Statistica Sinica **33** (2023), 431-451
doi:https://doi.org/10.5705/ss.202020.0464

DETECTING MULTIPLE CHANGE POINTS: THE PULSE CRITERION

Wenbiao Zhao[1], Xuehu Zhu[2] and Lixing Zhu[3,4]

[1]*Renmin University of China,* [2]*Xi'an Jiaotong University,*
[3]*Beijing Normal University and* [4]*Hong Kong Baptist University*

Abstract: Exhaustive search-based optimization algorithms can be computationally intensive and hypothesis testing-based procedures may encounter the false positive problem. To avoid these problems, we revisit change point detection of means and variances in a sequence of observations. We also propose a novel criterion, using a signal statistic to define a consistent estimation, even when the number of change points can go to infinity at a certain rate as the sample size goes to infinity. The signal statistic exhibits a useful "PULSE" pattern near change points, such that we can simultaneously identify all change points. The estimation consistency holds for the number of change points and for locations, in a certain sense. Furthermore, its visual nature means the locations can be more easily identified using plots than when using existing methods in the literature. The method can also detect weak signals in the sense that those changes go to zero. As a generic methodology, it may be extendable to handle other models. Numerical studies validate its good performance of the proposed method.

Key words and phrases: Double average ratios, multiple change-points detection, threshold, visualization.

1. Introduction

When there is a sequence of observations available, change point detection has attracted significant attention in a variety of research fields. For example Wu and Zhao (2007) detected mean changes in time series data for financial modeling, and Muggeo and Adelfio (2011) identified genes associated with diseases by applying a method of change point detection for means. There are a number of methods available in the literature; see, for example, Niu, Hao and Zhang (2016) for a comprehensive review.

Here, we focus on detecting mean changes, and as an adoption of the method, detecting variance changes. The following brief review stimulates us to consider a new way of investigating this issue, which has the potential to handle more

Corresponding author: Lixing Zhu, Department of Mathematics, Hong Kong Baptist University, Hong Kong, China. E-mail: lzhu@hkbu.edu.hk.

complex data structures. Several objective function-based criteria with optimization algorithms for exhaustive searches have been proposed for problems with fixed numbers of change points. Yao (1988)) suggested a BIC-Hyphenate type criterion. Frick, Munk and Sieling (2014) suggested a simultaneous multiscale change-point estimator (SMUCE) by solving an optimization problem, and Yao and Au (1989)) proposed a penalized least squares-based approach for mean changes. A weighted least squares function-based method was suggested by Gao et al. (2019). Harchaoui and Levy-Leduc (2010) proposed a lasso-based approach. The estimation consistency can be ensured under certain regularity conditions. One of the main concerns about these methods is their computational complexity. See the comments by Niu, Hao and Zhang (2016). When the number of change points goes to infinity as the sample size tends to infinity, the methods incur a greater computational cost. In contrast, cumulative sum (CUSUM)-based approaches are popular because of their lower computational cost. Here, relevant methods are based on hypothesis testing, and in many cases are efficient in terms of detection. The seminal paper by Page (1954)) had a significant influence on later developments. Vostrikova (1981) designed tests for multiple changes using binary segmentation procedures. To alleviate the difficulty caused by short spacings between change points or small jump magnitudes, Fryzlewicz (2014, 2020) introduced an additional randomization step in the algorithms called WBS and WBS2, where WBS2 is shown to be computationally more efficient. Using the moving sum (MOSUM) or "scan" statistic to construct the test statistic is also a popular technique; see Bauer and Hackl (1980) and Chu, Hornik and Kuan (1995). Wu and Zhao (2007) and Cao and Wu (2015) discussed the limiting distributions of the maxima of MOSUM. Hao, Niu and Zhang. (2013) considered a MOSUM-based test statistic, called the screening and ranking algorithm (SaRa), to simultaneously detect multiple change points. A further development is by Fang, Li and Siegmund (2020), who also used a hypothesis testing-based method to detect multiple changes, and gave a good way to control false positives in the study of large deviation theory. To handle the case with a diverging number of change points, for the independent and identically distributed (*i.i.d.*) normal errors, Baranowski, Chen and Fryzlewicz (2019) extended the CUSUM-based procedure. Wang, Yu and Rinaldo (2020) extended the WBS procedure. Eichinger and Kirch (2018) also suggested a MOSUM-based statistic to simultaneously determine changes when the number of change points goes to infinity as the sample size tends to infinity. They used the maximum of local MOSUMs over all possible local intervals such that all local changes with sufficiently large magnitudes can be detected, as in Hao, Niu and Zhang. (2013). This is also a computationally efficient approach.

Note that controlling false positives (Benjamini and Hochberg (1995)) affects the threshold determination for these test statistics and, thus, the detection of change points. This affects the estimation consistency of existing CUSUM-based or MOSUM-based methods in two ways, although family-wise test procedures can alleviate this issue. First, owing to the effect of noise, with a nonzero probability, test statistics in the intervals with no changes could also be larger than the designed threshold. Thus, in theory, the estimation consistency for the number of change points, if no extra conditions are assumed, would need further study. Second, these methods can well detect those change points with magnitudes that are larger than a designed threshold. We call this detection under the global alternative hypothesis. However, the detectability of weak signals that converge to zero at a certain rate as the sample size tends to infinity, under local alternatives, needs careful study. When stepwise procedures are applied, Shao and Zhang (2010) and Tewes (2017) provided theoretical results related to weak signal detection. However, in stepwise procedures, the convergence rates of detectable weak signals to zero slow down because of increasingly smaller sample sizes owing to the stepwise segmentations. In other words, it becomes more diffficult to detect weak signals in sequential testing procedures. This is particularly the case with a diverging number of change points. To the best of our knowledge, there are no relevant results on the estimation consistency in the literature. Niu, Hao and Zhang (2016) provide a comprehensive review.

We propose a novel signal statistic, as well as a criterion. The key feature of the approach is that it involves neither an optimization algorithm nor a hypothesis testing procedure. As such we can alleviate the computational complexity and deal with the issue of false positives, while still achieving estimation consistency for the number of change points and their locations. The defined signal statistic, which uses a sequence of MOSUM-based ridge ratios of double moving averages, can meet these requirements. Note that to identify change points, we need to make the values at the true change points (or nearby in a certain sense) stand out. To the best of our knowledge, we have not seen a similar idea in the literature. The distinguishing feature of the new criterion is that the defined signal statistic is discontinuous, with a useful "PULSE" pattern near all change points: at the population level, any change point plus two times the segment length of the moving average attains a local minimum tending to zero, followed by a local maximum going to infinity. This feature makes change points stand out significantly, and thus provides an efficient way to simultaneously identify them. We give a toy example to show this pattern in Section 2 when we describe the criterion construction. Note that its visual nature means the plot of the signal statistic is easy to implement. We call this criterion the PULSE criterion. To show its usefulness, we check how sensitive the criterion is to "weak changes" in the sense that some changes in the sequence of local means converge to a sequence without mean changes, at a certain rate. As a generic methodology, it can be extended to handle other change points detection problems, such as distributional changes (e.g., Pollak (1987)), changes in regression models (e.g., Qu and Perron (2007)), change points of functional data (e.g., Berkes et al. (2009)), and high-dimensional change point detection (e.g., Wang and Samworth (2018)). However, it has a limitation in handling the problems with short spacings between two change points. This is because, to guarantee estimation consistency, the segment length of the moving averages needs to be sufficiently long. As a result, there is more than one change point in short spacing scenarios. We will have a brief discussion in Section 6.

The remainder of the paper is organized as follows. In Section 2, we introduce the criterion construction, and investigate the estimation consistency. Section 3 investigates weak signals, where the magnitudes of changes converge to zero at a certain rate as the sample size tends to infinity. Section 4 examines detecting changes over variances. Some numerical studies are put in Section 5. Section 6 contains an illustrative application to the detection of mean changes. Section 7 concludes the paper. All technical proofs are presented in the online Supplementary Material.

2. Methodological Development

2.1. Notation

Let $X_1, \ldots, X_n$ be independent one-dimensional random variables decomposed as

$$X_i = \mu_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

where $\mu_i = E(X_i)$ are the means. Assume that there are K change points $1 < z_1 < z_2 < \cdots < z_K < n$, such that $\mu_{z_{k-1}+j} = \mu^{(k)}$, for $k = 1, \ldots, K+1$ and $0 \leq j \leq z_k - z_{k-1} - 1$, where $z_0 = 0$ and $z_{K+1} = n$. For $k = 1, \ldots, K$, write $\beta_k = |\mu^{(k+1)} - \mu^{(k)}|$ for the (nonzero) difference in means between consecutive segments. The number K can go to infinity as the sample size goes to infinity.

Write the minimum length of the segments as α_n^*:

$$\alpha_n^* := \min_{0 \leq k \leq K} \{z_{k+1} - z_k\}, \tag{2.1}$$

and the minimum magnitudes of the mean changes as ν:

$$\nu := \min_{1\le k\le K} \beta_k. \tag{2.2}$$

Denote by $1 < \hat{z}_1 \le \hat{z}_2 \le \cdots \le \hat{z}_K < n-1$ the estimated locations.

2.2. Criterion construction

Construct a signal statistic using the following steps. Consider the mean change detection problem first.

Difference of Moving Averages: To character the mean information, let $S(i)$ be the moving sum with window size α_n for every location i:

$$S(i) = \sum_{j=i}^{i+\alpha_n-1} \mu_j. \tag{2.3}$$

Because the difference between two successive moving sums at the population level shows the mean change at its location z_k, we define $D(i)$ as follows: for $1 \le i \le n-2\alpha_n$, if $2\alpha_n < \alpha_n^*$,

$$D(i) := \frac{1}{\alpha_n}(S(i) - S(i-\alpha_n)) = \frac{1}{\alpha_n}\left(\sum_{j=i}^{i+\alpha_n-1} \mu_j - \sum_{j=i-\alpha_n}^{i-1} \mu_j\right). \tag{2.4}$$

For any fixed k, we have:

$$D(i) = \begin{cases} \dfrac{i-(z_k-\alpha_n)}{\alpha_n}(\mu_{k+1}-\mu_k), & z_k-\alpha_n \le i < z_k, \\ \dfrac{z_k+\alpha_n-i}{\alpha_n}(\mu_{k+1}-\mu_k), & z_k \le i \le z_k+\alpha_n \\ 0, & z_{k-1}+\alpha_n \le i \le z_k-\alpha_n. \end{cases} \tag{2.5}$$

This is because, when $z_{k-1}+\alpha_n \le i \le z_k-\alpha_n$, $S(i) = S(i+\alpha_n)$. Here, $D(i)$ attains a local maximum/minimum at $i = z_k$, for any k, with $1 \le k \le K$ within the segment of length $2\alpha_n$. Figure 1 presents a plot visualizing the pattern. This is simply the idea of MOSUM. Identifying local minima would be a way to identify changes. Because we expect it to have too many local maxima/minima, owing to the randomness oscillation, we may have difficulty accurately determining the number of change points and their locations. To make the differences more smooth at the sample level, we consider a smoothing step by double averaging. Note that the second averaging step is not necessary in theory, but in practice, we found it is useful for better detection.

Double Averaging: The second round of averaging repeatedly uses the data points in every average. Note that at the population level, this step is not necessary, but at the sample level, it alleviates the oscillation of the sequence; see Remark 1. Denote $\tilde{D}(i)$ as the average of $D(i)$ within the window of size α_n:

$$\tilde{D}(i) = \frac{1}{\alpha_n} \sum_{j=i}^{i+\alpha_n-1} D(j). \tag{2.6}$$

Thus we have that

$$\tilde{D}(i) \begin{cases} > 0, \; z_k - 2\alpha_n \leq i \leq z_k + \alpha_n, \\ = 0, \text{ otherwise,} \end{cases}$$

with the following detail:

$$|\tilde{D}(i)| = \begin{cases} 0, & z_{k-1}+\alpha_n \leq i \leq z_k - 2\alpha_n; \\ \frac{(i-z_k+2\alpha_n+1)\cdot(i-z_k+2\alpha_n)}{\alpha_n^2}\beta_k, & z_k - 2\alpha_n < i \leq z_k - \alpha_n; \\ \frac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+(\alpha_n^2-\alpha_n)/2}{\alpha_n^2}\beta_k, & z_k - \alpha_n < i < z_k - \frac{\alpha_n}{2} - \sqrt{\alpha}_n; \\ \left(\frac{3}{4} - \frac{\alpha_n - \sqrt{\alpha_n}}{\alpha_n^2}\right)\beta_k, & i = z_k - \frac{\alpha_n}{2} - \sqrt{\alpha}_n; \\ \frac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+(\alpha_n^2-\alpha_n)/2}{\alpha_n^2}\beta_k, & z_k - \frac{\alpha_n}{2} - \sqrt{\alpha}_n < i < z_k - \frac{\alpha_n}{2}; \\ \frac{3}{4}\beta_k, & i = z_k - \frac{1}{2}\alpha_n; \\ \frac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+(\alpha_n^2-\alpha_n)/2}{\alpha_n^2}\beta_k, & z_k - \frac{\alpha_n}{2} < i < z_k - \frac{\alpha_n}{2} + \sqrt{\alpha}_n; \\ \left(\frac{3}{4} - \frac{\alpha_n - \sqrt{\alpha_n}}{\alpha_n^2}\right)\beta_k, & i = z_k - \frac{\alpha_n}{2} + \sqrt{\alpha}_n; \\ \frac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+(\alpha_n^2-\alpha_n)/2}{\alpha_n^2}\beta_k, & z_k - \frac{\alpha_n}{2} + \sqrt{\alpha}_n < i \leq z_k; \\ \frac{(-i+z_k+\alpha_n+2)(-i+1+\alpha_n+z_k)}{\alpha_n^2}\beta_k, & z_k < i \leq z_k + \alpha_n; \\ 0, & z_k + \alpha_n < i \leq z_{k+1} - 2\alpha_n. \end{cases}$$

where $\beta_k = |\mu_{k+1} - \mu_k|$. Clearly, $\tilde{D}(i)$ attains a local maxima at $z_k - \alpha_n/2$, for each k, with $1 \leq k \leq K$. The local maximizers of $\tilde{D}(i)$ plus $\alpha_n/2$ are the locations of the change points. Similarly to $D(i)$, the sequence $\tilde{D}(i)$ cannot be used directly as a signal statistic either. Now, we construct a sequence of ridge ratios as a signal statistic with a "pulse" pattern that identifies change points well.

Signal function (we call it the signal statistic at the sample level). Consider the ratios between $\tilde{D}(i)$ and $\tilde{D}(i + 3\alpha_n/2)$. Define the ridge ratios $T(i)$ at the population level as

$$T(i) = \frac{|\tilde{D}(i)| + c_n}{|\tilde{D}(i + (3/2)\alpha_n)| + c_n}, \tag{2.7}$$

where $c_n \to 0$ as $n \to \infty$, to be selected later, to avoid the undefined terms 0/0. In addition, for $i \in (z_{k-1} + \alpha_n, z_k - 2\alpha_n)$, $|\tilde{D}(i)| = 0$ and $|\tilde{D}(i + 3\alpha_n/2)|$ monotonically increases. For $i \in (z_k - 2\alpha_n, z_k - \alpha_n/2)$, $|\tilde{D}(i))|$ monotonically increases, and $|\tilde{D}(i+3\alpha_n/2)|$ monotonically decreases. For $i \in (z_k - \alpha_n/2, z_k + \alpha_n)$, $|\tilde{D}(i + 3\alpha_n/2)| = 0$ and $|\tilde{D}(i)|$ monotonically decreases. Then, c_n could also play a role of making $T(i)$ monotonic, to avoid the scenario where there are too many points tending to zero. In summary, the following property can be easily justified. Let $\searrow$ and $\nearrow$ mean decreasing and increasing, respectively, with respect to the index i. In addition, $\to 0$ and $\to \infty$ mean going to zero and infinity, respectively, as $n \to \infty$. Then,

$$T(i) = \begin{cases} 1, & z_{k-1} + \alpha_n \le i \le z_k - \frac{7}{2}\alpha_n, \\ \frac{c_n}{|\tilde{D}(i+3\alpha_n/2)|+c_n} \searrow, & z_k - \frac{7}{2}\alpha_n < i < z_k - 2\alpha_n, \\ \frac{c_n}{|\tilde{D}(i+3\alpha_n/2)|+c_n} \to 0, & i = z_k - 2\alpha_n, \\ \frac{|\tilde{D}(i)|+c_n}{|\tilde{D}(i+3\alpha_n/2)|+c_n} \nearrow, & z_k - 2\alpha_n < i < z_k - \frac{\alpha_n}{2}, \\ \frac{|\tilde{D}(i)|+c_n}{c_n} \to \infty, & i = z_k - \frac{\alpha_n}{2}, \\ \frac{|\tilde{D}(i)|+c_n}{c_n} \searrow, & z_k - \frac{\alpha_n}{2} < i < z_k + \alpha_n, \\ 1, & z_k + \alpha_n \le i < z_{k+1} - \frac{7}{2}\alpha_n. \end{cases}$$

Any true change point is just the local minimizer plus $2\alpha_n$. Based on the signal function, using the local minimizers to identify the change points is convenient to implement. The toy example in Figure 1 shows the curve patterns of $D(i)$, $|\tilde{D}(i + \alpha_n/2)|$ and $T(i + 2\alpha_n)$, enabling us to better understand why the pulse pattern of the signal function $T(i)$, rather than that of $D(i)$ or $|\tilde{D}(i)|$, can be used to construct a useful criterion. Define their empirical versions.

Sample Version. To define the signal statistic at the sample level, which is called the signal function at the population level, we can use the sample averages to estimate $D(i)$ and $\tilde{D}(i)$. Let $\hat{S}(i) = \sum_{j=i}^{i+\alpha_n-1} X_j$ to estimate $S(i)$, $D_n(i) = (1/\alpha_n)(\hat{S}(i) - \hat{S}(i+\alpha_n))$ and $\tilde{D}_n(i) = (1/\alpha_n)\sum_{j=i}^{i+\alpha_n-1} D_n(j)$. The signal statistic is then defined as follows: for $i = 1, \ldots, n - 7\alpha_n/2$,

$$T_n(i) = \frac{|\tilde{D}_n(i)| + c_n}{|\tilde{D}_n(i + 3\alpha_n/2)| + c_n}, \tag{2.8}$$

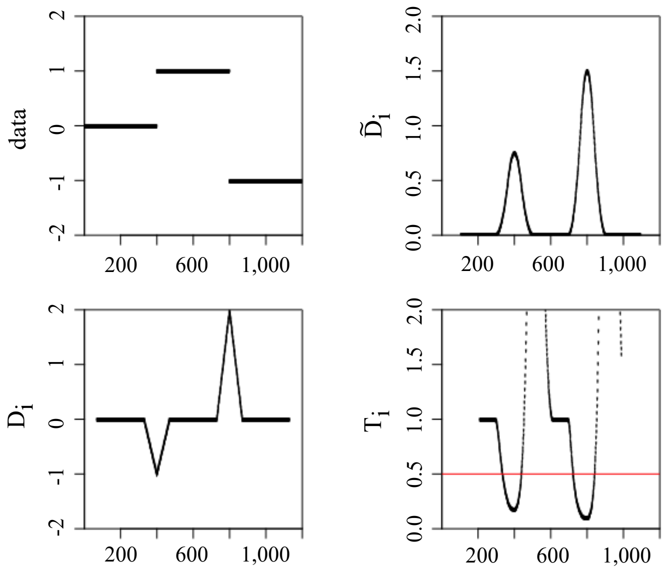

Figure 1. The plots at the population level.

and the ridge value c_n tends to zero at a certain rate specified later. We can see that $E\tilde{D}_n(i) = \tilde{D}(i)$.

Criterion: As we discussed above, the signal statistic should be highly oscillating, and there are too many local minima. Thus, we restrict our search, separately within each chosen interval, to find a local minimum of $T_n(i)$. We do this through a threshold τ, with $0 < \tau < 1$. That is,

$$\left\{ i, \alpha_n + 1 \le i \le n - \frac{5}{2}\alpha_n : T_n(i) < \tau \right\}.$$

From the properties of $T_n(i)$, which can also be seen from the plot of Figure 1 heuristically, all these indices can be separated into disjoint subsets, each containing only one change point asymptotically. Therefore, we can search, separately within the disjoint subsets, for local minima. To make the search easily implementable, we recommend $\tau = 0.5$ to avoid possible overestimation with large τ close to one and underestimation with small τ close to zero. From the definition of $T_n(i)$ and its pulse pattern, we can also search for the changes by identifying local maxima that, at the population level, tend to infinity. However, this is equivalent to using $1/T_n(i)$, and we do not discuss it further here. In addition, from the definition of $T(i)$ at the population level, the gap between two local minimizers must be larger than $2\alpha_n$. Owing to the consistency of the involved estimators, there are $\hat{K}$ pairs $\{m_k, M_k\}$, where m_k and M_k with $m_k < M_k$ are determined by $T_n(i) < 0.5$. In addition, m_k satisfies that $T_n(m_k - 1) \geq 0.5$ and $T_n(m_k) < 0.5$, and $T_n(M_k) < 0.5$ and $T_n(M_k + 1) \geq 0.5$. Write $\hat{z}_k - 2\alpha_n$ as the minimizer in each interval (m_k, M_k).

Theorem 1. *Assume that $X_i - EX_i$ with second moments are i.i.d. random variables. The tuning parameter c_n and the window size α_n satisfy $c_n/\sqrt{\log n/\alpha_n} \to \infty$, $n^{1/4}\log n/\sqrt{\alpha}_n \to 0$, and $\alpha_n^*/\alpha_n \to \infty$, where α_n^* is the minimum number of samples between any two change points.*

(1) *When K is known, then the estimators $\{\hat{z}_1, \ldots, \hat{z}_K\}$ have, for every $\epsilon > 0$, $\Pr\{\max_{1\leq k\leq \hat{K}} |(\hat{z}_k - z_k)/\alpha_n| < \epsilon\} \to 1$ as $n \to \infty$.*

(2) *When K is fixed but unknown, then $\hat{K} = K$ with probability going to one, and the estimators $\{\hat{z}_1, \ldots, \hat{z}_{\hat{K}}\}$ have the same consistency as the above.*

(3) *When $K = K_n$ grows unbounded at the rate satisfying $n/(\alpha^* K_n) \to \infty$ and is unknown, the results are the same as those in* (2)*.*

Remark 1. We have several issues to discuss.

(1) For selecting the window size α_n, we wish to use a small α_n such that we can detect changes within relatively short segments. On the other hand, we need a large α_n such that $\tilde{D}_n(\cdot)$ has a fast rate of convergence to make a wide range of the ridge c_n. This can make the signal statistic $T_n(\cdot)$ converge to its limit. In this sense, the optimal selection, if possible, should be different from the optimal tuning parameter selection in nonparametric estimation, which tries to balance between bias and variance. Because we need not discuss the limiting distribution and the bias and variance, we do not have an optimal selection of α_n for the estimation efficiency, and do not know whether an optimal choice exists.

(2) As previously discussed, the second averaging is mainly for practical use, because it makes the signal statistic less oscillating. The costs are as follows: 1) the technical proof becomes more complicated, but still manageable; and 2) the segment length for each change, if we consider even higher order averaging, should be increased to $(o + 1)\alpha_n$ from $2\alpha_n$ for the first-order averaging, where o is the order of the averaging. From the plot, double averaging ($o = 2$) makes the curve sufficiently smooth. Triple or higher order averaging may not be necessary any more, which requires an even longer segment to detect each change.

Remark 2. The conditions on the rates of divergence of α_n, α_n^*, and K_n are based on the following observations. First, we prove that $\tilde{D}_n(i) - \tilde{D}(i)$ converges in probability to zero at a rate of order $\sqrt{\log n/\alpha_n}$. Then, the ridge c_n should be a dominating term in every $T_n(i)$, which converges to zero at a rate slower than that of $\tilde{D}_n(i) - \tilde{D}(i)$. Such a ridge can help $T_n(i)$ hold the properties of $T(i)$ asymptotically. We show this in the proofs of the theorems in the Supplementary Material.

For the ridge c_n, we recommend a choice for practical use in the numerical studies. To guarantee the estimation consistency, α_n should not be too small such that the averages can be close to the corresponding means. Thus, for paradigms with short spacing, our method may not perform well. We present a discussion in Section 7.

Remark 3. Because one pulse corresponds to one change, the length of a segment cannot be longer than the minimum distance α_n^* between any two changes. Thus, we assume that the window size $\alpha_n = o(\alpha_n^*)$. On the other hand, α_n cannot be too small, otherwise the convergence rate of T_n to T will be slow. In our results, we assume that $n^{1/4}\log n/\sqrt{\alpha}_n \to 0$, although this condition can be weakened, which is beyond the scope of this study. These conditions also restrict the number of change points satisfying $K_n = o(n/\alpha_n^*)$.

3. The Case of Weak Signals

In this section, we extend the criterion to handle weak signal scenarios. The term "weak signals" means that the magnitudes of some changes converge to zero at a certain rate as the sample size goes to infinity. We also call such models local models. Consider the following sequence of models: for $1 \le k \le K$:

$$X_i = \mu + \beta_{z_k}\mathbb{I}\{i \ge z_k\} + \epsilon, \tag{3.1}$$

where z_k are the locations of the change points, and β_{z_k} are the change magnitudes, which converge to zero as $n \to \infty$. Denote $\beta_z = \min_{1\le k\le K}\beta_k$. We have the following results.

Theorem 2. *Under the conditions in Theorem* 1*, for the sequence of local models in* (3.1)*, when* $\log {\alpha_n}^{1/5}\beta_z \to \infty$*, we have* $\lim_{n\to\infty}\Pr\{\hat{K} = K\} = 1$ *and* $\lim_{n\to\infty}\Pr\{\max_{1\le k\le \hat{K}} |\hat{z}_k - z_k|/\alpha_n < \epsilon\} = 1$*, for every* $\epsilon > 0$.

4. Change Points in Variances

In this section, we adopt the criterion for detecting change points in variances. Consider the second moments of X_i that are generated from the following model:

$$X_i = \mu + \varepsilon_i, \quad 1 \le i \le n, \tag{4.1}$$

where μ is an unknown mean and $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2_{(i)}$. Similarly, we assume that $\sigma^2_{(i)}$ follows a piecewise constant structure with $K+1$ segments. In other words, there are K change points $1 < z_1 < z_2 < \cdots < z_K < n-1$ such that, for any k with $0 \le k \le K$,

$$\sigma^2_{z_k+1} = \cdots = \sigma^2_{z_{k+1}} = \sigma^2_k. \tag{4.2}$$

As before, define $z_0 = 0$ and $z_{K+1} = n$. At the population level, we similarly define $D(i)$ and $\tilde{D}(i)$:

$$D(i) = \log \sigma_{(i)} - \log \sigma_{(i-\alpha_n)} \quad \text{and} \quad \tilde{D}(i) = \frac{1}{\alpha_n} \sum_{j=i}^{i+\alpha_n-1} D(j).$$

We estimate μ by the sample mean and the variance by

$$\hat{\sigma}^2_{(i)} = \frac{1}{\alpha_n} \sum_{t=i}^{i+\alpha_n-1} \left(X_t - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^2, \tag{4.3}$$

and $D_n(i)$ and $\tilde{D}_n(i)$ are defined as the difference between the moving averages and the average of $D_n(j)$:

$$D_n(i) = \log \hat{\sigma}_{(i)} - \log \hat{\sigma}_{(i+\alpha_n)} \quad \text{and} \quad \tilde{D}_n(i) = \frac{1}{\alpha_n} \sum_{j=i}^{i+\alpha_n} D_n(j). \tag{4.4}$$

Finally, we take the ratios of $\tilde{D}(i)$ to acquire the required estimator of $T(i)$:

$$T_n(i) = \frac{|\tilde{D}_n(i)| + c_n}{|\tilde{D}_n(i + 3\alpha_n/2)| + c_n}. \tag{4.5}$$

The criterion is the same as before using

$$\left\{ i, 1 \le i \le n - \frac{7}{2}\alpha_n : T_n(i) < \tau \right\}.$$

Theorem 3. *Assume that $X_i - \mu$ with fourth moments are independent distributed random variables and that the conditions in Theorem 1 are still satisfied. Then, all the parallel results to those in Theorem 1 still hold.*

5. Simulations

To evaluate the performance of the proposed criterion, we run the simulations with other competing methods. We consider the methods CumSeg in Muggeo and Adelfio (2011), SMUCE in Frick, Munk and Sieling (2014), and WBS in Fryzlewicz (2014), and the MOSUM approach considered in Eichinger and Kirch (2018). Our theoretical results require that α_n should be of order higher than $n^{1/2}$. Thus, we tried to choose a small α_n with an order close to $n^{1/2}$. In our algorithm, we take $\alpha_n = O(n^{0.6})$, $\tau = 0.5$ and $c_n = O(\sqrt{\log n/\alpha_n})$. These choices follow a rule of thumb, and are not data driven. The steps are given in Algorithm 1 below. Note that choosing an appropriate threshold τ is useful, and is done in a data-driven manner. By the rule of thumb, 0.5 can be a good compromise, and is thus recommended.

Algorithm 1. How to estimate change points.

Input: $X \in \mathcal{R}^{n\times 1}$

1: Take $\alpha_n = O(n^{0.6})$, $\tau = 0.5$, $c_n = O(\sqrt{\log n/\alpha_n})$

2: Perform PULSE construction steps directly to acquire the estimation $\hat{z}_i$, for $i = 1,\ldots,k$.

3: Estimate $\hat{z}_i$ based on parameters we select in step 1.

Output: $\hat{z}_i$

For each example, 1,000 replications are used to approximate the distribution of $\hat{K}-K$, where $\hat{K}_n$ is the estimated number of change points. We also report the Rand Index (Rand (1971))), which represents a measure of similarity between two different partitions of the same observations. The Rand Index was also reported by Matteson and James (2014) to measure the quality of change point locations. Thus we refer to their definition. Specifically, suppose that the two clusters of n observations are given by $U = U_1,\ldots,U_a$ and $V = V_1,\ldots,V_b$, respectively, with a and b clusters. For these two clusters, the Rand Index is calculated by noting the relative cluster membership for all pairs of observations. Consider the pairs of observations that fall into one of the following two sets: $\{A\}$ pairs of observations in the same cluster under U and in the same cluster under V; $\{B\}$ pairs of observations in different clusters under U and in different clusters under V. Let $\#A$ and $\#B$ denote the respective number of pairs of observations in each of these two sets. The Rand Index is then defined as

$$\text{Rand}\quad\text{Index} = \frac{\#A + \#B}{\binom{n}{2}}.$$

5.1. Mean changes

We adopt the blocks setting used in the literature (Fryzlewicz (2014)). The model is

$$X_i = \mu_i + \varepsilon_i, \tag{5.1}$$

with four error distributions:

(i) $\varepsilon_i \overset{i.i.d.}{\sim} N(0, 1)$,

(ii) $\varepsilon_i \overset{i.i.d.}{\sim} N(0, 3)$,

(iii) $\varepsilon_i \overset{i.i.d.}{\sim} 7 \cdot Uniform(0, 1)$,

(iv) $\varepsilon_i \overset{i.i.d.}{\sim} 3 \cdot t_3$,

where t_v is the Student's t-distribution with v degree of freedom. The sample size is $n = 2048$. The number of change points is $K = 11$, and the change points are located at positions 171, 341, 511, 681, 851, 1021, 1191, 1361, 1531, 1701, and 1871; the means μ_i are, respectively, 1, 3, 2, -1, 1, 3, 2, 5, 1, -2, 3, and 0. We call this model the change point (CP) model. To check whether the method can detect weak signals, we consider another case with the respective means μ_i 0, 0.7, 0, -0.7, 0.7, 0, 2, 2.7, 0, -2.7, -2, and 0. We call this model the weak signal CP model. The results are reported in Tables 1 - 4 to examine the performance of the competitors. Tables 1 and 2 are for the estimated numbers of changes points, and Tables 3 and 4 are for the estimated locations.

There are several observations from Table 1. In the normal case (i) with $\sigma^2 = 1$, PULSE shows competitive performance with WBS and SMUCE, and better than that of the others. In the normal case (ii) with $\sigma^2 = 3$, all other competitors have a serious underestimation issue. Our method overestimates the number of change points, but the proportion of $|\hat{K} - K| \leq 1$ is still equal to 95%. In the uniform case (iii) and Student's t case (iv), the proportions of $|\hat{K} - K| \leq 1$ are equal to 98% and 78.1%, respectively, much higher than those of the competitors; particularly in case (iv), all the others have very inaccurate estimations.

For the quality of the estimated locations, Table 3 indicates that our method shows competitive performance compared with that of the others. Table 4 shows that in weak signal cases, our method outperforms the others, although it reasonably performs worse than it does in the strong signal cases.

Table 1. Distribution of $\hat{K} - K$ with $K = 11$ for various detection algorithms under the CP model.

		$\hat{K} - K$							
Scenarios	Procedures	≤ -3	-2	-1	0	1	2	≥ 3	MSE
(i)$\sigma = 1$	MOSUM	0	150	414	383	52	1	0	1.07
	cumSeg	0	0	1	875	119	5	0	0.14
	WBS	0	0	0	983	15	2	0	0.023
	SMUCE	0	0	0	996	4	0	0	0.004
	PULSE	0	0	3	994	3	0	0	0.006
(ii)$\sigma^2 = 3$	MOSUM	881	101	14	4	0	0	0	15.562
	cumSeg	702	257	35	5	1	0	0	9.241
	WBS	255	606	120	19	0	0	0	4.979
	SMUCE	475	497	27	1	0	0	0	6.409
	PULSE	0	7	72	643	250	30	3	0.477
(iii)	MOSUM	168	515	256	53	7	1	0	3.992
	cumSeg	21	703	234	33	9	0	0	3.244
	WBS	0	415	477	107	1	0	0	2.138
	SMUCE	0	347	496	157	0	0	0	1.884
	PULSE	0	5	115	833	47	0	0	0.182
(iv)	MOSUM	1,000	0	0	0	0	0	0	55.94
	cumSeg	996	1	1	0	2	0	0	61.69
	WBS	1	0	0	3	1	3	992	221.19
	SMUCE	0	0	0	0	0	0	1,000	505.54
	PULSE	1	8	66	331	384	164	46	4.63

5.2. Variance changes

Consider the model

$$X_i = \sigma_i \varepsilon_i. \tag{5.2}$$

The distributions of ε_i are the same as those in the mean changes detection above. Four methods are compared, including SUMCE introduced by Frick, Munk and Sieling (2014), BS introduced in Scott and Knott (1974)), and PELT introduced in Killick, Fearnhead and Eckley (2012). Again the sample size is $n = 2048$ and $K = 11$. The change points are located at positions 161, 323, 485, 638, 801, 967, 1132, 1299, 1465, 1632, and 1794; σ_i are equal to 1, 0.25, 1, 5, 1, 0.25, 1, 5, 1, 0.25, 1, and 5, respectively. Tables 5 and 6 present the comparison results. Clearly, the new method works robustly against all distributions, and much better than SUMCE and BS, which underestimate greatly the true number K in all cases. The performance of PELT varies. In the normal cases (i) and (ii), it performs well overall. In the uniform case (iii), it underestimates the number of changes, and in the Student's t case (iv), it tends to have an overestimation. Note that

Table 2. Distribution of $\hat{K}-K$ for various detection algorithms under the weak signal model.

Scenarios	Procedures	$\hat{K}-K$							MSE
		≤ -3	-2	-1	0	1	2	≥ 3	
(i)$\sigma = 1$	MOSUM	953	39	8	0	0	0	0	21.957
	cumSeg	245	254	266	152	82	1	0	4.167
	WBS	0	51	224	612	111	2	0	0.547
	SMUCE	12	113	554	321	0	0	0	1.114
	PULSE	0	0	37	889	73	1	0	0.114
(ii)$\sigma^2 = 3$	MOSUM	1,000	0	0	0	0	0	0	89.395
	cumSeg	1,000	0	0	0	0	0	0	50.469
	WBS	1,000	0	0	0	0	0	0	39.165
	SMUCE	1,000	0	0	0	0	0	0	39.561
	PULSE	6	17	43	93	163	211	467	8.813
(iii)	MOSUM	1,000	0	0	0	0	0	0	56.917
	cumSeg	1,000	0	0	0	0	0	0	46.107
	WBS	997	3	0	0	0	0	0	24.915
	SMUCE	998	2	0	0	0	0	0	25.035
	PULSE	158	156	223	203	147	78	35	3.71
(iv)	MOSUM	1,000	0	0	0	0	0	0	42.371
	cumSeg	992	7	0	1	0	0	0	41.275
	WBS	0	0	0	0	1	0	999	248.034
	SMUCE	0	0	0	0	0	0	1,000	628.123
	PULSE	18	56	97	221	224	198	168	3.699

Student's t distribution in case (iv) does not satisfy the required condition for our theoretical results. We conducted this simulation to keep continuity with the previous section when detecting change points in means, and checked its practical use, even when the condition is violated. The simulation results show that our "PULSE" procedure is still capable of estimating the number of change points well.

6. Real-Data Example

Consider an Array CGH data set, which shows aberrations in genomic DNA. The observations are normalized glioblastoma profiles from the data set of Bredel et al (2005). We now detect regions on which the observations jump from zero. Compute $T_n(i)$ about the array CGH profile of chromosome 13 in GBM31. The threshold $\tau = 0.5$ and ridge c_n are selected as before. In Figure 2, we plot the original data, $D_n(i)$, $\tilde{D}_n(i)$, and $T_n(i)$. From $D_n(i)$, the magnitudes of the changes are small, except for a point between 500 and 600, which is also smaller than 0.5. $\tilde{D}_n(i)$ also shows this pattern, but magnitude of the change of the latter

Table 3. Rand Index for various detection algorithms under the CP model.

Scenario	Procedure	Rand Index
(i)$\sigma = 1$	MOSUM	0.9835
	cumSeg	0.9851
	WBS	0.9868
	SMUCE	0.9861
	PULSE	0.9869
(i)$\sigma = 3$	MOSUM	0.8395
	cumSeg	0.9502
	WBS	0.9658
	SMUCE	0.9603
	PULSE	0.9614
(i)(iii)	MOSUM	0.9381
	cumSeg	0.9650
	WBS	0.9663
	SMUCE	0.9682
	PULSE	0.9470
(i)(iv)	MOSUM	0.5897
	cumSeg	0.7519
	WBS	0.9445
	SMUCE	0.9469
	PULSE	0.9426

point is 0.4, which is still small. The plot of $T_n(i)$ presents a curve clearly showing that it can be regarded as a change point. $T_n(i)$ also suggests the number 579 as the location of a change point. From all four plots, determining this location is reliable.

7. Conclusion

We have proposed a generic method for detecting change points of means and variances. Our approach achieves estimation consistency with less computational complexity. The construction of $\tilde{D}$ with the segment length α_n can be viewed from a nonparametric estimation perspective for a function with fixed designed points $t = 1, \ldots, n$. This is because moving averages can be regarded as a local smoothing procedure. Thus, the optimal selection of α_n, if we want to study the estimation efficiency of T_n, can be regarded as the optimal selection of a tuning parameter. Note that the optimal selection of the tuning parameter in a nonparametric estimation tries to balance the estimation bias and the variance. However, as we discussed in Remarks 1 and 3, the optimality, if it exists, is related to the rate of convergence of the signal statistic. Thus, this is an essential difference in methodology. It deserves further study to determine in what sense

Table 4. Rand Index for various detection algorithms under the weak signal model.

Scenario	Procedure	Rand Index
(i)$\sigma = 1$	MOSUM	0.9031
	cumSeg	0.9517
	WBS	0.9749
	SMUCE	0.9678
	PULSE	0.9775
(i)$\sigma = 3$	MOSUM	0.5753
	cumSeg	0.7724
	WBS	0.8294
	SMUCE	0.8412
	PULSE	0.9510
(i)(iii)	MOSUM	0.7543
	cumSeg	0.7775
	WBS	0.8917
	SMUCE	0.8928
	PULSE	0.9535
(i)(iv)	MOSUM	0.8154
	cumSeg	0.7877
	WBS	0.9469
	SMUCE	0.9466
	PULSE	0.9549

Table 5. Distribution of $\hat{K} - K$ using various detection algorithms under the CP model for variance changes.

		$\hat{K} - K$							
Scenarios	Procedures	≤ -3	-2	-1	0	1	2	≥ 3	MSE
(i)$\sigma = 1$	SMUCE	1,000	0	0	0	0	0	0	9
	BS	1,000	0	0	0	0	0	0	36
	PELT	11	4	0	985	0	0	0	0.412
	PULSE	0	0	0	998	2	0	0	0.002
(ii)$\sigma^2 = 3$	SMUCE	1,000	0	0	0	0	0	0	9
	BS	1,000	0	0	0	0	0	0	36
	PELT	11	0	308	681	0	0	0	0.471
	PULSE	0	0	0	992	8	0	0	0.008
(iii)	SUMCE	1,000	0	0	0	0	0	0	9
	BS	1,000	0	0	0	0	0	0	36
	PELT	1,000	0	0	0	0	0	0	36
	PULSE	0	0	0	1,000	0	0	0	0
(iv)	SMUCE	1,000	0	0	0	0	0	0	9
	BS	1,000	0	0	0	0	0	0	36
	PELT	9	2	4	67	84	173	661	23.867
	PULSE	0	0	33	962	5	0	0	0.038

Table 6. Rand Index using various detection algorithms under the CP model for variance changes.

Scenario	Procedure	Rand Index
(i)$\sigma = 1$	SMUCE	0.9275
	BS	0.8732
	PELT	0.9868
	PULSE	0.9636
(i)$\sigma = 3$	SMUCE	0.9430
	BS	0.8604
	PELT	0.9829
	PULSE	0.9625
(i)(iii)	SMUCE	0.9347
	BS	0.8875
	PELT	0.8882
	PULSE	0.9675
(i)(iv)	SMUCE	0.9173
	BS	0.8574
	PELT	0.9770
	PULSE	0.9598

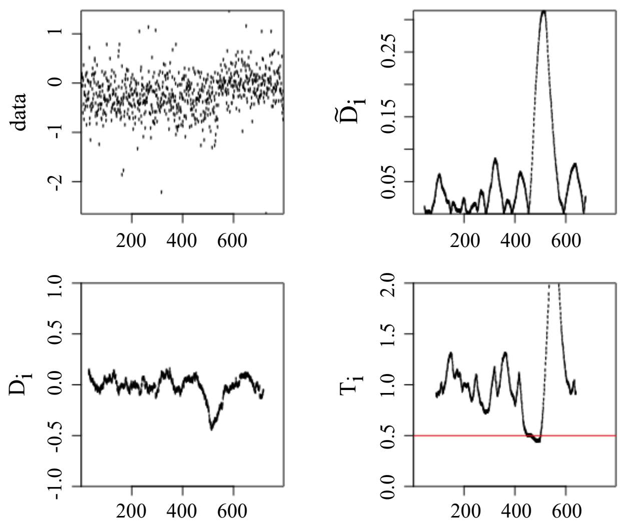

Figure 2. The plots for the Array CGH data.

we need an optimal selection, and whether the optimal α_n exists.

In addition, our approach can be extended to handle more general models than mean or variance changes. For example, it might be used to detect change points in distribution or regression functions. Our approach might also be applied to multivariate data, as in Matteson and James (2014), or, under certain regularity conditions, to high-dimensional data, as in Wang and Samworth (2018). A rough idea is to define a criterion that is the minimum of the signal statistics over all components. Note that such a minimum of component-based signal statistics will no longer have a pulse pattern, because we can check that the maximum value of this minimum signal statistic is one. However, the minima near the change points are still zero, which can be used to identify the changes. Our approach may also work for change point detection in functional data, mentioned in Berkes et al. (2009). In addition to the component-based method mentioned above, another possible way is to use projected variables.

Finally, a limitation of the proposed method is that it suffers from the short-spacing problem. More specifically, in our criterion, the segment length is larger than $3\alpha_n$, where $\alpha_n = O(n^{0.6})$. When the spacing between two change points is shorter than $3\alpha_n$, change points within the segment cannot be identified. A possible improvement of our method may be to incorporate the work of Fryzlewicz (2014, 2020), if we can derive the relevant asymptotic properties. This is left to further research.

Supplementary Material

Contain the brief description of the online supplementary materials.

Acknowledgments

The authors gratefully acknowledge the grants from the University Grants Council of Hong Kong (HKBU12303419), China Postdoctoral Science Foundation (2020M683456 and 2021T140547), and the National Natural Science Foundation of China (61877049). The authors also thank the editor, associate editor, and two referees for their generous help. In particular, we thank the referee who suggested using the logarithm of the variance to construct the signal statistic instead of using the variance itself.

References

Bauer, P. and Hackl, P. (1980). An extension of the MOSUM technique for quality control. *Technometrics* **22**, 1–7.

Berkes, I., Gabrys, R. Horvath, L. and Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 927–946.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2020). Narrowest-over-threshold detection of multiple change-points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**, 649–672.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300.

Bredel, M., Bredel, C., Juric, D., Harsh, G., Vogel, H., Recht, L. et al. (2005). High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Research* **65**, 4088–4096.

Cao, H. and Wu, W. B. (2015). Changepoint estimation: Another look at multiple testing problem. *Biometrika* **102**, 974–980.

Chu, C., Hornik, K. and Kuan, C. (1995). MOSUM tests for parameter constancy. *Biometrika* **82**, 603–617.

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli* **24**, 526–564.

Fang, X., Li, J. and Siegmund, D. (2020). Segmentation and estimation of change-point models: False postive control and confidence regions. *The Annals of Statistics* **48**, 1615–1647.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 495–580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* **42**, 2243–2281.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society* **49**, 1027–1070.

Gao, Z., Shang, Z., Du, P. and Robertson, J. L. (2019). Variance change point detection under a smoothly-changing mean trend with application to liver procurement. *Journal of the American Statistical Association* **114**, 773–781.

Harchaoui, Z. and Levy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association* **105**, 1480–1493.

Hao, N., Niu, Y. S. and Zhang, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statistica Sinica* **23**, 1553–1572.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**, 1590–1598.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* **109**, 334–345.

Niu, Y. S., Hao, N. and Zhang, H. (2016). Multiple change-point detection: A selective overview. *Statistical Science* **31**, 611–623.

Muggeo, V. M. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* **27**, 161–166.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100–115.

Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *The Annals of Statistics* **15**, 749–779.

Qu, Z. and Perron. P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica* **75**, 459–502.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30**, 507–512.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association* **105**, 1228–1240.

Tewes, J. (2017). Change-point tests under local alternatives for long-range dependent processes. *Electronic Journal of Statistics* **11**, 2461–2498.

Vostrikova, L. J. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathematics Doklady* **24**, 55–59.

Wang, T. and Samworth, R.J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 57–83.

Wang, D., Yu, Y. and Rinaldo, A. (2020). Univariate mean change point detection: Penalization, CUSUM and optimality. *Electronic Journal of Statistics* **14**, 1917–1961.

Wu, W. B. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 391–410.

Yao, Y, C.(1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters* **6**, 181–189.

Yao, Y, C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics, Series A* **51**, 370–381.

Wenbiao Zhao

School of Statistics, Renmin University, Beijing, China.

E-mail: wbzhao96@163.com

Xuehu Zhu

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

E-mail: zhuxuehu@xjtu.edu.cn

Lixing Zhu

School of Statistics, Beijing Normal University, Beijing, China.

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.

E-mail: lzhu@hkbu.edu.hk

(Received November 2020; accepted June 2021)