Electronic Journal of Statistics
Vol. 14 (2020) 1917–1961
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1710

Univariate mean change point detection: Penalization, CUSUM and optimality

Daren Wang

Department of Statistics, University of Chicago, 5747 S. Ellis Avenue, Jones 120A, Chicago, IL 60637 U.S.A.
e-mail: darenw@galton.uchicago.edu

Yi Yu

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
e-mail: yi.yu.2@warwick.ac.uk

Alessandro Rinaldo

Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213 U.S.A.
e-mail: arinaldo@cmu.edu

Abstract: The problem of univariate mean change point detection and localization based on a sequence of n independent observations with piecewise constant means has been intensively studied for more than half century, and serves as a blueprint for change point problems in more complex settings. We provide a complete characterization of this classical problem in a general framework in which the upper bound σ^2 on the noise variance, the minimal spacing Δ between two consecutive change points and the minimal magnitude κ of the changes, are allowed to vary with n. We first show that consistent localization of the change points is impossible in the low signal-to-noise ratio regime $\frac{\kappa\sqrt{\Delta}}{\sigma} \preceq \sqrt{\log(n)}$. In contrast, when $\frac{\kappa\sqrt{\Delta}}{\sigma}$ diverges with n at the rate of at least $\sqrt{\log(n)}$, we demonstrate that two computationally-efficient change point estimators, one based on the solution to an ℓ_0-penalized least squares problem and the other on the popular wild binary segmentation algorithm, are both consistent and achieve a localization rate of the order $\frac{\sigma^2}{\kappa^2}\log(n)$. We further show that such rate is minimax optimal, up to a $\log(n)$ term.

Keywords and phrases: Change point detection, minimax optimality, ℓ_0-penalization, CUSUM statistics, binary segmentation.

Received June 2019.

Contents

1 Introduction 1918
2 Phase transition and optimality minimax rates 1922
3 ℓ_0 penalization 1924
3.1 Optimal change point localization 1926
4 CUSUM 1928
5 Conclusions 1933

6 Acknowledgments 1934
A Proofs of the results in Section 2 1934
B Proofs of the Results in Section 3 1936
C Proofs of the results in Section 4 1948
C.1 Large probability events 1948
C.2 Technical details for Step 1 1949
C.3 Technical details for Step 2 1950
References 1959

1. Introduction

Research on change point detection in time series data has a relatively long history in modern statistics, covering both online (e.g. Wald, 1945; Page, 1954; James, James and Siegmund, 1987) and offline (e.g. Vostrikova, 1981; Yao and Au, 1989) search problems. It has been recently going through a renaissance due to the routinely collected complex and large amount of data sets in the 'Big Data' era. Change point detection problems in high-dimensional means (e.g. Cho and Fryzlewicz, 2015; Cho, 2015; Aston and Kirch, 2014; Jirak, 2015; Wang and Samworth, 2018), in covariance structures (e.g. Aue et al., 2009; Avanesov and Buzun, 2016; Wang, Yu and Rinaldo, 2017), in dynamic networks (e.g. Gibberd and Roy, 2017; Wang, Yu and Rinaldo, 2018), and in sequentially-correlated time series (e.g. Lavielle, 1999; Davis, Lee and Rodriguez-Yam, 2006; Aue et al., 2009) have been actively studied in recent years.

Arguably, the simplest and best-studied change point detection problem is on univariate mean from independent observations. It is fair to say that this is the most important ingredient in more complex problems. We formalize the model in Assumption 1.

Assumption 1 (Model)**.** *Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be independent sub-Gaussian random variables with continuous density such that $\mathbb{E}(Y_i) = f_i$ and $\|Y_i - f_i\|_{\psi_2} \leq \sigma$[1] for all $i \in \{1, \ldots, n\}$.*

Let $\{\eta_k\}_{k=0}^{K+1} \subset \{1, \ldots, n+1\}$ be a collection of change points such that $1 = \eta_0 < \eta_1 < \ldots < \eta_K \leq n < \eta_{K+1} = n+1$ and

$$f_t \neq f_{t-1} \quad \text{if and only if} \quad t \in \{\eta_1, \ldots, \eta_K\}.$$

Let the minimal spacing Δ and the jump size κ be

$$\min_{k=1,\ldots,K+1} \{\eta_k - \eta_{k-1}\} = \Delta$$

and

$$\min_{k=1,\ldots,K} |f_{\eta_k} - f_{\eta_k - 1}| = \min_{k=1,\ldots,K} \kappa_k = \kappa.$$

Assume $\Delta > 0$ and $\kappa > 0$.

[1]For any random variable X, let $\|X\|_{\psi_2}$ be its Orlicz-ψ_2 norm, i.e.

$$\|X\|_{\psi_2} = \inf \{t > 0; \mathbb{E} \exp(X^2/t^2) \leq 2\}.$$

See e.g. Definition 2.5.6. in Vershynin (2010).

Remark 1. *We do not need the condition that Y_i's have continuous densities. We include it here for simplicity, so that the event that two sets of independent random variables have the same sample mean has probability zero. This is the only time this condition is used.*

The model is completely characterized by the sample size n, the upper bound σ on the random fluctuations in terms of Orlicz-ψ_2-norm, the minimal spacing Δ between two consecutive change points and the lower bound κ of the jump size in terms of the absolute value of the difference between two consecutive population means. All three parameters σ, Δ and κ are allowed to change as n grows. Since the number of change points K is upper bounded by n/Δ, we will not keep track of K.

The goal of a change point detection problem is to obtain **consistent** change point estimators $\{\hat{\eta}_k\}_{k=1}^{\widehat{K}}$, with $\widehat{\eta}_1 < \ldots < \widehat{\eta}_{\widehat{K}}$, such that

$$\widehat{K} = K \quad \text{and} \quad \max_{k=1,\ldots,\widehat{K}} \left|\hat{\eta}_k - \eta_k\right| \leq \epsilon(n) = \epsilon, \tag{1}$$

where $\epsilon/\Delta \to 0$, with probability tending to 1 as $n \to \infty$. In the rest of the paper, we will refer to ϵ as the **localization error** and to the sequence $\{\epsilon/\Delta\}$ as the **localization rate**. Notice that the inequality in (1) can be seen as providing an upper bound on the bidirectional Hausdorff distance between $\{\eta_k\}_{k=1}^{K}$ and $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$, both viewed as subsets of $\{2, \ldots, n\}$; see (4) below.

In order to quantify the difficulty of the problem, we rely on the quantity

$$\kappa\sqrt{\Delta}/\sigma, \tag{2}$$

which can be thought of as measuring the **signal-to-noise ratio**. As we will see, the intrinsic statistical hardness of the change point detection and localization problems is fully captured by this quantity. In particular, the difficulty of the problem increases as κ and Δ decrease, and σ increases. The quantity (2) is rooted in two-sample mean testing (with common and known variance), resembling z-statistics used therein, and has counterparts in high-dimensional mean, covariance and network change point detection problems (e.g. Wang and Samworth, 2018; Wang, Yu and Rinaldo, 2017, 2018).

With the previously defined localization rate and signal-to-noise ratio, the optimality of the estimators possesses two aspects.

(i) Consistency. The first natural question one might ask is under what conditions localization is itself possible. We tackle this problem by identifying combinations of the model parameters, which we express using the signal-to-noise ratio (2), for which no estimator of the change points is guaranteed to be consistent, in a minimax sense.

(ii) Outside the region of impossibility identified in the previous step, the second natural question is to derive a lower bound on the localization rate that holds for any estimator. Once the information-theoretic lower bound is established, one may then proceed to demonstrate a computationally-efficient algorithm whose localization rate matches such lower bound. This algorithm is therefore minimax optimal.

We would like to point out that the phase transition phenomenon in terms of signal-to-noise ratio for the localization that we demonstrate below in Section 2 has been shown previously found in the literature. For instance, Theorem 1 in Chan and Walther (2013) showed a phase transition for testing the presence of a single change point that matches the one we obtain for localization. Frick, Munk and Sieling (2014) have further generalized this type of detection results to allow for an unbounded number of change points. Having said these, we would like to emphasize that testing and localization are two different statistical inference tasks, despite connection. In term of localization rate, Theorem 2.8 in Frick, Munk and Sieling (2014) has also provided a localization rate that match the minimax rate we derive in this paper. Similar results can also be found in other papers including Dümbgen and Spokoiny (2001), Dümbgen and Walther (2008), Li, Guo and Munk (2017), Jeng, Cai and Li (2012), Enikeeva, Munk and Werner (2018), to name but a few.

In this article we will be focusing on two types of change point estimators, one based on penalized least squares and the other on CUSUM statistics. Both types of estimator have been thoroughly studied.

- There exist several results and algorithms for change point detection using ℓ_0 penalization, including Liebscher and Winkler (1999), Friedrich et al. (2008), Boysen et al. (2009) and Killick, Fearnhead and Eckley (2012). It is worth comparing three papers providing theoretical results based on ℓ_0-penalization methods. Lavielle and Moulines (2000) studied the ℓ_0-penalization approach under general distributions, and showed that if one chooses the penalization parameter λ properly, then one would get similar asymptotic results to the case where the model assumes Gaussian noise. The closest-related result there is Theorem 9, which only showed asymptotic results. In this paper, we obtain the lower bounds based on Gaussian noise, but the upper bounds are achieved for sub-Gaussian noise, and provide non-asymptotic results. Boysen et al. (2009) studied consistent estimation of a general class of functions based on the solution to an ℓ_0 least squares problem (see (7) in Section 3 for details), which they referred to as the Potts functional. In particular, under the assumption that the mean function is piecewise-constant with a fixed number of change points, the authors showed that a solution to the Potts functional can consistently localize the change points if the minimal spacing satisfies $\Delta = cn$ for some $0 < c < 1$ and the change size κ is a constant. We extend such results by allowing all the parameters in the model – namely κ, Δ and σ – to change with n at a nearly minimax rate, and will demonstrate the existence of a phase transition in the space of model parameters. Furthermore, our analysis is non-asymptotic. Fan and Guan (2017) studied the ℓ_0-denoising on a general class of graphs including chains, i.e. piecewise-constant time series signals, and provided a number of information-theoretic results. Our paper and theirs have different targets – we focus on the change point localization but theirs focused on prediction, which are complementary to each other.

There are also a number of papers in 1980's studying the univariate mean change point detection problem from the least squares estimators perspective, for instance, Yao and Davis (1986), Yao (1988), Yao and Au (1989). The change point estimators are derived from least squares estimators, and the number of change points are chosen via the Schwarz information criterion. It can be shown (e.g. Tickle et al., 2018) that the Schwarz information criterion is asymptotically equivalent to the ℓ_0 penalization. Note that the results obtained there are asymptotic, while ours are non-asymptotic and allow all parameters to vary as the sample size n. Another related area is the reduced isotonic regression problem, which assumes the monotonic signal is piecewise-constant and which aims to recover the signal. Gao, Han and Zhang (2017) has shown an iterated logarithmic lower bound when there are multiple change points. Despite the close connection, the focus and results thereof are different from ours.

It is worth mentioning that ℓ_0-penalization method is appealing from the computational aspect, at least in the univariate case. Friedrich et al. (2008) showed that the Potts functional can be computed using dynamic programming and its computational cost is of order $O(n^2)$. Killick, Fearnhead and Eckley (2012) introduced the pruned exact linear time (PELT) method, which has the worst case computational cost of order $O(n^2)$; while in the situations where the number of change points increases linearly with n, the expected time of PELT is of order $O(n)$. There are also other algorithms, including Rigaill (2010) and Maidstone et al. (2017), which have been shown to have an expected cost which is smaller than that of PELT, but which have the worst case cost also of order $O(n^2)$.

- The CUSUM (see Definition 1 in Section 4) is short for the cumulative sums, proposed in Page (1954) for an online change point problem, and has been a cornerstone in numerous change point detection methods. We will show in Section 4 that in the univariate situation, it is identical to the likelihood ratio test statistics to test whether or not there exists a change point. Binary segmentation (BS) (e.g. Scott and Knott, 1974; Vostrikova, 1981) based on CUSUM statistics has been shown to be consistent, yet optimal, in localizing the change points. In the last few years, a considerable amount of efforts have been made into developing variants of BS in order to handle multiple change points scenarios, see e.g. Fryzlewicz (2014), Baranowski, Chen and Fryzlewicz (2016) and Eichinger and Kirch (2018).

  An important reference is Fryzlewicz (2014), who put forward the wild binary segmentation (WBS) algorithm, a variant of BS, and provided an analysis of its performance. Unfortunately, the proof of Theorem 3.2 in that reference suffers from critical errors. In this paper we rectify those issues and present a more comprehensive analysis of WBS that keeps track explicitly of all the relevant parameters and, in particular, allows to conclude that the localization rate afforded by WBS is nearly minimax rate optimal. Although our efforts in this regard are non-trivial, we acknowledge that the results we derive in Section 4 and the proofs in Appendix C

borrow heavily from Fryzlewicz (2014). As a result, we provide optimal results with all parameters being allowed to change with n and weaker conditions.

The univariate mean change point detection problem has been studied intensively, and we are aware that the results in this paper have been produced in different forms in existing literature. However, we still see the need to produce this paper merely focusing on this simple scenario, providing systematical analysis on various theoretical points, which can be served as benchmarks in more modern challenges.

We summarize our contributions as follows.

(i) We describe a phase transition in the space of the model parameter that separates parameter combinations for which consistent change point estimation is impossible (in a minimax sense) from those for which there exist algorithms that are provably consistent. Furthermore, we provide a global information-theoretic lower bound on the localization rate that holds over most of the region of the parameter space for which consistent estimation is possible. It is worth pointing out that this same phrase transition could be indirectly deduced from the existing literature on minimax change point detection and on change point localization for univariate piecewise signals. See, in particular, Chan and Walther (2013) and Frick, Munk and Sieling (2014). Here we provide a direct proof of this phenomenon based on formal minimax arguments.

(ii) We demonstrate that the ℓ_0-penalization method produces a minimax rate-optimal estimator of the change points. In addition, we demonstrate that the localization rate of ℓ_0-penalization method is locally adaptive to the jump size at each change point, a desirable feature both in theory and in practice (see Remark 5).

(iii) Among CUSUM-based methods, we show that the WBS algorithm (Fryzlewicz, 2014) is also minimax rate-optimal. While our analysis of the WBS is heavily inspired by the proof techniques in Fryzlewicz (2014), we are able to provide more refined results with optimal tracking of the underlying parameters, thus obtaining optimal rates. We also require weaker conditions than Fryzlewicz (2014).

The paper is organized as follows. The information-theoretic results are exhibited in Section 2. Matching upper bounds provided by en ℓ_0-penalization method and WBS can be found in Sections 3 and 4, respectively. Most of the proofs and technical details are in the Appendices.

2. Phase transition and optimality minimax rates

Recall the two aspects of optimality we describe in Section 1: to identify parameter combinations for which consistent localization is possible and to determine a minimax lower bound on the localization rate. In Lemma 1 we describe the low signal-to-noise ratio regime for which estimating the location of the change

points cannot be done. In detail, we show that if

$$\kappa\sqrt{\Delta}/\sigma < \sqrt{\log(n)}, \tag{3}$$

then no consistent estimator of the locations of the change points exists. On the other hand, when $\kappa\sqrt{\Delta}/\sigma \geq \sqrt{\log(n)}$, Lemma 2 demonstrates a minimax lower bound on the localization rate of the form $\frac{\sigma^2}{\kappa^2\Delta}$, for all n large enough. The analysis of the localization procedures described in Sections 3 and 4 will confirm that these results are in fact quite sharp. Specifically, we will verify both the existence of a phase transition for the localization task as the signal-to-noise ratio crosses the threshold $\sqrt{\log(n)}$, as prescribed by Lemma 1, and the near minimax optimality of the lower bound of Lemma 2.

Below, for two subsets E_1 and E_2 of $\{1, \ldots, n\}$, we let

$$H(E_1, E_2) = \max\left\{ \max_{x\in E_1} \min_{y\in E_2} |x-y|, \max_{y\in E_1} \min_{x\in E_2} |x-y| \right\} \tag{4}$$

denote their bidirectional Hausdorff distance.

Lemma 1. *Let $\{Y_i\}_{i=1}^n$ be a time series satisfying Assumption 1 and let $P^n_{\kappa,\Delta,\sigma}$ denote the corresponding joint distribution. For any $0 < c < 1$, consider the class of distributions*

$$\mathcal{P}^n_c = \left\{ P^n_{\kappa,\Delta,\sigma} : \Delta = \min\left\{ \left\lfloor c\frac{\log(n)}{\kappa^2/\sigma^2} \right\rfloor, \left\lfloor \frac{n}{4} \right\rfloor \right\} \right\}.$$

Then, there exists an $n(c)$, which depends on c, such that, for all n larger than $n(c)$,

$$\inf_{\widehat{\eta}} \sup_{P\in\mathcal{P}^n_c} \mathbb{E}_P\big(H(\widehat{\eta}, \eta(P))\big) \geq \frac{n}{8} \geq \frac{\Delta}{4},$$

where the infimum is over all estimators $\widehat{\eta} = \{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ of the change point locations and $\eta(P)$ is the set of locations of the change points of $P \in \mathcal{P}^n_c$.

In the above result, it is possible to let $c \to 0$ as $n \to \infty$ (and in fact, the value of $n(c)$ is increasing in c). Thus, we conclude that, if $\kappa\sqrt{\Delta}/\sigma < \lfloor\sqrt{\log(n)}\rfloor < \lfloor n/4\rfloor$, the localization rate is bounded away from 0, i.e. the estimator is not consistent.

In our next result we complement Lemma 1 by showing that if instead

$$\kappa\sqrt{\Delta}/\sigma \geq \zeta_n, \tag{5}$$

for any sequence $\{\zeta_n\}_{n=1,2,\ldots}$ of positive numbers diverging to infinity at an arbitrary pace as $n \to \infty$, then the corresponding lower bound is at least of order $\frac{\sigma^2}{\kappa^2}$, for all n large enough. Of course, in light of Lemma 1, this lower bound is interesting only when ζ_n is larger than $\sqrt{\log(n)}$. In the next sections, we will further show that, provided that ζ_n is of the order $\sqrt{\log^{1+\xi}(n)}$ or larger, for any $\xi > 0$, then, up to a logarithmic factor in n, $\frac{\sigma^2}{\kappa^2}$ yields the asymptotic minimax lower bound on the localization rate.

Lemma 2. *Let $\{Y_i\}_{i=1}^n$ be a time series satisfying Assumption 1 with one and only one change point. Let $P^n_{\kappa,\Delta,\sigma}$ denote the corresponding joint distribution. Consider the class of distributions*

$$\mathcal{Q}^n = \left\{ P^n_{\kappa,\Delta,\sigma} : \Delta < n/2,\, \kappa\sqrt{\Delta}/\sigma \geq \zeta_n \right\},$$

for any sequence $\{\zeta_n\}$ such that $\lim_{n\to\infty} \zeta_n = \infty$. Then, for all n large enough, it holds that

$$\inf_{\hat{\eta}} \sup_{P\in\mathcal{Q}^n} \mathbb{E}_P\big(|\hat{\eta} - \eta(P)|\big) \geq \max\left\{1, \frac{1}{2}\left\lceil \frac{\sigma^2}{\kappa^2} \right\rceil e^{-2}\right\},$$

where the infimum is over all estimators $\widehat{\eta}$ of the change point location and $\eta(P)$ denotes the change point location of $P \in \mathcal{Q}^n$.

The bounds in Lemma 1 and Lemma 2 are slightly sharper than the minimax lower bounds obtained by taking $p = 1$ in Proposition 3 in the supplementary material of Wang and Samworth (2018). Indeed, our analysis allows for a more refined characterization of the phase transition for the localization task by exhibiting the threshold value of $\sqrt{\log(n)}$ describing the transition from the low to high signal-to-noise ratio regime.

3. ℓ_0 penalization

In this section we describe an estimator of the change point locations based on the ℓ_0 penalty and demonstrate that it is minimax rate optimal.

We first formalize the ℓ_0-penalized optimization problem, and define the change point estimators generated therefrom. For the sake of analysis, we will provide an alternative objective function, which, we will show, generates identical change point estimators.

For fixed tuning parameter $\lambda > 0$ and data $\{Y_i\}_{i=1}^n$, define the ℓ_0-penalized sum of squares objective function as

$$H(u, \{Y_i\}_{i=1}^n, \lambda) = \sum_{i=1}^n \big(Y_i - u_i\big)^2 + \lambda \|Du\|_0, \tag{6}$$

where $\|\cdot\|_0$ is the ℓ_0-norm of a vector, $D \in \{\pm 1, 0\}^{(n-1)\times n}$ satisfies $(Du)_j = u_{j+1} - u_j$, for $j \in \{1, \ldots, n-1\}$. Let

$$\widehat{u}(\lambda) = \operatorname*{arg\,min}_{u\in\mathbb{R}^n} H(u, \{Y_i\}_{i=1}^n, \lambda). \tag{7}$$

Let $\big\{\hat{\eta}_k\big\}_{k=1}^{\widehat{K}}$ be the collection

$$\mathcal{J}\big(\widehat{u}\big) = \big\{i \in \{2, \ldots, n\} : \widehat{u}_i\big(\lambda\big) \neq \widehat{u}_{i-1}\big(\lambda\big)\big\}.$$

We thus call $\big\{\hat{\eta}_k\big\}_{k=1}^{\widehat{K}}$ the change point estimator induced by $\widehat{u}\big(\lambda\big)$, or the change point estimator from the optimization problem (7). If one replaces the penalty

term $\|Du\|_0$ with the ℓ_1-norm $\|Du\|_1$, then (6) is the fused Lasso objective function, see e.g. Tibshirani et al. (2005) and Rinaldo (2009).

Alternatively, let $\mathcal{P}$ be any *interval partition* of $\{1,\dots,n\}$, i.e. a collection of $\mathcal{P}_k$ disjoint subsets of $\{1,\dots,n\}$ of the form

$$\mathcal{P} = \big\{\{1,\dots,i_1-1\},\{i_1,\dots,i_2-1\},\dots,\{i_{\mathcal{P}_k},\dots,n\}\big\},$$

for some integers $1 < i_1 < \cdots < i_{\mathcal{P}_k} \le n$, where $\mathcal{P}_k \ge 1$. In particular, if $\mathcal{P}_k = 1$, then $\mathcal{P} = \big\{\{1,\dots,n\}\big\}$. For a fixed positive tuning parameter $\lambda > 0$ and data $\{Y_i\}_{i=1}^n$, let

$$\widehat{\mathcal{P}}(\lambda) = \arg\min_{\mathcal{P}} G\big(\mathcal{P}, \{Y_i\}_{i=1}^n, \lambda\big). \tag{8}$$

where the minimum ranges over all interval partitions of $\{1,\dots,n+1\}$ and, for any such partition $\mathcal{P}$,

$$G\big(\mathcal{P}, \{Y_i\}_{i=1}^n, \lambda\big) = \sum_{I\in\mathcal{P}}\sum_{i\in I}\big(Y_i - \overline{Y}_I\big)^2 + \lambda\big(|\mathcal{P}| - 1\big), \tag{9}$$

with $\overline{Y}_I = |I|^{-1}\sum_{i\in I} Y_i$. The optimization problem (8) is known as the minimal partition problem and can be solved using dynamic programming in polynomial time (e.g. Algorithm 1 in Friedrich et al., 2008). The change point estimator resulting from the solution to (8) is simply obtained from taking all the left endpoints of the intervals $I \in \widehat{\mathcal{P}}$ except 1. In general, without assuming any conditions on the inputs, there is no guarantee that the minimizers are unique.

We now make the simple observation that the optimization problems (7) and (8) with the same inputs yield the same change point estimators. To see this equivalence we will introduce some notation that we will be using throughout. For any vector $v \in \mathbb{R}^n$ and any $i \in \{2,\dots,n\}$, if $v_i \ne v_{i-1}$, one calls i an induced change point of v, and the collection of all the induced change points of v is denoted as $J(v)$. The set $J(v)$ yields an interval partition, i.e., if $J(v) = \{i_1,\dots,i_N\}$, then one can define the interval partition induced by v as

$$\mathcal{P} = \big\{\{1,\dots,i_1-1\},\{i_1,\dots,i_2-1\},\dots,\{i_N,\dots,n\}\big\}.$$

Conversely, for any interval partition $\mathcal{P}$ and a sequence $\{Y_i\}_{i=1}^n$, define their induced piecewise-constant vector v as $v_i = \overline{Y}_I$, for any $i \in I$ and $I \in \mathcal{P}$. Since for $I \subset \{1,\dots,n\}$,

$$\overline{Y}_I = \arg\min_{x\in\mathbb{R}} \sum_{i\in I}(Y_i - x)^2,$$

it follows that with the same inputs $\{Y_i\}_{i=1}^n$ and $\lambda > 0$, the solutions to (7) and (8) induce each other in the sense specified above.

Remark 2 (Tuning parameter)**.** *If we view any vector $u \in \mathbb{R}^n$ as a step function with at most $n-1$ jumps, then the tuning parameter λ penalizes the number of jumps in u. For an integer interval $I \subset \{1,\dots,n\}$, the tuning parameter λ works*

in the following way. If an integer interval I is to be split into two integer sub-intervals I_1 and I_2, then it follows from Lemma 5 that the sum of squares will decrease by

$$\frac{|I_1||I_2|}{|I_1|+|I_2|}\big(\overline{Y}_{I_1}-\overline{Y}_{I_2}\big)^2, \tag{10}$$

but, at the same time, the penalty term will increase by λ. Therefore the trade-off guiding the choice between refining a candidate integral partition of $\{1,\ldots,n\}$ by introducing one additional split and leaving it unchanged (so that this partition must then provide an optimal solution to (8)), is described by comparing (10) to λ. In Theorem 3 we will provide a theoretically optimal choice for λ.

Remark 3. *In the rest of this paper, when there is no ambiguity, we allow the following abuse of notation. If $s<e$, $s,e\in\mathbb{Z}$, we sometimes refer $\{s,\ldots,e\}$ and $\{s,\ldots,e-1\}$ as $[s,e]$ and $[s,e)$, respectively.*

3.1. Optimal change point localization

Recall in Lemma 1 we have shown that if $\kappa\sqrt{\Delta}/\sigma<\sqrt{\log(n)}$, then no algorithm is guaranteed to produce consistent change point estimators. To demonstrate the performances of (7), we thus require the signal-to-noise ratio $\kappa\sqrt{\Delta}/\sigma$ to be larger than a diverging function of n, which we take to be of the form $\log^{(1+\xi)/2}(n)$. As remarked in the previous section, such choice is consistent with Lemma 2, which in principle allows for a vanishing localization rate.

Assumption 2. *There exists a sufficiently large absolute constant $C_{\mathrm{SNR}}>0$ such that for any $\xi>0$,*

$$\kappa\sqrt{\Delta}/\sigma\geq C_{\mathrm{SNR}}\sqrt{\log^{1+\xi}(n)}.$$

We introduce the parameter $\xi>0$ in order to guarantee that even if $\Delta\asymp n$, the resulting estimator remains consistent. We do not know whether the above assumption can be relaxed by allowing for a rate of increase for $\kappa\sqrt{\Delta}/\sigma$ slower than $\sqrt{\log^{1+\xi}(n)}$. In our proofs, this seems to be the slowest rate that we can afford.

Theorem 3. *Let $\{Y_i\}_{i=1}^n$ be generated from a distribution described in Assumption 1 and, for any $\lambda>0$, set*

$$\widehat{u}(\lambda)=\operatorname*{arg\,min}_{u\in\mathbb{R}^n}H(u,\{Y_i\}_{i=1}^n,\lambda). \tag{11}$$

Let $\{\widehat{v}_k(\lambda)\}_{k=1}^{\widehat{K}(\lambda)}$ be the collection of change points induced by $\widehat{u}(\lambda)$. Under Assumption 2, for any choice of $c>3$, there exists a constant $C_\lambda>0$, which depends on c such that, for $\lambda=C_\lambda\sigma^2\log(n)$, it holds that

$$\mathbb{P}\big\{\widehat{K}(\lambda)=K,\ \text{and } |\widehat{\nu}_k(\lambda)-\nu_k|=\epsilon_k\leq C_\epsilon\sigma^2\log(n)/\kappa_k^2,\ \forall k\in\{1,\ldots,K\}\big\}\geq 1-e\cdot n^{3-c},$$

where $C_\epsilon>0$ is a constant depending on C_λ and C_{SNR}.

Recalling Lemma 2, we see now that the localization error bound we derived in Theorem 3 is minimax rate optimal aside from possibly a $\log(n)$ factor. Theorem 3 shows that with a proper choice of the tuning parameter, (7) provides consistent change point estimators in the sense that with probability tending to 1 as $n \to \infty$, it holds that $\widehat{K}(\lambda) = K$ and for all $k \in \{1, \dots, K\}$,

$$\epsilon_k / n \le C_\epsilon \frac{\sigma^2}{\kappa^2} \frac{\log(n)}{n} \le C_\epsilon \frac{\Delta}{n \log^\xi(n)} \to 0,$$

as $n \to \infty$. It is important to realize that Theorem 3 yields a *family of rates*, depending on how the parameters κ, Δ and σ scale with n. The logarithmic upper bound on the localization rate exhibited in the last display corresponds to the worst case scenario in which the weakest possible signal-to-noise ratio compatible with Assumption 2 is in effect. In fact, in most other cases, the localization rate is significantly faster. For instance, when σ and κ are constants (the classic scenario studied in the earlier change point detection literature), then the localization error is, with high probability, of order $\log(n)$ provided that Δ scales in n at a rate at least as fast as $\log^{1+\xi}(n)$. As a second example, assuming again σ to be constant, the estimator we study remains consistent even as κ vanishes, as long as its rate of decay is no faster that $\sqrt{\frac{\log(n)}{n}}$ and Δ satisfies the high signal-to-noise ratio condition in Assumption 2.

Remark 4 (Uniqueness)**.** *We mentioned earlier that the minimizers of the optimization problems* (7) *and* (8) *need not be unique. However, if the independent errors have a continuous distribution, as assumed in Assumption 1, the minimizer is unique almost surely, for each n and each λ; if not, then any two solutions, say $\mathcal{P}$ and $\mathcal{P}'$, are such that*

$$\sum_{I \in \mathcal{P}} \sum_{i \in I} \big(Y_i - \overline{Y}_I\big)^2 - \sum_{I' \in \mathcal{P}'} \sum_{i \in I'} \big(Y_i - \overline{Y}_{I'}\big)^2 = \lambda\big(|\mathcal{P}'| - \mathcal{P}|\big).$$

This is a quadratic polynomial in the $\{Y_i\}_{i=1}^n$. The set of its real solutions (if exists any) has n-dimensional Lebesgue measure 0. In general, if there are multiple solutions (so that Assumption 1 does not hold), Assumption 2 guarantees that, with large probability, the minimizer is unique almost surely.

Remark 5. *Theorem 3 provides a separate localization error for each change point. It is natural to expect that change point localization should be locally adaptive in a sense that, if the jump size κ_k gets larger, then it is easier to estimate the location of the change point η_k. In fact, the error rate ϵ_k derived in Theorem 3 matches this feature.*

Proof of Theorem 3. Define the event

$$\mathcal{B} = \Bigg\{ \max_{1 \le a < b < c \le n} \sqrt{\frac{(b-a)(c-b)}{c-a}} \big| \overline{Y}_{(a+1,b]} - \overline{f}_{(a+1,b]} + \overline{Y}_{(b+1,c]} - \overline{f}_{(b+1,c]} \big| \le \sigma \sqrt{C_\lambda \log(n)} \Bigg\},$$

where $C_\lambda > 0$ is a large enough constant only depending on c, and a, b, c are integers. In the remainder of the proof we work on the event $\mathcal{B}$. By Lemma 6 in Appendix B, this occurs with probability at least $1 - e \cdot n^{3-c}$.

For simplicity we will remove the dependence on λ in our notation as it will implicitly understood that $\lambda = C_\lambda \sigma^2 \log(n)$. Let $\widehat{\mathcal{P}}$ be the interval partition induced by $\widehat{u}$ (see (11)), and let $\{s, \ldots, e-1\}$ be any member of $\widehat{\mathcal{P}}$. The proof is completed by showing the following four steps.

Step 1 The interval $[s, e)$ contains no more than two true change points. This is shown in Lemma 7.

Step 2 If $[s, e)$ contains exactly two true change points, say η_k, η_{k+1}, then

$$\eta_k - s \leq C_\epsilon \sigma^2 \log(n)/\kappa_k^2, \text{ and } e - \eta_{k+1} \leq C_\epsilon \sigma^2 \log(n)/\kappa_{k+1}^2.$$

This is shown in Lemma 8.

Step 3 If $[s, e)$ contains only one true change point, say η_k, without loss of generality, let $\eta_k - s \leq e - \eta_k$, then it must hold that

$$s \leq \eta_k \leq e \leq \eta_{k+1}$$

and

$$\eta_k - s \leq C_\epsilon \sigma^2 \log(n)/\kappa_k^2, \text{ and } \eta_{k+1} - e \leq C_\epsilon \sigma^2 \log(n)/\kappa_{k+1}^2.$$

This is shown in Lemmas 9, 10 and 11.

Step 4 If $[s, e)$ contains no true change point, then there exist two true change points η_k and η_{k+1} satisfying

$$\eta_k \leq s < e \leq \eta_{k+1}$$

and

$$s - \eta_k \leq C_\epsilon \sigma^2 \log(n)/\kappa_k^2, \text{ and } \eta_{k+1} - e \leq C_\epsilon \sigma^2 \log(n)/\kappa_{k+1}^2.$$

This is shown in Lemma 12. □

4. CUSUM

As for the univariate mean change point detection problem, the ℓ_0-penalization estimator is not the only one which achieves the minimax optimality. Binary segmentation (BS) (e.g. Scott and Knott, 1974) based on CUSUM statistics is arguably the most popular change point detection method. It has been shown that BS is consistent yet not optimal (e.g. Venkatraman, 1992). Fryzlewicz (2014) proposed a variant of BS, namely wild binary segmentation (WBS), which is shown to lead to a better localization rate than the BS algorithm. In this section, we will recall the WBS algorithm, and give refined results on its performance, with a proof which has more careful tracking of all parameters and all the constants involved. As a result, we prove that WBS, just like the method studied in the previous section, also guarantees a localization error rate that is rate minimax optimal. However, compared to the ℓ_0-penalization methods, WBS is computationally more expensive and involves more tuning parameters.

Definition 1 (CUSUM statistics)**.** *For a sequence* $\{Y_i\}_{i=1}^n$*, any pair of time points* $(s, e) \subset \{0, \ldots, n\}$ *with* $s < e-1$*, and any time point* $t = s+1, \ldots, e-1$*, let the CUSUM statistics be*

$$\widetilde{Y}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} Y_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} Y_i.$$

For a collection of independent Gaussian random variables $\{Y_i\}_{i=1}^n$ with $\mathbb{E}(Y_i) = f_i$ and same variance, one can easily derive that

$$\max_{t=1,\ldots,n-1} \left|\widetilde{Y}_t^{0,n}\right|$$

is the generalized likelihood ratio statistic, which can be used to test the hypothesis:

$$\begin{aligned} &H_0 : f_1 = \cdots = f_n \quad \text{vs.} \\ &H_1 : \text{there exists } t_* \text{ such that } f_1 = \cdots = f_{t_*} \neq f_{t_*+1} = \cdots f_n. \end{aligned} \tag{12}$$

In particular, the BS algorithm searches for the time point which has the largest absolute CUSUM statistics value, i.e.

$$\widehat{t} = \operatorname*{arg\,max}_{t=1,\ldots,n-1} \left|\widetilde{Y}_t^{0,n}\right|.$$

However, as noted in Fryzlewicz (2014), when there are potentially multiple change points, their combined effect might cancel out and the BS is guaranteed to be effective only when applied to intervals containing at most one change point. WBS improves on BS by performing multiple CUSUM tests over randomly chosen sub-intervals in such a manner that each change point will, with high probability, be the only change point deep inside some selected interval and can be identified using the BS algorithm within that interval. See Algorithm 1 for a formal description of WBS.

It has been shown under a set of slightly stronger conditions, Fryzlewicz (2014) originally put forward the WBS algorithm and provided an analysis of its performance. Below we refine such analysis and formally prove that WBS is minimax rate-optimal in terms of the required signal-to-noise ratio and the localization rate.

Theorem 4. *For WBS algorithm detailed in Algorithm 1, assume the inputs are as follows:*

- *the sequence* $\{Y_i\}_{i=1}^n$ *satisfies Assumptions 1 and 2;*
- *the collection of intervals* $\{(\alpha_m, \beta_m)\}_{m=1}^M \subset \{1, \ldots, n\}$*, whose endpoints are drawn independently and uniformly from* $\{1, \ldots, n\}$*, satisfy*

$$\max_{m=1,\ldots,M} (\beta_m - \alpha_m) \leq C_R \Delta,$$

almost surely, for an absolute constant $C_R > 3/2$*;*

Algorithm 1 Wild Binary Segmentation. WBS$((s,e),\{(\alpha_m,\beta_m)\}_{m=1}^M,\tau)$

INPUT: Independent samples $\{X_i\}_{i=1}^n$, collection of intervals $\{(\alpha_m,\beta_m)\}_{m=1}^M$, tuning parameters $\tau>0$.
  for $m=1,\ldots,M$ **do**
    $(s_m,e_m)\leftarrow[s,e]\cap[\alpha_m,\beta_m]$
    if $e_m-s_m>1$ **then**
      $b_m\leftarrow\arg\max_{s_m+1\le t\le e_m-1}|\widetilde{Y}_t^{s_m,e_m}|$
      $a_m\leftarrow\left|\widetilde{Y}_{b_m}^{s_m,e_m}\right|$
    else
      $a_m\leftarrow-1$
    end if
  end for
  $m^*\leftarrow\arg\max_{m=1,\ldots,M}a_m$
  if $a_{m^*}>\tau$ **then**
    add b_{m^*} to the set of estimated change points
    WBS$((s,b_{m*}),\{(\alpha_m,\beta_m)\}_{m=1}^M,\tau)$
    WBS$((b_{m*}+1,e),\{(\alpha_m,\beta_m)\}_{m=1}^M,\tau)$
  end if
OUTPUT: The set of estimated change points.

- *the tuning parameters τ satisfies*

$$c_{\tau,1}\sigma\sqrt{\log(n)}<\tau<c_{\tau,2}\kappa\sqrt{\Delta}, \tag{13}$$

  where $c_{\tau,1},c_{\tau,2}>0$ are sufficiently large and small absolute constants.

Let $\{\hat{\eta}_k\}_{k=1}^{\widehat{K}}$ be the corresponding output of the WBS algorithm. Then,

$$\begin{aligned}\mathbb{P}\left\{\widehat{K}=K \quad \text{and} \quad \epsilon_k\le C_\epsilon\sigma^2\log(n)\kappa_k^{-2},\forall k\in\{1,\ldots,K\}\right\}\\ \ge 1-e\cdot n^{3-c}-e\cdot n^{2-c}-\exp\left\{\log\left(\frac{n}{\Delta}\right)-\frac{M\Delta}{4C_Rn}\right\},\end{aligned} \tag{14}$$

where $c>3$ is an absolute constant and $C_\epsilon>0$ is a sufficiently large constant.

Remark 6. *For simplicity, we require Assumption 1 in Theorem 4, but in fact we do not need continuous density functions condition.*

Theorem 4 shows that with suitable choice for the tuning parameters, WBS is optimal in the sense that:

- under the signal-to-noise ratio regime detailed in Assumption 2, it yields consistent estimators of the change point locations that with probability tending to 1: $\widehat{K}=K$, and for all $k=1,\ldots,K$,

$$\epsilon_k/\Delta\le C_\epsilon\sigma^2\log(n)\kappa_k^{-2}/\Delta\le\frac{C_\epsilon}{C_{\mathrm{SNR}}^2}\frac{1}{\log^\xi(n)}\to0,$$

  as $n\to\infty$; and
- it possesses a localization rate

$$\epsilon/\Delta=\Delta^{-1}\max_{k=1,\ldots,K}C_\epsilon\sigma^2\log(n)\kappa_k^{-2}\le\Delta^{-1}C_\epsilon\sigma^2\log(n)\kappa^{-2},$$

which is minimax rate optimal, save for a $\log(n)$ factor, according to Lemma 2.

Remark 7. *To guarantee that* (14) *tends to 1 as* $n \to \infty$*, the number of random intervals* M *needs to satisfy*

$$M \gtrsim \frac{n}{\Delta} \log\left(\frac{n}{\Delta}\right).$$

Remark 8 (Tracking constants)**.** *For readability, we refrain the pursuit of explicitly expressing all constants, and only show the hierarchy of the constants. One would first choose* $c > 3$ *in* (14) *to make sure that the consistency result holds. This will determine* $c_{\tau,1}$*, which is the same as* C_γ *used in the proof, and consequently* $c_{\tau,2}$*, which also depends on* C_{SNR} *and* C_R*. All these constants finally determine* C_ϵ*.*

Remark 9 (Tuning parameters)**.** *The range displayed in Equation* (13) *is used in* **Step 1** *in the proof. Notice that, by Assumption 2 and with properly chosen constants, such range is not an empty set for* τ*. As shown in the proof, over an event of probability tending to* 1*, the lower bound of* (13) *serves as an upper bound of the maximum CUSUM statistics when there are no change points, and the upper bound serves as a lower bound of the maximum CUSUM statistics when there exists a change point.*

Remark 10 (Comparisons with Theorem 3)**.** *In Theorem 3, the only tuning parameter is the penalization level* λ*, while in Theorem 4, we rneed to specify* τ *and the number of random intervals* M*. In practice, Fryzlewicz (2014) suggest an AIC-based method for picking these parameters.*

Proof of Theorem 4. Since ϵ is the desired order of localization error rate, by induction, it suffices to consider any generic interval $(s, e) \subset (0, T)$ that satisfies

$$\eta_{k-1} \le s \le \eta_k \le \ldots \le \eta_{k+q} \le e \le \eta_{k+q+1}, \quad q \ge -1,$$

and

$$\max\{\min\{\eta_k - s,\, s - \eta_{k-1}\},\, \min\{\eta_{k+q+1} - e,\, e - \eta_{k+q}\}\} \le \epsilon,$$

where $q = -1$ indicates that there is no change point contained in (s, e).

Under Assumption 2, it holds that

$$\epsilon = C_\epsilon \sigma^2 \log(n) \kappa^{-2} \le \frac{C_\epsilon}{C_{\mathrm{SNR}}^2} \frac{\Delta}{\log^\xi(n)} \le \Delta/4,$$

with sufficiently large C_{SNR}. It, therefore, has to be the case that for any change point $\eta_k \in (0, T)$, either $|\eta_k - s| \le \epsilon$ or $|\eta_k - s| \ge \Delta - \epsilon \ge 3\Delta/4$. This means that $\min\{|\eta_k - e|, |\eta_k - s|\} \le \epsilon$ indicates that η_k is a detected change point in the previous induction step, even if $\eta_k \in (s, e)$. We refer to $\eta_k \in (s, e)$ as an undetected change point if $\min\{\eta_k - s, \eta_k - e\} \ge 3\Delta/4$.

In order to complete the induction step, it suffices to show that WBS (i) will not detect any new change point in (s, e) if all the change points in that interval

have been previous detected, and (ii) will find a point $b \in (s, e)$ – in fact in $(s + \delta(e - s), e - \delta(e - s))$ – such that $|\eta_k - b| \leq \epsilon$ if there exists at least one undetected change point in (s, e).

We will consider the events $\mathcal{A}_1(\gamma)$, $\mathcal{A}_2(\gamma)$ and $\mathcal{M}$ defined in (40), (41) and (42), respectively. Set γ to be $C_\gamma \sigma \sqrt{\log(n)}$, with a sufficiently large C_γ. The rest of the proof assumes the event

$$\mathcal{A}_1(C_\gamma \sigma \sqrt{\log(n)}) \cap \mathcal{A}_2(C_\gamma \sigma \sqrt{\log(n)}) \cap \mathcal{M}.$$

which, in light of Lemma 13 from Appendix C, has probability tending to 1.

Step 1. In this step, we will show that WBS will consistently detect or reject the existence of undetected change points within (s, e).

Let a_m, b_m and m^* be defined as in Algorithm 1. Suppose there exists a change point $\eta_k \in (s, e)$ such that $\min\{\eta_k - s, e - \eta_k\} \geq 3\Delta/4$. In the event $\mathcal{M}$, there exists an interval (α_m, β_m) selected by WBS such that $\alpha_m \in [\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\beta_m \in [\eta_k + \Delta/2, \eta_k + 3\Delta/4]$.

Following Algorithm 1, $[s_m, e_m] = [\alpha_m, \beta_m] \cap [s, e]$. We have that $\min\{\eta_k - s_m, e_m - \eta_k\} \geq (1/4)\Delta$ and $[s_m, e_m]$ contains at most one true change point.

By choosing $c_1 = 1/2$ in Lemma 14, it holds that

$$\max_{s_m < t < e_m} \left| \widetilde{f}_t^{s_m, e_m} \right| \geq \kappa_k \sqrt{\Delta}/4,$$

where $e_m - s_m \leq 2\Delta$ is used in the last inequality. Therefore

$$a_m = \max_{s_m < t < e_m} \left| \widetilde{Y}_t^{s_m, e_m} \right| \geq \max_{s_m < t < e_m} \left| \widetilde{f}_t^{s_m, e_m} \right| - \gamma \geq \kappa_k \sqrt{\Delta}/4 - \gamma.$$

Thus for any undetected change point $\eta_k \in (s, e)$, it holds that

$$a_{m^*} = \sup_{1 \leq m \leq M} a_m \geq \kappa_k \sqrt{\Delta}/4 - \gamma \geq c_{\tau,2} \kappa_k \sqrt{\Delta}, \tag{15}$$

where the last inequality is from the choice of γ and $c_{\tau,2} > 0$ is achievable with a sufficiently large C_{SNR} in Assumption 1. Then, WBS correctly accepts the existence of undetected change points.

Suppose there does not exist any undetected change point within (s, e), then for any $(s_m, e_m) = (\alpha_m, \beta_m) \cap (s, e)$, one of the following situations must hold.

(a) There is no change point within (s_m, e_m);
(b) there exists only one change point $\eta_k \in (s_m, e_m)$ and either $\min\{\eta_k - s_m, e_m - \eta_k\} \leq \epsilon_k$; or
(c) there exist two change points $\eta_k, \eta_{k+1} \in (s_m, e_m)$ and $\eta_k - s_m \leq \epsilon_k$, $e_m - \eta_{k+1} \leq \epsilon_{k+1}$.

Since cases (a) and (b) are similar and in fact simpler to the case (c), we will only provide the analysis for (c) in the proof. Observe that if (c) holds, by Lemma 15

$$\sup_{s_m \leq t \leq e_m} |\widetilde{f}_t^{s_m, e_m}| \leq \sqrt{e_m - \eta_{k+1}} \kappa_{k+1} + \sqrt{\eta_k - s_m} \kappa_k \leq 2 C_\epsilon \sigma \sqrt{\log(n)}.$$

As a result,

$$\sup_{s_m \le t \le e_m} |\widetilde{Y}_t^{s_m,e_m}| \le \sup_{s_m \le t \le e_m} |\widetilde{f}_t^{s_m,e_m} - \widetilde{Y}_t^{s_m,e_m}| + \sup_{s_m \le t \le e_m} |\widetilde{f}_t^{s_m,e_m}| \le 2C_\epsilon \sigma \sqrt{\log(n)} + C_\gamma \sigma \sqrt{\log(n)} < \tau$$

where event $\mathcal{A}_1(C_\gamma \sigma \sqrt{\log(n)})$ is used in the first inequality and (13) is used in the last inequality. Therefore WBS will always correctly reject the existence of undetected change points.

Step 2. Assume that there exists a change point $\eta_k \in (s, e)$ such that $\min\{\eta_k - s, \eta_k - e\} \ge 3\Delta/4$. Let s_m, e_m and m^* be defined as in Algorithm 1. To complete the proof it suffices to show that there exists a change point $\eta_k \in (s_{m*}, e_{m*})$ such that $\min\{\eta_k - s_{m*}, \eta_k - e_{m*}\} \ge \Delta/4$ and $|b_{m*} - \eta_k| \le \epsilon$.

To that end, we are to ensure that the assumptions of Lemma 22 are verified. The proof of Lemma 22 relies on a number of results, the relationship of which is shown in Figure 2. Observe that (52) is straightforward from Assumption 2, (50) and (51) follow from the definition of $\mathcal{A}_1$ and $\mathcal{A}_2$, and that (49) follows from (15).

Thus, all the conditions in Lemma 22 are met, and we therefore conclude that there exists a change point η_k, satisfying

$$\min\{e_{m^*} - \eta_k, \eta_k - s_{m^*}\} > \Delta/4 \tag{16}$$

and

$$|b_{m*} - \eta_k| \le C_3 \gamma^2 \kappa^{-2} \le \epsilon,$$

where the last inequality holds from the choice of γ and Assumption 2.

The proof is complete by noticing the fact that (16) and $(s_{m^*}, e_{m^*}) \subset (s, e)$ imply that

$$\min\{e - \eta_k, \eta_k - s\} > \Delta/4 > \epsilon.$$

As discussed in the argument before **Step 1**, this implies that η_k must be an undetected change point. □

5. Conclusions

In this paper we have provided a complete characterization of the classical problem of univariate mean change point localization for a sequence of independent sub-Gaussian random variables with piecewise-constant means. We have considered the most general setting in which all the parameters of the problems are allowed to change with the length n of the sequence. We have identified a critical function of the model parameters that is able to discriminate the portion of the parameter space in which consistent localization is impossible from the part in which it is feasible. We have further specified, up to a $\log(n)$ term, the minimax optimal localization rate for this problem and showed that two computationally-efficient methods achieve such a rate.

We would like to point out that the ℓ_0-penalization methods can also be used in handling change point detection for more complex data types, such as high-dimensional mean, covariance and networks. The developments rely on feasible algorithms for their corresponding problems, but we conjecture that ℓ_0-penalization methods on complex data types would also enjoy the same optimality with fewer tuning parameters than those in CUSUM-based methods.

Finally, we conjecture that the upper bounds on the localization rate exhibited in both Sections 3 and 4 can be sharpened by replacing the $\log(n)$ term with a smaller quantity of order $\log\log(n)$, thus further reducing the gap with the lower bound in Lemma 2.

6. Acknowledgments

We thank Zhou Fan, Paul Fearnhead, Rebecca Killick, Housen Lin, Axel Munk, Richard J. Samworth, Ryan Tibshirani and Tengyao Wang for constructive conversations. We also thank the Editor, the Associate Editor and the Referee for helpful comments.

Appendix A: Proofs of the results in Section 2

Proof of Lemma 1. Without loss of generality, suppose that $n/4$ is an integer. For $l \in \{1, \ldots, n/4\}$, let $\widetilde{u}_l \in \mathbb{R}^n$ be such that the ith coordinate of $\widetilde{u}_l(i)$, $i = 1, \ldots, n$, satisfies

$$\widetilde{u}_l(i) = \begin{cases} \sqrt{c\sigma^2 \log(n)}, & i = l; \\ 0, & \text{otherwise}, \end{cases}$$

where $0 < c < 1$. Let $\widetilde{v}_l \in \mathbb{R}^n$ be such that $\widetilde{v}_l(i) = \widetilde{u}_l(n - i + 1)$, $i = 1, \ldots, n$. Let $\widetilde{P}_l$ and $\widetilde{Q}_l$ be the multivariate Gaussian distributions $\mathcal{N}(\widetilde{u}_l, \sigma^2 I_n)$ and $\mathcal{N}(\widetilde{v}_l, \sigma^2 I_n)$, respectively and set

$$\widetilde{P} = \frac{1}{n/4} \sum_{l=1}^{n/4} \widetilde{P}_l \quad \text{and} \quad \widetilde{Q} = \frac{1}{n/4} \sum_{l=1}^{n/4} \widetilde{Q}_l.$$

Note that for each $l \in \{1, \ldots, n/4\}$, $\widetilde{P}_l$ has two change points, at locations $l-1$ and l, and therefore, $\Delta = 1$. Furthermore, the jump size is $\kappa = \sqrt{c\sigma^2 \log(n)}$ and the fluctuation is σ^2. As a result,

$$\kappa\sqrt{\Delta}/\sigma = \sqrt{c \log(n)},$$

which implies that all $\widetilde{P}_l \in \mathcal{P}_c^n$. The same arguments show that $\widetilde{Q}_l \in \mathcal{P}_c^n$, for all l. For each l and l' in $\{1, \ldots, n/4\}$, we have that, by constructions, $H(\eta(\widetilde{P}_l), \eta(\widetilde{Q}_{l'}) \geq \frac{n}{2}$, where $\eta(\widetilde{P}_l)$ and $\eta(\widetilde{Q}_{l'})$ denote the sets of change point

locations of $\widetilde{P}_l$ and $\widetilde{Q}_{l'}$, respectively. Then it follows from Le Cam's lemma (e.g. Yu, 1997) that

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}_c^n} \mathbb{E}_P\big(H(\hat{\eta}, \eta(P))\big) \geq \frac{n}{4}\big\{1 - d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q})\big\}, \tag{17}$$

where $d_{\mathrm{TV}}(\cdot, \cdot)$ is the total variation distance between two probability measures and the infimum is over all estimators $\widehat{\eta} = \{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ of the change point locations. Above, $\eta(P)$ is the set of locations of all the change points of $P \in \mathcal{P}_c^n$.

Let $u_l \in \mathbb{R}^{n/2}$ be a sub-vector of $\widetilde{u}_l$ consisting of the first $n/2$ entries of $\widetilde{u}_l$. Let P_l and P_0 be the multivariate Gaussian distributions $\mathcal{N}(u_l, \sigma^2 I_{n/2})$ and $\mathcal{N}(0, \sigma^2 I_{n/2})$, receptively. Due to the symmetry between $\widetilde{u}_l$ and $\widetilde{v}_l$, it holds that

$$d_{\mathrm{TV}}(\widetilde{P}, \widetilde{Q}) \leq 2 d_{\mathrm{TV}}(P, P_0), \tag{18}$$

where $P = \frac{1}{n/4}\sum_{l=1}^{n/4} P_l$. Since $d_{\mathrm{TV}}(P, P_0) \leq \sqrt{\chi^2(P, P_0)}$, where $\chi^2(\cdot, \cdot)$ is the χ^2-divergence between two probability measures (see, e.g., Equation 2.27 in Tsybakov, 2009), it suffices to provide an upper bound for $\chi^2(P, P_0)$. We have

$$\begin{aligned}
\chi^2(P, P_0) &= \left(\frac{1}{n/4}\right)^2 \sum_{l,m=1}^{n/4} \mathbb{E}_{P_0}\left(\frac{dP_l dP_m}{dP_0 dP_0}\right) - 1 \\
&= \left(\frac{1}{n/4}\right)^2 \sum_{l,m=1}^{n/4} \exp\left(\frac{u_l^\top u_m}{\sigma^2}\right) - 1 \\
&= \left(\frac{4}{n}\right)^2 \left[\sum_{l=1}^{n/4}\big\{\exp\big(c\log(n)\big)\big\} + (n/4)(n/4-1)\right] - 1 \\
&= 4n^{-1}(n^c - 1),
\end{aligned}$$

where the third identity follows from the observation that for $l, m = 1, \ldots, n/4$,

$$u_l^\top u_m = \mathbb{1}\{l = m\} c\sigma^2 \log(n).$$

Therefore for any $0 < c < 1$, there exists a sufficiently large $n(c)$ such that for any $n \geq n(c)$, $4n^{-1}(n^c - 1) \leq 1/16$. This combining with (17) and (18) provides the desired result. □

Proof of Lemma 2. Let P_0 denote the joint distribution of the independent random variables $\{Y_i\}_{i=1}^n$, where

$$Y_1, \ldots, Y_\Delta \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{and} \quad Y_{\Delta+1}, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\kappa, \sigma^2);$$

and, similarly, let P_1 be the joint distribution of the independent random variables $\{Z_i\}_{i=1}^n$ such that

$$Z_1, \ldots, Z_{\Delta+\delta} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad \text{and} \quad Z_{\Delta+\delta+1}, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(\kappa, \sigma^2),$$

where δ is a positive integer no larger than $n-1-\Delta$. Observe that $\eta(P_0)=\Delta$ and $\eta(P_1)=\Delta+\delta$. By Le Cam's Lemma (e.g. Yu, 1997) and Lemma 2.6 in Tsybakov (2009), it holds that

$$\inf_{\hat{\eta}}\sup_{P\in\mathcal{Q}}\mathbb{E}_P\big(|\hat{\eta}-\eta|\big)\geq\delta\big\{1-d_{\mathrm{TV}}(P_0,P_1)\big\}\geq\frac{\delta}{2}\exp\left(-KL(P_0,P_1)\right),$$

where $KL(\cdot,\cdot)$ is the Kullback–Leibler divergence between two probability measures.

Since both P_0 and P_1 are product measures, it holds that

$$KL(P_0,P_1)=\sum_{i\in\{\Delta+1,\ldots,\Delta+\delta\}}KL(P_{0,i},P_{1,i})=\delta\frac{\kappa^2}{\sigma^2},$$

where $P_{0,i}$ and $P_{1,i}$ are the distributions of Y_i and Z_i, respectively and the last identity follows from the fact that, if P and Q are the normal distributions with common variance σ^2 and means μ_1 and μ_2, respectively, then $K(P,Q)=\frac{(\mu_1-\mu_2)^2}{\sigma^2}$. Thus,

$$\inf_{\hat{\eta}}\sup_{P\in\mathcal{Q}^n}\mathbb{E}_P\big(|\hat{\eta}-\eta|\big)\geq\frac{\delta}{2}\exp\left(-\delta\frac{\kappa^2}{\sigma^2}\right).\tag{19}$$

Next, set $\delta=\min\{\lceil\frac{\sigma^2}{\kappa^2}\rceil,n-1-\Delta\}$. By the assumption on ζ_n, for all n large enough we must have that $\delta=\lceil\frac{\sigma^2}{\kappa^2}\rceil$. Indeed, if $n-1-\Delta\leq\lceil\frac{\sigma^2}{\kappa^2}\rceil$ then, as $\Delta<n/2$, we must have that $\frac{\kappa^2}{\sigma^2}\leq\frac{1}{n-2-\Delta}<\frac{1}{n/2-2}$, and, therefore, that

$$\frac{\kappa^2\Delta}{\sigma^2}<\frac{\Delta}{n/2-2}<\frac{n}{n/2-2}<10,$$

where we may assume that $n>4$. Since $\frac{\kappa^2\Delta}{\sigma^2}\geq\zeta_n^2$ by assumption and ζ_n is diverging as $n\to\infty$, the above bound can only hold for finitely many n. The claimed bound now follows from (19), for all n large enough. □

Appendix B: Proofs of the Results in Section 3

In this section, we provide technical details of the proof of Theorem 3. Recalling Assumption 1, for any change point η_k, observe that the interval $I=\{\eta_{k-1}+1,\ldots,\eta_k\}$ contains one change point, but the signal $\{f_i\}_{i=1}^n$ is unchanged in I. For convenience, in this section, any interval I is said to contain a true change point if there exists $k\in\{1,\ldots,K\}$ such that $\{\eta_k,\eta_k+1\}\subset I$, where $|I|\geq 2$. This convention ensures that if I contains a true change point, then it is necessary that there exist $i,j\in I$ satisfying $f_i\neq f_j$.

Lemma 5. *Let I_1 and I_2 denote any two disjoint intervals of $\{1,\ldots,n\}$ and $I=I_1\cup I_2$. For any sequences $\{X_i\}_{i=1}^n,\{Y_i\}_{i=1}^n\subset\mathbb{R}$, it holds that*

$$\sum_{i\in I}\big(Y_i-\overline{Y}_I\big)^2=\sum_{i\in I_1}\big(Y_i-\overline{Y}_{I_1}\big)^2+\sum_{i\in I_2}\big(Y_i-\overline{Y}_{I_2}\big)^2+\frac{|I_1||I_2|}{|I_1|+|I_2|}\big(\overline{Y}_{I_1}-\overline{Y}_{I_2}\big)^2,\tag{20}$$

and

$$\begin{aligned}&\sum_{i\in I}(X_i-\overline{X}_I)(Y_i-\overline{Y}_I)\\&=\sum_{i\in I_1}(X_i-\overline{X}_{I_1})(Y_i-\overline{Y}_{I_1})+\sum_{i\in I_2}(X_i-\overline{X}_{I_2})(Y_i-\overline{Y}_{I_2})\\&\qquad+\frac{|I_1||I_2|}{|I_1|+|I_2|}(\overline{X}_{I_1}-\overline{X}_{I_2})(\overline{Y}_{I_1}-\overline{Y}_{I_2}).\end{aligned}\tag{21}$$

Proof. Without loss of generality, let $I_1=\{1,\dots,n_1\}$ and $I_2=\{n_1+1,\dots,n=n_1+n_2\}$. For simplicity, denote $\overline{X}=\overline{X}_I$, $\overline{X}_1=\overline{X}_{I_1}$ and $\overline{X}_2=\overline{X}_{I_2}$. The results (20) and (21) can be proved by similar arguments. We will only show (21) here.

Observe that

$$\begin{aligned}&\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\\&=\sum_{i=1}^{n_1}\left\{X_i-\overline{X}_1+\frac{n_2(\overline{X}_1-\overline{X}_2)}{n_1+n_2}\right\}\left\{Y_i-\overline{Y}_1+\frac{n_2(\overline{Y}_1-\overline{Y}_2)}{n_1+n_2}\right\}\\&\qquad+\sum_{i=n_1+1}^{n}\left\{X_i-\overline{X}_2-\frac{n_1(\overline{X}_1-\overline{X}_2)}{n_1+n_2}\right\}\left\{Y_i-\overline{Y}_2-\frac{n_1(\overline{Y}_1-\overline{Y}_2)}{n_1+n_2}\right\}\\&=\sum_{i=1}^{n_1}(X_i-\overline{X}_1)(Y_i-\overline{Y}_1)+\sum_{i=n_1+1}^{n_2}(X_i-\overline{X}_2)(Y_i-\overline{Y}_2)\\&\qquad+\frac{n_1n_2}{n_1+n_2}(\overline{X}_1-\overline{X}_2)(\overline{Y}_1-\overline{Y}_2).\end{aligned}$$

□

Lemma 6. *Assume that the sequence* $\{Y_i\}_{i=1}^n\subset\mathbb{R}$ *satisfies Assumption 1. It holds that*

$$\mathbb{P}\Bigg\{\sup_{1\le a<b<c\le n}\sqrt{\frac{(b-a)(c-b)}{c-a}}\big|\overline{Y}_{(a+1,b]}-\overline{f}_{(a+1,b]}+\overline{Y}_{(b+1,c]}-\overline{f}_{(b+1,c]}\big|\\\le C_{\mathcal{B}}\sigma\sqrt{\log(n)}\Bigg\}\ge e\cdot n^{3-c_{\mathcal{B}}},$$

where $c_{\mathcal{B}}$ *is an absolute constant chosen to satisfy* $c_{\mathcal{B}}>3$ *and* $C_{\mathcal{B}}>0$ *only depends on* $c_{\mathcal{B}}$.

Proof. It follows from Assumption 1 that for all $i\in\{1,\dots,n\}$, Y_i-f_i is a centred sub-Gaussian random variable with $\max_i\|Y_i-f_i\|_{\psi_2}\le\sigma$. Due to Hoeffding inequality (see e.g. Vershynin, 2010), it holds that for any non-empty set $I\subset\{1,\dots,n\}$ and any $\varepsilon>0$,

$$\mathbb{P}\big\{\big|\overline{Y}_I-\overline{f}_I\big|>\varepsilon\big\}\le e\cdot\exp\left(-\frac{c|I|\varepsilon^2}{\sigma^2}\right),$$

and for any triple $i_1 < i_2 < i_3$ chosen in $\{1, \ldots, n\}$

$$\mathbb{P}\left\{\sqrt{\frac{(i_2 - i_1)(i_3 - i_2)}{i_3 - i_1}}\left|\overline{Y}_{(i_1+1,i_2]} - \overline{f}_{(i_1+1,i_2]} + \overline{Y}_{(i_2+1,i_3]} - \overline{f}_{(i_2+1,i_3]}\right| \geq \varepsilon\right\} \leq e \cdot \exp\left(-\frac{c\varepsilon^2}{\sigma^2}\right),$$

where $c > 0$ is an absolute constant only depending on σ. The result follows from a union bound. □

For simplicity, in the rest of the proof, we will let $C_{\mathcal{B}} = 1$ and set $c_{\mathcal{B}} > 3$. This will only affect the constant C_λ, and in the statement of Theorem 3, we require $C_\lambda > 0$ to be large enough.

Since the change points of $\widehat{u}$ are our change point estimators, with the error rate

$$\epsilon_k = C_\epsilon \sigma^2 \log(n)/\kappa_k^2,$$

we refer to η_k as an undetected change point, if $\eta_k \in (s, e] \in \widehat{\mathcal{P}}(\widehat{u})$ and

$$\epsilon_k - s = \epsilon_k - \epsilon_{k-1} - (s - \epsilon_{k-1}) \geq \Delta - C_\epsilon \sigma^2 \log(n)/\kappa_k^2 > \Delta/3, \tag{22}$$

and similarly $e - \epsilon_k > \Delta/3$. The first and second inequalities of (22) follow from Assumptions 1 and 2, respectively. In the rest of this section, let $\widehat{\mathcal{P}} = \widehat{\mathcal{P}}(\widehat{u})$.

Step 1: no more than two true change points

In order to show that no $I \in \widehat{\mathcal{P}}$ contains more than two true change points, it suffices to show that no $I \in \widehat{\mathcal{P}}$ contains undetected change points, due to the minimal spacing Δ condition in Assumption 1.

Lemma 7. *Let $\{Y_i\}_{i=1}^n$ satisfy Assumptions 1 and 2, and λ satisfy the condition*

$$\sigma^2 \log(n) \leq \lambda \leq \kappa^2 \Delta/48. \tag{23}$$

Then, in the event $\mathcal{B}$, it holds that no $I \in \widehat{\mathcal{P}}$ contains any undetected change point.

Proof. We first point out that due to Assumption 2, (23) is not an empty set.

For the sake of contradiction, suppose that there exists $I \in \widehat{\mathcal{P}}$ containing an undetected change point η_k, i.e.,

$$\min\{e - (\eta_k + 1), \eta_k - s\} > \Delta/3. \tag{24}$$

Denote

$$I_1 = (s, \eta_k - \Delta/3], \quad I_2 = (\eta_k - \Delta/3, \eta_k], \quad I_3 = (\eta_k, \eta_k + \Delta/3]$$

and

$$I_4 = (\eta_k + \Delta/3, e],$$

none of which is empty due to (24).

Let $\widetilde{\mathcal{P}}$ be such that

$$\widetilde{\mathcal{P}} = \widehat{\mathcal{P}} \cup \{I_1, I_2, I_3, I_4\} \setminus \{I\},$$

and $\widetilde{u}$ be the piecewise constant vector induced by $\widetilde{\mathcal{P}}$. By the definition of $\widehat{u}$, it holds that

$$H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) \le H\big(\widetilde{u}, \{Y_i\}_{i=1}^n, \lambda\big).$$

Since $\widetilde{\mathcal{P}}$ is a refinement of $\widehat{\mathcal{P}}$ and we have assumed in Assumption 1 that the distributions of Y_i's have continuous density functions, it follows that

$$\lambda(\|D\widehat{u}\|_0 - \|D\widetilde{u}\|_0) = -3\lambda.$$

Then

$$\begin{aligned}
0 &\ge H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widetilde{u}, \{Y_i\}_{i=1}^n, \lambda\big) \\
&= -3\lambda + \sum_{i\in I}(Y_i - \overline{Y}_I)^2 - \sum_{i\in I_1}(Y_i - \overline{Y}_{I_1})^2 - \sum_{i\in I_2}(Y_i - \overline{Y}_{I_2})^2 - \sum_{i\in I_3}(Y_i - \overline{Y}_{I_3})^2 \\
&\qquad - \sum_{i\in I_4}(Y_i - \overline{Y}_{I_4})^2 \\
&\ge -3\lambda + \frac{|I_2||I_3|}{|I_2| + |I_3|}(\overline{Y}_{I_2} - \overline{Y}_{I_3})^2 \\
&= -3\lambda + \frac{|I_2||I_3|}{|I_2| + |I_3|}\big\{(\overline{Y}_{I_2} - f_{\eta_k}) - (\overline{Y}_{I_3} - f_{\eta_{k+1}}) + (f_{\eta_k} - f_{\eta_{k+1}})\big\}^2 \\
&\ge -3\lambda + \frac{\Delta}{12}(f_{\eta_k} - f_{\eta_{k+1}})^2 - \frac{|I_2||I_3|}{|I_2| + |I_3|}\big\{(\overline{Y}_{I_2} - f_{\eta_k}) - (\overline{Y}_{I_3} - f_{\eta_{k+1}})\big\}^2 \\
&\ge -4\lambda + \frac{\Delta}{12}\kappa_k^2 \\
&> 0,
\end{aligned} \tag{25}$$

where the second inequality follows from (20) by first splitting $I = \{I_1, I_2, I_3\} \cup \{I_4\}$, then $\{I_1, I_2, I_3\} = \{I_1\} \cup \{I_2, I_3\}$ and $\{I_2, I_3\} = \{I_2\} \cup \{I_3\}$; the third inequality follows from the observation that $(x+y)^2 \ge x^2/2 - y^2$ and letting $x = f_{\eta_k} - f_{\eta_{k+1}}$, $y = (\overline{Y}_{I_2} - f_{\eta_k}) - (\overline{Y}_{I_3} - f_{\eta_{k+1}})$; the fourth inequality follows from the definition of $\mathcal{B}$ and (23); and the last inequalities is due to (23).

Since (25) is a contradiction, we conclude that there is no interval containing undetected change point. □

Step 2: exactly two true change points

Lemma 8. *Let $\{Y_i\}_{i=1}^n$ satisfy Assumptions 1 and 2 and set $\lambda = C_\lambda \sigma^2 \log(n)$, where $C_\lambda > 1$. In the event $\mathcal{B}$, it holds that if $I = (s, e] \in \widehat{\mathcal{P}}$ contains exactly*

two change points, say η_k *and* η_{k+1}*, then*

$$\eta_k - s + 1 \le 12\lambda/\kappa_k^2, \text{ and } e - \eta_{k+1} \le 12\lambda/\kappa_{k+1}^2.$$

Proof. Let $I_1 = (s, \eta_k]$, $I_2 = (\eta_k, \eta_{k+1}]$ and $I_3 = (\eta_{k+1}, e]$. Since I contains exactly two true change points, none of I_1, I_2 or I_3 is an empty set, and $\{f_i\}_{i=1}^n$ is constant on I_1, I_2 and I_3. Denote by $\widehat{u}$ the solution of (6) with inputs $\{Y_i\}_{i=1}^n$ and λ, and by $\widehat{\mathcal{P}}$ the interval partition induced by $\widehat{u}$.

Let $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$ be

$$\widetilde{\mathcal{P}}_1 = \widehat{\mathcal{P}} \cup \{I_1, I_2 \cup I_3\} \setminus \{I\}, \text{ and } \widetilde{\mathcal{P}}_2 = \widehat{\mathcal{P}} \cup \{I_1, I_2, I_3\} \setminus \{I\},$$

respectively; and let $\widetilde{u}_1$ and $\widetilde{u}_2$ be the piecewise-constant vectors induced by $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$, respectively.

It follows from Lemma 5 that

$$H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) = \lambda - \frac{|I_1||I_2 \cup I_3|}{|I_1| + |I_2 \cup I_3|}\big(\overline{Y}_{I_1} - \overline{Y}_{I_2 \cup I_3}\big)^2$$

and

$$H\big(\widetilde{u}_2, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) = \lambda - \frac{|I_2||I_3|}{|I_2| + |I_3|}\big(\overline{Y}_{I_2} - \overline{Y}_{I_3}\big)^2.$$

Then,

$$\begin{aligned}
0 &\le H\big(\widetilde{u}_2, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) \\
&\le 2\lambda - \frac{|I_2||I_3|}{|I_2| + |I_3|}\big(\overline{Y}_{I_2} - \overline{Y}_{I_3}\big)^2 \\
&\le 2\lambda - \frac{1}{2}\frac{|I_2||I_3|}{|I_2| + |I_3|}\big\{\big(f_{\eta_{k+1}} - f_{\eta_{k+2}}\big)^2 - 2\big(\overline{Y}_{I_2} - f_{\eta_{k+1}} - \overline{Y}_{I_3} + f_{\eta_{k+2}}\big)^2\big\} \\
&\le 2\lambda - \frac{1}{2}\frac{|I_2||I_3|}{|I_2| + |I_3|}\kappa_{k+1}^2 + \lambda,
\end{aligned}$$

where the third inequality uses the same argument in the third inequality of (25), and the last inequality follows from the definition of $\mathcal{B}$ and Assumption 1.

If $|I_2| \le |I_3|$, then

$$\Delta/2 \le |I_2|/2 \le \frac{|I_2||I_3|}{|I_2| + |I_3|} \le 6\lambda/\kappa_{k+1}^2,$$

which contradicts Assumption 2. Therefore it must hold that $|I_2| > |I_3|$, which implies

$$|I_3|/2 \le \frac{|I_2||I_3|}{|I_2| + |I_3|} \le 6\lambda/\kappa_{k+1}^2.$$

Then $e - \eta_{k+1} \le 12\lambda/\kappa_{k+1}^2$. It can be shown similarly that $\eta_k - s + 1 \le 12\lambda/\kappa_k^2$. □

Step 3: one and only one change point

Let $I_1 = (s, e_1] \in \widehat{\mathcal{P}}$ contain exactly one true change point, namely η_k. With our convention set at the beginning of Appendix B, it holds that

$$\eta_{k-1} + 1 \leq s \leq \eta_k < \eta_k + 1 \leq e_1 \leq \eta_{k+1}. \tag{26}$$

Denote $\delta = e_1 - \eta_k$ and $\epsilon = \eta_k - (s-1)$. Without loss of generality, we assume that

$$0 < \epsilon \leq \delta. \tag{27}$$

We are to show that there exists an absolute constant $C > 8$ such that

$$\epsilon = |\eta_k - s + 1| \leq C\lambda/\kappa_k^2 \tag{28}$$

and

$$\epsilon_1 = |\eta_{k+1} - e_1| \leq C\lambda/\kappa_{k+1}^2. \tag{29}$$

Equation (28) will be shown in Lemma 9. To show (29), we rely on the following arguments (see Figure 1 for an illustration):

(i) Let $I_2 = (e_1, e_2]$ be the interval to the immediate right of I_1 in $\widehat{\mathcal{P}}$. It must hold that
$$e_1 \leq \eta_{k+1} < \eta_{k+1} + 1 \leq e_2. \tag{30}$$
This will be shown in Lemma 10.

(ii) It follows from Appendix B that there are at most two true change points in $(e_1, e_2]$. If there are exactly two true change points, then due to Appendix B, (29) holds.

(iii) If $e_2 \leq \eta_{k+2}$, then we let $\epsilon_1 = \eta_{k+1} - (e-1)$ and $\delta_1 = e_2 - \eta_{k+1}$. Lemma 11 shows that $\delta_1 < \epsilon_1$ is impossible. Thus, $\epsilon_1 \leq \delta_1$ and we then rely on Lemma 9.

Lemma 9. *Let $\{Y_i\}_{i=1}^n$ satisfy Assumptions 1 and 2 and set $\lambda \geq C_\lambda \sigma^2 \log(n)$, with $C_\lambda \geq 1$. In the event $\mathcal{B}$, it holds that if $I_1 = (s, e_1] \in \widehat{\mathcal{P}}$ contains exactly one change point, say η_k, then*

$$\min\{|J_1|, |J_2|\} \leq 8\lambda/\kappa_k^2,$$

where $J_1 = (s, \eta_k]$ and $J_2 = (\eta_k, e_1]$.

Proof. Observe that neither J_1 nor J_2 is empty by definition, and that $\{f_i\}_{i=1}^n$ is constant within J_1 and J_2, respectively. Let $\widetilde{\mathcal{P}}$ be such that

$$\widetilde{\mathcal{P}} = \widehat{\mathcal{P}} \cup \{J_1, J_2\} \setminus \{I_1\},$$

and let $\widetilde{u}$ be the piecewise-constant vector induced by $\widetilde{\mathcal{P}}$.

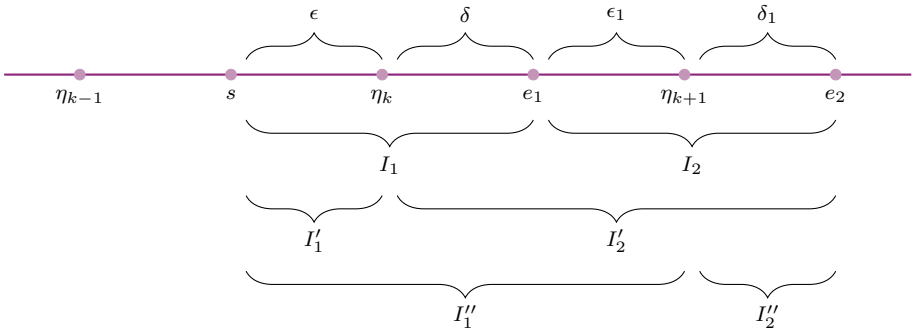

FIG 1. *Illustrations of the interval constructions used in the Step 3 in the proof of Theorem 3.*

Recall that $\mathbb{E}(\overline{Y}_{J_1}) = f_{\eta_k}$ and $\mathbb{E}(\overline{Y}_{J_2}) = f_{\eta_{k+1}}$. Without loss of generality, assume $f_{\eta_{k+1}} = f_{\eta_k} + \kappa_k$. Thus,

$$\begin{aligned}
0 &\geq H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widetilde{u}, \{Y_i\}_{i=1}^n, \lambda\big) \\
&= -\lambda + \sum_{i\in I_1}(Y_i - \overline{Y}_{I_1})^2 - \sum_{i\in J_1}(Y_i - \overline{Y}_{J_1})^2 - \sum_{i\in J_2}(Y_i - \overline{Y}_{I_2})^2 \\
&= -\lambda + \frac{|J_1||J_2|}{|I_1|}(\overline{Y}_{J_1} - \overline{Y}_{J_2})^2 \\
&= -\lambda + \frac{|J_1||J_2|}{|I_1|}\big\{(\overline{Y}_{J_1} - f_{\eta_k}) - (\overline{Y}_{J_2} - f_{\eta_k} - \kappa_k) - \kappa_k\big\}^2 \\
&\geq -\lambda + \frac{|J_1||J_2|}{2|I_1|}\big\{\kappa_k^2 - 2(\overline{Y}_{J_1} - f_{\eta_k} - \overline{Y}_{J_2} + f_{\eta_{k+1}})^2\big\} \\
&\geq -2\lambda + \frac{|J_1||J_2|}{2|I_1|}\kappa_k^2,
\end{aligned}$$

where the second identity follows from (20), the second inequality from the fact that $(x-y)^2 \geq y^2/2 - x^2$ with $x = \kappa_k$ and $y = (\overline{Y}_{J_1} - f_{\eta_k}) - (\overline{Y}_{J_2} - f_{\eta_k} - \kappa_k)$, and the last inequality from the definitions of the event $\mathcal{B}$ and the choice of λ. Therefore,

$$\min\{|J_1|, |J_2|\}\kappa_k^2/8 \leq \frac{|J_1||J_2|}{|I_1|}\kappa_k^2/4 \leq \lambda. \qquad \square$$

Lemma 10. *Let $\{Y_i\}_{i=1}^n$ satisfy Assumptions 1 and 2 and set $\lambda = C_\lambda \sigma^2 \log(n)$, with $C_\lambda > 85$. Assume that $I_1 = (s, e_1] \in \widehat{\mathcal{P}}$ contains exactly one change point namely η_k. Denote $\delta = e_1 - \eta_k$ and $\epsilon = \eta_k - (s-1)$. Assume that $\epsilon \leq \delta$. In the event $\mathcal{B}$, if $I_2 = (e_1, e_2] \in \widehat{\mathcal{P}}$, then it must hold that*

$$e_1 \leq \eta_{k+1} < \eta_{k+1} + 1 \leq e_2.$$

Proof. Let $I_1' = (s, \eta_k]$ and $I_2' = (\eta_k, e_2]$. Then $I_1 \cup I_2 = I_1' \cup I_2'$. Let $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$

be

$$\widetilde{\mathcal{P}}_1 = \widehat{\mathcal{P}} \cup \{I_1 \cup I_2\} \setminus \{I_1, I_2\}$$

and

$$\widetilde{\mathcal{P}}_2 = \widehat{\mathcal{P}} \cup \{I_1', I_2'\} \setminus \{I_1, I_2\},$$

respectively. Let $\widetilde{u}_1$ and $\widetilde{u}_2$ be the piecewise-constant vectors induced by $\widetilde{\mathcal{P}}_1$ and $\widetilde{\mathcal{P}}_2$, respectively.

We proceed by contradiction. We assume that $e_2 \leq \eta_{k+1}$. Without loss of generality, assume $f_{\eta_{k+1}} = f_{\eta_k} + \kappa_k$. Due to Assumption 1, it holds that $\mathbb{E}(\overline{Y}_{I_1'}) = f_{\eta_k}$, $\mathbb{E}(\overline{Y}_{I_2'}) = f_{\eta_{k+1}} = f_{\eta_k} + \kappa_k$, $\mathbb{E}(\overline{Y}_{I_1}) = f_{\eta_k} + \delta\kappa_k/|I_1|$ and $\mathbb{E}(\overline{Y}_{I_2}) = f_{\eta_{k+1}} = f_{\eta_k} + \kappa_k$. Then,

$$\begin{aligned}
0 &\leq H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) \\
&= -\lambda + \frac{|I_1||I_2|}{|I_1| + |I_2|}(\overline{Y}_{I_1} - \overline{Y}_{I_2})^2 \\
&= -\lambda + \frac{|I_1||I_2|}{|I_1| + |I_2|}\big\{\overline{Y}_{I_1} - \mathbb{E}(\overline{Y}_{I_1}) - \overline{Y}_{I_2} + f_{\eta_{k+1}} + (\delta/|I_1| - 1)\kappa_k\big\}^2 \\
&\leq -\lambda + \frac{|I_1||I_2|}{|I_1| + |I_2|}\big\{5(\overline{Y}_{I_1} - \mathbb{E}(\overline{Y}_{I_1}) - \overline{Y}_{I_2} + f_{\eta_{k+1}})^2 + \frac{5}{4}(\delta/|I_1| - 1)^2\kappa_k^2\big\} \\
&\leq -\lambda + 5\sigma^2 \log(n) + \frac{5}{4}\frac{|I_1||I_2|}{|I_1| + |I_2|}\frac{\epsilon^2\kappa_k^2}{|I_1|^2} \qquad (31)
\end{aligned}$$

where the second inequity follows form the fact that $(x + y)^2 \leq 5x^2 + (5/4)y^2$ and the last inequality follows from the definition of the event $\mathcal{B}$.

In addition, we have

$$\begin{aligned}
&H\big(\widetilde{u}_2, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) \\
&= \lambda - \frac{|I_1'||I_2'|}{|I_1'| + |I_2'|}(\overline{Y}_{I_1'} - \overline{Y}_{I_2'})^2 \\
&= \lambda - \frac{|I_1'||I_2'|}{|I_1'| + |I_2'|}\big\{\overline{Y}_{I_1'} - f_{\eta_k} - \overline{Y}_{I_2'} + f_{\eta_k} + \kappa_k - \kappa_k\big\}^2 \\
&\leq \lambda - \frac{|I_1'||I_2'|}{|I_1'| + |I_2'|}\left\{\frac{3}{4}\kappa_k^2 - 3(\overline{Y}_{I_1'} - f_{\eta_k} - \overline{Y}_{I_2'} + f_{\eta_k} + \kappa_k)^2\right\} \\
&\leq \lambda - \frac{3}{4}\frac{|I_1'||I_2'|}{|I_1'| + |I_2'|}\kappa_k^2 + 3\sigma^2 \log(n),
\end{aligned}$$

where the first inequality follows from the fact that $(x - y)^2 \geq (3/4)y^2 - 4x^2$, and the last inequality follows from the definition of the event $\mathcal{B}$.

Then

$$\begin{aligned}
0 \leq H\big(\widetilde{u}_2, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) &\leq 8\sigma^2 \log(n) \\
&+ \frac{5}{4}\frac{|I_1||I_2|}{|I_1| + |I_2|}\frac{\epsilon^2\kappa_k^2}{|I_1|^2} - \frac{3}{4}\frac{|I_1'||I_2'|}{|I_1'| + |I_2'|}\kappa_k^2
\end{aligned}$$

$$= 8\sigma^2 \log(n) + \frac{\kappa_k^2 \epsilon}{4(|I_1| + |I_2|)} \left\{ \frac{5|I_2|\epsilon}{\epsilon + \delta} - 3(\delta + |I_2|) \right\}$$
$$\leq 8\sigma^2 \log(n) - \frac{\kappa_k^2 \epsilon}{4(|I_1| + |I_2|)} (3\delta + |I_2|/2),$$

therefore

$$\kappa_k^2 \epsilon \leq 64\sigma^2 \log(n).$$

Combined with (31), this implies that

$$\lambda \leq 5\sigma^2 \log(n) + 64\sigma^2 \log(n) \frac{5|I_2|\epsilon}{4(|I_1| + |I_2|)|I_1|} \leq 85\sigma^2 \log(n),$$

which contradicts with the assumption that $\lambda > 85\sigma^2 \log(n)$. □

Lemma 11. *Let $\{Y_i\}_{i=1}^n$ satisfy Assumptions 1 and 2 and set $\lambda = C_\lambda \sigma^2 \log(n)$ with a sufficiently large $C_\lambda > 0$. Assume that there exists an interval partition $\mathcal{P}$ with induced piecewise constant vector u such that $I_1 = (s, e_1] \in \mathcal{P}$ and $I_2 = (e_1, e_2] \in \mathcal{P}$, where I_1 and I_2 satisfy (26) and (30). Let $\epsilon = \eta_k - s + 1$, $\delta = e_1 - \eta_k + 1$, $\epsilon_1 = \eta_{k+1} - e + 1$ and $\delta_1 = e_2 - \eta_{k+1} + 1$. Assume $\epsilon < \delta$ and $\epsilon_1 > \delta_1$. Then in the event $\mathcal{B}$, u is not a minimizer of (6).*

Proof. For notational simplicity, let

$$f_{\eta_k} = \mu + \omega_1, \quad f_{\eta_{k+1}} = \mu \quad \text{and} \quad f_{\eta_{k+2}} = \mu + \omega_2.$$

Let $I_1' = (s, \eta_k]$, $I_2' = (\eta_k, e_2]$, $I_1'' = (s, \eta_{k+1}]$ and $I_2'' = (\eta_{k+1}, e_2]$. Then $I_1 \cup I_2 = I_1' \cup I_2' = I_1'' \cup I_2''$. Let $\widetilde{\mathcal{P}}_1$, $\widetilde{\mathcal{P}}_2$ and $\widetilde{\mathcal{P}}_3$ be such that

$$\widetilde{\mathcal{P}}_1 = \mathcal{P} \cup \{I_1 \cup I_2\} \setminus \{I_1, I_2\}, \widetilde{\mathcal{P}}_2 = \mathcal{P} \cup \{I_1', I_2'\} \setminus \{I_1, I_2\}$$

and

$$\widetilde{\mathcal{P}}_3 = \mathcal{P} \cup \{I_1'', I_2''\} \setminus \{I_1, I_2\}.$$

Let $\widetilde{u}_1$, $\widetilde{u}_2$ and $\widetilde{u}_3$ be the piecewise constant vectors induced by $\widetilde{\mathcal{P}}_1$, $\widetilde{\mathcal{P}}_2$ and $\widetilde{\mathcal{P}}_3$. The population means are

$$\mathbb{E}(\overline{Y}_{I_1}) = \mu + \frac{\epsilon\omega_1}{\epsilon + \delta} \text{ and } \mathbb{E}(\overline{Y}_{I_2}) = \mu + \frac{\delta_1\omega_2}{\epsilon_1 + \delta_1}.$$

Let $0 < \alpha < 1$ be a fixed constant to be specified later. We have the following:

$$H(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda) - H(u, \{Y_i\}_{i=1}^n, \lambda) = -\lambda + \frac{|I_1||I_2|}{|I_1| + |I_2|} (\overline{Y}_{I_1} - \overline{Y}_{I_2})^2$$
$$= -\lambda + \frac{|I_1||I_2|}{|I_1| + |I_2|} \{\overline{Y}_{I_1} - \mathbb{E}(\overline{Y}_{I_1}) - \overline{Y}_{I_2} + \mathbb{E}(\overline{Y}_{I_2}) + \mathbb{E}(\overline{Y}_{I_1}) - \mathbb{E}(\overline{Y}_{I_2})\}^2$$
$$\leq -\lambda + 2(1 + \alpha)\alpha^{-1}\sigma^2 \log(n) + (1 + \alpha) \frac{|I_1||I_2|}{|I_1| + |I_2|} \left(\frac{\epsilon\omega_1}{\epsilon + \delta} - \frac{\delta_1\omega_2}{\epsilon_1 + \delta_1} \right)^2,$$

$$
\begin{aligned}
&H\big(\widetilde{u}_2,\{Y_i\}_{i=1}^n,\lambda\big)-H\big(\widetilde{u}_1,\{Y_i\}_{i=1}^n,\lambda\big)=\lambda-\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}(\overline{Y}_{I_1'}-\overline{Y}_{I_2'})^2\\
=&\lambda-\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}\big\{\overline{Y}_{I_1'}-\mathbb{E}(\overline{Y}_{I_1'})-\overline{Y}_{I_2'}+\mathbb{E}(\overline{Y}_{I_2'})+\mathbb{E}(\overline{Y}_{I_1'})-\mathbb{E}(\overline{Y}_{I_2'})\big\}^2\\
\le&\lambda-(1-\alpha)\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}\left(\omega_1-\frac{\omega_2\delta_1}{\delta+\epsilon_1+\delta_1}\right)^2+\frac{2(1-\alpha)}{\alpha}\sigma^2\log(n),
\end{aligned}
$$

and

$$
\begin{aligned}
&H\big(\widetilde{u}_3,\{Y_i\}_{i=1}^n,\lambda\big)-H\big(\widetilde{u}_1,\{Y_i\}_{i=1}^n,\lambda\big)=\lambda-\frac{|I_1''||I_2''|}{|I_1''|+|I_2''|}(\overline{Y}_{I_1''}-\overline{Y}_{I_2''})^2\\
=&\lambda-\frac{|I_1''||I_2''|}{|I_1''|+|I_2''|}\big\{\overline{Y}_{I_1''}-\mathbb{E}(\overline{Y}_{I_1''})-\overline{Y}_{I_2''}+\mathbb{E}(\overline{Y}_{I_2''})+\mathbb{E}(\overline{Y}_{I_1''})-\mathbb{E}(\overline{Y}_{I_2''})\big\}^2\\
\le&\lambda-(1-\alpha)\frac{|I_1''||I_2''|}{|I_1''|+|I_2''|}\left(\omega_2-\frac{\omega_1\epsilon}{\epsilon+\delta+\epsilon_1}\right)^2+\frac{2(1-\alpha)}{\alpha}\sigma^2\log(n).
\end{aligned}
$$

For the rest of the proof, we proceed by contradiction by assuming that u is the minimizer of (6). We will consider the cases $\omega_1\omega_2>0$ and $\omega_1\omega_2<0$ separately in **Steps 1** and **2**, respectively.

Step 1. Suppose $\omega_1\omega_2>0$. Without loss of generality, assume $\omega_1,\omega_2>0$ and for some $0<\beta\le 1$, it holds that

$$
\frac{\delta_1\omega_2}{\epsilon_1+\delta_1}=\beta\frac{\epsilon\omega_1}{\epsilon+\delta}.
$$

We have

$$
\begin{aligned}
0\le&\, H\big(\widetilde{u}_1,\{Y_i\}_{i=1}^n,\lambda\big)-H\big(u,\{Y_i\}_{i=1}^n,\lambda\big)\\
\le&-\lambda+2(1+\alpha)\alpha^{-1}\sigma^2\log(n)+(1+\alpha)(1-\beta)^2\frac{|I_1||I_2|}{|I_1|+|I_2|}\left(\frac{\epsilon\omega_1}{\epsilon+\delta}\right)^2, \qquad (32)
\end{aligned}
$$

and

$$
\begin{aligned}
&H\big(\widetilde{u}_2,\{Y_i\}_{i=1}^n,\lambda\big)-H\big(\widetilde{u}_1,\{Y_i\}_{i=1}^n,\lambda\big)\\
\le&\lambda-(1-\alpha)\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}\left(\omega_1-\frac{\omega_2\delta_1}{\delta+\epsilon_1+\delta_1}\right)^2+\frac{2(1-\alpha)}{\alpha}\sigma^2\log(n)\\
\le&\lambda+\frac{2(1-\alpha)}{\alpha}\sigma^2\log(n)-(1-\alpha)(1-\beta)^2\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}\omega_1^2, \qquad (33)
\end{aligned}
$$

where the last inequality of (33) follow from the observation that

$$
\frac{\omega_2\delta_1}{\delta+\epsilon_1+\delta_1}=\frac{\omega_2\delta_1}{\epsilon_1+\delta_1}\frac{\epsilon_1+\delta_1}{\delta+\epsilon_1+\delta_1}=\beta\omega_1\frac{\epsilon}{\epsilon+\delta}\frac{\epsilon_1+\delta_1}{\delta+\epsilon_1+\delta_1}\le\beta\omega_1/2.
$$

Equations (32) and (33) lead to that

$$
0\le H\big(\widetilde{u}_2,\{Y_i\}_{i=1}^n,\lambda\big)-H\big(u,\{Y_i\}_{i=1}^n,\lambda\big)
$$

$$\leq \frac{4}{\alpha}\sigma^2\log(n) + \frac{\omega_1^2\epsilon(1-\beta)^2\{(1+\alpha)\epsilon(\epsilon_1+\delta_1)-(1-\alpha)(\epsilon+\delta)(\delta+\epsilon_1+\delta_1)\}}{(\epsilon+\delta+\epsilon_1+\delta_1)(\epsilon+\delta)}$$
$$\leq \frac{4}{\alpha}\sigma^2\log(n) - \frac{\omega_1^2(1-\beta)^2\epsilon}{4},. \tag{34}$$

Plugging in (34) into (32) with a choice of $\alpha = 1/4$ yields that

$$\lambda \leq 50\sigma^2\log(n),$$

which is a contradiction.

Step 2. Suppose $\omega_1\omega_2 < 0$. Without loss of generality assume that with $\gamma \geq 1$ it holds that

$$\left|\frac{\epsilon\omega_1}{\epsilon+\delta}\right| = \gamma\left|\frac{\delta_1\omega_2}{\epsilon_1+\delta_1}\right|.$$

Since $\delta + \epsilon_1 = \eta_{k+1} - \eta_k + 1 > \Delta$, we have $\max\{\delta, \epsilon_1\} > \Delta/2$.

case 1. Suppose $\epsilon_1 > \Delta/2$. It follows from Lemma 9 that $\delta_1 < 8\lambda/\kappa_{k+1}^2$. Then,

$$\begin{aligned}
0 &\leq H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(u, \{Y_i\}_{i=1}^n, \lambda\big) \\
&\leq -\lambda + 2(1+\alpha)\alpha^{-1}\sigma^2\log(n) + (1+\alpha)\frac{|I_1||I_2|}{|I_1|+|I_2|}\left(\frac{\epsilon\omega_1}{\epsilon+\delta} - \frac{\delta_1\omega_2}{\epsilon_1+\delta_1}\right)^2 \\
&\leq -\lambda + 2(1+\alpha)\alpha^{-1}\sigma^2\log(n) + (\gamma+1)^2(1+\alpha)\frac{|I_1||I_2|}{|I_1|+|I_2|}\left(\frac{\delta_1\omega_2}{\epsilon_1+\delta_1}\right)^2
\end{aligned} \tag{35}$$

and

$$\begin{aligned}
&H\big(\widetilde{u}_3, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) \\
&\leq \lambda - (1-\alpha)\frac{|I_1''||I_2''|}{|I_1''|+|I_2''|}\left(\omega_2 - \frac{\omega_1\epsilon}{\epsilon+\delta+\epsilon_1}\right)^2 + \frac{2(1-\alpha)}{\alpha}\sigma^2\log(n) \\
&\leq \lambda - (1-\alpha)\frac{|I_1''||I_2''|}{|I_1''|+|I_2''|}\left(1 - \gamma\frac{\delta_1}{\epsilon_1+\delta_1}\frac{\epsilon+\delta}{\epsilon+\delta+\epsilon_1}\right)^2\omega_2^2 + \frac{2(1-\alpha)}{\alpha}\sigma^2\log(n).
\end{aligned} \tag{36}$$

Equations (35) and (36) lead to

$$\begin{aligned}
0 &\leq H\big(\widetilde{u}_3, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(u, \{Y_i\}_{i=1}^n, \lambda\big) \\
&\leq \frac{4\sigma^2\log(n)}{\alpha} + (\gamma+1)^2(1+\alpha)\frac{|I_1||I_2|}{|I_1|+|I_2|}\left(\frac{\delta_1\omega_2}{\epsilon_1+\delta_1}\right)^2 \\
&\qquad - (1-\alpha)\frac{|I_1''||I_2''|}{|I_1''|+|I_2''|}\left(1 - \gamma\frac{\delta_1}{\epsilon_1+\delta_1}\frac{\epsilon+\delta}{\epsilon+\delta+\epsilon_1}\right)^2\omega_2^2.
\end{aligned} \tag{37}$$

Then there exists a sufficiently small $c > 0$ such that (37) yields

$$c\omega_2^2\delta_1 < \sigma^2\log(n),$$

which can be plugged into (35) and shows that for a sufficiently large $C_1 > 0$,

$$\lambda \le C_1 \sigma^2 \log(n).$$

This contradicts the assumed condition on λ.

case 2. Suppose $\epsilon_1 \le \Delta/2$. It follows from Lemma 9 that $\epsilon_1 < 8\lambda/\kappa_{k+1}^2$. Then,

$$\begin{aligned} & H\big(\widetilde{u}_2, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widetilde{u}_1, \{Y_i\}_{i=1}^n, \lambda\big) \\ \le & \lambda - (1-\alpha)\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}\left(\omega_1 - \frac{\omega_2\delta_1}{\delta+\epsilon_1+\delta_1}\right)^2 + \frac{2(1-\alpha)}{\alpha}\sigma^2\log(n) \\ \le & \lambda - (1-\alpha)\frac{|I_1'||I_2'|}{|I_1'|+|I_2'|}\left(1 - \frac{1}{\gamma}\frac{\epsilon}{\epsilon+\delta}\frac{\epsilon_1+\delta_1}{\delta+\epsilon_1+\delta_1}\right)^2 \omega_1^2 + \frac{2(1-\alpha)}{\alpha}\sigma^2\log(n). \end{aligned} \tag{38}$$

Equations (35) and (38) lead to that there exists a sufficiently small $c > 0$ such that (37) yields

$$c\omega_2^2\epsilon_1 < \sigma^2\log(n),$$

which can be plugged into (35) and shows that for a sufficiently large $C_1 > 0$,

$$\lambda \le C_1\sigma^2\log(n).$$

This again contradicts the assumed condition on λ. □

Step 4: no changes

Suppose $I = (s_1, e] \in \widehat{\mathcal{P}}$ contains no true change point. By symmetry, it suffices to show that there exists a large enough constant $C > 0$ such that

$$s_1 - \eta_k + 1 \le C\lambda/\kappa_k^2. \tag{39}$$

Assume $I_0 = (s_0, s_1] \in \widehat{\mathcal{P}}$. We are to show the following.

(i) It is impossible that there is no true change point in $I_0 \cup I$. This will be shown in Lemma 12.
(ii) If there exist exactly two true change points in I_0, then (39) follows from Lemma 8.
(iii) If there exists one and only one change point $\eta_k \in I_0$ and $s_1 - \eta_k < \eta_k - s_0$, then (39) follows from Lemma 9.
(iv) If there exists one and only one change point $\eta_k \in I_0$ and $s_1 - \eta_k \ge \eta_k - s_0$, it follows from Lemma 10 that this is impossible in the event of $\mathcal{B}$.

Lemma 12. *Assume the inputs $\{Y_i\}_{i=1}^n$ satisfying Assumptions 1 and 2 and $\lambda = C_\lambda\sigma^2\log(n)$ with a sufficiently large $C_\lambda > 0$. Assume that $I = (s_1, e] \in \widehat{\mathcal{P}}$ contains no change point. Assume that $I_0 = (s_0, s_1] \in \widehat{\mathcal{P}}$. Then in the event $\mathcal{B}$, there must exist a change point in I_0.*

Proof. Let $J = I_0 \cup I$, $\widetilde{\mathcal{P}}$ be the interval partition such that

$$\widetilde{\mathcal{P}} = \widehat{\mathcal{P}} \cup \{J\} \setminus \{I_0, I\},$$

and $\widetilde{u}$ be the piecewise-constant vector induced by $\widetilde{\mathcal{P}}$.

Prove by contradiction, assuming that J contains no change points. Denote $\mu = \mathbb{E}(\overline{Y}_{I_0}) = \mathbb{E}(\overline{Y}_I)$. Then

$$\begin{aligned} 0 &\le H\big(\widetilde{u}, \{Y_i\}_{i=1}^n, \lambda\big) - H\big(\widehat{u}, \{Y_i\}_{i=1}^n, \lambda\big) = -\lambda + \frac{|I_0||I|}{|I_0|+|I|}(\overline{Y}_{I_0} - \overline{Y}_I)^2 \\ &\le -\lambda + \frac{|I_0||I|}{|I_0|+|I|}(\overline{Y}_{I_0} - \mu - \overline{Y}_I + \mu)^2 \\ &\le -\lambda + \sigma^2 \log(n), \end{aligned}$$

where the last inequality follows from the definition of the event $\mathcal{B}$, and results in a contradiction with the condition on λ. □

Appendix C: Proofs of the results in Section 4

C.1. Large probability events

Define the events

$$\mathcal{A}_1(\gamma) = \left\{ \sup_{0 \le s < t < e \le n} \left| \widetilde{Y}_t^{s,e} - \widetilde{f}_t^{s,e} \right| \le \gamma \right\}, \tag{40}$$

$$\mathcal{A}_2(\gamma) = \left\{ \sup_{0 \le s < e \le n} \frac{\left| \sum_{i=s+1}^e (Y_i - f_i) \right|}{\sqrt{e-s}} \le \gamma \right\}, \tag{41}$$

and

$$\mathcal{M} = \bigcap_{k=1}^K \big\{ s_m \in \mathcal{S}_k, e_m \in \mathcal{E}_k, \text{ for some } m \in \{1, \ldots, M\} \big\}, \tag{42}$$

where $\{s_m\}_{m=1}^M$ and $\{e_m\}_{m=1}^M$ are two sequences independently selected at random in (s, e) and satisfying $e_m - s_m \le C_R\Delta$, $\mathcal{S}_k = [\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\mathcal{E}_k = [\eta_k + \Delta/2, \eta_k + 3\Delta/4]$, $k = 1, \ldots, K$.

Lemma 13. *For $\{Y_i\}_{i=1}^n$ satisfying Assumption 1, it holds that*

$$\begin{aligned} \mathbb{P}\big\{\mathcal{A}_1(\gamma)\big\} &\ge 1 - e \cdot n^3 \exp\left(-c\gamma^2/\sigma^2\right), \\ \mathbb{P}\big\{\mathcal{A}_2(\gamma)\big\} &\ge 1 - e \cdot n^2 \exp\left(-c\gamma^2/\sigma^2\right) \end{aligned}$$

and

$$\mathbb{P}\big\{\mathcal{M}\big\} \ge 1 - \exp\left\{ \log\left(\frac{n}{\Delta}\right) - \frac{M\Delta}{4C_R n} \right\}.$$

Proof. Since for any suitable triples (s,t,e), both $\left|\widetilde{Y}_t^{s,e} - \widetilde{f}_t^{s,e}\right|$ and

$$(e-s)^{-1/2}\left|\sum_{i=s+1}^{e}(Y_i - f_i)\right|$$

can be written in the form $\left|\sum_{i=s+1}^{e} w_i X_i\right|$, where X_i's are centred sub-Gaussian random variables and w_i's satisfy $\sum_{i=s+1}^{e} w_i^2 = 1$.

It follows from Hoeffding inequality that there exists an absolute constant $c > 0$ only depending on σ such that

$$\mathbb{P}\{\mathcal{A}_1^c(\gamma)\} \le e \cdot n^3 \exp\left(-c\gamma^2/\sigma^2\right) \text{ and } \mathbb{P}\{\mathcal{A}_2^c(\gamma)\} \le e \cdot n^2 \exp\left(-c\gamma^2/\sigma^2\right).$$

Since the number of change points are bounded by n/Δ, it holds that

$$\begin{aligned}\mathbb{P}\{\mathcal{M}^c\} &\le \sum_{k=1}^{K}\prod_{m=1}^{M}\{1 - \mathbb{P}(s_m \in \mathcal{S}_k, e_m \in \mathcal{E}_k)\} \le K\{1-\Delta/(4C_R n)\}^M \\ &\le n/\Delta(1-\Delta/(4C_R n))^M \le \exp\left\{\log\left(\frac{n}{\Delta}\right) - \frac{M\Delta}{4C_R n}\right\}.\end{aligned}$$

□

C.2. Technical details for Step 1

Lemma 14. *Under Assumption 1, let $0 \le s < \eta_k < e \le n$ be any interval satisfying*

$$\min\{\eta_k - s, e - \eta_k\} \ge c_1\Delta,$$

with $c_1 > 0$. Then,

$$\max_{s<t<e}\left|\widetilde{f}_t^{s,e}\right| \ge (c_1/2)\kappa\Delta(e-s)^{-1/2}.$$

Proof. See Lemma 2.4 in Venkatraman (1992). □

Lemma 15. *Let $[s,e]$ contain only two change points η_k, η_{k+1}. Then*

$$\sup_{s\le t\le e}|\widetilde{f}_t^{s,e}| \le \sqrt{e-\eta_{k+1}}\kappa_{k+1} + \sqrt{\eta_k - s}\kappa_k.$$

Proof. Consider the sequence $\{g_t\}_{t=s+1}^{e}$ be such that

$$g_t = \begin{cases} f_{\eta_k}, & \text{if} \quad s+1 \le t < \eta_k, \\ f_t, & \text{if} \quad \eta_k \le t \le e. \end{cases}$$

For any $t \ge \eta_k$,

$$\widetilde{f}_t^{s,e} - \widetilde{g}_t^{s,e}$$

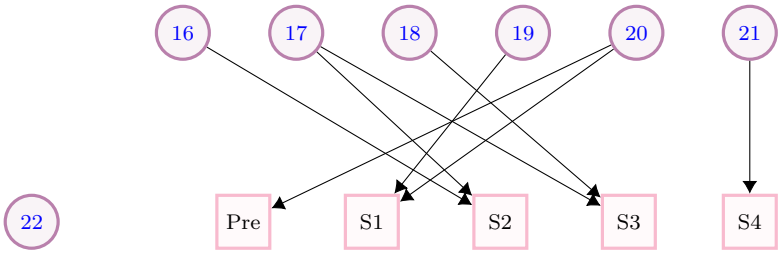

FIG 2. *Road map to complete the Step 2 in the proof of Theorem 4. The circles are lemmas, and the squares are the steps in the proof of Lemma 22. The directed edges mean the heads of the edges are used in the tails of the edges.*

$$
\begin{aligned}
&= \sqrt{\frac{e-t}{(e-s)(t-s)}} \left(\sum_{i=s+1}^{t} f_i - \sum_{i=s+1}^{\eta_k} f_{\eta_k} - \sum_{i=\eta_k+1}^{t} f_i \right) \\
&- \sqrt{\frac{t-s}{(e-s)(e-t)}} \left(\sum_{i=t+1}^{e} f_i - \sum_{i=t+1}^{e} f_i \right) \\
&= \sqrt{\frac{e-t}{(e-s)(t-s)}} (\eta_k - s)(f_{\eta_k} - f_{\eta_{k-1}}).
\end{aligned}
$$

So for $t \geq \eta_k$, $|\widetilde{f}_t^{s,e} - \widetilde{g}_t^{s,e}| \leq \sqrt{\eta_k - s}\kappa_k$. Since $\sup_{s \leq t \leq e} |\widetilde{f}_t^{s,e}| = \max\{|\widetilde{f}_{\eta_k}^{s,e}|, |\widetilde{f}_{\eta_{k+1}}^{s,e}|\}$, and that

$$
\begin{aligned}
\max\{|\widetilde{f}_{\eta_k}^{s,e}|, |\widetilde{f}_{\eta_{k+1}}^{s,e}|\} &\leq \sup_{s \leq t \leq e} |\widetilde{g}_t^{s,e}| + \sqrt{\eta_k - s}\kappa_k \\
&\leq \sqrt{e - \eta_{k+1}}\kappa_{k+1} + \sqrt{\eta_r - s}\kappa_k
\end{aligned}
$$

where the last inequality follows form the fact that g_t has only one change point in $[s, e]$. □

C.3. Technical details for Step 2

In this section, eight results will be provided. Before we go into details, we show the road map leading to complete the proof of Theorem 4 in Figure 2.

Lemma 16. *Suppose $(s, e) \subset (0, n)$ is a generic interval satisfying*

$$\eta_{k-1} \leq s \leq \eta_k \leq \ldots \leq \eta_{k+q} \leq e \leq \eta_{k+q+1}, \quad q \geq 0.$$

Then there exists a continuous function $\widetilde{F}_t^{s,e} : [s, e] \to \mathbb{R}$ such that $\widetilde{F}_t^{s,e} = \widetilde{f}_t^{s,e}$ for every $t \in [s, e] \cap \mathbb{Z}$ with the following additional properties.

(i) $|\widetilde{F}_t^{s,e}|$ is maximized at the change points within $[s, e]$. In other words,

$$\arg\max_{s \leq t \leq e} |\widetilde{F}_t^{s,e}| \cap \{\eta_k, \ldots, \eta_{k+q}\} \neq \emptyset.$$

(ii) If $\widetilde{F}_t^{s,e} > 0$ for some $t \in (s,e)$, then $\widetilde{F}_t^{s,e}$ is either monotonic or decreases and then increases within each of the interval $(s,\eta_k),\ldots,(\eta_{k+q},e)$.

The proof of Lemma 16 can be found in Lemmas 2.2 and 2.3 of Venkatraman (1992). We remark that if $\widetilde{F}_t^{s,e} \leq 0$ for all $t \in (s,e)$, then it suffices to consider the time series $\{-f_i\}_{i=1}^n$ and a similar result as in the second part of Lemma 16 still holds.

Our next lemma is an adaptation of a result first obtained by Venkatraman (1992), which quantifies how fast the CUSUM statistics decays around a good change point.

Lemma 17 (Venkatraman (1992) Lemma 2.6)**.** *Let $[s,e] \subset [1,n]$ be any generic interval. For some $c_1, c_2 > 0$ and $\gamma > 0$ such that*

$$\min\{\eta_k - s, e - \eta_k\} \geq c_1\Delta, \tag{43}$$

$$\widetilde{f}_{\eta_k} \geq c_2\kappa\Delta(e-s)^{-1/2}, \tag{44}$$

and suppose there exists a sufficiently small constant $c_3 > 0$ such that

$$\max_{s\leq t\leq e} |\widetilde{f}_t^{s,e}| - \widetilde{f}_{\eta_k}^{s,e} \leq 2\gamma \leq c_3\kappa\Delta^3(e-s)^{-5/2}. \tag{45}$$

Then there exists an absolute constant $c > 0$ such that if the point $d \in [s,e]$ is such that $|d - \eta_k| \leq c_1\Delta/16$, then

$$\widetilde{f}_{\eta_k}^{s,e} - \widetilde{f}_d^{s,e} > c\widetilde{f}_{\eta_k}^{s,e}|\eta_k - d|\Delta(e-s)^{-2}.$$

Proof. Without loss of generality, assume that $d \geq \eta_k$. Following the argument of Venkatraman (1992) Lemma 2.6, it suffices to consider two cases: (1) $\eta_{k+1} > e$, and (2) $\eta_{k+1} \leq e$.

Case 1. Let E_l be defined as in the case 1 in Venkatraman (1992) Lemma 2.6. There exists a $c' > 0$ such that, for every $d \in [\eta_k, \eta_k + c_1\Delta/16]$, $\widetilde{f}_{\eta_k}^{s,e} - \widetilde{f}_d^{s,e}$ (which in the notation of Venkatraman (1992) is the term E_l) can be written as

$$\widetilde{f}_{\eta_k}^{s,e}|d-\eta_k|\frac{e-s}{\sqrt{e-\eta_k}\sqrt{\eta_k - s + (d-\eta_k)}} \times \frac{1}{\sqrt{(\eta_k - s + (d-\eta_k))(e-\eta_k)} + \sqrt{(\eta_k - s)(e - \eta_k - (d-\eta_k))}}.$$

Using the inequality $(e-s) \geq 2c_1\Delta$, the previous expression is lower bounded by

$$c'|d-\eta_k|\widetilde{f}_{\eta_k}^{s,e}\Delta(e-s)^{-2}.$$

Case 2. Let $h = c_1\Delta/8$ and $l = d - \eta_k \leq h/2$. Then, following closely the initial calculations for case 2 of Lemma 2.6 of Venkatraman (1992), we obtain that

$$\widetilde{f}_{\eta_k}^{s,e} - \widetilde{f}_d^{s,e} \geq E_{1l}(1 + E_{2l}) + E_{3l},$$

where

$$E_{1l} = \frac{\widetilde{f}^{s,e}_{\eta_k} l(h-l)}{\sqrt{(\eta_k - s + l)(e - \eta_k - l)}} \times \frac{1}{\left(\sqrt{(\eta_k - s + l)(e - \eta_k - l)} + \sqrt{(\eta_k - s)(e - \eta_k)}\right)},$$

$$E_{2l} = \frac{((e-\eta_k-h)-(\eta_k-s))((e-\eta_k-h)-(\eta_k-s)-l)}{\sqrt{(\eta_k - s + l)(e - \eta_k - l)} + \sqrt{(\eta_k - s + h)(e - \eta_k - h)}} \times \frac{1}{\sqrt{(\eta_k - s)(e - \eta_k)} + \sqrt{(\eta_k - s + h)(e - \eta_k - h)}},$$

and

$$E_{3l} = -\frac{(\widetilde{f}^{s,e}_{\eta_k+h} - \widetilde{f}^{s,e}_{\eta_k})l}{h} \sqrt{\frac{(\eta_k - s + h)(e - \eta_k - h)}{(\eta_k - s + l)(e - \eta_k - l)}}.$$

Since $h = c_1\Delta/8$ and $l \le h/2$, it holds that

$$E_{1l} \ge (c_1/16)\widetilde{f}^{s,e}_{\eta_k} |d - \eta| \Delta (e-s)^{-2}.$$

Observe that

$$\eta_k - s \le \eta_k - s + l \le \eta_k - s + h \le 9(\eta_k - s)/8 \tag{46}$$

and

$$e - \eta_k \ge e - \eta_k - l \ge e - \eta_k - h \ge 7(e - \eta_k)/8. \tag{47}$$

Thus

$$E_{2l} = \frac{((e-\eta_k-h)-(\eta_k-s))^2 + l(h+\eta_k-s) - l(e-\eta_k)}{\left(\sqrt{(\eta_k - s + l)(e - \eta_k - l)} + \sqrt{(\eta_k - s + h)(e - \eta_k - h)}\right)} \times \frac{1}{\left(\sqrt{(\eta_k - s)(e - \eta_k)} + \sqrt{(\eta_k - s + h)(e - \eta_k - h)}\right)}$$

$$\ge \frac{-l(e-\eta_k)}{(\eta_k - s + h)(e - \eta_k - h)} \ge \frac{-l(e-\eta_k)}{(\eta_k - s)(7/8)(e - \eta_k)} \ge -1/2,$$

where (46) and (47) are used in the second inequality and the fact that $l \le h/2 \le c_1\Delta/16 \le (\eta_k - s)/16$ is used in the last inequality.

For E_{3l}, observe that

$$\widetilde{f}^{s,e}_{\eta_k+h} - \widetilde{f}^{s,e}_{\eta_k} \le |\widetilde{f}^{s,e}_{\eta_k+h}| - \widetilde{f}^{s,e}_{\eta_k} \le \max_{s \le t \le e} |\widetilde{f}^{s,e}_t| - \widetilde{f}^{s,e}_{\eta_k} \le 2\gamma.$$

This combines with (43) and that $l/2 \le h = c_1\Delta/8$, implying that

$$\eta_k - s \le \eta_k - s + l \le \eta_k - s + h \le 9(\eta_k - s)/8$$

and

$$e-\eta_k \geq e-\eta_k-l \geq e-\eta_k-h \geq 7(e-\eta_k)/8.$$

Therefore, with a sufficiently small constant $c''>0$, it holds that

$$\begin{aligned} E_{3l} &\geq -\frac{2(d-\eta_k)\gamma}{c_1\Delta/8}\sqrt{\frac{(9/8)(\eta_k-s)(e-\eta_k)}{(\eta_k-s)(7/8)(e-\eta_k)}} \geq -\frac{32(d-\eta_k)\gamma}{c_1\Delta} \\ &\geq -(c''/4)\widetilde{f}^{s,e}_{\eta_k}(d-\eta_k)\Delta(e-s)^{-2}, \end{aligned}$$

where the first inequality follows from (46) and (46), and the last inequality follows from (44) and (45). Thus,

$$\widetilde{f}^{s,e}_{\eta_k}-\widetilde{f}^{s,e}_{d} \geq E_{1l}(1+E_{2l})+E_{3l} \geq (c''/4)\widetilde{f}^{s,e}_{\eta_k}|\eta_k-d|\Delta(e-s)^{-2}. \qquad \square$$

Lemma 18. *Suppose $[s,e]\subset[1,n]$ such that $e-s\leq C_R\Delta$, and that*

$$\eta_{k-1}\leq s\leq \eta_k\leq\ldots\leq\eta_{k+q}\leq e\leq\eta_{k+q+1}, \quad q\geq 0.$$

Denote

$$\kappa^{s,e}_{\max}=\max\{\eta_p-\eta_{p-1}:\ k\leq p\leq k+q\}.$$

Then for any $k-1\leq p\leq k+q$, it holds that

$$\left|\frac{1}{e-s}\sum_{i=s}^{e}f_i-f_{\eta_p}\right|\leq C_R\kappa^{s,e}_{\max}.$$

Proof. Since $e-s\leq C_R\Delta$, the interval $[s,e]$ contains at most C_R+1 change points. Observe that

$$\begin{aligned} &\left|\frac{1}{e-s}\sum_{i=s}^{e}f_i-f_{\eta_p}\right| \\ =&\frac{1}{e-s}\left|\sum_{i=s}^{\eta_k}(f_{\eta_{k-1}}-f_{\eta_p})+\sum_{i=\eta_k+1}^{\eta_{k+1}}(f_{\eta_k}-f_{\eta_p})+\ldots+\sum_{i=\eta_{k+q}+1}^{e}(f_{\eta_{k+q}}-f_{\eta_p})\right| \\ \leq&\frac{1}{e-s}\sum_{i=s}^{\eta_k}|p-k|\kappa^{s,e}_{\max}+\sum_{i=\eta_k+1}^{\eta_{k+1}}|p-k-1|\kappa^{s,e}_{\max}+\ldots \\ &\qquad+\sum_{i=\eta_{k+q}+1}^{e}|p-k-q-1|\kappa^{s,e}_{\max} \\ \leq&\frac{1}{e-s}\sum_{i=s}^{e}(C_R+1)\kappa^{s,e}_{\max}, \end{aligned}$$

where $|p_1-p_2|\leq C_R+1$ for any $\eta_{p_1},\eta_{p_2}\in[s,e]$ is used in the last inequality. $\square$

Lemma 19. *If η_k is the only change point in (s,e), then*

$$|\widetilde{f}^{s,e}_{\eta_k}|=\sqrt{\frac{(\eta_k-s)(e-\eta_k)}{e-s}}\kappa_k\leq\sqrt{\min\{\eta_k-s,e-\eta_k\}}\kappa_k.$$

Lemma 20. *Let $(s, e) \subset (0, n)$ contains two or more change points such that*

$$\eta_{k-1} \leq s \leq \eta_k \leq \ldots \leq \eta_{k+q} \leq e \leq \eta_{k+q+1}, \quad q \geq 1.$$

If $\eta_k - s \leq c_1\Delta$, for $c_1 > 0$, then

$$|\widetilde{f}^{s,e}_{\eta_k}| \leq \sqrt{c_1}|\widetilde{f}^{s,e}_{\eta_{k+1}}| + 2\kappa_k\sqrt{\eta_k - s}.$$

Proof. Consider the sequence $\{g_t\}_{t=s+1}^{e}$ be such that

$$g_t = \begin{cases} f_{\eta_{r+1}}, & s+1 \leq t \leq \eta_k, \\ f_t, & \eta_k + 1 \leq t \leq e. \end{cases}$$

For any $t \geq \eta_r$, it holds that

$$\widetilde{f}^{s,e}_{\eta_k} - \widetilde{g}^{s,e}_{\eta_k} = \sqrt{\frac{(e-s)-t}{(e-s)(t-s)}}(\eta_k - s)(f_{\eta_{k+1}} - f_{\eta_k}) \leq \sqrt{\eta_k - s}\kappa_k.$$

Thus,

$$\begin{aligned} |\widetilde{f}^{s,e}_{\eta_k}| &\leq |\widetilde{g}^{s,e}_{\eta_k}| + \sqrt{\eta_k - s}\kappa_k \leq \sqrt{\frac{(\eta_k - s)(e - \eta_{k+1})}{(\eta_{k+1} - s)(e - \eta_k)}}|\widetilde{g}^{s,e}_{\eta_{k+1}}| + \sqrt{\eta_k - s}\kappa_k \\ &\leq \sqrt{\frac{c_1\Delta}{\Delta}}|\widetilde{g}^{s,e}_{\eta_{k+1}}| + \sqrt{\eta_k - s}\kappa_k \leq \sqrt{c_1}|\widetilde{f}^{s,e}_{\eta_{k+1}}| + 2\sqrt{\eta_k - s}\kappa_k, \end{aligned}$$

where the first inequality follows from the observation that the first change point of g_t in (s, e) is at η_{k+1}. □

For a pair (s, e) of positive integers with $s < e$, let $\mathcal{W}^{s,e}_d$ be the two dimensional linear subspace of $\mathbb{R}^{(e-s)}$ spanned by the vectors

$$u_1 = (\underbrace{1, \ldots, 1}_{d-s}, \underbrace{0, \ldots, 0}_{e-d})^\top \text{ and } u_2 = (\underbrace{0, \ldots, 0}_{d-s}, \underbrace{1, \ldots, 1}_{e-d})^\top.$$

For clarity, in the lemma below, we will use $\langle \cdot, \cdot \rangle$ to denote the inner product of two vectors in the Euclidean space.

Lemma 21. *For $x = (x_{s+1}, \ldots, x_e)^\top \in \mathbb{R}^{(e-s)}$, let $\mathcal{P}^{s,e}_d(x)$ be the projection of x onto $\mathcal{W}^{s,e}_d$.*

(i) The projection $\mathcal{P}^{s,e}_d(x)$ satisfies

$$\mathcal{P}^{s,e}_d(x) = \frac{1}{e-s}\sum_{i=s+1}^{e} x_i + \langle x, \psi^{s,e}_d\rangle\psi^{s,e}_d,$$

where $\langle \cdot, \cdot \rangle$ is the inner product in Euclidean space, and

$$\psi^{s,e}_d = ((\psi^{s,e}_d)_s, \ldots, (\psi^{s,e}_d)_{e-s})^\top$$

with

$$(\psi_d^{s,e})_i = \begin{cases} \sqrt{\frac{e-d}{(e-s)(d-s)}}, & i = s+1, \ldots, d, \\ -\sqrt{\frac{d-s}{(e-s)(e-d)}}, & i = d+1, \ldots, e, \end{cases}$$

i.e. the i-th entry of $\mathcal{P}_d^{s,e}(x)$ *satisfies*

$$\mathcal{P}_d^{s,e}(x)_i = \begin{cases} \frac{1}{d-s}\sum_{j=s+1}^{d} x_j, & i = s+1, \ldots, d, \\ \frac{1}{e-d}\sum_{j=d+1}^{e} x_j, & i = d+1, \ldots, e. \end{cases}$$

(ii) Let $\bar{x} = \frac{1}{e-s}\sum_{i=s+1}^{e} x_i$. *Since* $\langle \bar{x}, \psi_d^{s,e} \rangle = 0$, *it holds that*

$$\|x - \mathcal{P}_d^{s,e}(x)\|^2 = \|x - \bar{x}\|^2 - \langle x, \psi_d^{s,e} \rangle^2. \tag{48}$$

Proof. The results hold following the fact that the projection matrix of subspace $\mathcal{W}_d^{s,e}$ is

$$P_{\mathcal{W}_d^{s,e}}^{s,e} = \begin{pmatrix} 1/(d-s) & \cdots & 1/(d-s) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/(d-s) & \cdots & 1/(d-s) & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/(e-d) & \cdots & 1/(e-d) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 1/(e-d) & \cdots & 1/(e-d) \end{pmatrix}. \quad \square$$

Lemma 22. *Under Assumption 1, let* (s_0, e_0) *be an interval with* $e_0 - s_0 \le C_R\Delta$ *and contain at lest one change point* η_k *such that*

$$\eta_{k-1} \le s_0 \le \eta_k \le \ldots \le \eta_{k+q} \le e_0 \le \eta_{k+q+1}, \quad q \ge 0.$$

Suppose that there exists k' *such that* $\min\{\eta_{k'} - s_0,\, e_0 - \eta_{k'}\} \ge \Delta/16$. *Let* $\kappa_{\max}^{s,e} = \max\{\kappa_p : \min\{\eta_p - s_0, e_0 - \eta_p\} \ge \Delta/16\}$. *Consider any generic* $[s, e] \subset [s_0, e_0]$, *satisfying*

$$\min\{\eta_k - s_0, e_0 - \eta_k\} \ge \Delta/16 \quad \textit{for all } \eta_k \in [s, e].$$

Let $b \in \arg\max_{s<t<e} |\widetilde{Y}_t^{s,e}|$. *For some* $c_1 > 0$ *and* $\gamma > 0$, *suppose that*

$$|\widetilde{Y}_b^{s,e}| \ge c_1 \kappa_{\max}^{s,e} \sqrt{\Delta}, \tag{49}$$

$$\sup_{s<t<e} |\widetilde{Y}_t^{s,e} - \widetilde{f}_t^{s,e}| \le \gamma, \tag{50}$$

and

$$\sup_{s_1<t<e_1} \frac{1}{\sqrt{e_1 - s_1}} \left| \sum_{t=s_1+1}^{e_1} (Y_t - f_t) \right| \le \gamma. \tag{51}$$

If there exists a sufficiently small $0 < c_2 < c_1/2$ *such that*

$$\gamma \le c_2 \kappa_{\max}^{s,e} \sqrt{\Delta}, \tag{52}$$

then there exists a change point $\eta_k \in (s, e)$ *such that*

$$\min\{e - \eta_k, \eta_k - s\} > \Delta/4 \quad \textit{and} \quad |\eta_k - b| \le C_3 \gamma^2 \kappa_k^{-2}.$$

Proof. Without loss of generality, assume that $\widetilde{f}_b^{s,e} > 0$ and that $\widetilde{f}_t^{s,e}$ is locally decreasing at b. Observe that there has to be a change point $\eta_k \in [s, b]$, or otherwise $\widetilde{f}_b^{s,e} > 0$ implies that $\widetilde{f}_t^{s,e}$ is decreasing, as a consequence of Lemma 20. Thus, if $s \le \eta_k \le b \le e$, then

$$\widetilde{f}_{\eta_k}^{s,e} \ge \widetilde{f}_b^{s,e} \ge |\widetilde{Y}_b^{s,e}| - \gamma \ge c_1 \kappa_{\max}^{s,e} \sqrt{\Delta} - c_2 \kappa_{\max}^{s,e} \sqrt{\Delta} \ge (c_1/2) \kappa_{\max}^{s,e} \sqrt{\Delta}. \quad (53)$$

Observe that $e - s \le e_0 - s_0 \le C_R \Delta$ and that (s, e) has to contain at least one change point or otherwise $|\widetilde{f}_{\eta_k}^{s,e}| = 0$ which contradicts (53).

We decompose the rest of the proof in four steps. **Step 1** shows that η_k is far enough away from end points s and e. **Step 2** utilizes Lemma 17 – the machinery originally developed for BS in Venkatraman (1992) – to show that b is not far away from η_k. This is actually a consistent estimator, but not optimal. **Step 3** brings in the WBS techniques to refine the error bound, which is *de facto* optimal. The proof is completed in **Step 4**.

Step 1. In this step, we are to show that $\min\{\eta_k - s, e - \eta_k\} \ge \min\{1, c_1^2\}\Delta/16$.

Suppose η_k is the only change point in (s, e). So $\min\{\eta_k - s, e - \eta_k\} \ge \min\{1, c_1^2\}\Delta/16$ must hold or otherwise it follows from Lemma 19, we have

$$|\widetilde{f}_{\eta_k}^{s,e}| < \frac{c_1}{4} \kappa_k \sqrt{\Delta} \le \frac{c_1}{2} \kappa_{\max}^{s,e} \sqrt{\Delta},$$

which contradicts (53).

Suppose (s, e) contains at least two change points. Then $\eta_k - s \le \min\{1, c_1^2\}\Delta/16$ implies that η_k is the first change point in $[s, e]$. Therefore,

$$\begin{aligned} |\widetilde{f}_{\eta_k}^{s,e}| &\le \frac{1}{4} |\widetilde{f}_{\eta_{k+1}}^{s,e}| + 2\kappa_k \sqrt{\eta_k - s} \le \frac{1}{4} \max_{s<t<e} |\widetilde{f}_t^{s,e}| + \frac{c_1}{2} \kappa_k \sqrt{\Delta} \\ &\le \frac{1}{4} |\widetilde{Y}_b^{s,e}| + \gamma + \frac{c_1}{2} \kappa_{\max}^{s,e} \sqrt{\Delta} \le \frac{3}{4} |\widetilde{Y}_b^{s,e}| + \gamma < |\widetilde{Y}_b^{s,e}| - \gamma, \end{aligned}$$

where the first inequality follows from Lemma 20, the fourth inequality follows from (49), and the last inequality holds when c_2 is sufficiently small. This contradicts with (53).

Step 2. By Lemma 17 there exists d such that $d \in [\eta_k, \eta_k + \gamma\sqrt{\Delta}(\kappa_{\max}^{s,e})^{-1}]$ and that $\widetilde{f}_{\eta_k}^{s,e} - \widetilde{f}_d^{s,e} > 2\gamma$. For the sake of contradiction, suppose $b \ge d$. Then

$$\widetilde{f}_b^{s,e} \le \widetilde{f}_d^{s,e} < \widetilde{f}_{\eta_k}^{s,e} - 2\gamma \le \max_{s<t<e} |\widetilde{f}_t^{s,e}| - 2\gamma \le \max_{s<t<e} |\widetilde{Y}_t^{s,e}| + \gamma - 2\gamma = |\widetilde{Y}_b^{s,e}| - \gamma,$$

where the first inequality follows from Lemma 16, which ensures that $\widetilde{f}_t^{s,e}$ is decreasing on $[\eta_k, b]$ and $d \in [\eta_k, b]$. This is a contradiction to (53). Thus $b \in [\eta_k, \eta_k + \gamma\sqrt{\Delta}(\kappa_{\max}^{s,e})^{-1}]$.

Step 3. Let $f^{s,e} = (f_{s+1}, \ldots, f_e)^\top \in \mathbb{R}^{(e-s)}$ and $Y^{s,e} = (Y_{s+1}, \ldots, Y_e)^\top \in \mathbb{R}^{(e-s)}$. By the definition of b, it holds that

$$\left\|Y^{s,e} - \mathcal{P}_b^{s,e}(Y^{s,e})\right\|^2 \le \left\|Y^{s,e} - \mathcal{P}_{\eta_k}^{s,e}(Y^{s,e})\right\|^2 \le \left\|Y^{s,e} - \mathcal{P}_{\eta_k}^{s,e}(f^{s,e})\right\|^2.$$

For the sake of contradiction, throughout the rest of this argument suppose that, for some sufficiently large constant $C_3 > 0$ to be specified,

$$\eta_k + \max\{C_3\gamma^2\kappa_k^{-2}, \delta\} < b. \tag{54}$$

(This will of course imply that $\eta_k + \max\{C_3\gamma^2(\kappa^{s,e}_{\max})^{-2}, \delta\} < b$). We will show that this leads to the bound

$$\left\|Y^{s,e} - \mathcal{P}_b^{s,e}(Y^{s,e})\right\|^2 > \left\|Y^{s,e} - \mathcal{P}_{\eta_k}^{s,e}(f^{s,e})\right\|^2, \tag{55}$$

which is a contradiction.

To derive (55) from (54), we note that $\min\{e-\eta_k, \eta_k - s\} \geq \min\{1, c_1^2\}\Delta/16$ and that $|b - \eta_k| \leq \gamma\sqrt{\Delta}(\kappa^{s,e}_{\max})^{-1}$ implies that

$$\min\{e-b, b-s\} \geq \min\{1, c_1^2\}\Delta/16 - \gamma\sqrt{\Delta}(\kappa^{s,e}_{\max})^{-1} \geq \min\{1, c_1^2\}\Delta/32, \tag{56}$$

where the last inequality follows from (52) and holds for an appropriately small $c_2 > 0$.

Equation (55) is in turn implied by

$$2\langle\varepsilon^{s,e}, \mathcal{P}_b(Y^{s,e}) - \mathcal{P}_{\eta_k}(f^{(s,e)})\rangle < \|f^{s,e} - \mathcal{P}_b(f^{s,e})\|^2 - \|f^{s,e} - \mathcal{P}_{\eta_k}(f^{s,e})\|^2, \tag{57}$$

where $\varepsilon^{s,e} = Y^{s,e} - f^{s,e}$. By (48), the right hand side of (57) satisfies the relationship with sufficiently small absolute constants $c, c' > 0$,

$$\begin{aligned}
&\|f^{s,e} - \mathcal{P}_b(f^{s,e})\|^2 - \|f^{s,e} - \mathcal{P}_{\eta_k}(f^{s,e})\|^2 = \langle f^{s,e}, \psi_{\eta_k}\rangle^2 - \langle f^{s,e}, \psi_b\rangle^2 \\
=&(\widetilde{f}^{s,e}_{\eta_k})^2 - (\widetilde{f}^{s,e}_b)^2 \geq (\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_b)|\widetilde{f}^{s,e}_{\eta_k}| \geq c|d-\eta_k|(\widetilde{f}^{s,e}_{\eta_k})^2\Delta^{-1} \\
\geq& c'|d-\eta_k|(\kappa^{s,e}_{\max})^2,
\end{aligned}$$

where Lemma 17 and (53) are used in the second and third inequalities. The left hand side of (57) can in turn be rewritten as

$$\begin{aligned}
2\langle\varepsilon^{s,e}, \mathcal{P}_b(X^{s,e}) - \mathcal{P}_{\eta_k}(f^{s,e})\rangle = &\, 2\langle\varepsilon^{s,e}, \mathcal{P}_b(X^{s,e}) - \mathcal{P}_b(f^{s,e})\rangle \\
&+2\langle\varepsilon^{s,e}, \mathcal{P}_b(f^{s,e}) - \mathcal{P}_{\eta_k}(f^{s,e})\rangle.
\end{aligned} \tag{58}$$

The second term on the right hand side of the previous display can be decomposed as

$$\begin{aligned}
\langle\varepsilon^{s,e}, \mathcal{P}_b(f^{s,e}) - \mathcal{P}_{\eta_k}(f^{s,e})\rangle = &\left(\sum_{i=s+1}^{\eta_k} + \sum_{i=\eta_k+1}^{b} + \sum_{i=b+1}^{e}\right) \\
&\times \varepsilon_i^{s,e}\left(\mathcal{P}_b(f^{s,e})_i - \mathcal{P}_{\eta_k}(f^{s,e})_i\right) \\
&= I + II + III.
\end{aligned}$$

In order to bound the terms I, II and III, observe that, since $e - s \leq e_0 - s_0 \leq C_R\Delta$, the interval $[s, e]$ must contain at most $C_R + 1$ change points.

Step 3.1. We can write

$$I = \sqrt{\eta_k - s}\left(\frac{1}{\sqrt{\eta_k - s}}\sum_{i=s+1}^{\eta_k}\varepsilon_i^{s,e}\right)\left(\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - \frac{1}{\eta_k - s}\sum_{i=s+1}^{\eta_k} f_i\right).$$

Thus,

$$\begin{aligned}
&\left|\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - \frac{1}{\eta_k - s}\sum_{i=s+1}^{\eta_k} f_i\right| \\
=&\left|\frac{(\eta_k - s)(\sum_{i=s+1}^{\eta_k} f_i + \sum_{i=\eta_k+1}^{b} f_i) - (b-s)\sum_{i=s+1}^{\eta_k} f_i}{(b-s)(\eta_k - s)}\right| \\
=&\left|\frac{(\eta_k - b)\sum_{i=s+1}^{\eta_k} f_i + (\eta_k - s)\sum_{i=\eta_k+1}^{b} f_i)}{(b-s)(\eta_k - s)}\right| \\
=&\left|\frac{(\eta_k - b)\sum_{i=s+1}^{\eta_k} f_i + (\eta_k - s)(b-\eta_k) f_{\eta_k+1})}{(b-s)(\eta_k - s)}\right| \\
=&\frac{b-\eta_k}{b-s}\left|-\frac{1}{\eta_k - s}\sum_{i=s+1}^{\eta_k} f_i + f_{\eta_{k+1}}\right| \le \frac{b-\eta_k}{b-s}(C_R+1)\kappa_{\max}^{s,e}
\end{aligned}$$

where Lemma 18 is used in the last inequality. It follows from Equation (51) that

$$|I| \le \sqrt{\eta_k - s}\gamma\frac{|b-\eta_k|}{b-s}(C_R+1)\kappa_{\max}^{s,e} \le \frac{4\sqrt{2}}{\min\{1, c_1\}}|b-\eta_k|\Delta^{-1/2}\gamma(C_R+1)\kappa_{\max}^{s,e},$$

where (56) is used in the last inequality.

Step 3.2. For the second term II, we have that

$$\begin{aligned}
|II| =& \left|\sqrt{b-\eta_k}\left(\frac{1}{\sqrt{b-\eta_k}}\sum_{i=\eta_k+1}^{d}\varepsilon_i^{s,e}\right)\left(\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - \frac{1}{e-\eta_k}\sum_{i=\eta_k+1}^{e} f_i\right)\right| \\
\le& \sqrt{b-\eta_k}\gamma\left(\left|f_{\eta_k} - f_{\eta_{k+1}}\right| + \left|\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - f_{\eta_k}\right| + \left|\frac{1}{e-\eta_k}\sum_{i=\eta_k+1}^{e} f_i - f_{\eta_{k+1}}\right|\right) \\
\le& \sqrt{b-\eta_k}(\kappa_{\max}^{s,e} + (C_R+1)\kappa_{\max}^{s,e} + (C_R+1)\kappa_{\max}^{s,e}),
\end{aligned}$$

where the first inequality follows from (56) and (51), and the second inequality from Lemma 18.

Step 3.3. Finally, we have that

$$III = \sqrt{e-b}\left(\frac{1}{e-b}\sum_{i=b+1}^{e}\varepsilon_i^{s,e}\right)\left(\frac{1}{e-\eta_k}\sum_{i=\eta_k+1}^{e} f_i - \frac{1}{e-b}\sum_{i=b+1}^{e} f_i\right).$$

Therefore,

$$|III| \leq \sqrt{e-b}\gamma \frac{b-\eta_k}{e-b}(C_R+1)\kappa^{s,e}_{\max} \leq \frac{4\sqrt{2}}{\min\{1,c_1\}}|b-\eta_k|\Delta^{-1/2}\gamma(C_R+1)\kappa^{s,e}_{\max}.$$

Step 4. Using the first part of Lemma 21, the first term on the right hand side of (58) can be bounded as

$$\langle \varepsilon^{s,e}, \mathcal{P}_d(X^{s,e}) - \mathcal{P}_d(f^{s,e})\rangle \leq \gamma^2.$$

Thus (57) holds if

$$|b-\eta_k|(\kappa^{s,e}_{\max})^2 \geq C \max\left\{|b-\eta_k|\Delta^{-1/2}\gamma\kappa^{s,e}_{\max}, \quad \sqrt{b-\eta_k}\gamma\kappa^{s,e}_{\max}, \quad \gamma^2\right\}.$$

Since $\gamma \leq c_3\sqrt{\Delta}\kappa$, the first inequality holds. The second inequality follows from $|b-\eta_k| \geq C_3\gamma^2(\kappa_k)^{-2} \geq C_3\gamma^2(\kappa^{s,e}_{\max})^{-2}$, as assumed in (54). This completes the proof. □

References

Aston, J. A. D. and Kirch, C. (2014). Efficiency of change point tests in high dimensional settings. *arXiv preprint arXiv:1409.1771*.

Aue, A., Hömann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate nonlinear time series models. *The Annals of Statistics* **37** 4046-4087. MR2572452

Avanesov, V. and Buzun, N. (2016). Change-point detection in high-dimensional covariance structure. *arXiv preprint arXiv:1610.03783*. MR3861282

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016). Narrowest-Over-Threshold detection of multiple change-points and change-point-like feature. *arXiv preprint arXiv:1609.00293*. MR3961502

Boysen, L., Kempe, A., Liebscher, V., Munk, A. and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics* **37** 157–183. MR2488348

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica* **1** 409-428. MR3076173

Cho, H. (2015). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics* in press. MR3522667

Cho, H. and Fryzlewicz, P. (2015). Multiple change-point detection for high-dimensional time series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 475-507. MR3310536

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association* **101** 223-239. MR2268041

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics* 124–152. MR1833961

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *The Annals of Statistics* **36** 1758–1785. MR2435455

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli* **24** 526–564. MR3706768

Enikeeva, F., Munk, A. and Werner, F. (2018). Bump detection in heterogeneous Gaussian regression. *Bernoulli* **24** 1266–1306. MR3706794

Fan, Z. and Guan, L. (2017). Approximate l_0-penalized estimation of piecewise-constant signals on graphs. *arXiv preprint arXiv:1703.01421*. MR3852650

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. MR3210728

Friedrich, F., Kempe, A., Liebscher, V. and Winkler, G. (2008). Complexity penalized M-estimation: Fast computation. *Journal of Computational and Graphical Statistics* **17** 201–204. MR2424802

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* **42** 2243–2281. MR3269979

Gao, C., Han, F. and Zhang, C. H. (2017). Minimax risk bounds for piecewise constant models. *arXiv preprint arXiv:1705.06386*.

Gibberd, A. J. and Roy, S. (2017). Multiple changepoint estimation in high-dimensional Gaussian graphical models. *arXiv preprint arXiv:1712.05786*.

James, B., James, K. L. and Siegmund, D. (1987). Tests for a change-point. *Biometrika* **74** 71-83. MR0885920

Jeng, X. J., Cai, T. T. and Li, H. (2012). Simultaneous discovery of rare and common segment variants. *Biometrika* **100** 157–172. MR3034330

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics* **43** 2451-2483. MR3405600

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. MR3036418

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications* **83** 79-102. MR1705601

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis* **21** 33-59. MR1766173

Li, H., Guo, Q. and Munk, A. (2017). Multiscale change-point segmentation: Beyond step functions. *arXiv preprint arXiv:1708.03942*. MR4010980

Liebscher, V. and Winkler, G. (1999). A Potts model for segmentation and jump-detection. In *Proceedings S4G International Conference on Stereology, Spatial Statistics and Stochastic Geometry, Prague* **21**.

Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and Computing* **27** 519–533. MR3599687

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100-115. MR0088850

Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics* **37** 2292-2952. MR2541451

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91-108. MR2136641

Tickle, S. O., Eckley, I. A., Fearnhead, P. and Haynes, K. (2018). Parallelisation of a Common Changepoint Detection Method. *arXiv preprint arXiv:1810.03591*. MR4085871

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer. MR2724359

Venkatraman, E. S. (1992). Consistency results in multiple change-point problems, PhD thesis, Stanford University. MR2687536

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*. MR2963170

Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR* **259** 270–274. MR0625215

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* **16** 117-186. MR0013275

Wang, T. and Samworth, R. J. (2018). High-dimensional changepoint estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. MR3744712

Wang, D., Yu, Y. and Rinaldo, A. (2017). Optimal Covariance Change Point Detection in High Dimension. *arXiv preprint arXiv:1712.09912*.

Wang, D., Yu, Y. and Rinaldo, A. (2018). Optimal Change Point Detection and Localization in Sparse Dynamic Networks. *arXiv preprint arXiv:1809.09602*.

Yao, Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters* **6** 181–189. MR0919373

Yao, Y.-C. and Au, S.-T. (1989). Least-squares estimation of a stop function. *Sankhyā: The Indian Journal of Statistics, Series A* 370-381. MR1175613

Yao, Y. C. and Davis, R. A. (1986). The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhyā: The Indian Journal of Statistics, Series A* 339–353. MR0905446

Yu, B. (1997). *Festschrift for Lucien Le Cam* **423** Assouad, Fano, and Le Cam, 435. Springer Science & Business Media. MR1462963