Bernoulli **28**(1), 2022, 525–547
https://doi.org/10.3150/21-BEJ1355

Testing and estimation for clustered signals

HONGYUAN CAO[1,2] and WEI BIAO WU[3]

[1]*School of Mathematics, Jilin University, 2699 Qianjing Street, Changchun, 130012, China*
[2]*Department of Statistics, Florida State University, 117 N. Woodward Avenue, Tallahassee, FL, 32306, USA. E-mail: hongyuancao@gmail.com*
[3]*Department of Statistics, University of Chicago, 5747 South Ellis Avenue, Chicago, IL, 60637, USA. E-mail: wbwuchicago@gmail.com*

We propose a change-point detection method for large scale multiple testing problems with data having clustered signals. Unlike the classic change-point setup, the signals can vary in size within a cluster. The clustering structure on the signals enables us to effectively delineate the boundaries between signal and non-signal segments. New test statistics are proposed for observations from one and/or multiple realizations. Their asymptotic distributions are derived. We also study the associated variance estimation problem. We allow the variances to be heteroscedastic in the multiple realization case, which substantially expands the applicability of the proposed method. Simulation studies demonstrate that the proposed approach has a favorable performance. Our procedure is applied to an array based Comparative Genomic Hybridization (aCGH) dataset.

Keywords: Change-point inference; clustered signal; high dimension; multiple testing; signal aggregation; variance estimation

1. Introduction

Signal detection and multiple testing in a data rich environment have been important research topics in natural and social sciences. Typical examples include detecting anomalous traffic in computer networks [32], identifying voxels that correlate with certain activities [16] in functional Magnetic Resonance Imaging (fMRI), and associating single nucleotide polymorphisms (SNPs) with clinical outcomes [27]. The predominant framework in these research is via individual analysis–testing each hypothesis separately and declaring statistical significance if the p-value is less than certain threshold [2] or the two-sample t-statistic falls into the rejection region [6,13]. Various approaches were proposed to improve the power by incorporating structured or prior information. For example, [4] studied group hypothesis testing; [14,17] investigated p-value weighting; [7] considered p-value aggregating; [11, 19] utilized prior experimental information on each hypothesis in the inference stage with data from a new experiment; and [21,22] harnessed the sparsity of mean vectors with student's t-statistics.

For data with signals having clustered structure, multiple testing approaches currently in use fall into two general classes. The first approach defines possible regions of interest in advance, either by field knowledge or an independent experiment. [31] proposed a spatial testing procedure with pre-specified regions of interest in a compound theoretical framework; [16] developed an algorithm specifically tailored for brain imaging data where a preliminary scan is used to select clusters by grouping highly correlated voxels; [24] used the supreme statistic in a random field to construct confidence envelopes for the proportion of false discoveries and [1] used a two-stage hierarchical testing procedure to test predefined clusters first followed by a trimming stage to clean locations in which the signal is absent. The second approach is to adaptively identify a collection of differentially behaved regions with proven false discovery rate control. For example, [28] mapped the data in the wavelet domain first and removed redundant hypotheses to reduce the number of hypotheses tested and improve power; [39] studied multiple testing via false discovery rate control for large scale imaging data; [29] treated each cluster

as a testing unit and defined the false discovery as the clusters that are falsely declared among all declared clusters under the assumption that the number of false discoveries is approximately Poisson. The Poisson approximation requires the sparsity assumption on the signals. [10] gave a summary of literature in this area and developed new tools for spatial multiple testing. In this line of research, a cluster is defined to be a true discovery if it has non-zero overlap with the support of the signal. Methods that try to incorporate cluster size to improve power were also explored in [10].

In this paper, we study multiple testing problems for data with clustered signals. We propose a new test statistic that adaptively recognizes such clusters. Our test statistic aggregates information along a sliding window to boost signal noise ratio. At the boundary between signal and non-signal segments, the test statistic can be much larger than it is within the non-signal cluster. We investigate the asymptotic distribution of the proposed test statistic and set up rejection criterion controlling certain type I errors. A new algorithm is proposed to locate signal clusters for followup studies. We do not require signals to be sparse, which may be especially valuable given the current conjecture of polygenic effects on complex disease [40]. Furthermore, we allow signals to vary within a cluster, which differs from the popular assumption that the means are identical within the same cluster [37,38]. Numerical studies show that when signals have varied sizes, the proposed method has increased detection accuracy compared to method that assumes same signal size within a cluster [35–37]. Computationally, the speed of our algorithm is linear with number of tests while the algorithm used in [37] is quadratic. Unlike [7], we present a new approach for variance estimation under the setup of multiple testing. This is accomplished through the order statistics of the average squares of the original data across a sliding window, which is consistent under certain regularity conditions for the one realization case. In addition, we consider the multiple realization scenario and allow the variance to be heteroscedastic. New test statistics are proposed with unknown parameters consistently estimated with available data to conduct statistical inference. Moreover, the newly proposed algorithms are more accurate in detecting break-points than algorithms proposed in [7] as an additional turning parameter δ is used in the maximization to locate break-points. Numerical studies show improved detection precision compared to methods that did not utilize the clustering structure [13].

Recent multiple testing procedures that incorporate covariates require estimation of the prior probability that the ith test corresponds to a null, $i = 1, \ldots, m$. These weights are then estimated adaptively from available data. In particular, [33] uses an empirical Bayesian two group mixture model and proposes to minimize a penalized likelihood function where fused lasso type of penalty is used to have spatial smoothing [34]. OrderShapeEM proposed by [5] imposes a monotone increasing constraint on the prior probability of being null and a monotone decreasing constraint on the density function of p-value under alternative distribution. The implementation is achieved through combination of EM algorithm and pool-adjacent-violator-algorithm (PAVA). AdaPT [18] requires an order of the p-value to incorporate external information to boost power. SABHA [20] modifies the BH procedure by incorporating the probability that the ith test corresponds to a null, $i = 1, \ldots, m$. SABHA further suggests different ways to estimate such probabilities, including ordering, grouping, and low total variation. AdaPT and SABHA achieves finite sample control of FDR.

In our work, we impose block signal structure to improve power. Different from covariate adjusted multiple testing, we do not use covariate for individual test, instead, we treat clustered signals through aggregation of individual p-values. We do not require external covariate, such as ordering. Our results are asymptotic in terms of number of test m.

An important method for spatial cluster detection is based on scan statistics [23,32]. Scan statistic is defined as the maximum number of points in a fixed window as the window is shifted across the domain. The p-value is computed under the uniform distribution on the domain and the threshold is designed to control the familywise type I error. This statistic is used for an omnibus test of the null hypothesis that there is no clustering. If the test rejects the null hypothesis, then it leaves open the

question of where and how much clustering exists. Our test statistic is devised to compare the observed information with its expected value under the null hypothesis that there is no signal and then take the maximum across the domain. If the omnibus test detects signals, our proposed algorithm can locate such signals which is of special interest for followup studies.

The rest of the paper is organized as follows. In Section 2, we introduce a structured hypothesis testing problem with one realization. Section 3 studies the case that there are multiple realizations. In Section 4, we examine the performance of the proposed procedure via simulation; we see that our procedure is better able to detect clustered signals and the variance estimate has a good performance. Section 5 presents an application of the methodology to an array based Comparative Genomic Hybridization (aCGH) dataset.

2. Test and estimation with one realization

In this section we shall first present a structured hypothesis testing problem with locally clustered signals. Suppose we are given noisy data of the form

$$X_j = \mu_j + Z_j, \quad 1 \le j \le p, \tag{1}$$

where Z_j are i.i.d. with mean 0 and variance σ^2, and μ_j are means or signals. We say that a signal is present at location j if $\mu_j \neq 0$. In the study of aCGH data, we let X_j be the $\log_2$ ratio between the test and the reference sample intensities at locus j. Then $X_j > 0$ (resp. $X_j < 0$) means copy number duplication (resp. deletion). In this section we assume that one realization $(X_j)_{j=1}^p$ is available. In Section 3 we shall deal with the situation that multiple realizations are available with possibly non-i.i.d. Z_j. Based on the observation $(X_j)_{j=1}^p$, we test the null hypothesis of no signal

$$H_0 : \mu_1 = \cdots = \mu_p = 0 \tag{2}$$

versus the alternative hypothesis that signals are clustered: there exist *break-points* $1 = \tau_0 \le \tau_1 < \cdots < \tau_l \le \tau_{l+1} = p$ such that

$$\begin{aligned} H_1 : \mu_1 = \cdots = \mu_{\tau_1 - 1} = 0, &\qquad \mu_{\tau_1}, \ldots, \mu_{\tau_2 - 1} \neq 0, \\ \mu_{\tau_2} = \cdots = \mu_{\tau_3 - 1} = 0, &\qquad \mu_{\tau_3}, \ldots, \mu_{\tau_4 - 1} \neq 0, \ldots. \end{aligned} \tag{3}$$

Let $\mathcal{S}_f = \{\tau_f, \ldots, \tau_{f+1} - 1\}$, $f = 1, 2, \ldots$. We call sets $\mathcal{S}_1, \mathcal{S}_3, \ldots$, *signal clusters* on which μ_js are non-zero and let $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_3 \cup \ldots$ be the signal set. Let $\mathcal{N} = \mathcal{S}_0 \cup \mathcal{S}_2 \cup \ldots$ be the non-signal set. Note that our definition of break-points is different from change-points that are used in change-point analysis, where the alternative hypothesis is typically formulated as

$$H_c : \mu_1 = \cdots = \mu_{\tau_1 - 1} \neq \mu_{\tau_1} = \cdots = \mu_{\tau_2 - 1} \neq \mu_{\tau_2} = \cdots = \mu_{\tau_3 - 1} \neq \cdots.$$

For example, if there exists a j in the signal cluster $\mathcal{S}_1 = \{\tau_1, \ldots, \tau_2 - 1\}$ of (3) such that $\mu_{\tau_1} = \cdots = \mu_j \neq \mu_{j+1} = \cdots = \mu_{\tau_2 - 1}$, then this j is a change-point while it is not a break-point in our sense. While providing a very general framework, our setting of allowing unequal μ_js in the signal clusters substantially complicates the related statistical inference. The primary goal of the paper is to test H_0 vs H_1 and to locate those break-points.

2.1. One-sided Test

If in the signal sets $\mathcal{S}_1, \mathcal{S}_3, \ldots$, all nonzero μ_i are positive, namely

$$\begin{aligned} H_1' : \mu_1 = \cdots = \mu_{\tau_1 - 1} = 0, \qquad & \mu_{\tau_1}, \ldots, \mu_{\tau_2 - 1} > 0, \\ \mu_{\tau_2} = \cdots = \mu_{\tau_3 - 1} = 0, \qquad & \mu_{\tau_3}, \ldots, \mu_{\tau_4 - 1} > 0, \ldots, \end{aligned} \tag{4}$$

then we can use the following test statistic

$$R_i^\circ = \frac{1}{k} \sum_{j=i+1}^{i+k} X_j, \tag{5}$$

where k is the window size parameter. Note that the mean is $E R_i^\circ = k^{-1} \sum_{j=i+1}^{i+k} \mu_j$. Intuitively, i can be classified in the signal cluster if R_i° is big. The cutoff values can be computed based on Theorem 2.1, which provides a uniform Gaussian approximation of the distribution of R_i°. Theorem 2.1 follows from Theorem 3.1 with $n = 1$. For completeness, we state it here. It asserts that under H_0, R_i° can be uniformly approximated by the Gaussian process

$$\sigma G_i^\circ = \frac{1}{k} \sum_{j=i+1}^{i+k} \sigma \eta_j, \quad \text{where } \eta_j \overset{\text{i.i.d.}}{\sim} N(0,1). \tag{6}$$

We shall quantify the closeness by the coupling distance

$$\Delta^\circ = \sqrt{k} \max_{0 \le j \le p-k} \left| R_j^\circ / \sigma - G_j^\circ \right| \tag{7}$$

and the distributional distance

$$\rho^\circ = \sup_u \left| P\left(\sqrt{k} \max_{0 \le j \le p-k} R_j^\circ / \sigma \ge u \right) - P\left(\sqrt{k} \max_{0 \le j \le p-k} G_j^\circ \ge u \right) \right|. \tag{8}$$

We first introduce a moment condition.

Condition 2.1. $Z_1, Z_2, \ldots$, are i.i.d. with mean 0 and variance σ^2, and the θth norm $\|Z_i\|_\theta := (E|Z_i|^\theta)^{1/\theta} < \infty$, where $\theta > 2$. Write $K_\theta := \|Z_i\|_\theta$.

Theorem 2.1. *Assume Condition 2.1 and $\mu_i = 0$, $1 \le i \le p$.* (i) *Let $\theta > 2$. Then there exists a possibly larger probability space on which one can define $(Z_j)_j$ and $(\eta_j)_j$ such that, for all $u > 0$ and any positive integer k,*

$$P\left[k^{1/2} \Delta^\circ \ge c_0 u\right] \le \frac{p K_\theta^\theta}{u^\theta \sigma^\theta}, \tag{9}$$

where c_0 is a constant only depending on θ. (ii) *Let $\theta > 3$. The distributional distance*

$$\rho^\circ \lesssim k^{-1/6} (\log p)^{7/6} + \left(p k^{-\theta/2}\right)^{1/(\theta+1)} (\log p)^{(3\theta-2)/(2+2\theta)} \tag{10}$$

where $a \lesssim b$ means $a = O(b)$ and the multiplicative constant in $\lesssim$ only depends on θ, σ^2 and K_θ. Namely there exists a constant $C > 0$ depending on θ, σ^2 and K_θ such that $\rho^\circ \le C(k^{-1/6}(\log p)^{7/6} + (pk^{-\theta/2})^{1/(\theta+1)} (\log p)^{(3\theta-2)/(2+2\theta)})$.

Theorem 2.1 implies that, if the window size k satisfies $p^{2/\theta} = o(k)$, then $\Delta^\circ = o_P(1)$ by letting $u = (pk)^{1/(\theta+2)}$. Under the slightly stronger condition

$$p^{2/\theta}(\log p)^{3-2/\theta} = o(k), \tag{11}$$

we have $\rho^\circ = o(1)$, suggesting that R_j° and σG_j° are uniformly close.

Let $\hat{\sigma}^2$ be an estimate of σ^2 and $g_{1-\alpha}$ be the $(1-\alpha)$th quantile of $\max_{0\le j\le p-k} G_j^\circ$, $\alpha \in (0,1)$. The latter can be computed by Monte Carlo simulations. In Section 2.4.1, we shall propose a consistent estimate of σ^2 when $(\mu_j)_j$ has form (3). Theorem 2.1 suggests rejecting H_0 and accepting the alternative hypothesis H_1' of (4) at level α if $\max_{0\le j\le p-k} R_j^\circ > \hat{\sigma} g_{1-\alpha}$. Alternatively, let $T = p/k$, by Corollary A1 in [3] we can also have the Gumbel convergence

$$P\left[\frac{\max_{0\le j\le p-k}\sqrt{k}G_j^\circ}{\sqrt{2\log T}} - 1 - \frac{\log\log T - \frac{1}{2}\log(4\pi)}{4\log T} \le v\right] \to e^{-e^{-v}}, \tag{12}$$

which gives an approximate solution for $g_{1-\alpha}$ by letting $v = -\log\log(1-\alpha)^{-1}$. We do not recommend the latter since the convergence of (12) is very slow. A bootstrap calibration procedure is proposed in Section 3.2 which has better finite sample properties.

2.2. Two-sided Test

Under the general alternative H_1 of (3), the test statistic (5) is no longer applicable since the μ_js in the signal clusters can potentially cancel each other out. As a simple remedy, assuming at the outset that σ^2 is known, we define the modified version

$$R_i^\dagger = \frac{1}{k}\sum_{j=i+1}^{i+k}\left(X_j^2 - \sigma^2\right), \tag{13}$$

which, since $\epsilon_j = Z_j^2 - \sigma^2 + 2\mu_j Z_j$ has mean 0 under H_0, mimics R_i° in (5) in view of

$$X_j^2 - \sigma^2 = \mu_j^2 + \left(Z_j^2 - \sigma^2 + 2\mu_j Z_j\right) = \mu_j^2 + \epsilon_j. \tag{14}$$

Hence, a location i with a big value of $R_i^\dagger$ will likely be in signal clusters, regardless of signs of μ_j. Then we can apply the one-sided test procedure in Section 2.1. Note that the other modified version $R_i^\star := k^{-1}\sum_{j=i+1}^{i+k}(|X_j| - m_1)$, where $m_1 = E|Z_j|$, does not have the property that $E(|X_j| - m_1) = E(|\mu_j + Z_j| - |Z_j|) > 0$ for non-zero μ_j. So in general $R_i^\star$ cannot be used in the two-sided test. Assume that $E(Z_j^4) < \infty$. Similar to (7) and (8), we define

$$\Delta^\dagger = \sqrt{k}\max_{0\le j\le p-k}\left|R_j^\dagger/\kappa - G_j^\circ\right|, \quad \text{where } \kappa = \left\|Z_j^2 - \sigma^2\right\|_2 = \left[E\left(Z_j^2 - \sigma^2\right)^2\right]^{1/2}, \tag{15}$$

and the distributional distance

$$\rho^\dagger = \sup_u\left|P\left(\sqrt{k}\max_{0\le j\le p-k} R_j^\dagger/\kappa \ge u\right) - P\left(\sqrt{k}\max_{0\le j\le p-k} G_j^\dagger \ge u\right)\right|.$$

Note that under $\mu_j = 0$, we have $\mathrm{Var}(\epsilon_j) = E(Z_j^2 - \sigma^2 + 2\mu_j Z_j)^2 = \kappa$.

Corollary 2.1. *Assume Condition 2.1 hold with $\theta > 4$ and $\mu_i = 0$, $1 \le i \le p$. Then there exists a larger probability space on which one can define $(Z_j)_j$ and $(\eta_j)_j$ such that for all $u > 0$,*

$$P\big[k^{1/2}\Delta^\dagger \ge c_0 u\big] \le \frac{pK_\theta^\theta}{(\kappa u)^{\theta/2}}, \tag{16}$$

where c_0 is a constant only depending on θ, and the distributional distance

$$\rho^\dagger \lesssim k^{-1/6}(\log p)^{7/6} + \big(pk^{-\theta/4}\big)^{2/(\theta+2)}(\log p)^{(3\theta-4)/(4+2\theta)}, \tag{17}$$

where the constant in $\lesssim$ only depends on θ, κ and K_θ.

Corollary 2.1 follows from Theorem 2.1 by replacing θ in the latter by $\theta/2$ in view of (14). In comparison with (11), Corollary 2.1 requires the stronger condition $p^{4/\theta}(\log p)^{3-4/\theta} = o(k)$ to ensure that $\rho^\dagger = o(1)$.

Estimation of σ^2 and κ is discussed in Section 2.4.1. Recall that $g_{1-\alpha}$ is the $(1-\alpha)$th quantile of $\max_{0\le j\le p-k} G_j^\circ$, $\alpha \in (0,1)$. Corollary 2.1 suggests rejecting H_0 and accepting the alternative hypothesis H_1 of (4) at level α if $\max_{0\le j\le p-k} R_j^\dagger > \hat\kappa g_{1-\alpha}$.

Remark 1. Denote by Δ_k° the quantity Δ° in (7). A careful analysis of the proof of Theorem 3.1 (which implies Theorem 2.1 with $n = 1$) indicates that Theorem 2.1 is still valid with Δ_k° (resp. $R_k^\bullet := \sqrt{k}\max_{0\le j\le p-k} R_j^\circ/\sigma$ and $G_k^\bullet := \sqrt{k}\max_{0\le j\le p-k} G_j^\circ$) therein replaced by the uniform version $\max_{k\le m\le p}\Delta_m^\circ$ (resp. $\max_{k\le m\le p} R_m^\bullet$ and $\max_{k\le m\le p} G_m^\bullet$). Similarly, for the two-sided test, Corollary 2.1 also holds with the uniform version $\max_{k\le m\le p} m^{-1/2}\max_{0\le j\le p-m}\sum_{i=j+1}^{j+m}(X_i^2-\sigma^2)$. The latter quantity has an interesting connection with the adaptive Neyman's high dimensional multivariate normal mean test which has the form $\max_{1\le m\le p}(2m)^{-1/2}\sum_{i=1}^m (X_i^2-\sigma^2)$, which was considered in Section 2.1 in [12] in the setting that large values of μ concentrate on the first m dimensions and $X_i \overset{\text{i.i.d.}}{\sim} N(\mu_i, 1)$. Here m is estimated by the maximizer $\hat m = \operatorname{argmax}_{1\le m\le p}(2m)^{-1/2}\sum_{i=1}^m(X_i^2-\sigma^2)$.

2.3. An Algorithm for Locating Break-points

Once the null hypothesis is rejected, we need to locate break-points. We propose Algorithms 2.1 and 2.2 for locating break-points based on the one- and the two-sided tests, respectively. Theoretical properties of Algorithm 2.1 (resp. 2.2) are given in Theorem 2.2 (resp. 2.3).

2.3.1. *Locating break-points based on one-sided test*

We first present an algorithm based on the one-sided test.

Algorithm 2.1. Step 1. Let $L_j^\circ = R_{j-k}^\circ$, $j = k,\dots,p$. Compute $Q_j^\circ = 1(R_j^\circ > \gamma) + 1(L_j^\circ > \gamma)$ for a pre-specified cutoff value γ, $j = k,\dots,p-k$. We use a majority vote approach to smooth Q_j°. Specifically, denote $j_0 = \sum_{j=i-k}^{i+k} I\{Q_j^\circ = 0\}$, $j_1 = \sum_{j=i-k}^{i+k} I\{Q_j^\circ = 1\}$, and $j_2 = \sum_{j=i-k}^{i+k} I\{Q_j^\circ = 2\}$. Let $\tilde Q_j^\circ = \{k, \text{such that} j_k = \max_{l\in\{0,1,2\}} j_l\}$.

Step 2. Decompose $\{1,\dots,p\} = W_0 \cup W_1 \cup W_2$, where $j \in W_0$ if $\tilde Q_j^\circ = 0$, $j \in W_1$ if $\tilde Q_j^\circ = 1$ and $j \in W_2$ if $\tilde Q_j^\circ = 2$. Let $\mathcal{M}_1,\dots,\mathcal{M}_{\hat l}$ be connected components of W_1.

Step 3. Given $\delta < \gamma$, the break-points are defined as $\hat{\tau}_i = \text{argmax}_{j \in \mathcal{M}_i}\{R_j^\circ : L_j^\circ \le \delta\}$ if $\mathcal{M}_i$ is the transition region from W_0 to W_2. If $\mathcal{M}_i$ is the transition region from W_2 to W_0, $\hat{\tau}_i = \text{argmax}_{j \in \mathcal{M}_i}\{L_j^\circ : R_j^\circ \le \delta\}$.

The estimated signal sets are $\hat{\mathcal{S}}_1 = \{\hat{\tau}_1, \ldots, \hat{\tau}_2 - 1\}$, $\hat{\mathcal{S}}_3 = \{\hat{\tau}_3, \ldots, \hat{\tau}_4 - 1\}, \ldots$. The rationale behind Algorithm 2.1 is that if $\mu_j = 0$, then R_j° is close to 0; on the other hand, in the signal clusters, R_j° tends to be large. By locally averaging the data, we can reduce the variability, which has the effect of boosting the signal noise ratio. If there are many weak signals, we are able to detect them by the aggregation. On the other hand, if sporadic large values of X_j arise, they can be smoothed out through R_j° to avoid false discoveries. Therefore, we can effectively de-noise the data to achieve better inference. In the signal cluster, Q_j° is most likely to be 2 and in the non-signal cluster, Q_j° is most likely to be 0. In the boundary between signal and non-signal cluster, Q_j° is most likely to be 1. After Step 1, we get smoothed $\tilde{Q}_j^\circ$ that are in clusters of 0, 1 and 2. Step 2 focuses on the signal and non-signal cluster boundary regions, where $\tilde{Q}_j^\circ = 1$. Step 3 locates break-points. The basic idea is that without noise at the true break-points, R_j° reaches the maximum as there is no noise to dilute the summation if we are transiting from non-signal cluster to signal cluster. The constraint $L_j^\circ \le \delta$ prevents the detected break-points to be too far from the true break-points when μ_j increases in the signal cluster. Similarly, without noise, L_j° obtains the maximum if we are transiting from signal to non-signal cluster at the true break-points. The constraint $R_j^\circ \le \delta$ prevents the detected break-points to be too far from the true break-points when μ_j decreases in the signal cluster. With two thresholds $\delta < \gamma$, Algorithm 2.1 has more flexibility and produces more accurate estimates of the break-points than the procedure in [7] which only uses one threshold γ.

Our method depends on the choice of window size k and thresholds γ and δ. Theoretically speaking, the allowable range of k is specified in (11). Our simulation studies show that the proposed method is relatively robust to different choices of k. In practice, following the idea of the adaptive Neyman's high dimensional multivariate normal mean test mentioned in Remark 1, as a simple rule of thumb choice we can let $\hat{m} = \text{argmax}_{m \ge \sqrt{p}} R_m^\bullet$ and $k = \lfloor \hat{m}/2 \rfloor$. For a data-driven selection of γ and δ, we can choose $\gamma = \hat{\sigma} g_{1-\alpha}$ and $\delta = \hat{\sigma} g_{1,1-\alpha}$, where $g_{1-\alpha}$ and $g_{1,1-\alpha}$ are the $(1-\alpha)$th quantiles of $\max_{0 \le j \le p-k} G_j^\circ$ and $\max_{j \in W_1} G_j^\circ$, respectively, with α close to 0. They can be obtained by simulations. Section 2.4.1 gives an estimate $\hat{\sigma}$ of σ.

Condition 2.2. Recall $\mathcal{S}_f = \{\tau_f, \ldots, \tau_{f+1} - 1\}$ and $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_3 \cup \ldots$ is the signal set. Assume $d := \min_{i \in \mathcal{S}} \mu_i > 0$ and $2k < \min_f(\tau_{1+f} - \tau_f)$.

Condition 2.3. We say that a random variable Z is σ^2-*sub-Gaussian* if $E(\exp(uZ/\sigma)) \le \exp(u^2/2)$ for all $u \in \mathbb{R}$. Note that $N(0, \sigma^2)$ is σ^2-sub-Gaussian.

To state Theorem 2.2, we need to introduce truncated moment functions. For a random variable X with $E(X^2) < \infty$, define the truncated moment

$$\mathcal{M}_\upsilon(X) = E \min\big(|X|^\upsilon, X^2\big) < \infty, \quad \upsilon > 2. \tag{18}$$

If X has finite θth moment with $2 < \theta < \upsilon$, then $\mathcal{M}_\upsilon(X) \le E(|X|^\theta)$. Theorem 2.2(i) (resp. (ii)) concerns sub-Gaussian (resp. polynomial-tailed) noises.

Theorem 2.2. *Assume Condition 2.2 and $d/2 > \gamma > \delta$. (i) Assume Condition 2.3 holds for Z_j. Denote by $\hat{l}$ the estimated number of break points. Then*

$$1 - P\left[\hat{l} = l, \max_{j \le l} |\hat{\tau}_j - \tau_j| \le \frac{2k\delta}{d}\right] \le c_3\left(\frac{p}{k} e^{-c_1 k\gamma^2/\sigma^2} + l e^{-c_2 k\delta^2/\sigma^2}\right), \tag{19}$$

where c_1, c_2, c_3 are absolute constants. (ii) Assume Conditions 2.1 and let $\upsilon \ge \theta$. Then

$$\begin{aligned} 1 - P\left[\hat{l} = l, \max_{j \le l} |\hat{\tau}_j - \tau_j| \le \frac{2k\delta}{d}\right] &\lesssim p\mathcal{M}_\upsilon\big(Z_1/(k\gamma)\big) + \frac{p}{k} e^{-c_1 k\gamma^2/\sigma^2} \\ &\quad + lk\mathcal{M}_\upsilon\big(Z_1/(k\delta)\big) + l e^{-c_2 k\delta^2/\sigma^2} \\ &\le K_\theta^\theta\big(p\gamma^{-\theta} + lk\delta^{-\theta}\big)k^{-\theta} \\ &\quad + pk^{-1} e^{-c_1 k\gamma^2/\sigma^2} + l e^{-c_2 k\delta^2/\sigma^2}, \end{aligned} \tag{20}$$

where c_1 and c_2 are absolute constants and the constant in $\lesssim$ only depends on θ and υ.

Theorem 2.2 is proved in the Supplementary Material [8]. In comparison with (19), the extra term $K_\theta^\theta(p\gamma^{-\theta} + lk\delta^{-\theta})k^{-\theta}$ in (20) is due to polynomial tails, which are heavier than the sub-Gaussian ones. In the sub-Gaussian case (i) with unbounded l (namely $l \to \infty$), choose $\gamma = C_1(k^{-1}\log p)^{1/2}$, and $\delta = C_2(k^{-1}\log l)^{1/2}$, where C_1 and C_2 are sufficiently large constants, we have the uniform bound $\max_{j \le l} |\hat{\tau}_j - \tau_j| = O_P(d^{-1}(k\log l)^{1/2})$. The condition $d/2 > \gamma$ requires that $k \ge C_3 d^{-2}\log p$ for a sufficiently large constant C_3. When l is bounded, by letting $k = \lfloor Cd^{-2}\log p \rfloor$ for a sufficiently large C, we can similarly obtain the uniform bound $\max_{j \le l} |\hat{\tau}_j - \tau_j| = O_P(d^{-2}(\log p)^{1/2})$. In the context of detecting a deterministic signal with unknown spatial extent in the univariate sampled data model with standard white Gaussian noises, [9] dealt with the special case $\mu_j = d\mathbf{1}_{\tau_1 \le j < \tau_2}$ and considered the consistency of detection based on scan statistics under the condition $\tau_2 - \tau_1 \ge c_p d^{-2}\log p$, where $c_p = 2 + \iota_p$ and $\iota_p^2 \log p \to \infty$. The latter observation has a similar flavor as our condition $k \ge C_3 d^{-2}\log p$.

The polynomial-tailed case (20) is more involved. To ensure that the right-hand side of (20) goes to 0, we can choose $\gamma = C_1(k^{-1}p^{1/\theta} + (k^{-1}\log p)^{1/2})$ and $\delta = C_2(k^{-1}(lk)^{1/\theta} + (k^{-1}\log l)^{1/2})$, where $C_1, C_2 > 0$ are sufficiently large constants. Assume $k \ge C_3(d^{-2}\log p + d^{-1}p^{1/\theta})$ for a sufficiently large constant C_3, we have the uniform consistency $\max_{j \le l} |\hat{\tau}_j - \tau_j| \le O_P(k\delta/d)$. Thus the numbers of false discoveries and missed discoveries are bounded by $lO_P(k\delta/d)$. If l is bounded, then the latter bound becomes $O_P[(k\log p)^{1/2})/d]$.

2.3.2. *Locating break-points based on two-sided test*

We next present an algorithm based on the two-sided test. It is similar to Algorithm 2.1. With the square form (13), it can pick up signals with alternating positive and negative signs. Same simulation assisted choice of γ and δ as in the one-sided test can be used.

Algorithm 2.2. Step 1: Calculate $R_i^\dagger$ and let $L_i^\dagger = R_{i-k}^\dagger$, $i = k, \ldots, p$. For a pre-specified γ, let $Q_i^\dagger = 1(R_i^\dagger > \gamma) + 1(L_i^\dagger > \gamma)$, $i = k, \ldots, p-k$. The same majority vote approach as in Algorithm 2.1 is used to smooth $Q_i^\dagger$, denoted as $\tilde{Q}_i^\dagger$.

Step 2: Decompose $\{1, \ldots, p\} = W_0 \cup W_1 \cup W_2$, where $i \in W_0$ if $\tilde{Q}_i^\dagger = 0$, $i \in W_1$ if $\tilde{Q}_i^\dagger = 1$ and $i \in W_2$ if $\tilde{Q}_i^\dagger = 2$. Let $\mathcal{M}_1, \ldots, \mathcal{M}_{\hat{l}}$ be connected components of W_1.

Step 3. Given $\delta < \gamma$, the break-points are estimated as $\hat{\tau}_i = \text{argmax}_{j\in\mathcal{M}_i}\{R_j^\dagger : L_j^\dagger \le \delta\}$ if $\mathcal{M}_i$ is the transition region from W_0 to W_2. If $\mathcal{M}_i$ is the transition region from W_2 to W_0, $\hat{\tau}_i = \text{argmax}_{j\in\mathcal{M}_i}\{L_j^\dagger : R_j^\dagger \le \delta\}$.

Condition 2.4. Recall Condition 2.2 for $\mathcal{S}$. Let $d = \min_{i\in\mathcal{S}}|\mu_i| > 0$ and assume $2k < \min_f(\tau_{1+f} - \tau_f)$.

Theorem 2.3. *Assume Conditions 2.3, 2.4, and* $(k^{-1}\log p)^{1/2} = o(d^2)$. *Let* $\gamma = c_1(k^{-1}\log p)^{1/2}$ *and* $\delta = c_2(k^{-1}\log l)^{1/2}$, *where* c_1 *and* c_2 *are sufficiently large constants. Then there exists a constant* $c > 0$ *independent of* k *and* p *such that*

$$P\left[\hat{l} = l, \max_{j\le l}|\hat{\tau}_j - \tau_j| \le \frac{ck\delta}{d^2}\right] \to 1. \tag{21}$$

Theorem 2.3 provides a bound for uniform deviations of the estimated break-points. It is proved in the Supplementary Material, where the polynomial-tailed case is also studied. Same choice of γ and δ can be used as in the one-sided test scenario.

2.4. Variance Estimation

2.4.1. *Estimation of* σ^2

To apply Theorem 2.1 and Corollary 2.1 for computing the cutoff values based on R_j° and $R_j^\dagger$, we need to deal with the key issue of estimating the variance σ^2. Furthermore, to use $R_j^\dagger$, we need to estimate κ^2. For the nonparametric regression model $X_i = \mu_i + Z_i = f(i/p) + Z_i$, $1 \le i \le p$, where $\mu_i = f(i/p)$, f is a smooth function and Z_i are i.i.d. with mean 0 and variance σ^2, the problem of estimating σ^2 has a long history; see [15] and references therein. However, the difference-based method in the latter paper does not work here. Due to the presence of the nonzero μ_js, the problem of estimating σ^2 is highly nontrivial. The latter problem is further complicated by the fact that the nonzero μ_js in the signal segments can change wildly. Here we shall use order statistics and obtain a consistent estimator. Let

$$\hat{\sigma}_i^2 = \frac{1}{m}\sum_{j=i}^{i+m-1} X_j^2, \quad 1 \le i \le p', \text{ where } p' = p - m + 1. \tag{22}$$

Let $\hat{\sigma}_{(1)}^2 \le \hat{\sigma}_{(2)}^2 \le \cdots \le \hat{\sigma}_{(p')}^2$ be the order statistics of $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{p'}^2$. Theorem 2.4 shows that, for any $k \le p'/2$, $\hat{\sigma}_{(k)}^2$ is a consistent estimator of σ^2 under suitable conditions of m. The intuition is as follows: for large m, we expect that $\hat{\sigma}_i^2 \approx E\hat{\sigma}_i^2 = \sigma^2 + m^{-1}\sum_{j=i}^{i+m-1}\mu_j^2$. The latter uniform closeness relation will be made rigorous in the proof of Theorem 2.4, which is proved in the Supplementary Material. Under Condition 2.5 below, we expect that majority of $\sum_{j=i}^{i+m-1}\mu_j^2$ will be 0. Thus, the median or any lower quantile of $E\hat{\sigma}_i^2$ is σ^2.

In practice, we choose the sample median estimate with $k = p'/2$.

Condition 2.5. There exists a constant $c > 0$ such that the length of non-signal clusters $\tau_{i+1} - \tau_i \ge cp$ for all even i, and the total length $\sum_{i\text{ is even}}(\tau_{1+i} - \tau_i) \ge \lambda p$ with constant $\lambda > 1/2$.

Condition 2.5 implies the natural requirement that the proportion of non-signals (namely j with $\mu_j \neq 0$) is larger than $1/2$.

Theorem 2.4. *Assume* (μ_j) *satisfies* (3), *Condition* 2.5, $Z_i \in \mathcal{L}^\theta$, $\theta > 4$, $p^{2/\theta} = o(m)$ *and* $m = o(p)$. *Then we have for any* $k \leq p'/2$ *that*

$$\hat{\sigma}^2_{(k)} = \sigma^2 + O_P(\gamma_p), \quad \text{where } \gamma_p = \left(\frac{\log p}{m}\right)^{1/2} + \frac{p^{2/\theta}}{m}. \tag{23}$$

If Condition 2.3 *holds*, $\log p = o(m)$ *and* $m = o(p)$, *then* $\hat{\sigma}^2_{(k)} = \sigma^2 + O_P((m^{-1}\log p)^{1/2})$.

2.4.2. *Estimation of* κ

The estimation of κ in (15) is much more involved. The key issue is to estimate the fourth order moment $E(Z_i^4)$. Unlike (22), we cannot simply use order statistics of the moving window sample averages $\Xi_i := m^{-1}\sum_{j=i}^{i+m-1} X_j^4$, $1 \leq i \leq p-m+1$, to estimate $E(Z_i^4)$, since the median or lower quantile of $E(\Xi_i) = E(Z_i^4) + m^{-1}\sum_{j=i}^{i+m-1}(\mu_j^4 + 6\mu_j^2\sigma^2 + 4\mu_j^3 E(Z_j^3))$ is generally *not* $E(Z_i^4)$ if $E(Z_j^3) \neq 0$. $E(\Xi_i)$ can be greater or less than $E(Z_i^4)$ depending on what μ_j, $j = i, \ldots, i+m-1$ and $E(Z_j^3)$ are. The reason is that the function $E(\mu + Z_j)^4$ may not be minimized at $\mu = 0$. For example, if $Z_j = E_j - 1$ with $E_j \sim \exp(1)$, then $E(\mu + Z_j)^4$ is minimized at $\mu \approx -0.596072$. To circumvent the latter problem, we introduce

$$\hat{\nu}_i = \frac{1}{m}\sum_{j=i}^{i+m-1}(X_j - X_{j-1})^4, \quad 2 \leq i \leq p-m+1, \tag{24}$$

and $\nu = E(Z_1 - Z_0)^4 = 2\kappa^2 + 8\sigma^4$. Note that $E(\mu + Z_1 - Z_0)^4$ is indeed minimized at $\mu = 0$. The above estimate resembles the first order difference based estimate; see [15]. However, the setting and the motivation are quite different. Let $\hat{\nu}_{(2)} \leq \cdots \leq \hat{\nu}_{(p-m+1)}$ be the order statistics. Corollary 2.2 below concerns asymptotics for $\hat{\nu}_{(k)}$. It is proved in the Supplementary Material. Then we can estimate κ^2 by $\hat{\kappa}^2 = \hat{\nu}_{(k)}/2 - 4\hat{\sigma}^4_{(k)}$. In practice, we can choose $k = p'/2$, which corresponds to the sample median. By Theorems 2.4 and Corollary 2.2, we have $\hat{\kappa}^2 = \kappa^2 + O_P(\phi_{p,m})$, where $\phi_{p,m}$ is a function of p and m, given in the following corollary.

Corollary 2.2. *Assume* (3), *Condition* 2.5 *and that* $Z_i \in \mathcal{L}^q$, $q > 4$, $p^{4/q} = o(m)$ *and* $m = o(p)$. *Then we have for any* $k \leq p'/2$ *that*

$$\hat{\nu}^2_{(k)} = \nu + O_P(\phi_{p,m}), \quad \text{where } \phi_{p,m} = \left(\frac{\log p}{m}\right)^{1/2} + \frac{p^{4/q}}{m}. \tag{25}$$

If Condition 2.3 *holds*, $(\log p)^2 = o(m)$ *and* $m = o(p)$, *then* $\hat{\nu}^2_{(k)} = \nu + O_P((m^{-1}\log p)^{1/2})$.

3. Test and estimation with multiple realizations

In Section 2, only one realization $(X_j)_{j=1}^p$ is available, under the assumption that the errors Z_j are i.i.d. When we have more than one realization, we will be able to detect clustered signals even if the variances change along the sequence. The allowance of heteroscedasticity substantially expands the

application of our methods. Let $n(\geq 2)$-realizations $Y_i = (Y_{i1}, \ldots, Y_{ip})^T$ be observed, $i = 1, \ldots, n$, with

$$Y_{ij} = \mu_j + Z_{ij}, \quad 1 \leq j \leq p, \tag{26}$$

where Z_{ij} has mean 0, variance σ_j^2 and independent across both i and j. We are interested in testing (4) and (3). To this end, we propose a new test statistic and derive an omnibus test under the global null hypothesis H_0 in (2). Let $\hat{\mu}_j = n^{-1} \sum_{i=1}^n Y_{ij}$.

3.1. One-sided Test

Given a window size k, define

$$R_j^\star = \frac{\sum_{l=j+1}^{j+k} \sqrt{n}\hat{\mu}_l}{v_j^{1/2}}, \quad \text{where } v_j = \sum_{l=j+1}^{j+k} \sigma_l^2, 0 \leq j \leq p-k. \tag{27}$$

Let $(G_j^\star)_{0\leq j\leq p-k}$ be a mean zero Gaussian vector which has the same covariance structure as $(R_j^\star)_{0\leq j\leq p-k}$. As a stochastic realization, we can let

$$G_j^\star = \frac{W_j}{v_j^{1/2}}, \quad \text{where } W_j = \sum_{l=j+1}^{j+k} \sigma_l \eta_l, v_j = E\big(W_j^2\big) \text{ and } \eta_l \overset{\text{i.i.d.}}{\sim} N(0,1). \tag{28}$$

Let $\sigma = (\sigma_1, \ldots, \sigma_p)$. Then $G_j^\star$ has marginal variance 1 and covariance matrix $\Gamma(\sigma) = (\gamma_{j,j'}(\sigma))_{0\leq j,j'\leq p-k}$ with $\gamma_{j,j'}(\sigma) = v_j^{-1/2} v_{j'}^{-1/2} E(W_j W_{j'})$. Note that $\gamma_{j,j'}(\sigma) = 0$ if $|j - j'| \geq k$ and (W_j) are $(k-1)$-dependent. Let the coupled distance

$$\Delta^\star = \max_{0\leq j\leq p-k} \big|R_j^\star - G_j^\star\big|. \tag{29}$$

Theorem 3.1 below concerns the Gaussian approximation in terms of the closeness of $R_j^\star$ and $G_j^\star$ with various metrics. It is proved in the Supplementary Material. Relation (31) is a coupling statement which provides a tail probability inequality for the maximum distance $\Delta^\star$ on some common probability space, while (32) is for the distributional distance

$$\rho^\star := \sup_u \Big| P\Big(\max_{0\leq j\leq p-k} R_j^\star \geq u\Big) - P\Big(\max_{0\leq j\leq p-k} G_j^\star \geq u\Big)\Big|. \tag{30}$$

We shall impose the following regularity condition.

Condition 3.1. Let $\theta > 2$. Assume that there exist positive constants σ_*, σ^* and K_θ such that, for all $1 \leq j \leq p$, $\sigma_* \leq \sigma_j \leq \sigma^*$, and $E|Z_{ij}|^\theta \leq K_\theta^\theta$.

Theorem 3.1. *Assume Condition 3.1 and $\mu_i = 0$, $1 \leq i \leq p$.* (i) *Let $\theta > 2$. Then there exists a Gaussian process $(G_j^\star)_{0\leq j\leq p-k}$ such that on a possibly larger probability space, for all $u > 0$,*

$$P\big[(nk)^{1/2}\Delta^\star \geq c_0 u\big] \leq \frac{npK_\theta^\theta}{u^\theta \sigma_*^\theta}, \tag{31}$$

where c_0 is a constant only depending on θ. (ii) *Let $\theta > 3$. Then the distributional distance*

$$\rho^* \lesssim (nk)^{-1/6}(\log p)^{7/6} + \left(np/(nk)^{\theta/2}\right)^{1/(\theta+1)}(\log p)^{(3\theta-2)/(2+2\theta)}, \tag{32}$$

where the constant in $\lesssim$ only depends on θ, σ_*, σ^* and K_θ.

We emphasize that our theorem does not require $n \to \infty$ and it is also applicable when n is finite. For example, when $n = 2$ observations are available, if we choose the window size k be sufficiently large such that $p(\log p)^{3\theta/2-1} = o(k^{\theta/2})$, then by (32) and elementary manipulations we still have $\rho^* \to 0$. Under the slightly weaker condition $p^{2/\theta} = o(k)$, $R_j^\star$ and $G_j^\star$ are uniformly close to each other in the sense of $\max_{k\le j\le p-k} |R_j^\star - G_j^\star| = o_P(1)$ in view of (31).

3.2. Calculating cutoff values

If the variances σ_j^2 are known, given the level $0 < \alpha < 1$, we can choose the cutoff value $u = u_{1-\alpha}$ such that

$$P\Big(\max_{0\le j\le p-k} G_j^\star \ge u_{1-\alpha}\Big) = \alpha. \tag{33}$$

The above can be done by Monte Carlo simulations. Assuming that p, n, k satisfy the relation $(np)^{2/\theta}(\log p)^{3-2/\theta} = o(nk)$. Then the right-hand side of (32) goes to 0. By Theorem 3.1, the test $\max_{0\le j\le p-k} R_j > u_{1-\alpha}$ has the asymptotically correct size α.

3.2.1. *Estimation of block sum variances*

In general, however, the variances σ_j^2 are not known. Since we have multiple realizations, we can estimate them by the classical unbiased variance estimate

$$\hat\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \hat\mu_j)^2. \tag{34}$$

Correspondingly, our test statistic $R_j^\star$ in (27) now becomes

$$\hat R_j = \frac{\sum_{l=j+1}^{j+k}\sqrt{n}\hat\mu_l}{(\sum_{l=j+1}^{j+k}\hat\sigma_l^2)^{1/2}}, \quad 0 \le j \le p-k. \tag{35}$$

At first glance, if n is small, $\hat\sigma_j^2$ may deviate substantially from σ_j^2. For example, if $n = 2$, then $\hat\sigma_j^2 = (Y_{1j} - Y_{2j})^2/2$, which may be quite different from σ_j^2. This difference might suggest that replacing σ_l^2 in R_j by $\hat\sigma_l^2$ can be problematic. However, interestingly, under suitable conditions on p, k, n, $R_j^\star$ and $\hat R_j$ can still be uniformly close. This can be intuitively explained by the fact that, in R_j, it is the block sum variance $v_j = \sum_{l=j+1}^{j+k}\sigma_l^2$ that is directly involved, not just a single σ_l^2. The sum $\hat v_j = \sum_{l=j+1}^{j+k}\hat\sigma_l^2$ can still be a good estimate of v_j, despite that individually the difference $\hat\sigma_j^2 - \sigma_j^2$ can be big due to a small n. The convergence rate is given in the following Proposition 3.1. It implies that, under Condition 3.1, if $p = o(n^{\theta/2-1}k^{\theta/2})$, then $\hat v_j/v_j$ is uniformly close to 1. It is proved in the Supplementary Material.

Proposition 3.1. *Let Condition 3.1 be satisfied. If $\theta > 4$, we have*

$$P\Big(n \max_{0\le j\le p-k} |\hat{v}_j - v_j| > u\Big) \lesssim \frac{npK_\theta^\theta}{u^{\theta/2}} + \frac{p}{k}\exp\Big(-c_3\frac{u^2}{nkK_4^4}\Big), \tag{36}$$

where the constant in $\lesssim$ and $c_3 > 0$ only depend on θ. If $2 < \theta \le 4$, then

$$P\Big(n \max_{0\le j\le p-k} |\hat{v}_j - v_j| > u\Big) \lesssim \frac{npK_\theta^\theta}{u^{\theta/2}}. \tag{37}$$

Note that (36) of Proposition 3.1 implies that we have the uniform convergence rate

$$n \max_{0\le j\le p-k} |\hat{v}_j - v_j| = O_P\big((np)^{2/\theta} + (nk)^{1/2}\log p\big).$$

Under Condition 3.1, $k\sigma_*^2 \le v_j \le k(\sigma^*)^2$. Thus the term $n\max_{0\le j\le p-k}|\hat{v}_j - v_j|$ in (36) can be replaced by the ratio normalized version $nk\max_{0\le j\le p-k}|\hat{v}_j/v_j - 1|$ so that (36) is still valid with the constants in $\lesssim$ and c_3 therein depending on θ, σ_* and σ^*. By elementary calculations, if $p = o(n^{\theta/2-1}k^{\theta/2})$, the ratios $\hat{v}_j/v_j$ are uniformly close to 1 in the sense that $\max_{0\le j\le p-k}|\hat{v}_j/v_j - 1| = o_P(1)$.

3.2.2. *A bootstrap calibration procedure*

To perform the test for $H_0 : \mu_j \equiv 0$ based on $\hat{R}_j$ with σ_j^2 replaced by their estimates $\hat{\sigma}_j^2$, we need to estimate the corresponding cutoff value $u_{1-\alpha}$ based on (33). Recall that $\Gamma(\sigma) = (\gamma_{j,j'}(\sigma))_{k\le j,j'\le p-k}$ is the covariance matrix for the vector $(Z_j)_{k\le j\le p-k}$. Write $u_{1-\alpha} = q_\alpha(\sigma)$ as a function of $\sigma = (\sigma_1,\dots,\sigma_p)$. Write $\hat{\sigma} = (\hat{\sigma}_1,\dots,\hat{\sigma}_p)$ and $\hat{u}_{1-\alpha} = q_\alpha(\hat{\sigma})$ which satisfies

$$P^*\Big(\max_{0\le j\le p-k} G_j^* \ge \hat{u}_{1-\alpha}\Big) = \alpha, \tag{38}$$

where P^* is the probability measure given $Y = (Y_1,\dots,Y_n)$ and, given $\hat{\sigma}$, $(G_j^*)_{k\le j\le p-k}$ is mean 0 Gaussian vector with covariance matrix $\Gamma(\hat{\sigma})$. In particular, as (28), we can define

$$G_j^* = \frac{W_j^*}{\hat{v}_j^{1/2}}, \quad \text{where } W_j^* = \sum_{l=j+1}^{j+k} \hat{\sigma}_l\eta_l \tag{39}$$

and $\eta_l, l\in\mathbb{Z}$, are i.i.d. $N(0,1)$ random variables that are independent of $Y = (Y_1,\dots,Y_n)$. Given $\hat{\sigma}$, the cutoff value $\hat{u}_{1-\alpha}$ in (38) can also be computed by extensive simulations.

The following theorem shows the validity of the above plug-in method in the sense that the size of our test is close to α. It is proved in the Supplementary Material.

Theorem 3.2. *Let $t_1 = (\log p)^{1/2}(nk)^{-1/2}$, $t_2 = (pn(nk)^{-\theta/2}(\log p)^{-2/3})^{1/(1/3+\theta/2)}$ and $t_* = \max(t_1,t_2)$. Recall (32) for ρ^*. Let $0<\alpha<1$. Then under Condition 3.1, we have*

$$\Big|P\Big(\max_{0\le j\le p-k}\hat{R}_j \ge \hat{u}_{1-\alpha}\Big) - \alpha\Big| \lesssim \rho^* + t_*^{1/3}(\log p)^{2/3}, \tag{40}$$

where the constant in $\lesssim$ only depends on $\sigma_*, \sigma^*, \theta$ and K_θ. In particular, the right-hand side of (40) is $o(1)$ if $pn(\log p)^{3\theta/2-1} = o((nk)^{\theta/2})$.

3.3. Estimating break-points based on one-sided test

Algorithm 3.1 shows estimating break-points based on the one-sided test. It uses $R_j^\star$ assuming that σ_j, $1 \le j \le p$, are known. If not known, we shall use the estimates σ_j^2 in (34). Same simulation assisted γ and δ can be used as in the one realization one-sided test case. Theorem 3.3 provides theoretical properties of the break-point estimates.

Algorithm 3.1. Step 1. Let $L_i^\star = R_{i-k}^\star$, $i = k, \dots, p-k$ and denote $Q_i^\star = 1(R_i^\star > \gamma) + 1(L_i^\star > \gamma)$ for a pre-specified cutoff value γ. We use a majority vote approach to smooth $Q_i^\star$. Specifically, denote $j_0^\star = \sum_{i=j-k}^{j+k} I\{Q_i^\star = 0\}$, $j_1^\star = \sum_{i=j-k}^{j+k} I\{Q_i^\star = 1\}$, and $j_2^\star = \sum_{i=j-k}^{j+k} I\{Q_i^\star = 2\}$. Let $\tilde{Q}_j^\star = \{k, \text{such that} j_k^\star = \max_{l\in\{0,1,2\}} j_l^\star\}$.

Step 2. Decompose $\{1, \dots, p\} = W_0 \cup W_1 \cup W_2$, where $i \in W_0$ if $\tilde{Q}_i^\star = 0$, $i \in W_1$ if $\tilde{Q}_i^\star = 1$ and $i \in W_2$ if $\tilde{Q}_i^\star = 2$. Let $\mathcal{M}_1, \dots, \mathcal{M}_{\hat{l}}$ be connected components of W_1.

Step 3. Let $R_j^\flat = \sum_{f=j+1}^{j+k} \sqrt{n}\hat{\mu}_f/\sqrt{k}$ and $L_j^\flat = R_{j-k}^\flat$. Given $\delta < \gamma$, the break-points are defined as $\hat{\tau}_i = \text{argmax}_{j\in\mathcal{M}_i}\{R_j^\flat : L_j^\star \le \delta\}$ if $\mathcal{M}_i$ is the transition region from W_0 to W_2. If $\mathcal{M}_i$ is the transition region from W_2 to W_0, $\hat{\tau}_i = \text{argmax}_{j\in\mathcal{M}_i}\{L_j^\flat : R_j^\star \le \delta\}$.

Differently from Algorithm 2.1, in Step 3 of Algorithm 3.1 we use $R_j^\flat$ instead of $R_j^\star$ in the argmax function. The reason is for technical convenience: one has monotonicity $E(R_j^\flat) < E(R_i^\flat)$ for $\tau_1 - k < j < i \le \tau_1$, which tends to make the estimated break-point closer to τ_1. In comparison $E(R_j^\star)$ is generally not monotone, since the variances σ_j^2 can be unequal.

Theorem 3.3. *Assume Conditions 2.2, 2.5, Z_{ij} are σ^2-sub-Gaussian, and $2\sigma\gamma \le d\sqrt{nk}$. Let $m = \lfloor 2k^{1/2}\delta\sigma n^{-1/2}d^{-1}\rfloor$. Then*

$$1 - P\Big[\hat{l} = l, \max_{j\le l} |\hat{\tau}_j - \tau_j| \le m\Big] \lesssim \frac{p}{k}\exp\big(-c_1\gamma^2\big) + l\exp\big(-c_2\delta^2\big), \tag{41}$$

where the constant in $\lesssim$ and $c_1, c_2 > 0$ are independent of k, d, n and p.

Theorem 3.3 is proved in the Supplementary Material. Assume that $(\log p)(\log l) = o(n^2d^4)$ and k satisfies $(d^2n)^{-1}\log p = o(k)$ and $k = o(nd^2/\log l)$. Let $\gamma = C_1(\log p)^{1/2}$, $\delta = C_2(\log l)^{1/2}$, where $C_1, C_2 > 0$ are constants. Then the right hand side of (41) can be arbitrarily small by letting C_1, C_2 sufficiently large. Theorem 3.3 implies that we can have exact recovery with probability $P[\hat{l} = l, \max_{j\le l}|\hat{\tau}_j - \tau_j| = 0] \to 1$ since $k^{1/2}\delta n^{-1/2}d^{-1} \to 0$.

3.4. Two-sided test: A U-statistic approach

In the one-realization case, we use (13) to test the two-sided alternative H_1 of (3). If multiple realizations $Y_i = (Y_{i1}, \dots, Y_{ip})^T$, $1 \le i \le n$, are available, we can use the U-statistic

$$W_j = \frac{2}{n(n-1)} \sum_{1\le i<i'\le n} Y_{ij}Y_{i'j}, \tag{42}$$

which is an unbiased estimate of μ_j^2. This is different from (13) in that X_j^2 in the latter is not an unbiased estimate of μ_j^2. As an important consequence, we remark that unlike the two-sided test in Section 2.2, here we do not need to use κ of form (15). Under H_0, the variance of W_j is $2\sigma_j^4/(n(n-1))$. Define

$$R_{j,4} = \big((n^2-n)/2\big)^{1/2} \frac{W_{j+1}+\cdots+W_{j+k}}{(\sigma_{j+1}^4+\cdots+\sigma_{j+k}^4)^{1/2}}. \tag{43}$$

Let $\eta_i, i \in \mathbb{Z}$, be i.i.d. $N(0,1)$. Define the Gaussian process

$$G_{j,4} = \frac{\sigma_{j+1}^2\eta_{j+1}+\cdots+\sigma_{j+k}^2\eta_{j+k}}{(\sigma_{j+1}^4+\cdots+\sigma_{j+k}^4)^{1/2}}. \tag{44}$$

Theorem 3.4. *Assume Condition 3.1 and $\mu_i = 0$, $1 \le i \le p$. Then the distributional distance*

$$\begin{aligned}&\sup_u \Big| P\Big(\max_{0\le j\le p-k} R_{j,4} \ge u\Big) - P\Big(\max_{0\le j\le p-k} G_{j,4} \ge u\Big)\Big| \\ &\quad \lesssim k^{-1/6}(\log p)^{7/6} + \big(pk^{-\theta/2}\big)^{1/(\theta+1)}(\log p)^{(3\theta-2)/(2+2\theta)}.\end{aligned} \tag{45}$$

In $R_{j,4}$, the quantity σ_j^4 is typically unknown. Here we shall propose an unbiased estimate. Note that the natural estimate $(\hat\sigma_j^2)^2$ with $\hat\sigma_j^2$ given in (34) is not unbiased. Let

$$\hat\omega_j = \frac{1/4}{n(n-1)(n-2)(n-3)} \sum (Y_{ij}-Y_{i'j})^2(Y_{hj}-Y_{h'j})^2, \tag{46}$$

where the sum is over mutually different indexes $i, i', h, h' \in \{1,\dots,n\}$. Clearly $E(\hat\omega_j) = \sigma_j^4$. Similar to (35), consider the realized version

$$R_{j,4}^* = \big((n^2-n)/2\big)^{1/2} \frac{W_{j+1}+\cdots+W_{j+k}}{(\hat\omega_{j+1}+\cdots+\hat\omega_{j+k})^{1/2}}. \tag{47}$$

To test H_0 vs H_1 in (3), we reject H_0 at level $\alpha \in (0,1)$ if $\max_{0\le j\le p-k} R_{j,4}^* \ge q_{1-\alpha}$ for some cutoff value $q_{1-\alpha}$. As in (38), $q_{1-\alpha}$ can be approximated by $\hat q_{1-\alpha}$, which satisfies $P^*(\max_{0\le j\le p-k} G_{j,4}^* \ge \hat q_{1-\alpha}) = \alpha$, where

$$G_{j,4}^* = \frac{\hat\omega_{j+1}^{1/2}\eta_{j+1}+\cdots+\hat\omega_{j+k}^{1/2}\eta_{j+k}}{(\hat\omega_{j+1}+\cdots+\hat\omega_{j+k})^{1/2}} \tag{48}$$

a Gaussian process conditioning on $(Y_1,\dots,Y_n)$. As a slightly modified version, noting that for i.i.d. $N(0,1)$ random variables $Z_1,\dots Z_n$, the U-statistic $2\sum_{1\le i<i'\le n} Z_iZ_{i'} = (n^2-n)\bar Z_n^2 - \sum_{i=1}^n (Z_i - \bar Z_n)^2$ is identically distributed as $\zeta = (n-1)\chi_1^2 - \chi_{n-1}^2$, where the χ^2 random variables χ_1^2 and χ_{n-1}^2 are independent, we can use

$$G_{j,4}^\diamond = \frac{\hat\omega_{j+1}^{1/2}\zeta_{j+1}+\cdots+\hat\omega_{j+k}^{1/2}\zeta_{j+k}}{(\hat\omega_{j+1}+\cdots+\hat\omega_{j+k})^{1/2}}, \tag{49}$$

where ζ_i are independent and identically distributed as $\zeta/\sqrt{2n(n-1)}$. If n is big, $G_{j,4}^\diamond$ gives a better approximation.

In the definition of $\hat{\omega}_j$ in (46), it involves a 4-fold summation with $O(n^4)$ computation complexity. Interestingly, we can have the following expression which allows computing $\hat{\omega}_j$ within only $O(n)$ steps: elementary but tedious calculations show

$$\hat{\omega}_j = \frac{(n-1)(4S_{j,3}S_{j,1} - nS_{j,4} - 3S_{j,2}^2) + (nS_{j,2} - S_{j,1}^2)^2}{n(n-1)(n-2)(n-3)}, \quad \text{where } S_{j,l} = \sum_{i=1}^{n} Y_{ij}^l. \tag{50}$$

To compute W_j in (42), we use the well-known formula $W_j = S_{j,1}^2 - S_{j,2}/(n(n-1))$.

3.4.1. *Estimating break-points based on two-sided test*

Similar to Algorithm 2.2, we can adjust Algorithm 3.1 for locating break-points based on two-sided test. Theorem 3.5 shows theoretical properties of Algorithm 3.2 and it is proved in the Supplementary Material.

Algorithm 3.2. Step 1. Let $Q_{j,4} = 1(R_{j,4} > \gamma) + 1(L_{j,4} > \gamma)$ for a pre-specified cutoff value γ, $j = k, \ldots, p-k$. We use a majority vote approach to smooth $Q_{j,4}$. Specifically, denote $l_0 = \sum_{i=j-k}^{j+k} I\{Q_{j,4} = 0\}$, $l_1 = \sum_{i=j-k}^{j+k} I\{Q_{j,4} = 1\}$, and $l_2 = \sum_{i=j-k}^{j+k} I\{Q_{j,4} = 2\}$. Let $\tilde{Q}_{j,4} = \{k, \text{such that} l_k = \max_{j\in\{0,1,2\}} l_j\}$.

Step 2. Decompose $\{1, \ldots, p\} = W_0 \cup W_1 \cup W_2$, where $j \in W_0$ if $\tilde{Q}_{j,4} = 0$, $j \in W_1$ if $\tilde{Q}_{j,4} = 1$ and $j \in W_2$ if $\tilde{Q}_{j,4} = 2$. Let $\mathcal{M}_1, \ldots, \mathcal{M}_{\hat{l}}$ be connected components of W_1.

Step 3. Let $R_j^{\ddagger} = ((n^2-n)/2)^{1/2} \sum_{h=1}^{k} W_{j+h}/k^{1/2}$ and $L_j^{\ddagger} = R_{j-k}^{\ddagger}$. Given δ, the break-points are defined as $\hat{\tau}_f = \text{argmax}_{j\in\mathcal{M}_f}\{R_j^{\ddagger} : L_{j,4} \le \delta\}$ if $\mathcal{M}_f$ is the transition region from W_0 to W_2. If $\mathcal{M}_f$ is the transition region from W_2 to W_0, $\hat{\tau}_f = \text{argmax}_{j\in\mathcal{M}_f}\{L_j^{\ddagger} : R_{j,4} \le \delta\}$.

Theorem 3.5. *Assume Condition 2.4 and Z_{ij} are σ^2-sub-Gaussian. Let $\gamma = c_1(\log p)^{1/2}$, $\delta = c_2(\log l)^{1/2}$, where $c_1, c_2 > 0$ are sufficiently large constants. Assume that $(\log p)^{1/2} = o(d^2 n\sqrt{k})$. Then there exists a constant $c > 0$ independent of n, k and p such that*

$$P\left[\hat{l} = l, \max_{j\le l} |\hat{\tau}_j - \tau_j| \le \frac{ck^{1/2}(\log l)^{1/2}}{nd^2}\right] \to 1. \tag{51}$$

4. Numerical studies

In this section, we present simulation studies, assess the finite sample performance of the proposed methods and compare them with competing methods [2,13,37]. We look at one- and two-sided tests with one realization in Section 4.1 and Section 4.2, respectively. One- and two-sided tests with more realizations are presented in the Supplementary Material.

4.1. Simulation study 1

Consider the model $X_i = \mu_i + Z_i$, $1 \le i \le p$. The number of tests are $p = 600, 2000$ and 6000. There are 2 break-points $\tau_1 = 1 + 0.4p$, $\tau_2 = 0.6p$, 1 signal cluster $[\tau_1, \tau_2]$ and the configuration is displayed in Table 1 and Figure 3. We compare it with the change point detection for epidemic alternative proposed in [37]. We simulate data with three different error terms: standard normal distribution,

Table 1. Signal configuration for the one-sided test. seq 1: the linear sequence from 0.4 to 1.6; seq 2: the linear sequence from 1.6 to 0.4. Segment means percentage of the sequence.

Segment	40	10	10	40
Signal	0	seq 1	seq 2	0

rescaled student t distribution with 6 degree of freedom ($t(6)/1.5^{0.5}$) and rescaled Laplace distribution ($\mathrm{LP}(0,1)/2^{0.5}$) so that their variances are all 1.

The sliding window length $k = \lfloor p^{1/2} \rfloor$ is used in the calculation of $R_i^\circ = k^{-1} \sum_{j=i+1}^{i+k} X_j$. We also show results for other choices of k. In order to estimate the variance σ^2, we choose the tuning parameter $m = k$. Let $p' = p - m + 1$, $\hat{\sigma}_i^2 = m^{-1} \sum_{j=i}^{i+m-1} X_i^2$, $1 \le i \le p'$. Theoretically speaking, any statistics $\hat{\sigma}_{(j)}^2$ with $j \le p'/2$ are consistent and we use $\hat{\sigma}_{(\lfloor p'/2 \rfloor)}^2$ as the estimate.

We implemented algorithm 2.1. Thresholding values γ and δ are chosen as 0.95th quantile of $\hat{\sigma} \max_{0 \le j \le p-k} G_j^\circ$ and and $\hat{\sigma} \max_{j \in W_1} G_j^\circ$, respectively, where $G_j^\circ = \sum_{i=j+1}^{j+k} \eta_i / k$, $\eta_i, i \in \mathbb{Z}$, are i.i.d. $N(0,1)$ and W_1 are the major connected components which include indices j such that $Q_j^\circ = 1(R_j^\circ > \gamma) + 1(L_j^\circ > \gamma) = 1$.

In implementing [37], we use L_1, the likelihood ratio statistic as an example for illustration. Similar results can be obtained for other test statistics. Specifically,

$$L_1 = \max_{1 \le i < j \le p} \left\{ \sum_{k=i+1}^{j} X_k - \frac{j-i}{p} \sum_{k=1}^{p} X_k - \frac{1}{2} \delta_0 (j-i) \right\}, \tag{52}$$

where δ_0 is the signal magnitude, which is assumed to be the same within a cluster in [37]. In our setup, we take $\delta_0 = 1$, which is the average of signal magnitude within the cluster $[\tau_1, \tau_2]$. We identify the region $[\hat{I}, \hat{J}]$ as the epidemic alternative, where $\sum_{k=\hat{I}+1}^{\hat{J}} X_k - p^{-1}(\hat{J} - \hat{I}) \sum_{k=1}^{p} X_k - \frac{1}{2}\delta_0(\hat{J} - \hat{I})$ is the obtained maximum value in (52). Note that the computational speed is quadratic with number of tests p. Our evaluation criterion is the combined error rate (CER), which is the expected value of the ratio of the number of falsely rejected hypotheses and falsely accepted hypotheses over total number of tests, the estimated number of break points $\hat{l}$ and the average difference between the estimated break points and true break points. For the proposed method, we also look at false discovery rate (FDR), which is the expected value of the ratio of false rejections over total rejections and the power, which is the expected value of the the ratio of true rejections over total number of non-nulls.

Table 2 summarizes results based on 10^3 replications. We can see that across different error distributions, the variance estimate $\hat{\sigma}^2$ has a decent performance and, as expected from our asymptotic theory, it is close to the true ones. The proposed method has smaller CER compared to method based

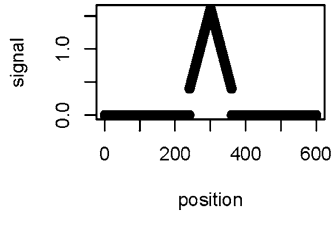

Figure 1. Signal configuration when $p = 600$.

Table 2. Summary statistics for one-sided test with 1000 simulations. $N(0,1)$: standard normal; $t(6)/1.5^{0.5}$: rescaled student t distribution with df 6; $\mathrm{LP}(0,1)/2^{0.5}$: rescaled Laplace distribution; k is the window size; CER is computed based on the proposed method; CER_Y is based on [37]; $\hat{l}$ is estimated number of break points based on the proposed method; $\hat{l}_Y$ is estimated number of change points based on [37]; Diff is the average distance between estimated break points and true break points based on the proposed method; Diff_Y is the average distance between estimated change points and true change points based on [37]; FDR is the expected value of the ratio of false rejections over total rejections and Power is the expected value of the the ratio of true rejections over total number of non-nulls

k	$\hat{\sigma}^2$	CER	CER_Y	$\hat{l}$	$\hat{l}_Y$	Diff	Diff_Y	FDR	Power
$p = 600$									
					$N(0,1)$				
24	1.0533	0.0503	0.0538	2	2	15.35	16.63	0.0016	0.75
30	1.0665	0.0475	0.0508	2	2	14.30	15.68	0.0029	0.77
36	1.0605	0.0528	0.0492	2	2	15.85	15.24	0.0021	0.74
					$t(6)/1.5^{0.5}$				
24	1.0363	0.0489	0.0513	2	2	14.66	15.84	0.0015	0.76
30	1.0312	0.0511	0.0533	2	2	15.33	16.44	0.0019	0.75
36	1.0425	0.0554	0.0543	2	2	16.61	16.75	0.0020	0.73
					$\mathrm{LP}(0,1)/2^{0.5}$				
24	1.0128	0.0517	0.0528	2	2	18.50	16.33	0.0033	0.74
30	1.0377	0.0532	0.0548	2	2	17.08	16.85	0.0051	0.74
36	1.0630	0.0528	0.0497	2	2	15.84	15.36	0.0010	0.74
$p = 2000$									
					$N(0,1)$				
44	1.0469	0.0262	0.0495	2	2	29.54	50.05	0.0021	0.87
55	1.0535	0.0244	0.0514	2	2	25.96	51.90	0.0016	0.88
66	1.0420	0.0251	0.0499	2	2	25.15	50.36	0.0025	0.88
					$t(6)/1.5^{0.5}$				
44	1.0342	0.0279	0.0505	2	2	31.46	50.97	0.0015	0.86
55	1.0394	0.0248	0.0503	2	2	26.64	50.82	0.0019	0.88
66	1.0355	0.0265	0.0518	2	2	26.52	52.28	0.0032	0.87
					$\mathrm{LP}(0,1)/2^{0.5}$				
44	1.0382	0.0278	0.0524	2	2	32.84	52.91	0.0006	0.86
55	1.0569	0.0228	0.0500	2	2	22.77	50.46	0.0018	0.89
66	1.0475	0.0111	0.0498	2	2	26.15	50.33	0.0017	0.87
$p = 6000$									
					$N(0,1)$				
60	1.0433	0.0170	0.0495	2	2	51.51	148.85	0.0007	0.92
77	1.0396	0.0116	0.0489	2	2	40.67	147.34	0.0009	0.94
100	1.0395	0.0108	0.0509	2	2	32.34	153.22	0.0015	0.95
					$t(6)/1.5^{0.5}$				
60	1.0308	0.0168	0.0505	2	2	66.74	152.02	0.0008	0.92
77	1.0330	0.0127	0.0505	2	2	43.87	151.99	0.0014	0.94
100	1.0345	0.0103	0.0507	2	2	30.94	152.71	0.0011	0.95
					$\mathrm{LP}(0,1)/2^{0.5}$				
60	1.0429	0.0181	0.0493	2	2	65.60	148.27	0.0004	0.91
77	1.0465	0.0136	0.0499	2	2	46.42	150.07	0.0009	0.93
100	1.0452	0.0111	0.0498	2	2	33.16	150.05	0.0014	0.95

Table 3. BH procedure with Gaussian error term. The definition of FDR, Power and CER are the same as that in Table 2

p	FDR_{BH}	Power_{BH}	CER_{BH}
600	0.0435	0.0011	0.0724
6000	0.0433	0.0003	0.1101

on [37], especially with large number of tests. Both procedures correctly identified 2 break points. The difference between estimated break points and true ones are smaller based on the proposed method especially with large samples. Our results are robust to different error terms and the sliding window length k. For different error distributions the respective values of CER are quite close, as expected from our theoretical result.

Per the request of a referee, we implement the BH procedure [2] with Gaussian error term and summarize the results in Table 3. The simulation set up is the same as that in Table 2. At FDR level 5%, we can see that the BH procedure always controls FDR but with low power for clustered signals.

We also conduct simulation studies to check the empirical type-I error rates under the global null with Gaussian error term. The results are summarized in Table 4. At significance level 5%, the proposed method has a similar type-I error rate to BH procedure under the global null as evidenced from Table 4.

4.2. Simulation study 2

In this section, we examine the two-sided test procedure. Data is generated through model (2.1). Let $p = 600$, 2000 and 6000. The signal configuration is summarized in Table 5. We look at the robustness of our procedure with different error terms ($N(0,1)/2^{0.25}$, $t(10)/(75/16)^{0.25}$ and $LP(0,1)/20^{0.25}$), which are standardized to have $\kappa = 1$. Window size $k = \lfloor p^{1/2} \rfloor$ and $m = \lfloor p^{1/2} \rfloor$ are used for illustration. The calculation of $\hat{\sigma}^2$, γ and δ are the same as that in simulation study 1 except that $\hat{\kappa}$ is used instead of $\hat{\sigma}$ and the calculation of $\hat{\kappa}$ is through $\hat{\kappa}^2 = \hat{\nu}_{(k)}/2 - 4\hat{\sigma}^4_{(k)}$. We follow Algorithm 2.2 to implement our method. As a comparison, results based on true values of σ^2 and κ are presented as well. From Table 6, we can see that procedures using the estimated parameters and the true ones have a comparable performance in terms of CER, FDR, power, estimated number of break points and the difference between estimated break points and true break points. This is consistent with our large sample theory. The results are relatively robust across different error terms. As numbers of tests increase, CER and FDR decrease and power and the difference between estimated break points and true break points increase.

Table 4. Type-I error rates under the global null with Gaussian error term. FDR represents FDR based on the proposed method and FDR_{BH} represents FDR based on BH procedure

p	k	FDR	FDR_{BH}
600	36	0.0594	0.0495
6000	60	0.0396	0.0495

Table 5. Signal configuration for the two-sided test. "−1 and 1 alternating": μ_i is −1 if i is odd and 1 if i is even, seq(0.5, 1.5): a linear sequence from 0.5 to 1.5 and seq(1.5, 0.5): a linear sequence from 1.5 to 0.5

Segment (%)	30	10	20	5	5	30
Signal strength	0	−1 and 1 alternating	0	seq(0.5, 1.5)	seq(1.5, 0.5)	0

5. Applications to real data

We now apply our procedure to an array-based Comparative Genomic Hybridization (array CGH) data. Array CGH is a powerful technology for measuring copy numbers at thousands of loci simultaneously. The output of array CGH experiment is usually a long vector, spanning each chromosome, recording the $\log_2$ ratios of the normalized probe intensities from the test samples vs. the reference samples. These ratios of intensities are used to approximate the ratios of DNA copy numbers in the test samples vs. the reference samples. A $\log_2$ ratio far from 0 (either positive or negative) indicates a possible DNA copy number amplification or deletion for the probe. Identification of chromosomal alteration regions will provide valuable information to elucidate disease etiology and to discover novel disease related genes.

In the study conducted by [26], cDNA microarray CGH was profiled across 6691 mapped human genes in 44 breast tumor samples and 10 breast cancer cell lines. The raw data can be downloaded from the PNAS website (https://www.pnas.org/content/suppl/2002/09/23/162471999.DC1/4719CopyNoDatasetLegend.html). We picked the breast cancer cell line BT474 as an example, and applied our method to detect DNA copy number amplification. Details of one realization are in the Supplementary Material and the results are presented in Table 7

For multiple realization analysis, we consider the one-sided test using cell line 1 in addition to BT474 for analysis. We compute $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ for $i = 1, \ldots, p$, and test statistics $\hat{R}_j = \sum_{l=j}^{j+k-1} \sqrt{2}\hat{\mu}_l / (\sum_{l=k}^{j+k-1} \hat{\sigma}_l^2)^{1/2}$, $j = 1, \ldots, p-k+1$. We use the same window length as in the one realization case $k = \lfloor p^{1/2} \rfloor = 78$, and compute $Q_j^\star = 1(R_j^\star > \gamma) + 1(L_j^\star > \gamma)$ following algorithm 3.1. Critical values $\gamma = 3.8907$ and $\delta = 1.0992$ are obtained through the 0.95th quan-

Table 6. Summary statistics for two-sided test with 1000 simulations. Underscore e is based on estimated σ^2 and κ, and underscore t is based on true σ^2 and κ

p	CER_e	CER_t	FDR_e	FDR_t	$Power_e$	$Power_t$	$\hat{l}_e$	$\hat{l}_t$	$Diff_e$	$Diff_t$
$N(0,1)/2^{0.25}$										
600	0.0822	0.0743	0.0207	0.0323	0.61	0.65	4	4	19.67	21.53
2000	0.0390	0.0354	0.0091	0.0142	0.81	0.84	4	4	19.47	19.81
6000	0.0223	0.0202	0.0048	0.0101	0.89	0.91	4	4	35.64	38.68
$t(10)/(75/16)^{0.25}$										
600	0.0817	0.0753	0.0266	0.0367	0.61	0.65	4	4	17.62	18.84
2000	0.0378	0.0339	0.0129	0.0203	0.82	0.85	4	4	36.37	30.40
6000	0.0208	0.0192	0.0073	0.0109	0.90	0.91	4	4	51.26	64.54
$LP(0,1)/20^{0.25}$										
600	0.0758	0.0642	0.0217	0.0295	0.64	0.71	4	4	18.53	13.03
2000	0.0350	0.0327	0.0140	0.0173	0.84	0.85	4	4	35.70	41.91
6000	0.0199	0.0178	0.0078	0.0103	0.91	0.92	4	4	79.76	79.63

Table 7. Results based on one sequence and multiple sequence with one-sided test

One realization			Multiple realizations		
Chromosome number	Beginning loci	Ending loci	Chromosome number	Beginning loci	Ending loci
11	68434309	81603744	11	46512342	81,603,744
			14	16522721	106,822,024
			15	17156123	18,891,425
17	28552955	82172608	17	28552955	42,040,770
20	43585793	66314778	20	44457372	66,314,778
21	12430025	15830914	21	12430025	15,889,676

Note: Chromosome 14 and 15 are connected as one cluster with very short segments in chromosome 15 with multiple realizations analysis, chromosome 20 and 21 are connected as one cluster with both one realization and multiple realizations analysis.

tile of the empirical distribution of $\max_{1\le j\le p-k+1} G_j^\star$ and $\max_{j\in W_1} G_j^\star$, respectively, where $G_j^\star = \sum_{l=j}^{j+k-1} \hat{\sigma}_l \eta_l / (\sum_{l=j}^{j+k-1} \hat{\sigma}_l^2)^{1/2}$, $j \in \mathbb{Z}$, η_j are i.i.d. $N(0,1)$ random variables and W_1 is the transition region which includes indices j such that the smoothed $\tilde{Q}_j^\star = 1$.

The results are summarized in Table 7. We can see that four clustered regions are detected by the multiple realizations analysis, three of which overlap with those detected by one realization analysis, which shows that amplifications in these genome regions are shared among the two breast cancer patients. The identified chromosomal amplification regions are implicated in the literature to harbor genes associated with breast cancer [25]. In cancer studies, "passenger" mutations tend to occur more or less randomly throughout the genome, and "driver" mutations tend to cluster and favor certain genome positions containing functionally relevant genes. An important goal in the analysis of tumor cell lines is to find the "driver" mutations, which play a functional role in driving tumor progression [30]. Thus our analysis can suggest followup studies and intervention strategies. We choose to conduct our data analysis at the genome level, rather than at the chromosome level because genome scale analysis allows the detection of copy number aberrations involving entire chromosome arms, which might be missed in chromosome-level analyses for which no actual changepoints exist.

Acknowledgements

We are grateful to two referees for their many helpful comments. The research is partially supported by NSF and NIH.

Supplementary Material

Supplement to "Testing and estimation for clustered signals" (DOI: 10.3150/21-BEJ1355SUPP; .pdf). Supplementary material Section 1 provides proofs for some results in Sections 2 and 3. Supplementary material Section 2 presents some additional simulation studies and real data analysis

References

[1] Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549 https://doi.org/10.1198/016214507000000941

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[3] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. MR0348906

[4] Cai, T.T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. MR2597000 https://doi.org/10.1198/jasa.2009.tm08415

[5] Cao, H., Chen, J. and Zhang, X. (2021). Optimal false discovery rate control for large scale multiple testing with auxiliary information. Available at arXiv:2103.15311.

[6] Cao, H. and Kosorok, M.R. (2011). Simultaneous critical values for t-tests in very high dimensions. *Bernoulli* **17** 347–394. MR2797995 https://doi.org/10.3150/10-BEJ272

[7] Cao, H. and Wu, W.B. (2015). Changepoint estimation: Another look at multiple testing problems. *Biometrika* **102** 974–980. MR3431567 https://doi.org/10.1093/biomet/asv031

[8] Cao, H. and Wu, W.B. (2022). Supplement to "Testing and estimation for clustered signals." https://doi.org/10.3150/21-BEJ1355SUPP

[9] Chan, H.P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statist. Sinica* **23** 409–428. MR3076173

[10] Chouldechova, A. (2014). *False Discovery Rate Control for Spatial Data*. Ann Arbor, MI: ProQuest LLC. Thesis (Ph.D.)–Stanford University. MR4144738

[11] Du, L. and Zhang, C. (2014). Single-index modulated multiple testing. *Ann. Statist.* **42** 30–79. MR3226157 https://doi.org/10.1214/14-AOS1222

[12] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91** 674–688. MR1395735 https://doi.org/10.2307/2291663

[13] Fan, J., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.* **102** 1282–1288. MR2372536 https://doi.org/10.1198/016214507000000969

[14] Genovese, C.R., Roeder, K. and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439 https://doi.org/10.1093/biomet/93.3.509

[15] Hall, P., Kay, J.W. and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842 https://doi.org/10.1093/biomet/77.3.521

[16] Heller, R., Stanley, D., Yekutieli, D., Rubin, N. and Benjamini, Y. (2006). Cluster-based analysis of FMRI data. *NeuroImage* **33** 599–608. https://doi.org/10.1016/j.neuroimage.2006.04.233

[17] Hu, J.X., Zhao, H. and Zhou, H.H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.* **105** 1215–1227. MR2752616 https://doi.org/10.1198/jasa.2010.tm09329

[18] Lei, L. and Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 649–679. MR3849338 https://doi.org/10.1111/rssb.12253

[19] Li, A. and Barber, R.F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Amer. Statist. Assoc.* **112** 837–849. MR3671774 https://doi.org/10.1080/01621459.2016.1180989

[20] Li, A. and Barber, R.F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. MR3904779

[21] Liu, W. and Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025. MR3262475 https://doi.org/10.1214/14-AOS1249

[22] Liu, W.-D. (2015). Incorporation of sparsity information in large-scale multiple two-sample t tests. ArXiv e-prints. 2015. Available at arXiv:1410.4282.

[23] Patil, G.P. and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection. *Statist. Sci.* **18** 457–465. MR2109372 https://doi.org/10.1214/ss/1081443229

[24] Perone Pacifico, M., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* **99** 1002–1014. MR2109490 https://doi.org/10.1198/0162145000001655

[25] Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23** 41–46.

[26] Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.-L. and Brown, P.O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* **99** 12963–12968.

[27] Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **1** 123–137.

[28] Shen, X., Huang, H.-C. and Cressie, N. (2002). Nonparametric hypothesis testing for a spatial signal. *J. Amer. Statist. Assoc.* **97** 1122–1140. MR1951265 https://doi.org/10.1198/016214502388618933

[29] Siegmund, D.O., Zhang, N.R. and Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98** 979–985. MR2860337 https://doi.org/10.1093/biomet/asr057

[30] Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009). The cancer genome. *Nature* **458** 719–724.

[31] Sun, W., Reich, B.J., Cai, T.T., Guindani, M. and Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 59–83. MR3299399 https://doi.org/10.1111/rssb.12064

[32] Szor, P. (2005). *The Art of Computer Virus Research and Defense*. Reading: Addison-Wesley.

[33] Tansey, W., Koyejo, O., Poldrack, R.A. and Scott, J.G. (2018). False discovery rate smoothing. *J. Amer. Statist. Assoc.* **113** 1156–1171. MR3862347 https://doi.org/10.1080/01621459.2017.1319838

[34] Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9** 18–29.

[35] Wu, W. and Zhou, Z. (2018). Gradient-based structural change detection for nonstationary time series M-estimation. *Ann. Statist.* **46** 1197–1224. MR3798001 https://doi.org/10.1214/17-AOS1582

[36] Wu, W.-C. and Zhou, Z. (2020). Multiscale jump testing and estimation under complex temporal dynamics. In ArXiv E-prints, 2020. Available at arXiv:1909.06307.

[37] Yao, Q.W. (1993). Tests for change-points with epidemic alternatives. *Biometrika* **80** 179–191. MR1225223 https://doi.org/10.1093/biomet/80.1.179

[38] Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373 https://doi.org/10.1016/0167-7152(88)90118-6

[39] Zhang, C., Fan, J. and Yu, T. (2011). Multiple testing via FDR_L for large-scale imaging data. *Ann. Statist.* **39** 613–642. MR2797858 https://doi.org/10.1214/10-AOS848

[40] Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9** e1003264. https://doi.org/10.1371/journal.pgen.1003264

Received July 2020 and revised April 2021