{N}\left(0, \frac{1+\rho}{w\rho^2}\left(I_k - uu^\top\right)\right).$$

Consequently,

$$\mathbb{E}_0\left[\left(\hat{u}_{t+w}^\top x_t\right)^2\right] = \sigma^2(1+\rho)\left(1 - \frac{k-1}{w\rho} + o\left(\frac{1}{w}\right)\right), \tag{4.11}$$

with the $o(\cdot)$ term being negligible compared to the other two when $k/w \ll 1$, where a=o(b) denotes that $a/b \to 0$.

Consider now the case where ρ is *unknown* but exceeds some preset minimal SNR $\rho_{\min}$. From the above derivation, given the worst-case SNR and an estimation for the noise variance $\hat{\sigma}^2$, we can give a lower bound for $\mathbb{E}_0[(\hat{u}_{t+w}^\top x_t)^2]$. Consequently, the drift d can be anything between $\hat{\sigma}^2$ and $\hat{\sigma}^2(1+\rho_{\min})(1-(k-1)/(w\rho_{\min}))$ where we observe that the latter quantity exceeds $\hat{\sigma}^2$ when $w > (k-1)(1+\rho_{\min})/\rho_{\min^2}$. Below, for simplicity, for d we use the average of the two bounds. It is worthwhile mentioning that the lower bound (4.10) and upper bound (4.11) are derived based on the assumption that the window size w is large enough.

Remark 4.2 (Monte Carlo simulation to choose the threshold). Alternatively, and in particular when w does not satisfy $w \gg k$, where $\gg$means much greater than. we can estimate $\mathbb{E}_0[(\hat{u}_{t+w}^\top x_t)^2]$ by Monte Carlo simulation. This method requires (i) estimating the noise level $\hat{\sigma}^2$, which can be obtained from training data without a change point; (ii) the preset worst-case SNR $\rho_{\min}$; and (iii) a unit norm vector u_0 that is generated randomly. Under the nominal regime we have $\mathbb{E}_\infty[(\hat{u}_{t+w}^\top x_t)^2] = \hat{\sigma}^2$. Under the alternative, $\mathbb{E}_0[(\hat{u}_{t+w}^\top x_t)^2]$ depends only on the SNR ρ as shown in (4.11). We can therefore simulate the worst-case scenario $\rho_{\min}$ using the randomly generated vector u_0 by generating samples from the distribution $\mathcal{N}(0, \hat{\sigma}^2 I_k + \rho_{\min} u_0 u_0^\top)$.

Even though the average of the update in (3.6) does not depend on true subspace u, the computation of the test statistic $\mathcal{S}_t$ (3.6) requires the estimate $\hat{u}_{t+w}$ of the eigenvector. This can be accomplished by applying singular value decomposition (or the power method; Mises and Pollaczek-Geiringer 1929) on the unnormalized sample covariance matrix $\hat{\Sigma}_{t+w,w}$.

5. SIMULATION STUDY

In this section, numerical results are presented to compare the proposed detection procedures. The tests are first applied to synthetic data, where the performance of the subspace-CUSUM and largest-eigenvalue Shewhart chart are compared against the CUSUM optimum performance. Then the performance of subspace-CUSUM is optimized by selecting the most appropriate window size.

5.1. Performance Comparison

Simulation studies are performed to compare the largest-eigenvalue Shewhart chart and the subspace CUSUM procedure. The exact CUSUM procedure with all parameters known is chosen as the baseline and gives the minimal detection delay to all detection procedures.

Figure 3 depicts EDD versus log-ARL for parameter values $k = 5$, $\sigma^2 = 1$, $w = 50$ and three different levels of signal strength (SNR): $\theta = 0.5$, $\theta = 1$, and $\theta = 1.5$. For fair comparison, the SNR lower bound is set to be a constant $\rho_{\min} = 0.5$ in all three scenarios.

The threshold for each procedure is determined using the preset lower bound $\rho_{\min}$ as discussed in Remark 4.2. In Figure 3, the black line corresponds to the exact CUSUM procedure, which is clearly the best, and it lies below the other curves. Subspace-CUSUM has much smaller EDD than the largest-eigenvalue Shewhart chart, and the difference increases with increasing ARL for SNR $\theta = 0.5$ and $\theta = 1$. However, when the signal is stronger ($\theta = 1.5$), the largest-eigenvalue Shewhart chart outperforms the subspace-CUSUM as shown in Figure 3(c). This is consistent with previous research findings in Neuburger et al. (2017) that Shewhart charts are more efficient when detecting strong signals, whereas the CUSUM-type chart can detect weak signals more quickly due to its cumulative structure.

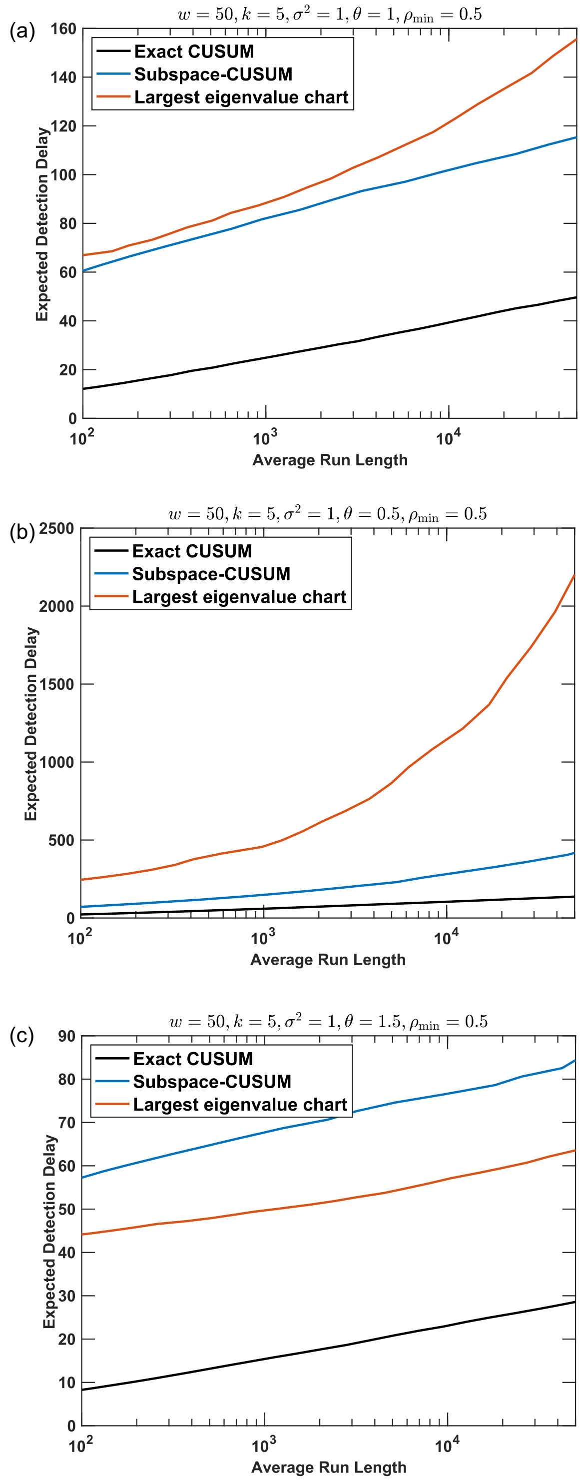

Figure 3. Comparison of subspace-CUSUM and the largest-eigenvalue Shewhart chart, fixed window size $w = 50$. Baseline: Exact CUSUM (optimal).

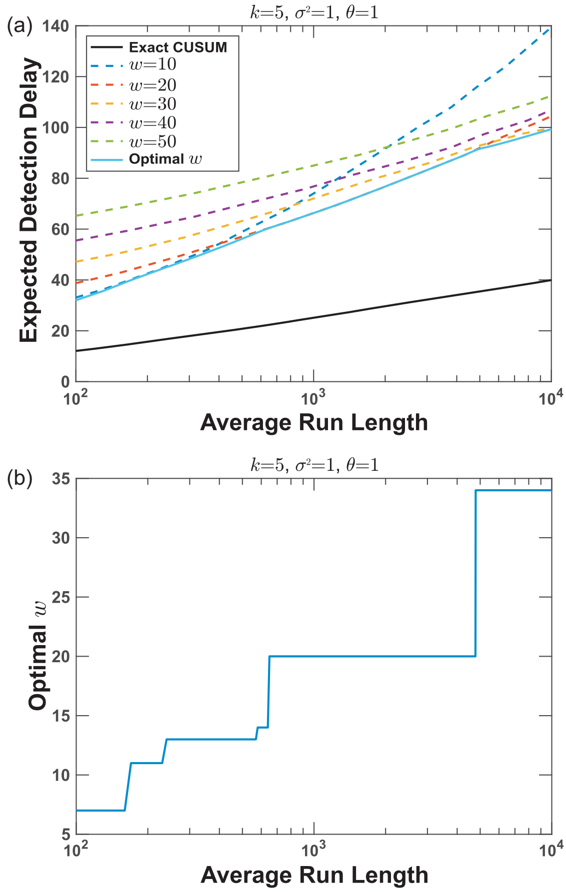

Figure 4. (a) Minimal EDD vs ARL among window sizes from 1 to 50 and (b) optimal window size resulting in smallest EDD.

5.2. Optimal Window Size

We also consider the EDD-ARL curve, where w is optimized to minimize the detection delay at every ARL. We first compute the EDD for window sizes $w = 1, 2, \ldots, 50,$ given each ARL value. Then we plot in Figure 4(a) the lower envelope of EDDs corresponding to the optimal EDD achieved by varying w. We also plot the optimal value of w as a function of ARL in Figure 4(b). Even though the best EDD of the subspace-CUSUM is diverging from the performance enjoyed by CUSUM, this divergence we believe is slower than the increase of the optimum CUSUM EDD. One of the goals in the future publication regarding the analysis of subspace-CUSUM is to show that this is indeed the case, which in turn will demonstrate that this detection structure is first-order asymptotically optimum.

6. REAL DATA EXAMPLES

In this section, we show how to apply the proposed methods to real data problems and demonstrate the performance using two real data sets: a human gesture detection data set and a seismic data set. It is worth mentioning that the model formulation is a fundamental problem in high-dimensional problems, and the proposed methods are widely applicable to a variety of applications.

6.1. Human Gesture Detection

We apply the proposed method to the sequential posture detection problem using a real data set: the Microsoft Research Cambridge-12 Kinect gesture data set (Fothergill et al. 2012). The cross-correlation structure of such multivariate functional data may change over time due to the posture change. Zhang et al. (2018) studied the same data set from the dynamic subspace learning perspective in the off-line setting; our goal is to detect the change point from sequential observations. This data set contains 18 sensors. At each time t, each sensor records the coordinates in the three-dimensional Cartesian coordinate system. Therefore, there are 54 attributes in total.

We select a subsequence with a posture change from "bow" to "throw" and we use the first 250 training samples to estimate the subspace before the change. Figure 5(a) shows the eigenvalues of the principal component analysis. We select r leading eigenvectors of the sample covariance matrix as our estimate of the prechange subspace. For

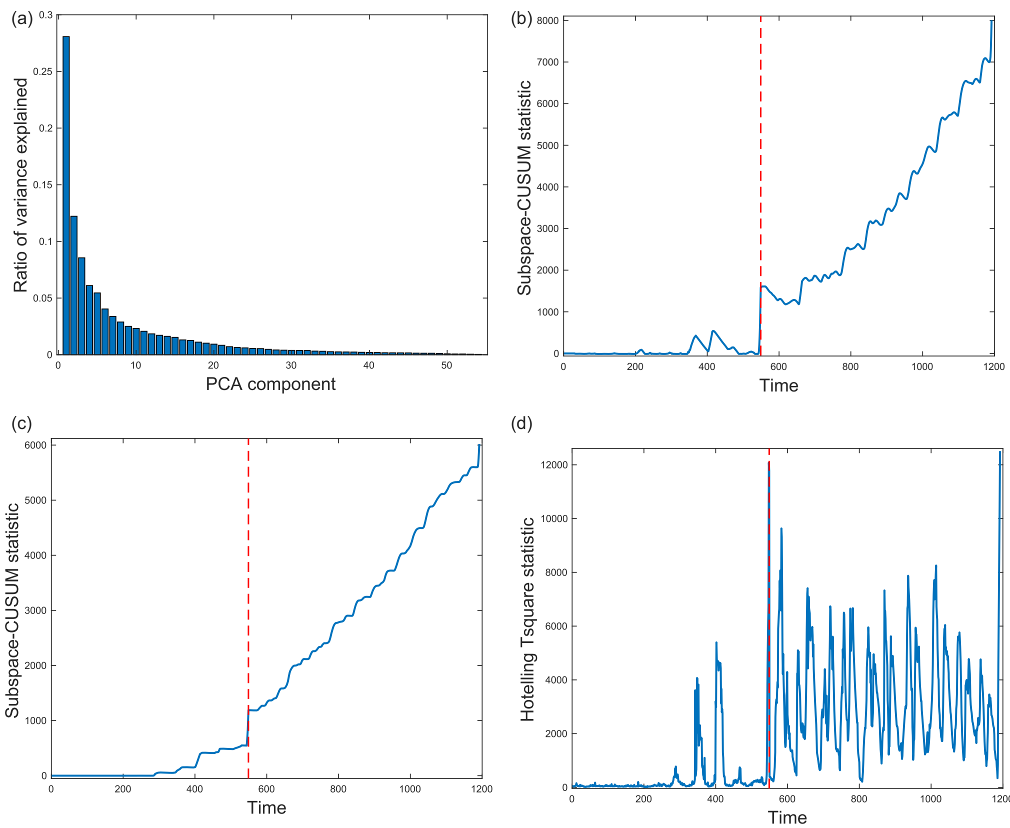

Figure 5. (a) Principal component analysis eigenvalues; (b), (c) subspace-CUSUM statistic over time; and (d) Hotelling T^2 statistic. True change point indicated by red line.

example, when $r = 1$, we estimate the prechange subspace to be a rank 1 space characterized by the leading eigenvector of the sample covariance matrix of training samples. Then we normalize the observations by multiplying them with a matrix Q that is orthogonal to the prechange subspace, as discussed in Section 2. This enables us to approximate the covariance of prechange observations by an *identity matrix*. Then we apply the proposed subspace-CUSUM procedure to detect the change.

The detection statistic is shown in Figures 5(b) and 5(c) for different r; the detection statistic indeed increased significantly at the true change point time (indicated by the red dash line). It also shows that the proposed test performs well not only when $r = 1$ but also for $r > 1$, which means that although we focus on the rank 1 case in the previous discussion, the proposed method can be widely used in many problems that involve such low-rank change. We also compare the proposed method with Hotelling's T^2 control chart (Hotelling 1947). We use the same training data to estimate the prechange mean $\bar{\mu}$ and covariance matrix $\bar{\Sigma}$ and then construct the Hotelling T^2 statistics $(x_t - \bar{\mu})^T \bar{\Sigma}^{-1}(x_t - \bar{\mu})$. As shown in Figure 5(d), the detection statistic has a much larger vibration than the subspace-CUSUM procedure and the performance is sensitive to the estimation of $\bar{\mu}$ and $\bar{\Sigma}$.

6.2. Seismic Event Detection

In this example, we consider a seismic signal detection problem. The goal is to detect micro-earthquakes and tremor-like signals, which are weak signals caused by minor subsurface changes in the Earth. The tremor signal may be seen at a subset of sensors, and the affected sensors observe a similar waveform corrupted by noise. The tremor signals are not earthquakes, but they are useful for geophysical study and prediction of potential earthquakes. Usually, the tremor signals are challenging to detect using an individual sensor's data; therefore, network detection methods have been developed, which mainly use covariance information of the data for detection (Z. Li et al. 2018). We will show that network-based detection can be cast as a subspace detection problem.

Assume that we have N sensors. At an unknown onset, the tremor signal may affect all sensors. Let $s(t)$ be the unknown signal waveform, then the signal observed at sensors can be represented as

$$x_i(t) = u_i s(t - \tau) + w_i(t), \; i = 1, 2, ..., n, \tag{6.1}$$

where $w_i(t) \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ denotes the random noise, $u_i > 0$ is the unknown *deterministic* magnitude of the signal, and τ is the unknown change point or the time when the seismic event happens. Here the waveform function $s(t)$ is assumed to be causal; that is, $s(t) = 0, \forall t < 0$. Moreover, we suppose that the signal waveform at time t follows a zero-mean normal distribution with time-varying variance (vibration); that is, $s(t) \sim \mathcal{N}(0, \sigma_t^2)$.

Denote the observation vector $X(t) := [x_1(t), ..., x_n(t)]^\top$ and magnitudes $u := [u_1, ..., u_n]^\top$. Following (6.1), we can formulate the problem as follows:

$$\begin{aligned} X(t) &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_n), & t &= 1, 2, ..., \tau, \\ X(t) &\stackrel{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2 I_n + \sigma_t^2 u u^\top\right), & t &= \tau + 1, \tau + 2, ... \end{aligned} \tag{6.2}$$

We apply the proposed methods to a real seismic data set recorded at Parkfield,

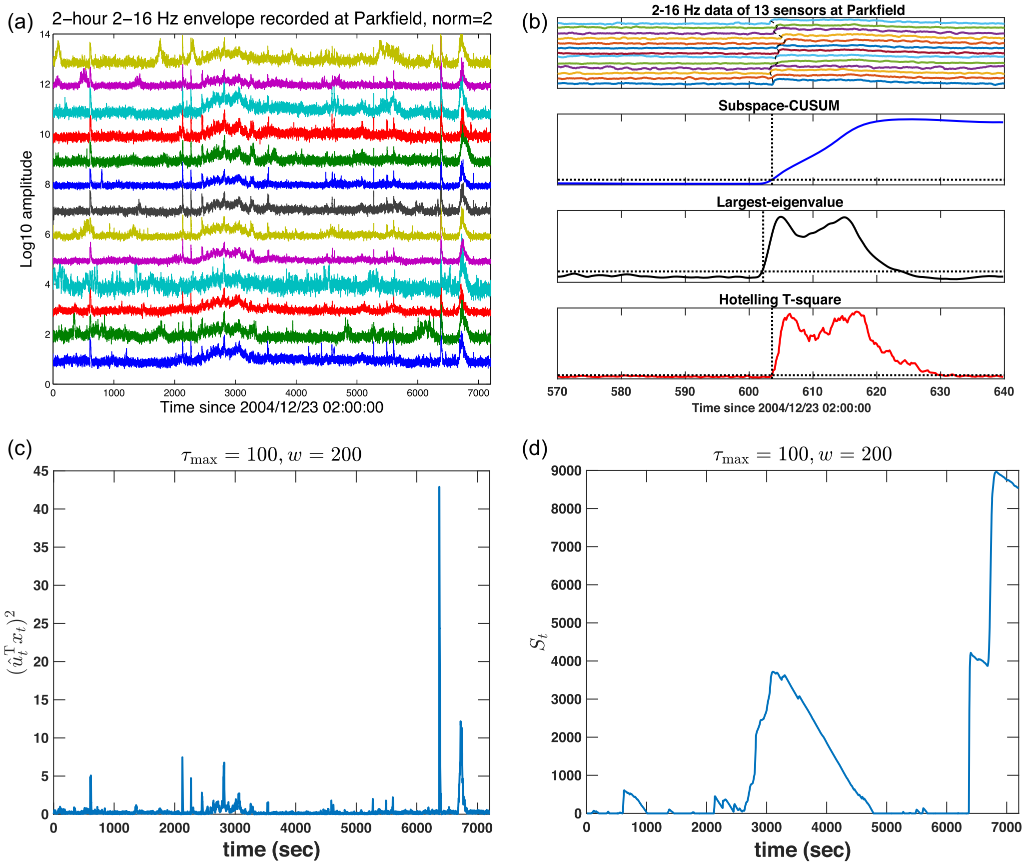

Figure 6. (a) Raw data; (b) comparison of different detection procedures; (c) increment term; and (d) subspace-CUSUM statistic.

California, from 2 a.m. to 4 a.m. on 23 December 2004. The raw data contain records at 13 seismic sensors that simultaneously record a continuous stream of data. The frequency of the raw data is 250 Hz. In this example, we set the window size $w = 200$, which corresponds to a 0.8 s time window. For each procedure, we use the data within the first 600 s to find the threshold by controlling the false alarm rate.

We apply the proposed largest-eigenvalue Shewhart chart and the subspace-CUSUM procedure. We further compare them with the classic Hotelling T^2 procedure based on the estimated sample mean and sample covariance. The results are shown in Figure 6(b). Using the detection statistics in Figures 6(c), and 6(d), we find three main events at 615, 2,127, and 6,371 s, as well as some continuous vibration during 2,500–3,200 s. By comparing the detection results with the true seismic event catalog that can be found online at the Northern California Earthquake Data Center, we found that our findings match the three true events at 594, 2,124, and 6,369 s, along with a tremor catalog around 2,500–3,180 s. The comparison shows that all detection delays are within 20 s.

Both the largest-eigenvalue Shewhart chart and the subspace-CUSUM procedure work effectively for this data set.

7. DISCUSSION AND CONCLUSION

In this article, we study two online detection procedures for detecting the emergence of a spiked covariance model: the largest-eigenvalue Shewhart chart and the subspace-CUSUM control chart. For subspace-CUSUM, we perform a simultaneous estimate of the required subspace in parallel with its sequential detection. We avoid estimating all unknown parameters by following a worst-case analysis with respect to the subspace power. We were able to derive theoretical expressions for the ARL and an interesting lower bound for the EDD of the largest-eigenvalue Shewhart chart. In particular, we were able to handle the correlations resulting from the usage of a sliding window, which is an issue that is not present in the off-line version of the same procedure. For the comparisons of the two proposed detection procedures, we discuss how it is necessary to calibrate each detector so that comparisons are fair. Comparisons were performed using simulated data and real data about human gesture detection and seismic event detection.

APPENDIX A: PROOF OF LEMMA 4.1

We have the following asymptotic distribution (Anderson 1963):

$$\frac{1}{\sqrt{w}}(\hat{\varphi}_w - u) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=2}^{k} \frac{\lambda_1 \lambda_j}{(\lambda_1 - \lambda_j)^2} \nu_j \nu_j^\top\right),$$

where λ_j is the jth largest eigenvalue of the true covariance matrix and ν_j is the corresponding eigenvector. In our case the true covariance matrix is $\sigma^2 I_k + \theta u u^\top$; therefore, $\lambda_1 = \sigma^2 + \theta$ and $\lambda_j = \sigma^2$ for $j \geq 2$, and $\{\nu_j, j \geq 2\}$ is a basis of the orthogonal space of u. Thus, we have

$$\sum_{j=2}^{k} \frac{\lambda_1 \lambda_j}{(\lambda_1 - \lambda_j)^2} \nu_j \nu_j^\top = \frac{\sigma^2(\sigma^2 + \theta)}{\theta^2} \sum_{j=2}^{k} \nu_j \nu_j^\top = \frac{\sigma^2(\sigma^2 + \theta)}{\theta^2}\left(I_k - uu^\top\right) = \frac{(1+\rho)}{\rho^2}\left(I_k - uu^\top\right).$$

This completes the proof. □

APPENDIX B: PROOF OF THEOREM 4.2

In order to prove Theorem 4.2, we need the following lemma to characterize the local correlation between largest eigenvalue statistics.

Lemma B.1 (Approximation of local correlation). *Let* $c_1 = \mathbb{E}[W_1] = -1.21, c_2 = \sqrt{\mathrm{Var}(W_1)} = 1.27$ *and*

$$\beta_{k,w} = 1 + \frac{\left(1 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)\left(2 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)}{c_2^2 k^{-\frac{1}{3}}/w}.$$

Then we have

$$\mathrm{corr}(Z_t, Z_{t+\delta}) \leq 1 - \beta_{k,w}\vartheta + o(\vartheta),$$

where $\vartheta = \delta/w$ *and* $\mathrm{corr}(X, Y)$ *stands for the Pearson's correlation*

$$\mathrm{corr}(X,Y)=\frac{\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Proof. Under the prechange measure, $x_t \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2 I_k)$. For $\delta\in\mathbb{Z}^+$, let

$$P=\sum_{i=t-w+1}^{t-w+\delta} x_i x_i^\top,\quad Q=\sum_{i=t-w+\delta+1}^{t} x_i x_i^\top,\quad R=\sum_{i=t+1}^{t+\delta} x_i x_i^\top.$$

Then P, Q, and R are independent random matrices. Now we also want to give a general upper bound for the covariance between Z_t and $Z_{t+\delta}$. Then we have

$$\begin{aligned}\mathbb{E}[Z_tZ_{t+\delta}]&=\mathbb{E}\left[\lambda_{\max}\left(\hat{\Sigma}_{t,w}\right)\lambda_{\max}\left(\hat{\Sigma}_{t+\delta,w}\right)\right]=\mathbb{E}[\lambda_{\max}(P+Q)\lambda_{\max}(Q+R)]\\&\le\mathbb{E}\left\{[\lambda_{\max}(P)+\lambda_{\max}(Q)][\lambda_{\max}(Q)+\lambda_{\max}(R)]\right\}\\&=\mathbb{E}[\lambda_{\max^2}(Q)]+\mathbb{E}[\lambda_{\max}(Q)]\mathbb{E}[\lambda_{\max}(R)]\\&\quad+\mathbb{E}[\lambda_{\max}(P)]\{\mathbb{E}[\lambda_{\max}(Q)]+\mathbb{E}[\lambda_{\max}(R)]\},\end{aligned}$$

where the inequality is due to the fact that the largest eigenvalue of the sum of two nonnegative definite matrices is upper bounded by the sum of the corresponding largest eigenvalues of the two matrices. The mean and second-order moments can be computed using the Tracy-Widom law depicted in (4.4).

Because k is a fixed constant, we just write μ_n and σ_n instead of $\mu_{n,k}$ and $\sigma_{n,k}$ to simplify our notation. We first consider the covariance term $\mathbb{E}[Z_tZ_{t+\delta}]$ and decompose it into four parts as follows:

$$\frac{1}{w^2}\mathbb{E}[Z_tZ_{t+\delta}]\le A+B+C+D,$$

where

$$\begin{aligned}A&=\left(\frac{\mu_{w(1-\vartheta)}+c_1\sigma_{w(1-\vartheta)}}{w}\right)^2,\\B&=\left(\frac{c_2\sigma_{w(1-\vartheta)}}{w}\right)^2,\\C&=2\left[\frac{\mu_{w(1-\vartheta)}+c_1\sigma_{w(1-\vartheta)}}{w}\right]\left(\frac{\mu_{w\vartheta}+c_1\sigma_{w\vartheta}}{w}\right),\\D&=\left(\frac{\mu_{w\vartheta}+c_1\sigma_{w\vartheta}}{w}\right)^2.\end{aligned}$$

First, the common terms $\mu_{w(1-\vartheta)}/w$ and $\sigma_{w(1-\vartheta)}/w$ can be written as

$$\begin{aligned}\frac{\mu_{w(1-\vartheta)}}{w}&=\frac{1}{w}\left[\sqrt{w(1-\vartheta)-1}+\sqrt{k}\right]^2=\frac{w(1-\vartheta)-1}{w}\left[1+\sqrt{\frac{k}{w(1-\vartheta)-1}}\right]^2\\&\doteq\frac{w(1-\vartheta)-1}{w}\doteq 1-\vartheta,\end{aligned}$$

where the second term in the brackets was ignored because $k/w=o(1)$. Moreover, we have

$$\frac{\sigma_{w(1-\vartheta)}}{w}=\frac{1}{w}\left(\sqrt{w(1-\vartheta)-1}+\sqrt{k}\right)\cdot\left(\frac{1}{\sqrt{w(1-\vartheta)-1}}+\frac{1}{\sqrt{k}}\right)^{1/3}.$$

After extracting the term $\sqrt{w(1-\vartheta)-1}$ from the first bracket and $1/\sqrt{k}$ from the second bracket, we obtain

$$\frac{\sigma_{w(1-\vartheta)}}{w}=\frac{k^{-\frac{1}{6}}}{w}\sqrt{w(1-\vartheta)-1}\left(1+\sqrt{\frac{k}{w(1-\vartheta)-1}}\right)^{\frac{4}{3}}\doteq\sqrt{\frac{1-\vartheta}{w}}k^{-\frac{1}{6}},$$

where the second term in the brackets was ignored because $k/w=o(1)$. Plugging these two terms into the first part, we have

$$\begin{aligned}A=&\left(\frac{\mu_{w(1-\vartheta)}+c_1\sigma_{w(1-\vartheta)}}{w}\right)^2\doteq\left(1-\vartheta+c_1\sqrt{\frac{1-\vartheta}{w}}k^{-\frac{1}{6}}\right)^2\\=&(1-\vartheta)\left(1-\vartheta+2c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\sqrt{1-\vartheta}+c_1^2\frac{k^{-\frac{1}{3}}}{w}\right).\end{aligned}$$

Because ϑ is relatively small, $\sqrt{1-\vartheta}=1-\frac{1}{2}\vartheta+o(\vartheta)$ by Taylor expansion. Then the term is approximately

$$\begin{aligned}A\doteq&(1-\vartheta)\left(1-\vartheta+2c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\left(1-\frac{1}{2}\vartheta+o(\vartheta)\right)+c_1^2\frac{k^{-\frac{1}{3}}}{w}\right)\\=&\left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2-\left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\left(2+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\vartheta+o(\vartheta),\end{aligned}$$

where the higher order terms of ϑ are included in $o(\vartheta)$. Parts C and D can be considered negligible, because C is order $\mathcal{O}(\vartheta)$ and D is order $o(\vartheta)$. In summary, we have

$$\begin{aligned}\mathrm{corr}(Z_t,Z_{t+\delta})=&\frac{\mathbb{E}[Z_tZ_{t+\delta}]-\mathbb{E}[Z_t]\mathbb{E}[Z_{t+\delta}]}{\sqrt{\mathrm{Var}(Z_t)}\sqrt{\mathrm{Var}(Z_{t+\delta})}}\\ \lesssim&\frac{1}{\left(\frac{c_2k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2}\{\left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2+c_2^2\frac{1-\vartheta}{w}k^{-\frac{1}{3}}-\left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\left(2+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\vartheta\\&-\left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2+o(\vartheta)\}\\=&1-\beta_{k,w}\vartheta+o(\vartheta)\end{aligned}$$

This completes the proof. □

The key to proving Theorem 4.2 is to quantify the tail probability of the detection statistic. However, this probability is very small when the threshold is large (S. Li et al. 2015). Therefore, we use the change-of-measure technique in Siegmund, Yakir, and Zhang (2010) to recenter the process mean to the threshold, so that the tail probability becomes much higher. First, the detection statistic is standardized by

$$Z_t'=\frac{Z_t-\mathbb{E}_\infty[Z_t]}{\mathrm{Var}_\infty(Z_t)}.$$

Here, $\mathbb{E}_\infty[Z_t]$ and $\mathrm{Var}_\infty(Z_t)$ depend only on k and w but do not depend on t. Then Z_t' has zero mean and unit variance under the $\mathbb{P}_\infty$ measure. We are interested in finding the probability $\mathbb{P}_\infty(T_{\mathrm{E}}\le M)=\mathbb{P}_\infty\left(\max_{1\le t\le M}Z_t'>b\right)$. We now prove our proposition in four steps.

Step 1. Exponential tilting. Denote the cumulant generating function of Z'_t by

$$\psi(a) = \log \mathbb{E}_\infty[e^{aZ'_t}].$$

Define a family of new measures

$$\frac{d\mathbb{P}_t}{d\mathbb{P}_\infty} = \exp\{aZ'_t - \psi(a)\},$$

where $\mathbb{P}_t$ denotes the new measure after the transformation. The new measure takes the form of the exponential family, and a can be viewed as the natural parameter. It can be verified that $\mathbb{P}_t$ is indeed a probability measure because

$$\int d\mathbb{P}_t = \int \exp\{aZ'_t - \psi(a)\} d\mathbb{P} = 1.$$

It can also be shown that $\dot{\psi}(a)$ is the expected value of Z'_t under $\mathbb{P}_t$, because

$$\dot{\psi}(a) = \frac{\mathbb{E}_\infty\left[Z'_t e^{aZ'_t}\right]}{\mathbb{E}_\infty\left[e^{aZ'_t}\right]} = \mathbb{E}_\infty\left[Z'_t e^{aZ'_t - \psi(a)}\right] = \mathbb{E}_t\left[Z'_t\right],$$

and, similarly, $\ddot{\psi}(a)$ is the variance under the tilted measure. We use the Gaussian approximation for Z'_t, then its log moment generating function is $\psi(a) = a^2/2$. We set $a = b$ such that $\dot{\psi}(a) = \mathbb{E}_t\left[Z'_t\right] = b$; therefore, the tail probability after measure transformation will become much larger. Given this choice, the transformed measure is given by $d\mathbb{P}_t = \exp\left(bZ'_t - b^2/2\right)d\mathbb{P}_\infty$. We also define, for each t, the log-likelihood ratio $\log\left(d\mathbb{P}_t/d\mathbb{P}_\infty\right)$ of the form

$$\ell_t = bZ'_t - \frac{1}{2}b^2.$$

Step 2. Change-of-measure by the likelihood ratio identity. Now we convert the original problem of finding the small probability that the maximum of a random field exceeds a large threshold to another problem: finding an alternative measure under which the event happens with a much higher probability. By likelihood ratio identity, we have

$$\begin{aligned}
\mathbb{P}_\infty\left(\max_{1\le m\le M} Z'_m \ge b\right) &= \mathbb{E}_\infty\left[1_{\left\{\max\limits_{1\le m\le M} Z'_m \ge b\right\}}\right] = \mathbb{E}_\infty\left[\frac{\sum_{t=1}^M e^{\ell_t}}{\sum_{n=1}^M e^{\ell_n}} \cdot 1_{\left\{\max\limits_{1\le m\le M} Z'_m \ge b\right\}}\right] \\
&= \sum_{t=1}^M \mathbb{E}_\infty\left[\frac{e^{\ell_t}}{\sum_n e^{\ell_n}} \cdot 1_{\left\{\max\limits_{1\le m\le M} Z'_m \ge b\right\}}\right] \\
&= \sum_{t=1}^M \mathbb{E}_t\left[\frac{1}{\sum_n e^{\ell_n}} \cdot 1_{\left\{\max\limits_{1\le m\le M} Z'_m \ge b\right\}}\right] \\
&= e^{-b^2/2}\sum_{t=1}^M \mathbb{E}_t\left[\frac{M_t}{S_t} e^{-\left(\tilde{\ell}_t + \log M_t\right)} \cdot 1_{\{\tilde{\ell}_t + \log M_t \ge 0\}}\right],
\end{aligned} \tag{B.1}$$

where M_t and S_t in the last step are defined as the maximum and sum of likelihood ratio differences as

$$M_t = \max_{m\in\{1,\dots,M\}} e^{\ell_m - \ell_t}, \quad S_t = \sum_{m\in\{1,\dots,M\}} e^{\ell_m - \ell_t}.$$

And $\tilde{\ell}_t$ is defined as the recentered likelihood ratio, or the so-called global term:

$$\tilde{\ell}_t = b\left(Z'_t - b\right).$$

The last equation in (B.1) converts the tail probability to a product of two terms: a deterministic term $e^{-b^2/2}$ associated with the large deviation rate and a sum of expectations under the transformed measures. The expectation involves a product of the ratio M_t/S_t and an exponential function that depends on $\tilde{\ell}_t$, which plays the role of a weight. Under the new measure $\mathbb{P}_t, \tilde{\ell}_t$ has zero mean and variance equal to b^2 and it dominates the other term $\log M_t$; hence, the probability of exceeding zero is much higher. Next, we characterize the limiting ratio and the other factors precisely by the localization theorem.

Step 3. Establish properties of local and global terms. In (B.1), our target probability has been decomposed into terms that only depend on (i) the local field $\{\ell_m - \ell_t\}, 1 \le m \le M,$ which are the differences between the log-likelihood ratios with parameter t and m, and (ii) the global term $\tilde{\ell}_t$, which is the centered and scaled likelihood ratio with parameter t. We need to first establish some useful properties of the local field and global term before applying the localization theorem. We will eventually show that the local field and the global term are asymptotically independent.

For the local field $\{\ell_m - \ell_t\}$, let $r_{m,t}$ denote the correlation between Z'_m and Z'_t, then we have

$$\begin{aligned}\mathbb{E}_t(\ell_m - \ell_t) &= -b^2(1 - r_{m,t}),\\ \mathrm{Var}_t(\ell_m - \ell_t) &= 2b^2(1 - r_{m,t}),\\ \mathrm{Cov}_t(\ell_{m_1} - \ell_t, \ell_{m_2} - \ell_t) &= b^2(1 + r_{m_1,m_2} - r_{m_1,t} - r_{m_2,t}).\end{aligned}$$

We have Lemma B.1 to characterize the local correlation, which offers a reasonably good approximation for $\mathbb{E}[Z_t Z_{t+\delta}]$ and leads to $r_{m,t} \approx 1 - |m - t|\beta_{k,w}/w$.

Because we assume that Z'_t is approximately Gaussian, the local field $\ell_m - \ell_t$ and the global term $\tilde{\ell}_t$ are also approximately Gaussian. Therefore, when $|\delta|$ is small (i.e., in the neighborhood of zero), we can approximate the local field using a two-sided Gaussian random walk with drift $b^2\beta_{k,w}/w$ and variance of the increment equal to $2b^2\beta_{k,w}/w$:

$$\ell_{t+\delta} - \ell_t \overset{\Delta}{=} b\sqrt{\frac{2\beta_{k,w}}{w}}\sum_{i=1}^{|\delta|}\xi_i - b^2\frac{\beta_{k,w}}{w}\delta, \delta = \pm 1, \pm 2, \ldots,$$

where ξ_i are independent and identically distributed standard normal random variables.

Step 4. Approximation using localization theorem. From the argument in Siegmund and Yakir (2000), let $\hat{M}_t$ and $\hat{S}_t$ denote the maximization and summation restricted to the small neighborhood of t. Then they are asymptotically independent of the global term $\tilde{\ell}_t$. Moreover, under the tilted measure,

$$\mathbb{E}_t\left[\tilde{\ell}_t\right] = 0, \quad \mathrm{Var}_t\left[\tilde{\ell}_t\right] = b^2.$$

Therefore, the density $\mathbb{P}_t(\tilde{\ell}_t)$ can be approximated by $1/\sqrt{2\pi b^2}$ in a neighborhood of radius $o(1/b)$ of zeros. The inner expectation in (B.1) can be approximated as

$$\mathbb{E}_t\left[\frac{M_t}{S_t}e^{-\left(\tilde{\ell}_t + \log M_t\right)} \cdot 1_{\{\tilde{\ell}_t + \log M_t \ge 0\}}\right] \doteq \frac{\mathbb{E}_t\left(\hat{M}_t/\hat{S}_t\right)}{b\sqrt{2\pi b^2}}.$$

By Siegmund and Yakir (2000), the expectation $\mathbb{E}_t\left(\hat{M}_t/\hat{S}_t\right)$ does not depend on t and equals $b^2\beta_{k,w}\nu\left(b\sqrt{2\beta_{k,w}/w}\right)/w$ in the asymptotic regime. Substituting into (B.1) we have

$$\begin{aligned}\mathbb{P}_\infty(T \le M) =& \mathbb{P}_\infty\left(\max_{1\le t\le M} Z'_t > b\right)\\ =& e^{-b^2/2}\sum_{t=1}^{M}\mathbb{E}_t\left[\frac{M_t}{S_t}e^{-\left[\tilde{\ell}_t + \log M_t\right]} \cdot 1_{\{\tilde{\ell}_t + \log M_t \ge 0\}}\right] \doteq Mb\phi(b)\beta_{k,w}\nu\left(b\sqrt{2\beta_{k,w}/w}\right)/w,\end{aligned}$$

where $\nu(\cdot)$ is the function defined in (4.6). From the above cumulative distribution function, we can approximate T as exponential distribution, yielding the mean value $1/[b\phi(b)\beta_{k,w}\nu(b\sqrt{2\beta_{k,w}/w})/w]$.

Because Z'_t is standardized, here the threshold b needs to be converted to the original threshold using a simple formula

$$b' = \left[b - (\mu_{w,k} + c_1\sigma_{w,k})\right]/(c_2\sigma_{w,k}).$$

This completes the proof.

□

APPENDIX C: PROOF OF THEOREM 4.3

We first relate the largest eigenvalue procedure to a CUSUM procedure. Note that

$$\lambda_{\max}\left(\hat{\Sigma}_{t,w}\right) = \max_{||q||=1} q^\top \hat{\Sigma}_{t,w} q. \tag{C.1}$$

For each q, we have

$$q^\top \hat{\Sigma}_{t,w} q = \sum_{i=t-w+1}^{t} \left(q^\top x_i\right)^2.$$

According to the Grothendieck's inequality (Guédon and Vershynin 2016), the q that attains the maximum in equation (C.1) is very close to u under the alternative. Therefore, assuming that the optimal q always equals u will only cause a small error but will bring great convenience to our analysis.

Now we have under $\mathbb{P}_\infty, q^\top x_i \sim \mathcal{N}(0, \sigma^2)$ and under $\mathbb{P}_0, q^\top x_i \sim \mathcal{N}(0, \sigma^2 + \theta)$. Let f_∞ denote the probability density function of $\mathcal{N}(0, \sigma^2)$ and f_0 the probability density function of $\mathcal{N}(0, \sigma^2 + \theta)$. For each observation y, we can derive the one-sample log-likelihood ratio

$$\log \frac{f_0(y)}{f_\infty(y)} = -\frac{1}{2}\log(1+\rho) + \frac{1}{2\sigma^2}\left(1 - \frac{1}{1+\rho}\right)y^2.$$

Define the CUSUM procedure

$$\tilde{T} = \inf\left\{t : \max_{0\le k<t} \sum_{i=k+1}^{t} \left[\frac{1}{2\sigma^2}\left(1 - \frac{1}{1+\rho}\right)\left(q^\top x_i\right)^2 - \frac{\log(1+\rho)}{2}\right] \ge b'\right\},$$

where $b' = \frac{1}{2\sigma^2}\left(1 - \frac{1}{1+\rho}\right)\left(b - \frac{\sigma^2 \log(1+\rho)}{1-1/(1+\rho)}\right)w$. Then we have

$$\mathbb{E}_0[T_{\mathrm{E}}] \ge \mathbb{E}_0[\tilde{T}].$$

Because $\tilde{T}$ is a CUSUM procedure with

$$\int \log\left[\frac{f_0(y)}{f_\infty(y)}\right] f_0(y)dy = -\frac{1}{2}\log(1+\rho) + \frac{\rho}{2},$$

by Siegmund (1985) we have

$$\mathbb{E}_0[\tilde{T}] = \frac{e^{-b'} + b' - 1}{-\log(1+\rho)/2 + \rho/2}.$$

This completes the proof.

□

ACKNOWLEDGMENTS

The authors thank the Editor and the anonymous referees for the thoughtful comments and suggestions, which led to an improvement of the presentation.

DISCLOSURE

The authors have no conflicts of interest to report.

FUNDING

The work of Liyan Xie and Yao Xie was supported by the U.S. National Science Foundation (CCF-1442635, CMMI-1538746, DMS-1830210) and the Career Award (CCF-1650913). The work of George V. Moustakides was supported by the U.S. National Science Foundation (CIF-1513373) through Rutgers University.

ORCID

Liyan Xie http://orcid.org/0000-0003-3009-4514
Yao Xie http://orcid.org/0000-0001-6777-2951
George V. Moustakides http://orcid.org/0000-0002-6498-5860

REFERENCES

Alt, F. B. 1985. "Multivariate Quality Control." In *Encyclopedia of Statistical Sciences*, edited by S. Kotz and N. L. Johnson, 110–22. New York: Wiley.

Alt, F. B., and N. D. Smith. 1988. "Multivariate Process Control." In *Handbook of Statistics, Chapter* 17, edited by P. R. Krishnaiah and C. R. Rao, 333–51. Amsterdam: Elsevier.

Anderson, T. W. 1963. "Asymptotic Theory for Principal Component Analysis." *Annals of Mathematical Statistics* 34 (1):122–48. doi:10.1214/aoms/1177704248

Arias-Castro, E., S. Bubeck, and G. Lugosi. 2012. "Detection of Correlations." *Annals of Statistics* 40 (1):412–35. doi:10.1214/11-AOS964

Avanesov, V., and N. Buzun. 2018. "Change-Point Detection in High-Dimensional Covariance Structure." *Electronic Journal of Statistics* 12 (2):3254–94. doi:10.1214/18-EJS1484

Baik, J., and J. W. Silverstein. 2006. "Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models." *Journal of Multivariate Analysis* 97 (6):1382–408. doi:10.1016/j.jmva.2005.08.003

Balzano, L., Y. Chi, and Y. M. Lu. 2018. "Streaming PCA and Subspace Tracking: The Missing Data Case." *Proceedings of IEEE. Institute of Electrical and Electronics Engineers* 106 (8): 1293–310. doi:10.1109/JPROC.2018.2847041

Berthet, Q., and P. Rigollet. 2013. "Optimal Detection of Sparse Principal Components in High Dimension." *Annals of Statistics* 41 (4):1780–815. doi:10.1214/13-AOS1127

Bodnar, O., and W. Schmid. 2005. "Multivariate Control Charts Based on a Projection Approach." *Allgemeines Statistisches Archiv* 89 (1):75–93. doi:10.1007/s101820500193

Chan, L. K., and J. Zhang. 2001. "Cumulative Sum Control Charts for the Covariance Matrix." *Statistica Sinica* 11:767–90.

Chen, J., and A. Gupta. 2004. "Statistical Inference of Covariance Change Points in Gaussian Model." *Statistics* 38 (1):17–28. doi:10.1080/0233188032000158817

Edelman, A., and Y. Wang. 2013. "Random Matrix Theory and Its Innovative Applications." In *Advances in Applied Mathematics, Modeling, and Computational Science*, edited by R. Melnik and I. Kotsireas, 91–116. Boston: Springer.

Fothergill, S., H. Mentis, P. Kohli, and S. Nowozin. 2012. "Instructing People for Training Gestural Interactive Systems." Proceedings of SIGCHI Conference on Human Factors in Computing Systems: 1737–1746.

Geman, S. 1980. "A Limit Theorem for the Norm of Random Matrices." *Annals of Probability* 8 (2):252–61. doi:10.1214/aop/1176994775

Guédon, O., and R. Vershynin. 2016. "Community Detection in Sparse Networks via Grothendieck's Inequality." *Probability Theory and Related Fields* 165 (3-4):1025–49. doi:10.1007/s00440-015-0659-z
Healy, J. D. 1987. "A Note on Multivariate CUSUM Procedures." *Technometrics* 29 (4):409–12. doi:10.1080/00401706.1987.10488268
Hotelling, H. 1947. "Multivariate Quality Control." In *Techniques of Statistical Analysis*, edited by C. Eisenhart, H. Hastay, and W. A. Wallis, 111–84. New York: McGraw-Hill.
Huber, P. J. 1985. "Projection Pursuit." *Annals of Statistics* 13 (2):435–75. doi:10.1214/aos/1176349519
Jackson, J. E. 1959. "Quality Control Methods for Several Related Variables." *Technometrics* 1 (4): 359–77. doi:10.1080/00401706.1959.10489868
Jackson, J. E., and G. S. Mudholkar. 1979. "Control Procedures for Residuals Associated with Principal Component Analysis." *Technometrics* 21 (3):341–9. doi:10.1080/00401706.1979.10489779
Jiang, T., K. Leder, and G. Xu. 2017. "Rare-Event Analysis for Extremal Eigenvalues of White Wishart Matrices." *Annals of Statistics* 45 (4):1609–37. doi:10.1214/16-AOS1502
Jiao, Y., Y. Chen, and Y. Gu. 2018. "Subspace Change-Point Detection: A New Model and Solution." *IEEE Journal of Selected Topics in Signal Processing* 12 (6):1224–39. doi:10.1109/JSTSP.2018.2873147
Johnstone, I. M. 2001. "On the Distribution of the Largest Eigenvalue in Principal Components Analysis." *Annals of Statistics* 29 (2):295–327. doi:10.1214/aos/1009210544
Lai, T. L. 1995. "Sequential Change-Point Detection in Quality Control and Dynamical Systems." *Journal of Royal Statistical Society: Series B (Methodological)* 57 (4):613–44. doi:10.1111/j.2517-6161.1995.tb02052.x
Lai, T. L., and J. Z. Shan. 1999. "Efficient Recursive Algorithms for Detection of Abrupt Changes in Signals and Control Systems." *IEEE Transactions on Automatic Control* 44:952–66. doi:10.1109/9.763211
Li, S., Y. Xie, H. Dai, and L. Song. 2015. "M-Statistic for Kernel Change-Point Detection." *Advances in Neural Information Processing Systems* :3366–74.
Li, Z., Z. Peng, D. Hollis, L. Zhu, and J. McClellan. 2018. "High-Resolution Seismic Event Detection Using Local Similarity for Large-N Arrays." *Scientific Reports* 8 (1):1646 doi:10.1038/s41598-018-19728-w
Lorden, G. 1971. "Procedures for Reacting to a Change in Distribution." *Annals of Mathematical Statistics* 42 (6):1897–908. doi:10.1214/aoms/1177693055
Mises, R. V., and H. Pollaczek-Geiringer. 1929. "Praktische Verfahren Der Gleichungsauflösung." *ZAMM - Zeitschrift Für Angewandte Mathematik Und Mechanik* 9 (2):152–64. doi:10.1002/zamm.19290090206
Moustakides, G. V. 1986. "Optimal Stopping Times for Detecting Changes in Distributions." *Annals of Statistics* 14 (4):1379–87. doi:10.1214/aos/1176350164
Neuburger, J., K. Walker, C. Sherlaw-Johnson, J. van der Meulen, and D. A. Cromwell. 2017. "Comparison of Control Charts for Monitoring Clinical Performance Using Binary Data." *BMJ Quality & Safety* 26 (11):919–28. doi:10.1136/bmjqs-2016-005526
Page, E. S. 1954. "Continuous Inspection Schemes." *Biometrika* 41 (1-2):100–15. doi:10.1093/biomet/41.1-2.100
Paul, D. 2007. "Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model." *Statistica Sinica* 17:1617–42.
Pignatiello, J. J., Jr, and G. C. Runger. 1990. "Comparisons of Multivariate CUSUM Charts." *Journal of Quality Technology* 22 (3):173–86. doi:10.1080/00224065.1990.11979237
Rudin, W. 2006. *Real and Complex Analysis*, New York: McGraw-Hill.
Runger, G. C. 1996. "Projections and the U^2 Multivariate Control Chart." *Journal of Quality Technology* 28 (3):313–9. doi:10.1080/00224065.1996.11979681
Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6 (2):461–4. doi:10.1214/aos/1176344136
Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer.

Siegmund, D., and B. Yakir. 2000. "Tail Probabilities for the Null Distribution of Scanning Statistics." *Bernoulli* 6 (2):191–213. doi:10.2307/3318574

Siegmund, D., and B. Yakir. 2007. *Statistics of Gene Mapping*, New York: Springer.

Siegmund, D., B. Yakir, and N. Zhang. 2010. "Tail Approximations for Maxima of Random Fields by Likelihood Ratio Transformations." *Sequential Analysis* 29 (3):245–62. doi:10.1080/07474946.2010.487428

Sprint, G., D. J. Cook, and M. Schmitter-Edgecombe. 2016. "Unsupervised Detection and Analysis of Changes in Everyday Physical Activity Data." *J Biomed Inform* 63:54–65. doi:10.1016/j.jbi.2016.07.020

Sultani, W., C. Chen, and M. Shah. 2018. "Real-World Anomaly Detection in Surveillance Videos." Proceedings of the Conference on Computer Vision and Pattern Recognition: 6479–6488.

Tracy, C. A., and H. Widom. 1996. "On Orthogonal and Symplectic Matrix Ensembles." *Communications in Mathematical Physics* 177 (3):727–54. doi:10.1007/BF02099545

Wang, D., Y. Yu, and A. Rinaldo. 2017. "Optimal Covariance Change Point Localization in High Dimension." *arXiv:1712.09912*

Woodbury, M. A. 1950. "Inverting Modified Matrices." *Memorandum Report* 42, Statistical Research Group, Princeton University, Princeton, NJ.

Xie, Y., and D. Siegmund. 2013. "Sequential Multi-Sensor Change-Point Detection." *Annals of Statistics* 41 (2):670–92. doi:10.1214/13-AOS1094

Yin, Y. Q., Z. D. Bai, and P. R. Krishnaiah. 1988. "On the Limit of the Largest Eigenvalue of the Large Dimensional Sample Covariance Matrix." *Probability Theory and Related Fields* 78 (4):509–21. doi:10.1007/BF00353874

Zhang, C., H. Yan, S. Lee, and J. Shi. 2018. "Dynamic Multivariate Functional Data Modeling via Sparse Subspace Learning." *Technometrics*. doi:10.1080/00401706.2020.1800516