Journal of Multivariate Analysis 188 (2022) 104833

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

High dimensional change point inference: Recent developments and extensions

Bin Liu [a], Xinsheng Zhang [a], Yufeng Liu [b,c,d,*]

[a] School of Management, Fudan University, Shanghai, 200433, China
[b] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA
[c] Department of Genetics, Department of Biostatistics, University of North Carolina at Chapel Hill, USA
[d] Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, USA

ARTICLE INFO

Article history:
Received 16 August 2021
Received in revised form 7 September 2021
Accepted 8 September 2021
Available online 22 September 2021

AMS 2021 subject classifications:
primary 62H15
secondary 62E20

Keywords:
Alternative patterns
Change point detection
High dimensions
Hypothesis testing
Minimax optimality

ABSTRACT

Change point analysis aims to detect structural changes in a data sequence. It has always been an active research area since it was introduced in the 1950s. In modern statistical applications, however, high-throughput data with increasing dimensions are ubiquitous in fields ranging from economics, finance to genetics and engineering. For those problems, the earlier works are typically no longer applicable. As a result, the problem of testing a change point for high dimensional data sequences has been an important yet challenging task. In this paper, we first focus on models for at most one change point, and review recent state-of-art techniques for change point testing of high dimensional mean vectors and compare their theoretical properties. Based on that, we provide a survey of some extensions to general high dimensional parameters beyond mean vectors as well as strategies for testing multiple change points in high dimensions. Finally, we discuss some open problems for possible future research directions.

1. Introduction

Change point analysis has a long history since the seminal work of [50,51]. Since then, it has become an active research area in various scientific fields including finance, genetics, climatology, engineering, and astronomy. Generally speaking, suppose we have a sequence of ordered observations such as a time series. Change point analysis aims to answer the following two questions: [i] whether there is a change for the parameter of the underlying data distribution during the observations; [ii] If a change is detected, where is the position of the change point? The above two questions are referred to as the change point testing and estimation problems, which are two indispensable pillars in change point analysis. In this paper, we mainly focus on the former question. Classical methods for change point testing assume that the data dimension is fixed. In the last few decades, a rich literature has been developed in addressing different specific problems under various model settings. See the book in [11,22] for a summary of the classical methods and a recent review paper in [32] for some extensions.

With the rapid development of data collection and storage capacity, high dimensional data are ubiquitous, where the data dimension can be tens of thousands and are typically much larger than the sample size. In this case, the data generating mechanism can be complicated and heterogeneity often exists. Hence, for high dimensional data analysis,

* Correspondence to: Department of Statistics and Operations Research, 354 Hanes Hall, CB 3260, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
E-mail address: yfliu@email.unc.edu (Y. Liu).

https://doi.org/10.1016/j.jmva.2021.104833
0047-259X/

heterogeneity detection or change point detection is an important issue. While there are many earlier works obtained with good theoretical results, in high dimensions, the classical methods are often no longer applicable. As a result, it is desirable to design new methods suitable for modern statistical applications. Driven by this demand, rapid developments have been made in the literature over the last 5–10 years for change point analysis. In this paper, focusing on change point testing, we first review recent developments on single change point detection of high dimensional mean vectors. We believe this can reflect the distinctive challenges for high dimensional change point analysis. In addition, it can serve as a foundation for developing new methodologies for some other complex change point problems. With the concept of high dimensional efficiency, we compare different methods in terms of size and power, and show their optimality in terms of the minimax optimality separation rate. The latter one is more technically involved in high dimensions and usually exhibits a phase transition according to the change point alternative patterns. Beyond mean vectors, in the second part of this paper, we review some recent extensions to other high dimensional parameters such as variance, covariance matrices, or non-parametric testing of distributional changes. Lastly, based on the existing literature, we point out several possible research directions for more complex model settings and problems.

Note that there are two related problems in high dimensional change point analysis. The first one is change point estimation [17,49,56]. For this problem, it is usually assumed there are $K_0 \geq 1$ change points that exist in the model and the goal is to simultaneously estimate both the number and locations of the change points. This problem is also called data segmentation. Although change point testing and estimation are related, they are fundamentally different. For example, the former is concerned with proposing tests for controlling the type I and type II errors, while the latter one mainly focuses on developing algorithms for estimation consistency of numbers and locations. The second type of problem is called online or sequential change point detection [48], where data are observed sequentially. The goal is to detect a change point as soon as possible while controlling the false alarms. This is very different from our considered offline setting that we have observed all historical data at once. Although the above two problems are interesting and actively studied to date, it is not possible for us to review all related works in this paper. Hence, in this review, we only focus our attention on the problem of offline change point testing.

Throughout this paper, for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, define its ℓ_p-norm as $\|\boldsymbol{x}\|_p = (\sum_{j=1}^d |x_j|^p)^{1/p}$ for $1 \leq p \leq \infty$. For $p = \infty$, define $\|\boldsymbol{x}\|_\infty = \max_{1\leq j\leq d} |x_j|$. For any set $\mathcal{S}$, denote its cardinality by $|\mathcal{S}|$. For two real numbered sequences a_n and b_n, we set $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ for a sufficiently large n; $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$; $a_n \asymp b_n$ if there exist constants c and C such that $c|b_n| \leq |a_n| \leq C|b_n|$ for a sufficiently large n. For a sequence of random variables (r.v.s) $\{\xi_1, \xi_2, \ldots\}$, we set $\xi_n \xrightarrow{\mathbb{P}} \xi$ (or $\xi_n \xrightarrow{d} \xi$) if ξ_n converges to ξ in probability (or in distribution) as $n \to \infty$. We also denote $\xi_n = o_p(1)$ if $\xi_n \xrightarrow{\mathbb{P}} 0$. For a positive number x, we use $\lfloor x \rfloor$ to denote the largest integer less than or equal to x.

The rest of this paper is organized as follows. Section 2 introduces the formulations of high dimensional change point inference and its distinctive challenges from the low dimensional problems. Sections 3–4 review recent methods for change point inference of high dimensional mean vectors. Section 5 discussed their theoretical properties. Sections 6–7 provide some extensions to high dimensional change point inference for general parameters as well as techniques for testing multiple change points. We conclude this paper in Section 8.

2. High dimensional change point inference

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \sim F(\boldsymbol{x})$ and $\boldsymbol{\mu} := \mathbb{E}\boldsymbol{X} = (\mu_1, \ldots, \mu_d)^\top$ be its mean vector. Suppose we have n ordered but independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{id})^\top$ being its ith observation. Typically, a change point model for mean vectors has the following form: for $i = 1, \ldots, n$,

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{\delta}_{k_0} \mathbf{1}\{i \geq k_0\} + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean vector before change point, $k_0 \in \{1, \ldots, n-1\}$ is the possible but unknown change point location, $\boldsymbol{\delta}_{k_0} = (\delta_1, \ldots, \delta_d)^\top$ is the mean shift after the change point k_0 (if it exists), and $\boldsymbol{\epsilon}_i := (\epsilon_{i1}, \ldots, \epsilon_{id})^\top$ are i.i.d error terms with $\mathbb{E}\boldsymbol{\epsilon}_i = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$. Model (1) is usually called at most one change point (AMOC) model. A typical question for Model (1) is to test whether there is a change point. In other words, for Model (1), we have the following hypothesis:

$$\mathbf{H}_0 : \boldsymbol{\delta}_{k_0} = \mathbf{0} \quad \text{and} \quad k_0 = n \quad \text{vs.} \quad \mathbf{H}_1 : \exists k_0 \in \{1, \ldots, n-1\} \text{ such that } \boldsymbol{\delta}_{k_0} \neq \mathbf{0}. \tag{2}$$

More specifically, the data are homogeneous across the observations in terms of the mean vectors under $\mathbf{H}_0$, while there is a mean shift $\boldsymbol{\delta}_{k_0}$ at the unknown location k_0 under $\mathbf{H}_1$. For the simplicity of notations, we use $\boldsymbol{\delta}$ directly when it is appropriate. In the traditional low dimensional setting with fixed d and $d < n$, the cumulative sum (CUSUM) statistic [22] is typically adopted for Problem (2). In particular, for each candidate search location $k \in \{1, \ldots, n-1\}$, the CUSUM statistic is defined as:

$$\boldsymbol{C}(k) = \sqrt{\frac{k(n-k)}{n}}\Big(\frac{1}{n-k}\sum_{i=k+1}^{n} \boldsymbol{X}_i - \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{X}_i\Big). \tag{3}$$

Then based on (3), we can construct the following test statistic:

$$T_n = \max_{1\le k\le n-1} \boldsymbol{C}(k)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}(k). \tag{4}$$

Intuitively, T_n searches through all possible locations k for comparing the mean differences before and after k. Hence, a large value of T_n triggers the rejection of $\mathbf{H}_0$. It is worth mentioning that T_n is related to the log-ratio of the maximum likelihood based statistic. To see this, consider the one dimensional case with $d = 1$. Suppose $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ with known σ^2. For each fixed k, the log-ratio of the maximum likelihood statistic is defined as

$$\begin{aligned} H_k &= \log\left\{ \frac{\prod_{i=1}^k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{\big(X_i - \bar{X}(k)\big)^2}{2\sigma^2}\Big) \prod_{i=k+1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{\big(X_i - \bar{X}(n-k)\big)^2}{2\sigma^2}\Big)}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{\big(X_i - \bar{X}\big)^2}{2\sigma^2}\Big)} \right\} \\ &= \frac{1}{2\sigma^2}\left\{ \sqrt{\frac{n-k}{nk}} \sum_{i=1}^k X_i - \sqrt{\frac{k}{n(n-k)}} \sum_{i=k+1}^n X_i \right\}^2 = \frac{1}{2\sigma^2}\boldsymbol{C}^2(k), \end{aligned} \tag{5}$$

where $\bar{X}(k) := k^{-1}\sum_{i=1}^k X_i$, $\bar{X}(n-k) := (n-k)^{-1}\sum_{i=k+1}^n X_i$, and $\bar{X} := n^{-1}\sum_{i=1}^n X_i$. Hence, based on (5), the maximum likelihood ratio based test statistic for testing (2) is defined as:

$$\max_{1\le k\le n-1} 2H_k = \max_{1\le k\le n-1} \frac{1}{\sigma^2}\boldsymbol{C}^2(k),$$

which is equivalent to the use of (4). In the low dimensional case, the CUSUM statistic in (3) or (5) has been extensively studied for change point detection. See [3,22,28–30,47,52]. Typically, under some regular conditions, we have, as $n \to \infty$,

$$\max_{1\le k\le n-1} \frac{k}{n}\frac{n-k}{n}\boldsymbol{C}(k)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{C}(k) \xrightarrow{d} \sup_{0\le t\le 1} \sum_{j=1}^d B_j^2(t),$$

where $\{B_1(t), \ldots, B_d(t)\}$ are independent Brownian bridges with mean zero and covariance structure $\mathbb{E}(B_1(t), B_1(s)) = (\min(t, s) - ts)$. Formulas and critical values for the above limiting distribution can be found in [36].

In the high dimensional setting, however, the data dimension d can grow with the sample size n ($d \to \infty$) and even be much larger than n ($d \gg n$). For example, in finance, if we are interested in testing a change point for returns of companies in the S&P 500 index over a given period, we need to deal with data with the dimension of 500, see [35]. In genetics, we can use change point testing to detect chromosomal copy number abnormality which involves thousands of genes [58]. Other modern statistical applications include the detection of Denial of Service in internet traffic [39], functional Magnetic Resonance Imaging (fMRI) data studies, see [69] as well as the detection of distant galaxies in astronomy [25]. For those real applications, the high dimensionality of data brings great challenges in both implementation and theoretical studies [34]. More specifically, the methods designed for low dimensional cases are no longer applicable. One direct difficulty comes from the estimation of the covariance matrix $\boldsymbol{\Sigma}$ or its inverse $\boldsymbol{\Sigma}^{-1}$ in the construction of the CUSUM statistic. It is well known that the standard sample covariance matrix estimator performs poorly and can lead to invalid conclusions in high-dimensional settings with relatively low sample sizes. Thus, estimation of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ can be nontrivial in such high dimensional problems. To obtain estimation consistency, some structural assumptions are typically imposed in the literature in order to estimate $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$. These include banding [61], thresholding [7], or penalized likelihood based estimation [8,64]. See [9] for a recent development for this topic. In addition, using the above methods for constructing the CUSUM statistics as in (4) usually imposes strong structural assumptions for $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}^{-1}$) and involves the selection of tuning parameters [45]. This makes the use of (4) very difficult. Hence, new change point testing statistics for high dimensional data are needed. More importantly, since the dimension d can grow with the sample size n, it becomes more challenging to derive the limiting null distribution in high dimensions to make the corresponding testing procedure under a given significance level $\alpha \in (0, 1)$.

Another distinctive difficulty for high dimensional change point testing is the power analysis. Recall the mean jump $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_d)^\top$. Let $\Pi = \{j : \delta_j \neq 0\}$ be the set of coordinates having a change point and $s := |\Pi|$ be its cardinality. According to various structures of $\boldsymbol{\delta}$, high dimensional change point models can be typically summarized into the following two cases:

- Case 1: Sparse patterns with only a few number of non-zero elements in $\boldsymbol{\delta}$ and the corresponding magnitudes of changes are large.
- Case 2: Dense patterns where a large number of entries in $\boldsymbol{\delta}$ are non-zero and each with a small magnitude.

The above two cases are called alternative structures or patterns. For the sparse pattern, we usually require $s \ll \sqrt{p}$ while for the dense pattern we have $s \gg \sqrt{p}$ [25,41]. As shown later, the detection boundary on the signal strength $\boldsymbol{\delta}$ usually exhibits a phase transition according to s. Note that in high dimensions, both p and s scale with the sample size n. This is essentially different from the traditional setting with a fixed d. More specifically, for high dimensional data

analysis, the standard norms such as ℓ_2 or ℓ_∞-norms are no longer equivalent. In other words, a test statistic designed for the sparse pattern may lose its power under the dense pattern, and vice versa. Hence, for high dimensional change point detection, the main question is how to propose a powerful test statistic that effectively utilizes the underlying alternative structures. The above aspects make high dimensional change point inference a challenging problem and not much developments have been made until recently. Recall the d-dimensional CUSUM statistic $\boldsymbol{C}(k) = \big(C_1(k), \ldots, C_d(k)\big)^\top$, where

$$C_j(k) = \sqrt{\frac{k(n-k)}{n}}\Big(\frac{1}{n-k}\sum_{i=k+1}^{n} X_{ij} - \frac{1}{k}\sum_{i=1}^{k} X_{ij}\Big) \tag{6}$$

is the CUSUM statistic for each coordinate $j \in \{1, \ldots, d\}$. Hence, we can construct the CUSUM statistic-based $d \times (n-1)$ dimensional matrix $\mathcal{C} := \big(\boldsymbol{C}(1), \ldots, \boldsymbol{C}(n-1)\big)$:

$$\mathcal{C} = \begin{pmatrix} C_1(1) & \cdots & C_1(n-1) \\ C_2(1) & \ldots & C_2(n-1) \\ \vdots & \ddots & \vdots \\ C_d(1) & \ldots & C_d(n-1) \end{pmatrix}. \tag{7}$$

By definition, the columns of $\mathcal{C}$ contain information about the change point location, while the rows reflect the alternative structures. Note that for the AMOC model in (1), all coordinates in Π share a common change point at k_0. Hence, similar to the low dimensional setting, we can construct a test statistic such as:

$$T = \max_{1\le k\le n-1} \|\boldsymbol{C}(k)\|_p.$$

The biggest question for using T is how to adopt a proper ℓ_p-norm to take into account the underlying alternative patterns and benefit from the cross-sectional nature of the change-point that is shared across different coordinates. Furthermore, from a theoretical viewpoint, high dimensional change point inference is related to the concept of high dimensional efficiency (HDE) proposed in [2] as well as the minimax optimality studied in [25,41].

We first introduce the high dimensional efficiency.

Definition 2.1 (*High Dimensional Efficiency [2]*)**.** Suppose $d \to \infty$ as $n \to \infty$. Let $\mathcal{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a testing statistic for the problem in (2). Let $\mathcal{E}(\boldsymbol{\delta})$ be a mapping from $\mathbb{R}^d$ to $\mathbb{R}_+ := (0, \infty)$. We say the (absolute) high dimensional efficiency of $\mathcal{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is $\mathcal{E}(\boldsymbol{\delta})$ if it satisfies:

(i) $\mathcal{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \xrightarrow{d} L$ for some non-degenerate limiting distribution L under $\mathbf{H}_0$;
(ii) $\mathcal{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \xrightarrow{\mathbb{P}} \infty$ if $\sqrt{n}\mathcal{E}(\boldsymbol{\delta}) \to \infty$;
(iii) $\mathcal{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \xrightarrow{d} L$ if $\sqrt{n}\mathcal{E}(\boldsymbol{\delta}) \to 0$.

By definition, high dimensional efficiency characterizes the rate at which the cross-sectional size of change is allowed to converge to zero (e.g., $\|\boldsymbol{\delta}\|_p \to 0$) as $n \to \infty$ such that the power of the change-point test $\mathcal{T}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is strictly between the size and one. Hence, using high dimensional efficiency, we can compare different tests' efficiency (or relative efficiency) in terms of their power. For example, if we have a change point test $\mathcal{T}^{(1)}$ with efficiency $\mathcal{E}^{(1)}(\boldsymbol{\delta}) = \|\boldsymbol{\delta}\|_2$ and another $\mathcal{T}^{(2)}$ having efficiency $\mathcal{E}^{(2)}(\boldsymbol{\delta}) = \|\boldsymbol{\delta}\|_2/\log(n)$, then the relative high dimensional efficiency between $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$ is $\mathcal{E}^{(1)}(\boldsymbol{\delta})/\mathcal{E}^{(2)}(\boldsymbol{\delta}) = \log(n)$. In other words, for detecting a change point with power one, compared with $\mathcal{T}^{(1)}$, we require a stronger signal condition with an additional order $\log n$ for $\mathcal{T}^{(2)}$.

In addition to HDE, minimax optimality is also a way to show the optimality of a change point testing method. Before introducing that, some notations are needed. For Problem (2), let $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\delta}_{k_0}^\top)^\top$ be the parameters of interest. Under $\mathbf{H}_0$, define the parameter space of signals by

$$\Theta_0(d, n) = \Big\{(\boldsymbol{\mu}, \boldsymbol{\delta}_{k_0}) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\delta}_{k_0} = \mathbf{0}, k_0 \in \{1, \ldots, n-1\}\Big\}.$$

Next, we define the parameter space of signals under the alternative that a change point occurs at a known location $k_0 \in \{1, \ldots, n-1\}$ with a known sparsity level s by:

$$\Theta^{(k_0)}(d, n, s, \rho) = \Big\{(\boldsymbol{\mu}, \boldsymbol{\delta}_{k_0}) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\delta}_{k_0} \in \mathbb{R}^d, \|\boldsymbol{\delta}\|_0 = s, \frac{k_0(n-k_0)}{n}\big\|\boldsymbol{\delta}\big\|_2^2 \ge \rho^2\Big\}. \tag{8}$$

The alternative $\Theta^{(k_0)}(d, n, s, \rho)$ says that Model (1) has a change point at $k_0 \in \{1, \ldots, n-1\}$ with a signal jump $\boldsymbol{\delta}_{k_0}$ which has s non-zero entries and a magnitude $\frac{k_0(n-k_0)}{n}\big\|\boldsymbol{\delta}\big\|_2^2 \ge \rho^2$. Note that $\rho := \rho(n, d, s, k_0)$ may depend on n, d, s, and k_0. Since both k_0 and s are unknown, for Model (1), we define the final parameter space under the alternative as:

$$\Theta_1(d, n, \rho) := \bigcup_{k_0=1}^{n-1}\bigcup_{s=1}^{d} \Theta^{(k_0)}(d, n, s, \rho).$$

With the above notations, the testing problem in (2) is equivalent to:

$$\mathbf{H}_0 : \boldsymbol{\theta} \in \Theta_0(d, n) \ \text{ vs. } \ \mathbf{H}_1 : \boldsymbol{\theta} \in \Theta_1(d, n, \rho). \tag{9}$$

We are now ready to introduce the concept of minimax optimality for high dimensional change point inference.

Definition 2.2 (*Minimax Optimality*)**.** For Problem (9), let Ψ denote the class of possible test statistics, i.e., measurable functions $\psi(\mathbf{X})$ taking values in $\{0, 1\}$, where $\mathbf{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. We say there is a change point if $\psi(\mathbf{X}) = 1$. For any test $\psi(\mathbf{X})$, we also define its testing error by

$$\mathcal{R}(\psi, \rho) := \sup_{\boldsymbol{\theta} \in \Theta_0(d,n)} \mathbb{E}_{\boldsymbol{\theta}}\big[\psi(\mathbf{X})\big] + \sup_{\boldsymbol{\theta} \in \Theta_1(d,n,\rho)} \mathbb{E}_{\boldsymbol{\theta}}\big[1 - \psi(\mathbf{X})\big].$$

We say $\rho^* = \rho(n, d, s, k)$ is the minimax rate of change point testing if the following two conditions are satisfied:

1. For any $\epsilon \in (0, 1)$, there exist a constant $C_\epsilon > 0$ depending only on ϵ and a test ψ^* such that $\mathcal{R}(\psi^*, C\rho^*) \le \epsilon$ for all $C > C_\epsilon$ (or asymptotically $\limsup_{n\to\infty} \mathcal{R}(\psi^*, C\rho^*) \le \epsilon$).
2. For any $\epsilon \in (0, 1)$, there exists a constant $c_\epsilon > 0$ depending only on ϵ such that $\forall\ 0 < c < c_\epsilon$ and for any test $\psi \in \Psi$, the following holds:

$$\mathcal{R}(\psi, c\rho^*) \ge 1 - \epsilon \ \big(\text{or} \ \ \text{asymptotically} \liminf_{n\to\infty} \mathcal{R}(\psi, c\rho^*) \ge 1 - \epsilon\big).$$

Remark 1. By definition, $\rho^* = \rho(n, d, s, k)$ is the minimax rate of change point testing in the sense that there is no test that can control the overall errors (type I and type II errors) under $1 - \epsilon$ if the signal strength $\rho < c_\epsilon \rho^*$. Moreover, if we restrict our attention to the class of all α level tests $\Psi_\alpha = \{\psi : \sup_{\boldsymbol{\theta} \in \Theta_0(d,n)} \mathbb{E}_{\boldsymbol{\theta}}[\psi(\mathbf{X})] \le \alpha\}$, then $\rho^* = \rho(n, d, s, k)$ is said to obtain the minimax rate of testing in the sense that no α level test can detect a change point with overwhelming probability if the signal jump satisfies $\rho < c_\epsilon \rho^*$.

Next we give a comprehensive review on recent developments of high dimensional change point inference and compare their theoretical properties in terms of HDE or minimax optimality. Recall that the goal is to develop an efficient test via aggregating the CUSUM matrix (7) to account for the alternative structures. In the high dimensional setting, these tests are mainly classified into two categories: one class knows and utilizes the sparse or dense alternative patterns [4,16,31,35,58,59,63,67], and another class aims to construct a method that accounts for the unknown alternative structures in a data-driven way [2,15,25,41,43,53,68], which is also called the data-adaptive method. In what follows, we assume that the variance in each coordinate of $\boldsymbol{X}$, e.g., $\mathrm{Var}(X_j) = \sigma_j^2$, is known and without loss of generality, we assume $\sigma_j^2 = 1$ for $1 \le j \le d$. Moreover, we require all coordinates with a change point have the same order in the sense that

$$0 < \underline{c} \le \liminf_{n\to\infty} \frac{\delta_{\min}}{\delta_{\max}} \le \limsup_{n\to\infty} \frac{\delta_{\max}}{\delta_{\min}} \le \overline{C} < \infty, \tag{10}$$

where $\delta_{\min} = \min_{j \in \Pi} |\delta_j|$ and $\delta_{\max} = \max_{j \in \Pi} |\delta_j|$. This helps us to make a comparison between different methods under a relatively unified framework.

Remark 2. In change point analysis, variance estimation is an important problem. As shown in [54], inappropriate variance estimation may lead to non-monotonic power performance. In this review paper, to avoid unnecessary notations, we assume $\sigma_j^2 = 1$ for $1 \le j \le d$ for discussing the main ideas of different methods and theoretical properties.

3. High dimensional change point inference with known alternative structures

3.1. ℓ_∞-norm based methods in [35,63] for sparse alternatives

For high sparse change point alternatives, [35] proposed the following ℓ_∞-norm based test statistic:

$$T_{\text{Jirak}} = \max_{1\le k\le n-1} \max_{1\le j\le d} \sqrt{\frac{k}{n}\frac{n-k}{n}} \big|C_j(k)\big| = \max_{1\le k\le n-1} \max_{1\le j\le d} \sqrt{\frac{k}{n}\frac{n-k}{n}} \big\|\boldsymbol{C}(k)\big\|_\infty. \tag{11}$$

They proved that under some temporal conditions (Assumption 2.1 therein) for data observations and spatial conditions among coordinates (Assumption 2.2 therein), the test statistic has the following limiting null Gumbel distribution:

$$\lim_{n\to\infty} \mathbb{P}\big(\sqrt{2\log(2d)}(T_{\text{Jirak}} - f_d) \le x\big) = \exp(-e^{-x}), \tag{12}$$

where $f_d = \frac{1}{2}\sqrt{2\log(2d)} - \log(3\log(2d))/\sqrt{2\log(2d)}$. Hence, a critical value $c_\alpha = -\log(-\log(1-\alpha))$ can be used to implement an α level test. This result can be regarded as an extension of the low dimensional counterpart (Theorem 1.3.1 in [22]) to high dimensions. Note that [35] requires d grows with the sample size n with at most a polynomial rate ($d \ll n^C$ for some $C > 0$). It is well-known that the convergence to an extreme distribution in (12) is slow and

typically requires strong conditions on the covariance structures among coordinates (see Assumption 2.2 therein), which seems unreasonable in practice. To overcome this problem, [35] also proposed to use a parametric method as well as a multiplier bootstrap method respectively, to approximate the limiting null distribution of T_{jirak}. In terms of the power analysis, it is proved that if $\min_{1\le j\le d}|\delta_j| \gg \sqrt{\log(n)/n}$ holds, with probability tending to one, T_{jirak} can detect and identify a change point.

Note that the method in [35] requires strong assumptions on the covariance structure $\boldsymbol{\Sigma}$, and the bootstrap method in [35] also requires a consistent estimator for the relative change point location k_0/n to mimic the data under $\mathbf{H}_0$. This results in strong conditions for the signal strength. To relax the assumptions on $\boldsymbol{\Sigma}$ as well as the signal condition on $\boldsymbol{\delta}$, [63] proposed the following test statistic for detecting sparse alternatives:

$$T_{\text{Yu}} = \max_{\underline{k}\le k\le n-\underline{k}} \max_{1\le j\le d} \big|C_j(k)\big| = \max_{\underline{k}\le k\le n-\underline{k}} \big\|\boldsymbol{C}(k)\big\|_\infty, \tag{13}$$

where $\underline{k}$ and $n-\underline{k}$ are the lower and upper bounds for the candidate search location. Note that the main difference between T_{jirak} and T_{Yu} is that the former places higher weights on search locations closer to the centre of data observations, which is easier to detect a change point if it is located around the centre in the data sequence. Instead of deriving the limiting distribution directly, T_{Yu} considered a multiplier bootstrap method to approximate the limiting null distribution of T_{Yu}, which proceeds as follows. Let $e_1,\ldots,e_n \overset{i.i.d}{\sim} N(0,1)$. Define the bootstrap based CUSUM statistic as:

$$\boldsymbol{C}^*(k) = \sqrt{\frac{k(n-k)}{n}}\Big(\frac{1}{n-k}\sum_{i=k+1}^{n} e_i\big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}(n-k)\big) - \frac{1}{k}\sum_{i=1}^{k} e_i\big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}(k)\big)\Big), \tag{14}$$

where $\bar{\boldsymbol{X}}(k) = k^{-1}\sum_{i=1}^{k}\boldsymbol{X}_i$, $\bar{\boldsymbol{X}}(n-k) = (n-k)^{-1}\sum_{i=k+1}^{n}\boldsymbol{X}_i$. Based on $\boldsymbol{C}^*(k)$, the bootstrap based test statistic for T_{Yu} is defined as:

$$T^*_{\text{Yu}} = \max_{\underline{k}\le k\le n-\underline{k}} \max_{1\le j\le d} \big|C^*_j(k)\big| = \max_{\underline{k}\le k\le n-\underline{k}} \big\|\boldsymbol{C}^*(k)\big\|_\infty. \tag{15}$$

The main question for bootstrap is whether the conditional distribution of T^*_{Yu} given $\mathbf{X}$ can approximate the distribution of T_{Yu}? Let $c_{T^*_{\text{Yu}}|\mathbf{X}}(1-\alpha)$ be the $1-\alpha$ quantile of T^*_{Yu} given $\mathbf{X}$. [63] proved that for sub-exponential distributions of $\boldsymbol{\epsilon}$ with some other regularity conditions, under $\mathbf{H}_0$, we have, as $n\to\infty$,

$$\sup_{\alpha\in(0,1)}\Big|\mathbb{P}\big(T_{\text{Yu}} \le c_{T^*_{\text{Yu}}|\mathbf{X}}(1-\alpha)\big) - \alpha\Big| \to 0,$$

as long as $\log^7(dn) = o(\underline{k})$ holds. Hence, [63] allows the data dimension d to grow exponentially with the sample size n in the sense that $d = O(e^{n^c})$ for some $0 < c < 1/7$. For the power results, [63] proved that under $\mathbf{H}_1$, for any $\zeta\in(0,1)$, we have

$$\mathbb{P}\big(T_{\text{Yu}} \ge c_{T^*_{\text{Yu}}|\mathbf{X}}(1-\alpha)\big) \ge 1 - (nd)^{-C_1} - C_2\Big(\frac{\log^7(dn)}{\underline{k}}\Big)^{1/6} - \zeta$$

as long as the signal jump satisfies:

$$\big\|\boldsymbol{\delta}\big\|_\infty := \max_{1\le j\le d}|\delta_j| \ge C_3\sqrt{\frac{\log(\zeta^{-1})\log(dn) + \log(nd/\alpha)}{n\frac{k_0}{n}\big(1-\frac{k_0}{n}\big)}}, \tag{16}$$

where C_1, C_2, C_3 are some universal positive constants. Note that the above results also apply to data with uniform polynomial moments (see Assumption D therein), where some technical conditions need to be modified.

Remark 3. For detecting a change point with overwhelming probability, [35] requires $\delta_{\min} \gg \sqrt{\log(n)/n}$ and [63] requires $\delta_{\min} > C\sqrt{\log(d)/n}$ for some big enough constant $C > 0$. From the aspect of HDE, we see that the efficiency of [35] is $\frac{\delta_{\min}}{\sqrt{\log(n)}}$ while the efficiency of [63] is $\frac{\delta_{\max}}{\sqrt{\log(d)}}$. Moreover, we show in Section 5.4 that the signal strength requirement of [63] reaches the minimax optimality for detecting sparse change points.

3.2. ℓ_2-norm based methods in [31,59,67] for dense alternatives

It is known that for detecting dense change point alternatives with a large number of coordinates in $\boldsymbol{\delta}$ experiencing a change of a small signal jump, the ℓ_2-norm based test statistic typically has good performance than the ℓ_∞-norm based one. The intuition is that in high dimensions, the ℓ_2-norm is not equivalent to the ℓ_∞-norm. Once the signal jump for $\|\boldsymbol{\delta}\|_\infty$ is a smaller order than $\sqrt{\log(d)/n}$, the ℓ_∞-norm based method fails to capture such signal information while the ℓ_2 norm succeeds by adding up all the weak signals.

Assuming the errors $\{\epsilon_{ij}\}, i \in \{1, \ldots, n\}, j \in \{1, \ldots, d\}$ in Mode (1) are i.i.d $N(0, 1)$, [67] proposed the following test statistic for detecting dense change point patterns:

$$T_{\text{Zhang}} = \max_{\lfloor c_1 n \rfloor \le k \le \lfloor c_2 n \rfloor} \sum_{j=1}^{d} C_j^2(k) = \max_{\lfloor c_1 n \rfloor \le k \le \lfloor c_2 n \rfloor} \big\| \boldsymbol{C}(j) \big\|_2^2,$$

where $0 < c_1 < c_2 < 0.5$ are some prespecified constants. Under $\mathbf{H}_0$, $\|\boldsymbol{C}(j)\|_2^2$ follows a chi-squared distribution with a degree of freedom of d. To obtain the critical value that determines a change point, [67] derived an explicit approximation to the tail probability of T_{Zhang} under the null hypothesis. This leads to an approximation to the theoretical p-value. The authors showed that the approximation is very accurate at moderate to small p-values. Note that the derivation in [67] is non-asymptotic in the sense that the dimension d can be large but fixed.

In a similar way, [31] proposed a test statistic like:

$$T_{\text{HH}} = \max_{1 \le k \le n-1} \frac{k}{n} \frac{n-k}{n} \Big| \frac{1}{\sqrt{d}} \sum_{j=1}^{d} \big(C_j^2(k) - 1 \big) \Big| = \max_{1 \le k \le n-1} \frac{k}{n} \frac{n-k}{n} \Big| \frac{\|\boldsymbol{C}(k)\|_2^2 - d}{\sqrt{d}} \Big|. \tag{17}$$

For the above test statistic, [31] proved that under $\mathbf{H}_0$, the following holds, as $n \to \infty$,

$$T_{\text{HH}} \xrightarrow{d} \sup_{0 \le t \le 1} |\Gamma(t)|, \tag{18}$$

where $\{\Gamma(t), 0 \le t \le 1\}$ is a Gaussian process with $\mathbb{E}[\Gamma(t)] = 0$ and $\mathbb{E}[\Gamma(t)\Gamma(s)] = 2t^2(1 - s^2)$ for $0 \le t \le s \le 1$. Hence, a critical value can be obtained via Monte Carlo simulations for the limiting distribution. Note that [31] mainly requires $d = o(\sqrt{n})$, which can be regarded as a low dimensional problem. Moreover, the derivation in (18) requires all coordinates in $\boldsymbol{\epsilon}$ are independent with $\boldsymbol{\Sigma} = \mathbf{I}_d$. For the power analysis, [31] proved that if $\|\boldsymbol{\delta}\|_2 \gg \frac{d^{1/4}}{\sqrt{n}}$, with probability tending to one, we can detect a change point. Hence, the HDE for [31] is $d^{-1/4}\|\boldsymbol{\delta}\|_2$.

Note that in the dense case, the ℓ_2-norm of the signal jump $\|\boldsymbol{\delta}\|_2$ matters. Hence, a natural question is: is it possible to give an unbiased estimator for $\|\boldsymbol{\delta}\|_2^2$ and construct a change point test statistic based on that? [59] answered this question. Recall $\boldsymbol{\delta} = \mathbb{E}\boldsymbol{X}_{k_0} - \mathbb{E}\boldsymbol{X}_{k_0 - 1}$ is the mean difference before and after the change point k_0. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^d$. Define $h((\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{y}_1, \boldsymbol{y}_2)) = (\boldsymbol{x}_1 - \boldsymbol{y}_1)^\top (\boldsymbol{x}_2 - \boldsymbol{y}_2)$. Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ and $(\boldsymbol{X}_2, \boldsymbol{Y}_2)$ be independent copies of $(\boldsymbol{X}_{k_0 - 1}, \boldsymbol{X}_{k_0})$, respectively. It is shown in [12] that

$$\mathbb{E}\big[h\big((\boldsymbol{X}_1, \boldsymbol{X}_2), (\boldsymbol{Y}_1, \boldsymbol{Y}_2) \big) \big] = \big\| \mathbb{E}\boldsymbol{X}_{k_0} - \mathbb{E}\boldsymbol{X}_{k_0 - 1} \big\|_2^2 = \big\| \boldsymbol{\delta} \big\|_2^2.$$

Motivated by this observation, in a recent work, [59] proposed the following U-statistic based process as:

$$G_n(k) = \frac{1}{k(k-1)} \frac{1}{(n-k)(n-k-1)} \sum_{\substack{1 \le i_1, i_2 \le k \\ i_1 \ne i_2}} \sum_{\substack{(k+1) \le j_1, j_2 \le n \\ j_1 \ne j_2}} (\boldsymbol{X}_{i_1} - \boldsymbol{X}_{j_1})^\top (\boldsymbol{X}_{i_2} - \boldsymbol{X}_{j_2}). \tag{19}$$

Based on $G_n(k)$, [59] proposed the following self-normalized based test statistic:

$$T_{\text{Wang}} = \sup_{2 \le k \le n-3} \frac{G_n^2(k)}{W_n(k)},$$

where

$$W_n(k) = \frac{1}{n} \sum_{t=2}^{k-2} \Big[\sum_{\substack{1 \le i_1, i_2 \le t \\ i_1 \ne i_2}} \sum_{\substack{t+1 \le j_1, j_2 \le k \\ j_1 \ne j_2}} (\boldsymbol{X}_{i_1} - \boldsymbol{X}_{j_1})^\top (\boldsymbol{X}_{i_2} - \boldsymbol{X}_{j_2}) \Big]^2 + \frac{1}{n} \sum_{t=k+2}^{n-2} \Big[\sum_{\substack{k \le i_1, i_2 \le t \\ i_1 \ne i_2}} \sum_{\substack{t+1 \le j_1, j_2 \le n \\ j_1 \ne j_2}} (\boldsymbol{X}_{i_1} - \boldsymbol{X}_{j_1})^\top (\boldsymbol{X}_{i_2} - \boldsymbol{X}_{j_2}) \Big]^2.$$

Note that $W_n(k)$ in the denominator of T_{Wang} is to cancel out the asymptotical variance in the limiting distribution of $G_n^2(k)$, making the limit of T_{Wang} pivotal. This method is called self-normalization, which can avoid the problem of estimating the variance. In particular, it is shown in [59] that under $\mathbf{H}_0$, we have

$$T_{\text{Wang}} \xrightarrow{d} T^*, \quad \text{as} \quad n \to \infty, \tag{20}$$

where T^* is a distribution not depending on any unknown parameters (see Theorem 3.4 therein), and the critical value can be obtained via Monte Carlo simulations. Note that the derivation of (20) requires some uniform bounds on moments of $\boldsymbol{\epsilon}$ and some "short-range" dependence type conditions on the entries of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)^\top$.

In terms of the power analysis, [59] showed that under $\mathbf{H}_1$, as $n \to \infty$, the following results hold:

(i) If $\dfrac{\sqrt{n}\|\boldsymbol{\delta}\|_2}{\|\boldsymbol{\Sigma}\|_F^{1/2}} \to 0$, then $T_{\text{Wang}} \xrightarrow{d} T^*$;

(ii) If $\frac{\sqrt{n}\|\boldsymbol{\delta}\|_2}{\|\boldsymbol{\Sigma}\|_F^{1/2}} \to \infty$, then $T_{\text{Wang}} \xrightarrow{\mathbb{P}} \infty$;

(iii) If $\frac{\sqrt{n}\|\boldsymbol{\delta}\|_2}{\|\boldsymbol{\Sigma}\|_F^{1/2}} \to \gamma \in (0, \infty)$, then $T_{\text{Wang}} \xrightarrow{d} T^{**}$,

where $\|\boldsymbol{\Sigma}\|_F$ is the Frobenius norm of $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{\epsilon})$ and T^{**} is some distribution with an additive shift. Recall the high dimensional efficiency in Definition 2.1. The HDE of T_{Wang} for detecting dense signals is $\|\boldsymbol{\delta}\|_2/\|\boldsymbol{\Sigma}\|_F^{1/2}$. Considering the special case with $\boldsymbol{\Sigma} = \mathbf{I}_d$, the HDE reduces to $d^{-1/4}\|\boldsymbol{\delta}\|_2$, which is equivalent to [31].

4. High dimensional change point inference without knowing the alternative structures

So far, we have given a discussion about methods designed for high dimensional change point inference with known alternative patterns. In practice, however, the alternative patterns may be unknown. Methods with wrong pattern assumptions may result in inefficient power performance and lead to wrong conclusions. Hence, a natural question is: how to propose a change point test method that enjoys simultaneous high powers for any given unknown alternative patterns. These are called data-adaptive tests. Next, we give a detailed discussion on this research direction.

4.1. Projection based methods in [2,58]

Essentially, for making a powerful change point test, it is important to increase the signal to noise ratio for the test method. One way is to project the data into a low dimensional subspace (such as one dimension), which maximizes the signal to noise ratio after projection. For example, [2] proposed to project the data into a one-dimensional subspace. In particular, let $\boldsymbol{v} \in \mathbb{R}^d$ be a projection vector. Then, we can use $\boldsymbol{v}$ to project the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ into a one-dimensional data sequence $\boldsymbol{v}^\top \boldsymbol{X}_1, \ldots, \boldsymbol{v}^\top \boldsymbol{X}_n$ with the model:

$$\boldsymbol{v}^\top \boldsymbol{X}_i = \boldsymbol{v}^\top \boldsymbol{\mu} + \boldsymbol{v}^\top \boldsymbol{\delta} \mathbf{1}\{i \geq k_0\} + \boldsymbol{v}^\top \boldsymbol{\epsilon}_i, i \in \{1, \ldots, n\}. \tag{21}$$

Let $Z_i = \boldsymbol{v}^\top \boldsymbol{X}_i$, $i \in \{1, \ldots, n\}$. For the projected Model (21), we can use the low dimensional CUSUM based method for constructing a test statistic as:

$$T_{\text{Pro}} = \max_{1 \leq k \leq n-1} |U(k)|,$$

where

$$U(k) = \sqrt{n}\frac{k}{n}\frac{n-k}{n}\frac{1}{\tau(v)}\Big(\frac{1}{n-k}\sum_{i=k+1}^{n} Z_i - \frac{1}{k}\sum_{i=1}^{k} Z_i\Big)$$

is the CUSUM statistic and $\tau^2(\boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$ is the variance for errors after projection. For any given projection $\boldsymbol{v} \in \mathbb{R}^d$, it is shown in [2] that, under $\mathbf{H}_0$, with some regular spatial and temporal conditions, the following holds, as $n \to \infty$,

$$T_{\text{Pro}} \xrightarrow{d} \sup_{0 \leq t \leq 1} |B(t)|,$$

where $B(\cdot)$ is a standard Brownian bridge. Under $\mathbf{H}_1$, [2] showed that the test statistic T_{Pro} has a high dimensional efficiency as:

$$\mathcal{E}(\boldsymbol{\delta}, \boldsymbol{v}) = \frac{|\boldsymbol{v}^\top \boldsymbol{\delta}|}{\tau(\boldsymbol{v})} = \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}\right\|_2 \cos\big(\alpha_{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}}\big), \tag{22}$$

where $\alpha_{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}}$ denotes the angle between $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}$. Hence, the last step for the projection based method is to find $\boldsymbol{v}$ which maximizes $\mathcal{E}(\boldsymbol{\delta}, \boldsymbol{v})$. One can see that the best projection is $\boldsymbol{v}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, which is also called the oracle projection. Using $\boldsymbol{v}^*$, the HDE for T_{Pro} is $\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}\right\|_2$. As a special case, when the covariance structure of $\boldsymbol{\epsilon}$ is identity with $\boldsymbol{\Sigma} = \mathbf{I}_d$, the HDE of T_{Pro} reduces to $\|\boldsymbol{\delta}\|_2$.

According to [2], if we know $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$, we can construct an oracle projection based test, which does not rely on any pre-knowledge about the alternative patterns. As a result, it has the highest efficiency, which can be regarded as an upper benchmark for the existing methods theoretically. In practice, however, it is of great difficulty to estimate $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}^{-1}$ simultaneously and construct efficient tests based on them. To address this issue, [58] proposed new algorithms for estimating $\boldsymbol{\delta}$ and used projection based ideas for estimating change point locations. In particular, suppose $\mathbf{H}_1$ holds and assume $\boldsymbol{\epsilon}_i \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I}_d)$. [58] observed that $\boldsymbol{\delta}$ can be approximated by the s-sparse leading left singular vector of the CUSUM matrix $\mathcal{C}$ in (7), and the corresponding optimization problem can be solved efficiently using convex relaxation. Once the oracle projection's estimator $\widehat{\boldsymbol{\delta}}$ is obtained, [58] projected the data matrix $\mathbf{X}$ along the direction $\widehat{\boldsymbol{\delta}}$, and applied the existing one-dimensional change point localization technique on the projected data for estimating the change point location k_0. Note that [58] mainly focused on change point estimation instead of testing.

4.2. Thresholded ℓ_1-norm based method in [16]

In addition to projection, another possible way is to find an appropriate norm for extracting the signals as much as possible. Recall $\Pi = \{j : \delta_j \neq 0\}$ is the set of coordinates with a change point and $s := |\Pi|$ is its cardinality. Let

$$|\delta_{(1)}| \geq |\delta_{(2)}| \geq \cdots |\delta_{(s)}| \geq \underbrace{|\delta_{(s+1)}| \geq \cdots \geq |\delta_{(d)}|}_{0s} \tag{23}$$

be the ordered statistics of the true signals $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_d)^\top$. For high dimensional change point inference, it is the first s largest entries in $\boldsymbol{\delta}$ that distinguish $\mathbf{H}_1$ from $\mathbf{H}_0$. In other words, we are more primarily interested in coordinates with stronger signal jumps than those with smaller ones or zero. Motivated by this observation, [16] proposed the following thresholded ℓ_1-norm based test statistic:

$$T_{\mathrm{Cho}} = \max_{1\leq k\leq n-1} \sum_{j=1}^{d} \big|C_j(k)\big| \mathbf{1}\big\{|C_j(k)| \geq \pi_n\big\},$$

where π_n is some user prespecified threshold parameter.

Using π_n, we see that the individual CUSUM statistics with values larger than π_n reflect contributions in detecting and localizing the change points and are summed up to the final statistics T_{Cho}. This method is also called a sparsified step since the coordinates with small contributions are disregarded. Note that π_n is derived such that $\max_{1\leq k\leq n-1} \big|C_j(k)\big| \geq \pi_n$ for all $k \in \Pi$ and $\max_{1\leq k\leq n-1} \big|C_j(k)\big| < \pi_n$ for all $k \in \Pi^c$ hold with probability tending to one. Assuming at least one change point occurs in the data sequence, [16] proved that coupled with the binary segmentation algorithm, T_{Cho} can correctly identify the number and locations of multiple change points (see Theorem 1 therein) with assumptions that $\delta_{\min} \geq c_*$ for some $c_* > 0$ and $dn^{-\log n} \to 0$ as $n \to \infty$.

4.3. Double CUSUM based method in [15]

Note that the implementation of [16] involves a selection of the threshold parameter π_n, whose value depends on the unknown underlying data generation mechanism such as the number and locations of change points. Hence, from a practical viewpoint, the selection of π_n is not an easy task. To overcome this limitation, [15] proposed a Double CUSUM (DC) based testing procedure that aims to select π_n in a data-driven way. In particular, recall the CUSUM statistic $\boldsymbol{C}(k) = (C_1(k), \ldots, C_d(k))^\top$ as defined in (3). Let

$$|C_{(1)}(k)| \geq C_{(2)}(k) \geq \cdots \geq |C_{(d)}(k)| \tag{24}$$

be the ordered CUSUM statistics at each candidate search location $k \in \{1, \ldots, n-1\}$. For each fixed $m \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, n-1\}$, [16] proposed the DC based test statistic as:

$$C_{\mathrm{DC}}^{\varphi}(m, k) = \left\{\frac{m(2d-m)}{2d}\right\}^{\varphi} \Big(\frac{1}{m}\sum_{j=1}^{m} \big|C_{(j)}(k)\big| - \frac{1}{2d-m}\sum_{j=m+1}^{d} \big|C_{(j)}(k)\big|\Big),$$

where $\varphi \in [0, 1]$ is some user pre-specified parameter to account for the alternative patterns which is discussed later. Note that when $\varphi = 1/2$, it reduces to the classical CUSUM statistic except that the data are ordered CUSUM statistics $|C_{(1)}(k)|, \ldots, |C_{(d)}(k)|$. This is why it is called Double CUSUM. Then, for a fixed $\varphi \in [0, 1]$, the final test statistic in [15] is

$$T_{DC}^{\varphi} = \max_{1\leq k\leq n-1} \max_{1\leq m\leq d} C_{\mathrm{DC}}^{\varphi}(m, k). \tag{25}$$

The main idea for T_{DC}^{φ} is that if there are s non-zero elements in $\boldsymbol{\delta}$ having a change point, there is a big gap between the average for the first s largest CUSUM statistics and the last $d-s$ ones. Since s is typically unknown, it is natural to search all candidate $s \in \{1, \ldots, d\}$ maximizing the gap. Based on T_{DC}^{φ}, a change point is detected if $T_{DC}^{\varphi} \geq \pi_n^{\varphi}$, where π_n^{φ} is a test criterion. According to [15], under some regular conditions, if we choose π_n^{φ} properly in theory, under $\mathbf{H}_0$, we have $\mathbb{P}(T_{DC}^{\varphi} \geq \pi_n^{\varphi}) \to 0$ as $n \to \infty$. This controls the type I error asymptotically. As for the power analysis, suppose $k_0 \asymp n$ and let $\bar{\delta} := s^{-1}\sum_{j\in\Pi} |\delta_j|$ be the average of non-zero elements with a change point. According to [15], if the signal strength satisfies:

$$\frac{\sqrt{n}\bar{\delta}}{\left(\frac{d}{s}\right)^{\varphi} \log n} \to \infty, \tag{26}$$

with probability tending to one, a change point is detected.

According to (26), we see that the HDE for T_{DC}^{φ} is $s^{\varphi}\bar{\delta}/(d^{\varphi}\log n)$. As for the choice of φ, it is shown that $\varphi = 0$ corresponds to a test which is powerful for sparse alternatives while $\varphi > 0$ is sensitive to dense alternatives. To see this, consider the special case with high sparsity with $s = 1$. The HDE of T_{DC}^{0} is $\bar{\delta}/\log(n)$, which has an efficiency loss of $\log n$ compared to

the oracle projection based method [2] having an efficiency of $\|\boldsymbol{\delta}\|_2$. For the high dense case with $s \asymp d^{\beta}$ with $\beta \in (1/2, 1]$, the HDE of T_{DC}^{φ} is

$$\frac{\bar{\delta}}{d^{\varphi(1-\beta)}\log n}, \quad \text{for any } \varphi > 0.$$

In this case, combining (10) and $\|\boldsymbol{\delta}\|_2 \asymp \sqrt{s}\bar{\delta}$, we see that there is an efficiency loss of an order of $d^{\varphi-(\varphi-1/2)\beta}\log n$, compared to the oracle case with an efficiency of $\|\boldsymbol{\delta}\|_2$.

4.4. The (s_0, p)-norm based data-adaptive test in [43]

Note that [16] and [15] essentially used a thresholded ℓ_1-norm for aggregating the CUSUM statistics, which aims at selecting coordinates $\delta_{(1)}, \ldots, \delta_{(s)}$ in (23) that are more relevant to change points. Using a similar but more general framework, [43] proposed a class of the (s_0, p)-norm based testing statistics. In particular, for a vector $\boldsymbol{v} = (v_1, \ldots, v_d)^{\top} \in \mathbb{R}^d$, with $|v_{(1)}| \geq \cdots \geq |v_{(d)}|$ being the ordered statistic for $|v_1|, \ldots, |v_d|$, define its (s_0, p)-norm as follows:

$$\|\boldsymbol{v}\|_{(s_0,p)} := \Big(\sum_{j=1}^{s_0} |v_{(j)}|^p\Big)^{1/p}, \tag{27}$$

where $s_0 \in \{1, \ldots, d\}$ and $1 \leq p \leq \infty$. Essentially, the (s_0, p)-norm is the ℓ_p-norm for the first s largest entries (in absolute value) for $\boldsymbol{v}$. Hence, it can be regarded as an adjusted ℓ_p-norm since it uses the ordered statistics of $|v_{(1)}|, \ldots, |v_{(d)}|$. Note that by choosing a proper combination of s_0 and p, $\|\boldsymbol{v}\|_{(s_0,p)}$ reduces to some classical ℓ_p-norm adopted in the literature. For example, for $p = 2$, if we choose $s_0 = d$, it reduces to the traditional ℓ_2-norm adopted in [31,67]. For any given s_0, if we choose $p = \infty$, it reduces to the ℓ_∞-norm used in [35,63]. Moreover, for a given s_0 with $p = 1$, it can be regarded as the thresholded ℓ_1-norm proposed in [16]. Hence, the (s_0, p)-norm is a flexible generalization of the existing methods. Using (27), for a user-prespecified $s_0 \in \{1, \ldots, d\}$, [43] proposed a class of testing statistics with respect to different p as follows:

$$T_{(s_0,p)} = \max_{\underline{k} \leq k \leq n-\underline{k}} \sqrt{\frac{k}{n}\frac{n-k}{n}} \big\|\boldsymbol{C}(k)\big\|_{(s_0,p)}, \text{ with } 1 \leq p \leq \infty, \tag{28}$$

where $\underline{k}$ and $n - \underline{k}$ are the lower and upper search locations, respectively. Note that for a given s_0, the statistic $T_{(s_0,p)}$ with a small p (e.g., $p = 1, 2$) is more sensitive to dense alternatives while that with a large p (e.g., $p = \infty$) is more powerful under sparse alternatives. Hence, for any unknown alternative pattern, there exists at least one test in $\{T_{(s_0,p)}, 1 \leq p \leq \infty\}$ enjoying a powerful performance, which is called individual test statistics. For each $T_{(s_0,p)}$, it is very difficult to directly derive its limiting null distribution. Therefore, [43] proposed to use a multiplier bootstrap based procedure for obtaining a good approximation. In particular, for the bth bootstrap, $b \in \{1, \ldots, B\}$, let $e_1^b, \ldots, e_n^b \overset{i.i.d}{\sim} N(0, 1)$. Define the bth bootstrap version of the CUSUM statistic as:

$$\boldsymbol{C}^b(k) = \sqrt{\frac{k(n-k)}{n}}\Big(\frac{1}{n-k}\sum_{i=k+1}^{n} e_i^b\big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}(n-k)\big) - \frac{1}{k}\sum_{i=1}^{k} e_i^b\big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}(k)\big)\Big). \tag{29}$$

Then, for each $T_{(s_0,p)}$, conditional on $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, define its bth bootstrap based version as:

$$T_{(s_0,p)}^b = \max_{\underline{k} \leq k \leq n-\underline{k}} \sqrt{\frac{k}{n}\frac{n-k}{n}} \big\|\boldsymbol{C}^b(k)\big\|_{(s_0,p)}, \text{ with } 1 \leq p \leq \infty. \tag{30}$$

Liu et al. [43] proved that under $\mathbf{H}_0$, with some regular conditions, we can use $T_{(s_0,p)}^b$ to approximate $T_{(s_0,p)}$:

$$\sup_{z \in (0,\infty)} \big|\mathbb{P}(T_{(s_0,p)} \leq z) - \mathbb{P}(T_{(s_0,p)}^b \leq z | \mathbf{X})\big| = o_p(1), \text{ as } n, d \to \infty. \tag{31}$$

Note that the derivation of (31) allows the data dimension d grows exponentially with the sample size n in the sense that $s_0^2 \log(dn) = O(n^{\delta})$ for some $0 < \delta < 1/7$. Based on the bootstrap samples $\{T_{(s_0,p)}^1, \ldots, T_{(s_0,p)}^B\}$, we can obtain a $1 - \alpha$ critical value $\widehat{c}_{\alpha,(s_0,p)}$ and construct a test as $\Psi_{\alpha,(s_0,p)} := \mathbf{1}\{T_{s_0,p} \geq \widehat{c}_{\alpha,(s_0,p)}\}$. The result in (31) guarantees that the individual test $\Psi_{\alpha,(s_0,p)}$ has the asymptotic level α.

In terms of the power analysis, [43] requires the signal strength satisfies:

$$\sqrt{n}\Big\|\Big(\frac{k_0}{n}\frac{n-k_0}{n}\Big)\boldsymbol{\delta}\Big\|_{(s_0,p)} \geq C_0 s_0(1+\epsilon_n)\Big(\sqrt{2\log(d(n-2\underline{k}))} + \sqrt{2\log(\alpha^{-1})}\Big), \tag{32}$$

where $\epsilon_n = o(1)$ with $\epsilon_n\sqrt{\log nd} \to \infty$, $\alpha \in (0, 1)$ is the significance level, and C_0 is some big enough constant. Under $\mathbf{H}_1$, if (32) holds, [43] proved that with probability tending to one, a change point is detected:

$$\mathbb{P}(\Psi_{\alpha,(s_0,p)} = 1) \to 1, \quad \text{as } n, d, B \to \infty.$$

Remark 4. [43] requires the signal strength satisfies $\|\boldsymbol{\delta}\|_{(s_0,p)} \geq C_0 s_0 \sqrt{\frac{\log d}{n}}$. Hence, the HDE of $T_{(s_0,p)}$ is $\frac{\|\boldsymbol{\delta}\|_{(s_0,p)}}{s_0\sqrt{\log d}}$. Moreover, we show in Section 5.4 that for sparse alternatives, $T_{(s_0,p)}$ obtains the minimax optimality separation rate for a given but fixed s_0.

Once the individual test statistics $\{T_{(s_0,p)}, 1 \leq p \leq \infty\}$ are constructed, the remaining step is how to combine them to yield a powerful test that adapts to the unknown alternative patterns. Note that a small p-value leads to a rejection of $\mathbf{H}_0$. For a user prespecified s_0, [43] proposed a data-adaptive test statistic by taking the minimum p value among the individual tests:

$$T_{\text{ad}} = \min_{p\in\mathcal{P}} \widehat{P}_{(s_0,p)}, \tag{33}$$

where $\widehat{P}_{(s_0,p)}$ is the approximated p-value for each $T_{(s_0,p)}$, which is obtained using the bootstrap samples, and $\mathcal{P}$ is a candidate subset of p satisfying $|\mathcal{P}| < \infty$. In practice, it is recommended to choose $\mathcal{P} = \{1, 2, 3, 4, 5, \infty\}$ for enjoying simultaneous high powers across various alternative patterns. Moreover, note that $\widehat{P}_{(s_0,p)}$ is typically correlated and the distribution of T_{ad} is very difficult to derive. Hence, [43] proposed a low cost bootstrap based procedure for obtaining the p-value for the data-adaptive test T_{ad}, see Algorithm 2 therein.

Remark 5. As discussed by [43], the data-adaptive test T_{ad} is robust against the choice of s_0, given s_0 is not too small. In practice, it is recommended to use $s_0 = d/2$, even though it is theoretically required to be $s_0 = \log^{\delta_2} d$ for some $\delta_2 > 0$ under $\mathbf{H}_1$.

4.5. The ℓ_p-norm based data-adaptive test in [68]

We have discussed that the ℓ_2-norm is powerful for dense alternatives. Moreover, it is known that $\ell_p \to \ell_\infty$ as $p \to \infty$. Hence, for the general ℓ_p-norm, a larger value of p may yield powerful performance for sparse alternatives. Motivated by this observation, [68] proposed a class of the ℓ_p-norm based individual test statistics and combine them to construct a data-adaptive method. Specifically, for each fixed even number $p \in \{2, 4, \ldots, \}$, define the two-sample U-statistic based process

$$U_p(k) = \sum_{\ell=1}^{d} \sum_{1\leq i_1,\ldots,i_p\leq k} \sum_{k+1\leq j_1,\ldots j_p\leq n} \underbrace{(X_{i_1\ell} - X_{j_1\ell}) \times (X_{i_2\ell} - X_{j_2\ell}) \times \cdots \times (X_{i_p\ell} - X_{j_p\ell})}_{p},$$
$$\text{with } \ k \in \{2p, 2p+1, \ldots, n-2p\}. \tag{34}$$

Note that under $\mathbf{H}_1$, choosing $k = k_0$, we have

$$\mathbb{E}\Big[\frac{U_p(k)}{A_k^p A_{n-k}^p}\Big] = \|\boldsymbol{\delta}\|_p^p, \quad \text{for } \ p \in \{2, 4, \ldots\},$$

where $A_k^p = \frac{k!}{(k-p)!}$ and $A_{n-k}^p = \frac{(n-k)!}{(n-k-p)!}$. Hence, $U_p(k)$ can be regarded as an estimation for $\|\boldsymbol{\delta}\|_p^p$. For $p = 2$, it reduces to (19) up to some constants. Then, for a user prespecified $p \in 2N$, [68] proposed the following individual test statistic with respect to different p:

$$T_{n,p} = \max_{2p\leq k\leq n-2p} \frac{U_p^2(k)}{W_p(k)},$$

where

$$W_p(k) = \frac{1}{n}\sum_{t=p}^{k-p} U_p^2(t; 1, k) + \frac{1}{n}\sum_{t=k+p}^{n-p} U_p^2(t; k+1, n),$$

and

$$U_p(t; s, m) = \sum_{\ell=1}^{d} \sum_{s\leq i_1,\ldots,i_p\leq t} \sum_{t+1\leq j_1,\ldots j_p\leq m} (X_{i_1\ell} - X_{j_1\ell}) \times (X_{i_2\ell} - X_{j_2\ell}) \times \cdots \times (X_{i_p\ell} - X_{j_p\ell}).$$

Note that the test statistic T_p is a self-normalized test statistic using W_p in the denominator to cancel out the asymptotical variance in the limiting distribution of $U_p^2(k)$. This makes the limiting distribution of T_p becomes pivotal without any unknown parameters. Moreover, $T_{n,p}$ with a smaller value of p is more powerful for the dense case and that with a larger one is sensitive to sparse alternatives.

It is shown in [68] that under $\mathbf{H}_0$, for any given $p \in 2N$, we have, as $n \to \infty$, $T_{n,p} \xrightarrow{d} \widetilde{T}_p$, where $\widetilde{T}_p$ is the corresponding limiting null distribution whose critical value can be obtained using Monte Carlo simulations. More importantly, as shown

in [68], for any $p_1 \neq p_2$, the two limiting distributions $\widetilde{T}_{p_1}$ and $\widetilde{T}_{p_2}$ are asymptotically independent. Hence, in theory, it is possible to construct a family of asymptotically independent individual test statistics $\{T_{n,p}, p \in 2N\}$.

As for the power performance, for each $T_{n,p}$, [68] showed that under $\mathbf{H}_1$, as $n \to \infty$, the following results hold:

(i) If $\dfrac{\sqrt{n}\|\boldsymbol{\delta}\|_p}{\|\boldsymbol{\Sigma}\|_p^{1/2}} \to 0$, then $T_{n,p} \xrightarrow{d} \widetilde{T}_p$;

(ii) If $\dfrac{\sqrt{n}\|\boldsymbol{\delta}\|_p}{\|\boldsymbol{\Sigma}\|_p^{1/2}} \to \infty$, then $T_{n,p} \xrightarrow{\mathbb{P}} \infty$;

(iii) If $\dfrac{\sqrt{n}\|\boldsymbol{\delta}\|_p}{\|\boldsymbol{\Sigma}\|_p^{1/2}} \to \gamma \in (0, \infty)$, then $T_{n,p} \xrightarrow{d} \widetilde{\widetilde{T}}_p$,

where $\|\boldsymbol{\Sigma}\|_p$ denotes the element-wise ℓ_p-norm for $\boldsymbol{\Sigma}$, and $\widetilde{\widetilde{T}}_p$ is a pivotal limit with an additive shift compared to $\widetilde{T}_p$. Hence, from the above results, we know that the HDE for each $T_{n,p}$ is $\|\boldsymbol{\delta}\|_p/\|\boldsymbol{\Sigma}\|_p^{1/2}$. When $p = 2$ with $\boldsymbol{\Sigma} = \mathbf{I}_d$, it reduces to $d^{-1/4}\|\boldsymbol{\delta}\|_2$, which is equivalent to [31,59].

Once the individual tests are constructed, it is desirable to construct a data-adaptive method. Similar to the idea of [43,68] proposed a data-adaptive test statistic using the minimum p-value:

$$W_{\text{ad}} = \min_{p \in \mathcal{P}} P_p,$$

where P_p is the theoretical p-value for $T_{n,p}$, which can be obtained using Monte Carlo simulations, and $\mathcal{P}$ is a candidate subset of p satisfying $|\mathcal{P}| < \infty$. Since the individual tests are asymptotically independent, the p-value for W_{ad} can be calculated directly as $1 - (1 - W_{\text{ad}})^{|\mathcal{P}|}$. According to [68], choosing $\mathcal{P} = \{2, 6\}$ enjoys good size and power performance across various alternative patterns.

Remark 6. Both [43] and [68] constructed a data-adaptive method by combining the individual tests using the minimum p-value. There are some essential differences between them. First, [43] adopted the (s_0, p)-norm for the sample mean difference $\boldsymbol{C}(k)$. Typically, $\|\mathbf{C}(k_0)\|_{(s_0,p)}$ is not an unbiased estimator for $\|\boldsymbol{\delta}\|_{(s_0,p)}$. In contrast, [68] directly estimated the ℓ_p-norm of $\boldsymbol{\delta}$ using the two-sample U-statistic based process $U_p(k)$ as in (34). Second, the individual tests in [43] are not independent while those in [68] are asymptotically independent. Hence, the former introduced a low-cost bootstrap to obtain the data-adaptive test's p-value while the p-value in [68] can be calculated directly. Third, the proposed individual tests in [43] include the ℓ_∞-norm as a special case, while those in [68] include the ℓ_2-norm based method. The two special cases obtain the minimax optimality, under the sparse and dense cases, respectively. Lastly, the data-adaptive method in [43] can be extended to other high dimensional parameters such as variance, covariance matrix, or Kendall's tau correlation matrix (see Section 6). It appears unclear on how to use the idea of [68] for those general applications.

5. Minimax optimality for high dimensional change point inference

For high dimensional change point inference, one may be interested in whether the proposed method is optimal. In other words, it needs to show that the requirement on $\boldsymbol{\delta}$ for detecting a change point with overwhelming probability is the weakest. Formally, this question is equivalent to the concept of minimax optimality as introduced in Definition 2.2. Different from the traditional focus on the size and power performance, minimax optimality is a more refined result. Not much development on this topic has been made in the literature until recently under some special cases. To introduce these results, in this section, we assume the errors in Model (1) satisfy $\boldsymbol{\epsilon}_i \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I}_d)$.

5.1. Minimax lower bound in [25]

Recall $s =: |\Pi|$ is the sparsity level of $\boldsymbol{\delta}$ and $\rho(n, d, s, k_0)$ is the scaled signal strength for Model (1) defined in (8). Suppose $s \asymp d^{1-\beta}$ for $\beta \in [0, 1)$. Note that $\beta \in [0, 1/2)$ corresponds to the moderate sparse or dense level studied in [31,59], while $\beta \in (1/2, 1)$ corresponds to the highly sparse level studied in [35,63].

For Problem (9), [25] derived the minimax lower bound. In particular, [25] proved that for any $\epsilon \in (0, 1)$, if the signal strength $\rho := \rho(n, d, s, k_0)$ satisfies:

$$\limsup_{n\to\infty} \frac{\rho^2(n, d, s, k_0)}{\sqrt{d}} \leq \sqrt{2\log(1 + 4(1-\epsilon)^2)}, \text{ for } \beta \in [0, 1/2), \tag{35}$$

and

$$\limsup_{n\to\infty} \frac{\rho^2(n, d, s, k_0)}{s\log(d/s)} < 2 - \frac{1}{\beta}, \text{ for } \beta \in (1/2, 1), \tag{36}$$

then for any test ψ, we have $\liminf_{n\to\infty} \mathcal{R}(\psi, \rho) \geq 1 - \epsilon$. In other words, no tests can control the overall type I and type II errors with vanishing probability if (35) holds for dense alternatives and if (36) holds for sparse alternatives.

5.2. Minimax upper bound in [25]

To show the upper bounds, [25] proposed a linear and scan test statistic, which corresponds to the dense and sparse case, respectively. Specifically, recall the CUSUM statistic $\boldsymbol{C}(k)$ in (3). Then, for the moderate sparse or dense case with $\beta \in [0, 1/2)$, the linear statistic is defined as:

$$T_{\text{Linear}} = \max_{1\le k\le n-1} \frac{\|\boldsymbol{C}(k)\|_2^2 - d}{\sqrt{2d}}. \tag{37}$$

Based on (37), [25] proposed a linear type test $\psi_{\text{Linear}} = \mathbf{1}\{T_{\text{Linear}} \ge H\}$, where H is some critical value. For the highly sparse case with $\beta \in (1/2, 1)$, [25] proposed a scan type test statistic as:

$$\begin{aligned} T_{\text{Scan}} &= \max_{1\le k\le n-1} \max_{1\le m\le d} \frac{1}{H_m} \max_{\ell\in\mathcal{M}(d,m)} \left\{ \frac{\left\|\Pi_\ell \boldsymbol{C}(k)\right\|_2^2 - m}{\sqrt{2m}} \right\} \\ &= \max_{1\le k\le n-1} \max_{1\le m\le d} \frac{1}{H_m} \frac{\sum_{j=1}^m C_{(j)}^2(k) - m}{\sqrt{2m}}, \end{aligned} \tag{38}$$

where $\mathcal{M}(d, m)$ denotes the collection of all subsets of $\{1, \ldots, d\}$ of cardinality m, $\Pi_\ell \boldsymbol{v}$ is the projection of a vector $\boldsymbol{v} \in \mathbb{R}^d$ onto a subspace indexed by $\ell \in \mathcal{M}(d, m)$, H_m is some critical value to account for different sparsity m, and $C_{(j)}(k)$ is the ordered CUSUM statistic in (24). Based on (38), a scan type test is proposed as $\psi_{\text{Linear}} = \mathbf{1}\{T_{\text{Scan}} \ge 1\}$. Moreover, to account for the unknown alternative patterns, [25] proposed a combined test as:

$$\psi_{\text{adaptive}} = \max\left(\psi_{\text{Linear}}, \psi_{\text{Scan}}\right).$$

For the above data-adaptive test ψ_{adaptive}, it is proved by [25] that if $\rho := \rho(n, d, s, k_0)$ satisfies

$$\liminf_{n\to\infty} \min_{1\le k_0\le n-1} \frac{\rho^2(n, d, s, k_0)}{\sqrt{d\log(d\log n)}} \ge \sqrt{2}, \quad \text{for } \beta \in [0, 1/2), \tag{39}$$

or

$$\liminf_{n\to\infty} \min_{1\le k_0\le n-1} \min_{1\le s\le d} \frac{\rho^2(n, d, s, k_0)}{s\log(d/s)} \ge 2, \quad \text{for } \beta \in (1/2, 1), \tag{40}$$

then for any $0 < \epsilon < 1$, we have $\limsup_{n\to\infty} \mathcal{R}(\psi_{\text{adptive}}, \rho) \le \epsilon$. In other words, if the signal strength is large enough such that (39) or (40) holds, the combined test has vanishing type I and type II errors.

Remark 7. For the highly sparse case, we see that the lower bound (36) matches the upper bound (40). Hence, the minimax optimal rate is $\rho^2(n, d, s, k_0) \asymp s\log(d/s)$. For the moderate sparse or dense case, we see that the change point is not detectable if $\rho^2(n, d, s, k_0) \asymp \sqrt{d}$, and can be detected with overwhelming probability for the linear type test in (37) if $\rho^2(n, d, s, k_0) \asymp \sqrt{d\log(d\log n)}$. Hence, there is a gap of an order of $\sqrt{\log(d\log n)}$ between the lower and upper bounds derived in [25].

5.3. Exact asymptotic constants for the minimax optimality separation rate in [41]

As an extension of [25,41] further proved the exact minimax separation rate for the dense case. In particular, according to [41], we have the following results:

- Dense case: Assume $s^2/(d\log\log n) \to \infty$ as $n \to \infty$. Suppose $\rho^2(n, d, s, k_0) = \xi\sqrt{d\log\log n}$. Then, there exists a test ψ^* such that $\lim_{n\to\infty} \mathcal{R}(\psi^*, \rho) \to 0$ if $\xi > 2$, and for any test ψ, $\lim_{n\to\infty} \mathcal{R}(\psi, \rho) \to 1$ holds if $\xi < 2$.
- Sparse case: Assume $s^2/d \to 0$ and $s/(\log\log n) \to \infty$. Suppose $\rho^2(n, d, s, k_0) = \xi s\log\big(\frac{d\log\log n}{s^2}\big)$. Then, there exists a test ψ^* such that $\lim_{n\to\infty} \mathcal{R}(\psi^*, \rho) \to 0$ if $\xi > 2$, and for any test ψ, $\lim_{n\to\infty} \mathcal{R}(\psi, \rho) \to 1$ holds if $\xi < 1$.

Remark 8. According to [41], the minimax optimality separation rate for the dense regime is $\rho^2(n, d, s, k_0) \asymp \sqrt{d\log\log n}$ with a sharpest constant 2. Moreover, for the sparse regime, both [25] and [41] derived the minimax separation rate with an order of $\rho^2(n, d, s, k_0) \asymp s\log(d/s)$.

5.4. Minixmax optimality for the existing methods

In this section, we show that the rates derived in [43,63] are minimax for detecting sparse change point alternatives, and those in [59,68] are optimal for the dense alternatives (up to a logarithmic factor). Recall $\rho^2(n, d, s, k) \asymp \frac{k(n-k)}{n}\|\boldsymbol{\delta}\|_2^2$.

Table 1
Signal requirements and main model assumptions of the existing methods for detecting a change point with power tending to one. The change point location k_0 is assumed to be $c_1 n \le k_0 \le c_2 n$ for some constants $0 < c_1 < c_2 < 1$.

Method	$\boldsymbol{\delta}$	$\boldsymbol{\epsilon}$	$\boldsymbol{\Sigma}$	Allow temporal dependence?	d and n
ℓ_∞-norm based [35]	$\|\boldsymbol{\delta}\|_\infty \gg \sqrt{\frac{\log n}{n}}$	Uniform Polynomial Moment	Strong	Yes	$d \ll n^c$
ℓ_∞-norm based [63]	$\|\boldsymbol{\delta}\|_\infty \ge C_0\sqrt{\frac{\log d}{n}}$	Sub-exponential Distribution (or Uniform Polynomial Moment)	Mild	No	$d \asymp \exp(n^c)$
Scan type test [25]	$\|\boldsymbol{\delta}\|_2 \ge \sqrt{2\frac{s\log(d/s)}{n}}$	Gaussian Distribution	Strong	No	$\frac{\log n}{s\log(d/s)} \to 0$
ℓ_2-norm based [31]	$\|\boldsymbol{\delta}\|_2 \gg \frac{d^{1/4}}{\sqrt{n}}$	Uniform Polynomial Moment	Strong	Yes	$d = o(\sqrt{n})$
ℓ_2-norm based [59]	$\|\boldsymbol{\delta}\|_2 \gg \frac{\|\boldsymbol{\Sigma}\|_F^{1/2}}{\sqrt{n}}$	Uniform Polynomial Moment	Mild	Yes	$d \to \infty$
Linear type test [25]	$\|\boldsymbol{\delta}\|_2 \ge \frac{(2d\log(d\log n))^{1/4}}{\sqrt{n}}$	Gaussian Distribution	Strong	No	$\frac{\log n}{s\log(d/s)} \to 0$
Oracle projection based [2]	$\|\boldsymbol{\delta}\|_2 \gg \frac{1}{\sqrt{n}}$	Uniform Polynomial Moment	Mild	No	$d \to \infty$
Double CUSUM [15]	$\bar{\delta} \gg \frac{\left(\frac{d}{s}\right)^{\varphi}\log n}{\sqrt{n}}, \varphi \in [0,1]$ $(\bar{\delta} := s^{-1}\sum_{j\in\Pi}\|\delta_j\|)$	Sub-exponential Distribution	Mild	Yes	$d \asymp n^c$
Thresholded ℓ_1-norm [16]	$\|\boldsymbol{\delta}\|_\infty \ge C_0$	Chi-squared distribution	Mild	Yes	$\frac{d}{n^{\log n}} \to 0$
(s_0, p)-norm based [43]	$\|\boldsymbol{\delta}\|_{(s_0,p)} \ge C_0 s_0\sqrt{\frac{\log d}{n}}$	Sub-exponential Distribution	Mild	No	$d \asymp \exp(n^c)$
ℓ_p-norm based method [68]	$\|\boldsymbol{\delta}\|_p \gg \frac{\|\boldsymbol{\Sigma}\|_p^{1/2}}{\sqrt{n}}, p \in \{2,4,\dots,\}$	Uniform Polynomial Moment	Mild	Yes	$d \to \infty$

For $\boldsymbol{\delta}$, suppose (10) holds. Then, we have $\rho^2(n,d,s,k_0) \asymp \frac{k_0(n-k_0)}{n}\|\boldsymbol{\delta}\|_2^2 \asymp \frac{k_0(n-k_0)}{n}s\delta_{\max}^2$. According to [25], we know that the detection boundary for the sparse alternative is an order of

$$\delta_{\max} \asymp \sqrt{\frac{\log d}{nh(k_0)}}, \quad \text{with} \quad h(k_0) = \frac{k_0}{n}\Big(1 - \frac{k_0}{n}\Big).$$

Combining the above results with (16) and (32), we see that the ℓ_∞-norm based method in [63] and the (s_0, p)-norm based individual test (with a fixed given s_0) in [43] are optimal for detecting sparse alternatives.

For the dense case, according to [41], the optimal rate is an order of

$$\|\boldsymbol{\delta}\|_2 \asymp \frac{(d\log\log n)^{1/4}}{n^{1/2}h(k_0)}.$$

Hence, for the ℓ_2-norm based methods in [31,59], they are rate optimal up to the logarithmic factor.

To end this section, for the existing state-of-art techniques, we summarize the signal conditions for detecting a change point in Model (1) with probability tending to one. The results are provided in Table 1. In addition to that, we also report the corresponding model assumptions in terms of the moment condition on the underlying errors $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_d)^\top$, the spatial condition $\boldsymbol{\Sigma}$ among coordinates of $\boldsymbol{X} = (X_1, \dots, X_d)^\top$ (strong or mild), the temporal condition between observations, as well as the scaling relationship between d and n.

6. High dimensional change point inference for general parameters

In addition to change point inference for high dimensional mean vectors as in Problem (2), change point detection can be extended to other high dimensional parameters. In particular, let $F_i(\boldsymbol{x})$ be the cumulative distribution function for a high dimensional vector $\boldsymbol{X}_i = (X_{i1}, \dots, X_{id})^\top \in \mathbb{R}^d$ at time point $i \in \{1, \dots, n\}$. Let $\boldsymbol{\Gamma}$ be a function mapping the probability distribution $F_i(\boldsymbol{x})$ to some parameter space $\mathcal{V}$. Then, we can consider a general hypothesis problem:

$$\begin{aligned} &\mathbf{H}_0 : \boldsymbol{\Gamma}(F_1) = \cdots = \boldsymbol{\Gamma}(F_n) \ \text{ vs. the alternative that} \\ &\mathbf{H}_1 : \exists k_0 \in \{1, \dots, n-1\} \text{ such that } \boldsymbol{\Gamma}(F_1) = \cdots = \boldsymbol{\Gamma}(F_{k_0}) \ne \boldsymbol{\Gamma}(F_{k_0+1}) = \cdots = \boldsymbol{\Gamma}(F_n). \end{aligned} \tag{41}$$

For Problem (41), we can take $\boldsymbol{\Gamma}(F_i) = \mathbb{E}\boldsymbol{X}_i := \boldsymbol{\mu}_i$ for detecting high dimensional mean vectors and set $\boldsymbol{\Gamma}(F_i) = \mathbb{E}(\boldsymbol{X}_i - \mathbb{E}\boldsymbol{X}_i)(\boldsymbol{X}_i - \mathbb{E}\boldsymbol{X}_i)^\top := \boldsymbol{\Sigma}_i$ for detecting high dimensional covariance matrix, etc. Change point testing for general

high dimensional parameters is more technically involved and has gained increasing interest in some real applications. In recent years, there has been some developments for Problem (41) by taking a special function Γ. Here, we give a brief review on this topic.

For high dimensional variance vectors with $\boldsymbol{\Gamma}(F_i) := (\sigma_{i1}^2, \ldots, \sigma_{id}^2)^\top \in \mathbb{R}^d$ with $\sigma_{ij}^2 = \mathrm{Var}[X_{ij}]$, [16] proposed a sparsified binary segmentation algorithm, which aggregates the cumulative sum statistics using the thresholded ℓ_1-norm. For high dimensional covariance matrix with $\boldsymbol{\Gamma}(F_i) = \boldsymbol{\Sigma}_i \in \mathbb{R}^{d\times d}$, [59] proposed an ℓ_2-norm based self-normalized test statistic; [4] used an ℓ_∞-norm based test statistic for testing sparse changes in $\boldsymbol{\Sigma}$, and their method is based on the CUSUM matrix obtained from a de-biased lasso estimator. [55] adopted a projection based technique for testing and estimating change points in the covariance structure of a high-dimensional linear time series, including vector autoregressive moving average (VARMA) models and spiked covariance models as special cases. Dette et al. [23] applied a dimension reduction technique for estimating the single change point (suppose $\mathbf{H}_1$ holds) and [60] used the matrix ℓ_2-norm for localizing multiple change points (suppose $\mathbf{H}_1$ holds). Furthermore, [6,40] considered change point detection for large contemporaneous covariance matrices of high dimensional time series satisfying an approximate factor model.

In addition to mean vectors or covariance structures, another recent research development is to consider change point inference for more general parameters in the sense that $\boldsymbol{\Gamma}(F) = (\theta_1, \ldots, \theta_q)^\top$ with $\theta_s = \mathbb{E}[\Phi_s(\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')]$, where $\Phi_s(\boldsymbol{x}_1', \ldots, \boldsymbol{x}_m') : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$ is a measurable, symmetric (or anti-symmetric) kernel with order m, and $\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m'$ are independent copies with the same distribution as $\boldsymbol{X}$. This problem is known as U-statistic based change point inference since the parameter θ_s can be estimated using U-statistic. Note that by choosing a special kernel $\Phi_s(\cdot)$, the high dimensional parameter $\boldsymbol{\Gamma}(F)$ reduces to some specific problems. This includes the mean vectors, the covariance matrix, the Kendall's tau correlation matrix, and the Wilcoxon–Mann–Whitney based change point tests as special cases, where the latter two cases are known as robust change point testing methods. For testing the changes of $\boldsymbol{\Gamma}(F) = (\theta_1, \ldots, \theta_q)^\top$, similar to mean vectors, there are still some concerns about the alternative patterns (sparse or dense) of $\boldsymbol{\Gamma}(F_{k_0+1}) - \boldsymbol{\Gamma}(F_{k_0})$ in the high dimensional setting. Recently, several papers made progress on this issue. For example, based on U-statistics, [62] proposed an ℓ_∞-norm based change point test for location parameters; Constructing U-statistic based CUSUM, [43] considered the (s_0, p)-norm based method for detecting general parameters and a data-adaptive test statistic as in (33) was also proposed.

For high dimensional change point inference, another interesting problem is to consider the following non-parametric test:

$$\begin{aligned} &\mathbf{H}_0 : F_1(\boldsymbol{x}) = \cdots = F_n(\boldsymbol{x}) \ \text{ for all } \ \boldsymbol{x} \in \mathbb{R}^d \ \text{ vs. the alternative that} \\ &\mathbf{H}_1 : \exists k_0 \in \{1, \ldots, n-1\} \text{ such that } F_1(\boldsymbol{x}) = \cdots = F_{k_0}(\boldsymbol{x}) \neq F_{k_0+1}(\boldsymbol{x}) = \cdots = F_n(\boldsymbol{x}), \end{aligned} \tag{42}$$

where $F_{k_0}(\boldsymbol{x})$ and $F_{k_0+1}(\boldsymbol{x})$ differ on sets of non-zero measures.

Note that Problem (42) does not rely on any distributional assumption or is limited to a particular parameter such as mean vectors or covariance matrices. Hence, it can detect general types of changes in the data generating distribution. In the low dimensional setting with a fixed d, there are some works based on the empirical cumulative functions [21], the empirical characteristic functions [33], the energy distance [44], or some kernel based procedures [1] for testing (42). In high dimensions with $d \gg n$, the traditional methods are neither no longer applicable nor lack theoretical justifications. Developing non-parametric change point tests in high dimensions is a challenging task. Several papers appeared in the literature and we name a few here. Chen et al. [14] proposed a graph based scan test statistic for detecting a single change-point or two change points. Their method can be applied to non-Euclidean data. Chu and Chen [18] improved [14] for detecting changes of scale alternatives while maintaining high powers for changes in the middle of data observations. In a recent work, [10] observed that in high dimensions, the energy distance as used in [44] is only able to capture the information of the differences of the first two moments (mean and variance) between two distributions. Hence, the method in [44] fails to detect general distributional changes beyond the first two moments in high dimensions. To overcome this problem, using a generalized homogeneity metric, [10] proposed a nonparametric change point test for the homogeneity of two high-dimensional distributions. Their method is based on a cumulative sum based process in an embedded Hilbert space, which can be regarded as an extension of [44] to the high dimensional settings.

7. Extensions to multiple change points inference

So far, we have reviewed methods for change point inference for AMOC in (1) and discussed their theoretical properties. One natural extension for Model (1) is to investigate inference for high dimensional multiple change points. Specifically, by letting $\boldsymbol{\mu}_i = \mathbb{E}\boldsymbol{X}_i$, $i \in \{1, \ldots, n\}$, we can consider the following hypothesis:

$$\begin{aligned} &\mathbf{H}_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_n \ \text{ vs. the alternative that} \\ &\mathbf{H}_1 : \exists\ 1 < k_1 < \cdots < k_{m^*} < n \text{ such that } \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_{k_1-1} \neq \boldsymbol{\mu}_{k_1} = \cdots = \boldsymbol{\mu}_{k_{m^*}-1} \neq \boldsymbol{\mu}_{k_{m^*}} = \cdots = \boldsymbol{\mu}_n. \end{aligned} \tag{43}$$

In other words, there are m^* unknown change points $k_1 < \cdots < k_{m^*}$ that divide the data into $m^* + 1$ segments with different constant mean vectors across the segments. Note that for multiple change point models, both the number of change points m^* and the locations $\{k_1, \ldots, k_{m^*}\}$ are typically unknown. This is essentially different from AMOC in (1)

where $m^* \leq 1$ is assumed. Moreover, as discussed in the previous sections, the high dimensionality also brings great challenges for change point inference. The above two aspects make the testing of (43) a non-trivial task. For detecting high dimensional multiple change points, the common strategy for the existing techniques is to combine the test statistic designed for a single change point as in Sections 3–4, with the binary segmentation (BS) algorithm in [57], the wild binary segmentation (WBS) algorithm in [26], the moving sum (MOSUM) based procedure in [19,24], or the scan based method in [66]. Next, we give a summary for this research direction.

The binary segmentation algorithm is one of the most commonly adopted methods for (43). Let $(s, e) \subset \{1, \ldots, n-1\}$ be a candidate search interval. Let

$$\boldsymbol{C}(s,k,e) = \sqrt{\frac{(k-s+1)(e-k)}{e-s+1}}\Big(\frac{1}{e-k}\sum_{i=k+1}^{e}\boldsymbol{X}_i - \frac{1}{k-s+1}\sum_{i=s}^{k}\boldsymbol{X}_i\Big), \tag{44}$$

be the CUSUM statistic calculated using the samples $\{\boldsymbol{X}_s, \ldots, \boldsymbol{X}_e\}$. Suppose $T_{(s,e)}$ is a test statistic with $T_{(s,e)} = \max_{s\leq k\leq e} H\big(\boldsymbol{C}(s,k,e)\big)$, where $H(\cdot)$ denotes a general aggregation for $\boldsymbol{C}(s,k,e)$. The choice of $H(\cdot)$ may depend on the alternative pattern. For example, we can take $H(\cdot) = \|\cdot\|_\infty$ as in [35], $H(\cdot) = \|\cdot\|_2$ in [31], or $H(\cdot) = \|\cdot\|_{(s_0,p)}$ in [43]. Suppose $c_{(s,e)}$ is a critical value (either obtained from a limiting null distribution or using bootstrap). The main idea of BS is that for the candidate search interval (s, e), we use $T_{(s,e)}$ and $c_{(s,e)}$ to detect the existence of a change point. If $\mathbf{H}_0$ is rejected, we identify a new change point b by taking the location at which $H\big(\boldsymbol{C}(s,k,e)\big)$ maximizes. Then the interval (s, e) is split into two subintervals (s, b) and (b, e) and we conduct the above procedure on (s, b) and (b, e) separately. This algorithm is stopped until no subinterval can detect a change point. At the beginning of BS, we may choose $(s, e) = (1, n-1)$. In high dimensions, several papers used BS for solving (43) [15,16,43,63].

Note that along with the iteration of the BS algorithm, a search interval that contains more than one change point may be used. Hence, it is sub-optimal under some unfavourable settings where changes between different segments exhibit a non-monotonic pattern. In that case, BS based method may lose its power. To enhance the performance of BS, [26] introduced the wild binary segmentation algorithm. The main idea of WBS is that suppose (s, e) is the current candidate search interval. Instead of using the whole samples (s, e) for calculating a single test statistic $T_{(s,e)}$, WBS generates many random subintervals $\{(s_m, e_m) \subset (s, e)\}_{m=1}^M$, to allow at least one of them contains only one single change point (with high probability). Then, WBS mainly proceeds as follows:

(1) Compute the CUSUM statistic $\boldsymbol{C}(s_m, k, e_m)$ on each subinterval (s_m, e_m), then maximize each CUSUM by calculating $T_{(s_m,e_m)} = \max_{k\in(s_m,e_m)} H\big(\boldsymbol{C}(s_m,k,e_m)\big)$, for $m = 1, \ldots, M$.
(2) Find (m^*, b^*) such that $m^* = \operatorname{argmax}_{1\leq m\leq M} T_{(s_m,e_m)}$ and $b^* = \operatorname{argmax}_{k\in(s_{m^*},e_{m^*})} H\big(\boldsymbol{C}(s_{m^*},k,e_{m^*})\big)$.
(3) Use the critical value $c_{(s_{m^*},e_{m^*})}$ and $T_{(s_{m^*},e_{m^*})}$ to decide whether we can identify b^* as a new change point. If b^* is considered to be significant, then split the interval (s, e) into two subintervals (s, b^*) and (b^*, e).
(4) Conduct the above procedure on (s, b^*) and (b^*, e) separately until some stopping rule is reached.

It is proved [26] that WBS enjoys better performance than BS both in theory and application. In high dimensions, several papers used WBS for multiple change point detection. See [10,58,68] for mean vectors as well as [60] for the covariance matrix.

It is worth mentioning that both BS and WBS are heuristic algorithms since the next iteration depends on the results from the previous step. Hence, it is very difficult to control the overall significance level for the whole procedure. Different from BS or WBS, the moving sum (MOSUM) based procedure is another popular way that directly constructs a test statistic for multiple change points. In particular, let G_n denote some user pre-specified bandwidth. For each candidate search location $G_n \leq k \leq n - G_n$, using the samples $\{\boldsymbol{X}_{k-G_n}, \ldots, \boldsymbol{X}_{k+G_n}\}$, calculate the CUSUM statistics $\boldsymbol{C}(k-G_n, k, k+G_n)$ as in (44). Then, we can construct a MOSUM based test statistic as:

$$T_{G_n} = \max_{G_n\leq k\leq n-G_n} H\big(\boldsymbol{C}(k-G_n,k,k+G_n)\big). \tag{45}$$

By construction, we know that the main idea of MOSUM is to search subintervals with a length of $2G_n + 1$ in a moving procedure. Hence, it can detect the existence of multiple change points if one chooses G_n properly. Moreover, in the low dimensional setting, the limiting null distribution of T_{G_n} can be typically obtained [24]. This results in a test of (43) under any prespecified significance level. In high dimensions, deriving the limiting null distribution for T_{G_n} is a challenging task, and not much research exists. In a recent work, by choosing $H(\cdot) = \|\cdot\|_\infty$, [13] studied the ℓ_∞-norm based MOSUM test statistic for testing multiple change points with sparse alternatives. They used the Gaussian approximation technique for approximating the limiting null distribution of the corresponding test statistic. Along this research direction, it is still an open question to investigate the theoretical properties of T_{G_n} for other types of aggregation methods such as $H(\cdot) = \|\cdot\|_2$ or $H(\cdot) = \|\cdot\|_{(s_0,p)}$.

In addition to the above mentioned techniques, we note that there are some recent extensions of methods for multiple change points detection, especially in the low dimensional setting. These are the multiple scale MOSUM [46] procedure, the narrowest-over-threshold method [5], the tail-greedy unbalanced Haar (TGUH) transform based technique [27], the seeded binary segmentation based method [37], among others. We believe that it is meaningful to combine these methods with the high dimension based techniques for solving multiple change point detection arising from modern high-throughput data sequences.

8. Concluding remarks

Focused on mean vectors, this paper provides a comprehensive review on recent developments of high dimensional change point inference. This includes the motivations and challenges for high dimensional change point analysis, the methodologies designed for known alternative patterns, or those in a data-driven fashion. Using the concept of high dimensional efficiency, we compare theoretical properties of different methods. We also demonstrate the detection boundary for this problem in terms of the minimax optimality separation rates. In addition, we list several recent extensions from high dimensional mean vectors to more complex problems such as change point inference for variances, covariance matrices, U-statistic based parameters, as well as non-parametric change point tests for distributional changes. Moreover, we provide some commonly adopted techniques for high dimensional multiple change point inference and some possible research generalizations.

In spite of recent rapid developments on this topic, there are many interesting and open research directions. One problem is to consider change point inference for more complex statistical models with $\{(Y_i, \boldsymbol{X}_i)\}, i \in \{1, \ldots, n\}$ being observed, where $Y_i \in \mathbb{R}^1$ is the dependent variable and $\boldsymbol{X}_i \in \mathbb{R}^d$ is the independent vector. The goal is to test whether the underlying data generating mechanism between Y and $\boldsymbol{X}$ has a change point. Typically, in high dimensions, we can consider the following conditional linear regression model

$$\theta(Y_i|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(1)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(2)} \mathbf{1}\{i \geq k_0\}, i \in \{1, \ldots, n\}, \tag{46}$$

where $\theta(Y_i|\boldsymbol{X}_i)$ is the parameter of Y_i conditional on $\boldsymbol{X}_i$ that is of interest, $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \in \mathbb{R}^d$ are the underlying regression coefficients, and $k_0 \in \{1, \ldots, n-1\}$ is the possible change point location. For example, if we choose $\theta(Y_i|\boldsymbol{X}_i) = \mathbb{E}(Y_i|\boldsymbol{X}_i)$, Model (46) reduces to the classical linear regression model. To test a change point in (46), we can consider the hypothesis:

$$\mathbf{H}_0 : \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)} \quad \text{vs.} \quad \mathbf{H}_1 : \text{There exists } k_0 \in \{1, \ldots, n-1\} \text{ s.t. } \boldsymbol{\beta}^{(1)} \neq \boldsymbol{\beta}^{(2)}. \tag{47}$$

For linear regression models with a fixed dimension d, Problem (47) has been extensively studied. See [20,32] for a summary of classical methods. In high dimensions with $d \gg n$, a few papers exist in the literature and the majority of the existing techniques such as [38,65] mainly focuses on the estimation of k_0. In contrast, there are limited developments for the testing of (47). The challenge comes from two aspects. One difficulty is the fact that, unlike mean vectors, there is no natural testing statistic such as CUSUM as in (3) for high dimensional regression coefficients. Essential modifications are needed in terms of the construction of a testing statistic. Another difficulty is how to derive or approximate the limiting null distribution of the testing statistic once it is constructed. In a recent paper, [42] used the debiased-lasso technique for constructing an ℓ_∞-norm based Wald-type statistic for testing a single change point. They used a novel Gaussian multiplier bootstrap procedure to approximate the limiting null distribution. Hence, along this research direction, we can consider multiple change point inference for high dimensional linear regression models. Moreover, beyond linear models, it is possible to investigate more complicated models such as high dimensional generalized linear models by setting $\theta(Y_i|\boldsymbol{X}_i) = g\big(\mathbb{E}(Y_i|\boldsymbol{X}_i)\big)$, where $g(\cdot)$ is the link function, or consider the quantile regression models by setting $\theta(Y_i|\boldsymbol{X}_i) = \text{Quant}_\tau(Y_i|\boldsymbol{X}_i)$. To our knowledge, change point inference for the latter two cases has not been considered yet in high dimensions.

Another open problem is the issue of robustness. Since the CUSUM statistics in (3) is constructed using the sample means, it is not robust against outliers or data with heavy-tailed distributions. In other words, the CUSUM based method fails to control the size under $\mathbf{H}_0$ or cannot detect a change point under $\mathbf{H}_1$ because of the outliers. Therefore, it is necessary to consider robust change point inference in high dimensions. In a recent paper, [62] proposed a rank based ℓ_∞-norm typed testing statistic to detect sparse changes of mean vectors. Along this research direction, it can be interesting to construct robust tests for dense alternatives or tests that are both robust to outliers and adaptive to the unknown alternative patterns.

CRediT authorship contribution statement

Bin Liu: Writing of paper, Original draft preparation, Reviewing and editing. **Xinsheng Zhang:** Writing of paper, Original draft preparation, Reviewing and editing. **Yufeng Liu:** Writing of paper, Original draft preparation, Reviewing and editing.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China Grant 11971116 (Zhang), 12101132 (Bin Liu), and US National Institute of Health Grant R01GM126550 and National Science Foundation, USA Grants DMS1821231 and DMS2100729 (Yufeng Liu).

References

[1] S. Arlot, A. Celisse, Z. Harchaoui, A kernel multiple change-point algorithm via model selection, J. Mach. Learn. Res. 20 (162) (2019) 1–56.
[2] J.A. Aston, C. Kirch, High dimensional efficiency with applications to change point tests, Electron. J. Stat. 12 (2018) 1901–1947.
[3] A. Aue, S. Hörmann, L. Horváth, M. Reimherr, Break detection in the covariance structure of multivariate time series, Ann. Statist. 37 (6) (2009) 4046–4087.
[4] V. Avanesov, N. Buzun, Change-point detection in high-dimensional covariance structure, Electron. J. Stat. 12 (2018) 3254–3294.
[5] R. Baranowski, Y. Chen, P. Fryzlewicz, Narrowest-over-threshold detection of multiple change points and change-point-like features, J. R. Stat. Soc. Ser. B Stat. Methodol. 81 (2019) 649–672.
[6] M. Bours, A. Steland, Large-sample approximations and change testing for high-dimensional covariance matrices of multivariate linear time series and factor models, Scand. J. Stat. 48 (2) (2021) 610–654.
[7] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 672–684.
[8] T. Cai, W. Liu, X. Luo, A constrained ℓ_1 minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 594–607.
[9] T.T. Cai, Z. Ren, H.H. Zhou, et al., Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation, Electron. J. Stat. 10 (1) (2016) 1–59.
[10] S. Chakraborty, X. Zhang, High-dimensional change-point detection using generalized homogeneity metrics, 2021, arXiv preprint arXiv:2105.08976.
[11] J. Chen, A.K. Gupta, Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance, Springer Science and Business Media, New York, 2012.
[12] S.X. Chen, Y.-L. Qin, A two-sample test for high-dimensional data with applications to gene-set testing, Ann. Statist. 38 (2) (2010) 808–835.
[13] L. Chen, W. Wang, W.B. Wu, Inference of breakpoints in high-dimensional time series, J. Amer. Statist. Assoc. (just-accepted) (2021) 1–33, (in press).
[14] H. Chen, N. Zhang, Graph-based change-point detection, Ann. Statist. 43 (1) (2015) 139–176.
[15] H. Cho, Change-point detection in panel data via double CUSUM statistic, Electron. J. Stat. 10 (2) (2016) 2000–2038.
[16] H. Cho, P. Fryzlewicz, Multiple change point detection for high dimensional time series via sparsified binary segmentation, J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (2) (2015) 475–507.
[17] H. Cho, C. Kirch, Data segmentation algorithms: Univariate mean change and beyond, 2020, arXiv preprint arXiv:2012.12814.
[18] L. Chu, H. Chen, Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data, Ann. Statist. 47 (1) (2019) 382–414.
[19] C.-S.J. Chu, K. Hornik, C.-M. Kuan, Mosum tests for parameter constancy, Biometrika 82 (3) (1995) 603–617.
[20] D.R. Cox, D.V. Hinkley, Theoretical Statistics, CRC Press, New York, 1979.
[21] M. Csörgő, L. Horváth, Nonparametric tests for the changepoint problem, J. Statist. Plann. Inference 17 (none) (1987) 1–9.
[22] M. Csörgő, L. Horváth, Limit Theorems in Change-Point Analysis, John Wiley and Sons Inc, New York, 1997.
[23] H. Dette, G. Pan, Q. Yang, Estimating a change point in a sequence of very high-dimensional covariance matrices, J. Amer. Statist. Assoc. (2020) 1–11.
[24] B. Eichinger, C. Kirch, A MOSUM procedure for the estimation of multiple random change points, Bernoulli 24 (1) (2018) 526–564.
[25] F. Enikeeva, Z. Harchaoui, High-dimensional change-point detection under sparse alternatives, Ann. Statist. 47 (2019) 2051–2079.
[26] P. Fryzlewicz, Wild binary segmentation for multiple change-point detection, Ann. Statist. 42 (6) (2014) 2243–2281.
[27] P. Fryzlewicz, Tail-greedy bottom-up data decompositions and fast multiple change-point detection, Ann. Statist. 46 (6B) (2018) 3390–3421.
[28] E. Gombay, L. Horváth, Asymptotic distributions of maximum likelihood tests for change in the mean, Biometrika 77 (2) (1990) 411–414.
[29] D.M. Hawkins, Testing a sequence of observations for a shift in location, J. Amer. Statist. Assoc. 72 (357) (1977) 180–186.
[30] D.V. Hinkley, Inference about the change-point in a sequence of random variables, Biometrika 57 (1970) 1–17.
[31] L. Horváth, M. Hušková, Change-point detection in panel data, J. Time Series Anal. 33 (4) (2012) 631–648.
[32] L. Horváth, G. Rice, Extensions of some classical methods in change point analysis, Test 23 (2) (2014) 219–255.
[33] M. Hušková, S.G. Meintanis, Change point analysis based on empirical characteristic functions, Metrika 63 (2) (2006) 145–168.
[34] M. Jirak, Change-point analysis in increasing dimension, J. Multivariate Anal. 111 (2012) 136–159.
[35] M. Jirak, Uniform change point tests in high dimension, Ann. Statist. 43 (6) (2015) 2451–2483.
[36] J. Kiefer, K-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests, Ann. Math. Stat. 30 (2) (1959) 420–447.
[37] S. Kovács, H. Li, P. Bühlmann, A. Munk, Seeded binary segmentation: A general methodology for fast and optimal change point detection, 2020, arXiv preprint arXiv:2002.06633.
[38] S. Lee, M.H. Seo, Y. Shin, The lasso for high dimensional regression with a possible change point, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (1) (2016) 193–210.
[39] C. Lévy-Leduc, F. Roueff, et al., Detection and localization of change-points in high-dimensional network traffic data, Ann. Appl. Stat. 3 (2) (2009) 637–662.
[40] Y.-N. Li, D. Li, P. Fryzlewicz, Detection of multiple structural breaks in large covariance matrices, 2021, Preprint on webpage at stats.lse.ac.uk/fryzlewicz/wbscov/wbscov.pdf.
[41] H. Liu, C. Gao, R.J. Samworth, Minimax rates in sparse, high-dimensional change point detection, Ann. Statist. 49 (2) (2021) 1081–1112.
[42] B. Liu, X. Zhang, Y. Liu, Simultaneous Change Point Detection and Identification for High Dimensional Linear Models, Technical Report, 2020.
[43] B. Liu, C. Zhou, X.-S. Zhang, Y. Liu, A unified data-adaptive framework for high dimensional change point detection, J. R. Stat. Soc. Ser. B Stat. Methodol. 82 (4) (2020) 933–963.
[44] D.S. Matteson, N.A. James, A nonparametric approach for multiple change point analysis of multivariate data, J. Amer. Statist. Assoc. 109 (505) (2014) 334–345.
[45] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (3) (2006) 1436–1462.
[46] M. Messer, S. Albert, G. Schneider, The multiple filter test for change point detection in time series, Metrika 81 (6) (2018) 589–607.
[47] R. Miller, D. Siegmund, Maximally selected chi square statistics, Biometrics (1982) 1011–1016.
[48] G.V. Moustakides, Optimal stopping times for detecting changes in distributions, Ann. Statist. 14 (4) (1986) 1379–1387.
[49] Y.S. Niu, N. Hao, H. Zhang, Multiple change-point detection: a selective overview, Statist. Sci. (2016) 611–623.
[50] E.S. Page, Continuous inspection schemes, Biometrika 41 (1/2) (1954) 100–115.
[51] E. Page, Control charts with warning lines, Biometrika 42 (1–2) (1955) 243–257.
[52] E.S. Page, A test for a change in a parameter occurring at an unknown point, Biometrika (3–4) (1955) 3–4.
[53] E. Pilliat, A. Carpentier, N. Verzelen, Optimal multiple change-point detection for high-dimensional data, 2020, arXiv preprint arXiv:2011.07818.
[54] X. Shao, X. Zhang, Testing for change points in time series, J. Amer. Statist. Assoc. 105 (491) (2010) 1228–1240.
[55] A. Steland, Testing and estimating change-points in the covariance matrix of a high-dimensional time series, J. Multivariate Anal. 177 (2020) 104582.
[56] C. Truong, L. Oudre, N. Vayatis, Selective review of offline change point detection methods, Signal Process. 167 (2020) 1–20.

[57] L.Y. Vostrikova, Detecting disorder in multidimensional random process, Sov. Math. Dokl. 24 (1981) 55–59.
[58] T. Wang, R.J. Samworth, High-dimensional changepoint estimation via sparse projection, J. R. Stat. Soc. Ser. B Stat. Methodol. 80 (2018) 57–83.
[59] R. Wang, S. Volgushev, X. Shao, Inference for change points in high dimensional data, 2019, arXiv preprint arXiv:1905.08446.
[60] D. Wang, Y. Yu, A. Rinaldo, Optimal covariance change point localization in high dimensions, Bernoulli 27 (1) (2021) 554–575.
[61] W.B. Wu, M. Pourahmadi, Banding sample autocovariance matrices of stationary processes, Statist. Sinica (2009) 1755–1768.
[62] M. Yu, X. Chen, A robust bootstrap change point test for high-dimensional location parameter, 2019, arXiv preprint arXiv:1904.03372.
[63] M. Yu, X. Chen, Finite sample change point inference and identification for high-dimensional mean vectors, J. R. Stat. Soc. Ser. B Stat. Methodol. 83 (2) (2021) 247–270.
[64] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (1) (2007) 19–35.
[65] B. Zhang, J. Geng, L. Lai, Change-point estimation in high dimensional linear regression models via sparse group lasso, in: 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015, pp. 815–821.
[66] T. Zhang, L. Lavitas, Unsupervised self-normalized change-point testing for time series, J. Amer. Statist. Assoc. 113 (522) (2018) 637–648.
[67] N.R. Zhang, D.O. Siegmund, H. Ji, J.Z. Li, Detecting simultaneous changepoints in multiple sequences, Biometrika 97 (3) (2010) 631–645.
[68] Y. Zhang, R. Wang, X. Shao, Adaptive inference for change points in high-dimensional data, J. Amer. Statist. Assoc. (2021) 1–12, (in press).
[69] P.-S. Zhong, J. Li, P. Kokoszka, Multivariate analysis of variance and change points estimation for high-dimensional longitudinal data, Scand. J. Stat. 48 (2021) 375–405.