Taylor & Francis
Taylor & Francis Group

Sequential Analysis
Design Methods and Applications

ISSN: 0747-4946 (Print) 1532-4176 (Online) Journal homepage: https://www.tandfonline.com/loi/lsqa20

Scan *B*-statistic for kernel change-point detection

Shuang Li, Yao Xie, Hanjun Dai & Le Song

To cite this article: Shuang Li, Yao Xie, Hanjun Dai & Le Song (2019) Scan *B*-statistic for kernel change-point detection, Sequential Analysis, 38:4, 503-544, DOI: 10.1080/07474946.2019.1686886

To link to this article: https://doi.org/10.1080/07474946.2019.1686886

Published online: 29 Jan 2020.

Submit your article to this journal

Article views: 42

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lsqa20

SEQUENTIAL ANALYSIS
2019, VOL. 38, NO. 4, 503–544
https://doi.org/10.1080/07474946.2019.1686886

Taylor & Francis
Taylor & Francis Group

Check for updates

Scan *B*-statistic for kernel change-point detection

Shuang Li[a], Yao Xie[a], Hanjun Dai[b], and Le Song[b]

[a]School of Industrial and Systems Engineering (ISyE), Georgia Institute of Technology, Atlanta, Georgia, USA; [b]School of Computer Science and Engineering (CSE), Georgia Institute of Technology, Atlanta, Georgia, USA

ABSTRACT

Detecting the emergence of an abrupt change-point is a classic problem in statistics and machine learning. Kernel-based nonparametric statistics have been used for this task, which enjoys fewer assumptions on the distributions than the parametric approach and can handle high-dimensional data. In this article, we focus on the scenario when the amount of background data is large and propose a computationally efficient kernel-based statistics for change-point detection, inspired by the recently developed *B*-statistics. A novel theoretical result of the article is the characterization of the tail probability of these statistics using the change-of-measure technique, which focuses on characterizing the tail of the detection statistics rather than obtaining its asymptotic distribution under the null distribution. Such approximations are crucial to controlling the false alarm rate, which corresponds to the average run length in online change-point detection. Our approximations are shown to be highly accurate. Thus, they provide a convenient way to find detection thresholds for online cases without the need to resort to the more expensive simulations. We show that our methods perform well on both synthetic data and real data.

ARTICLE HISTORY
Received 11 January 2019
Revised 13 September 2019
Accepted 21 October 2019

KEYWORDS
Change-point detection; false-alarm control; kernel-based statistics; online algorithm

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

1. Introduction

Given a sequence of samples, $x_1, x_2, \ldots, x_t$, from a domain $\mathcal{X}$, we are interested in detecting a possible change-point τ, such that before the change samples x_i are independent and identically distributed (i.i.d.) with a null distribution P, and after the change samples x_i are i.i.d. with a distribution Q. Here, we consider two scenarios: the time horizon t is fixed, $t = T_0$, which we call the offline or fixed-sample change-point detection, or the time horizon t is not fixed, meaning that one can keep getting new samples, which we call the online or sequential change-point detection. In the offline setting, our goal is to detect the existence of a change. In the online setting, our goal is to detect the emergence of a change as soon as possible after it occurs. Here, we restrict our attention to detecting one change point. One such instance is seismic event detection as studied by Ross and Ben-Zion (2014), where one would like to either detect

CONTACT: Yao Xie yao.c.xie@gmail.com School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsqa.
Recommended by Nitis Mukhopadhyay

© 2019 Taylor & Francis Group, LLC

the presence of a weak event in retrospect to better understand the geophysical structure or detect the event as quickly as possible for online monitoring.

Ideally, the detection algorithm should be free of distributional assumptions to be robust when applied to real data. To achieve this goal, various kernel-based nonparametric statistics have been proposed in the statistics and machine learning literature; see, for example, Harchaoui et al. (2008), Enikeeva and Harchaoui (2014), S. Zou et al. (2017), Kifer et al. (2004), Liu et al. (2013), and Desobry et al. (2005), which typically work well with multidimensional real data because they are distributional free. Kernel approaches are distribution free and more robust because they provide consistent results over larger classes of data distributions, although they can be less powerful in settings where a clear distributional assumption can be made. However, most kernel-based statistics cost $\mathcal{O}(n^2)$ to compute over n samples. In the online change-point detection setting, the number of samples grows with time and hence we cannot directly use the naive approach. Recently, Zaremba et al. (2013) developed the so-called B-test statistic to reduce computational complexity. The B-test statistic samples N pairs of blocks of size B from the two-sample data, computes the unbiased estimates of the kernel-based statistic between each pair, and then takes an average. The computational complexity of the B-test statistic reduces to $\mathcal{O}(nB^2)$ instead of $\mathcal{O}(n^2)$.

In this article, we present two scan statistics related to B-test statistics customized for offline and online change-point detection, which we call *scan B-statistics*. The proposed statistics are based on kernel maximum mean discrepancy (MMD) in Gretton et al. (2012) and Harchaoui et al. (2013). They are inspired by the B-test statistic but differ in various ways to tailor to the need of change-point detection. Typically, there is a small number of post change samples (for instance, seismic events are relatively rare, and in online change-point detection, one would like to detect the change quickly). However, there is a large amount of reference data. So when constructing the detection statistic, we reuse the post change samples for the test block and construct multiple and disjoint reference blocks. This leads to a non negligible dependence between the MMD statistics being averaged over. Hence, we cannot use the existing approach based on the central limit theorem to analyze them. Moreover, the scanning nature of the proposed statistic also introduces non negligible dependence. We construct the reference and test blocks in a structured way so that analytical expressions for false alarm can be obtained.

Our main theoretical contribution includes accurate theoretical approximations to the false-alarm rate of scan B-statistics. Controlling false alarms is a key challenge in change-point detection. Specifically, this means to quantify the significance level for offline change-point detection and the average run length (ARL) for online change-point detection. Here, we cannot directly rely on the null property of the B-test statistic established in the existing work, because the scan statistics take the maximum of multiple statistics computed over overlapping data blocks, which causes strong correlations. Hence, one cannot use the central limit theorem or even the martingale central limit theorem. Instead, we adopt a recently developed change-of-measure technique by Yakir (2013) for scan statistics that is capable of dealing with the more challenging situation here.

Our contribution also includes (1) obtaining a closed-form variance estimator, which allows easy calculation of the scan B-statistics, and (2) further improving the accuracy

of our approximations by taking into account the skewness of the kernel-based statistics. The accuracy of our approximations is validated by numerical examples. Finally, we demonstrate the good performance of our method using real data, including speech and human activity data.

1.1. Related work

Classic parametric approaches for change-point detection can be found in Siegmund (1985) and Tartakovsky et al. (2014). There is an array of nonparametric change-point detection methods. Notable nonparametric schemes for change-point detection include Gordon and Pollak (1994) and Picard (1985), which are designed for scalar observations and not suitable for vector observations. Brodsky and Darkhovsky (2013) provided a comprehensive introduction to the methodologies and applications of nonparametric change-point detection. Bibinger et al. (2017) constructed a nonparametric minimax-optimal test to discriminate continuous paths with volatility jumps and prove weak convergence of the test statistic to an extreme value distribution. In the online setting, Kifer et al. (2004) presented a meta-algorithm that compares data in some "reference window" to the data in the current window, using empirical distance measures that are not kernel based; Desobry et al. (2005) detected abrupt changes by comparing two sets of descriptors extracted online from the signal at each time instant (the immediate past set and the immediate future set) and then used a soft margin single-class support vector machine to build a dissimilarity measure in the feature space between those sets without estimating densities as an intermediate step, which is asymptotically equivalent to the Fisher ratio in the Gaussian case; Song et al. (2013) presented a density ratio estimation method to detect change points, fitting the density ratio using a nonparametric Gaussian kernel model whose parameters are updated online via stochastic gradient descent approach. Another important branch of nonparametric change-point detection methods is based on the Kolmogorov-Smirnov test, as in Massey (1951) and Lilliefors (1967), which has been used in Wang et al. (2014). The generalization of the Kolmogorov-Smirnov test from the univariate setting to the multidimensional setting is given by Fasano and Franceschini (1987), which, however, is less convenient to use than the kernel-based statistic test.

Seminal works by Csörgő and Horváth (1989) studied a kernel-based U-statistic for change-point detection. They showed that the statistic indexed by the assumed change-point location parameter τ, after proper standardization and rescaling of time and magnitude, converges in distribution to a Gaussian process under the null and converges to a deterministic path in probability under the alternative distribution when the number of samples goes to infinity. These results are useful for bounding the detection statistics under the null with high probability (hence controlling the false detection) and for studying the consistency of tests. Csörgő and Horváth (1997) and Serfling (2009) contain comprehensive discussions on asymptotic theory of nonparametric statistics including U-statistics. Our scan B-statistic can also be viewed as a form of U-statistic using an appropriate definition of the kernel. The main differences between these classic works from our proposed scan B-statistic are that (1) our statistic uses B-test block decomposition and averaging to make the test statistic more computationally efficient; (2) our statistic is

more challenging to analyze due to the block structure and correlation introduced by scan statistics; and (3) our analytical approach is different: Csörgő and Horváth (1989) leveraged the invariance principle to establish convergence of the entire sample path; we focus on characterizing the tail probability of the statistic under the null and use the change-of-measure technique to achieve good approximation accuracy.

Other existing works that also focus on establishing asymptotic distribution of the detection statistic under the null for controlling the false alarm rate include the following: Harchaoui et al. (2008) presented a maximum kernel Fisher discriminant ratio statistic and studied its asymptotic null distribution; Dehling et al. (2015) investigated the two-sample test U-statistic for dependent data. Our approach is different from those above in that we focus on directly approximating the tail of the detection statistic under the null, rather than trying to obtain its asymptotic distribution. Moreover, traditional analyses are usually done for offline change-point detection, whereas our analytical framework based on change-of-measure can be applied to both offline and online change-point detection.

Change-point detection problems are related to the classical statistical two-sample test. However, they are usually more challenging than the two-sample test because the change-point location τ is unknown. Hence, when forming the detection statistic, one has to "take the maximum" of the detection statistics. The statistics being maxed over are usually highly correlated because they are computed using overlapping data.

Our techniques for approximating false alarm rates differ from large-deviation techniques in Dembo and Zeitouni (2010), which establish the exponential rate by which the probability converges to zero. In certain scenarios, the first-order approximation obtained from large-deviation techniques may not be sufficient for choosing a threshold. Our method provides more refined approximations to include polynomial terms and constants.

Finally, there are also works taking different approaches rather than hypothesis test for change-point detection. For instance, Harchaoui and Cappé (2007) developed a kernel-based multiple-change-point detection approach, where the optimal location to segment the data was obtained by dynamic programming; Arlot et al. (2012) estimated multiple change points by developing a kernelized linear model, and they provided a nonasymptotic oracle inequality for the estimation error. In the offline setting, S. Zou et al. (2017) studied a problem when there are s anomalous sequences out of n sequences to be detected, and the test statistic was constructed using MMD; Matteson and James (2014) proposed a nonparametric approach based on U-statistics and adopted hierarchical clustering, which is capable of consistently estimating an unknown number of multiple change-point locations; C. Zou et al. (2014) proposed a nonparametric maximum likelihood approach, with the number of change points determined from the Bayesian information criterion and the locations of the change points estimated via dynamic programming.

Our notations are standard. Let I_k denote the identity matrix of size $k \times k$. Let $\mathbb{E}[\mathcal{A}; \mathcal{B}] = \mathbb{E}[\mathcal{A}\mathbf{1}_{\mathcal{B}}]$ denote the expectation conditioned on event $\mathcal{B}$, where $\mathbf{1}_{\mathcal{B}}$ represents the indicator function that takes a value of 1 when the event $\mathcal{B}$ happens and takes a value of 0, otherwise. Let $\mathrm{Var}(\cdot)$ and $\mathrm{Cov}(\cdot)$ denote the variance and the covariance. Let 0 and $\mathbf{e}$ denote vectors of all zeros and all ones, respectively. Let $[\boldsymbol{\Sigma}]_{ij}$ denote the ijth

element of a matrix Σ. In Section 4.1, $\mathbb{E}_B, \mathrm{Var}_B$, and Cov_B denote the values computed under the new probability measure $\mathbb{P}_B$ after the change-of-measure, where B is the block size. Similarly, in Section 4.2, $\mathbb{E}_t, \mathrm{Var}_t$, and Cov_t denote the values obtained under the new probability measure $\mathbb{P}_t$ after the change-of-measure, where t is the time index.

2. Background

We first briefly review the reproducing kernel Hilbert space (RKHS) and the MMD. An RKHS $\mathcal{F}$ on $\mathcal{X}$ with a kernel $k(x, x')$ is a Hilbert space of functions $f(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Its element $k(x, \cdot)$ satisfies the reproducing property $\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{F}} = f(x)$ and, consequently, $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{F}} = k(x, x')$, meaning that we can view the evaluation of a function f at any point $x \in \mathcal{X}$ as an inner product. Commonly used RKHS kernel functions include the Gaussian radial basis function (RBF) $k(x, x') = \exp\left(-||x - x'||^2 / 2\sigma^2\right)$, where $\sigma > 0$ is the kernel bandwidth, and polynomial kernel $k(x, x') = (\langle x, x' \rangle + a)^d$, where $a > 0$ and $d \in \mathbb{N}$ (see Schölkopf and Smola, 2001). RKHS kernels can also be defined for sequences, graphs, and other structured objects (see Schölkopf et al., 2004). In this article, if not otherwise stated, we will assume that a Gaussian RBF kernel is used.

Assume that there are two sets X and Y, each with n samples taking value on a general domain $\mathcal{X}$, where $X = \{x_1, x_2, ..., x_n\}$ are i.i.d. with a distribution P, and $Y = \{y_1, y_2, ..., y_n\}$ are i.i.d. with a distribution Q. The MMD is defined as (Gretton et al., 2012)

$$\mathrm{MMD}[\mathcal{F}, P, Q] := \sup_{f \in \mathcal{F}} \{\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]\}.$$

An unbiased estimator of MMD^2 can be obtained using the U-statistic (Gretton et al., 2012)

$$\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} h(x_i, x_j, y_i, y_j), \tag{2.1}$$

where $h(\cdot)$ is the kernel for the U-statistic and it can be defined using an RKHS kernel as

$$h(x_i, x_j, y_i, y_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i). \tag{2.2}$$

Intuitively, the empirical test statistic MMD_u^2 is expected to be small (close to zero) if $P = Q$, and large if P and Q are "far" apart. The complexity of evaluating MMD_u^2 is $\mathcal{O}(n^2)$, because we have to form the so-called Gram matrix for the data, which is of size $n \times n$. Under the null hypothesis, $P = Q$, the U-statistic is degenerate and has the same distribution as an infinite sum of chi-square variables.

To improve computational efficiency, an alternative approach to eatimate MMD^2, called the B-test, was presented by Zaremba et al., (2013). The key idea is to partition the n samples from P and Q into N nonoverlapping blocks, $X_1, ..., X_N$ and $Y_1, ..., Y_N$, each of size B. Then one computes $\mathrm{MMD}_u^2[\mathcal{F}, X_i, Y_i]$ for each pair of blocks and takes an average:

$$\mathrm{MMD}_B^2[\mathcal{F}, X, Y] = \frac{1}{N}\sum_{i=1}^{N} \mathrm{MMD}_u^2[\mathcal{F}, X_i, Y_i].$$

Because B is constant and N is on the order of $\mathcal{O}(n)$, the computational complexity of $\mathrm{MMD}_B^2[\mathcal{F}, X, Y]$ is $\mathcal{O}(nB^2)$, which is significantly lower than the $\mathcal{O}(n^2)$ complexity of $\mathrm{MMD}_u^2[\mathcal{F}, X, Y]$. Furthermore, by averaging $\mathrm{MMD}_u^2[\mathcal{F}, X_i, Y_i]$ over blocks, when blocks are independent, the B-test statistic is asymptotically normal under the null using central limit theorem. This property allows a simple threshold to be derived for the B-test.

3. Scan *B*-statistics

Now we present our change-point detection procedure based on the *scan B-statistic*. Consider a sequence of data $\{..., x_{-2}, x_{-1}, x_0, x_1, ..., x_t\}$, each taking values on a general domain $\mathcal{X}$. Let $\{..., x_{-2}, x_{-1}, x_0\}$ denote the reference data that we know to follow a given prechange distribution. Assume there is a large amount of reference data.

In offline change point detection, the number of samples is fixed, and our goal is to detect the *existence* of a change point τ, such that before the change point, the samples are i.i.d. with a distribution P and after the change point, the samples are i.i.d. with a different distribution Q. The location τ where the change point occurs is unknown. In other words, we are concerned with testing the null hypothesis

$$H_0 : x_i \sim P, \quad i = 1, ..., t,$$

against the single change-point alternative

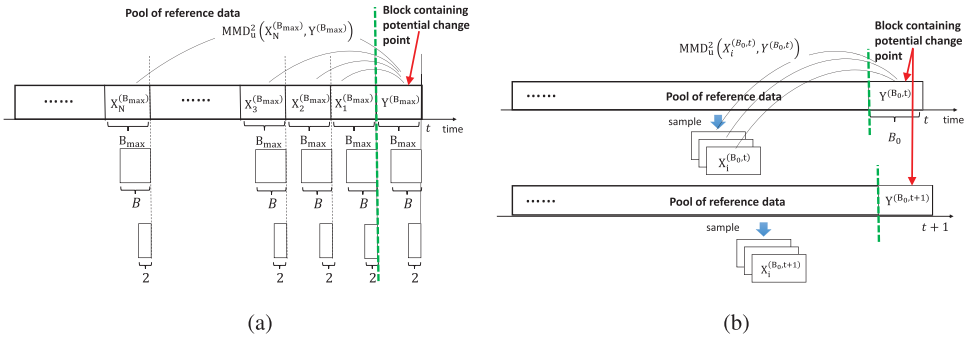
$$H_1 : \exists 1 \le \tau < t \quad x_i \sim \begin{cases} Q, & i > \tau \\ P, & \text{otherwise.} \end{cases}$$

Note that we are interested in the case of a sustained change: before the change, all samples follow one distribution, and after the change all samples follow another distribution and never switch back. In online change-point detection, the number of samples is not fixed, and the goal is to detect the *emergence* of a change point as quickly as possible. In various change-point detection settings, the number of post change samples is small, but the number of reference samples is large. Therefore, when constructing MMD statistics over blocks, we will use a common post change block and multiple disjoint pre change reference blocks.

3.1. Offline change-point detection

For each possible change location τ, the post change block consists of the most recent samples indexed from τ to t. Because we do not know the change-point location, we scan all possible change-point locations τ. This corresponds to considering a range of post change block sizes B ranging from two (i.e., the most recent two samples are post change samples) to $B_{\max}$. Here, we exclude $B = 1$ because the corresponding MMD is unable to be computed.

The detection statistic is constructed as follows, also illustrated in Figure 1(a). Data are split into N reference blocks and one test block, each block is size of $B_{\max}$. Then we select data from each block to form smaller subblocks of various size B, $2 \le B \le B_{\max}$.



(a) (b)

Figure 1. Illustration of (a) offline change-point detection: data are initially split into blocks of size $B_{\max}$; we select data from each block to form smaller sub blocks of various size B, $2 \le B \le B_{\max}$; (b) online change-point detection: the most recent B_0 samples constitute the test block, which is constantly updated with new data; the reference blocks of the same size B_0 are sampled from the reference pool of data.

The reference blocks are denoted as $X_i^{(B)}, i = 1, \ldots, N,$ and the test block as $Y^{(B)}$. We compute MMD_u^2 for each reference subblock with respect to the *common* postchange block and take an average:

$$Z_B = \frac{1}{N}\sum_{i=1}^{N} \mathrm{MMD}_u^2(X_i^{(B)}, Y^{(B)}). \tag{3.1}$$

Because the estimator MMD_u^2 is unbiased, under the null hypothesis $P = Q$, $\mathbb{E}[Z_B] = 0$. Let $\mathrm{Var}[Z_B]$ denote the variance of Z_B under the null. The variance of Z_B depends on the block size B and the number of blocks N. To have a fair comparison, we normalize each Z_B by their standard deviation

$$Z'_B = Z_B/(\mathrm{Var}[Z_B])^{1/2}$$

and take the maximum over all B to form the *offline scan B-statistic*. The variance $\mathrm{Var}[Z_B]$ is given in Lemma 3.1. The closed-form expression facilitates the estimation of the variance of the statistic. A change point is detected whenever the offline scan B-statistic exceeds a pre specified threshold b:

$$\max_{2\le B\le B_{\max}} Z'_B > b \quad \{\text{offline change point detection}\}. \tag{3.2}$$

3.2. Online change-point detection

In the online setting, new samples are sequential and we constantly test whether the incoming samples come from a different distribution. To reduce computational burden, in the online setting we fix the block size and adopt a *sliding window* approach. The resulted sliding window procedure can be viewed as a type of Shewhart chart by Shewhart (1939).

The detection statistic is constructed as follows, also illustrated in Figure 1(b). At each time t, we treat the most recent B_0 samples as the post change block. In online change-point detection, we want to detect the change as quickly as possible. Hence, typically we will not wait until many post change samples are collected. On the other

hand, there is a large amount of reference data. To utilize data efficiently, we utilize a common test block consisting of the most recent samples to form the statistic with N different reference blocks. The reference blocks are formed by taking NB_0 samples without replacement from the reference pool. We compute MMD_u^2 between each reference block with respect to the common post change block and take an average:

$$Z_{B_0,t} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{MMD}_u^2(X_i^{(B_0,t)}, Y^{(B_0,t)}), \tag{3.3}$$

where B_0 is the fixed block size, $X_i^{(B_0,t)}$ is the ith reference block at time t, and $Y^{(B_0,t)}$ is post change block at time t. When there are new samples, we append them to the post-change block and purge the oldest samples. We show later that this construction allows for an explicit characterization of the false alarm rate. We divide each statistic by its standard deviation to form the *online scan B-statistic*:

$$Z'_{B_0,t} = Z_{B_0,t}/(\mathrm{Var}[Z_{B_0,t}])^{1/2}.$$

The calculation of $\mathrm{Var}[Z_{B_0,t}]$ can also be achieved using Lemma 3.1. The online change-point detection procedure is a stopping time: an alarm is raised whenever the detection statistic exceeds a prespecified threshold $b > 0$:

$$T = \inf\{t : Z'_{B_0,t} > b\} \quad \{\text{online change-point detection}\}. \tag{3.4}$$

The online scan B-statistic can be computed efficiently. Note that the variance of the $Z_{B_0,t}$ only depends on the block size B_0 but is independent of t. Hence, it can be pre-computed. Moreover, there is a simple way to compute the online B-statistic recursively, as specified in Appendix A.

3.3. Analytical expression for $\mathrm{Var}[Z_B]$

We obtain an analytical expression for $\mathrm{Var}[Z_B]$, which is useful when forming the detection statistic in (3.2) and (3.4).

Lemma 3.1. (Variance of Z_B under the null). *Given block size $B \geq 2$ and the number of blocks N, under the null hypothesis,*

$$\mathrm{Var}\ [Z_B] = \binom{B}{2}^{-1}\left(\frac{1}{N}\mathbb{E}\left[h^2(x,x',y,y')\right] + \frac{N-1}{N}\mathrm{Cov}[h(x,x',y,y'),h(x'',x''',y,y')]\right), \tag{3.5}$$

where x, x', x'', $x''\prime$, y, and y' are i.i.d. random variables with the null distribution P.

The lemma is proved by making a connection between MMD_u^2 and the U-statistic in Serfling (2009) and utilizing the properties of the U-statistic. A detailed proof is provided in Appendix B.1.

3.4. Examples of detection statistics

Below, we present a few examples to demonstrate that the B-statistics is quite robust in various settings with different distributions.

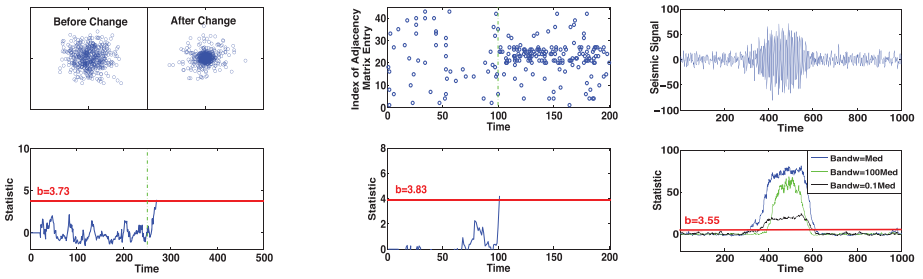


Figure 2. Examples of scan B-statistics with $N = 5$. All thresholds (red lines) are theoretical values obtained from Theorem 4.1 (offline) and Theorem 4.2 (online). (a) Gaussian to GMM, $\tau = 250$; (b) graphs, $\tau = 100$; and (c) real seismic signal.

Gaussian to Gaussian mixture. In Figure 2(a), $P = \mathcal{N}(0, I_2)$, Q is a mixture Gaussian: $0.3\mathcal{N}(0, I_2) + 0.7\mathcal{N}(0, 0.1I_2)$, and $\tau = 250$. The online procedure stops at time 270, meaning that the change is detected with a small delay of 20 unit times.

Sequence of graphs. In Figure 2(b), we consider detecting the emergence of a community inside a network, which is modeled using a stochastic block model, as considered by Maragoni-Simonsen and Xie (2015). Assume that before the change, each sample is a realization of an Erdős-Rényi random graph, with the probability of forming an edge $p_0 = 0.1$ uniformly across the graph. After the change, a "community" emerges, which is a subset of nodes where the edges are formed in between these nodes with much higher probability $p_1 = 0.3$. The postchange distribution models a community where the members of the community interact more often. Our online procedure stops at time 102, meaning that the change is detected with a small delay of 2 unit times.

Real seismic signal; effect of kernel bandwidth. In Figure 2(c), we consider a segment of a real seismic signal that contains a change point. Using the seismic signal, we illustrate the effect of different kernel bandwidths. For Gaussian RBF kernel $k(Y, Y') = \exp(-||Y - Y'||^2/2\sigma^2)$, the kernel bandwidth $\sigma > 0$ is typically chosen using a "median trick" in Schölkopf and Smola (2001) and Ramdas et al. (2015), where σ is set to be the median of the pairwise distances between data points.

4. Theoretical approximations

4.1. Theoretical approximation for significance level of offline scan B-statistic

In the offline setting, the choice of the threshold b involves a tradeoff between two standard performance metrics: (1) significance level (SL), which is the probability that the statistic exceeds the threshold b when the null hypothesis is true (i.e., when there is no change), and (2) power, which is the probability that the statistic exceeds the threshold when the alternative hypothesis is true.

We present an accurate approximation to the SL of the offline scan B-statistic, assuming that the detection threshold b tends to infinity and the number of blocks N is fixed. The following theorem is our main result.

Theorem 4.1. (SL of offline scan B-statistic). *When $b \to \infty$, and $B_{\max} \to \infty$, with $b/(B_{\max})^{1/2}$ held as a fixed positive constant, the significance level of the offline B-statistic defined in (3.2) is given by*

$$\mathbb{P}\left\{\max_{2\le B\le B_{\max}} Z'_B > b\right\} = be^{-\frac{1}{2}b^2}\cdot\sum_{B=2}^{B_{\max}}\frac{(2B-1)}{2\sqrt{2\pi}B(B-1)}\nu\left(b\sqrt{\frac{2B-1}{B(B-1)}}\right)\cdot[1+o(1)], \tag{4.1}$$

where the special function

$$\nu(\mu)\approx\frac{(2/\mu)(\Phi(\mu/2)-0.5)}{(\mu/2)\Phi(\mu/2)+\phi(\mu/2)}, \tag{4.2}$$

and $\phi(x)$ and $\Phi(x)$ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively.

Although the approximation (4.1) is derived in the asymptotic regime and under the assumption that the collection of random variables $\{Z'_B\}_{B=2,\ldots,B_{\max}}$ form a Gaussian random field, we can show numerically that (4.1) is quite accurate in the nonasymptotic regime. Consider synthetic data that are i.i.d. normal $P=\mathcal{N}(0,I_{20})$. We set $B_{\max}$ to be 50, 100, 150, and in each case, $N=5$. We compare the thresholds obtained by (4.1) and by simulation for a prescribed SL α. To obtain the threshold by simulation, we generate Monte Carlo trials for offline B-statistics and find the $(1-\alpha)$-quantile as the estimated threshold. Table 1 shows that for various choices of $B_{\max}$, the thresholds predicted by Theorem 4.1 match quite well with those obtained by simulation. The accuracy can be further improved for smaller α values by skewness correction as shown in Section 6.

The complete proof of Theorem 4.1 can be found in Appendix 8, which leverages the change-of-measure technique. In a nutshell, we aim to find the probability of a rare event: under the null the distribution, the boundary exceeding event $\{\max_{2\le B\le B_{\max}} Z'_B > b\}$ for a large threshold b is rare (so that false alarm remains low). Because quantifying such a small probability is hard under the null distribution, we consider an alternative probability measure under which this boundary-exceeding event happens with much higher probability. Under the new measure, one can use the local central limit theorem to obtain an analytical expression for the probability. In the end, the original small probability will be related to the probability under the alternative measure using Mill's ratio in Yakir (2013).

The proof assumes that the collection of random variables $\{Z'_B\}_{B=2,\ldots,B_{\max}}$ forms a Gaussian random field (as an approximation). This means that the finite-dimensional joint distributions of the collection of random variables are all Gaussian, and they are completely specified by the mean and the covariance functions, which we characterize below (this is useful for establishing Theorem 4.1). These results will be used when we quantify the tail probability of the scan B-statistics. Under the null distribution, the expectation $\mathbb{E}[Z'_B]$ is zero due to the unbiased property of the MMD estimator. The covariance under the null distribution is given by the following lemma.

Lemma 4.1. (Covariance structure of Z'_B in the offline setting). *Under the null distribution, the covariance of $\{Z'_B\}_{B=2,\ldots,B_{\max}}$ is given by*

Table 1. Thresholds for the offline scan B-statistics using ***synthetic data***, obtained by simulation, theory (Theorem 4.1), and theory with skewness correction (Section 6), respectively, for various SL values α.

	$B_{\max} = 50$			$B_{\max} = 100$			$B_{\max} = 150$		
α	b (simulation)	b (theory)	b (skewness correction)	b (simulation)	b (theory)	b (skewness correction)	b (simulation)	b (theory)	b (skewness correction)
0.10	2.41	**2.38**	2.57	2.43	**2.50**	2.76	2.53	**2.56**	2.89
0.05	2.77	**2.67**	2.97	2.76	**2.78**	3.17	2.97	**2.83**	3.22
0.01	3.54	3.23	**3.64**	3.47	**3.32**	3.82	3.64	3.37	**3.89**

$$r_{u,v} = \mathrm{Cov}(Z'_u, Z'_v) = \sqrt{\binom{u}{2}\binom{v}{2}} \Big/ \binom{u \vee v}{2}, \quad 2 \le u, v \le B_{\max}, \tag{4.3}$$

where $u \vee v = \max\{u, v\}$.

The proof can be found in Appendix B.2.

4.2. Theoretical approximation for ARL of online scan B-statistic

In the online setting, two commonly used performance metrics are (see, e.g., Y. Xie and Siegmund, 2013): (1) the ARL, which is the expected time before incorrectly announcing a change of distribution when none has occurred, and (2) the expected detection delay (EDD), which is the expected time to fire an alarm when a change occurs immediately at $\tau = 0$. The EDD considers the worst case and provides an upper bound on the expected delay to detect a change point when the change occurs later in the sequence of observations.

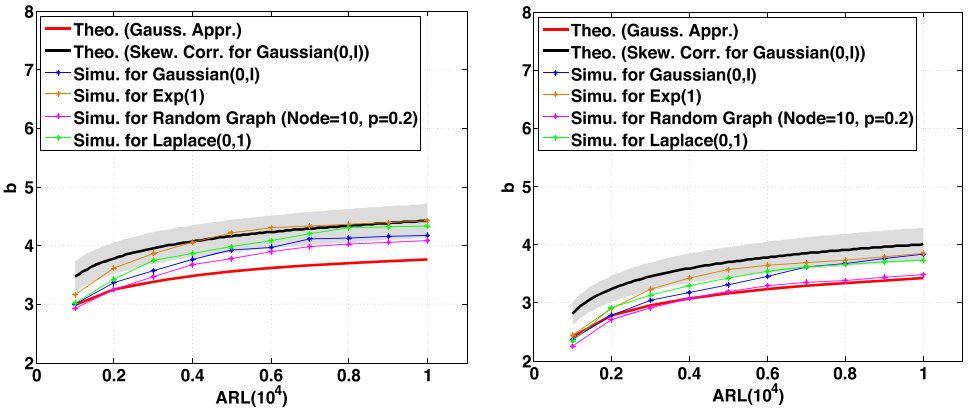
We present an accurate approximation to the ARL of online scan B-statistics. The approximation is quite useful in setting the threshold. As a result, given a target ARL, one can determine the corresponding threshold value b from the analytical approximation, avoiding the more expensive numerical simulations. Our main result is the following theorem.

Theorem 4.2. (ARL in online scan B-statistic). *Let* $B_0 \ge 2$. *When* $b \to \infty$, *the ARL of the stopping time* T *defined in (3.4) is given by*

$$\mathbb{E}[T] = \frac{e^{b^2/2}}{b} \cdot \left\{ \frac{(2B_0 - 1)}{\sqrt{2\pi} B_0 (B_0 - 1)} \cdot \nu\left(b\sqrt{\frac{2(2B_0 - 1)}{B_0(B_0 - 1)}} \right) \right\}^{-1} \cdot [1 + o(1)]. \tag{4.4}$$

The complete proof of Theorem 4.2 is given in Appendix C.

We verify the accuracy of the approximation numerically, by comparing the thresholds obtained by Theorem 4.2 with those obtained from Monte Carlo simulation. Consider several cases of null distributions: standard normal $\mathcal{N}(0, 1)$, exponential distribution with mean 1, Erdős-Rényi random graph with 10 nodes and probability of 0.2 of forming random edges, as well as a Laplace distribution with zero mean and unit variance. The simulation results are obtained from 5,000 direct Monte Carlo trials. As shown in Figure 3, the thresholds predicted by Theorem 4.2 are quite accurate. Figure 3 also demonstrates that theory is quite accurate for various block sizes (especially for



(a): $B_0 = 50$ (b): $B_0 = 200$

Figure 3. For a range of target ARL values, thresholds determined from simulation, from Theorem 4.2, and from theory with the skewness correction (Section 6) under various null distributions are compared. Shaded areas represent standard deviations for skewness-corrected thresholds.

larger B_0). However, we also note that theory tends to underestimate the thresholds. This is especially pronounced for small B_0; for example, $B_0 = 50$. The accuracy of the theoretical results can be improved by skewness correction, shown by black lines in Figure 3, which is discussed later in Section 6.

Theorem 4.2 shows that ARL is $\mathcal{O}(e^{b^2})$ and, hence, b is $\mathcal{O}((\log \text{ARL})^{1/2})$. Note that EDD is typically on the order of b/Δ due to Wald's identity (Siegmund, 1985), where Δ is the Kullback-Leibler divergence between the null and the alternative distributions (a constant). Hence, given the desired ARL (typically on the order of 5,000 or 10,000), the error in the estimated threshold will only be translated linearly to EDD. This is a blessing because it means that typically a reasonably accurate b will cause little performance loss in EDD. Similarly, Theorem 4.1 shows that SL is $\mathcal{O}(e^{-b^2})$ and a similar argument can be made for the offline case.

5. Detection power study

In this section, we study the detection power and the expected detection delay of the offline and online scan B-statistics, respectively, and compare them with classic methods.

5.1. Offline change-point detection: comparison with parametric statistics

We compare the offline scan B-statistic with two commonly used parametric test statistics: Hotelling's T^2 and the generalized likelihood ratio (GLR) statistics. Assume samples $\{x_1, x_2, \ldots, x_n\}$.

Hotelling's T^2 statistic. For a hypothetical change-point location τ, we can define the Hotelling's T^2 statistic for samples in two segments $[1, \tau]$ and $[\tau + 1, t]$ as

$$T^2(\tau) = \frac{\tau(n-\tau)}{n}(\bar{x}_\tau - \bar{x}^*_\tau)^T \widehat{\Sigma}^{-1}(\bar{x}_\tau - \bar{x}^*_\tau),$$

where $\bar{x}_\tau = \sum_{i=1}^{\tau} x_i/\tau, \bar{x}^*_\tau = \sum_{i=\tau+1}^{n} x_i/(n-\tau)$ and the pooled covariance estimator

$$\widehat{\Sigma} = (n-2)^{-1}\left(\sum_{i=1}^{\tau}(x_i-\bar{x}_i)(x_i-\bar{x}_i)^T + \sum_{i=\tau+1}^{n}(x_i-\bar{x}_i^*)(x_i-\bar{x}_i^*)^T\right).$$

Hotelling's T^2 test detects a change whenever $\max_{1\le\tau\le n}\max T^2(\tau)$ exceeds a threshold. The **GLR** statistic can be derived by assuming that the null and alternative distributions are two multivariate normal distributions and both the mean and the covariance matrix are all unknown. For a hypothetical change-point location τ, the GLR statistic is given by

$$\ell(\tau) = n\log|\widehat{\Sigma}_n| - \tau\log|\widehat{\Sigma}_\tau| - (n-\tau)\log|\widehat{\Sigma}_\tau^*|,$$

where $\widehat{\Sigma}_\tau = \tau^{-1}(\sum_{i=1}^{\tau}(x_i-\bar{x}_i)(x_i-\bar{x}_i)^T)$, and $\widehat{\Sigma}_\tau^* = (n-\tau)^{-1}\sum_{i=\tau+1}^{n}(x_i-\bar{x}_i^*)(x_i-\bar{x}_i^*)^T$. The GLR statistic detects a change whenever $\max_{1\le\tau\le n}\ell(\tau)$ exceeds a threshold.

For our examples, we set $n = B_{\max} = 200$ for the Hotelling's T^2 and the scan B-statistics, respectively. Let the change point occur at $\tau = 100$ and choose the significance level $\alpha = 0.05$. The thresholds for the offline scan B-statistic are obtained from Theorem 4.1, and for the other two methods the thresholds are obtained from simulations. Consider the following cases.

Case 1. (mean shift): Observe a sequence of observations in $\mathbb{R}^{20}$, whose distribution shifts from $\mathcal{N}(0, I_{20})$ to $\mathcal{N}(0.1\mathbf{e}, I_{20})$.

Case 2. (mean shift with larger magnitude): Observe a sequence of observations in $\mathbb{R}^{20}$, whose distribution shifts from $\mathcal{N}(0, I_{20})$ to $\mathcal{N}(0.2\mathbf{e}, I_{20})$.

Case 3. (mean and local covariance change): Observe a sequence of observations in $\mathbb{R}^{20}$, whose distribution shifts from $\mathcal{N}(\mathbf{e}, I_{20})$ to $\mathcal{N}(0.2\mathbf{e}, \Sigma)$, where $[\Sigma]_{11} = 2$ and $[\Sigma]_{ii} = 1$, $i = 2, \ldots, 20$.

Case 4. (Gaussian to Laplace): Observe a sequence of one-dimensional observations, whose distribution shifts from $\mathcal{N}(0,1)$ to Laplace distribution with zero mean and unit variance. Note that the mean and variance remain the same after the change.

We estimate the power for each case using 100 Monte Carlo trials. Table 2 shows that the scan B-statistic achieves higher power than the Hotelling's T^2 statistic as well as the GLR statistic in all cases. The GLR statistic performs poorly, because when τ is small or closer to the end point, it estimates the prechange and postchange sample covariance matrix using a very limited number of samples.

5.2. Online change-point detection: comparison with hotelling's T^2 Statistics

Now consider the online scan B-statistic with a fixed block-size $B_0 = 20$. We compare the online scan B-statistic with a Shewhart chart based on Hotelling's T^2 statistic. Here we made no comparison of the online scan B-statistic with the GLR statistic, because in our experiments, Hotelling's T^2 consistently outperformed GLR when the dimension was high. At each time t, we form a Hotelling's T^2 statistic using the immediately past B_0 samples in $[t - B_0 + 1, t]$,

$$T^2(t) = B_0(\bar{x}_t - \widehat{\mu})^T\widehat{\Sigma}_0^{-1}(\bar{x}_t - \widehat{\mu}_0),$$

where $\bar{x}_t = (\sum_{i=t-B_0+1}^{t} x_i)/B_0$, and $\widehat{\mu}_0$ and $\widehat{\Sigma}_0$ are estimated from reference data. The

Table 2. Comparison of detection power for offline change-point detection. Thresholds for all methods are calibrated so that the significance level is $\alpha = 0.05$.

	Case 1	Case 2	Case 3	Case 4
B-statistic	**0.71**	**1.00**	**1.00**	**0.44**
Hotelling's T^2	0.18	0.88	0.87	0.03
GLR	0.03	0.05	0.12	0.04

Table 3. Comparison of EDD in online change-point detection. The parameter is $B_0 = 20$ and thresholds for all methods are calibrated so that ARL $= 5,000$. Dashes indicate that the procedure failed to detect the change; that is, EDD was longer than 50.

	Case 1	Case 2	Case 3	Case 4	Case 5
B-statistic	4.20	**9.10**	**1.00**	**23.38**	**23.03**
Hotelling's T^2	**2.47**	25.46	1.27	–	–

procedure detects a change point whenever $T^2(t)$ exceeds a threshold for the first time. The threshold for the online scan B-statistic is obtained from Theorem 4.2, and from simulations for the Hotelling's T^2 statistic. To simulate EDD, let the change occur at the first point of the testing data. Consider the following cases:

Case 1. (mean shift): Distribution shifts from $\mathcal{N}(0, I_{20})$ to $\mathcal{N}(0.31, I_{20})$.

Case 2. (covariance change): Distribution shifts from $\mathcal{N}(0, I_{20})$ to $\mathcal{N}(0, \Sigma)$, where $[\Sigma]_{ii} = 2, i = 1, 2, ..., 5,$ and $[\Sigma]_{ii} = 1, \ i = 6, ..., 20.$

Case 3. (covariance change): Distribution shifts from $\mathcal{N}(0, I_{20})$ to $\mathcal{N}(0, 2I_{20})$.

Case 4. (Gaussian to Gaussian mixture): Distribution shifts from $\mathcal{N}(0, I_{20})$ to mixture Gaussian $0.3\mathcal{N}(0, I_{20}) + 0.7\mathcal{N}(0, 0.1I_{20})$.

Case 5. (Gaussian to Laplace): Distribution shifts from $\mathcal{N}(0, 1)$ to a Laplace distribution with zero mean and unit variance. For these difficult situations, we report the EDD comparisons based on the selected 500 sequences where B-statistics successfully detected the changes, defined as crossing the threshold within 50 steps from the time that the change occurs. Hotelling's T^2 failed to detect the changes for all sequences.

We evaluated the EDD for each case using 500 Monte Carlo trials. The results are summarized in Table 3. Note that in detecting changes in either Gaussian mean or covariance, the online scan B-statistic performed competitively with Hotelling's T^2, which is tailored to the Gaussian distribution. In the more challenging scenarios such as case 4 and case 5, the Hotelling's T^2 failed to detect the change point, whereas the online scan B-statistic detected the change fairly quickly.

6. Skewness correction

We have shown that approximations to the significance level and ARL, assuming that random variables $\{Z'_B\}_{B=2,3,...}$ form a Gaussian random field, are reasonably accurate. However, Z'_B does not converge to a normal distribution even when B is large (see Appendix 8) and it has a non-vanishing skewness, as illustrated by the following numerical example. Form 10,000 instances of Z_B computed using samples from $\mathcal{N}(0, I_{20})$. Figures 4(a)–(b) show the

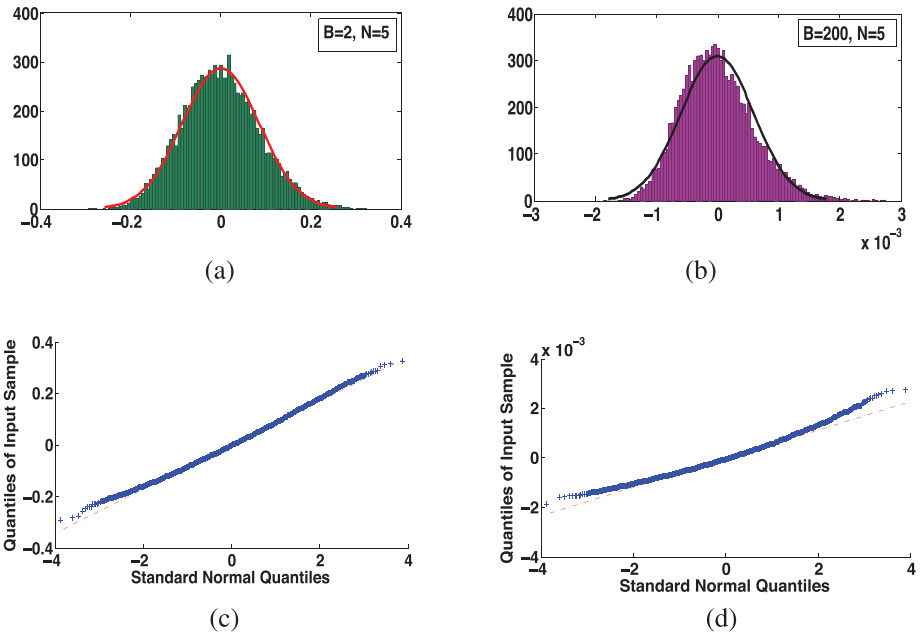


Figure 4. Empirical distributions of Z_B when $B=2$ and $B=200$, respectively. Note that although Gaussian distribution seems to be a reasonable fit to the statistic, the skewness becomes larger for larger values of B. (a) $B=2$, $N=5$, empirical distribution; (b) $B=200$, $N=5$, empirical distribution; (c) $B=2$, $N=5$, Q-Q plot; and (d) $B=200$, $N=5$, Q-Q plot.

empirical distributions of Z_B when $N=5$ and $B=2$ or $B=200$, respectively. Also plotted are the Gaussian probability density functions with mean equal to the sample mean and the variance predicted by Lemma 3.1. Note that the empirical distributions of Z_B match with the Gaussian distributions to a certain extent but the skewness becomes larger for larger B. Figures 4(c)–(d) show the corresponding Q-Q plots.

To incorporate the skewness of Z_B, one can improve the accuracy of the approximations for significance level in Theorem 4.1 and for ARL in Theorem 4.2. Note that the log moment generating function $\psi(\theta)$ defined in (C.1) corresponds to the cumulant generating function (McCullagh and Kolassa, 2009) and it has an expansion for θ close to zero:

$$\psi(\theta) = \kappa_1\theta + \frac{\kappa_2}{2}\theta^2 + \frac{\kappa_3}{3!}\theta^3 + o(\theta^3).$$

Because $\mathbb{E}[Z'_B] = 0$, the cumulants take values $\kappa_1 = \mathbb{E}[Z'_B] = 0, \kappa_2 = \mathrm{Var}[Z'_B] = 1, \kappa_3 = \mathbb{E}[(Z'_B)^3] - 3\mathbb{E}[(Z'_B)^2]\mathbb{E}[Z'_B] + 2(\mathbb{E}[Z'_B])^3 = \mathbb{E}[(Z'_B)^3]$. Recall that when deriving approximations using change-of-measurement, we choose parameter θ such that $\dot{\psi}(\theta) = b$. If Z'_B is a standard normal, $\psi(\theta) = \theta^2/2$, and hence $\theta = b$. Now with skewness correction, we approximate $\psi(\theta)$ as $\theta^2/2 + \kappa_3\theta^3/6$ when solving for θ. Hence, we solve for

$$\dot{\psi}(\theta) \approx \theta + \mathbb{E}\left[(Z'_B)^3\right]\theta^2/2 = b$$

and denote the solution to be θ_B (note that this time the solution depends on B). Moreover, with skewness correction, we will change the leading exponent term in (4.1) and (D.4) from $e^{-b^2/2}$ to be $e^{\psi(\theta'_B)-\theta'_B b}$.

Table 4. Thresholds for the offline scan B-statistics using ***real speech data***, obtained by simulation, theory (Theorem 4.1), and theory with skewness correction, respectively, for various significance levels α.

	$B_{\max} = 50$			$B_{\max} = 100$			$B_{\max} = 150$		
α	b (bootstrap)	b (theory)	b (skewness correction)	b (bootstrap)	b (theory)	b (skewness correction)	b (bootstrap)	b (theory)	b (skewness correction)
0.10	2.96	2.38	**3.23**	3.16	2.50	**3.59**	3.21	**2.56**	3.94
0.05	3.62	2.67	**3.68**	3.82	2.78	**4.06**	3.86	2.83	**4.43**
0.01	4.85	3.23	**4.61**	5.20	3.32	**5.03**	5.42	3.37	**5.45**

From numerical experiments, we find that the skewness correction is especially useful when the significance level is small (e.g., $\alpha = 0.01$) for the offline case, when block size B_0 is small (see Table 1 and Figure 3) and can be important for real data where the data are noisy and the null distribution is more difficult to characterize.

For example, we consider real speech data from the CENSREC-1-C data set (more details in Section 7). Here, the null distribution P corresponds to the unknown distribution of the background signal, and we are interested in detecting the onset of speech signals. This case is more challenging because the true distribution can be arbitrary. In the data set, there are 3,000 reference samples. We bootstrap these reference samples to generate 10,000 resamples to estimate the tail of the detection statistic. Table 4 demonstrates that the thresholds predicted by the expensive bootstrapping (by Theorem 4.1) and by theory with skewness correction, respectively, for various SL values α. Note that in this case, the accuracy improves significantly by skewness correction.

The remaining task is to estimate the skewness of the scan B-statistic. Because Z_B is zero-mean, the skewness of Z'_B is related to the variance and third moment of Z_B via

$$\kappa_3 = \mathbb{E}\left[(Z'_B)^3\right] = \mathrm{Var}[Z_B]^{-3/2}\mathbb{E}\left[Z_B^3\right].$$

We already know how to estimate the variance of Z_B from Lemma 3.1. The following lemma shows the third-order moment $\mathbb{E}[Z_B^3]$ in terms of the moments of the kernel h defined in (2.2):

Lemma 6.1. (Third-order moment of Z_B).

$$\begin{aligned}
\mathbb{E}\left[Z_B^3\right] = {} & \frac{8(B-2)}{B^2(B-1)^2}\left\{\frac{1}{N^2}\mathbb{E}[h(x,x',y,y')h(x',x'',y',y'')h(x'',x,y'',y)]\right. \\
& + \frac{3(N-1)}{N^2}\mathbb{E}[h(x,x',y,y')h(x',x'',y',y'')h(x''',x'''',y'',y)] \\
& \left. + \frac{(N-1)(N-2)}{N^2}\mathbb{E}[h(x,x',y,y')h(x'',x''',y',y'')h(x'''',x''''',y'',y)]\right\} \\
& + \frac{4}{B^2(B-1)^2}\left\{\frac{1}{N^2}\mathbb{E}[h(x,x',y,y')^3]\right. \\
& + \frac{3(N-1)}{N^2}\mathbb{E}[h(x,x',y,y')^2h(x'',x''',y,y')] \\
& \left. + \frac{(N-1)(N-2)}{N^2}\mathbb{E}[h(x,x',y,y')h(x'',x''',y,y')h(x'''',x''''',y,y')]\right\}.
\end{aligned} \tag{6.1}$$

The proof can be found in Appendix E. Lemma 6.1 enables us to estimate the skewness efficiently, by reducing it to evaluating simpler terms in (6.1) that only requires estimating the statistic of the kernel function $h(\cdot, \cdot, \cdot, \cdot)$ with tuples of samples.

Finally, although Z'_B does not converge to Gaussian, the difference between its moment generating functions and that of the standard normal distribution can be bounded, as we show below. By applying an argument on page 220 of Yakir (2013), we obtain that

$$\left|\mathbb{E}[e^{\theta Z'_B}] - \left(1 + \frac{\theta^2}{2}\right)\right| \leq \min\left\{\frac{|\theta|^3}{6}\mathbb{E}\left[|Z'_B|^3\right], \theta^2\mathbb{E}\left[|Z'_B|^2\right]\right\}.$$

If considering the skewness κ_3 of Z'_B, we have a better estimation

$$\left|\mathbb{E}[e^{\theta Z'_B}] - \left(1 + \frac{\theta^2}{2} + \frac{\theta^3\kappa_3}{6}\right)\right| \leq \min\left\{\frac{\theta^4}{24}\mathbb{E}\left[|Z'_B|^4\right], \frac{1}{3}|\theta|^3\mathbb{E}\left[|Z'_B|^3\right]\right\}.$$

7. Real data

We test the performance of the scan B-statistics for change-point detection on real data. Our data sets include (1) CENSREC-1-C: a real-world speech data set in the Speech Resource Consortium corpora provided by National Institute of Informatics (Available from http://research.nii.ac.jp/src/en/CENSREC-1-C.html) and (2) Human Activity Sensing Consortium (HASC) Challenge 2011 data (Available from http://hasc.jp/hc2011). We compare our proposed scan B-statistics with a baseline algorithm, the relative density-ratio (RDR) estimate (Song et al., 2013). One limitation of the RDR algorithm, however, is that it is not suitable for high-dimensional data because estimating the density ratio in the high-dimensional setting is an ill-posed problem. To achieve reasonable performance for the RDR algorithm, we adjust the bandwidth and the regularization parameter at each time step and, hence, the RDR algorithm is computationally more expensive than using the scan B-statistics. We adopt the standard area under curve (AUC) as in Song et al. (2013) for our performance metric. The larger the AUC, the better.

Our scan B-statistics demonstrate competitive performance compared with the baseline RDR algorithm on the real data. Here we only report the main results and leave the details in Appendix 8. For speech data, our goal is to online detect the emergence of a speech signal from the background. The backgrounds are taken from real acoustic signals, such as noise recorded on highways, in airports, and in subway stations. The overall AUC for the scan B-statistic is ***0.8014*** and for the baseline algorithm it is ***0.7578***. For human activity detection data, our goal is to detect a transition from one activity to another as quickly as possible. Each instance consists of six possible human activity signals collected by portable three-axis accelerometers. The overall AUC for the scan B-statistic is ***0.8871*** and for the baseline algorithm it is ***0.7161***.

8. Discussion

There are a few possible directions to extend our work. 1. Thus far, we have assumed that data are i.i.d. from a null distribution P and, when the change happens, data are i.i.d. from an alternative distribution Q. Under these assumptions, we have developed the offline and online change-point detection algorithms based on the two-sample nonparametric test statistic MMD. One may relax the temporal independence assumption and extend scan B-statistics for dependent data by incorporating ideas from Chwialkowski and Gretton (2014). 2. We have demonstrated how the number of blocks and block size affect the performance of scan B-statistics. One can also explore how kernel bandwidth, as well as the dimensionality of the data, would affect the performance. An empirical observation is that the performance of the MMD statistic degrades with the increasing dimensions of data. Some recent results for the kernel-based test can be found in Ramdas et al. (2015). We may adopt the idea of Ramdas et al. (2015) to extend our scan B-statistics for detecting a change in high dimensions. 3. For an exceedingly high-dimensional data set with large Gram matrix, one can perform random subsampling to reduce complexity similar to B. Xie et al. (2015).

Acknowledgment

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

Appendix A: Recursive implementation of online scan B-statistic

The online scan B-statistic can be computed recursively via a simple update scheme. By its construction, when time elapses from t to $(t+1)$, a new sample is added to the postchange block and the oldest sample is moved to the reference pool. Each reference block is updated similarly by adding one sample randomly drawn from the pool of reference data, and the oldest sample is purged. Hence, only a limited number of entries in the Gram matrix due to the new

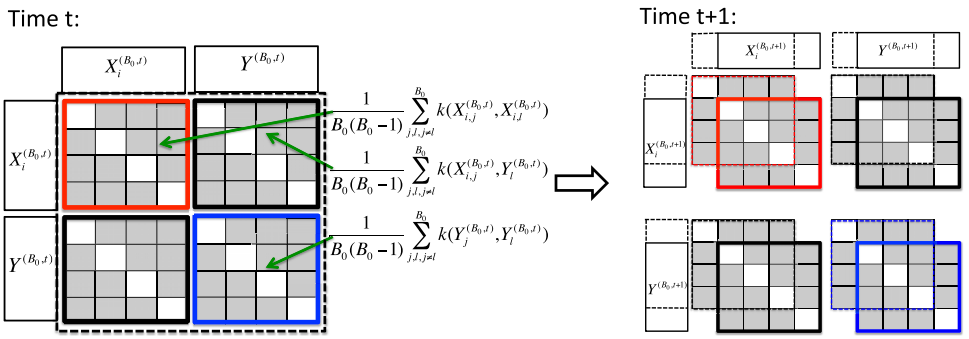


Figure A.1. Recursive updating scheme to compute the online scan B-statistics. The online B-statistic is formed with N background blocks and one testing block and, hence, we keep track of N Gram matrices. For illustration purposes, we partition the Gram matrix into four windows (in red, black, and blue, as shown on the left panel). At time t, to obtain $\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)})$, we compute the shaded elements and take an average within each window. The diagonal entries in each window are removed to obtain an unbiased estimate. At time $t+1$, we update $X_i^{(B_0,t)}$ and $Y^{(B_0,t)}$ with the new data point and purge the oldest data point, and update the Gram matrix by moving the colored window as shown in the right panel. We compute the elements within the new windows and take an average. Note that we only need to compute the right-most column and the bottom row.

sample will be updated. The update scheme is illustrated in Figure A.1 and explained in more details therein. Similarly, the offline scan B-statistic can also be computed recursively by utilizing the fact that Z_B for $B \in \{2, \ldots, B_{\max}\}$ shares many common terms.

Appendix B: Variance and covariance calculation

Below, $X_{i,j}^{(B)}$, where $i = 1, \ldots, N$, and $j = 2, \ldots, B_{\max}$, denotes the jth sample in the ith block $X_i^{(B)}$, and $Y_j^{(B)}$ denotes the jth sample in $Y^{(B)}$. The superscript B denotes the block size. We start with proving Lemma B.1 and Lemma B.2, which are useful in proving Lemma 3.1.

Lemma B.1. (Variance of MMD, under the null.). *Under the null hypothesis,*

$$\mathrm{Var}\left[\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)})\right] = \binom{B}{2}^{-1} \mathbb{E}[h^2(x, x', y, y')], \quad i = 1, \ldots, N. \tag{B.1}$$

Proof. For notational simplicity, below we drop the superscript B, which denotes the block size. Furthermore, we use x, x', y, and y' to denote generic samples; that is, $X_{i,l} =^d x, X_{i,j} =^d x', Y_l =^d y, Y_j =^d y'$ and they are mutually independent of each other. Here the notation $=^d$ means two random variables have the same distribution. Below, we follow the same convention. For any $i = 1, 2, \ldots, n$, by definition of the U-statistic, we have

$$\begin{aligned}
\mathrm{Var}[\mathrm{MMD}^2(X_i, Y)] &= \mathrm{Var}\left[\binom{B}{2}^{-1} \sum_{l<j} h(X_{i,l}, X_{i,j}, Y_l, Y_j)\right] \\
&= \binom{B}{2}^{-2} \left[\binom{B}{2}\binom{2}{1}\binom{B-2}{2-1} \mathrm{Var}[\mathbb{E}_{x,y}[h(x, x', y, y')]] \right. \\
&\quad \left. + \binom{B}{2}\binom{2}{2}\binom{B-2}{2-2} \mathrm{Var}[h(x, x', y, y')]\right].
\end{aligned} \tag{B.2}$$

Under a null distribution, $\mathbb{E}_{x,y}[h(x, x', y, y')] = 0$. Thus, $\mathrm{Var}[\mathbb{E}_{x,y}[h(x, x', y, y')]] = 0$, and

$$\mathrm{Var}[h(x, x', y, y')] = \mathbb{E}\left[h^2(x, x', y, y')\right] - \mathbb{E}\left[h(x, x\prime, y, y\prime)\right]^2 = \mathbb{E}\left[h^2(x, x', y, y')\right].$$

Substitute these results into (B.2), and we obtain the desired result (B.1). □

Lemma B.2. (Covariance of MMD, under the null, different block index.). *For $s \neq 0$, under a null hypothesis*

$$\begin{aligned}
&\mathrm{Cov}[\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)}), \mathrm{MMD}^2(X_{i+s}^{(B)}, Y^{(B)})] \\
&= \binom{B}{2}^{-1} \mathrm{Cov}[h(x, x', y, y'), h(x'', x''\prime, y, y')].
\end{aligned}$$

Proof. For $i = 1, 2, \ldots, N$, and $s = (1-i), (2-i), \ldots, (N-i), s \neq 0$,

$$\begin{aligned}
&\mathrm{Cov}[\mathrm{MMD}^2(X_i, Y), \mathrm{MMD}^2(X_{i+s}, Y)] \\
&= \mathrm{Cov}\left[\binom{B}{2}^{-1} \sum_{l<j} h(X_{i,l}, X_{i,j}, Y_l, Y_j), \binom{B}{2}^{-1} \sum_{p<q} h(X_{i+s,p}, X_{i+s,q}, Y_p, Y_q)\right] \\
&= \binom{B}{2}^{-2} \binom{B}{2}\binom{2}{1}\binom{B-2}{2-1} \mathrm{Cov}[h(x, x', y, y'), h(x'', x''\prime, y, y'')] \\
&\quad + \binom{B}{2}^{-2} \binom{B}{2}\binom{2}{2}\binom{B-2}{2-2} \mathrm{Cov}[h(x, x', y, y'), h(x'', x''\prime, y, y')].
\end{aligned}$$

Under a null distribution,

$$\begin{aligned}
&\mathrm{Cov}[h(x,x',y,y'),h(x'',x''\prime,y,y'')]\\
&=\int h(x,x',y,y')h(x'',x''\prime,y,y'')d\mathbb{P}(x,x',x'',x''\prime,y,y',y'')\\
&=\int\Big(\underbrace{\int h(x,x',y,y')d\mathbb{P}(x',y')}_{=0}\Big)d\mathbb{P}(x)\cdot\int\Big(\underbrace{\int h(x'',x''',y,y'')d\mathbb{P}(x'',y'')}_{=0}\Big)d\mathbb{P}(x''')=0.
\end{aligned}$$

Above, with a slight abuse of notation, we use $d\mathbb{P}(\cdot)$ to denote the probability measure of appropriate arguments. Finally, we have the desired results as shown in Lemma B.2. □

B.1. Variance of scan B-statistics

Proof. for Lemma 3.1. Using results in Lemma B.1 and Lemma B.2, we have

$$\begin{aligned}
\mathrm{Var}[Z_B]&=\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\mathrm{MMD}^2(X_i,Y)\right]\\
&=\frac{1}{N^2}\left[N\mathrm{Var}\left[MMD^2(X_i,Y)\right]+\sum_{i\neq j}\mathrm{Cov}[\mathrm{MMD}^2(X_i,Y;B),\mathrm{MMD}^2(X_j,Y)]\right]\\
&=\binom{B}{2}^{-1}\left[\frac{1}{N}\mathbb{E}\left[h^2(x,x',y,y')\right]+\frac{N-1}{N}\mathrm{Cov}[h(x,x',y,y'),h(x'',x''\prime,y,y')]\right].
\end{aligned}$$

□

Next, we introduce Lemma B.3 and Lemma B.4, which are useful in proving Lemma 4.1.

Lemma B.3. (Covariance of MMD, different block sizes, same block index). *For blocks with the same index i but with distinct block sizes, under the null hypothesis we have*

$$\mathrm{Cov}\left[\mathrm{MMD}^2(X_i^{(B)},Y^{(B)}),\mathrm{MMD}^2(X_i^{(B+\nu)},Y^{(B+\nu)})\right]=\binom{B\vee(B+\nu)}{2}^{-1}\mathbb{E}[h^2(x,x',y,y')]. \tag{B.3}$$

Proof. Note that

$$\begin{aligned}
&\mathrm{Cov}\left[\mathrm{MMD}^2(X_i^{(B)},Y^{(B)}),\mathrm{MMD}^2(X_i^{(B+\nu)},Y^{(B+\nu)})\right]\\
&=\mathrm{Cov}\left[\binom{B}{2}^{-1}\sum_{l<j}^{B}h(X_{i,l},X_{i,j},Y_l,Y_j),\binom{B+\nu}{2}^{-1}\sum_{p<q}^{B+\nu}h(X_{i,p},X_{i,q},Y_p,Y_q)\right]\\
&=\binom{B}{2}^{-1}\binom{B+\nu}{2}^{-1}\mathrm{Cov}\left[\sum_{l<j}^{B}h(X_{i,l},X_{i,j},Y_l,Y_j),\sum_{p<q}^{B+\nu}h(X_{i,p},X_{i,q},Y_p,Y_q)\right]\\
&=\binom{B}{2}^{-1}\binom{B+\nu}{2}^{-1}\binom{B\wedge(B+\nu)}{2}\mathrm{Var}\left[h(x,x',y,y')\right]\\
&=\binom{B\vee(B+\nu)}{2}^{-1}\mathbb{E}\left[h^2(x,x',y,y')\right],
\end{aligned}$$

where the second to last equality is due to a similar argument as before to drop block indices as they are i.i.d. under the null. □

Lemma B.4. (Covariance of MMD, different block sizes, different block indices). *Under the null we have*

$$\mathrm{Cov}\left[\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)}), \mathrm{MMD}^2(X_{i+s}^{(B+v)}, Y^{(B+v)})\right] = \binom{B\vee(B+v)}{2}^{-1} \cdot \mathrm{Cov}[h(x,x',y,y'), h(x'',x''',y,y')].$$

Proof. Note that

$$\begin{aligned}
&\mathrm{Cov}\left[\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)}), \mathrm{MMD}^2(X_{i+s}^{(B+v)}, Y^{(B+v)})\right] \\
&= \mathrm{Cov}\left[\binom{B}{2}^{-1}\sum_{l<j}^{B} h(X_{i,l}^{(B)}, X_{i,j}^{(B)}, Y_l^{(B)}, Y_j^{(B)}), \binom{B+v}{2}^{-1}\sum_{p<q}^{B+v} h(X_{i+s,p}^{(B+v)}, X_{i+s,q}^{(B+v)}, Y_p^{(B+v)}, Y_q^{(B+v)})\right] \\
&= \binom{B}{2}^{-1}\binom{B+v}{2}^{-1}\mathrm{Cov}\left[\sum_{l<j}^{B} h(X_{i,l}^{(B)}, X_{i,j}^{(B)}, Y_l^{(B)}, Y_j^{(B)}), \sum_{p<q}^{B+v} h(X_{i+s,p}^{(B+v)}, X_{i+s,q}^{(B+v)}, Y_p^{(B+v)}, Y_q^{(B+v)})\right] \\
&= \binom{B}{2}^{-1}\binom{B+v}{2}^{-1}\binom{B\wedge(B+v)}{2}\mathrm{Cov}[h(x,x',y,y'), h(x'',x''\prime,y,y')] \\
&= \binom{B\vee(B+v)}{2}^{-1}\mathrm{Cov}[h(x,x',y,y'), h(x'',x''\prime,y,y')],
\end{aligned}$$

where the second to last equality is due to a similar argument as before to drop block indices as they are i.i.d. under the null. □

B.2. Covariance of offline scan B-statistics

Proof of Lemma 4.1. For the offline case, we have that the correlation

$$r_{B,B+v} := \frac{1}{\sqrt{\mathrm{Var}[Z_B]}}\frac{1}{\sqrt{\mathrm{Var}[Z_{B+v}]}}\mathrm{Cov}[Z_B, Z_{B+v}],$$

where

$$\begin{aligned}
\mathrm{Cov}(Z_B, Z_{B+v}) &= \mathrm{Cov}\left[\frac{1}{N}\sum_{i=1}^{N}\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)}), \frac{1}{N}\sum_{j=1}^{n}\mathrm{MMD}^2(X_j^{(B+v)}, Y^{(B+v)})\right] \\
&= \frac{1}{N}\mathrm{Cov}[\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)}), \mathrm{MMD}^2(X_i^{(B+v)}, Y^{(B+v)})] \\
&\quad + \frac{1}{N^2}\sum_{i\neq j}\mathrm{Cov}[\mathrm{MMD}^2(X_i^{(B)}, Y^{(B)}), \mathrm{MMD}^2(X_j^{(B+v)}, Y^{(B+v)})].
\end{aligned}$$

Using results from Lemma B.3 and Lemma B.4, we have

$$\mathrm{Cov}(Z_B, Z_{B+v}) = \binom{B\vee(B+v)}{2}^{-1}\left[\frac{1}{N}\mathbb{E}[h^2(x,x',y,y')] + \frac{N-1}{N}\mathrm{Cov}[h(x,x',y,y'), h(x'',x''',y,y')]\right].$$

Finally, plugging in the expressions for $\mathrm{Var}[Z_B]$ and $\mathrm{Var}[Z_{B+v}]$, we have (4.3) for the offline case.

B.3. Covariance of online scan B-statistic

Similarly, for the online case we need to analyze $\rho_{t,t+s} := \mathrm{Cov}(Z_{B_0,t}', Z_{B_0,t+s}')$. We adopt the same strategy as the above for a fixed block size B_0 to obtain

$$\begin{aligned}
&\mathrm{Cov}\left(\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)}), \mathrm{MMD}^2(X_i^{(B_0,t+s)}, Y^{(B_0,t+s)})\right) \\
&= \mathrm{Cov}\left[\binom{B_0}{2}^{-1}\sum_{l<j}^{B_0} h(X_{i,l}^{(t)}, X_{i,j}^{(t)}, Y_l^{(t)}, Y_j^{(t)}), \binom{B_0}{2}^{-1}\sum_{p<q}^{B_0} h(X_{i,p}^{(t+s)}, X_{i,q}^{(t+s)}, Y_p^{(t+s)}, Y_q^{(t+s)})\right] \\
&= \binom{B_0}{2}^{-2}\binom{(B_0-s)\vee 0}{2}\mathrm{Var}[h(x,x',y,y')].
\end{aligned} \tag{B.4}$$

Figure B.1 (a) demonstrates how $\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)})$ and $\mathrm{MMD}^2(X_i^{(B_0,t+s)}, Y^{(B_0,t+s)})$ are constructed. The shaded areas represent the overlapping data.

Similarly, we have

$$\begin{aligned}
&\mathrm{Cov}\left(\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)}), \mathrm{MMD}^2(X_j^{(B_0,t+s)}, Y^{(B_0,t+s)})\right) \\
&= \mathrm{Cov}\left[\binom{B_0}{2}^{-1}\sum_{l<k}^{B_0} h(X_{i,l}^{(t)}, X_{i,k}^{(t)}, Y_l^{(t)}, Y_k^{(t)}), \binom{B_0}{2}^{-1}\sum_{p<q}^{B_0} h(X_{j,p}^{(t+s)}, X_{j,q}^{(t+s)}, Y_p^{(t+s)}, Y_q^{(t+s)})\right] \\
&= \binom{B_0}{2}^{-2}((B_0-s)\vee 02)\mathrm{Cov}(h(x,x',y,y'), h(x'',x''',y,y')),
\end{aligned} \tag{B.5}$$

Figure B.1 (b) demonstrates how $\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)})$ and $\mathrm{MMD}^2(X_j^{(B_0,t+s)}, Y^{(B_0,t+s)}), j \neq i$, are constructed. The shaded areas represent the overlapping data. Thus,

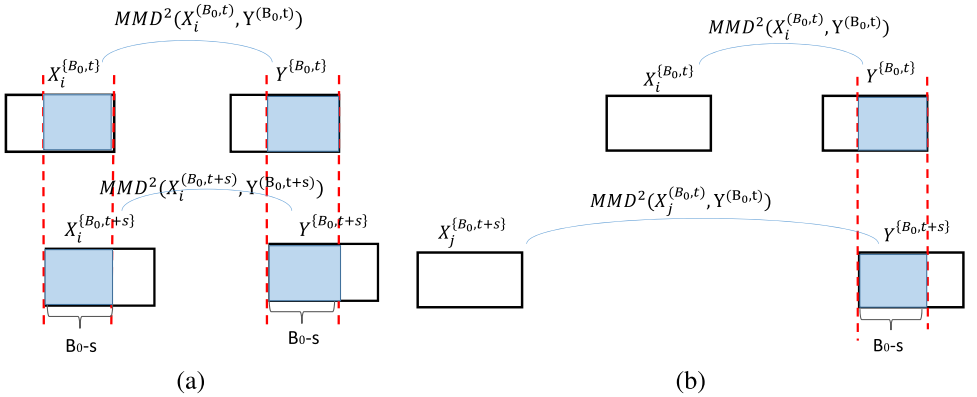


Figure B.1. (a) Illustration of how $\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)})$ and $\mathrm{MMD}^2(X_i^{(B_0,t+s)}, Y^{(B_0,t+s)})$ are constructed in the online change-point detection, where the shaded areas represent the overlapping data. (b) Illustration of how $\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)})$ and $\mathrm{MMD}^2(X_j^{(B_0,t+s)}, Y^{(B_0,t+s)}), j \neq i$, are constructed in the online change-point detection, where the shaded areas represent the overlapping data.

$$
\begin{aligned}
&\mathrm{Cov}(Z_{B_0,t}, Z_{B_0,k+s}) \\
&= \mathrm{Cov}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{MMD}^2(X_i^{(B_0,t)}, Y^{(B_0,t)}), \frac{1}{N}\sum_{j=1}^{N}\mathrm{MMD}^2(X_j^{(B_0,t+s)}, Y^{(B_0,t+s)})\right) \\
&= \binom{B_0}{2}^{-2}\binom{(B_0-s)\vee 0}{2}\left[\frac{1}{N}\mathrm{Var}(h(x,x',y,y')) \quad + \frac{N-1}{N}\mathrm{Cov}(h(x,x',y,y'),h(x'',x''',y,y'))\right].
\end{aligned}
$$

Finally, plugging in the expressions for $\mathrm{Var}[Z_{B_0,t}]$ and $\mathrm{Var}[Z_{B_0,t+s}]$, we have (D.2) for the online case. □

Appendix C: Proof of Theorem 4.1

Below, we present the main steps in proving Theorem 4.1, including (1) exponential tilting; (2) change-of-measure by the likelihood identity; (3) establishing properties of the local field and the global term; and (4) performing asymptotic approximation using the localization theorem (theorem 5.1 in Siegmund et al. [2010] and section 3.4 in Yakir [2013]) by showing that the "global" log likelihood and the "local process" are asymptotically independent. Finally, we collect terms together to obtain the result.

C.1. Step 1: Exponential tilting

We first introduce exponential tilting, which creates a family of distributions that is related to the original distribution of Z'_B. Let the log moment generating function of Z'_B be

$$\psi(\theta) = \log \mathbb{E}[e^{\theta Z'_B}]. \tag{C.1}$$

Define a family of new measures

$$d\mathbb{P}_B = \exp\{\theta Z'_B - \psi(\theta)\}d\mathbb{P}, \tag{C.2}$$

where $\mathbb{P}$ represents the original probability measure of Z'_B under the null distribution P, $\mathbb{P}_B$ is the new measure after the transformation, and θ parameterizes the family of the new measures. Note that the new measures take the form of exponential family, with θ being the parameter.

Recall that, under the null distribution, Z'_B has zero mean and unit variance. Given the assumption that Z'_B is a standard Gaussian random variable, the corresponding log moment generating function is given by $\psi(\theta) = \theta^2/2$. One has the freedom to select the value of θ to determine the new measure. We will set θ such that the mean under the tilted measure is equal to a given threshold b. This means that the new measure peaks at the threshold b, which enables us to use the local central limit theorem later on. This can be done by choosing θ such that $\dot{\psi}(\theta) = b$, and therefore $\theta = b$. Note that the solution θ does not depend on B. Hence, we can set the mean under the transformed measure to b, by uniformly choosing $\theta = b$ for any B. Given such a choice, the transformed measure is given by $d\mathbb{P}_B = \exp\{bZ'_B(x) - b^2/2\}d\mathbb{P}$. We also define, for each B, the log-likelihood ratio $\log(d\mathbb{P}_B/d\mathbb{P})$ of the form

$$\ell_B = bZ'_B - b^2/2. \tag{C.3}$$

This way, we have associated the detection statistic Z'_B with a likelihood ratio, even if Z'_B itself does not come out of a likelihood ratio.

The following lemma shows that Z'_B under the new measure has the same unit variance and its mean has been shifted to b. This key fact will lead to the desired exponential tail.

Lemma C.1. (Mean and variance under tilted measure). *Define* $\mathbb{E}_B$ *and* Var_B *as the expectation and variance under the transformed measures*

$$\mathbb{E}_B[U] = \mathbb{E}[Ue^{\ell_B}], \tag{C.4}$$

$$\mathrm{Var}_B[U] = \mathbb{E}[U^2 e^{\ell_B}] - \mathbb{E}_B^2[U]. \tag{C.5}$$

We have $\mathbb{E}_B[Z'_B] = b$, *and* $\mathrm{Var}_B[Z'_B] = 1$.

Proof. First, $\mathbb{E}_B[Z'_B] = \dot{\psi}(b) = b$ by construction. To show $\mathrm{Var}_B[Z'_B] = 1$, note that $\log \mathbb{E}[e^{bZ'_B}] = b^2/2$. Taking the derivative of $\psi(\theta)$ with respect to b twice gives $\mathbb{E}[(Z'_B)^2 e^{bZ'_B}] = e^{b^2/2} + b^2 e^{b^2/2}$. Hence, $\mathbb{E}_B[(Z'_B)^2] = \mathbb{E}[(Z'_B)^2 e^{\theta Z'_B - \psi(b)}] = 1 + b^2$, and $\mathrm{Var}_B[Z'_B] = \mathbb{E}_B[(Z'_B)^2] - b^2 = 1$. □

The following lemma shows that Z'_B under the new measure has the same unit variance with the mean shifted to b. This key fact will lead to the desired exponential tail.

Lemma C.2. (Mean and variance under tilted measure). *Define* $\mathbb{E}_B$ *and* Var_B *as the expectation and variance under the transformed measures*

$$\mathbb{E}_B[U] = \mathbb{E}[Ue^{\ell_B}], \tag{C.6}$$

$$\mathrm{Var}_B[U] = \mathbb{E}[U^2 e^{\ell_B}] - \mathbb{E}_B^2[U]. \tag{C.7}$$

We have $\mathbb{E}_B[Z'_B] = b$ *and* $\mathrm{Var}_B[Z'_B] = 1$.

Proof. First, $\mathbb{E}_B[Z'_B] = \dot{\psi}(b) = b$ by construction. To show $\mathrm{Var}_B[Z'_B] = 1$, note that $\log \mathbb{E}[e^{bZ'_B}] = b^2/2$. Taking the derivative of $\psi(\theta)$ with respect to b twice gives $\mathbb{E}[(Z'_B)^2 e^{bZ'_B}] = e^{b^2/2} + b^2 e^{b^2/2}$. Hence, $\mathbb{E}_B[(Z'_B)^2] = \mathbb{E}[(Z'_B)^2 e^{\theta Z'_B - \psi(b)}] = 1 + b^2$, and $\mathrm{Var}_B[Z'_B] = \mathbb{E}_B[(Z'_B)^2] - b^2 = 1$. □

C.2. Step 2: Change-of-Measure

Now we are ready to analyze the tail probability $\mathbb{P}\{\max_{2\le B\le B_{\max}} Z'_B > b\}$. The basic idea is to 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.

Here, the alternative measure will be a mixture of simple exponential tilted measures. Define the maximum and the sum for likelihood ratio differences relative to a particular parameter value B:

$$M_B = \max_{s\in\{2,\ldots,B_{\max}\}} e^{\ell_s - \ell_B}, \quad S_B = \sum_{s\in\{2,\ldots,B_{\max}\}} e^{\ell_s - \ell_B}. \tag{C.8}$$

Also define a re-centered likelihood ratio, which we call the *global term*

$$\tilde{\ell}_B = b(Z'_B - b).$$

With the definitions above and the log likelihood ratios ℓ_B in (C.3), we have the following:

$$\begin{aligned}
\mathbb{P}\Big\{\max_{2\le B\le B_{\max}} Z'_B > b\Big\} &= \mathbb{E}\Big[1;\ \max_{2\le B\le B_{\max}} Z'_B > b\Big] = \mathbb{E}\Big[\underbrace{\frac{\sum_{B=2}^{B_{\max}} e^{\ell_B}}{\sum_{s=2}^{B_{\max}} e^{\ell_s}}}_{=1};\ \max_{2\le u\le B_{\max}} Z'_u > b\Big] \\
&= \sum_{B=2}^{B_{\max}} \mathbb{E}\Big[\frac{e^{\ell_B}}{\sum_s e^{\ell_s}};\ \max_{2\le u\le B_{\max}} Z'_u > b\Big] \overset{(C.6)}{=} \sum_{B=2}^{B_{\max}} \mathbb{E}_B\Big[\frac{1}{\sum_s e^{\ell_s}};\ \max_{2\le u\le B_{\max}} Z'_u > b\Big] \\
&= e^{-b^2/2} \sum_{B=2}^{B_{\max}} \mathbb{E}_B\Big[\frac{M_B}{S_B} e^{-(\tilde{\ell}_B + \log M_B)};\ \tilde{\ell}_B + \log M_B \ge 0\Big],
\end{aligned} \tag{C.9}$$

where an intermediate step is done by changing the measure to $\mathbb{P}_B$, and the last equality can be verified by simple algebra. Recall our notation $\mathbb{E}_B[\mathcal{A};\mathcal{B}] = \mathbb{E}_B[\mathcal{A}1\{\mathcal{B}\}]$ for a random quantity $\mathcal{A}$ and event $\mathcal{B}$; 1 denotes an indicator function.

In a nutshell, the last equation in (C.9) 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 conditional expectations under the transformed measures. A close examination of the conditional expectations of the form $\mathbb{E}_B[\cdots;[\cdots]\geq 0]$ reveals that it involves a product of the ratio M_B/S_B and an exponential function that depends on $\tilde{\ell}_B$, which plays the role of weight. Under the new measure $\mathbb{P}_B$, $\tilde{\ell}_B$ has zero mean and variance equal to b^2 (shown in Lemma C.3) and it dominates the other term $\log M_B$ and, hence, the probability of exceeding zero will happen with much higher probability. Next, we characterize the limiting ratio and the other factors precisely, by the localization theorem.

C.3. Step 3: Establish properties of local and global terms

In (C.9), our target probability has been decomposed into terms that only depend on (i) the *local field* $\{\ell_s - \ell_B\}, 2\leq s\leq B_{\max}$, which are the differences between the log-likelihood ratio with parameter B and with other parameter values s, $2\leq s\leq B_{\max}$, and (ii) the *global term* $\tilde{\ell}_B$, which is the centered and scaled likelihood ratio with parameter B. We need to first establish some useful properties of the local field and the global term under the tilted measure. We will eventually show that the local field and the global term are asymptotically independent.

The following property for the global term can be derived from Lemma C.2. The result shows that under the tilted measure, the global term $\tilde{\ell}_B$ has zero mean for any B, with variance diverging with b.

Lemma C.3. (Global term for offline scan B-statistic). *The mean and variance of the global term $\tilde{\ell}_B = b(Z'_B - b)$, for $2\leq B\leq B_{\max}$, are given by*

$$\mathbb{E}_B[\tilde{\ell}_B] = 0, \quad \mathrm{Var}_B[\tilde{\ell}_B] = b^2. \tag{C.10}$$

Assuming Z'_B is approximately normal, the local field $\ell_s - \ell_B$ (or, equivalently, $b(Z'_s - Z'_B)$) and the global term $\tilde{\ell}_B$ (or, equivalently, $b(Z'_B - b)$) are also approximately normally distributed.

Lemma C.4. (Local field for offline scan B-statistic). *The mean and variance of the local field $\{\ell_s - \ell_B\}$, for $|s-B| = 0, 1, 2, \ldots$, are given by*

$$\mathbb{E}_B[\ell_s - \ell_B] = -b^2(1 - r_{s,B}), \quad \mathrm{Var}_B[\ell_s - \ell_B] = 2b^2(1 - r_{s,B}),$$

with $r_{s,B}$ defined in (4.3). For any s_1 and s_2, the covariance between two local field terms is given by

$$\mathrm{Cov}_B(\ell_{s_1} - \ell_B, \ell_{s_2} - \ell_B) = b^2(1 + r_{s_1,s_2} - r_{s_1,B} - r_{s_2,B}).$$

Proof. Note that $\ell_s - \ell_B = b(Z'_s - Z'_B)$, $\mathbb{E}_B[Z'_B] = b, \mathrm{Var}_B[Z'_B] = 1$. Moreover, due to the normal assumption of Z'_B, we have the following decomposition $\mathbb{E}_B[\ell_s - \ell_B] = \mathbb{E}_B[b(Z'_s - Z'_B)] = \mathbb{E}_B[b(r_{s,B}Z'_B + (1 - r^2_{s,B})^{1/2}W - Z'_B)] = -b^2(1 - r_{s,B})$, where W is a zero-mean random variable and independent of Z'_B, representing residual of regression. The variance and covariance can be found using similar decompositions. □

Remark. C.1 (Consequence of Lemma C.4). *From the expression of the covariance in (4.3), we have that for $s - B > 0$,*

$$r_{s,B} = [1 + (s-B)/B]^{-1/2}[1 + (s-B)/(B-1)\Big)]^{-1/2},$$

and for $s - B < 0$,

$$r_{s,B} = [1 + (s-B)/B]^{1/2}[1 + (s-B)/(B-1)]^{1/2}.$$

Consequently,

1. *When* $|s - B| \to \infty$, $r_{s,B} \to 0$. *Therefore, when* $|s - B| \to \infty$, $\mathbb{E}_B[\ell_s - \ell_B]$ *converges to* $-b^2$ *and* $\mathrm{Var}_B[\ell_s - \ell_B]$ *converges* $2b^2$.
2. *When* $|s - B|$ *is small, assume* $s = B + j$, $j = 0, \pm 1, \pm 2, \ldots$. *Performing the Taylor expansion of* $r_{B+j,B}$ *around 0, we have that*

$$r_{B+j,B} = 1 - \frac{1}{2}\frac{2B-1}{B(B-1)}|j| + o(|j|). \tag{C.11}$$

Define

$$\mu = b\{(2B-1)/[B(B-1)]\}^{1/2}. \tag{C.12}$$

Note that μ *depends on the threshold as well as B, the block size parameter. Using (C.11), we have*

$$\lim_{|j|\to 0} \mathbb{E}_B\big[\ell_{B+j} - \ell_B\big] = -\frac{\mu^2}{2}|j|,$$
$$\lim_{|j|\to 0} \mathrm{Var}_B\big[\ell_{B+j} - \ell_B\big] = \mu^2|j|,$$
$$\lim_{|j_1|\to 0, |j_2|\to 0} \mathrm{Cov}_B(\ell_{B+j_1} - \ell_B, \ell_{B+j_2} - \ell_B) = \mu^2(|j_1| \wedge |j_2|).$$

Therefore, when $|j|$ *is small (i.e., in the neighborhood of zero), we can approximate the local field using a two-sided Gaussian random walk with drift* $\mu^2/2$ *and the variance of the increment being* μ^2*:*

$$\ell_{B+j} - \ell_B \overset{d}{=} \mu \sum_{i=1}^{|j|} \vartheta_i - \mu^2 j/2, \quad j = \pm 1, \pm 2, \ldots \tag{C.13}$$

where ϑ_i *are i.i.d. standard normal random variables.*

C.4. Step 4: Approximation using localization theorem

The remaining work is to compute the conditional expectations $\mathbb{E}_B[\cdots; (\cdots) \geq 0]$ for each B in (C.9). In the following, we drop the subscript B in $\mathbb{E}_B$ for simplicity, and the approximation results hold for each B. We assume $b \to \infty, B_{\max} \to \infty$, and $b^2/B_{\max}$ is held to a fixed positive constant. Introduce an abstract index κ and let $\kappa = b^2$; this choice is because $\kappa^{1/2}$ is the multiplicative factor that balances the rate of convergence of the global term under the transformed measure. Typically, κ is equal to the variance of the global term $\tilde{\ell}_B = b(Z'_B - b)$, which is b^2 as shown in Lemma C.3; κ is also associated with the drift and the variance of the incremental of the local field $\{\ell_s - \ell_B\}$ for $|s - B| = 0, 1, 2, \ldots$, as shown in Lemma C.4.

Using a powerful localization theorem (see theorem 3.1 in Siegmund et al. [2010] or theorem 5.2 in Yakir [2013]), we can obtain the limit for each term in the summand of (C.9), rewritten as (by changing the index to κ)

$$\mathbb{E}\left[\frac{M_\kappa}{S_\kappa} e^{-(\tilde{\ell}_\kappa + \log M_\kappa)}; \tilde{\ell}_\kappa + \log M_\kappa \geq 0\right], \tag{C.14}$$

when $\kappa \to \infty$. Basically, the localization theorem states that (C.14) scaled by $\kappa^{\frac{1}{2}}$ converges under mild conditions when $\kappa \to \infty$.

The statement of the theorem involves a local σ-algebra denoted as $\widehat{\mathcal{F}}_\kappa$:

$$\widehat{\mathcal{F}}_\kappa = \sigma\{\ell_s - \ell_B : |s - B| \le g(\kappa)\}, \tag{C.15}$$

where a function $g(\kappa)$ specifies the size of the local region. The choice of $g(\kappa)$ is critical and it guarantees subsequent convergence. Following the analysis of scan statistics in Yakir [2013], we choose $g(\kappa) = cb^{-2}$ for some large constant c. This local σ-field is asymptotically independent of $\tilde{\ell}_\kappa$, and it carries all information needed to construct the local field.

Define $\widehat{M}_\kappa$ and $\widehat{S}_\kappa$ as the maximization and summation restricted to a smaller subset of parameter values $\{s : |s - B| \le g(\kappa)\}$, and they are measurable with respect to $\widehat{\mathcal{F}}_\kappa$. Note that $\widehat{M}_\kappa$ and $\widehat{S}_\kappa$ serve as approximations to M_κ and S_κ. In the limit, the local random field converges to a Gaussian random field, and the ratio $\mathbb{E}[\widehat{M}_\kappa/\widehat{S}_\kappa]$ converges to a limit that can be determined with the parameters of the Gaussian random field.

The localization theorem (theorem 5.1 in Siegmund et al. [2010] and section 3.4 in Yakir [2013]) consists of the five conditions as follows.

Theorem C.1. (Localization Theorem). *Given $\epsilon > 0$, if for all large κ, all fo the following conditions hold*:

I. *Both $0 < M_\kappa \le S_\kappa < \infty$ and $0 < \widehat{M}_\kappa \le \widehat{S}_\kappa < \infty$ hold in probability one.*

II. *Denote $A^c = \{|\log M_\kappa - \log \widehat{M}_\kappa| > \epsilon\} \cup \{|\widehat{S}_\kappa/S_\kappa - 1| > \epsilon\}$. For some $0 < \delta$ that does not depend on ϵ:*

$$\max_{|x| \le 3g(\kappa)} \mathbb{P}[A^c \cap \{\tilde{\ell}_\kappa + \log \widehat{M}_\kappa \in x + (0, \delta]\} \cap \{|\widehat{m}| \le g(\kappa)\}] \le \epsilon\kappa^{-1/2},$$

where $\widehat{m}_\kappa = \min\{\log \widehat{M}_\kappa, g(\kappa)\} - \log(1 - \epsilon)$.

III. *$\mathbb{E}[\widehat{M}_\kappa/\widehat{S}_\kappa]$ converges to a finite and positive limit denoted by $\mathbb{E}[M/S]$.*

IV. There exist $\mu_\kappa \in \mathbb{R}$ and $\sigma_\kappa \in \mathbb{R}^+$ such that for every $0 < \epsilon'$, δ, for any event $E \in \widehat{\mathcal{F}}_\kappa$ and for all large enough κ,

$$\sup_{|x| \le \epsilon\kappa^{1/2}} \left| \kappa^{1/2}\mathbb{P}(\tilde{\ell}_\kappa \in x + (0, \delta], E) - \frac{\delta}{\sigma}\phi\left(\frac{\mu}{\sigma}\right)\mathbb{P}(E) \right| \le \epsilon'.$$

V. *$\mathbb{P}(|\log M_\kappa| > \epsilon\kappa^{1/2}), \mathbb{P}(|\log \widehat{M}_\kappa| > \epsilon\kappa^{1/2})$ and $\mathbb{P}(\log M_\kappa - \log \widehat{M}_\kappa < -\epsilon)$ are all $o(\kappa^{-1/2})$. Then*

$$\lim_{\kappa \to \infty} \kappa^{1/2}\mathbb{E}\left[\frac{M_\kappa}{S_\kappa} e^{-[\tilde{\ell}_\kappa + \log M_\kappa]}; \tilde{\ell}_\kappa + \log M_\kappa \ge 0\right] = \sigma^{-1}\phi\left(\frac{\mu}{\sigma}\right)\mathbb{E}[M/S], \tag{C.16}$$

where $\phi(\cdot)$ is the density of the standard normal distribution.

Intuitively, the localization theorem says the following. To find the desired limit of (C.14) as $\kappa \to \infty$, one first approximates M_κ and S_κ by their localized versions, which are obtained by restricting the maximization and summation in a neighborhood of parameter values. Then one can show that the localized ratio M_κ/S_κ is asymptotically independent of the global term $\tilde{\ell}_\kappa$ as $\kappa \to \infty$. The asymptotic analysis is then performed on the local field and the global term separately. The expected value of the localized ratio $\mathbb{E}[M_\kappa/S_\kappa]$ converges to a constant independent of κ, and the limiting conditional distribution of $\tilde{\ell}_\kappa$ can be found using the local central limit theorem. Thus, one can calculate the remaining conditional expectation involving $\tilde{\ell}_\kappa$.

Checking conditions. Let us now verify the validity of the conditions in our setting. First, Condition I is met because for Gaussian random variables, $M_\kappa > 0, S_\kappa > 0$ with probability 1, and the maximization of a collection of nonnegative numbers is smaller than or equal to the summation. Similar arguments hold for their counterparts $\widehat{M}_\kappa > 0$ and $\widehat{S}_\kappa > 0$ when the maximization and summation are over a smaller set.

Condition II describes that the localized versions $\widehat{M}_\kappa$ and $\widehat{S}_\kappa$ are good approximations of M_κ and S_κ when κ is sufficiently large, for properly defined $\widehat{\mathcal{F}}_\kappa$. In section 3.4.4 of Yakir (2013), the corresponding Condition II has been rigorously checked, assuming a local region defined in the same form of our local region and assuming Gaussian random field. Thus, checking Condition II for our case will follow the same steps, using the properties established in Section 8. We omit the details here.

Condition III is checked by applying the distributional approximations to the localized version of M_κ/S_κ. We can show that the expectation of the ratio $\mathbb{E}[\widehat{M}_\kappa/\widehat{S}_\kappa]$ converges to a finite and positive limit denoted by $\mathbb{E}[M/S]$, which does not depend on κ. Because the increment $\ell_{B+j} - \ell_B$ has negative mean as shown in Lemma C.4, the values of M_κ and S_κ will be determined by values j close to 0, as is the ratio M_κ/S_κ. This implies that a relatively small local region centered on B is sufficient.

From Remark C.1, the local field when the index is close to the shifted measure parameter B can be approximated as a two-sided Gaussian random walk with drift $-\mu^2/2$ and variance μ^2 (with μ defined in (C.12)), which is denoted as $W(\mu^2 j)$ below. Therefore, we have that with high probability,

$$\mathbb{E}\left[\widehat{M}_\kappa/\widehat{S}_\kappa\right] = \mathbb{E}\left[\frac{\max_{|j|\le cb^{-2}} e^{W(\mu^2 j)}}{\sum_{|j|\le cb^{-2}} e^{W(\mu^2 j)}}\right].$$

When $c \to \infty$, it approaches to a limit known as *Mill's ratio*,

$$\mathbb{E}[M/S] = \mathbb{E}\left[\frac{\max_{|j|} e^{W(\mu^2 j)}}{\sum_{|j|} e^{W(\mu^2 j)}}\right],$$

with maximization and summation extending to the entire collection of negative and positive integers. Mill's ratio is related to the Laplace transform of the overshoot of the maxima of a Gaussian random field over a threshold b, and an expression has been obtained based on nonlinear renewal theory (see Siegmund [1985] and ch. 2.2 of (Yakir [2013]): $\mathbb{E}[M/S] = \exp\left(-2\sum_{j=1}^{\infty}\Phi(-j^{1/2}\mu/2)\right)$. An easier numerical evaluation is given by $\mathbb{E}[M/S] \approx (\mu^2/2)\nu(\mu)$ for a special function $\nu(\mu)$ defined in (4.2).

Condition IV can be checked via a local multivariate central limit theorem that is local in one component and nonlocal in others (theorem 5.3 in Yakir, [2013]). The theorem says the following: assume that ξ_i are independent and identically distributed random vector of dimension $d+1$. Assume that the mean of each vector is zero, the variance of the first component converges to a finite σ, the covariance matrix of the last d components converges a finite matrix Σ, and the correlation between these components and the first one converges to zero (hence, the overall covariance matrix is block-diagonal). Define $S_\gamma = \sum_{i=1}^{\gamma}\xi_{i,1}$ and a d-dimensional vector with element $h_{\gamma,j} = \gamma^{-1/2}\xi_{i,j}$, for $1 \le j \le d$. Then, under mild conditions,

$$\lim_{\gamma\to\infty} \gamma^{1/2}\mathbb{P}(S_\gamma \in [l,u], h_\gamma \in \mathcal{A}) = \frac{l-u}{(2\pi)^{1/2}\sigma}\mathbb{P}(h \in \mathcal{A}) \tag{C.17}$$

for any interval $[l,u]$ and an arbitrary set $\mathcal{A}$.

Our setting matches exactly to the above distribution when we set the global term as the first component and the local field as the remaining components. Using the properties in Section 8, we have shown the finite mean and variance (covariance) of the global and local field terms. We only need to show the global term, and the local fields are independent of each other

asymptotically. It suffices to prove that the conditional covariance of $\{\ell_{B+j} - \ell_B\}$ given $\tilde{\ell}_B$ converges to the unconditional covariance, and the conditional means converges to the unconditional one. With a slight abuse of notation, $r_1 = r_{B+j_1,B}$ and $r_2 = r_{B+j_2,B}$ and using the linear regression decomposition, when conditioning on Z'_B (which is proportional to $\tilde{\ell}_B$), the two local field terms are independent of each other:

$$\begin{aligned}&\mathrm{Cov}(b(Z'_{B+j_1} - Z'_B), b(Z'_{B+j_2} - Z'_B)|Z'_B)\\ &= \mathrm{Cov}(b(r_1 Z'_B + (1-r_1^2)^{1/2} W_1 - Z'_B), b(r_2 Z'_B + (1-r_2^2)^{1/2} W_2 - Z'_B)|Z'_B) = 0,\end{aligned}$$

where W_1 and W_2 are two mutually independent zero-mean random variables that represent the regression residuals (they are also independent of Z'_B).

On the other hand, using the same decomposition, we can show that without conditioning, the covariance is given by

$$\mathrm{Cov}(b(Z'_{B+j_1} - Z'_B), b(Z_{B+j_2}{}' - Z'_B)) = b^2(1-r_1)(1-r_2).$$

Hence, when $b \to \infty$, due to the property of local field in equation (C.11), for $|j_1| \le cb^{-2}, |j_2| \le cb^{-2}$, the unconditioned covariance converges to zero given (C.11), which is equal to the conditioned covariance. Similarly, we can show that the conditional means of $\{Z'_{B+j} - Z'_B\}$ conditioning on Z'_B converge to the unconditional ones.

Now we invoke the local central limit theorem. Because the density of the global term $\tilde{\ell}_B$ is approximately normal, we can calculate a desired form of the probability. From (C.10), the variance of the global term increases with b. The density of $\tilde{\ell}_B$ can be uniformly approximated by $1/(2\pi b^2)^{1/2}$ within a small region around the origin $|x| \le 3(4/ + 1 + \epsilon)\log b$ (Yakir, 2013). Such an approximation also holds for $\tilde{\ell}_B - x$ given any value x that is not too large. Furthermore, notice that $\log \widehat{M}_\kappa$ is very close to 0 and therefore is negligible; this is because $e^{\ell_s - \ell_B}$ should attain its maximal value when $|s - B|$ close to 0 as analyzed before. Let $\mu_\kappa = \mathbb{E}_B[\tilde{\ell}_\kappa/b] = 0$ and $\sigma_\kappa^2 = \mathrm{Var}_B[\tilde{\ell}_\kappa/b] = 1$. When $\kappa = b^2 \to \infty$, using local central limit theorem (C.17), we have that

$$\kappa^{1/2}\mathbb{P}\left(\tilde{\ell}_\kappa \in x - \log \widehat{M}_\kappa + (0,\delta]\right) \to \frac{\delta}{\sigma_\kappa}\phi\left(\frac{\mu_\kappa}{\sigma_\kappa}\right). \tag{C.18}$$

Condition V is checked as follows. Note that the terms inside the M_κ are likelihood ratios with unit expectation because $\mathbb{E}_B[\exp(\ell_B)] = 1$. Thus, $\exp(\ell_s - \ell_B)$ is a martingale and by a standard martingale inequality, $\mathbb{P}(\log M_\kappa > \epsilon\kappa^{1/2}) \le \exp(-\epsilon\kappa^{1/2})$. Then using a similar argument as in Siegmund et al. (2010), one can show the other two inequalities, because $\widehat{M}_\kappa$ is an approximation to M_κ.

Finally, because all conditions are met, we can now apply the localization theorem for $b \to \infty$ and put things together to obtain

$$\mathbb{E}_B\left[\frac{M_B}{S_B}e^{-\left[\tilde{\ell}_B + \log M_B\right]}; \tilde{\ell}_B + \log M_B \ge 0\right] = \frac{\mu^2}{2}\nu(\mu)\frac{1}{\sqrt{2\pi b^2}}(1 + o(1)). \tag{C.19}$$

Substitute (C.19) back in to the likelihood ratio identity (C.9), and we arrive at the approximation in Theorem 4.1.

Appendix D: Proof of Theorem 4.2

The method for approximating the ARL is related to that used to analyze the offline scan B-statistic. In addition, we need the following lemma.

Lemma D.1. (Asymptotic null distribution of T). *Under the null, when $b \to \infty$, the stopping time T defined in (3.4) is uniformly integrable and asymptotically exponentially distributed; that is,*

$$|\mathbb{P}\{T \geq m\} - \exp(-\lambda_0 m)| \to 0,$$

in the range where $m\lambda_0$ is bounded away from 0.

Proof. The proof is based on adapting arguments in Siegmund and Venkatraman (1995), Siegmund and Yakir (2008), and Yakir (2009). The main idea is to show that the number of boundary cross-events for the detection statistic over disjoint intervals converges to a Poisson random variable in the total variation norm, resulting from the Poisson limit theorem (theorem 1 in Arratia et al., 1989) for dependent samples. First, we show that the stopping time T is asymptotically exponentially distributed. The analysis of the distribution of the stopping time is based on Poisson approximation. Define an indicator of the event $\mathbf{1}_j$ such that the event $\mathbf{1}\{\max_{(j-1)m\leq t\leq jm} Z'_{B_0,t} > b\}$. Consider the time interval $[0,x]$. Note that the stopping time is not activated in the interval $[0,x]$, if and only if all of the relevant indicators are zero. For simplicity, we assume that x is divisible by m. Define the random variable $\widehat{W} = \sum_{j=1}^{x/m} \mathbf{1}_j$. Hence, $\{\widehat{W} = 0\} = \{T_b > x\}$. Thus, to characterize the tail probability of the stopping time $\mathbb{P}\{T_b > x\}$, we show that the sum of the indicator functions converges to a Poisson distribution. □

Using Lemma D.1, we know that for large m, $\mathbb{P}\{T \leq m\}$ is approximately $1 - \exp(-\lambda_0 m) \approx \lambda_0 m$, and $\mathbb{E}\{T\}$ is equal to λ_0^{-1} asymptotically when $b \to \infty$. So the remaining question is to find the probability and the corresponding λ_0. Consider $\mathbb{P}\{T \leq m\} = \mathbb{P}\{\max_{2\leq t\leq m} Z_{B_0,t}{}' > b\}$. Suppose $m > B_0$ and $\log b \ll m \ll b^{-1}e^{\frac{1}{2}b^2}$. We will adopt a similar strategy to approximate this probability using the change-of-measure technique.

Note that the covariance structures for online and offline scan B-statistics are different, so there will be different drift parameters when we invoke the localization theorem. Using exponential tilting, we introduce a likelihood ratio

$$\zeta_t = bZ'_{B_0,t} - b^2/2.$$

Again using the change-of-measure by likelihood ratio identity, we obtain

$$\mathbb{P}\{\max_{2\leq t\leq m} Z'_{B_0,t} > b\} = e^{-b^2/2} \sum_{t=2}^{m} \mathbb{E}_t \left[\frac{M'_t}{S'_t} e^{-[\tilde{\zeta}_t + \log M_t]}; \tilde{\zeta}_t + \log M'_t \geq 0 \right], \tag{D.1}$$

where

$$M'_t = \max_{2\leq s\leq m} e^{\zeta_s - \zeta_t}, \quad S'_t = \sum_{2\leq s\leq m} e^{\zeta_s - \zeta_t}, \quad \text{and} \quad \tilde{\zeta}_t = b(Z'_{B_0,t} - b).$$

Hence, one can again apply the localization theorem to find the approximation when $b \to \infty$, and the only differences are in the definition and characterization of global and local field terms.

Lemma D.2. (Local field of online scan B-statistic). *The mean, variance, and covariance of the local field $\{\zeta_s - \zeta_t\}$ are given by*

$$\mathbb{E}_t[\zeta_s - \zeta_t] = -b^2(1 - \rho_{s,t}), \quad \mathrm{Var}_t[\zeta_s - \zeta_t] = 2b^2(1 - \rho_{s,t}),$$
$$\mathrm{Cov}_t(\zeta_{s_1} - \zeta_t, \zeta_{s_2} - \zeta_t) = b^2(1 + \rho_{s_1,s_2} - \rho_{s_1,t} - \rho_{s_2,t}),$$

where

$$\rho_{s,t} = \mathrm{Cov}(Z'_{B_0,s}, Z'_{B_0,t}) = \frac{\binom{(B_0 - |t-s|)\vee 0}{2}}{\binom{B_0}{2}}. \tag{D.2}$$

The proof can be found in Appendix B.3. Note that when $|t-s|$ is close to 0, $\mathbb{E}_t[\zeta_s - \zeta_t]$ is close to 0. With an increasing $|t-s|, \mathbb{E}_t[\zeta_s - \zeta_t]$ decreases until $|t-s| > B_0$ (when there are no

overlapping test data in the sliding block) and then $\mathbb{E}_t[\zeta_s - \zeta_t]$ becomes $-b^2$. The values of M_κ and S_κ as in the localization theorem will be determined by the values of $|j|$ close to 0.

Now, again, we will use an argument based on Taylor expansion to find the drift term of the local field. When $|s-t|$ is close to 0, we can approximate $\{\zeta_s - \zeta_t\}$ as a two-sided random walk. Using Taylor expansion, we have

$$\rho_{t+j,t} = 1 - \frac{2B_0 - 1}{B_0(B_0-1)}|j| + o(|j|). \tag{D.3}$$

Let $\lambda = b[2(2B_0-1)]/[B_0(B_0-1)]^{1/2}$. Hence, we can show that the mean, variance, and covariance of the local field are approximately

$$\lim_{|j|\to 0} \mathbb{E}_t\left[\zeta_{t+j} - \zeta_t\right] = -\frac{\lambda^2}{2}|j|,$$

$$\lim_{|j|\to 0} \mathrm{Var}_t\left[\zeta_{t+j} - \zeta_t\right] = \lambda^2|j|,$$

$$\lim_{|j_1|\to 0, |j_2|\to 0} \mathrm{Cov}_t(\zeta_{t+j_1} - \zeta_t, \zeta_{t+j_2} - \zeta_t) = \lambda^2(|j_1| \wedge |j_2|).$$

As a result, by invoking the localization theorem through a similar set of steps, we obtain

$$\mathbb{P}\{T \le m\} = m \cdot \frac{be^{-\frac{1}{2}b^2}}{\sqrt{2\pi}} \frac{(2B_0-1)}{B_0(B_0-1)} \cdot \nu\left(b\sqrt{\frac{2(2B_0-1)}{B_0(B_0-1)}}\right)(1+o(1)). \tag{D.4}$$

Matching this to the above, we know that λ_0 is the factor that multiplies m, and this leads to the desired result.

For online scan B-statistics, the standard Poisson limit cannot be directly applied, because the events $\{\mathbf{1}_j\}, j = 1, \ldots, x/m,$ are not independent, and we need the generalized Poisson limit theorem (Arratia et al., 1989), which allows for dependence between the variables. The setup for the theorem is as follows. Let I be an arbitrary index set, and for $\alpha \in I,$ let X_α be a Bernoulli random variable with $p_\alpha = \mathbb{P}(X_\alpha = 1) > 0$. Let $W = \sum_{\alpha\in I} X_\alpha$. For each $\alpha \in I,$ suppose that we choose $B_\alpha \subset I$ with $\alpha \in B_\alpha$. Think of B_α as a "neighborhood of dependence" for each α, such that X_α is independent or nearly independent of all of the X_β for $\beta \notin B_\alpha$. Define $p_1 = \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} p_\alpha p_\beta, p_2 = \sum_{\alpha_l}\sum_{\alpha\neq\beta\in B_\alpha}\mathbb{E}(X_\alpha X_\beta), p_3 = \sum_{\alpha\in I}\mathbb{E}|\mathbb{E}(X_\alpha - p_\alpha|\sigma(X_\beta : \beta \in I - B_\alpha))|,$ where $\sigma(\cdot)$ represents the σ-field generated by the corresponding random field. Loosely speaking, p_1 measures the neighborhood size, p_2 measures the expected number of neighbors of a given occurrence, and p_3 measures the dependence between an event and the number of occurrences outside its neighborhood. Then, we have the following theorem.

Theorem D.1. (Poisson approximation, theorem 1 in Arratia et al., 1989). *Let W be the number of occurrences of dependent events, and let Z be a Poisson random variable with $\mathbb{E}Z = \mathbb{E}W = \lambda > 0$. Then the total variation distance between the distributions of W and Z is bounded by*

$$\sup_{||h||=1} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \le p_1 + p_2 + p_3.$$

where $h: \mathbb{Z}^+ \to \mathbb{R}, ||h|| = \sup_{k\ge 0}|h(k)|$.

The theorem is a consequence of the powerful Chen-Stein method.

Invoking the above theorem in our online scan B-statistics setting, we can bound the total variation distance between the random variable, defined as the number of boundary cross-events for the statistic over disjoint intervals and a Poisson random variable with the same rate. In our setting, let $I = \{1, 2, \ldots, x/m\}$ and $\mathcal{N}(j) = \{j-1, j, j+1\}$ where $j = 2, \ldots (x/m - 1)$ (with obvious modifications for $j = 1$ and $j = x/m$). Then we can specify

$$p_1 = \sum_{j\in I} \sum_{i\in\mathcal{N}(j)\setminus\{j\}} \mathbb{P}\{\mathbf{1}_j = 1\}\mathbb{P}\{\mathbf{1}_i = 1\} = 2(x/m - 2)\mathbb{P}\{\mathbf{1}_1 = 1\}^2 + 2\mathbb{P}\{\mathbf{1}_1 = 1\}, \tag{D.5}$$

$$p_2 = \sum_{j\in I} \sum_{i\in\mathcal{N}(j)\setminus\{j\}} \mathbb{P}\{\mathbf{1}_j = \mathbf{1}, 1_i = 1\} = 2(x/m - 1)\mathbb{P}\{\mathbf{1}_1 = 1, \mathbf{1}_2 = 1\}, \tag{D.6}$$

$$p_3 = \sum_{j\in I} \mathbb{E}\{|\mathbb{E}\{1_j|\sigma\{\mathbf{1}_i : i \notin \mathcal{N}(j)\}\} - \mathbb{E}\{\mathbf{1}_j\}|\}. \tag{D.7}$$

We will show that p_1, p_2, and p_3 converge to 0 as $b \to \infty$. For p_1, the last summand in (D.5) is associated with the two edge elements. It follows that p_1 is asymptotically to $(2C + 2)\mathbb{P}\{1_1 = 1\}$, which will converge to zero as $b \to \infty$ because $\mathbb{P}\{1_1 = 1\}$ converges to zero when m is sub-exponential; that is, $\log b \ll m \ll b^{-1}e^{\frac{1}{2}b^2}$. Next, let us examine p_2 in (D.6). Redefine the parameter sub-regions

$$S_1 = [0, m - B_0/2], \quad S_2 = [m - B_0/2, m + B_0/2], \quad S_3 = [m + B_0/2, 2m],$$

and denote Y_i, $i = 1, 2, 3$ as $\{Y_i = 1\} = \{\max_{t\in S_i} Z'_{B_0,t} > b\}$, which are the indicator functions of crossings of the threshold in the approximate sub-regions. Notice that the indicator functions Y_1 and Y_3 are independent of each other and they share the same distribution. We use the fact that unless the crossing occurs in a shared subregion, it must simultaneously occur in two disjoint subregions in order to have double crossing. As a consequence, we obtain the upper bound $\mathbf{1}_1 \cdot \mathbf{1}_2 \le Y_2 + Y_1 \cdot Y_3$, and

$$\mathbb{P}\{\mathbf{1}_1 = 1, \mathbf{1}_2 = 1\} \le \mathbb{P}\{Y_2 = 1\} + \mathbb{P}\{Y_1 = 1\}^2 \le \mathbb{P}\{Y_2 = 1\} + \mathbb{P}\{\mathbf{1}_1 = 1\}^2.$$

The probability $\mathbb{P}\{Y_2 = 1\}$ is proportional to $B_0 \cdot be^{-\frac{1}{2}b^2}$. Consequently, p_2 is asymptotically bounded by $2C(B_0/m + \mathbb{P}\{\mathbf{1}_1 = 1\})$. Hence, p_2 converges to zero if $\log b \ll m \ll b^{-1}e^{\frac{1}{2}b^2}$ whenever $b \to \infty$. For p_3 in (D.7), $\mathbf{1}_j$ and $\mathbf{1}_i$ are computed over nonoverlapping observations and are therefore independent. Thus, the term p_3 vanishes.

Next prove that the collection of stopping times $\{T_b\}$ indexed by b is *uniformly integrable*. Again consider the sequence of indicators $\{\mathbf{1}_j\}, j = 2k$ and $k = 1, 2, \ldots$. Define the random variable τ that identifies the index of the first indicator in the sequence that obtains the value one: $\tau = \inf\{k : \mathbf{1}_{2k} = 1\}$. Note that τ has a geometric distribution. Moreover, because $T_b \le 2m\tau$, we obtain that

$$\mathbb{P}\{T_b > x\} \le \mathbb{P}\{\tau > x/(2m)\} = (1 - \mathbb{P}(\mathbf{1}_2 = 1))^{\lfloor x/(2m)\rfloor}.$$

The conclusion then follows from that $1/m \cdot \mathbb{P}(\mathbf{1}_2 = 1)$ converges to 0.

Appendix E: Skewness correction

In the following, Lemma E.1, Lemma E.2, and Lemma E.3 are used to derive the final expression for the skewness of the scan B-statistic.

Lemma E.1. *Under the null hypothesis,*

$$\mathbb{E}\left[\left(\mathrm{MMD}^2(X_i, Y)\right)^3\right]$$
$$= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')h(x',x'',y',y'')h(x'',x,y'',y)] + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')^3].$$

Proof. Note that

$$\mathbb{E}\left[\left(\mathrm{MMD}^2(X_i, Y)\right)^3\right] = \binom{B}{2}^{-3} \mathbb{E}\left[\left(\sum_{a<b} h(X_{i,a}, X_{i,b}, Y_a, Y_b)\right)^3\right] = \binom{B}{2}^{-3} \sum_k C_k \mathbb{E}[h_{ab} h_{cd} h_{ef}],$$

where for simplicity we write $h_{ab} = h(X_{i,a}, X_{i,b}, Y_a, Y_b)$ and define C_k as the corresponding number of combinations under specific structure. Most of the terms in $\mathbb{E}[h_{ab}h_{cd}h_{ef}]$ vanish under the null. By enumerating all of the combinations, only two terms are nonzero: $\mathbb{E}[h_{ab}h_{bc}h_{ca}]$ and $\mathbb{E}[h_{ab}h_{ab}h_{ab}]$. Then,

$$\begin{aligned}
\mathbb{E}\left[\left(\mathrm{MMD}^2(X_i, Y)\right)^3\right] &= \binom{B}{2}^{-3}\binom{B}{2} 2(B-2)\mathbb{E}[h_{ab}h_{bc}h_{ca}] + \binom{B}{2}^{-3}\binom{B}{2}\mathbb{E}[h_{ab}h_{ab}h_{ab}] \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(X_{i,a}, X_{i,b}, Y_a, Y_b)h(X_{i,b}, X_{i,c}, Y_b, Y_c)h(X_{i,c}, X_{i,a}, Y_c, Y_a)] \\
&\quad + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(X_{i,a}, X_{i,b}, Y_a, Y_b)^3] \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x', x'', y', y'')h(x'', x, y'', y)] + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')^3].
\end{aligned}$$

□

Lemma E.2. *Under the null hypothesis,*

$$\begin{aligned}
&\mathbb{E}[\left(\mathrm{MMD}^2(X_i, Y)\right)^2 \mathrm{MMD}^2(X_j, Y)]_{i\neq j} \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x', x'', y', y'')h(x''', x'''', y'', y)] \\
&\quad + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')^2 h(x'', x''', y, y')].
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
&\mathbb{E}[\left(\mathrm{MMD}^2(X_i, Y)\right)^2 \mathrm{MMD}^2(X_j, Y)]_{i\neq j} \\
&= \binom{B}{2}^{-3} \mathbb{E}\left[\left(\sum_{a<b} h(X_{i,a}, X_{i,b}, Y_a, Y_b)\right)^2 \left(\sum_{a<b} h(X_{j,a}, X_{j,b}, Y_a, Y_b)\right)\right] \\
&= \binom{B}{2}^{-3} \sum_k C_k \mathbb{E}[h_{i,ab} h_{i,cd} h_{j,ef}],
\end{aligned}$$

where for simplicity we write $h_{i,ab} = h(X_{i,a}, X_{i,b}, Y_a, Y_b)$ and define C_k as the corresponding number of combinations under specific structure. Similarly, most of the terms in $\mathbb{E}[h_{i,ab}h_{i,cd}h_{j,ef}]$ vanish under the null. By enumerating all of the combinations, only two terms are nonzero: $\mathbb{E}[h_{i,ab}h_{i,bc}h_{j,ca}]$ and $\mathbb{E}[h_{i,ab}h_{i,ab}h_{j,ab}]$. Then,

$$
\begin{aligned}
&\mathbb{E}[\Big(\mathrm{MMD}^2(X_i, Y)\Big)^2 \mathrm{MMD}^2(X_j, Y)]_{i\neq j} \\
&= \binom{B}{2}^{-3}\binom{B}{2}2(B-2)\mathbb{E}[h_{i,ab}h_{i,bc}h_{j,ca}] + \binom{B}{2}^{-3}\binom{B}{2}\mathbb{E}[h_{i,ab}h_{i,ab}h_{j,ab}] \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(X_{i,a}, X_{i,b}, Y_a, Y_b)h(X_{i,b}, X_{i,c}, Y_b, Y_c)h(X_{j,c}, X_{j,a}, Y_c, Y_a)] \\
&\quad + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(X_{i,a}, X_{i,b}, Y_a, Y_b)^2 h(X_{j,a}, X_{j,b}, Y_a, Y_b)] \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x', x'', y', y'')h(x''', x'''', y'', y)] \\
&\quad + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')^2 h(x'', x''', y, y')].
\end{aligned}
$$

Lemma E.3. *Under the null hypothesis,*

$$
\begin{aligned}
&\mathbb{E}[\mathrm{MMD}^2(X_i, Y)\mathrm{MMD}^2(X_j, Y)\mathrm{MMD}^2(X_r, Y)]_{i\neq j\neq r} \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x'', x''\prime, y', y'')h(x'''', x''''', y'', y)] \\
&\quad + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x'', x''', y, y')h(x'''', x''''', y, y')].
\end{aligned}
$$

Proof. Note that

$$
\begin{aligned}
&\mathbb{E}[\mathrm{MMD}^2(X_i, Y)\mathrm{MMD}^2(X_j, Y)\mathrm{MMD}^2(X_r, Y)]_{i\neq j\neq r} \\
&= \binom{B}{2}^{-3}\mathbb{E}\left[\Big(\sum_{a<b}h(X_{i,a}, X_{i,b}, Y_a, Y_b)\Big)\Big(\sum_{c<d}h(X_{j,c}, X_{j,d}, Y_c, Y_d)\Big)\Big(\sum_{e<f}h(X_{r,e}, X_{r,f}, Y_e, Y_f)\Big)\right] \\
&= \binom{B}{2}^{-3}\sum_k C_k\mathbb{E}[h_{i,ab}h_{j,cd}h_{r,ef}].
\end{aligned}
$$

Similarly, most of the terms in $\mathbb{E}[h_{i,ab}h_{j,cd}h_{r,ef}]$ vanish under the null. By enumerating all of the combinations, only two terms are nonzero: $\mathbb{E}[h_{i,ab}h_{j,bc}h_{r,ca}]$ and $\mathbb{E}[h_{i,ab}h_{j,ab}h_{r,ab}]$. Then,

$$
\begin{aligned}
&\mathbb{E}[\mathrm{MMD}^2(X_i, Y)\mathrm{MMD}^2(X_j, Y)\mathrm{MMD}^2(X_r, Y)]_{i\neq j\neq r} \\
&= \binom{B}{2}^{-3}\binom{B}{2}2(B-2)\mathbb{E}[h_{i,ab}h_{j,bc}h_{r,ca}] + \binom{B}{2}^{-3}\binom{B}{2}\mathbb{E}[h_{i,ab}h_{j,ab}h_{r,ab}] \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(X_{i,a}, X_{i,b}, Y_a, Y_b)h(X_{j,b}, X_{j,c}, Y_b, Y_c)h(X_{r,c}, X_{r,a}, Y_c, Y_a)] \\
&+ \frac{4}{B^2(B-1)^2}\mathbb{E}[h(X_{i,a}, X_{i,b}, Y_a, Y_b)h(X_{j,a}, X_{j,b}, Y_a, Y_b)h(X_{r,a}, X_{r,b}, Y_a, Y_b)] \\
&= \frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x'', x''', y', y'')h(x'''', x''''', y'', y)] \\
&+ \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x, x', y, y')h(x'', x''\prime, y, y')h(x'''', x''''', y, y')].
\end{aligned}
$$

□

Using results from Lemma E.1, Lemma E.2, and Lemma E.3, we can derive the final expression for the skewness of the scan B-statistic, as summarized in Lemma 6.1.

Proof. We can write the raw third-order moment as

$$\begin{aligned}
\mathbb{E}\left[Z_B^3\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{MMD}^2(X_i, Y)\right)^3\right] \\
&= \frac{1}{N^3}\mathbb{E}\left[\left(\sum_{i=1}^{N}\mathrm{MMD}^2(X_i, Y)\right)\left(\sum_{j=1}^{N}\mathrm{MMD}^2(X_j, Y)\right)\left(\sum_{r=1}^{N}\mathrm{MMD}^2(X_r, Y)\right)\right] \\
&= \frac{1}{N^3}N\mathbb{E}\left[\left(\mathrm{MMD}^2(X_i, Y)\right)^3\right] + \frac{1}{N^3}\binom{3}{2}\binom{N}{1}\binom{N-1}{1}\mathbb{E}[\left(\mathrm{MMD}^2(X_i, Y)\right)^2\mathrm{MMD}^2(X_j, Y)]_{i\neq j} \\
&\quad + \frac{1}{N^3}\binom{N}{1}\binom{N-1}{1}\binom{N-2}{1}\mathbb{E}[\mathrm{MMD}^2(X_i, Y)\mathrm{MMD}^2(X_j, Y)\mathrm{MMD}^2(X_r, Y)]_{i\neq j\neq r} \\
&= \frac{1}{N^2}\left\{\frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')h(x',x'',y',y'')h(x'',x,y'',y)] + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')^3]\right\} \\
&\quad + \frac{3(N-1)}{N^2}\left\{\frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')h(x',x'',y',y'')h(x''',x'''',y'',y)]\right. \\
&\quad \left. + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')^2h(x'',x''',y,y')]\right\} \\
&\quad + \frac{(N-1)(N-2)}{N^2}\left\{\frac{8(B-2)}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')h(x'',x''',y',y'')h(x'''',x''''',y'',y)]\right. \\
&\quad \left. + \frac{4}{B^2(B-1)^2}\mathbb{E}[h(x,x',y,y')h(x'',x''',y,y')h(x'''',x''''',y,y')]\right\}.
\end{aligned}$$

□

Appendix F: Z_B does not converge to gaussian

Note that the third-order moment of Z_B scales as $\mathcal{O}(B^{-3})$ (due to (6.1)), but when dividing by its variance, which scales as $\mathcal{O}(B^{-2})$, the skewness becomes a constant with respect to B. Furthermore, examining the Taylor expansion of the moment generating function at $\theta = 0$, we have

$$\mathbb{E}[e^{\theta Z_B'}] = 1 + \underbrace{\mathbb{E}[Z_B']}_{0}\theta + \frac{\theta^2}{2}\underbrace{\mathbb{E}\left[(Z_B')^2\right]}_{1} + \frac{\theta^3}{6}\mathbb{E}\left[(Z_B')^3 e^{\theta Z_B'}\right] + o(\theta^3).$$

Recall that the moment generating function of a standard normal Z is given by $\mathbb{E}[e^{\theta Z}] = 1 + \theta^2/2 + o(\theta^3)$. The difference between the two moment generating functions is given by

$$\left|\mathbb{E}[e^{\theta Z_B'}] - \mathbb{E}[e^{\theta Z}]\right| = \frac{|\theta|^3}{6}|\mathbb{E}\left[(Z_B')^3 e^{\theta' Z_B'}\right]| + o(\theta^3) > \frac{|\theta|^3}{6}c|\mathbb{E}\left[(Z_B')^3\right]| + o(\theta^3), \tag{F.1}$$

where the inequality is due to the fact that $e^{\theta' Z_B'} > 0$ and we may assume that it is larger than an absolute constant c. Note that the first term on the right-hand side of (F.1) is given by $(c\theta^3/6)\mathrm{Var}[Z_B]^{-3/2}|\mathbb{E}[Z_B{}^3]|$, which is clearly bounded away from zero. Hence,

$$\left|\mathbb{E}[e^{\theta Z'_B}] - \left(1 + \frac{\theta^2}{2}\right)\right| > \frac{|\theta|^3}{6}\gamma + o(\theta^3)$$

for some constant $\gamma > 0$. This shows that the difference between the moment generating functions of Z'_B and a standard normal is always nonzero and, hence, Z'_B does not converge to a standard normal in any sense. This explains why incorporating the skewness of Z_B can improve the accuracy of the approximations for SL in Theorem 4.1 and for ARL in Theorem 4.2.

Appendix G: More details for real data experiments

G.1. CENSREC-1-C Speech Data Set

CENSREC-1-C is a real-world speech data set in the Speech Resource Consortium corpora provided by National Institute of Informatics. This data set contains two categories of data:

1. Simulated data. The simulated speech data are constructed by concatenating several utterances spoken by one speaker. Each concatenated sequence is then added with seven different levels of noise from eight different environments. So there are totally 56 different types of noise. Each noise setting contains 104 sequences from 52 males and 52 females speakers.
2. Recording data. The recording data are from two real noise environments (in university restaurant and in the vicinity of highway), and with two signal-noise ratio (SNR) settings (lower

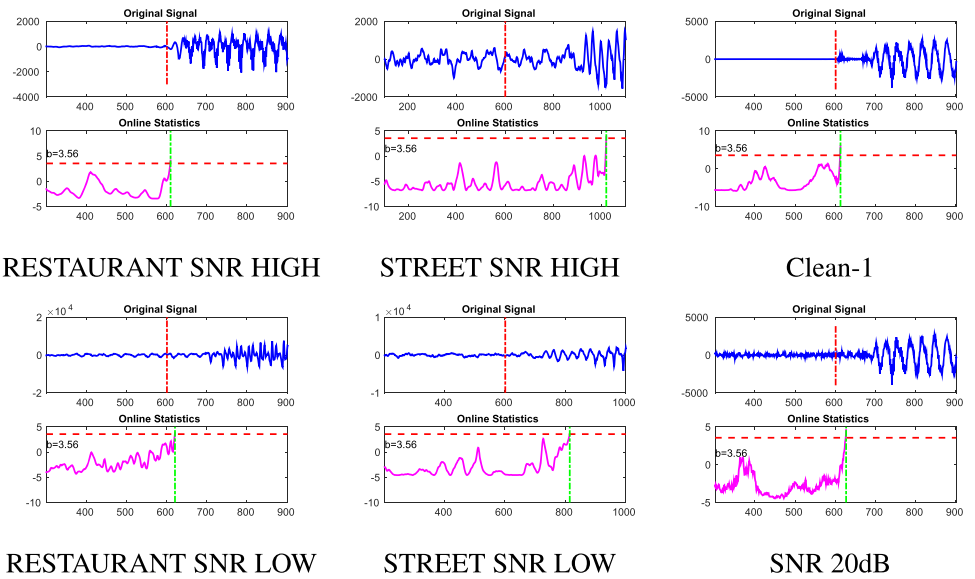


Figure G.1. Examples of speech data set. The red vertical bar shown in the upper part of each figure is the ground truth of the change point. The green vertical bar shown in the lower part is the change point detected by our algorithm (the point where the statistic exceeds the threshold). We also plot the threshold as a red dashed horizontal line in each figure. Once the statistics touch the threshold, the detection is stopped.

Table G.1. AUC results in CENSREC-1-C speech data set. Recording data are from RESTAURANT_SNR_HIGH (RH), RESTAURANT_SNR_LOW (RL), STREET_SNR_HIGH (SH), and STREET_SNR_LOW (SL).

	RH	RL	SH	SL
Ours	**0.7800**	**0.7282**	**0.6507**	**0.6865**
Baseline	0.7503	0.6835	0.4329	0.6432

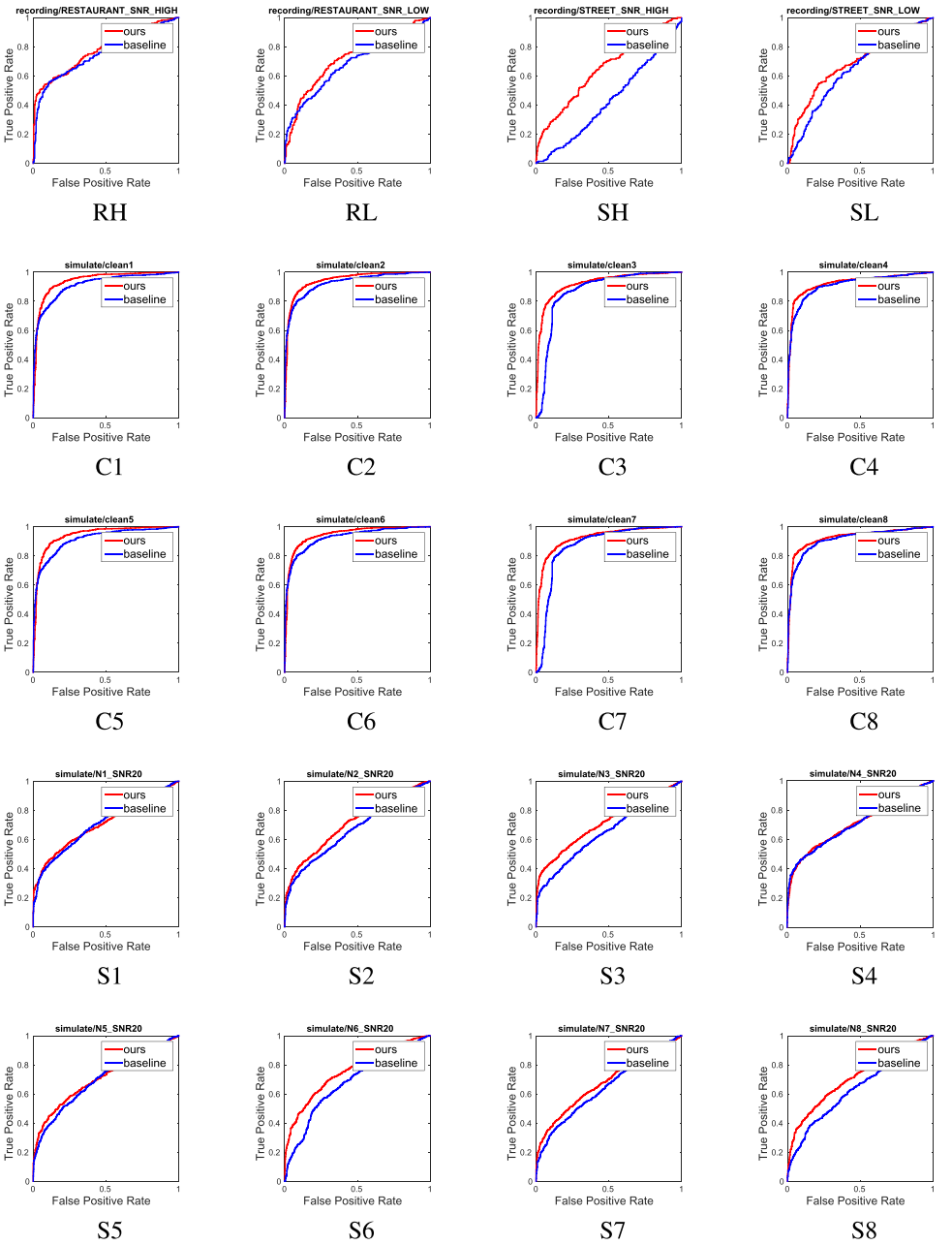


Figure G.2. Curves comparison of ROC curves for speech data set.

and higher). Ten subjects were employed for recording, and each one has four speech sequence data.

Experiment Settings. We will compare our algorithm with the baseline algorithm from Song et al. (2013). Song et al. (2013) only utilized 10 sequences from the "STREET_SNR_HIGH" setting in recording data. Here we will use all of the settings in recording data, the SNR level 20 dB, and clean signals from simulated data. See Figure G.1 for some examples of the testing data, as well as the statistics computed by our algorithm. For each sequence, we decompose it into several segments. Each segment consists of two types of signals (noise vs. speech). Given the reference data from noise, we want to detect the point where the signal changes from noise to speech.

Evaluation Metrics. We use the AUC to evaluate the computed statistics, as in Song et al. (2013). Specifically, for each test sequence that consists of two signal distributions, we will mark the points as change points whose statistics exceed the given threshold. If the distance between the detected point and true change point is within the size of detection window, then we consider it a true alarm (true positive). Otherwise, it is a false alarm (false positive).

We use 10% of the sequences to tune the parameters of both algorithms and use the remaining 90% for reporting AUC. The kernel bandwidth is tuned in $\{0.1d_{\text{med}}, 0.5d_{\text{med}}, d_{\text{med}}, 2d_{\text{med}}, 5d_{\text{med}}\}$, where d_{med} is the median of pairwise distances of reference data. The block size is fixed to be 50, and the number of blocks is simply tuned in {10, 20, 30}.

Results. Table G.1 shows the AUC of two algorithms on different background settings. Our algorithm outperforms the baseline on most cases. Both algorithms are performing quite well on the simulated clean data, because the difference between speech signals and background is more significant than the noisy ones. The averaged AUC of our algorithm on all of these settings is ***0.8014***, compared to ***0.7578*** achieved by the baseline algorithm. See the ROC curves in Figure G.2 for a complete comparison.

G.2. HASC Human activity data set

This data set is from HASC Challenge 2011. Data consist of human activity information collected by portable three-axis accelerometers. Following the setting in Song et al. (2013), we use the ℓ_2-norm of three-dimensional data (i.e., the magnitude of acceleration) as the signals.

We use the "RealWorldData" from HASC Challenge 2011, which consists of six kinds of human activities:

walk/jog, stairUp/stairDown, elevatorUp/elevatorDown,
escalatorUp/escalatorDown, movingWalkway, stay.

We make pairs of signal sequences from different activity categories and remove the sequences that are too short. We finally get 381 sequences. We tune the parameters the same way as in the CENSREC-1-C experiment. The AUC of our algorithm is ***0.8871***, compared to ***0.7161*** achieved by the baseline algorithm, which greatly improved the performance.

Examples of the signals are shown in Figure G.3. For some sequences it is easy to find the change point; see Figures G.3, and (d). Some pairs of the signals are hard to distinguish visually; see Figures G.3(b) and (c). The examples show that our algorithm can tell the change point from walk to stairUp/stairDown or from stairUp/stairDown to escalatorUp/escalatorDown. There are some cases when our algorithm raises a false alarm; see Figure G.3(h). It finds a change point during the activity "elevatorUp/elevatorDown." This is reasonable, because this type of action contains the phase from acceleration to uniform motion, and the phase from uniform motion to acceleration.

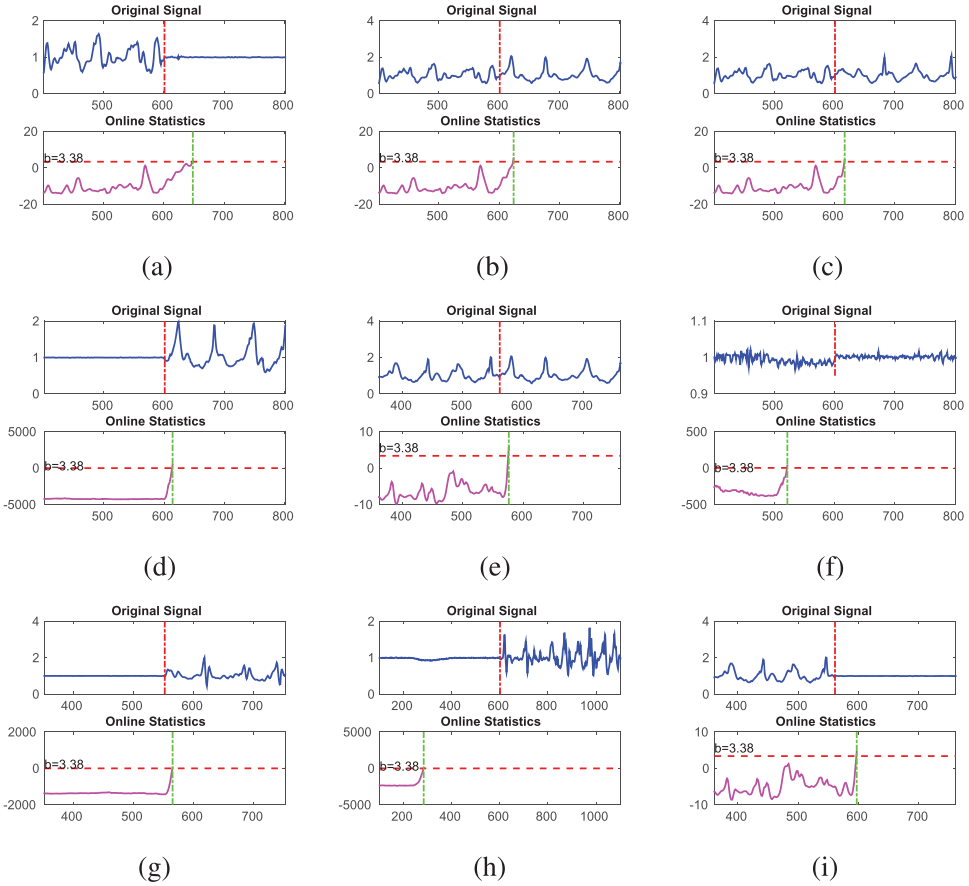


Figure G.3. Examples of the HASC data set. The meaning of the markers in this figure is the same as in Figure G.1. (a) A1 vs. A6, (b) A1 vs. A4, (c) A1 vs. A2, (d) A6 vs. A2, (e) A2 vs. A4, (f) A4 vs. A3, (g) A6 vs. A4, (h) A3 vs. A1, and (i) A2 vs. A6.

Table G.2. Simulate data with low SNR, with noise from different environments.

	C1	C2	C3	C4	C5	C6	C7	C8
Ours	**0.9413**	**0.9446**	**0.9236**	**0.9251**	**0.9413**	**0.9446**	**0.9236**	**0.9251**
Baseline	0.9138	0.9262	0.8691	0.9128	0.9138	0.9216	0.8691	0.9128

Table G.3. Simulated data with SNR = 20 dB, with noise from different environments.

	S1	S2	S3	S4	S5	S6	S7	S8
Ours	0.7048	**0.7160**	**0.7126**	**0.7129**	**0.7094**	**0.7633**	**0.6796**	**0.7145**
Baseline	**0.7083**	0.6681	0.6490	0.7119	0.6994	0.6815	0.6487	0.6541

Funding

This research was supported in part by NSF CMMI-1538746, NSF CCF-1442635, NSF CAREER CCF-1650913, DMS-1830210, a grant from the Atlanta Police Foundation, and a gift donation from Adobe Research to Yao Xie and NSF/NIH BIGDATA 1R01GM108341, ONR N00014-15-1-

2340, NSF IIS-1218749, NSF IIS-1639792, NSF CAREER IIS-1350983, and grants from Intel and NVIDIA to Le Song.

References

Arlot, S., Celisse, A., and Harchaoui, Z. (2012). Kernel Change-Point Detection, *arXiv:1202.3878*.

Arratia, R., Goldstein, L., and Gordon, L. (1989). Two Moments Suffice for Poisson Approximations: The Chen-Stein Method, *Annals of Probability* 17: 9–25. doi:10.1214/aop/1176991491

Bibinger, M., Jirak, M., and Vetter, M. (2017). Nonparametric Change-Point Analysis of Volatility, *Annals of Statistics* 45: 1542–1578. doi:10.1214/16-AOS1499

Brodsky, E. and Darkhovsky, B. S. (2013). *Nonparametric Methods in Change Point Problems*, New York: Springer.

Chwialkowski, K. and Gretton, A. (2014). A Kernel Independence Test for Random Processes, *International Conference on Machine Learning* 32: II-1422–II-1430.

Csörgő, M. and Horváth, L. (1989). Invariance Principles for Changepoint Problems, *Journal of Multivariate Analysis* 27: 151–168.

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*, New York: Wiley.

Dehling, H., Fried, R., Garcia, I., and Wendler, M. (2015). Change-Point Detection under Dependence Based on Two-Sample *U*-Statistics, in D. Dawson, R. Kulik, M. Ould Haye, B. Szyszkowicz, and Y. Zhao, Y., eds., *Asymptotic Laws and Methods in Stochastics*, vol. 76, pp. 195–220, Fields Institute Communications.

Dembo, A. and Zeitouni, O. (2010). *Large Deviations Techniques and Applications*, New York: Springer.

Desobry, F., Davy, M., and Doncarli, C. (2005). An Online Kernel Change Detection Algorithm, *IEEE Transactions on Signal Processing* 53: 2961–2974. doi:10.1109/TSP.2005.851098

Enikeeva, F., and Harchaoui, Z. (2014). High-Dimensional Change-Point Detection with Sparse Alternatives, *arXiv:1312.1900*.

Fasano, G. and Franceschini, A. (1987). A Multidimensional Version of the Kolmogorov-Smirnov Test, *Monthly Notices of Royal Astronomical Society* 225: 155–170. doi:10.1093/mnras/225.1.155

Gordon, L. and Pollak, M. (1994). An Efficient Sequential Nonparametric Scheme for Detecting a Change of Distribution, *Annals of Statistics* 22: 763–804. doi:10.1214/aos/1176325495

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test, *Journal of Machine Learning Research* 13: 723–773.

Harchaoui, Z., Bach, F., Cappe, O., and Moulines, E. (2013). Kernel-Based Methods for Hypothesis Testing: A Unified View, *IEEE Signal Processing Magazine* 30: 87–97. doi:10.1109/MSP.2013.2253631

Harchaoui, Z., Bach, F., and Moulines, E. (2008). Kernel Change-Point Analysis, *Advances in Neural Information Processing Systems* 609–616.

Harchaoui, Z. and Cappé, O. (2007). Retrospective Multiple Change-Point Estimation with Kernels, in *Proceedings of 14th IEEE Workshop on Statistical Signal Processing*, pp. 768–772.

Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting Change in Data Streams, *Proceedings of 30th International Conference on Very Large Data Bases* 30: 180–191.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown, *Journal of American Statistical Association* 62: 399–402. doi:10.1080/01621459.1967.10482916

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013). Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation, *Neural Networks* 43: 72–83. doi:10.1016/j.neunet.2013.01.012

Maragoni-Simonsen, D. and Xie, Y. (2015). Sequential Changepoint Approach for Online Community Detection, *IEEE Signal Processing Letters* 22: 1035–1039. doi:10.1109/LSP.2014.2381553

Massey, F. J., Jr. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit, *Journal of American Statistical Association* 46: 68–78. doi:10.1080/01621459.1951.10500769

Matteson, D. and James, N. (2014). A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data, *Journal of American Statistical Association* 109: 334–345. doi:10.1080/01621459.2013.849605

McCullagh, P. and Kolassa, J. (2009). Cumulants, *Scholarpedia* 4: 4699. doi:10.4249/scholarpedia.4699

Picard, D. (1985). Testing and Estimating Change-Points in Time Series, *Advances in Applied Probability* 17: 841–867. doi:10.2307/1427090

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the Decreasing Power of Kernel and Distance Based Nonparametric Hypothesis Tests in High Dimensions, in *Proceedings of Twenty-ninth AAAI Conference on Artificial Intelligence*, pp. 3571–3577.

Ross, Z. E. and Ben-Zion, Y. (2014). Automatic Picking of Direct *P*, *S* Seismic Phases and Fault Zone Head Waves, *Geophysical Journal International* 199: 368–381. doi:10.1093/gji/ggu267

Schölkopf, B. and Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge: MIT Press.

Schölkopf, B., Tsuda, K., and Vert, J. P. (2004). *Kernel Methods in Computational Biology*, Cambridge: MIT Press.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Shewhart, W. (1939). Statistical Method from the Viewpoint of Quality Control, Department of Agriculture, University of Michigan.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer.

Siegmund, D. and Venkatraman, E. S. (1995). Using the Generalized Likelihood Ratio Statistic for Sequential Detection of a Change-Point, *Annals of Statistics* 23: 255–271. doi:10.1214/aos/1176324466

Siegmund, D. and Yakir, B. (2008). Detecting the Emergence of a Signal in a Noisy Image, *Statistics and Its Interface* 1: 3–12. doi:10.4310/SII.2008.v1.n1.a1

Siegmund, D., Yakir, B., and Zhang, N. (2010). Tail Approximations for Maxima of Random Fields by Likelihood Ratio Transformations, *Sequential Analysis* 29: 245–262. doi:10.1080/07474946.2010.487428

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, Chapman and Hall/CRC.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System, *Science* 343: 80–84. doi:10.1126/science.1246981

Xie, B., Liang, Y., and Song, L. (2015). Scale Up Nonlinear Component Analysis with Doubly Stochastic Gradients, *Advances in Neural Information Processing Systems* 2341–2349.

Xie, Y. and Siegmund, D. (2013). Sequential Multi-Sensor Change-Point Detection, *Annals of Statistics* 41: 670–692. doi:10.1214/13-AOS1094

Yakir, B. (2009). Multi-Channel Change-Point Detection Statistic with Applications in DNA Copy-Number Variation and Sequential Monitoring, in *Proceedings of Second International Workshop in Sequential Methodologies*, pp. 15–17.

Yakir, B. (2013). *Extremes in Random Fields: A Theory and Its Applications*, New York: Wiley.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). *B*-Test: A Non-parametric, Low Variance Kernel Two-Sample Test, *Advances in Neural Information Processing Systems* 755–763.

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric Maximum Likelihood Approach to Multiple Change-Point Problems, *Annals of Statistics* 42: 970–1002. doi:10.1214/14-AOS1210

Zou, S., Liang, Y., Poor, H. V., and Shi, X. (2017). Nonparametric Detection of Anomalous Data Streams, *IEEE Transactions on Signal Processing* 65: 5785–5797. doi:10.1109/TSP.2017.2733472