Robust Change Point Detection for High-dimensional Linear Models with Tolerance for Heavy Tails

Zhi Yang[a], Liwen Zhang[a], Bin Liu[b,*]

[a]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China
[b]School of Management, Fudan University, Shanghai, 200433, China

Abstract

This paper focuses on localizing change points in high-dimensional linear regression models with piecewise constant regression coefficients. The existing literature primarily concentrates on change point estimation under strict Gaussian or sub-Gaussian noise assumptions. However, in practical situations, the noise distribution is often uncertain or exhibits heavy-tailed properties. To address these challenges, we propose two algorithms: one based on dynamic programming and the other on binary segmentation techniques, for detecting change points. These algorithms offer flexibility in accommodating growing sample sizes and data dimensions. Without requiring the existence of the first or second moments of the noise distribution, we establish the consistency of the estimated change point numbers and locations for both algorithms. The dynamic programming algorithm (DPA) achieves enhanced localization accuracy, while the binary segmentation algorithm (BSA) offers computational efficiency. Finally, extensive simulation studies and real data applications further demonstrate the competitive performance of our proposed methods.

Keywords: High dimensions, Change point detection, Linear regression, Heavy tail, Dynamic programming, Binary segmentation

1. Introduction

In the last two decades, the big data has gained prominence across diverse scientific disciplines, including biology, neuroscience, genomics, finance, and social sciences, owing to advancements in data collection and storage capacity. However, one common challenge associated with the big data, as emphasized in [19], is the issue of heterogeneity. Heterogeneity frequently manifests as non-stationarity in sequentially acquired data, where the data-generating process undergoes structural breaks over time. To tackle these problems, various change point analysis methods have been developed for detecting structural breaks.

Change point detection and localization are classical problems in which we collect a series of data points and aim to determine whether and at what point changes have occurred in the underlying generative model. The literature on change point detection has significantly expanded since the pioneering work of [24], with applications spanning multiple disciplines, including genomics [21], social sciences [27], and even recent studies on the COVID-19 pandemic [9]. Change-point detection is mostly investigated within the context of mean models, encompassing scenarios in both low and high dimensions, as exemplified by [3, 5, 13, 30, 35]. [20] further proposed a tail adaptive approach to locate the change-point in a multivariate data with heavy-tailed distribution. More recently, [21] introduced an adjusted ℓ_q-aggregation strategy based data-adaptive test.

In this paper, we mainly consider change point detection and localization for high-dimensional linear regression models. Specifically, we consider a scenario where we have n independent, but ordered, realizations $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$. We formally introduce our model settings next. Let the data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ satisfy the model

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_i^0 + \varepsilon_i, \tag{1.1}$$

where $\boldsymbol{\beta}_i^0 = (\beta_{i1}^0, \ldots, \beta_{ip}^0)^\top$ is the unknown regression coefficient vector, $\boldsymbol{X}_i$ are independently drawn from a distribution $\mathbb{P}_X$ with $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{X}_i^\top \boldsymbol{X}_i] = \boldsymbol{\Sigma}$, and ε_i are i.i.d. random noises independent of $\boldsymbol{X}_i$. In addition, there exists $M \geq 0$ change points $\{\eta_m\}_{m=1}^M \subset \{1, \ldots, n-1\}$ such that

*Corresponding author.
Email address: liubin0145@gmail.com (Bin Liu)

$$\boldsymbol{\beta}^0_{\eta_m} \neq \boldsymbol{\beta}^0_{\eta_m+1}, \quad m = 1, \ldots, M, \text{ and } \quad \boldsymbol{\beta}_{\eta_m+1} = \cdots = \boldsymbol{\beta}_{\eta_{m+1}}, \quad m = 0, \ldots, M,$$

where by convention we define $\eta_0 = 0$ and $\eta_{M+1} = n$.

Extensive investigations have been conducted for change point analysis in low-dimensional ($p \ll n$) linear regression models. Methods include like partial sums of regression residuals [1, 8], the maximum likelihood ratio approach [7], and the empirical likelihood ratio method [22]. Additionlly, change point detection in quantile regression models has been explored with significant contributions from[38, 39].

It's important to note that, unlike the low-dimensional case, high-dimensional ($p \gg n$) change point analysis poses a more challenging problem due to the curse of dimensionality. In the high-dimensional single change-point setting, [16] extended the Lasso method [28] and demonstrated consistent estimation of both the change point and the regression parameters. Building upon this work, [15] further extended the results to the high-dimensional quantile regression model. In terms of high-dimensional linear models with multiple structural breaks, [37] explored the use of the Sparse Group Lasso (SGL) algorithm Additionally, [17] introduced fast algorithms that leverage dynamic programming and binary search techniques. Expanding on this, [34] introduced a projection-based algorithm incorporating wild binary segmentation (WBS, [5]) for the estimation of multiple change points. Furthermore, [26] made significant progress by offering a more comprehensive proof, utilizing dynamic programming, and introducing an additional refinement technique reminiscent of group lasso. In a distinct context, [36] extended the method by [17] to high-dimensional generalized linear models for multiple change points estimation.

It is essential to acknowledge that the aforementioned works predominantly rely on the assumption that noises follow Gaussian or sub-Gaussian distributions. However, this assumption may not be suitable for scenarios where the data is contaminated or corrupted by outliers. For instance, [16] assumed Gaussian noise, while [11, 26, 34] considered sub-Gaussian distributed noise. Nevertheless, when the noise fails to meet the finite variance assumption, most existing theories based on least squares (LS) losses become inapplicable. Many real-world situations exhibit heavy-tailed behavior, such as Insurance Payouts and Financial Returns. However, the existing literature on change-point detection in high-dimensional linear regression models does not address the challenging issue of the unknown or heavy-tailed noises, which is the primary focus of this paper.

When dealing with heavy-tailed noises or violations of the underlying distribution assumptions, an alternative solution is to utilize robust methods such as median regression, as proposed in [15]. However, it is important to consider that in the context of statistical inference for homogeneous linear models, the asymptotic relative efficiency of median regression-based estimators compared to least squared-based estimators is approximately 64% in both low and high dimensions, as discussed in [6, 42]. Inference based on quantile regression (QR), thoughrobust, can have arbitrarily small relative efficiency compared to the least-square approach. This phenomenon has also been observed in [18] for high dimensional change point testing. Hence, an interesting question is how to propose efficient multiple change point detection algorithms for high dimensional linear regression models that is robust to the heavy-tailed distribution while not losing detection efficiency for the light-tailed errors.

To safeguard quantile regression against potential efficiency loss, another method known as composite quantile regression (CQR, [6, 41, 42]) enhances asymptotic relative efficiency to 70% in the worst case by combining information over different quantiles via a mix of quantile loss functions. The relative efficiency lower bound is further improved to 86.4% in [10]. When the noise is actually Gaussian distributed, the relative efficiency is approximately 95%. Importantly, this method retains its validity even in situations where the first two moments of the noise distribution are not well-defined. Inspired by the nice statistical properties of the CQR estimation, in this paper, we explore efficient multiple change point detection in the context of high dimensional linear regression models. The goal is to consider a flexible and versatile framework for analyzing high-dimensional data with multiple structural breaks when the noise distribution is unknown and potentially heavy-tailed.

The main contributions of our study are summarized as follows. (**i**) We offer consistent change point estimators for Model (1.1) in scenarios involving unknown or heavy-tailed noises, allowing for model parameters to vary with the sample size n. These parameters encompass the dimensionality of the data, the sparsity of coefficient vectors on an entry-wise basis, the number of change points, the minimum distance between two consecutive change points, and the smallest difference between two successive distinct regression coefficients. (**ii**) We propose two algorithms for estimating the number and location of change points, leveraging the Lasso estimator of the CQR coefficients. The core concept is as follows: when considering any interval, if the sum of the loss functions on the two sub-intervals after partitioning, in addition to a threshold parameter, is smaller than the loss function before partitioning, the algorithms will identify a partition and detect a change point. These two algorithms are variants of dynamic programming and binary segmentation, exhibiting computational complexities of $O(n^2\ \text{CQRlasso}(n, K))$ and $O(n \log n\ \text{CQRlasso}(n, K))$, respectively. Here, $O(\text{CQRlasso}(n, K))$ represents the cost associated with computing the CQR Lasso estimator for a given n and the number of quantiles K. (**iii**) We examined several theoretical properties of the change point estimators computed by the two algorithms without requirement on the moments

of the noise distribution. In the proof provided detailed proofs in Appendix B, we tackled the most challenging part, obtaining a consistent lasso estimator within sub-intervals containing multiple change points, even when dealing with non-differentiable quantile loss. Without requiring the existence of the first or second moment of the noise distribution, both the two algorithms yield consistent estimators for the number and locations of true change points. DPA achieves a more accurate localization error of $O_p(d_0^2 \log(n \vee p)/n)$ for sparsity d_0 in scenarios with gradual change points. Importantly, the theoretical results presented in scenarios involving unknown or heavy-tailed noises, aligning with those obtained in the sub-Gaussian setting [26]. Additionally, BSA with a stronger signal-to-noise ratio condition, results in a localization error of $O_p(d_0^{3/2}\sqrt{\log(n \vee p)/n})$. (**iv**) Finally, to corroborate our theoretical findings and demonstrate the practicality of our procedures, we present analyses of simulated data and real-world datasets. To visually showcase the competitiveness of our algorithms, Figure 1 presents a comparison between our algorithms and the state-of-the-art algorithms under various noise distributions for Model (1.1) with three change points. It is evident that our method exhibits robustness, even when noise exhibits heavy tails or lacks first and second moments.

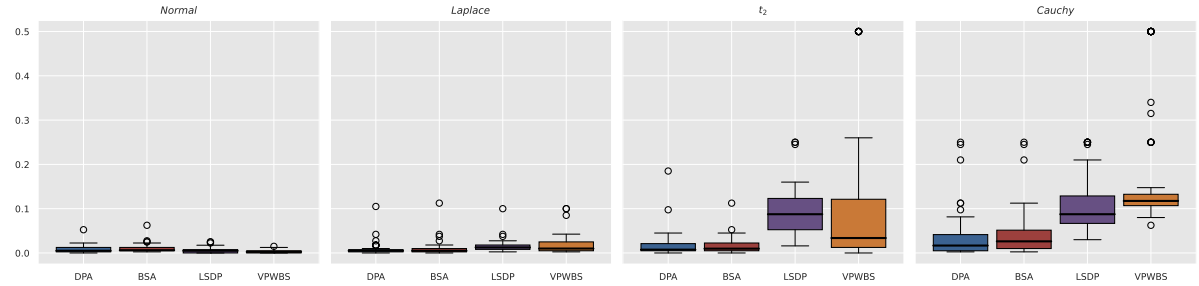

Fig. 1: Boxplots of the scaled Hausdorff distance (5.1) of different noise distributions for multiple change points estimation based on 100 replications. Our introduced methods, DPA and BSA, are contrasted with the state-of-the-art algorithms for change-point detection in high-dimensional linear regression settings, LSDP ([26]) and VPWBS ([34]). Comprehensive configuration details can be located in Section 5.

To provide a comprehensive understanding of our work, it is important to highlight its relationship with several related papers. In comparison to the study by [26] and [34], we tackle the challenge of estimating change points in the presence of unknown or heavy-tailed noise without requiring the existence of the first or second moment. As for the data without change points, [6] and [41] can effectively obtain the consistent estimator of the regression coefficient in unknown or heavy-tailed scenarios. However, their results cannot be directly applied to situations where change points exist in the data. Moreover, the proof of the lasso properties within sub-intervals is nontrivial, as the pseudo-true parameters are not explicitly expressed in CQR, which really differs from [17].

The structure of the remaining sections is as follows. In Section 2, we establish the model discussed in this paper and provide preliminary knowledge along with symbol definitions. In Section 3, we describe the construction of our algorithms and provide detailed execution steps. Section 4 contains the useful lemma and theoretical results related to different algorithms. To assess the performance of our methods, we offer extensive numerical results and a real data application in Sections 5 and 6. A summary of the article is presented in Section 7. For interested readers, detailed proofs of the main theorems and important lemmas can be found in the Appendix.

Throughout this paper, for $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, we define its ℓ_p norm as $\|\boldsymbol{v}\|_p = (\sum_{j=1}^d |v_j|^p)^{1/p}$ for $1 \le p \le \infty$. For $p = \infty$, define $\|\boldsymbol{v}\|_\infty = \max_{1\le j\le d} |v_j|$. For any set $\mathcal{S}$, denote its cardinality by $|\mathcal{S}|$. For two positive sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n = O(b_n)$ if there exists $C > 0$ such that $\limsup_{n\to\infty} a_n/b_n < C$ and write $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We write $a_n \succeq b_n$ if $\liminf_{n\to\infty} a_n/b_n = \infty$. Let $\{x_n\}_{n=1}^\infty$ be a sequence of random variables. We write $x_n = O_p(b_n)$ if $x_n/b_n = O_p(1)$ and write $x_n = o_p(b_n)$ if $x_n/b_n = o_p(1)$, where $O_p(1)$ and $o_p(1)$ follow the standard probability notation of big O (stochastic boundedness) and small o (convergence to zero in probability). For an $n \times p$ matrix $\mathbf{X}$, we define $\boldsymbol{X}_i$ as its i-th row and X_{ij} as the j-th value in the i-th row.

2. Preliminary

Driven by the motivations outlined in Section 1, our paper focuses on investigating change point detection in the context of high-dimensional linear regression. Given data sampled from Model (1.1), our main task is to develop computationally-efficient algorithms that can consistently estimate both the unknown number M of change points and the change points location, at which the coefficient vector change. More precisely, we aim to construct an estimator of $\{\eta_m\}_{m=1}^M$ of the form $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}} \subset \{1, \ldots, n-1\}$ with $\widehat{\eta}_1 < \widehat{\eta}_2 < \ldots < \widehat{\eta}_{\widehat{M}}$ satisfying the following notion of consistent localization. Any change point estimators $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}$ are deemed to be consistent if,

with probability tending to 1 as $n \to \infty$,

$$\widehat{M} = M \quad \text{and} \quad \max_{m=1,\ldots,M} \left|\widehat{\eta}_m - \eta_m\right| \leq \epsilon,$$

where ϵ represents the **localization error** of the estimator, meeting $\epsilon/n \to 0$, and ϵ/n is termed the **localization rate**. In cases involving unknown or heavy-tailed noise, we have developed algorithms introduced in Section 3 to obtain consistent change-point estimators under relatively weak conditions. The utilization of coefficient estimators within sub-intervals is a common theme in these algorithms. Subsequently, we will discuss the selection of loss function and the construction of lasso estimators.

Moving forward, we explain why our algorithms consider using CQR, initially introduced in [42], instead of LS and QR to analyze the change-point problem in Model (1.1). Before proceeding, let's clarify the notation we will be using. Assuming that the random noise ε follows cumulative distribution function $F_\varepsilon(\cdot)$ and probability density function $f_\varepsilon(\cdot)$. Given an ordered sequence of quantile levels $\tau_1 < \tau_2 < \ldots < \tau_K \in (0,1)$, let $b_k^0 := F^{-1}(\tau_k)$, where $F^{-1}(\tau_k) = \inf\{t : F(t) \geq \tau_k\}$ represents the τ_k-th quantile for $1 \leq k \leq K$. Then, as for Model (1.1), it becomes clear that the τ_k-th conditional quantile of $Y_i|\boldsymbol{X}_i$ is

$$Q_{\tau_k}(Y_i|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}_i^0 + b_k^0, \quad \text{for } k = 1, \ldots, K.$$

Let's begin by discussing $K = 1$, i.e., the QR estimator. While the QR estimator is expected to be more efficient than the LS estimator under certain noises (e.g. Laplace distribution), it only considers one quantile at a time and may not fully utilize the distributional information for consistently efficient estimation. As discussed in Section 1, if the error density at a specific quantile approaches zero, the asymptotic variance of the corresponding QR estimator becomes infinity, resulting in an estimator with extremely low efficiency. Such situations can cause large localization error of $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}$ when change points occur under specific extreme quantile conditions.

To mitigate potential efficiency loss in QR, we employ the CQR technique to obtain a lasso estimator that integrates information from multiple quantiles. The underlying concept is intuitive: utilizing more quantiles provides access to a more comprehensive distributional perspective, potentially resulting in more efficient estimation. Hence, if there are multiple change points in Model (1.1), the various conditional quantiles of $Y_i|\boldsymbol{X}_i$ also exhibit multiple breaks. This observation allows us to use CQR models to capture the structural breaks of $\boldsymbol{\beta}^0$.

Next, we will define the loss function and the lasso estimator related to CQR within a generic sub-interval $I = (s, e] \subseteq (0, n]$. In the remainder of this paper, when there is no ambiguity, we use the following notation. Given two natural numbers $s < e$, we denote $(s, e] := \{i \in \mathbb{N} \mid s < i \leq e\}$. The composite quantile loss function, defined in equation (2.1), comprises the sum of K check loss functions, each corresponding to K different quantile values of τ. Each term of the sum involves the same vector $\boldsymbol{\beta} \in \mathbb{R}^p$, but a different scalar $b \in \mathbb{R}$. To facilitate further discussion, we introduce the following notation: for each pair $(\boldsymbol{\beta}, \boldsymbol{b}_K)$, where $\boldsymbol{b}_K = (b_1, b_2, \ldots, b_K)^\top \in \mathbb{R}^K$, we define $\underset{\sim}{\boldsymbol{\beta}} := (\boldsymbol{\beta}^\top, \boldsymbol{b}_K^\top)^\top \in \mathbb{R}^{p+K}$. For brevity, we may omit the subscript K as needed in the subsequent discussion. Now, we formally introduce composite quantile loss function for sub-interval I as follows:

$$\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}) := \sum_{i \in I} \frac{1}{K} \sum_{k=1}^{K} \rho_{\tau_k}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} - b_k\right), \tag{2.1}$$

where ρ_{τ_k} represents the check loss function defined by [12] as $\rho_{\tau_k}(t) = \tau_k t_+ + (1 - \tau_k) t_-$ with $t_+ = \max\{t, 0\}$ and $t_- = \max\{-t, 0\}$. In order to estimate the model coefficients across different sub-intervals, we introduce the subscript notation with interval indices I, denoted as $\underset{\sim}{\boldsymbol{\beta}}_I = (\boldsymbol{\beta}_I^\top, \boldsymbol{b}_I^\top)^\top = (\boldsymbol{\beta}_I^\top, b_{I,1}, \ldots, b_{I,K})^\top$. Specifically, given a dataset $\{(\boldsymbol{X}_i, Y_i)\}_{i \in I}$, the Lasso-based CQR method (see e.g. [6, 41]) solves the following problem with an ℓ_1 penalty [28] term:

$$\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I = \underset{\underset{\sim}{\boldsymbol{\beta}} \in \mathbb{R}^{p+K}}{\operatorname{argmin}} \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}) + \lambda \sqrt{\max\{|I|, \log(n \vee p)\}}\, \|\boldsymbol{\beta}\|_1, \tag{2.2}$$

where $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I = (\widehat{\boldsymbol{\beta}}_I^\top, \widehat{\boldsymbol{b}}_I^\top)^\top \in \mathbb{R}^{p+K}$ is the estimated coefficient for sub-interval I, $\lambda \geq 0$ is the non-negative regularization parameter. Note that the choice of $\{\tau_k\}_{k=1}^K$ is user specified. A typical choice is to take equally spaced τ_k's: $\tau_k = k/(K+1), 1 \leq k \leq K$.

Remark 1. Note that we scale the penalty term by $\max\{|I|, \log(n \vee p)\}$ in (2.2). This adjustment is made to satisfy specific large deviation inequalities required for consistency, as utilized in [17, 26]. Incorporating this penalty term enables our proposed method to effectively estimate change points and CQR coefficients with high-dimensional data. Additional discussion regarding the quantity $\max\{|I|, \log(n \vee p)\}$ is available in the appendix of this paper.

3. Methodology

3.1. Dynamic Programing Algorithm

We present a search approach based on the dynamic programming approach, which is highly suitable for tackling our change-point problem. To attain the objective of obtaining consistent change-point estimators, we employ a DPA with a composite quantile loss function, which we summarize next. Let $\mathcal{I}$ be an integer interval partition of $\{1, \ldots, n\}$ into $M_{\mathcal{I}}$ intervals, i.e.

$$\mathcal{I} = \left\{(0, i_1 - 1], (i_1 - 1, i_2 - 1], \ldots, (i_{M_{\mathcal{I}}-1} - 1, i_{M_{\mathcal{I}}} - 1]\right\},$$

for some integers $1 < i_1 < \cdots < i_{M_{\mathcal{I}}} = n + 1$, where $M_{\mathcal{I}} \geq 0$. For a positive tuning parameter $\gamma > 0$, let

$$\widehat{\mathcal{I}} \in \arg\min_{\mathcal{I}} \Big\{ \sum_{I \in \mathcal{I}} \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) + \gamma |\mathcal{I}| \Big\}, \tag{3.1}$$

where $\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)$ is the composite quantile loss function defined by (2.1) with $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I$ solved by (2.2), $|\mathcal{I}|$ is the cardinality of $\mathcal{I}$, and the minimization is taken over all possible interval partitions of $\{1, \ldots, n\}$.

Algorithm 1 Dynamic Programming Algorithm. DPA$\left(\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n, \lambda, \gamma\right)$

Input: Data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$, tuning parameters $\lambda > 0, \gamma > 0$.
  Initialize $(C, s, e, \text{FLAG}) \leftarrow (\emptyset, 0, 2, 0)$
  while $s < n - 1$ **do**
    $s \leftarrow s + 1$
    while $e < n$ and FLAG $= 0$ **do**
      $e \leftarrow e + 1$ ▷ $\mathcal{D}(\cdot, \cdot)$ is defined in (3.2), λ and γ are involved thereof
      if $\min_{t=s+1,\ldots,e-1} \{\mathcal{D}(s, t) + \mathcal{D}(t, e) \leq \mathcal{D}(s, e)\}$ **then**
        $s \leftarrow \min \arg\min_{t=s+1,\ldots,e-1} \{\mathcal{D}(s, t) + \mathcal{D}(t, e) \leq \mathcal{D}(s, e)\}$
        $C \leftarrow C \cup \{s\}$
        FLAG $\leftarrow 1$
      end if
    end while
  end while
Output: The set of estimated change points C.

The optimization problem (3.1) is known as the *minimal partition problem* and can be solved using dynamic programming (e.g. [4]). Firstly, we estimate the high-dimensional sub-interval coefficient $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I$ adopting the Lasso procedure in (2.2), where the tuning parameter λ is used to obtain sparse CQR estimates. Secondly, with the estimated coefficient vector , we summon the minimal partitioning setup in (3.1) to obtain change point estimators, where the tuning parameter γ is deployed to penalize over-partitioning, as discussed in Section 4.2. More discussions about tuning parameters will be provided later, encompassing both theoretical assurances and practical guidance. The change-point estimator $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}$ resulting from the solution to (3.1) is obtained by taking all the left endpoints of the intervals $I \in \widehat{\mathcal{I}}$, except for 1. We design Algorithm 1 for the purpose of executing DPA.

Algorithms based on dynamic programming are frequently used in the literature on change point detection. [33] described a novel dynamic programming approach ohto localize changes in the high-dimensional autoregressive processes. Under high-dimensional linear regression with sub-Gaussian noises, [17, 26] proposed a variant of the dynamic programming approach, but the latter showed better localization error rate. [36] also extended the dynamic programming method by [17] to high-dimensional generalized linear models for multiple change points estimation. Compared to the literature mentioned earlier, our CQR-based dynamic programming algorithm is primarily tailored for high-dimensional linear models. Additionally, it has the capability to detect changes of $\boldsymbol{\beta}_i^0$ even when dealing with highly heavy-tailed distributions. However, DPA is known to have an overall computational cost of order $O(n^2\text{CQRlasso}(n, K))$, which can become computationally expensive when dealing with a large value of n. In the following section, we introduce an efficient computational method rooted in binary segmentation.

3.2. Binary segmentation algorithm

We introduce a novel technique based on CQR-based binary segmentation algorithm (BSA), as analyzed in [3, 17], and has been shown to be considerably more efficient than DPA. The key concept behind using BSA to address the change point problem is that, for each candidate search interval $(s, e]$, we apply a penalized composite quantile loss function to determine whether a new change point can be incorporated. In particular, for any given integers $0 < s < e < n$, we define

$$\mathcal{D}(s,e) = \begin{cases} \mathcal{L}_n((s,e], \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{(s,e]}) + \gamma, & \text{if } e - s \geq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{3.2}$$

where $\mathcal{L}_n((s,e], \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{(s,e]})$ is defined by (2.1) with $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_{(s,e]}$ solved by (2.2) when $I = (s,e]$, and define

$$v = \underset{t \in \{s\} \cup [s+\zeta, e-\zeta]}{\arg\min} \{\mathcal{D}(s,t) + \mathcal{D}(t,e)\}, \tag{3.3}$$

where ζ is positive tuning parameter. λ and γ are defined in a manner that is similar to that employed in the DPA algorithm. For clarity, we provide pseudocode for the BSA in Algorithm 2, which includes additional details.

Algorithm 2 Binary Segmentation Algorithm. BSA$((s,e], \lambda, \gamma, \zeta)$

Input: Data $\{\boldsymbol{X}_i, Y_i\}_{i=s+1}^{e}$, tuning parameters $\lambda > 0, \gamma > 0, \zeta > 0$.
  Initialize FLAG $\leftarrow 0$
  while $e - s > 2\zeta$ and FLAG $= 0$ **do**
    $v \leftarrow \underset{t \in \{s\} \cup [s+\zeta, e-\zeta]}{\arg\min} \{\mathcal{D}(s,t) + \mathcal{D}(t,e)\}$
    if $v = s$ **then**
      FLAG $\leftarrow 1$
    else
      Add v to the set of estimated change points
      BSA$((s,v], \lambda, \gamma, \zeta)$
      BSA$((v,e], \lambda, \gamma, \zeta)$
    end if
  end while
Output: The set of estimated change points.

The key distinction between BSA and the standard binary segmentation approach lies in BSA's focus on optimizing the sum of the composite quantile loss function exclusively for time points within the interval $(s,e]$ that are at least $\zeta = O(d_0 \log(n \vee p))$ away from the interval endpoint. In contrast to DPA, this approach examines a significantly reduced number of potential change points, leading to enhanced computational efficiency. More precisely, as evidenced in [5, 32], BSA entails a computational cost of only $O(n \log n \, \text{CQRlasso}(n, K))$ operations.

Remark 2. As for the tuning parameter ζ used in Algorithm 2, intuitively, for small subsamples, the estimation error of $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_{(s,e]}$ become difficult to control, which can lead to inaccurate interval segmentation. The sample size demand in CQR is remarkably akin to scale condition of [41]. In change-point linear model setting, similar requirements that are used to ensure the validity of restricted eigenvalue condition can also be found in [11, 17].

4. Main results

In this section, we present theoretical results for our proposed methods. Section 4.1 introduces the basic model assumptions and useful lemma. In Sections 4.2 and 4.3, we discuss the signal conditions for DPA and BSA, as well as the consistency of change point estimation in terms of both number and location. In Section 4.4, we compare the current state-of-the-art for multiple change-point detection in high-dimensional linear model.

4.1. Basic assumptions and useful lemma

To initiate, we introduce certain vital notations for the elucidation of the basic assumptions that will be presented subsequently. These assumptions are indispensable for the derivation of the necessary lemma and the principal theorem. Define the population version of $\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}})$,

$$\mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}) := \mathbb{E}[\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}})] = \mathbb{E}\big[\sum_{i \in I} \frac{1}{K} \sum_{k=1}^{K} \rho_{\tau_k}\big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} - b_k\big)\big].$$

For a generic sub-interval I of $(0,n]$, we define $\underset{\sim}{\boldsymbol{\beta}}_I^* = (\boldsymbol{\beta}_I^{*\top}, \boldsymbol{b}_I^{*\top})^\top \in \mathbb{R}^{p+K}$, where $\boldsymbol{\beta}_I^* \in \mathbb{R}^p$ and $\boldsymbol{b}_I^* = (b_{I,1}^*, \ldots, b_{I,K}^*)^\top \in \mathbb{R}^K$. Here, $\underset{\sim}{\boldsymbol{\beta}}_I^*$ is defined as the solution to the following optimization problem:

$$\underset{\sim}{\boldsymbol{\beta}}_I^* := \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{b} \in \mathbb{R}^K}{\arg\min} \mathbb{E}\big[\sum_{i \in I} \frac{1}{K} \sum_{k=1}^{K} \rho_{\tau_k}\big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} - b_k\big)\big]. \tag{4.1}$$

It is worth noting that $\widetilde{\boldsymbol{\beta}}_I^*$ is unique minimizer of (4.1) and thus can be viewed as the population version of $\widehat{\boldsymbol{\beta}}_I$. If there are no change points in the sub-interval I, then $\widetilde{\boldsymbol{\beta}}_I^*$ can be considered as the true parameters. In the sub-interval I with change points, $\widetilde{\boldsymbol{\beta}}_I^*$ is typically a weighted combination of the parameters before and after the change points, as seen in (C.2). For further discussion of $\widetilde{\boldsymbol{\beta}}_I^*$ under our sub-interval composite quantile loss function, please refer to Appendix C.1. With the above notations, we are ready to introduce our assumptions as follows:

Assumption 1 (Sparsity support)**.** Let β_{ij}^0 be j-th entry of $\boldsymbol{\beta}_i^0$. There exist a subset $\mathcal{S} \subset \{1, \ldots, p\}$ with $|\mathcal{S}| = d_0$ such that

$$\beta_{ij}^0 = 0 \quad i = 1, \ldots, n, \quad j \in \mathcal{S}^c = \{1, \ldots, p\} \backslash \mathcal{S}.$$

In addition, there exists an absolute constant $C_\beta > 0$ such that $\max_{i=1,\ldots,n} \|\boldsymbol{\beta}_i^0\|_\infty \leq C_\beta < \infty$.

Assumption 2 (Design matrix)**.** The design matrix $\mathbf{X}$ has i.i.d rows $\{\boldsymbol{X}_i\}_{i=1}^n$. In addition,
(i) Assume that there are positive constants $\underline{\rho}$ and $\bar{\rho}$ such that $\lambda_{\min}(\boldsymbol{\Sigma}) \geq \underline{\rho} > 0$ and $\lambda_{\max}(\boldsymbol{\Sigma}) \leq \bar{\rho} < \infty$ hold.
(ii) There exists some constant $\mathfrak{M} \geq 1$ such that $\left|X_{ij}\right| \leq \mathfrak{M}$ almost surely for all $1 \leq i \leq n, 1 \leq j \leq p$.

Assumption 3 (Underlying distribution)**.** The random variable ε has a continuously differentiable density function $f_\varepsilon(t)$ whose derivative is denoted by $f_\varepsilon'(t)$. Furthermore, suppose there exist some constants $\bar{f}$, $\underline{f}$ and $\bar{f}'$ such that
(i) $\sup_{t\in\mathbb{R}} f_\varepsilon(t) \leq \bar{f}$; (ii) $\sup_{t\in\mathbb{R}} |f_\varepsilon'(t)| \leq \bar{f}'$;
(iii) $\inf_{i=1,\ldots,n} \inf_{1\leq k\leq K} f_\varepsilon(\boldsymbol{X}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_i^0) + b_k^0) \geq \underline{f}$ for any $\boldsymbol{\beta} \in \mathcal{B}(\boldsymbol{\beta}_i^0, 2d_0C_\beta)$, where $\mathcal{B}(\boldsymbol{\beta}_i^0, r) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}_i^0\|_1 \leq r\}$.

Assumption 1 is a standard assumption for high-dimensional linear regression models, and it has been applied in the works of [11, 17, 26]. Assumption 2 stipulates that the design matrix $\mathbf{X}$ possesses a nondegenerate covariance matrix $\boldsymbol{\Sigma}$ in terms of its eigenvalues. Additionally, it mandates that the covariates of $\mathbf{X}$ exhibit good behavior. Concerning the assumptions regarding the design matrix, we initially delve into Assumption 2(i). This assumption is pivotal for establishing the Lasso property of high-dimensional CQR for any sub-interval. Furthermore, the fulfillment of Assumption 2(ii) also requires that X_{ij} be bounded by a substantial constant $\mathfrak{M} > 0$. This is a commonplace assumption in the literature, as evidenced by works such as [17, 18]. Lastly, Assumption 3 dictates that the noise term ε possesses a bounded density function, accompanied by bounded derivatives. Many other studies on quantile regression, including [15, 41], share a similar assumption to that of Assumption 3. Assumptions 3(i)-(iii) impose moderate constraints on the density function of ε, thereby encompassing a wide spectrum of distributions. This inclusivity extends to heavy-tailed distributions lacking the first two moments, mixtures of distributions, and distributions featuring outliers. It's noteworthy that Assumption 3(iii) additionally mandates that the density function at $\boldsymbol{X}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_i^0) + b_k^0$ remains strictly bounded away from zero.

Remark 3. Our algorithm that utilizes composite quantile loss is designed to cover any interval, regardless of whether it contains change point or not. To obtain the necessary error bounds for Lasso estimation, we must require that $\inf_{i\in I} \inf_{1\leq k\leq K} f_\varepsilon(\boldsymbol{X}_i^\top(\widetilde{\boldsymbol{\beta}}_I^* - \boldsymbol{\beta}_i^0) + b_k^0) \geq \underline{f}$. This is distinct from the classical assumption [41] that $\inf_{1\leq k\leq K} f_\varepsilon(b_k^0) \geq \underline{f}$. It's worth noting that our assumption is relatively mild, as it only requires that the density function be non-degenerate in a neighborhood of $b_{I,k}^*$, which is shown to hold under Assumptions 1 and 2(ii).

Lemma 4 (Lasso property under mixture)**.** For Model (1.1), under Assumptions 1-3, for a generic sub-interval $I = (s, e]$, with $|I| > C_I d_0 \log(n \vee p)$ and $\lambda \geq C_\lambda \sqrt{\log(n \vee p)}$, we have with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$,

$$\|\widehat{\boldsymbol{\beta}}_I - \widetilde{\boldsymbol{\beta}}_I^*\|_q \leq C_1 \lambda d_0^{1/q} / \sqrt{|I|}, \quad \|\widehat{\boldsymbol{b}}_I - \boldsymbol{b}_I^*\|_q \leq C_2 \lambda K^{1/q} \sqrt{d_0/|I|}, \text{ and } \mathcal{M}(\widehat{\boldsymbol{\beta}}_I) \leq C_3 d_0,$$

for $q = 1, 2$, where $C_I, C_\lambda, C_1, C_2, C_3, c_0$, and c_1 are positive absolute constants, depending solely on K and the constants in Assumptions 1 through 3. Additionally, $\mathcal{M}(\widehat{\boldsymbol{\beta}}_I)$ denotes the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_I$.

Lemma 4 provides a uniform consistency result for the CQR Lasso estimator $\widehat{\boldsymbol{\beta}}_I$ defined in (2.2) with respect to its population quantity $\widetilde{\boldsymbol{\beta}}_I^*$ across all sub-intervals $I \subseteq (0, n]$, which is of independent interest. We note that when we estimate $\boldsymbol{\beta}_I^0$ in the sub-interval I, we would need to assume that effiective sample size I is large, usually of order $O(d_0 \log(n \vee p))$ (see e.g. [17, 26, 36]). Lemma 4 is used extensively in the technical proof of DPA and BSA.

4.2. The consistency of the DPA

In this section, we will show consistency for the DPA estimator as defined in Algorithm 1. Let $\kappa = \min_{m=1,\ldots,M} \|\boldsymbol{\beta}_{\eta_m}^0 - \boldsymbol{\beta}_{\eta_{m-1}}^0\|_2$ and $\Delta = \min_{m=1,\ldots,M} (\eta_m - \eta_{m-1})$ be the minimal jump size and minimal spacing, respectively. In our first result, we establish consistency of change-point estimator from DPA.

Theorem 5. In Model (1.1), suppose that Assumptions 1-3 hold and for any $\xi > 0$,

$$\Delta\kappa^2 \geq C_{snr} M d_0^2 \log^{1+\xi}(n \vee p), \tag{4.2}$$

where $C_{snr} > 0$ isthe fixed positive constant. In addition, the change point estimators $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}$ from the DPA detailed in Algorithm 1 with tuning parameters $\lambda = C_\lambda \sqrt{d_0 \log(n \vee p)}$ and $\gamma = C_\gamma (M+1) d_0^2 \log(n \vee p)$, Then, with probability $1 - C(n \vee p)^{-c}$, $\widehat{M} = M$ and

$$\max_{m=1,\ldots,M} \left|\widehat{\eta}_m - \eta_m\right| \leq C_\epsilon M d_0^2 \log(n \vee p)/\kappa^2 \tag{4.3}$$

where $C_\lambda, C_\gamma, C_\epsilon, C, c > 0$ are absolute constants depending only on K and constants in Assumptions 1-3.

The condition (4.2) can be interpreted as a signal-to-noise ratio condition, which plays a crucial role in accurately detecting and estimating the change point locations. We notice that the lower bound in (4.2) is consistent with [26](sub-Gaussian noise) without imposing constraints on the moments of the noise distribution. We observe that the lower bound in (4.2) and the localization error in (4.3) aligns with [26] and exhibits a more lenient requirement on the moments of the noise distribution. By Assumption 1, we can provide the bound $\kappa^2 \leq d_0(2C_\beta)^2 = 4C_\beta^2 d_0$. Hence, (4.2) also implies that

$$\Delta \gtrsim M d_0 \log^{1+\xi}(n \vee p). \tag{4.4}$$

If $\Delta = \Theta(n)$ and $M = O(1)$, then (4.4) becomes $n \gtrsim d_0 \log^{1+\xi}(n \vee p)$, which can be thought of as a standard sample size condition commonly found in the high-dimensional Lasso estimation literature. When the minimal spacing Δ diverges with n, arguably a very natural asymptotic regime, one can take ξ in (4.2) to be set arbitrarily small. To be more precisely, the constant ξ is needed to guarantee consistency when Δ is of the same order as n but can be set to zero if $\Delta = o(n)$. [31] contains further details about ξ that may be of interest.

The proof of Theorem 5 is deferred to Appendix B.1, where it can be seen that the order of the estimation error is of the form $(\lambda^2 d_0 + \gamma)/\kappa^2$. By the SNR condition (4.2) , we have that

$$\max_{m=1,\ldots,M} |\widehat{\eta}_m - \eta_m|/\Delta \leq C_\epsilon M d_0^2 \log(n \vee p)/(\Delta\kappa^2) \leq C_\epsilon/[C_{snr} \log^{\xi}(n \vee p)] \to 0.$$

This explains the role of the quantity ξ in (4.2) and shows the consistency of the DPA algorithm. It is worth emphasizing that the bound in (4.3) along with Model (1.1) provide a *family* of rates, depending on how the model parameters (p, d_0, κ, Δ, M and f_ε) scale with n.

The performance of the CQR Lasso estimator is affected by the tuning parameter λ. Additionally, the second tuning parameter, γ, is employed to prevent overfitting during the search for the optimal partition as a solution to problem (3.1). Specifically, the requirement for γ is essentially that $\gamma \asymp \lambda^2 d_0$. This condition can be intuitively explained as an upper bound on the difference between $\mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2})$ and $\mathcal{L}_n(I_1 \cup I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1 \cup I_2})$, where I_1 and I_2 are two relatively long, non-overlapping, and adjacent intervals, and there is no true change point near the shared endpoint of I_1 and I_2. In this case, one would not wish to partition $I_1 \cup I_2$ into I_1 and I_2. Over-estimating will result in that

$$\mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) < \mathcal{L}_n(I_1 \cup I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1 \cup I_2}).$$

Thus, we avoid the over-partitioning by impose the penalty γ, and further details will be provided in Appendix A.

Note that we do not assume that (3.1) possesses only a single optimal solution, namely $\widehat{\mathcal{I}}$. Instead, the consistency outcome in Theorem 5 holds for any minimizer of the DPA. Moreover, the choice of K significantly impacts computational expenses and the precision of localizing errors. Simulation results indicate that a larger value of K leads to more accurate change point estimates. We suggest selecting $K = 9$, as it achieves a satisfactory balance between computational costs and localization accuracy.

4.3. The consistency of the BSA

Next, we present theoretical results of change point estimators computed by BSA.

Theorem 6. In Model (1.1), suppose that Assumptions 1-3 hold and

$$\Delta\kappa d_0^{-1/2} \geq C_\alpha n^{\Theta}, \text{ and } d_0 \log(n \vee p) \leq c_\alpha n^{4\Theta - 3}, \tag{4.5}$$

where $\Theta \in (3/4, 1]$, $C_\alpha > 0$ is a sufficiently large constant and $c_\alpha > 0$ is sufficient small constant. In addition, the change point estimator $\{\widetilde{\eta}_m\}_{m=1}^{\widetilde{M}}$ from the BSA detailed in Algorithm 2 with input tuning parameters $\lambda =$

$C_\lambda \sqrt{d_0 \log(n \vee p)}$, $\zeta = C_\zeta d_0 \log(n \vee p)$, and $C_\gamma^* \lambda^2 d_0 < \gamma < c_\gamma(\Delta\kappa^2 - \lambda^2 d_0)$. Then, with probability $1 - C(n \vee p)^{-c}$, $\widehat{M} = M$ and

$$\max_{m=1,\ldots,M} \left|\widetilde{\eta}_m - \eta_m\right| \leq C_\epsilon d_0^{3/2} \sqrt{n \log(n \vee p)}/\kappa^2, \tag{4.6}$$

where $C_\lambda, C_\gamma^*, C_\zeta, C_\epsilon, C, c_\gamma, c > 0$ are absolute constants depending only on K and constants in Assumptions 1-3.

SNR condition (4.5) can be seen as a high-dimensional linear model version of Assumption 2 in [5], which was used to establish the consistency of binary segmentation in the univariate mean change point detection problem. Note that the constant C_α is not dependent on n. When the parameters κ and d_0 are fixed, the assumption requires that Δ is of slightly smaller order than n, the size of the data. It is well-known that for the binary segmentation to perform well, Δ cannot be too small compared to n (see, e.g., [23]). We also impose an upper bound on the dimension d_0 in the (4.5) , which is restricted to be at most $n^{4\Theta-3}/\log(n \vee p)$. This means that the sparity d_0 is allowed to diverge as $n \to \infty$, offering flexibility in high-dimensional scenarios.

Theorem 6 implies, that with high probability, the BSA algorithm will estimate their locations with an error that is bounded by

$$\max_{m=1,\ldots,M} \left|\widetilde{\eta}_m - \eta_m\right| / n \lesssim (\kappa^2/d_0)^{-1} \sqrt{d_0 \log(n \vee p)/n},$$

where κ^2/d_0 can be regarded as the standard jump size. The results regarding localization error (4.6) in Theorem 6 are consistent with recent research in the field of multiple change-point detection, including high-dimensional covariance change-point detection [32] and sparse dynamic networks change-point detection [31]. The above bound yields a family of rates of consistency, depending on the scaling of each of the quantities involved in it.

As for the first tuning parameter λ, it affects the performance of the CQR Lasso estimator on the sub-interval I with detectable change points, which is consistent with the DPA. Regarding the parameter γ, its lower bound remains as $\gamma \gtrsim \lambda^2 d_0$, consistent with the gamma used in DPA. However, there is an upper bound requirement for γ. If there are change points to be detected within the interval, an appropriate upper bound of γ using BSA will result in that

$$\mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) + \gamma < \mathcal{L}_n(I_1 \cup I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1 \cup I_2}).$$

Finally, with regard to the tuning parameter ζ, it is important to note that when dealing with small subsamples, controlling the estimation error of (2.1) becomes increasingly challenging. The parameter ζ is specifically designed to address this situation and determines the minimum length required for a subsample $(s, e]$ to be considered for change-point detection. Similar requirements, ensuring the validity of the restricted eigenvalue condition, can be found in other studies such as [17, 36]. In practical applications, setting $\zeta = d_0 \log(n \vee p)$ is generally sufficient.

4.4. Comparisons with related work

Currently, there is a significant scarcity of research on high-dimensional change-point linear regression models with unknown or heavy-tailed error distributions. Most existing models in this domain assume a sub-Gaussian error distribution. Therefore, our objective is to compare the strengths and weaknesses of our proposed method against existing approaches. We now discuss how our contributions compared with the results of [17, 26, 34] , which investigate the same high-dimensional change point linear regression model.

[17] analysed two algorithms, one based on a dynamic programming approach, and the other on binary segmentation, and claimed that they both yield the same localization, which is, in our notation,

$$\max_{m=1,\ldots,M} \left|\widehat{\eta}_m - \eta_m\right| \lesssim d_0^2 \sqrt{n \log(np)}/\kappa^2. \tag{4.7}$$

Note that, the localization error in [17] is originally of the form $\sum_{m=1}^M \left|\widehat{\eta}_m - \eta_m\right| \lesssim d_0 \sqrt{n \log(np)}/\kappa^2$ under a slightly stronger assumption than ours, refered in [26]. Moreover, in the proof of [17], the actual approach used takes the form of $\max_{m=1,\ldots,M} \left|\widehat{\eta}_m - \eta_m\right|$ instead of $\sum_{m=1}^M \left|\widehat{\eta}_m - \eta_m\right|$, and we have made modifications to this aspect. As long as $M \lesssim \sqrt{n/\log(np)}$ using DPA, our localization rates are better than the one implied by (4.7). Furthermore, regarding BSA, our localization rate still surpasses that of [17], even under the more relaxed assumption of noise distribution, making it more applicable to heavy-tailed noise distributions.

[26] introduced an advanced detection methods LS-based Dynamic Programming (LSDP) and Local Refinement for high-dimensional linear regression models under sub-Gaussian noise, incorporating dynamic programming and transformed group lasso penalty. Compared to [17], the major of [26] contributions lie in providing proofs for improved localization accuracy using dynamic programming and establishing a localization lower bound. Our proof approach is derived from framework of LSDP, leading to DPA's localization error being consistent with it. LSDP's technique yields superior results compared to [17], as evident from simulations of LSDP. As a result, in our numerical simulations, we only provide a comparison between LSDP and our methods.

[34] proposed the variance projected wild binary segmentation (VPWBS) based on group lasso estimator for high-dimensional regression models. In particular, they projected the high-dimensional time series $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ onto the univariate time series $\{z_i(u)\}_{i=1}^n$. The optimal projection direction u is obtained by local group Lasso screening(LGS). Then they conducted mean change point detection by WBS on the univariate time series $\{z_i(u)\}_{i=1}^n$. Our findings reveal that under the assumption of sub-Gaussian noise, VPWBS can effectively detect variations in underlying coefficients, contingent on certain conditions being met by the signal. However, VPWBS shows limitations when dealing with heavy-tailed noise. Simulations indicate that the projection structure of VPWBS remains highly competitive, even when the data follows sub-Gaussian or slightly more relaxed heavy-tailed distributions.

5. Simulation studies

In this section, we examine the numerical performance of our proposed methods and compare these with existing techniques in terms of change point detection and identification. Implementation details, such as the algorithm settings and estimation accuracy metrics, are discussed in Section 5.1. Subsequently, Section 5.2 presents the simulation results for different model settings.

5.1. Implementation details

Model settings: To demonstrate the broad applicability of our method, we conduct data generation using various model settings. For the design matrix $\mathbf{X}$, we independently and identically generate $\boldsymbol{X}_i$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$ under two distinct cases: (**i**) $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$; (**ii**) $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$, where $\boldsymbol{\Sigma}^* = (\sigma_{ij}^*) \in \mathbb{R}^{p\times p}$ with $\sigma_{ij}^* = 0.8^{|i-j|}$ for $1 \le i, j \le p$. Moreover, to demonstrate how our methods perform under different error distributions, we generate the error term ε_i from a wide range of distributions:

- the **Normal** distribution, $\varepsilon \sim N(0, 1)$,
- the **Laplace** distribution, where $\varepsilon \sim f_\varepsilon(t) = \frac{1}{2}\exp(-|t|)$,
- the Student's $\boldsymbol{t_\nu}$-distribution with 3 degrees of freedom $\varepsilon \sim t_3$,
- the Student's $\boldsymbol{t_\nu}$-distribution with 2 degrees of freedom $\varepsilon \sim t_2$,
- the **Cauchy** distribution, $\varepsilon \sim f_\varepsilon(t) = 1/[\pi(1 + t^2)]$.

It is important to note that t_2 and Cauchy represent the error without second moments and first moments, respectively. Lastly, we present our simulated results in two distinct scenarios: $M = 1$, and $M = 3$, which correspond to data with single change point, and multiple change points, respectively. The reason behind this is that our proposed method does not rely on prior knowledge of the number of change points and can automatically account for the underlying data generation mechanism.

Evaluation metrics: To ensure an accurate and comprehensive evaluation of our algorithms, we assess them from three perspectives: the estimated number of change points, the accuracy of change point localization, and the computational time. To begin with, we measure the performance of an estimator $\widehat{M}$ in estimating the true number of change points using *mean square error* (MSE). Instead of relying on the localization rate, we adopt the *Hausdorff distance* between the true change points $\{\eta_m\}_{m=1}^M$ and the corresponding estimators $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}$ to evaluate the algorithm. Moreover, we use the scaled Hausdorff distance to evaluate the performance in change point estimation, which is defined as: $D\big(\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}, \{\eta_m\}_{m=1}^M\big) = d\big(\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}, \{\eta_m\}_{m=1}^M\big)/n$, where $d(\cdot,\cdot)$ is the Hausdorff distance between two compact sets S_1, S_2 in $\mathbb{R}$, defined by

$$d(S_1, S_2) = \max\{\max_{s_1\in S_1}\min_{s_2\in S_2}|s_1 - s_2|, \max_{s_2\in S_2}\min_{s_1\in S_1}|s_1 - s_2|\}. \tag{5.1}$$

Our theories suggest that the scaled Hausdorff distance decreases as n increases. Note that scaled Hausdorff distance is a number between 0 and 1, and a smaller one indicates better change point estimation. Lastly, to establish a fair comparison in terms of computational time between DPA and BSA, we implemented the corresponding algorithms on a 2.50 GHz CPU (Linux) with 8 cores and 16 GB of RAM.

Tuning parameters: As discussed in Section 3, there are three tuning parameters (λ, γ, ζ) in our algorithms. We remark that the performance of our algorithms is robust to the choices of ζ, and the key tuning parameters are (λ, γ). Throughout the simulation section, we set $\zeta = d_0 \log(n \vee p)$ across all simulation settings in Section 5.2.

In the following, we provide a sample splitting based cross-validation procedure that selects (λ, γ) in a fully data-driven fashion. Specifically, given the original sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$, let samples with odd indices $\{(Y_{2i-1}, \boldsymbol{X}_{2i-1})\}_{i=1}^{n/2}$ be the training set and even ones $\{(Y_{2i}, \boldsymbol{X}_{2i})\}_{i=1}^{n/2}$ be the validation set. without loss of generality, n is even. Note that the training data and test data share the same number and locations of change-points.

For each pair of $(\lambda, \gamma) \in \Lambda \times \Gamma$ with $\Lambda, \Gamma \subset \mathbb{R}^+$, we conduct our procedure on the training set, obtain the estimated change point, and further estimate the regression coefficients $\{\widehat{\boldsymbol{\beta}}_i\}_{i=1}^{n/2}$ conditional on the estimated change points as in Model (1.1). We then compute the prediction error of $\{\widehat{\widetilde{\boldsymbol{\beta}}}_i\}_{i=1}^{n/2}$ using the test data via $er_i = \frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}\left(Y_{2i} - \boldsymbol{X}_{2i}^{\top}\widehat{\boldsymbol{\beta}}_i - \widehat{b}_{i,k}\right)$. The tuning parameters (λ, γ) are then selected as the pair of $(\lambda, \gamma) \in \Lambda \times \Gamma = \{0.5, 1, 2, 4\} \times \{1, 6, \ldots, 31\}$ that achieves the minimum prediction error $\sum_{i=1}^{n/2} er_i$ on the test data. Based on our extensive numerical simulations, we find that our methods are stable over a certain range of tuning parameters. Hence, we use an empirical choice of the parameters λ and γ to save computational cost.

Competing methods: We further compare our methods with LSDP and VPWBS discussed in Section 4.4. For LSDP and VPWBS, we use the R package named `glmnet` coupled with segmentation technique with parameters through cross-validation procedure.

Others: In regard to all the approaches we have proposed, the regression coefficients $\widehat{\widetilde{\boldsymbol{\beta}}}$ are computed using the R package named `cqrReg`, which introduced by [25]. Theoretical results from [42] demonstrate that the approximate relative efficiency is close to 1 when $K \geq 9$, and choosing $K = 19$ yields even better outcomes. As a result, we offer four quantile sequences $\{\tau_k = k/(K+1)\}_{k=1}^{K}$ with $K \in \{1, 5, 9, 19\}$. To differentiate between algorithms for different values of K, we indicate the specific value of K as a superscript in the abbreviated algorithm name. For instance, DPA^9 denotes the DPA algorithm with $K = 9$.

5.2. Simulation results

5.2.1. Single change point scenario

Let's consider an alternative scenario where there is only one change point present. We set the sample size $n = 400$, the data dimension $p = 100$, the sparsity $d_0 = 5$ and the jump size $\kappa = 5$. The change point occurs at $\eta_1 = 120$ or $\eta_1 = 200$. In addition, we assume the regression coefficients $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^{\top}$,with $\beta_j^0 = \kappa/(2\sqrt{d_0})$, $j \in \{1, \ldots, d_0\}$, and zero otherwise. Let $\boldsymbol{\beta}_1^0 = \cdots = \boldsymbol{\beta}_{\eta_1}^0 = \boldsymbol{\beta}^0$ and $\boldsymbol{\beta}_{\eta_1+1}^0 = \cdots = \boldsymbol{\beta}_n^0 = -\boldsymbol{\beta}^0$.

Tables 1 and 2 respectively present the scaled Hausdorff distance of $\{\widehat{\eta}_m\}_{m=1}^{\widehat{M}}$ and the MSE of $\widehat{M}$ for five noise distributions with various covariance matrices and different η_1. Figure 2 displays the average execution time of DPA and BSA in the single change point scenario, considering different values of K and n.

Table 1: Scaled Hausdorff distances for various methods under different covariance matrix and η_1 in the single change point scenario, based on 100 repetitions. Each highlighted and italicized number represents the best and worst performance within the corresponding setting.

		Change point detection for $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$							**Change point detection for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$**					
η_1	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**		η_1	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**
120	DPA^1	0.013	0.010	0.020	0.030	0.068		120	DPA^1	0.012	0.014	0.021	0.032	0.064
	BSA^1	*0.014*	0.010	0.022	0.030	0.068			BSA^1	*0.014*	0.014	0.021	0.032	0.068
	DPA^5	0.010	0.008	0.014	0.021	0.054			DPA^5	0.007	0.012	0.014	0.024	0.055
	BSA^5	0.010	0.007	0.017	0.022	0.054			BSA^5	0.007	0.011	0.014	0.024	0.055
	DPA^9	0.004	0.006	**0.010**	**0.010**	**0.049**			DPA^9	0.005	0.008	0.011	0.017	0.054
	BSA^9	0.005	0.006	0.011	0.014	0.050			BSA^9	0.005	0.009	0.012	0.017	0.059
	DPA^{19}	0.005	**0.005**	**0.010**	**0.010**	0.051			DPA^{19}	0.004	**0.005**	**0.011**	**0.016**	**0.052**
	BSA^{19}	0.005	**0.005**	**0.010**	0.012	0.050			BSA^{19}	0.005	**0.005**	**0.011**	0.017	0.054
	LSDP	0.004	*0.011*	0.042	0.131	0.262			LSDP	0.003	*0.015*	0.041	0.142	0.242
	VPWBS	**0.002**	*0.011*	*0.046*	*0.172*	*0.315*			VPWBS	**0.002**	*0.015*	*0.048*	*0.185*	*0.312*
200	DPA^1	0.010	0.008	0.019	0.026	0.062		200	DPA^1	0.010	0.012	0.018	0.031	0.060
	BSA^1	*0.012*	0.008	0.024	0.030	0.065			BSA^1	*0.012*	0.012	0.022	0.030	0.060
	DPA^5	0.009	0.006	0.012	0.018	0.049			DPA^5	0.006	0.007	0.012	0.019	0.052
	BSA^5	0.008	0.006	0.016	0.019	0.050			BSA^5	0.006	0.007	0.012	0.020	0.054
	DPA^9	0.004	0.003	**0.006**	0.010	0.042			DPA^9	0.004	0.006	0.008	0.014	0.048
	BSA^9	0.004	0.004	0.008	0.012	0.045			BSA^9	0.004	0.006	0.009	0.016	0.050
	DPA^{19}	0.004	**0.003**	0.008	**0.009**	**0.040**			DPA^{19}	0.004	**0.003**	**0.007**	**0.012**	**0.046**
	BSA^{19}	0.005	0.004	0.008	0.011	0.042			BSA^{19}	0.005	0.004	0.008	**0.012**	**0.046**
	LSDP	0.003	0.008	0.036	0.126	0.210			LSDP	0.002	*0.014*	0.040	0.132	0.214
	VPWBS	**0.002**	*0.010*	*0.042*	*0.154*	*0.241*			VPWBS	**0.001**	0.012	*0.042*	*0.142*	*0.238*

Firstly, we can observe that in heavy-tailed distributions, particularly the t_v and Cauchy distribution, DPA and BSA demonstrate significantly higher accuracy compared to LSDP and VPWBS. However, under the standard normal distribution, LSDP and VPWBS still maintain their advantages. Regarding DPA compared to BSA, it demonstrates better estimation accuracy. DPA may suffer from a tendency to overestimate the number of change points, resulting in an increase in the Hausdorff distance, as shown in Table 2. Nevertheless, it is evident from simulation of [26] that providing prior information about the number of change points could lead to improved performance for DPA. It is worth noting that VPWBS performs catastrophically in the Cauchy distribution. This

Table 2: MSE of $\widehat{M}$ for various methods under different covariance matrix and η_1 in the single change point scenario, based on 100 repetitions. Each highlighted and italicized number represents the best and worst performance within the corresponding setting.

		Change point detection for $\Sigma = \mathbf{I}_{p\times p}$							Change point detection for $\Sigma = \Sigma^*$				
η_1	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**	η_1	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**
120	DPA[1]	*0.03*	0.02	0.06	0.09	0.20	120	DPA[1]	*0.03*	0.02	0.07	0.11	0.21
	BSA[1]	*0.03*	0.02	0.06	0.09	0.20		BSA[1]	*0.03*	0.02	0.06	0.11	0.21
	DPA[5]	0.02	0.02	0.04	0.08	0.17		DPA[5]	0.01	0.02	0.04	0.08	0.18
	BSA[5]	0.02	0.02	0.06	0.07	0.17		BSA[5]	0.01	0.02	0.03	0.07	0.16
	DPA[9]	0.01	**0.01**	**0.01**	0.05	0.17		DPA[9]	0.01	0.02	**0.01**	0.06	**0.14**
	BSA[9]	0.01	**0.01**	0.02	0.05	**0.16**		BSA[9]	0.01	**0.01**	0.02	**0.05**	0.18
	DPA[19]	0.01	**0.01**	**0.01**	**0.04**	**0.16**		DPA[19]	0.01	**0.01**	0.02	**0.05**	0.16
	BSA[19]	0.01	**0.01**	**0.01**	**0.04**	**0.16**		BSA[19]	0.01	**0.01**	0.01	**0.05**	0.16
	LSDP	**0.00**	*0.05*	0.12	0.15	0.78		LSDP	0.01	0.02	0.11	0.15	0.66
	VPWBS	**0.00**	0.02	*0.14*	*1.54*	*8.43*		VPWBS	**0.00**	*0.03*	*0.16*	*1.78*	*7.98*
200	DPA[1]	*0.03*	0.02	0.06	0.08	0.20	200	DPA[1]	*0.03*	0.02	0.06	0.10	0.20
	BSA[1]	*0.03*	0.02	0.08	0.09	0.20		BSA[1]	0.02	0.02	0.06	0.09	0.20
	DPA[5]	0.02	0.02	0.04	0.07	0.17		DPA[5]	0.02	0.02	0.04	0.06	0.18
	BSA[5]	0.01	0.02	0.07	0.07	0.17		BSA[5]	0.02	0.02	0.04	0.06	0.18
	DPA[9]	0.01	**0.00**	**0.00**	0.04	0.16		DPA[9]	0.01	0.02	0.01	0.05	0.14
	BSA[9]	**0.00**	**0.00**	**0.00**	**0.03**	0.16		BSA[9]	0.01	0.01	0.01	0.05	0.17
	DPA[19]	0.01	**0.00**	0.01	0.04	0.16		DPA[19]	0.01	**0.00**	0.01	**0.04**	**0.13**
	BSA[19]	0.01	**0.00**	**0.00**	**0.03**	**0.15**		BSA[19]	0.01	**0.00**	**0.00**	**0.04**	**0.13**
	LSDP	**0.00**	0.03	*0.10*	0.13	0.44		LSDP	**0.00**	*0.03*	0.10	0.15	0.48
	VPWBS	**0.00**	*0.05*	*0.10*	*1.29*	*4.72*		VPWBS	**0.00**	*0.03*	*0.12*	*1.16*	*4.32*

suggests that for data with extreme heavy tails the existing techniques may not be applicable. Note that all the proposed algorithms perform better the closer the change point location is to the middle of the data observations.

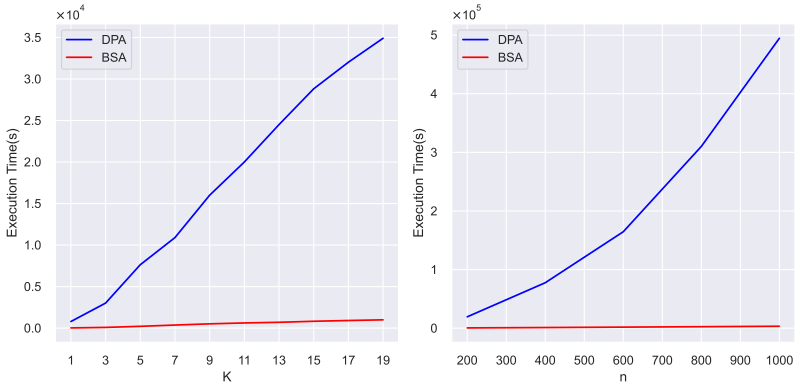

Fig. 2: Efficiency of change point estimation under the model with one change point. The left panel shows average execution time of BSA and DPA across different $K \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$ with fixed $n = 400$ and $p = 100$. The right panel shows average execution time of BSA and DPA across different $n \in \{200, 400, 600, 800, 1000\}$ with fixed $p = 100$, and $K = 9$.

As for the choice of K, we observe that when $K = 1$ (i.e., median regression), our method excels particularly well under Laplace noise but exhibits slightly inferior performance under other distribution types. It is evident that as K gradually increases, both DPA and BSA produce more precise change point estimates. Additionally, we observe that increasing K to 19 does not yield a significant improvement in performance compared to $K = 9$. However, the substantial computational cost associated with $K = 19$, as depicted in Figure 2 (left), led us to prioritize computational efficiency. Hence, in the forthcoming scenarios (multiple change point scenario in Section 5.2.2), considering both efficiency and accuracy, we will only focus on the case of $K = 9$. As shown in Figure 2 (right), the computational cost of BSA^9 grows moderately (514 - 3348 s) as the data sample size increases from 200 to 1000, while the computational cost of DPA^9 grows exponentially (19,000 - 500,000 s).

5.2.2. Multiple change points scenario

Finally, let's explore an alternative scenario where we introduce three change points. We set the sample size $n = 800$ and the data dimension $p = 200$. The change points are located at $\eta_1 = 200, \eta_2 = 400$, and $\eta_3 = 600$, respectively. The underlying regression coefficients $\{\boldsymbol{\beta}_i^0\}_{i=1}^n$ are structured in the following manner:

$$\boldsymbol{\beta}_i^* = \begin{cases} c_\kappa \cdot (\underbrace{1,\ldots,1}_{8},0,\ldots,0), & i \in \{1,\ldots,200\}, \\ c_\kappa \cdot (\underbrace{0,\ldots,0}_{8},\underbrace{1,\ldots,1}_{8},0,\ldots,0), & i \in \{201,\ldots,400\}, \\ c_\kappa \cdot (\underbrace{0,\ldots,0}_{16},\underbrace{1,\ldots,1}_{8},0,\ldots,0), & i \in \{401,\ldots,600\}, \\ c_\kappa \cdot (\underbrace{0,\ldots,0}_{24},\underbrace{1,\ldots,1}_{8},0,\ldots,0), & i \in \{601,\ldots,800\}. \end{cases}$$

In Table 3 and Table 4, we present the results for the scaled Hausdorff distance and MSE of $\widehat{M}$ with two signal strengths, $c_\kappa \in \{1, 2\}$, and five different noise distributions.

Table 3: Scaled Hausdorff distances for various methods under different covariance matrix and c_κ in the multiple change points scenario, based on 100 repetitions. Each highlighted and italicized number represents the best and worst performance within the corresponding setting.

	Change point detection for $\Sigma = \mathbf{I}_{p\times p}$							**Change point detection for $\Sigma = \Sigma^*$**					
c_κ	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**	c_κ	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**
1	DPA^9	*0.014*	**0.016**	**0.025**	**0.030**	**0.054**	1	DPA^9	*0.016*	**0.020**	**0.026**	**0.033**	**0.056**
	BSA^9	*0.014*	**0.016**	0.030	0.031	0.058		BSA^9	*0.016*	0.021	0.029	**0.033**	0.060
	LSDP	0.012	*0.029*	0.058	0.113	0.114		LSDP	0.012	0.028	0.061	0.115	0.110
	VPWBS	**0.009**	*0.028*	*0.086*	*0.119*	*0.192*		VPWBS	**0.011**	*0.029*	*0.089*	*0.120*	*0.184*
2	DPA^9	*0.006*	**0.010**	**0.013**	**0.016**	**0.040**	2	DPA^9	0.006	**0.008**	**0.014**	0.017	**0.039**
	BSA^9	*0.006*	0.011	0.014	**0.016**	0.043		BSA^9	*0.007*	0.010	**0.014**	**0.016**	0.043
	LSDP	**0.003**	0.015	0.042	0.102	0.115		LSDP	0.005	*0.017*	0.043	0.104	0.117
	VPWBS	**0.003**	*0.017*	*0.049*	*0.116*	*0.181*		VPWBS	**0.002**	0.016	*0.046*	*0.112*	*0.178*

Table 4: MSE of $\widehat{M}$ for various methods under different covariance matrix and c_κ in the multiple change points, based on 100 repetitions. Each highlighted and italicized number represents the best and worst performance within the corresponding setting.

	Change point detection for $\Sigma = \mathbf{I}_{p\times p}$							**Change point detection for $\Sigma = \Sigma^*$**					
c_κ	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**	c_κ	Method	**Normal**	**Laplace**	t_3	t_2	**Cauchy**
1	DPA^9	*0.02*	**0.02**	0.06	**0.07**	0.19	1	DPA^9	*0.02*	**0.02**	**0.05**	**0.08**	0.20
	BSA^9	*0.02*	**0.02**	**0.05**	**0.07**	**0.16**		BSA^9	*0.02*	**0.02**	**0.05**	**0.08**	**0.18**
	LSDP	*0.02*	*0.06*	0.20	0.52	0.94		LSDP	**0.01**	0.04	0.26	0.96	1.02
	VPWBS	**0.01**	0.05	*0.65*	*6.34*	*19.68*		VPWBS	**0.01**	*0.05*	*0.82*	*7.05*	*17.82*
2	DPA^9	*0.02*	**0.01**	**0.02**	**0.04**	0.14	2	DPA^9	*0.01*	**0.01**	**0.02**	0.05	0.16
	BSA^9	0.01	**0.01**	**0.02**	**0.04**	**0.13**		BSA^9	*0.01*	0.02	**0.02**	**0.04**	**0.15**
	LSDP	0.01	*0.04*	0.10	0.34	0.76		LSDP	*0.01*	*0.03*	0.11	0.38	0.91
	VPWBS	**0.00**	0.03	*0.22*	*1.94*	*17.21*		VPWBS	**0.00**	*0.03*	*0.19*	*2.24*	*14.64*

To facilitate a visual comparison, Figure 1 displays boxplots depicting the scaled Hausdorff distance for four methods across various noise distributions, excluding the normal distribution. Notably, for the normal noise distribution, both LSDP and VPWBS consistently demonstrate strong performance, as evidenced by the smaller Hausdorff distance and MSE. In contrast, when dealing with heavy-tailed distributions, DPA^9 and BSA^9 maintain a significant advantage. It's worth noting that as the signal strength decreases, the performance of all methods deteriorates due to the increased difficulty in estimation. Similar to the single change point scenario, DPA offers more precise localization but may occasionally overestimate the number of change points. Additionally, as both n and M increase, the performance of LSDP and VPWBS deteriorates further, as elaborated in the preceding subsection.

6. Real data applications

We analyze the air quality dataset from beijing-multisite-airquality-data-set, as discussed in [40]. This dataset consists of hourly data spanning from 2013 to 2017, containing 6 main pollutants and 6 relevant meteorological variables across 12 air quality monitoring sites in Beijing.

Specifically, we have chosen the time period from June 1, 2015, to June 1, 2016. The PM2.5 concentration levels at the Dongsi site are selected as the response variable, while the covariates include the SO_2 concentration ($\mu g/m^3$), the NO_2 concentration ($\mu g/m^3$), the CO concentration ($\mu g/m^3$), the O_3 concentration ($\mu g/m^3$), the temperature (°C), the pressure (hPa), the dew point temperature (°C), the precipitation (mm), the wind speed (m/s), and the PM2.5 concentration levels corresponding to the Aotizhongxin, Changping, Dingling, Guanyuan,

Gucheng, Huairou, Nongzhanguan, Shunyi, Tiantan, Wanliu, and Wanshouxigong air quality monitoring sites. In the subsequent step, to facilitate change point detection, we transformed the hourly data into daily data by calculating the averages across 24 hours and eliminating all dates containing missing values. Consequently, we obtained a finalized dataset $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ with $n = 360$ days and $p = 20$ covariates. Our objective is to identify potential change points in this dataset and assess their consistency with historical information.

We applied the DPA and BSA algorithms for $K = 9$. DPA identified a change point on December 8, 2015. Interestingly, BSA pinpointed December 13, 2015, as the change point. The visualization is presented in Figure 3.

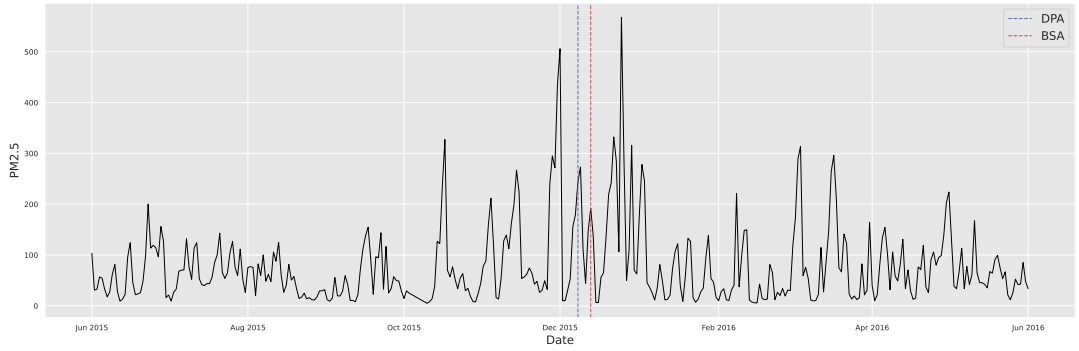

Fig. 3: PM2.5 concentration level ($\mu g/m^3$) detected at Dongsi air quality monitoring site in Beijing, from June 1, 2015, to June 1, 2016. The vertical dashed lines in blue and red denote the change point detected by DPA and BSA, respectively.

The change point derived from BSA and DPA are closely situated, potentially associated with the peak of the El Niño event (e.g. [2]) during the winter period. It is widely acknowledged that El Niño events can induce significant anomalies in atmospheric circulation and weather patterns. During the December peak of the 2015 El Niño, the combined influence of anomalous northwest Pacific low-level anticyclonic circulation and the negative-phase anomaly circulation of the Euro-Atlantic-Pacific teleconnection pattern suppressed the northerly airflow in Northern China. Consequently, there was a notable increase in PM2.5 concentrations, which reached the hazardous red alert on December 8th in Beijing. Simultaneously, intensified coal-based heating during the autumn and winter seasons exacerbated the frequent occurrence of haze.

7. Summary

In this paper, we present two change point detection algorithms in high-dimensional linear models, where the dimension p can be much larger than the sample size n. Our algorithms are based on composite quantile loss, allowing them to handle unknown or heavy-tailed noise. Notably, our proposed algorithms can automatically explain the underlying data generation mechanism without the need for specifying any prior knowledge about the number of change points M. Additionally, we propose two algorithms, DPA and BSA, based on dynamic programming and binary segmentation techniques to detect multiple change points. Furthermore, we investigate the theoretical properties of the estimated change point estimators computed by the two algorithms without requirement on the moments of the noise distribution. We establish the consistency of the estimated number and locations of the change points. Moreover, we have provided a detailed proof of convergence for both algorithms, overcoming the challenge posed by the non-differentiability of quantile loss. DPA achieves better localization accuracy, while BSA has lower computational costs. Finally, we demonstrate the efficiency and accuracy of our proposed methods through extensive numerical results under various model settings.

Acknowledgement

This research is supported in part by the National Social Science Fund of China 22BTJ031 (Liwen Zhang), and the National Natural Science Foundation of China Grant 12101132 (Bin Liu).

Appendix A. useful lemmas

Before the proofs, we give some notations. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, we denote $J(\boldsymbol{v}) = \{1 \le j \le p : v_j \neq 0\}$ as the set of non-zero elements of $\boldsymbol{v}$ and set $\mathcal{M} := |J(\boldsymbol{v})|$ as the number of non-zero elements of $\boldsymbol{v}$. For a set J and $\boldsymbol{v} \in \mathbb{R}^p$, we denote $\boldsymbol{v}_J$ as the vector in $\mathbb{R}^p$ that has the same coordinates as $\boldsymbol{v}$ on J and zero coordinates on the complement J^c of J. For any symmetric positive definite matrix $\mathbf{A}$, define $\|\boldsymbol{x}\|_{\mathbf{A}} := (\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x})^{1/2}$. We use $C_1, C_2, \ldots$ to denote

constants that may vary from line to line. We let $\underset{\sim}{\boldsymbol{\beta}}_k := (\boldsymbol{\beta}^\top, b_k)^\top \in \mathbb{R}^{p+1}$, for $k = 1, \ldots, K$. Following these rules, we can define $\widehat{\underset{\sim}{\boldsymbol{\beta}}}, \underset{\sim}{\boldsymbol{\beta}}^*$ and $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_k, \underset{\sim}{\boldsymbol{\beta}}^*_k$ for $k = 1, \ldots, K$. Without causing confusions, we use $\underset{\sim}{\boldsymbol{X}}$ to denote both the $\mathbb{R}^{p+1}$ vector $(\boldsymbol{X}^\top, 1)^\top$ and the $\mathbb{R}^{p+K}$ vector $(\boldsymbol{X}^\top, \mathbf{1}_K^\top)^\top$, where $\mathbf{1}_K = (1, \ldots, 1)^\top \in \mathbb{R}^K$. Lemmas 8-12 play a fundamental role in proving Theorem 5-6, serving as the cornerstone elements of the proof.

Lemma 7 (Choice of γ)**.** Suppose the same assumptions hold as Theorem 5 hold in Model (1.1), if there exists no true change point in $I = (s, e]$, and $\lambda \geq \lambda_1 := C_\lambda \sqrt{\log(n \vee p)}$, where $C_\lambda > 0$ being an absolute constant, it with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ holds that

$$\left| \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*_I) \right| \leq \max\{C_1 \lambda^2 d_0, C_4 \lambda d_0^{3/2} \sqrt{\log(n \vee p)}\},$$

where $C_4 > 0$ is an absolute constant depending on all the other constants.

Lemma 8 (One change point)**.** Suppose the same assumptions hold as Theorem 5 hold in Model (1.1), assume that $I = (s, e]$ has only one true change point η . Denote $I_1 = (s, \eta]$, $I_2 = (\eta, e]$. If, in addition, it holds that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) \leq \mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) + \gamma, \tag{A.1}$$

then with $\lambda \geq \lambda_2 := C_\lambda \sqrt{d_0 \log(n \vee p)}$, where $C_\lambda > 0$ being big enough constants, it holds with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that

$$\min\{|I_1|, |I_2|\} \leq C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2.$$

Lemma 9 (Two change points)**.** Suppose the same assumptions hold as Theorem 5 hold in Model (1.1), assume that $I = (s, e]$ containing exactly two change points η_1 and η_2. Denote $I_1 = (s, \eta_1]$, $I_2 = (\eta_1, \eta_2]$, $I_3 = (\eta_2, e]$. If, in addition, it holds that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) \leq \mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) + \mathcal{L}_n(I_3, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_3}) + 2\gamma, \tag{A.2}$$

then with $\lambda \geq \lambda_2 := C_\lambda \sqrt{d_0 \log(n \vee p)}$, where $C_\lambda > 0$ being big enough constants, it holds with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that

$$\min\{|I_1|, |I_3|\} \leq C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2.$$

Lemma 10 (Multiple change points)**.** Suppose the same assumptions hold as Theorem 5 hold in Model (1.1), if $I = (s, e]$ contains T true change points $\{\eta\}_{t=1}^T$, where $T \geq 3$. If $\lambda \geq \lambda_2 := C_\lambda \sqrt{d_0 \log(n \vee p)}$, where $C_\lambda > 0$ being big enough constants, then with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that, in addition, it holds that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) > \sum_{t=1}^{T+1} \mathcal{L}_n(I_t, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_t}) + T\gamma, \tag{A.3}$$

where $I_1 = (s, \eta_1]$, $I_t = (\eta_t, \eta_{t+1}]$ for any $2 \leq t \leq T$ and $I_{T+1} = (\eta_{T+1}, e]$.

Lemma 11 (No change point)**.** Suppose the same assumptions hold as Theorem 5 hold in Model (1.1), if there exists no true change point in $I = (s, e]$, with $\lambda > \lambda_2 := C_\lambda \sqrt{d_0 \log(n \vee p)}$ where $C_\lambda > \max\{\sqrt{12}K^{-1/2}, 128(2\rho^{-1/2} + 1)K^{-1/2}\}$, and $\gamma = C_\gamma d_0^2 \log(n \vee p)$ where $C_\gamma > \max\{3C_\lambda^2, 3C_4 C_\lambda\}$, it holds with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) \leq \min_{a=s+1,\ldots,e-1} \left\{ \mathcal{L}_n((s, a], \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{(s,a]}) + \mathcal{L}_n((a, e], \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{(a,e]}) \right\} + \gamma.$$

Lemma 12. Suppose the same assumptions hold as Theorem 5 hold in Model (1.1), if there exists no true change point in the interval I. For any interval $J \supset I$ with $|J| \geq C_I d_0 \log(n \vee p)$, with $\lambda \geq \lambda_2 := C_\lambda \sqrt{d_0 \log(n \vee p)}$ where $C_\lambda > 0$, then with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that, in addition, it holds that

$$\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*_I) - \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_J) \leq C_5 \lambda^2 d_0. \tag{A.4}$$

Appendix B. Proof of main results

Appendix B.1. Proof of Theorem 5

In this subsection, we present the proof for Theorem 5. For simplicity, we omit the subscript I whenever it is required. To begin, we provide two additional propositions. The proof of Theorem 5 is an immediate consequence of Propositions 13 and 14. For clarity and convenience. The first additional proposition shows error bound in the different case, The proof of Proposition 13 is given in Appendix D.1.

Proposition 13. Assuming the same conditions as in Theorem 5, and letting $\widehat{\mathcal{I}}$ be the solution to (3.1), the following results hold with probability at least $1 - C(n \vee p)^{-c}$.

(i) For each interval $\widehat{I} = (s, e] \in \widehat{\mathcal{I}}$ containing one and only one true change point η, it is guaranteed that $\min\{\eta - s, e - \eta\} \leq C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$, where $C_\epsilon > 0$ is an absolute constant;

(ii) for each interval $\widehat{I} = (s, e] \in \widehat{\mathcal{I}}$ containing two true change point $\eta_1 < \eta_2$, it must be the case that $\min\{\eta_1 - s, e - \eta_2\} \leq C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$, where $C_\epsilon > 0$ is an absolute constant;

(iii) no interval $\widehat{I} \in \widehat{\mathcal{I}}$ contains strictly more than two true change points; and

(iv) for all consecutive intervals $\widehat{I}$ and $\widehat{J}$ in $\widehat{\mathcal{I}}$, the interval $\widehat{I} \cup \widehat{J}$ contains at least one true change point.

The second additional proposition shows number of estimated change points is consistency. The proof of Proposition 14 is given in Section Appendix D.2.

Proposition 14. Under the same conditions in Theorem 5, with $\widehat{\mathcal{I}}$ being the the solution to (3.1), satisfying $M \leq |\widehat{\mathcal{I}}| \leq 3M$, then with probability at least $1 - C(n \vee p)^{-c}$, it holds that $\widehat{\mathcal{I}} = M$.

It follows from Proposition 13 that, $M \leq |\widehat{\mathcal{I}}| \leq 3M$. This combined with Proposition 14 completes the proof.

Appendix B.2. Proof of Theorem 6

The overall proof strategy is to apply the induction argument commonly used in the Binary Search algorithms. Denote $\epsilon_n = C_\epsilon d_0^{3/2}\sqrt{n \log(n \vee p)}/\kappa^2$. Let $I := (s, e] \subset (0, n]$ be the change point estimators returned by the Binary Search in the previous repetition. Since ϵ_n is the desired localization rate, by induction, it suffices to consider any generic $(s, e] \subset (0, n]$ that satisfies the following three conditions:

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq -1; \tag{B.1}$$

$$\text{either } \eta_r - s \leq \epsilon_n \quad \text{or} \quad s - \eta_{r-1} \leq \epsilon_n; \tag{B.2}$$

$$\text{either } \eta_{r+q+1} - e \leq \epsilon_n \quad \text{or} \quad e - \eta_{r+q} \leq \epsilon_n.$$

Here $q = -1$ indicates that there is no change point contained in $(s, e]$. Observe that under SNR condition (4.5), for sufficiently large constant C_α, it holds that $\epsilon_n < \Delta/4$. Therefore, ifor any true change point $\eta_p \in \{\eta_r, \ldots, \eta_{r+q}\}$, it is either the case that $|\eta_p - s| \leq \epsilon_n$, or that $|\eta_p - s| \geq \Delta - \epsilon_n \geq \Delta/4$. This means that $\min\{|\eta_p - e|, |\eta_p - s|\} \leq \epsilon_n$ indicates that η_p is a detected change point in the previous induction step, even if $\eta_p \in (s, e]$. We refer to $\eta_p \in (s, e]$ as an undetected change point if $\min\{|\eta_p - s|, |\eta_p - e|\} \geq \Delta/4$.

To complete the induction step, it suffices to show that $\mathrm{BSA}((s, e], \lambda, \gamma, \zeta)$ **(i)** will not detect any new change point in $(s, e]$ if all the change points in that interval have been previous detected, and **(ii)** will find a point v in $(s, e]$ such that $|\eta_p - v| \leq \epsilon_n$ if there exists at least one undetected change point in (s, e).

Step 1. Suppose there does not exist any undetected change points in $(s, e]$, one of the following situations must hold: (**a**) There is no change point within $(s, e]$; (**b**) there exists only one change point η_r within $(s, e]$ and $\min\{\eta_r - s, e - \eta_r\} \leq \epsilon_n$; (**c**) there exist two change points η_r, η_{r+1} within $(s, e]$ and $\eta_r - s \leq \epsilon_n$ and $e - \eta_{r+1} \leq \epsilon_n$. Note that $p = r$ in this case. Under SNR condition (4.5), for sufficiently large constant C_α, it holds that $\zeta \leq \Delta/4$ from (**a**). Furthermore, we can use similar methods as in the proof of Lemmas 11, 8 and 9 to derive conclusions (**a**), (**b**) and (**c**). As a result, $\mathrm{BSA}((s, e], \lambda, \tau, \zeta)$ correctly reject if $(s, e]$ contains no undetected change points.

Step 2. Let v be defined as in $\mathrm{BSA}((s, e], \lambda, \tau, \zeta)$. Assume that there exists a undetected change point $\eta_p \in \{\eta_r, \ldots, \eta_{r+q}\}$ closest to v. To complete the induction, it suffices to show that

$$\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4, \tag{B.3}$$

and that

$$|v - \eta_p| \leq \epsilon_n. \tag{B.4}$$

First, we will show why (B.3) holds. For the sake of contradiction, assume that $\min\{\eta_p - s, e - \eta_p\} < 3\Delta/4$. Suppose $\eta_p - s < 3\Delta/4$. Since $\eta_p - \eta_{p-1} \geq \Delta$, one has $s - \eta_{p-1} \geq \Delta/4$. This also means that η_p is the first change point within $(s, e]$. Therefore $p = r$ in (B.1). By (B.2), $\min\{\eta_r - s, s - \eta_{r-1}\} \leq \epsilon_n < \Delta/4$. Since $s - \eta_{p-1} = s - \eta_{r-1} \geq \Delta/4$, it must be the case that $\eta_p - s = \eta_r - s \leq \epsilon_n$. This is a contradiction to $\eta_p - s| \geq \Delta/4$. Therefore $\eta_p - s \geq 3\Delta/4$. The argument of $e - \eta_p \geq 3\Delta/4$ can be made analogously.

Next, we will show that $\mathrm{BSA}((s, e], \lambda, \tau, \zeta)$ can consistently detect the existence of undetected change points within $(s, e]$. We denote some intervals notation I_1, I_2, J_1, J_2 and L, as shown in Figure B.4. By definition of v, one then obtains the inequality:

$$\begin{aligned}
&\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \sum_{i\in\{1,2\}} \mathcal{L}_n(I_i, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_i}) \geq \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \sum_{i\in\{1,2\}} \mathcal{L}_n(I_i, \underset{\sim}{\boldsymbol{\beta}}^*_{I_i}) - 2C_1\lambda^2 d_0 \\
\geq &\sum_{i\in\{1,2\}} \big\{\mathcal{L}(I_i, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}(I_i, \underset{\sim}{\boldsymbol{\beta}}^*_{I_i}) - \big|\mathcal{L}_n(I_i, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}_n(I_i, \underset{\sim}{\boldsymbol{\beta}}^*_{I_i}) - \big[\mathcal{L}(I_i, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}(I_i, \underset{\sim}{\boldsymbol{\beta}}^*_{I_i})\big]\big|\big\} - 2C_1\lambda^2 d_0 \\
\geq &\underline{f}/(4K) \sum_{i\in\{1,2\}} |I_i| \|\underset{\sim}{\boldsymbol{\Delta}}_i\|_{\mathbf{S}}^2 - \lambda\sqrt{d_0} \sum_{i\in\{1,2\}} |I_i| \|\underset{\sim}{\boldsymbol{\Delta}}_i\|_{\mathbf{S}} - 2C_1\lambda^2 d_0 \geq \underline{f}/(8K) \sum_{i\in\{1,2\}} |I_i| \|\underset{\sim}{\boldsymbol{\Delta}}_i\|_{\mathbf{S}}^2 - (2C_1 + 4K\underline{f}^{-1})\lambda^2 d_0 \\
\geq &C_0 \|\boldsymbol{\beta}^*_{I_1} - \boldsymbol{\beta}^*_{I_2}\|_2^2 \min\{|I_1|, |I_2|\} - (2C_1 + 4K\underline{f}^{-1})\lambda^2 d_0 \geq C_0'\Delta\kappa^2 - (2C_1 + 4K\underline{f}^{-1})\lambda^2 d_0 \geq c_\gamma(\Delta\kappa^2 - \lambda^2 d_0),
\end{aligned}$$

where the penultimate inequality follows from the assumption that (B.3) and $\|\boldsymbol{\beta}^*_{I_1} - \boldsymbol{\beta}^*_{I_2}\|_2^2 \asymp \kappa^2$, the other inequalities from a similar argument as proof of Lemma 8. Therefore BSA$((s, e], \lambda, \tau, \zeta)$ will tend to favor split the interval that containing detectable change points.

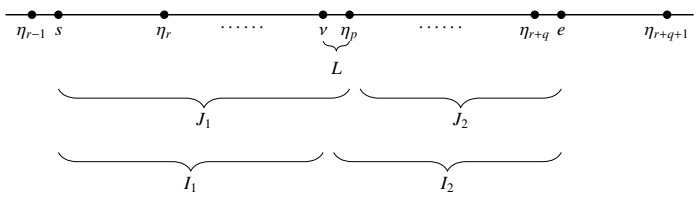

Fig. B.4: Illustrations of the interval constructions used in the **Step 2** in the proof of Theorem 6.

Finally, we will prove why (B.4) holds. Our proof begins by assuming that $\eta_p - v > \epsilon_p$ holds when $\eta_p \geq v$, in order to demonstrate a contradiction with (B.4). By definition of v, one then obtains the inequality:

$$\begin{aligned}
\mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) &\leq \mathcal{L}_n(J_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{J_1}) + \mathcal{L}_n(J_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{J_2}) - \mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) \leq \mathcal{L}_n(J_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{J_1}) + \mathcal{L}_n(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) - \mathcal{L}_n(I_1, \underset{\sim}{\boldsymbol{\beta}}^*_{I_1}) + 2C_1\lambda^2 d_0 \\
&\leq_{(1)} \mathcal{L}_n(J_1, \underset{\sim}{\boldsymbol{\beta}}^*_{I_1}) + \lambda\sqrt{|J_1|}(\|\boldsymbol{\beta}^*_{I_1} - \widehat{\boldsymbol{\beta}}_{J_1}\|_1) + \mathcal{L}_n(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) - \mathcal{L}_n(I_1, \underset{\sim}{\boldsymbol{\beta}}^*_{I_1}) + 2C_1\lambda^2 d_0 \\
&\leq_{(2)} \mathcal{L}_n(L, \underset{\sim}{\boldsymbol{\beta}}^*_{I_1}) + \lambda\sqrt{|J_1|}(\|\boldsymbol{\beta}^*_{I_1} - \boldsymbol{\beta}^*_{J_1}\|_1) + \mathcal{L}_n(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) + 3C_1\lambda^2 d_0 \\
&\leq_{(3)} \mathcal{L}_n(L, \underset{\sim}{\boldsymbol{\beta}}^*_{L}) + \mathcal{L}_n(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) + C_1' d_0^2 \log(n \vee p) + C_2' d_0^{3/2}\sqrt{n\log(n \vee p)} \\
&\leq_{(4)} \mathcal{L}_n(L, \underset{\sim}{\boldsymbol{\beta}}^*_{L}) + \mathcal{L}_n(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) + C d_0^{3/2}\sqrt{n\log(n \vee p)},
\end{aligned} \tag{B.5}$$

where (1) follows from definition of $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_{J_1}$ with $|J_1| \gtrsim \log(n \vee p)$, (2) follows from $\|\boldsymbol{\beta}^*_{I_1} - \widehat{\boldsymbol{\beta}}_{J_1}\|_1 \leq \|\boldsymbol{\beta}^*_{I_1} - \boldsymbol{\beta}^*_{J_1}\|_1 + \|\boldsymbol{\beta}^*_{J_1} - \widehat{\boldsymbol{\beta}}_{J_1}\|_1$, (3) follows from the facts that Lemma 4 with $\lambda = C_\lambda d_0^2 \log(n \vee p)$ and $|\mathcal{L}_n(L, \underset{\sim}{\boldsymbol{\beta}}^*_{L}) - \mathcal{L}_n(L, \underset{\sim}{\boldsymbol{\beta}}^*_{I_1})| \lesssim \lambda^2 d_0$, and (4) follow from $s\log(n \vee p) \leq c_\alpha n$ by (4.5). Then (B.5) leads to

$$\begin{aligned}
\mathcal{L}(L, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) - \mathcal{L}(L, \underset{\sim}{\boldsymbol{\beta}}^*_{L}) + \mathcal{L}(J_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) - \mathcal{L}(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) &\leq \big|\mathcal{L}_n(L, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) - \mathcal{L}_n(L, \underset{\sim}{\boldsymbol{\beta}}^*_{L}) - \big[\mathcal{L}(L, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) - \mathcal{L}(L, \underset{\sim}{\boldsymbol{\beta}}^*_{L})\big]\big| \\
&+ \big|\mathcal{L}_n(J_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) - \mathcal{L}_n(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2}) - \big[\mathcal{L}(J_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) - \mathcal{L}(J_2, \underset{\sim}{\boldsymbol{\beta}}^*_{J_2})\big]\big| + Cd_0\kappa\sqrt{n\log(n \vee p)},
\end{aligned}$$

the rest follows from the similar arguments as in proof of Lemma 8. Hence, $\min\{L, J_2\} = \eta_p - v \leq \epsilon_n$.

Appendix C. Proofs of useful lemmas in Appendix A

Appendix C.1. Proof of Lemma 4

In this section, we prove Lemma 4. For $q \geq 0$, we prove the results by assuming there is q change points, i.e., $I = (s, \eta_{r+1}] \cup \ldots \cup (\eta_{r+q}, e] = I_1 \cup \ldots \cup I_{q+1}$. Note that there is not change point in I if $q = 0$. The proof is very similar to the proof of Theorem 4.3 in [41], which is omitted. Hence, only the change point case is proofed, i.e. $q \geq 1$. Denote $\widetilde{\lambda} = \lambda\sqrt{\max\{|I|, \log(n \vee p)\}}$ and $\underset{\sim}{\boldsymbol{\beta}}^*_I = (\boldsymbol{\beta}_I^{*\top}, \boldsymbol{b}_I^{*\top})^\top \in \mathbb{R}^{p+K}$ defined in (4.1). By the first order condition, $\underset{\sim}{\boldsymbol{\beta}}^*_I$ satisfies the following equation for $k = 1, \ldots, K$:

$$\mathbb{E}\big[\sum_{i\in I}\sum_{k=1}^{K} \boldsymbol{X}_i\big(\mathbb{I}\{Y_i \leq \boldsymbol{X}_i^\top\boldsymbol{\beta}^*_I + b^*_k\} - \tau_k\big)\big] = \mathbf{0}_p, \text{ and } \mathbb{E}\big[\sum_{i\in I}\big(\mathbb{I}\{Y_i \leq \boldsymbol{X}_i^\top\boldsymbol{\beta}^*_I + b^*_k\} - \tau_k\big)\big] = 0. \tag{C.1}$$

By the fact that $Y_i = \boldsymbol{X}_i^\top\boldsymbol{\beta}_i^0 + \varepsilon_i, i \in I$, for $k = 1, \ldots, K$, (C.1) have:

$$\sum_{i\in I}\mathbb{E}\big[\sum_{k=1}^{K} \boldsymbol{X}_i\left(F_\varepsilon\big(\boldsymbol{X}_i^\top(\boldsymbol{\beta}^*_I - \boldsymbol{\beta}_i^0) + b^*_k\big) - F_\varepsilon(b_k^0)\right)\big] = \mathbf{0}_p, \text{ and } \sum_{i\in I}\mathbb{E}\big[\left(F_\varepsilon\big(\boldsymbol{X}_i^\top(\boldsymbol{\beta}^*_I - \boldsymbol{\beta}_i^0) + b^*_k\big) - F_\varepsilon(b_k^0)\right)\big] = 0.$$

Moreover, let $\underset{\sim}{\boldsymbol{\beta}}_i^0 := (\boldsymbol{\beta}_i^{0\top}, \boldsymbol{b}^{0\top})^\top \in \mathbb{R}^{p+K}, i \in I$, $\underset{\sim}{\boldsymbol{X}}_i := (\boldsymbol{X}_i^\top, \mathbf{1}_K^\top)^\top \in \mathbb{R}^{p+K}$, and $\boldsymbol{S}_k := \mathrm{diag}(\mathbf{1}_p, \mathbf{e}_k)$, where $\mathbf{e}_k \in \mathbb{R}^K$ is a vector with all zero entries except the k-th entry is 1, and $\mathbf{1}_K = (1, \ldots, 1)^\top \in \mathbb{R}^K$. With above notations, for the above equations, we have $\sum_{i\in I}\mathbb{E}\big[\sum_{k=1}^{K}\boldsymbol{S}_k\underset{\sim}{\boldsymbol{X}}_i\left(F_\varepsilon\big((\boldsymbol{S}_k\underset{\sim}{\boldsymbol{X}}_i)^\top(\underset{\sim}{\boldsymbol{\beta}}^*_I - \underset{\sim}{\boldsymbol{\beta}}_i^0) + b_k^0\big) - F_\varepsilon(b_k^0)\right)\big] = \mathbf{0}_{p+K}$, where $\mathbf{0}_{p+K} = (0, \ldots, 0)^\top \in \mathbb{R}^{p+K}$. Furthermore, by the Taylor's expansion, we have

$$\sum_{i\in I}\Big\{\sum_{k=1}^{K}\mathbb{E}\big[\int_0^1 (\boldsymbol{S}_k\underset{\sim}{\boldsymbol{X}}_i)(\boldsymbol{S}_k\underset{\sim}{\boldsymbol{X}}_i)^\top f_\varepsilon\big(b_k^0 + z(\boldsymbol{S}_k\underset{\sim}{\boldsymbol{X}}_i)^\top(\underset{\sim}{\boldsymbol{\beta}}^*_I - \underset{\sim}{\boldsymbol{\beta}}_i^0)\big)dz\big]\Big\}(\underset{\sim}{\boldsymbol{\beta}}^*_I - \underset{\sim}{\boldsymbol{\beta}}_i^0) := \sum_{i\in I}\widetilde{\boldsymbol{\Sigma}}_i(I)(\underset{\sim}{\boldsymbol{\beta}}^*_I - \underset{\sim}{\boldsymbol{\beta}}_i^0) = \mathbf{0}_{p+K},$$

Hence, for $\underset{\sim}{\boldsymbol{\beta}}^*_I$, by defining $\widetilde{\boldsymbol{\Sigma}}_i(I), i \in I$, it has the following explicit form

$$\underset{\sim}{\boldsymbol{\beta}}^*_I = \big(\sum_{i\in I}\widetilde{\boldsymbol{\Sigma}}_i(I)\big)^{-1}\big(\sum_{i\in I}\widetilde{\boldsymbol{\Sigma}}_i(I)\underset{\sim}{\boldsymbol{\beta}}_i^0\big). \tag{C.2}$$

For example, when $T = 2$, i.e., there is a change point in $I = (s, \eta] \cup (\eta, e] = I_1 \cup I_2$, we have

$$\underset{\sim}{\boldsymbol{\beta}}_I^* = \left[|I_1| \widetilde{\boldsymbol{\Sigma}}_\eta(I) + |I_2| \widetilde{\boldsymbol{\Sigma}}_{\eta+1}(I) \right]^{-1} \left[|I_1| \widetilde{\boldsymbol{\Sigma}}_\eta(I) \underset{\sim}{\boldsymbol{\beta}}_\eta^0 + |I_2| \widetilde{\boldsymbol{\Sigma}}_{\eta+1}(I) \underset{\sim}{\boldsymbol{\beta}}_{\eta+1}^0 \right].$$

So far, we have derived the explicit form for $\underset{\sim}{\boldsymbol{\beta}}_I^*$, which is very important for proving Lemma 4. Now, we are ready to give the detailed proof. For simplicity, we omit the subscript I whenever needed and denote $\mathcal{J} := J(\underset{\sim}{\boldsymbol{\beta}}_I^*)$ in this section. Note that $|\mathcal{J}| \le d_0$ by the Assumption 1 and (C.2). For any $\underset{\sim}{\boldsymbol{\beta}}_I \in \mathbb{R}^{p+K}$, let $\underline{\boldsymbol{\Delta}}_I = \underset{\sim}{\boldsymbol{\beta}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*$, we define

$$\mathcal{A}_I := \left\{ (\boldsymbol{\Delta}_I^\top, \boldsymbol{\delta}_I^\top)^\top : \|\boldsymbol{\Delta}_{I,\mathcal{J}^c}\|_1 \le 3\|\boldsymbol{\Delta}_{I,\mathcal{J}}\|_1 + \|\boldsymbol{\delta}_I\|_1 \right\}.$$

where $\underline{\boldsymbol{\Delta}}_I = (\boldsymbol{\Delta}_I^\top, \boldsymbol{\delta}_I^\top)^\top$ with $\boldsymbol{\Delta}_I \in \mathbb{R}^p$ and $\boldsymbol{\delta}_I \in \mathbb{R}^K$. Let $\widehat{\underline{\boldsymbol{\Delta}}}_I = \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*$, where $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I$ is minimizer of the empirical loss defined in (2.2). The proof of Lemma 4 relies on the following three lemmas. The first lemma shows that under a suitably chosen λ , $\widehat{\underline{\boldsymbol{\Delta}}}_I := \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*$ lies in a restricted set $\mathcal{A}_I$ with probability tending to 1.

Lemma 15. For Model (1.1), under Assumption 1-3, for any interval $I = (s, e]$, and $\lambda \ge C_\lambda \sqrt{\log(n \vee p)}$, with $C_\lambda \ge 0$ being a large enough absolute constant, we have

$$\mathbb{P}\{(\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*) \in \mathcal{A}_I\} \ge 1 - 2K(n \vee p)^{-c_1},$$

where $c_1 > 0$ is an absolute constant depending on C_λ.

Before presenting the next lemma, we first state its additional required assumption. As pointed out put by [15] need to hold only locally around "true" parameters $\underset{\sim}{\boldsymbol{\beta}}_I^*$. Assumption 4 is similar to the restricted nonlinearity assumption in [6, 15].

Assumption 4 (Restricted Nonlinearity)**.** There exists a constant r_* such that for all $\underset{\sim}{\boldsymbol{\beta}} = (\boldsymbol{\beta}^\top, \boldsymbol{b}^\top)^\top \in \mathbb{R}^{p+K}$, we have

$$\inf_{\underset{\sim}{\boldsymbol{\beta}} \in \mathcal{G}(\underset{\sim}{\boldsymbol{\beta}}^*, r_*)} \frac{\sum_{k=1}^K \mathbb{E}[|\boldsymbol{X}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + b_k - b_k^*|^2]^{3/2}}{\sum_{k=1}^K \mathbb{E}[|\boldsymbol{X}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + b_k - b_k^*|^3]} \ge r_* \frac{2\bar{f}'}{3\underline{f}},$$

where the "prediction balls" with radius r and corresponding centers as follows:

$$\mathcal{G}(\underset{\sim}{\boldsymbol{\beta}}^*, r) = \left\{ \underset{\sim}{\boldsymbol{\beta}} = (\boldsymbol{\beta}^\top, \boldsymbol{b}^\top)^\top \in \mathbb{R}^{p+K} : \sum_{k=1}^K \mathbb{E}[|\boldsymbol{X}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + b_k - b_k^*|^2] \le r^2 \right\}.$$

The second lemma shows that $\mathcal{L}(I, \cdot)$ has a local quadratic curvature in the a neighborhood around $\underset{\sim}{\boldsymbol{\beta}}_I^*$ in the terms of norm $\|\cdot\|_{\mathbf{S}}$. We first define the following notations:

$$\mathbf{S} := \sum_{k=1}^K \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \operatorname{diag}(\mathbf{e}_k) \end{pmatrix} \in \mathbb{R}^{(p+K)\times(p+K)}, \quad \|\underline{\boldsymbol{\Delta}}_I\|_{\mathbf{S}}^2 = \sum_{k=1}^K \left(\boldsymbol{\Delta}_I^\top \Sigma \boldsymbol{\Delta}_I + \delta_{I,k}^2 \right).$$

Lemma 16. Define $Q(I, \underline{\boldsymbol{\Delta}}_I) := \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^* + \underline{\boldsymbol{\Delta}}_I) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^*)$. For Model (1.1), under Assumptions 3 - 4, for any interval $I = (s, e]$, we have

$$Q(I, \underline{\boldsymbol{\Delta}}_I) \ge \underline{f}|I|/(4K) \min\left\{ \|\underline{\boldsymbol{\Delta}}_I\|_{\mathbf{S}}^2, r_* \|\underline{\boldsymbol{\Delta}}_I\|_{\mathbf{S}} \right\},$$

where r_* is some constant not depending on n and p.

The third lemma bounds the difference between $\mathcal{L}_n(I, \cdot)$ and $\mathcal{L}(I, \cdot)$. we use empirical process theory to obtain a tail probability for the maximum over neighborhood around $\underset{\sim}{\boldsymbol{\beta}}_I^*$.

Lemma 17. For Model (1.1), under Assumptions 1-3, for any interval $I = (s, e]$, and $\lambda \ge C_\lambda \sqrt{\log(n \vee p)}$, with $C_\lambda \ge 0$ being a large enough absolute constant, we have $\mathbb{P}\{\mathcal{B}_I\} \ge 1 - 6(n \vee p)^{-c_2}$, where

$$\mathcal{B}_I = \Big\{ \sup_{\underline{\boldsymbol{\Delta}}_I \in \mathcal{A}_I, \|\underline{\boldsymbol{\Delta}}_I\|_{\mathbf{S}} \le r} \left| \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^* + \underline{\boldsymbol{\Delta}}_I) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) - [\mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^* + \underline{\boldsymbol{\Delta}}_I) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^*)] \right| \le \widetilde{\lambda} \sqrt{d_0} r \Big\},$$

where $c_2 > 0$ is an absolute constant depending on $\underline{f}, \underline{\rho}, \mathfrak{M}, K$.

Define the following event $\mathcal{E}_I = \{(\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*) \in \mathcal{A}_I\}$. With the above lemmas, we are ready to prove Lemma 4. By Lemmas 15 and 17, when $\lambda \ge C_\lambda \sqrt{\log(n \vee p)}$, we have $\mathbb{P}(\mathcal{B}_I^c \cup \mathcal{E}_I^c) \le 6(n \vee p)^{-c_1} + 2K(n \vee p)^{-c_2}$. The following derivation is based on the condition that events $\mathcal{B}_I$ and $\mathcal{E}_I$. Let $\|\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_{\mathbf{S}} = r$, as

$$r^2 = \|\underline{\boldsymbol{\Delta}}_I\|_{\mathbf{S}}^2 = \sum_{k=1}^K (\|\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*\|_{\boldsymbol{\Sigma}}^2 + (\widehat{b}_{I,k} - b_{I,k}^*)^2) \ge (K\underline{\rho}\|\widehat{\boldsymbol{\Delta}}_I\|_2^2 + \|\widehat{\boldsymbol{\delta}}_I\|_2^2),$$

we have

$$\|\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*\|_2 \le r/(K\underline{\rho})^{1/2} \text{ and } \quad \|\widehat{\boldsymbol{b}}_I - \boldsymbol{b}_I^*\|_2 \le r.$$

Therefore, it holds on the event $\mathcal{E}_I$

$$\|\boldsymbol{\beta}_I^*\|_1 - \|\widehat{\boldsymbol{\beta}}_I\|_1 \le \|\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*\|_1 \le 4\|(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*)_{\mathcal{J}}\|_1 + \|\widehat{\boldsymbol{b}}_I - \boldsymbol{b}_I^*\|_1/K \le 4d_0^{1/2}\|\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*\|_2 + \|\widehat{\boldsymbol{b}}_I - \boldsymbol{b}_I^*\|_2/K^{1/2} \le C_k d_0^{1/2} r. \quad \text{(C.3)}$$

where $C_k > 0$ is some universal constant. By the optimality of $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I$ and $|I| \ge \log(n \vee p)$, we have $\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) + \lambda\sqrt{|I|}(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I^*) \le 0$, which leads to

$$\mathcal{L}(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) \le \left|\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) - [\mathcal{L}(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^*)]\right| + \lambda\sqrt{|I|}(\|\boldsymbol{\beta}_I^*\|_1 - \|\widehat{\boldsymbol{\beta}}_I\|_1).$$

Note that $r_n \le C_k\sqrt{s}r$ with $C_k = 2(2\rho^{-1/2} + 1)/\sqrt{K}$ if $\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^* \in \mathcal{A}_I$. On the events $\mathcal{B}_I$ and $\mathcal{E}_I$, we have

$$\left|\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) - [\mathcal{L}(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^*)]\right| \le \lambda\sqrt{d_0|I|}r,$$

it follows that

$$\underline{f}\min\left\{|I|r^2/4, |I|r_*r/4\right\} \le \lambda\sqrt{d_0|I|}r + \lambda\sqrt{|I|}(\|\boldsymbol{\beta}_I^*\|_1 - \|\widehat{\boldsymbol{\beta}}_I\|_1)C_k\lambda r\sqrt{d_0|I|},$$

where the second inequality follows from (C.3). For the above inequality, it implies either

$$\underline{f}r_*|I|r/4 \le C_k C_\lambda r\sqrt{d_0|I|\log(n \vee p)}, \quad \text{(C.4)}$$

or

$$\underline{f}|I|r^2/4 \le C_k C_\lambda r\sqrt{d_0|I|\log(n \vee p)}. \quad \text{(C.5)}$$

By $|I| > C_I d_0\log(n \vee p)$ with C_I is an absolute constant depending on all the other absolute constants. (C.4) cannot hold. Hence (C.5) must hold, which implies that

$$\|\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_{\mathbf{S}} \le C\lambda\sqrt{d_0}/\sqrt{|I|}.$$

Lastly, by the above inequality and some trivial calculations, we can directly derive Lemma 4.

Appendix C.2. Proof of lemma 7

For simplicity, we omit the subscript I whenever needed. **Case 1**: If $|I| \ge C_I d_0\log(n \vee p)$, then $|I| > \log(n \vee p)$. With probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$, we have that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*) \le \lambda\sqrt{|I|}(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \le \lambda\sqrt{|I|}\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 \le C_1\lambda^2 d_0,$$

where the first inequality follows from the definition of $\widehat{\underset{\sim}{\boldsymbol{\beta}}}$ and the second is due to Lemma 4.
Case 2: If $|I| < C_I d_0\log(n \vee p)$, then

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*) \le \lambda\sqrt{\max\{|I|, \log(n \vee p)\}}\|\boldsymbol{\beta}^*\|_1 \le C_4\lambda d_0^{3/2}\sqrt{\log(n \vee p)},$$

where the first inequality follows from the definition of $\widehat{\underset{\sim}{\boldsymbol{\beta}}}$ and the last is due to $|I| \lesssim d_0\log(n \vee p)$ and $\|\boldsymbol{\beta}^*\|_1 \le d_0 C_\beta$.

Appendix C.3. Proof of lemma 8

We prove by contradiction, assuming that $\min\{|I_1|, |I_2|\} > C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2 > C_I d_0\log(n \vee p)$, where the second inequality follows from the observation that $\kappa^2 \le 4d_0 C_\beta^2$. Then we also have $\min\{|I_1|, |I_2|\} > \log(n \vee p)$. It follows from Lemma 7 and (A.1) that, with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that

$$\mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) = \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I) \le \mathcal{L}_n(I_1, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_1}) + \mathcal{L}_n(I_2, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_2}) + \gamma \le \mathcal{L}_n(I_1, \underset{\sim}{\boldsymbol{\beta}}_{I_1}^*) + \mathcal{L}_n(I_2, \underset{\sim}{\boldsymbol{\beta}}_{I_2}^*) + \gamma + 2C_1\lambda^2 d_0. \quad \text{(C.6)}$$

Denote $\underset{\sim}{\boldsymbol{\Delta}}_i = (\boldsymbol{\Delta}_i^\top, \boldsymbol{\delta}_i^\top)^\top = \widehat{\underset{\sim}{\boldsymbol{\beta}}}_I - \underset{\sim}{\boldsymbol{\beta}}_{I_i}^*$ for $i = 1, 2$. Let $\|\underset{\sim}{\boldsymbol{\Delta}}_i\|_{\mathbf{S}} = r_i$, as

$$r_i^2 = \|\underset{\sim}{\boldsymbol{\Delta}}_i\|_{\mathbf{S}}^2 = \sum_{k=1}^K(\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}^2 + (\delta_{i,k})^2) \ge (K\underline{\rho}\|\boldsymbol{\Delta}_i\|_2^2 + \|\boldsymbol{\delta}_i\|_2^2),$$

we have

$$\|\boldsymbol{\Delta}_i\|_2 = \|\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_{I_i}^*\|_2 \le r_i/(K\underline{\rho})^{1/2} \quad \text{and} \quad \|\boldsymbol{\delta}_i\|_2 = \|\widehat{\boldsymbol{b}}_I - \boldsymbol{b}_{I_i}^*\|_2 \le r_i. \quad \text{(C.7)}$$

The proof of Lemma 8 relies on the following lemma. According to Lemma 18, when $\underset{\sim}{\boldsymbol{\Delta}}_i \notin \mathcal{A}_{I_i}$, we can control the difference between the excess risk and its empirical version by scaling λ with a $\sqrt{d_0}$ factor.

Lemma 18. For Model (1.1), under Assumption 1-4, suppose there exists no true change point in the interval I. For any interval $J \supset I$, with $\lambda \ge \lambda_2 := C_\lambda\sqrt{d_0\log(n \vee p)}$, with $C_\lambda \ge 0$ being a large enough absolute constant, we have at least with probability $1 - 6(n \vee p)^{-c_2}$ hold

$$\sup_{\|\widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_J\|_{\mathbf{S}} \le r} \left| \mathcal{L}_n(I, \widehat{\underset{\sim}{\beta}}_J) - \mathcal{L}_n(I, \underset{\sim}{\beta}^*_I) - [\mathcal{L}(I, \widehat{\underset{\sim}{\beta}}_J) - \mathcal{L}(I, \underset{\sim}{\beta}^*_I)] \right| \le \lambda \sqrt{d_0} r \sqrt{\max\{|I|, \log(n \vee p)\}}.$$

Then (C.6) leads to

$$\begin{aligned}
\sum_{i\in\{1,2\}} \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] &\le \sum_{i\in\{1,2\}} \left| \mathcal{L}_n(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}_n(I_i, \underset{\sim}{\beta}^*_{I_i}) - \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] \right| + \gamma + 2C_1\lambda^2 d_0 \\
&\le \lambda \sqrt{d_0} \sum_{i\in\{1,2\}} \sqrt{|I_i|} \|\underset{\sim}{\Delta}_I\|_{\mathbf{S}} + \gamma + 2C_1\lambda^2 d_0 \\
&\le \underline{f}/(8K) \sum_{i\in\{1,2\}} |I_i| \|\underset{\sim}{\Delta}_i\|_{\mathbf{S}}^2 + \gamma + (2C_1 + 4K\underline{f}^{-1})\lambda^2 d_0, \qquad \text{(C.8)}
\end{aligned}$$

where the second inequality is by (D.7) and Lemma 18 with $\lambda \ge \max\{\lambda_1, \lambda_2\}$, and the last inequality follows $ab \le a^2/8 + 2b^2$. In addition, due to Lemma 16 and $\min\{|I_1|, |I_2|\} \gtrsim \log(n \vee p)$, it holds that

$$\sum_{i\in\{1,2\}} \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] \ge \underline{f}/(4K) \sum_{i\in\{1,2\}} |I_i| \|\underset{\sim}{\Delta}_i\|_{\mathbf{S}}^2. \qquad \text{(C.9)}$$

Combining (C.8) and (C.9),

$$\gamma + (2C_1 + 4K\underline{f}^{-1})\lambda^2 d_0 \ge \underline{f}/(8K) \sum_{i\in\{1,2\}} |I_i| \|\underset{\sim}{\Delta}_i\|_{\mathbf{S}}^2 \ge \underline{\rho}\,\underline{f}/8 \sum_{i\in\{1,2\}} |I_i| \|\boldsymbol{\Delta}_i\|_2^2 \ge \underline{\rho}\,\underline{f}/16\kappa^2 \min\{|I_1|, |I_2|\},$$

where the second inequality follows from (C.7), the last inequality from the fact

$$|I_1| \|\boldsymbol{\Delta}_1\|_2^2 + |I_2| \|\boldsymbol{\Delta}_2\|_2^2 \ge \inf_{v\in\mathbb{R}^p} \{|I_1| \|\boldsymbol{\beta}^0_\eta - v\|_2^2 + |I_2| \|\boldsymbol{\beta}^0_{\eta+1} - v\|_2^2\} \ge \kappa^2 |I_1||I_2|/|I| \ge \kappa^2 \min\{|I_1|, |I_2|\}/2.$$

Therefore $\min\{|I_1|, |I_2|\} \le C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$, which is a contradiction with $\min\{|I_1|, |I_2|\} > C_I d_0 \log(n \vee p)$.

Appendix C.4. Proof of lemma 9

First we notice that with the choice of λ, it holds that $\lambda \ge \max\{\lambda_1, \lambda_2\}$, and therefore we can apply Lemmas 4, 7, and 18 when needed. By symmetry, it suffices to show that $|I_1| \le C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$. We prove by contradiction, assuming that

$$|I_1| > C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2 > C_I d_0 \log(n \vee p),$$

where the second inequality follows from the observation that $\kappa^2 \le 4d_0 C_\beta^2$. Therefore we have $|I_1| > \log(n \vee p)$. Denote $\underset{\sim}{\Delta}_i = (\boldsymbol{\Delta}_i^\top, \delta_i^\top)^\top = \widehat{\underset{\sim}{\beta}}_I - \underset{\sim}{\beta}^*_{I_i}$ for $i \in \{1, 2, 3\}$. We then consider the following two cases.

Case 1. If $|I_3| > C_I d_0 \log(n \vee p)$, then $|I_3| > \log(n \vee p)$. It follows from Lemma 7 and (A.2) that, with probability at least $1 - 2K(n \vee p)^{-c_0} - 6(n \vee p)^{-c_1}$ that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\beta}}_I) \le \sum_{i\in\{1,2,3\}} \mathcal{L}_n(I_i, \widehat{\underset{\sim}{\beta}}_{I_i}) + 2\gamma \le \sum_{i\in\{1,2,3\}} \mathcal{L}_n(I_i, \underset{\sim}{\beta}^*_{I_i}) + 2\gamma + 3C_1\lambda^2 d_0,$$

which implies that

$$\begin{aligned}
\sum_{i\in\{1,2,3\}} \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] &\le \sum_{i\in\{1,2,3\}} \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] + 2\gamma + 3C_1\lambda^2 d_0 \\
&\le \lambda \sum_{i\in\{1,2,3\}} \sqrt{d_0 |I_i|} \|\underset{\sim}{\Delta}_i\|_{\mathbf{S}} + 2\gamma + 3C_1\lambda^2 d_0,
\end{aligned}$$

where the last inequality is by Lemma 18. It follows from identical arguments in proof of Lemma 8 that, with probability at least $1 - C(n \vee p)^{-c}$ hold that $\min\{|I_1|, |I_2|\} \le C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$. Since $|I_2| > C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$ by assumption, it follows from (4.2) that $|I_1| \le C_\epsilon(\lambda^2 d_0 + \gamma)/\kappa^2$, which contradicts the fact that $|I_1| > C_I d_0 \log(n \vee p)$.

Case 2. If $|I_3| \le C_I d_0 \log(n \vee p)$, then it follows from Lemma 7 and (A.2) that the following holds with probability at least $1 - 2K(n \vee p)^{-c_1} - 6(n \vee p)^{-c_2}$ that

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\beta}}_I) \le \sum_{i\in\{1,2,3\}} \mathcal{L}_n(I_i, \widehat{\underset{\sim}{\beta}}_{I_i}) + 2\gamma \le \sum_{i\in\{1,2,3\}} \mathcal{L}_n(I_i, \underset{\sim}{\beta}^*_{I_i}) + 2\gamma + 2C_1\lambda^2 d_0 + C_4\lambda d_0^{3/2} \sqrt{\log(n \vee p)},$$

which implies that

$$\begin{aligned}
&\sum_{i\in\{1,2,3\}} \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] \le \sum_{i\in\{1,2,3\}} \left[\mathcal{L}(I_i, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_i, \underset{\sim}{\beta}^*_{I_i})\right] + 2\gamma + 2C_1\lambda^2 d_0 + C_4\lambda d_0^{3/2} \sqrt{\log(n \vee p)} \\
\le &\lambda \sum_{i\in\{1,2\}} \sqrt{d_0 |I_i|} \|\underset{\sim}{\Delta}_i\|_{\mathbf{S}} + \left| \mathcal{L}(I_3, \widehat{\underset{\sim}{\beta}}_I) - \mathcal{L}(I_3, \underset{\sim}{\beta}^*_{I_3}) \right| + 2\gamma + (4C_1 + C_5)\lambda^2 d_0 + C_4\lambda d_0^{3/2} \sqrt{\log(n \vee p)},
\end{aligned}$$

where the second inequality is by Lemma 17 and the last inequality is by Lemma 12. The above inequality leads to

$$\sum_{i\in\{1,2\}}\left[\mathcal{L}(I_i,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_i,\underset{\sim}{\boldsymbol{\beta}}^*_{I_i})\right]\le\lambda\sum_{i\in\{1,2,3\}}\sqrt{d_0|I_i|}\|\underset{\sim}{\boldsymbol{\Delta}}_i\|_{\mathbf{S}}+2\gamma+(4C_1+C_5)\lambda^2d_0+C_4\lambda d_0^{3/2}\sqrt{\log(n\vee p)}.$$

The rest follows from the same arguments as in **Case 1**.

Appendix C.5. Proof of lemma 10

First we notice that with the choice of λ, it holds that $\lambda\ge\max\{\lambda_1,\lambda_2\}$, and therefore we can apply Lemmas 4, 7, 16 and 18 when needed. We prove by contradiction, assuming that $\mathcal{L}_n(I,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)\le\sum_{t=1}^{T+1}\mathcal{L}_n(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_t})+T\gamma$. Denote $\underset{\sim}{\boldsymbol{\Delta}}_i=(\boldsymbol{\Delta}_i^\top,\delta_i^\top)^\top=\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I-\underset{\sim}{\boldsymbol{\beta}}^*_{I_i}$ for $i=1,\dots,T+1$. It then follows from Lemma 7 with probability at least $1-2K(n\vee p)^{-c_1}-6(n\vee p)^{-c_2}$ that

$$\mathcal{L}_n(I,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)\le\sum_{t=1}^{T+1}\mathcal{L}_n(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_{I_t})+T\gamma\le\sum_{t=1}^{T+1}\mathcal{L}_n(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})+T\gamma+(T+1)C_\gamma d_0^2\log(n\vee p),$$

which implies that

$$\sum_{t=1}^{T+1}\left\{\mathcal{L}(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})\right\}\le\sum_{t=1}^{T+1}\left|\mathcal{L}_n(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}_n(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})-\left[\mathcal{L}(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})\right]\right|+T\gamma+(T+1)C_\gamma d_0^2\log(n\vee p).\tag{C.10}$$

Step 1. For any $t\in\{2,\dots,T\}$, it follows from (4.2) that $|I_t|\ge C_I d_0\log(n\vee p)$. It holds that

$$\begin{aligned}&\left|\mathcal{L}_n(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}_n(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})-\left[\mathcal{L}(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})\right]\right|\\ \le&\lambda\sqrt{d_0|I_t|}\|\underset{\sim}{\boldsymbol{\Delta}}_t\|_{\mathbf{S}}+C_1\lambda^2d_0\le\underline{f}/(8K)\left(|I_t|\|\underset{\sim}{\boldsymbol{\Delta}}_t\|_{\mathbf{S}}^2\right)+(2K/\underline{f}+C_1)\lambda^2d_0,\end{aligned}\tag{C.11}$$

where the first inequality is by Lemma 17 and the last inequality follows $ab\le a^2/8+2b^2$. In addition, due to Lemma 16, it holds that

$$\mathcal{L}(I_1,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_1,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})\ge\underline{f}|I_t|\|\underset{\sim}{\boldsymbol{\Delta}}_t\|_{\mathbf{S}}^2/(4K).\tag{C.12}$$

Step 2. We then discuss the intervals I_1 and I_{T+1}. These two will be treated in the same way, and therefore for $J\in\{I_1,I_{T+1}\}$ and $j\in\{1,T+1\}$, we have the following arguments. If $|J|\ge C_I d_0\log(n\vee p)$, then due to the same arguments in **Step 1**, (C.11) and (C.12) hold. If instead, $|J|<C_I d_0\log(n\vee p)$ holds, then by Lemma 12,

$$\sum_{t\in\{1,T+1\}}\left|\mathcal{L}_n(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}_n(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})-\left[\mathcal{L}(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})\right]\right|\le\sum_{t\in\{1,T+1\}}\left|\mathcal{L}(I_t,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_I)-\mathcal{L}(I_t,\underset{\sim}{\boldsymbol{\beta}}^*_{I_t})\right|+2C_5\lambda^2d_0.$$

Therefore, it follows from (C.10) that

$$\underline{\rho}\,\underline{f}/8\sum_{t=2}^{T}|I_t|\|\boldsymbol{\Delta}_t\|_2^2\le T\gamma+2(T+1)C\max\{\lambda^2d_0,\lambda d_0^{3/2}\log(n\vee p)\}.$$

Step 3. Since for any $t\in\{2,\dots,T-1\}$, it holds that

$$|I_t|\|\boldsymbol{\Delta}_t\|_2^2+|I_{t+1}|\|\boldsymbol{\Delta}_{t+1}\|_2^2\ge\inf_{v\in\mathbb{R}^p}\left\{|I_t|\|\boldsymbol{\beta}^0_{\eta_t}-v\|^2+|I_{t+1}|\|\boldsymbol{\beta}^0_{\eta_{t+1}}-v\|^2\right\}\ge\kappa^2|I_t||I_{t+1}|/(|I_t|+|I_{t+1}|)\ge\kappa^2\min\{|I_t|,|I_{t+1}|\}/2.$$

It then follows from the same arguments in Lemma 8 that $\min_{t=2,\dots,T}|I_t|\le C_\epsilon\lambda^2d_0+\gamma/\kappa^2$, which is a contradiction to $|I_t|\ge C_I d_0\log(n\vee p)$.

Appendix C.6. Proof of lemma 11

First we notice that with the choice of λ, it holds that $\lambda\ge\lambda_1$, therefore we can apply Lemma 7 when needed. For any $a=s+1,\dots,e-1$, let $I_1=(s,a]$ and $I_2=(a,e]$. It follows from Lemma 7 that with probability at least $1-2K(n\vee p)^{-c_1}-6(n\vee p)^{-c_2}$,

$$\max_{J\in\{I_1,I_2,I\}}\left|\mathcal{L}_n(J,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J)-\mathcal{L}_n(J,\underset{\sim}{\boldsymbol{\beta}}^*_J)\right|\le\max\{C_1\lambda^2d_0,C_4\lambda d_0^{3/2}\log(n\vee p)\}\le\gamma/3.$$

Appendix C.7. Proof of lemma 12

Denote $\underset{\sim}{\boldsymbol{\Delta}}=(\boldsymbol{\Delta}^\top,\delta^\top)^\top=\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J-\underset{\sim}{\boldsymbol{\beta}}^*_I$. We then consider the following two cases. **Case 1.** If $|I|>C_I d_0\log(n\vee p)$, we then have that,

$$\begin{aligned}&\mathcal{L}_n(I,\underset{\sim}{\boldsymbol{\beta}}^*_I)-\mathcal{L}_n(I,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J)\le\left|\mathcal{L}_n(I,\underset{\sim}{\boldsymbol{\beta}}^*_I)-\mathcal{L}_n(I,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J)-\left[\mathcal{L}(I,\underset{\sim}{\boldsymbol{\beta}}^*_I)-\mathcal{L}(I,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J)\right]\right|-\left[\mathcal{L}(I,\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J)-\mathcal{L}(I,\underset{\sim}{\boldsymbol{\beta}}^*_I)\right]\\ \le&\lambda\sqrt{s|I|}\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}-\underline{f}|I|\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}^2/(4K)+C_1\lambda^2d_0\\ \le&K/\underline{f}\lambda^2d_0+\underline{f}|I|\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}^2/(4K)-\underline{f}|I|\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}^2/(4K)+C_1\lambda^2d_0\le C_5\lambda^2d_0,\end{aligned}$$

where the first inequality follows from simple algebraic transformation, the second inequality follows from Lemmas 4, 16 and 18, the third follows from $ab \le a^2 + b^2/4$, letting $a = \lambda\sqrt{d_0 K/f}$ and $b = \sqrt{f|I|/K}\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathfrak{S}}$.

Case 2. If $|I| \le C_I d_0 \log(n \vee p)$, then with probability at least $1 - 2K(n \vee p)^{-c_0} - 6(n \vee p)^{-c_1}$,

$$\begin{aligned}
&\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) - \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_J) \le \left|\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) - \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_J) - \left[\mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}_I^*) - \mathcal{L}(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}_J)\right]\right| \\
&\le \widetilde{\lambda}\|\underset{\sim}{\boldsymbol{\Delta}}\|_1 \le \widetilde{\lambda}\|\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J - \underset{\sim}{\boldsymbol{\beta}}_J^*\|_1 + \widetilde{\lambda}\|\underset{\sim}{\boldsymbol{\beta}}_J^* - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_1 \le \lambda|J|\|\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J - \underset{\sim}{\boldsymbol{\beta}}_J^*\|_1 + \lambda d_0^{1/2}\sqrt{\log(n \vee p)}\|\underset{\sim}{\boldsymbol{\beta}}_J^* - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_1 \\
&\le C_1\lambda^2 d_0 + C_4\lambda d_0^{3/2}\sqrt{\log(n \vee p)} \le 2\max\{C_1\lambda^2 d_0, C_4\lambda d_0^{3/2}\sqrt{\log(n \vee p)}\} \le C_5\lambda^2 d_0,
\end{aligned}$$

where the first inequality follows from simple algebraic transformation, the second inequality follows from Lemma 18, the third inequality follows from the fact $\|\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_1 \le \|\widehat{\underset{\sim}{\boldsymbol{\beta}}}_J - \underset{\sim}{\boldsymbol{\beta}}_J^*\|_1 + \|\underset{\sim}{\boldsymbol{\beta}}_J^* - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_1$, the fourth inequality follows from $\max\{|I|, \log(n \vee p)\} \le C_I d_0 \log(n \vee p) \le |J|$ and $\|\underset{\sim}{\boldsymbol{\beta}}_J^* - \underset{\sim}{\boldsymbol{\beta}}_I^*\|_1 \le 2d_0 C_\beta$.

Appendix D. Proof of additional lemmas and theorems

Appendix D.1. Proof of Proposition 13

Directly, Lemmas 8, 9, 10 and 11 respectively lead to four conclusions, namely *(i)*, *(ii)*, *(iii)*, and *(iv)*.

Appendix D.2. Proof of Proposition 14

The proof procedure closely aligns with the proof of Proposition 2 in [26], differing solely in our adoption of the loss function $\mathcal{L}_n(\cdot,\cdot)$ instead of squared loss. Consequently, we omit this part.

Appendix D.3. Proof of Lemma 15

For simplicity, we omit the subscript I whenever needed. Let us denote $\widetilde{\lambda} = \lambda\sqrt{\max\{|I|, \log(n \vee p)\}}$, $h_{ik}^* = \tau_k - \mathbb{I}\left(\varepsilon_i^* \le b_k^*\right)$, $\varepsilon_i^* = Y_i - X_i^\top \boldsymbol{\beta}^*$, $\boldsymbol{u} = (u_1, \ldots, u_p)^\top$ and $\boldsymbol{v} = (v_1, \ldots, v_K)^\top$, where

$$u_j = \partial\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*)/\partial\beta_j\Big|_{\underset{\sim}{\beta}_k = \underset{\sim}{\beta}_k^*} = \sum_{i\in I}\frac{1}{K}\sum_{k=1}^K X_{ij}h_{ik}^*, \quad \text{and } v_k = \partial\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*)/\partial b_k\Big|_{\underset{\sim}{\beta}_k = \underset{\sim}{\beta}_k^*} = \frac{1}{K}\sum_{i\in I}h_{ik}^*.$$

Note that $(\boldsymbol{u}^\top, \boldsymbol{v}^\top)^\top \in \partial\mathcal{L}_n(I, \boldsymbol{\beta}^*)$, where the subdifferential is taken with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$. In the following, we show that $\|\boldsymbol{u}\|_\infty$ and $\|\boldsymbol{v}\|_\infty$ are bounded with probability to 1. We first consider $\|\boldsymbol{v}\|_\infty$. In fact, we have

$$\begin{aligned}
\|\boldsymbol{v}\|_\infty &= \frac{1}{K}\max_{1\le k\le K}\Big|\sum_{i\in I}h_{ik}^*\Big| = \frac{1}{K}\max_{1\le k\le K}\Big|\sum_{i\in I}(h_{ik}^* - \mathbb{E}h_{ik}^*)\Big| \\
&= \frac{1}{K}\max_{1\le k\le K}\Big|\sum_{i\in I}\left(\mathbb{I}(\varepsilon_i \le X_i^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^*) - \mathbb{E}\left[F_\varepsilon\left(X_i^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^*\right)\right]\right)\Big|,
\end{aligned}$$

where second equality comes from the first order condition in (C.1). Let $z_{i,k} := \mathbb{I}(\varepsilon_i \le X_i^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^*) - \mathbb{E}\left[F_\varepsilon\left(X_i^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^*\right)\right]$. Note that $\mathbb{E}[z_{i,k}] = 0$ and $\left|z_{i,k}\right| \le 1$, by the Hoeffding's inequality, we have $\mathbb{P}\{\|\boldsymbol{v}\|_\infty \ge t_1/K\} \le 2K\exp(-2t_1^2/|I|)$. Taking $t_1 = 2\sqrt{\max\{|I|, \log(n \vee p)\}}\sqrt{\log(n \vee p)}$, we have

$$\mathbb{P}\{\|\boldsymbol{v}\|_\infty \ge \widetilde{\lambda}/(2K)\} \le 2K(n \vee p)^{-c_1}, \tag{D.1}$$

where c_1 is an absolute constant.

Next we consider $\|\boldsymbol{u}\|_\infty$. In fact, we have

$$\|\boldsymbol{u}\|_\infty = \max_{1\le j\le p}\Big|\sum_{i\in I}\frac{1}{K}\sum_{k=1}^K X_{ij}h_{ik}^*\Big| = \max_{1\le j\le p}\Big|\sum_{i\in I}\frac{1}{K}\sum_{k=1}^K\left(X_{ij}h_{ik}^* - \mathbb{E}\left[X_{ij}h_{ik}^*\right]\right)\Big| \le \max_{1\le j\le p}\max_{1\le k\le K}\Big|\sum_{i\in I}\left(X_{ij}h_{ik}^* - \mathbb{E}\left[X_{ij}h_{ik}^*\right]\right)\Big|,$$

where second equality comes from the first order condition in (C.1). Let $w_{ijk} := X_{ij}h_{ik}^* - \mathbb{E}\left[X_{ij}h_{ik}^*\right]$. Note that $\mathbb{E}[w_{ijk}] = 0$ and $\left|w_{ijk}\right| \le \mathfrak{M}$, by the Hoeffding's inequality, we have $\mathbb{P}\{\|\boldsymbol{u}\|_\infty \ge t_2/K\} \le 2pK\exp\left(-2t_2^2/(|I|\mathfrak{M})\right)$. Taking $t_2 = 2\mathfrak{M}\sqrt{\max\{|I|, \log(n \vee p)\}}\sqrt{\log(n \vee p)}$, we have

$$\mathbb{P}\{\|\boldsymbol{u}\|_\infty \ge \widetilde{\lambda}/(2K)\} \le 2K(n \vee p)^{-c_2}, \tag{D.2}$$

where c_2 is an absolute constant depending on $\mathfrak{M}$.

Therefore, Combining (D.1) and (D.2), and $\lambda \ge C_\lambda\sqrt{\max\{|I|, \log(n \vee p)\}}$, we have with probability at least $1 - 2K(n \vee p)^{-c_0}$ that

$$\|\boldsymbol{u}\|_\infty \le \widetilde{\lambda}/2, \quad \text{and} \quad \|\boldsymbol{v}\|_\infty \le \widetilde{\lambda}/(2K).$$

By convexity of $\mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}})$, we have

$$\mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*) \ge \boldsymbol{u}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \boldsymbol{v}^\top(\widehat{\boldsymbol{b}} - \boldsymbol{b}^*).$$

Combining the above inequality and by optimality of $\widehat{\underset{\sim}{\boldsymbol{\beta}}}$, we have with probability at least $1 - 2K(n \vee p)^{-c_0}$ that

$$
\begin{aligned}
0 \leq & \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*) - \mathcal{L}_n(I, \widehat{\underset{\sim}{\boldsymbol{\beta}}}) + \widetilde{\lambda}(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \\
\leq & \|\boldsymbol{u}\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \|\boldsymbol{v}\|_\infty \|\widehat{\boldsymbol{b}} - \boldsymbol{b}^*\|_1 + \widetilde{\lambda}(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \\
\leq & \widetilde{\lambda}/2 \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \widetilde{\lambda}/(2K) \|\widehat{\boldsymbol{b}} - \boldsymbol{b}^*\|_1 + \widetilde{\lambda}(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1).
\end{aligned}
$$

Canceling out $\widetilde{\lambda}$, we have

$$
0 \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1/2 + \|\widehat{\boldsymbol{b}} - \boldsymbol{b}^*\|_1/(2K) + (\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq 3/2 \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J(\boldsymbol{\beta}^*)}\|_1 + \|\widehat{\boldsymbol{b}} - \boldsymbol{b}^*\|_1/(2K) - \|\widehat{\boldsymbol{\beta}}_{J^c(\boldsymbol{\beta}^*)}\|_1/2,
$$

where the second inequality follows from $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J(\boldsymbol{\beta}^*)}\|_1 + \|\widehat{\boldsymbol{\beta}}_{J^c(\boldsymbol{\beta}^*)}\|_1$ and $\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{\beta}^*_{J(\boldsymbol{\beta}^*)}\|_1 - \|\widehat{\boldsymbol{\beta}}_{J(\boldsymbol{\beta}^*)}\|_1 - \|\widehat{\boldsymbol{\beta}}_{J^c(\boldsymbol{\beta}^*)}\|_1 \leq \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{J(\boldsymbol{\beta}^*)}\|_1 - \|\widehat{\boldsymbol{\beta}}_{J^c(\boldsymbol{\beta}^*)}\|$. This concludes the proof.

Appendix D.4. Proof of Lemma 16

For simplicity, we omit the subscript I whenever needed. By Knight identity, we have

$$
\begin{aligned}
Q(I, \underset{\sim}{\boldsymbol{\Delta}}) = \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}^* + \underset{\sim}{\boldsymbol{\Delta}}) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}^*) &= \sum_{i \in I} \frac{1}{K} \sum_{k=1}^K \mathbb{E}\big[(\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k) h_{ik}^* + \int_0^{\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k} \big(\mathbb{I}\{Y_i \leq \underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\beta}}_k^* + t\} - \mathbb{I}\{Y_i \leq \underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\beta}}_k^*\}\big) dt\big] \\
&=_{(1)} \sum_{i \in I} \frac{1}{K} \sum_{k=1}^K \mathbb{E}\big[\int_0^{\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k} F_\varepsilon\big(\boldsymbol{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^* + t\big) - F_\varepsilon\big(\boldsymbol{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^*\big) dt\big] \\
&=_{(2)} \sum_{i \in I} \frac{1}{K} \sum_{k=1}^K \mathbb{E}\big[\int_0^{\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k} t f_\varepsilon\big(\boldsymbol{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^*\big) + \frac{t^2}{2} f_\varepsilon'(\boldsymbol{X}^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i^0) + b_k^* + \widetilde{t}) dt\big] \\
&\geq_{(3)} \sum_{i \in I} \frac{1}{K} \sum_{k=1}^K \mathbb{E}\big[\underset{\sim}{f}/2 (\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k)^2 - \bar{f}'/6 |\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k|^3\big] = \sum_{i \in I} \frac{1}{K} \big[\sum_{k=1}^K \underset{\sim}{f}/2 \mathbb{E}|\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k|^2 - \sum_{k=1}^K \bar{f}'/6 \mathbb{E}|\underset{\sim}{\boldsymbol{X}}_i^\top \underset{\sim}{\boldsymbol{\Delta}}_k|^3\big], \quad \text{(D.3)}
\end{aligned}
$$

where equality (1) follows by the first order condition (C.1), equality (2) follows from Taylor's expansion with $0 < \widetilde{t} < t$, inequality (3) follows from Assumption 3.

We consider the following two cases: **(i)** when $\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}} \leq r_*$, it holds that (D.3)$\geq \underset{\sim}{f} |I| \|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}^2/(4K)$ from the fact

$$
\underset{\sim}{f}/4 \mathbb{E}[|\boldsymbol{X}^\top \boldsymbol{\Delta} + \delta_k|^2] \geq \bar{f}'/6 \mathbb{E}[|\boldsymbol{X}^\top \boldsymbol{\Delta} + \delta_k|^3].
$$

To see why the above inequality hold, note that by Assumption 4, for any $\underset{\sim}{\boldsymbol{\beta}} \in \mathcal{B}(\underset{\sim}{\boldsymbol{\beta}}^*, r_*)$,

$$
\sum_{k=1}^K \mathbb{E}[|\boldsymbol{X}^\top \boldsymbol{\Delta} + \delta_k|^2]^{3/2} / \sum_{k=1}^K \mathbb{E}[|\boldsymbol{X}^\top \boldsymbol{\Delta} + \delta_k|^3] \geq r_* 2\bar{f}'/(3\underset{\sim}{f}) \geq 2\bar{f}'/(3\underset{\sim}{f}) \sum_{k=1}^K \mathbb{E}[|\boldsymbol{X}^\top \boldsymbol{\Delta} + \delta_k|^2]^{1/2}.
$$

(ii) when $\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}} > r_*$, observe that Q is convex with respect to $\underset{\sim}{\boldsymbol{\Delta}}$ and $Q(I, 0) = 0$, moreover, $r_*/\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}} \in (0, 1)$. Therefore, by the convexity it holds that

$$
Q(I, r_* \underset{\sim}{\boldsymbol{\Delta}}/\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}) \leq (1 - r_*/\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}) Q(I, 0) + r_* \|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}} Q(I, \underset{\sim}{\boldsymbol{\Delta}}) = r_*/\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}} Q(I, \underset{\sim}{\boldsymbol{\Delta}}),
$$

if we let $\underset{\sim}{\boldsymbol{\Delta}}_0 = (r_*/\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}) \underset{\sim}{\boldsymbol{\Delta}}$, then $\|\underset{\sim}{\boldsymbol{\Delta}}_0\|_{\mathbf{S}} = r_*$. So by case **(i)**, we have $Q(I, \underset{\sim}{\boldsymbol{\Delta}}_0) \geq \underset{\sim}{f} |I| r_*^2/(4K)$. Then

$$
Q(I, \underset{\sim}{\boldsymbol{\Delta}}) \geq \|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}/r_* Q(I, \underset{\sim}{\boldsymbol{\Delta}}_0) \geq \underset{\sim}{f} r_* |I|/(4K) \|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}.
$$

Combining case **(i)** and **(ii)**, we get $Q(I, \underset{\sim}{\boldsymbol{\Delta}}) \geq \underset{\sim}{f} |I|/(4K) \min\big\{\|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}^2, r_* \|\underset{\sim}{\boldsymbol{\Delta}}\|_{\mathbf{S}}\big\}$, which concludes the proof.

Appendix D.5. Proof of Lemma 17

For simplicity, we omit the subscript I whenever needed. Define the following quantity:

$$
Q_i(\underset{\sim}{\boldsymbol{\Delta}}) = \frac{1}{K} \sum_{k=1}^K \big\{\rho_{\tau_k}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}^* - b_k^* - (\boldsymbol{X}_i^\top \boldsymbol{\Delta} + \delta_k)) - \rho_{\tau_k}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}^* - b_k^*)\big\}.
$$

It follows immediately that

$$
\big| \sum_{i \in I} [Q_i(\underset{\sim}{\boldsymbol{\Delta}}) - \mathbb{E} Q_i(\underset{\sim}{\boldsymbol{\Delta}})] \big| = \big| \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^* + \underset{\sim}{\boldsymbol{\Delta}}) - \mathcal{L}_n(I, \underset{\sim}{\boldsymbol{\beta}}^*) - [\mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}^* + \underset{\sim}{\boldsymbol{\Delta}}) - \mathcal{L}(I, \underset{\sim}{\boldsymbol{\beta}}^*)] \big|.
$$

Let us show that the check loss $\rho_{\tau_k}(\cdot)$ is Lipschitz continuous with Lipschitz constant $\max(\tau_k, 1 - \tau_k)$. To see it, note that for any $x, y \in \mathbb{R}$, we have $\rho_{\tau_k}(x) - \rho_{\tau_k}(y) \leq |x - y|$. Hence, for any $i \in I$,

$$
\mathrm{Var}[Q_i(\underset{\sim}{\boldsymbol{\Delta}}) - \mathbb{E} Q_i(\underset{\sim}{\boldsymbol{\Delta}})] \leq \mathrm{Var}[\frac{1}{K} \sum_{k=1}^K |\boldsymbol{X}_i^\top \boldsymbol{\Delta} + \delta_k|] \leq K \frac{1}{K^2} \sum_{k=1}^K \mathbb{E}[(\boldsymbol{X}_i^\top \boldsymbol{\Delta} + \delta_k)^2] \leq r^2/K.
$$

Now let $\xi_1, \ldots, \xi_n$ denote a sequence of *i.i.d* Rademacher random variables. By the symmetrization method (see, for example, Lemma 2.3.7 of [29]), we have

$$\mathbb{P}\Big\{ \sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} \Big| \sum_{i\in I} [Q_i(\underset{\sim}{\Delta}) - \mathbb{E}Q_i(\underset{\sim}{\Delta})] \Big| > t \Big\} \le 2\mathbb{P}\Big\{ \sup_{\|\underset{\sim}{\Delta}\|\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} \Big| \sum_{i\in I} \xi_i Q_i(\underset{\sim}{\Delta}) \Big| > t/4 \Big\} \Big/ \big(1 - 4\,|I|\, r^2/(Kt^2)\big) \Big\}, \quad \text{(D.4)}$$

For any positive sequences r_n satisfy $\frac{1}{K}\sum_{k=1}^K \|\underset{\sim}{\Delta}_k\|_1 = \|\underset{\sim}{\Delta}\|_1 = r_n$. Note that $r_n \le C_k \sqrt{d_0} r$ with $C_k = 2(2\underset{\sim}{\rho}^{-1/2} + 1)/\sqrt{K}$ if $\underset{\sim}{\Delta} \in \mathcal{A}$. We have for any $a > 0$,

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} \Big| \sum_{i\in I} \xi_i Q_i(\underset{\sim}{\Delta}) \Big| > t \Big\} \le_{(1)} e^{-at}\mathbb{E}\big[\exp\big(\sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} a \Big| \sum_{i\in I} \xi_i Q_i(\underset{\sim}{\Delta}) \Big| \big) \big] \\
\le_{(2)} & e^{-at}\mathbb{E}\big[\exp\big(\sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} 2a \Big| \sum_{i\in I} \frac{1}{K}\sum_{k=1}^K \xi_i (\underset{\sim}{X}_i \underset{\sim}{\Delta}_k) \Big| \big) \big] \le e^{-at}\mathbb{E}\big[\exp\big(\sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} 2a \| \sum_{i\in I} \xi_i \underset{\sim}{X}_i^\top \|_\infty \frac{1}{K}\sum_{k=1}^K \|\underset{\sim}{\Delta}_k\|_1 \big) \big] \\
\le \quad & e^{-at}\mathbb{E}\big[\exp\big(\sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} 2ar_n \| \sum_{i\in I} \xi_i \underset{\sim}{X}_i^\top \|_\infty \big) \big] \quad \le_{(3)} e^{-at}\mathbb{E}\big[\max_{1\le j\le p+1} \exp\big(2ar_n \Big| \sum_{i\in I} \xi_i X_{ij}^\top \Big| \big) \big] \\
\le_{(4)} & 2pe^{-at} \max_{1\le j\le p+1} \mathbb{E}\big[\exp\big(2ar_n \sum_{i\in I} \xi_i X_{ij}^\top \big) \big] \le_{(5)} 2p \exp\big(8a^2\mathfrak{M}^2 r_n^2\, |I| - at \big) \le 2p \exp(-t^2/(32a^2\mathfrak{M}^2 r_n^2\, |I|)), \quad \text{(D.5)}
\end{aligned}$$

where inequality (1) follows from Markov's inequality, (2) follows the contraction principle (Theorem 4.12 of [14]), (3) is from definition of $\|\cdot\|_\infty$, (4) comes from the fact that $\mathbb{E}[e^{|Z|}] = \mathbb{E}[\max\{e^Z, e^{-Z}\}] \le \mathbb{E}[e^Z] + \mathbb{E}[e^{-Z}] = 2\mathbb{E}[e^Z]$ for any symmetric random variable Z with $Z \overset{d}{=} -Z$, (5) follows Hoeffding's Lemma, $\mathbb{E}[e^{aZ}] \le \exp(a^2(d_2 - d_1)^2/2)$ for any random variable Z bounded between $[d_1, d_2]$, and the last inequality follow the fact that we take $a = t/(16\mathfrak{M}^2 r_n^2\, |I|)$ to minimize the $8a^2\mathfrak{M}^2 r_n^2\, |I| - at$.

Therefore, we have when $t = \max\{\sqrt{12} r(|I|/K)^{1/2}, 64\mathfrak{M} r_n (|I| \log(n \vee p))^{1/2}\}$,

$$\text{(D.4)} \le 3\mathbb{P}\Big\{ \sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} \Big| \sum_{i\in I} \xi_i Q_i(\underset{\sim}{\Delta}) \Big| > t/4 \Big\} \le 6p \exp\{-t^2/(256\mathfrak{M}^2 r_n^2\, |I|)\} \le 6(n \vee p)^{-c_2},$$

where the first inequality is by (D.4) and the fact $t > \sqrt{12} r(|I|/K)^{1/2}$ implies $4\,|I|\, r^2/(Kt^2) \le 1/3$ and the last inequality is by choice that $t > 64\mathfrak{M} r_n (|I| \log(n \vee p))^{1/2}$. Hence, taking $t = C_\lambda \sqrt{d_0} r \sqrt{\max\{|I|, \log(n \vee p)\}} \sqrt{\log(n \vee p)}$, yields that

$$\mathbb{P}\Big\{ \sup_{\underset{\sim}{\Delta}\in\mathcal{A}, \|\underset{\sim}{\Delta}\|_{\mathbb{S}}\le r} \Big| \sum_{i\in I} [Q_i(\underset{\sim}{\Delta}) - \mathbb{E}Q_i(\underset{\sim}{\Delta})] \Big| \le \widetilde{\lambda} \sqrt{d_0} r \Big\} > 1 - 6(n \vee p)^{-c_2}, \quad \text{(D.6)}$$

where $\lambda = C_\lambda \sqrt{\log(n \vee p)}$ with $C_\lambda \ge C_* := \mathfrak{M} \max\{\sqrt{12}, 128(2\underset{\sim}{\rho}^{-1/2} + 1)\}/K^{1/2}$ being a large enough constant.

Appendix D.6. Proof of Lemma 18

We denote $\underset{\sim}{\Delta} = \widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_J$ and can prove that $\underset{\sim}{\Delta}$ has a similar conclusion to Lemma 15, i.e.,

$$\|(\widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_J)_{\mathcal{S}^c}\|_1 = \|(\widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_J)_{\mathcal{S}^c}\|_1 \le 3\|(\widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_J)_{\mathcal{S}}\|_1 + \|\widehat{\underset{\sim}{b}}_J - \underset{\sim}{b}^*_J\|_1/K \le 3\|(\widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_J)_{\mathcal{S}}\|_1 + \|\widehat{\underset{\sim}{b}}_J - \underset{\sim}{b}^*_I\|_2/\sqrt{K} + 6d_0 C_\beta,$$

where first inequality follows from Lemma 15 with $\lambda > \max\{\lambda_1, \lambda_2\}$, and last inequality follows from the analysis of $\underset{\sim}{\beta}^*_J$ in proof of Lemma 4. As $r^2 = \|\underset{\sim}{\Delta}\|_{\mathbb{S}}^2 \ge (K\underset{\sim}{\rho}\|\underset{\sim}{\Delta}\|_2^2 + \|\underset{\sim}{\delta}\|_2^2)$,

$$\begin{aligned}
\|\underset{\sim}{\Delta}\|_1 &\le 4\|\underset{\sim}{\Delta}_{\mathcal{S}}\|_1 + (K^{1/2} + K^{-1/2})\|\underset{\sim}{\delta}\|_2 + 6d_0 C_\beta \\
&\le 4\sqrt{d_0}\|\underset{\sim}{\Delta}_{\mathcal{S}}\|_2 + (K^{1/2} + K^{-1/2})\|\underset{\sim}{\delta}\|_2 + 6d_0 C_\beta \\
&\le 4\sqrt{d_0} r/(K\underset{\sim}{\rho})^{1/2} + (K^{1/2} + K^{-1/2}) r + 6d_0 C_\beta \le C_k d_0 r. \quad \text{(D.7)}
\end{aligned}$$

To save space, the analysis of the following fact from a similar argument as proof of Lemma 17. If $t = C_\lambda d_0 r \sqrt{\max\{|I|, \log(n \vee p)\}} \sqrt{\log(n \vee p)}$, yields that inequality aligned with (D.6), with the only difference being $\underset{\sim}{\Delta} = \widehat{\underset{\sim}{\beta}}_J - \underset{\sim}{\beta}^*_I$, where $\lambda = C_\lambda \sqrt{d_0 \log(n \vee p)}$ with C_λ being a large enough absolute constant.

References

[1] J. Bai, P. Perron, Estimating and testing linear models with multiple structural changes, Econometrica 66 (1998) 47–78.

[2] L. Chang, J. Xu, X. Tie, J. Wu, Impact of the 2015 El Niño event on winter air quality in China, Scientific reports 6 (2016) 34275.

[3] H. Cho, P. Fryzlewicz, Multiple change point detection for high dimensional time series via sparsified binary segmentation, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77 (2015) 475–507.

[4] F. Friedrich, A. Kempe, V. Liebscher, G. Winkler, Complexity penalized M-estimation: Fast computation, Journal of Computational and Graphical Statistics 17 (2008) 201–224.

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

[6] Y. Gu, H. Zou, Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration, IEEE Transactions on Information Theory 66 (2020) 7132–7154.
[7] L. Horváth, Detecting changes in linear regressions, Statistics: A Journal of Theoretical and Applied Statistics 26 (1995) 189–208.
[8] L. Horváth, Q.-M. Shao, Limit theorems for the union-intersection test, Journal of Statistical Planning and Inference 44 (1995) 133–148.
[9] F. Jiang, Z. Zhao, X. Shao, Modelling the COVID-19 infection trajectory: A piecewise linear quantile trend model, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84 (2021) 1589–1607.
[10] B. Kai, R. Li, H. Zou, Local composite quantile regression smoothing: An efficient and safe alternative to local polynomial regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (2010) 49–69.
[11] A. Kaul, V. K. Jandhyala, S. B. Fotopoulos, An efficient two step algorithm for high dimensional change point regression models without grid search, Journal of Machine Learning Research 20 (2019) 1–40.
[12] R. Koenker, G. Bassett Jr, Regression quantiles, Econometrica 46 (1978) 33–50.
[13] S. Kovács, H. Li, P. Bühlmann, A. Munk, Seeded binary segmentation: A general methodology for fast and optimal change point detection, Biometrika 110 (2022) 249–256.
[14] M. Ledoux, M. Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, Springer Science and Business Media, 1991.
[15] S. Lee, Y. Liao, M. H. Seo, Y. Shin, Oracle estimation of a change point in high-dimensional quantile regression, Journal of the American Statistical Association 113 (2018) 1184–1194.
[16] S. Lee, M. H. Seo, Y. Shin, The lasso for high dimensional regression with a possible change point, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78 (2016) 193–210.
[17] F. Leonardi, P. Bühlmann, Computationally efficient change point detection for high-dimensional regression, arXiv preprint arXiv:1601.03704 (2016).
[18] B. Liu, Z. Qi, X. Zhang, Y. Liu, Change point detection for high-dimensional linear models: A general tail-adaptive approach, arXiv preprint arXiv:2207.11532 (2022).
[19] B. Liu, X. Zhang, Y. Liu, High dimensional change point inference: Recent developments and extensions, Journal of Multivariate Analysis 188 (2022) 104833.
[20] B. Liu, C. Zhou, X. Zhang, A tail adaptive approach for change point detection, Journal of Multivariate Analysis 169 (2019) 33–48.
[21] B. Liu, C. Zhou, X. Zhang, Y. Liu, A unified data-adaptive framework for high dimensional change point detection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82 (2020) 933–963.
[22] Y. Liu, C. Zou, R. Zhang, Empirical likelihood ratio test for a change-point in linear regression model, Communications in Statistics-Theory and Methods 37 (2008) 2551–2563.
[23] A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based dna copy number data, Biostatistics 5 (2004) 557–572.
[24] E. Page, Control charts with warning lines, Biometrika 42 (1955) 243–257.
[25] M. Pietrosanu, J. Gao, L. Kong, B. Jiang, D. Niu, Advanced algorithms for penalized quantile and composite quantile regression, Computational Statistics 36 (2021) 333–346.
[26] A. Rinaldo, D. Wang, Q. Wen, R. Willett, Y. Yu, Localizing changes in high-dimensional regression models, in: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 2089–2097.
[27] S. Roy, Y. Atchadé, G. Michailidis, Change point estimation in high dimensional Markov random-field models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79 (2017) 1187–1206.
[28] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58 (1996) 267–288.
[29] A. W. Van Der Vaart, J. A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer, 1996.
[30] D. Wang, Y. Yu, A. Rinaldo, Univariate mean change point detection: Penalization, cusum and optimality, Electronic Journal of Statistics 14 (2020) 1917 – 1961.
[31] D. Wang, Y. Yu, A. Rinaldo, Optimal change point detection and localization in sparse dynamic networks, The Annals of Statistics 49 (2021) 203–232.
[32] D. Wang, Y. Yu, A. Rinaldo, Optimal covariance change point localization in high dimensions, Bernoulli 27 (2021) 554–575.
[33] D. Wang, Y. Yu, A. Rinaldo, R. Willett, Localizing changes in high-dimensional vector autoregressive processes, arXiv preprint arXiv:1909.06359 (2019).
[34] D. Wang, Z. Zhao, K. Z. Lin, R. Willett, Statistically and computationally efficient change point localization in regression settings, Journal of Machine Learning Research 22 (2021) 1–46.
[35] T. Wang, R. J. Samworth, High dimensional change point estimation via sparse projection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80 (2018) 57–83.
[36] X. Wang, B. Liu, X. Zhang, Y. Liu, A. D. N. Initiative, Efficient multiple change point detection for high-dimensional generalized linear models, Canadian Journal of Statistics 51 (2023) 596–629.
[37] B. Zhang, J. Geng, L. Lai, Multiple change-points estimation in linear regression models via sparse group lasso, IEEE Transactions on Signal Processing 63 (2015) 2209–2224.
[38] L. Zhang, H. J. Wang, Z. Zhu, Testing for change points due to a covariate threshold in quantile regression, Statistica Sinica 24 (2014) 1859–1877.
[39] L. Zhang, H. J. Wang, Z. Zhu, Composite change point estimation for bent line quantile regression, Annals of the Institute of Statistical Mathematics 69 (2017) 145–168.
[40] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, S. X. Chen, Cautionary tales on air-quality improvement in Beijing, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 473 (2017) 20170457.
[41] T. Zhao, M. Kolar, H. Liu, A general framework for robust testing and confidence regions in high-dimensional quantile regression, arXiv preprint arXiv:1412.8724 (2014).
[42] H. Zou, M. Yuan, Composite quantile regression and the oracle model selection theory, The Annals of Statistics 36 (2008) 1108–1126.