**Statistica Sinica Preprint No: SS-2023-0337**

| | |
|---|---|
| **Title** | Change Point Detection for High-dimensional Linear Models: A General Tail-adaptive Approach |
| **Manuscript ID** | SS-2023-0337 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202023.0337 |
| **Complete List of Authors** | Bin Liu,<br>Zhengling Qi,<br>Xinsheng Zhang and<br>Yufeng Liu |
| **Corresponding Authors** | Yufeng Liu |
| **E-mails** | yfliu@email.unc.edu |

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337

Statistica Sinica

# Change Point Detection for High-dimensional Linear Models: A General Tail-adaptive Approach

Bin Liu[1], Zhengling Qi[2], Xinsheng Zhang[3], and Yufeng Liu[4]

[1,3] *Fudan University, China;*

[2] *George Washington University, U.S.A;*

[4] *University of North Carolina at Chapel Hill, U.S.A*

*Abstract:* We propose a novel approach for detecting change points in high-dimensional linear regression models. Unlike previous research that relies on strict Gaussian/sub-Gaussian error assumptions and has prior knowledge of change points, we propose a tail-adaptive method for change point detection and estimation. We use a weighted combination of composite quantile and least squared losses to build a new loss function, allowing us to leverage information from both conditional means and quantiles. For change point testing, we develop a family of individual testing statistics with different weights to account for unknown tail structures. These individual tests are further aggregated to construct a powerful tail-adaptive test for sparse regression coefficient changes. For change point estimation, we propose a family of argmax-based individual estimators. We provide theoretical justifications for the validity of these tests and change point estimators. Additionally, we introduce a new algorithm for detecting multiple change points in a tail-adaptive manner using the wild binary segmentation. Extensive

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337



numerical results show the effectiveness of our proposed method.

*Key words and phrases:* Binary segmentation; Bootstrap; Heterogeneity; Multiple change points

## 1. Introduction

With the advances of data collection and storage capacity, large scale/high-dimensional data are ubiquitous in many scientific fields ranging from genomics, finance, to social science. Due to the complex data generation mechanism, the heterogeneity, also known as the structural break, has become a common phenomenon for high-dimensional data, where the underlying model of data generation changes and the identically distributed assumption may not hold anymore. Change point analysis is a powerful tool for handling structural changes since the seminal work by Page (1955). It received considerable attentions in recent years and has a lot of real applications in various fields including genomics (Liu et al., 2020), social science (Roy et al., 2017), and even for the recent COVID-19 pandemic studies (Jiang et al., 2022). Motivated by this, in this paper, we study change point testing and estimation for high-dimensional linear regression models.

Specifically, suppose we have $n$ independent but (time) ordered realizations $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$. For each time point $i$, consider the following regression model:

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

where $Y_i \in \mathbb{R}$ is the response variable, $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^\top$ is the regression coefficient vector for the $i$-th observation, and $\epsilon_i$ is the error term. For the above model, our first question is whether there is a change point. This can be formulated as the following hypothesis testing problem:

$$\mathbf{H}_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n \ \text{ v.s. } \mathbf{H}_1 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_{k_1} \neq \boldsymbol{\beta}_{k_1+1} = \cdots = \boldsymbol{\beta}_n, \tag{1.2}$$

where $k_1$ is the possible but unknown change point location. According to (1.2), the linear regression structure between $Y$ and $\boldsymbol{X}$ remains homogeneous if $\mathbf{H}_0$ holds, and otherwise there is a change point $k_1$ that divides the data into two segments with different regression coefficients, $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$. Our second goal of this paper is to identify the change point location if we reject $\mathbf{H}_0$ in (1.2). Note that the above two goals are referred as change point testing and estimation, respectively. In the practical use, both testing and estimation are important since practitioners typically have no prior knowledge about either the existence of a change point or its location. Therefore, it is very useful to consider simultaneous change point detection and estimation. Furthermore, the tail structure of the error $\epsilon_i$ in Model (1.1) is typically unknown, which could significantly affect the performance of the change point detection and estimation. In the existing literature, the performance guarantee of most methods on change point estimation relies

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337



on the assumption that the error $\epsilon_i$ follows a Gaussian/sub-Gaussian distribution. Such an assumption could be violated in practice when the data distribution is heavy-tailed or contaminated by outliers. While some robust methods can address these issues, they may lose efficiency when errors are indeed sub-Gaussian distributed. It is also very difficult to estimate the tail structures and construct a corresponding change point method based on that. Hence, it is of great interest to construct a tail-adaptive change point detection and estimation method for high-dimensional linear models.

## 1.1 Contribution

Motivated by our previous discussion, in this paper, under the high-dimensional setup with $p \gg n$, we propose a tail-adaptive procedure for simultaneous change point testing and estimation in linear regression models. The proposed method relies on a new loss function in our change point estimation procedure, which is a weighted combination between the composite quantile loss proposed in Zou and Yuan (2008) and the least squared loss with the weight $\alpha \in [0, 1]$ for balancing the efficiency and robustness. Thanks to this new loss function with different $\alpha$, we are able to borrow information related to the possible change point from both the conditional mean and quantiles in Model (1.1). Therefore, besides controlling the type

I error to any desirable level when $\mathbf{H}_0$ holds, the proposed method simultaneously enjoys high power and accuracy for change point testing and identification across various underlying error distributions including both lighted and heavy-tailed errors when there exists a change point. By combining our single change point estimation method with the wild binary segmentation (WBS) technique (Fryzlewicz, 2014), we also generalize our method for detecting multiple change points in Model (1.1).

In terms of our theoretical contribution, for each given $\alpha$, a novel score-based $\mathbb{R}^p$-dimensional individual CUSUM process $\{\boldsymbol{C}_\alpha(t), t \in [0,1]\}$ is proposed. Based on this, we construct a family of individual-based testing statistics $\{T_\alpha, \alpha \in [0,1]\}$ via aggregating $\boldsymbol{C}_\alpha(t)$ using the $\ell_2$-norm of its first $s_0$ largest order statistics, known as the $(s_0, 2)$-norm proposed in Zhou et al. (2018). A high-dimensional bootstrap procedure is introduced to approximate $T_\alpha$'s limiting null distributions. The proposed bootstrap method only requires mild conditions on the covariance structures of $\boldsymbol{X}$ and the underlying error distribution $\epsilon$, and is free of tuning parameters and computationally efficient. Furthermore, combining the corresponding individual tests in $\{T_\alpha, \alpha \in [0,1]\}$, we construct a tail-adaptive test statistic $T_{\text{ad}}$ by taking the minimum $P$-values of $\{T_\alpha, \alpha \in [0,1]\}$. The proposed tail-adaptive method $T_{\text{ad}}$ chooses the best individual test according to the data and thus enjoys simultaneous high power across various tail structures. Theoretically, we adopt a low-cost bootstrap method for approximating the limiting distribution of $T_{\text{ad}}$. In terms of size and power, for both individual and tail-adaptive tests, we prove that the corresponding test can control the type I error for any given significance level if $\mathbf{H}_0$ holds, and reject the null hypothesis with probability tending to one otherwise.

As for the change point estimation, once $\mathbf{H}_0$ is rejected by our test, based on each individual test statistic, we can estimate its location via taking argmax with respect to different candidate locations $t \in (0,1)$ for the $(s_0, 2)$-norm aggregated process $\{\|\boldsymbol{C}_\alpha(t)\|_{(s_0,2)}, t \in [0,1]\}$. Under some regular conditions, for each individual based estimator $\{\widehat{t}_\alpha, \alpha \in [0,1]\}$, we can show that the estimation error is rate optimal up to a $\log(pn)$ factor. Hence, the proposed individual estimators for the change point location allow the signal jump size scale well with $(n,p)$ and are consistent.

## 1.2 Related literature

For the low dimensional setting with a fixed $p$ and $p < n$, change point analysis for linear regression models has been well-studied. Specifically, Quandt (1958) considered testing (1.2) for a simple regression model with $p = 2$. Other extensions include the maximum likelihood ratio tests

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337



(Horváth, 1995), partial sums of regression residuals (Bai and Perron, 1998), and so on. Other related methods include Qu (2008); Zhang et al. (2014); Oka and Qu (2011) and among others.

Due to the curse of dimensionality, only a few papers studied high-dimensional change point analysis, which mainly focused on the change point estimation. See Lee et al. (2016); Kaul et al. (2019); Lee et al. (2018); Leonardi and Bühlmann (2016); Wang et al. (2021, 2023). However, none of the aforementioned papers handle hypothesis testing, which is the prerequisite for the change point detection. Furthermore, most existing literature requires strong conditions on the underlying errors $\epsilon_i$ for deriving the desirable theoretical properties, which may not be applicable when the data are heavy-tailed or contaminated by outliers. One possible solution is to use the robust method such as median regression in Lee et al. (2018) for change point estimation. As discussed in Zou and Yuan (2008); Zhao et al. (2014), when making statistical inference for homogeneous linear models, the asymptotic relative efficiency of median regression-based estimators compared to least squared-based is only about 64% in both low and high dimensions. In addition, inference based on quantile regression can have arbitrarily small relative efficiency compared to the least squared based regression. Our proposed tail-adaptive method is the one that can perform simultaneous change point testing and estimation for high-dimensional linear regression models with different distributions. In addition to controlling the type I error to any desirable level, the proposed method enjoys simultaneously high power and accuracy for the change point testing and identification across various underlying error distributions when there exists a change point.

The rest of this paper is organized as follows. In Section 2, we introduce our new tail-adaptive methodology for detecting a single change point as well as multiple change points. In Section 3, we derive the theoretical results in terms of size and power as well as the change point estimation. In Section 4, we present a brief summary of the simulation results. The concluding remarks are provided in Section 5. Detailed proofs and the full numerical results as well as an application to the S&P 100 dataset are given in the online supplementary materials.

**Notations:** For $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, we define its $\ell_p$ 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 real numbered sequences $a_n$ and $b_n$, we set $a_n = O(b_n)$ if there exits a constant $C$ such that $|a_n| \le C|b_n|$ for a sufficiently large $n$; $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$; $a_n \asymp b_n$ if there exists constants $c$ and $C$ such that $c|b_n| \le |a_n| \le C|b_n|$ for a sufficiently large $n$. For a sequence of random variables (r.v.s) $\{\xi_1, \xi_2, \ldots\}$, we set $\xi_n \xrightarrow{\mathbb{P}} \xi$ if $\xi_n$ converges to $\xi$ in probability as $n \to \infty$. We also denote $\xi_n = o_p(1)$ if $\xi_n \xrightarrow{\mathbb{P}} 0$. We define $\lfloor x \rfloor$ as the largest integer less than or equal to $x$ for $x \geq 0$. Denote $(\mathcal{X}, \mathcal{Y}) = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$.

## 2. Methodology

### 2.1 Single change point detection

In this section, we introduce our new methodology for Problem (1.2). We first focus on detecting a single change point in Model (1.1) with

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(1)} \mathbf{1}\{i \leq k_1\} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(2)} \mathbf{1}\{i > k_1\} + \epsilon_i, \text{ for } i = 1, \ldots, n. \quad (2.1)$$

In this paper, we assume $k_1 = \lfloor nt_1 \rfloor$ for some constant $t_1 \in (0, 1)$. Note that $t_1$ is called the relative change point location. We assume the change point does not occur at the beginning or end of data observations. Specifically, suppose there exists a constant $q_0 \in (0, 0.5)$ such that $q_0 \leq t_1 \leq 1 - q_0$ holds, which is a common assumption in the literature (e.g., Dette et al., 2018; Jirak, 2015). For Model (2.1), the conditional mean of $Y_i$ is:

$$\mathbb{E}[Y_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(1)} \mathbf{1}\{i \leq k_1\} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(2)} \mathbf{1}\{i > k_1\}. \quad (2.2)$$

Moreover, let $0 < \tau_1 < \ldots < \tau_K < 1$ be $K$ candidate quantile indices. For each $\tau_k \in (0, 1)$, let $b_k^{(0)} := \inf\{t : \mathbb{P}(\epsilon \leq t) \geq \tau_k\}$ be the $\tau_k$-th theoretical quantile for the generic error term $\epsilon$ in Model (2.1). Then, conditional on $\boldsymbol{X}_i$, the $\tau_k$-th quantile for $Y_i$ becomes:

$$Q_{\tau_k}(Y_i|\boldsymbol{X}_i) = b_k^{(0)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(1)} \mathbf{1}\{i \leq k_1\} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(2)} \mathbf{1}\{i > k_1\}, k = 1, \ldots, K, \quad (2.3)$$

where $Q_{\tau_k}(Y_i|\boldsymbol{X}_i) := \inf\{t : \mathbb{P}(Y_i \leq t|\boldsymbol{X}_i) \geq \tau_k\}$. Hence, if there exists a change point in Model (2.1), both the conditional mean and the conditional quantile change after the change point. This suggests that we can make change point inference for $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ using either (2.2) or (2.3). To propose our new testing statistic, we first introduce the following weighted composite loss function. In particular, let $\alpha \in [0, 1]$ be some candidate weight. Define the weighted composite loss function as:

$$\ell_\alpha(\boldsymbol{x}, y; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta}) := (1-\alpha)\frac{1}{K}\sum_{k=1}^{K} \rho_{\tau_k}(y - b_k - \boldsymbol{x}^\top \boldsymbol{\beta}) + \frac{\alpha}{2}(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2, \quad (2.4)$$

where $\rho_\tau(t) := t(\tau - \mathbf{1}\{t \leq 0\})$ is the check loss function (Koenker and Bassett, 1978), $\widetilde{\boldsymbol{\tau}} := (\tau_1, \ldots, \tau_K)^\top$ are user-specified $K$ quantile levels, and $\boldsymbol{b} = (b_1, \ldots, b_K)^\top \in \mathbb{R}^K$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$. Note that we can regard $\ell_\alpha(\boldsymbol{x}, y; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta})$ as a weighted loss function between the composite quantile loss and the squared error loss. For example, for $\alpha = 1$, it reduces to the ordinary least squared-based loss function with $\ell_1(\boldsymbol{x}, y) = (y - \boldsymbol{x}^\top \boldsymbol{\beta})^2/2$. When $\alpha = 0$, it is the composite quantile loss function $\ell_0(\boldsymbol{x}, y) = \sum_{k=1}^{K} \rho_{\tau_k}(y - b_k - \boldsymbol{x}^\top \boldsymbol{\beta})/K$ proposed in Zou and Yuan (2008). It is known that the least squared loss-based method has the best statistical efficiency when errors follow Gaussian distributions and the composite quantile loss is more robust when the error distribution is heavy-tailed or contaminated by outliers. As discussed before, in practice, it is challenging to obtain the tail structure of the error distribution and construct a corresponding change point testing method based on the error structure. Hence, we propose a weighted loss function by borrowing the information related to the possible change point from both the conditional mean and quantiles. We use the weight $\alpha$ to balance the efficiency and robustness.

Our new testing statistic is based on a novel high-dimensional weighted score-based CUSUM process of the weighted composite loss function. In particular, for the composite loss function $\ell_\alpha(\boldsymbol{x}, y; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta})$, define its score (subgradient) function $\partial \ell_\alpha(\boldsymbol{x}, y; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$ as:

$$\boldsymbol{Z}_\alpha(\boldsymbol{x}, y; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta}) := \Big[\frac{1-\alpha}{K}\sum_{k=1}^{K} \boldsymbol{x}\big(\mathbf{1}\{y - b_k - \boldsymbol{x}^\top\boldsymbol{\beta} \le 0\} - \tau_k\big)\Big] - \alpha\big[\boldsymbol{x}(y - \boldsymbol{x}^\top\boldsymbol{\beta})\big]. \tag{2.5}$$

Using $\boldsymbol{Z}_\alpha(\boldsymbol{x}, y; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta})$, for each $\alpha \in [0,1]$ and $t \in (0,1)$, we first define the oracle score-based CUSUM as follows:

$$\widetilde{\boldsymbol{C}}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta}) = \frac{1}{\sqrt{n}\sigma(\alpha, \widetilde{\boldsymbol{\tau}})}\Big(\sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{Z}_\alpha(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta}) - \frac{\lfloor nt \rfloor}{n}\sum_{i=1}^{n} \boldsymbol{Z}_\alpha(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta})\Big), \tag{2.6}$$

where $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}}) := \mathrm{Var}[(1-\alpha)e_i(\widetilde{\boldsymbol{\tau}}) - \alpha\epsilon_i)]$ with $e_i(\widetilde{\boldsymbol{\tau}}) := K^{-1}\sum_{k=1}^{K}(\mathbf{1}\{\epsilon_i \le b_k^{(0)}\} - \tau_k)$. Note that we call $\widetilde{\boldsymbol{C}}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \boldsymbol{b}, \boldsymbol{\beta})$ oracle since we assume $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})$ is known. In Section 2.2, we will give the explicit form of $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})$ under

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337



various combinations of $\alpha$ and $\widetilde{\boldsymbol{\tau}}$ and introduce its consistent estimator under both $\mathbf{H}_0$ and $\mathbf{H}_1$. In the following, to motivate our test statistics, we study the behaviors of $\widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\boldsymbol{b},\boldsymbol{\beta})$ under $\mathbf{H}_0$ and $\mathbf{H}_1$ respectively. First, under $\mathbf{H}_0$, if we replace $\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{b}=\boldsymbol{b}^{(0)}$ in (2.6), the score based CUSUM becomes

$$\begin{aligned}&\widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\boldsymbol{b}^{(0)},\boldsymbol{\beta}^{(0)})\\&=\frac{1}{\sqrt{n}\sigma(\alpha,\widetilde{\boldsymbol{\tau}})}\Big(\sum_{i=1}^{\lfloor nt\rfloor}\boldsymbol{Z}_\alpha(\boldsymbol{X}_i,Y_i;\widetilde{\boldsymbol{\tau}},\boldsymbol{b}^{(0)},\boldsymbol{\beta}^{(0)})-\frac{\lfloor nt\rfloor}{n}\sum_{i=1}^{n}\boldsymbol{Z}_\alpha(\boldsymbol{X}_i,Y_i;\widetilde{\boldsymbol{\tau}},\boldsymbol{b}^{(0)},\boldsymbol{\beta}^{(0)})\Big).\end{aligned}$$

By noting that under $\mathbf{H}_0$, we have $Y_i=\boldsymbol{X}_i^\top\boldsymbol{\beta}^{(0)}+\epsilon_i$, the above CUSUM reduces to the following $\mathbb{R}^p$–dimensional random noise based CUSUM:

$$\begin{aligned}&\widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\boldsymbol{b}^{(0)},\boldsymbol{\beta}^{(0)})\\&=\frac{1}{\sqrt{n}\sigma(\alpha,\widetilde{\boldsymbol{\tau}})}\Big(\sum_{i=1}^{\lfloor nt\rfloor}\boldsymbol{X}_i((1-\alpha)e_i(\widetilde{\boldsymbol{\tau}})-\alpha\epsilon_i)-\frac{\lfloor nt\rfloor}{n}\sum_{i=1}^{n}\boldsymbol{X}_i((1-\alpha)e_i(\widetilde{\boldsymbol{\tau}})-\alpha\epsilon_i)\Big),\end{aligned}\tag{2.7}$$

whose asymptotic distribution can be easily characterized. Since both $\boldsymbol{b}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ are unknown, we need some proper estimators that can approximate them well under $\mathbf{H}_0$. In this paper, for each $\alpha\in[0,1]$, we obtain the estimators by solving the following penalized optimization problem:

$$(\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha)=\mathop{\arg\min}_{\substack{\boldsymbol{b}\in\mathbb{R}^K,\\ \boldsymbol{\beta}\in\mathbb{R}^p}}\Big[(1-\alpha)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=1}^{K}\rho_{\tau_k}(Y_i-b_i-\boldsymbol{X}_i^\top\boldsymbol{\beta})+\frac{\alpha}{2n}\sum_{i=1}^{n}(Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta})^2+\lambda_\alpha\|\boldsymbol{\beta}\|_1\Big],\tag{2.8}$$

where $\widehat{\boldsymbol{b}}_\alpha:=(\widehat{b}_{\alpha,1},\cdots,\widehat{b}_{\alpha,K})^\top$, $\widehat{\boldsymbol{\beta}}_\alpha:=(\widehat{\beta}_{\alpha,1},\ldots,\widehat{\beta}_{\alpha,p})^\top$, and $\lambda_\alpha$ is the non-negative tuning parameter controlling the overall sparsity of $\widehat{\boldsymbol{\beta}}_\alpha$. Note that the above estimators are obtained using all observations $(\mathcal{X},\mathcal{Y})$. After obtaining $(\widehat{\boldsymbol{b}}_\alpha^\top,\widehat{\boldsymbol{\beta}}_\alpha^\top)$, we plug them into the score function in (2.6) and obtain

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337



the $\mathbb{R}^p$–dimensional oracle score based CUSUM statistic as follows:

$$\widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha) = \big(\widetilde{C}_{\alpha,1}(t;\widetilde{\boldsymbol{\tau}},\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha),\ldots,\widetilde{C}_{\alpha,p}(t;\widetilde{\boldsymbol{\tau}},\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha)\big)^\top. \tag{2.9}$$

Using $(\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha)$, we can prove that under $\mathbf{H}_0$, for each $\alpha\in[0,1]$, (2.9) can approximate the random-noise based CUSUM process in (2.7) under some proper norm aggregations. Next, we investigate the behavior of (2.9) under $\mathbf{H}_1$. Observe that the score based CUSUM has the following decomposition:

$$\widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha) = \widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\boldsymbol{b}^{(0)},\boldsymbol{\beta}^{(0)}) + \boldsymbol{\delta}_\alpha(t) + \boldsymbol{R}_\alpha(t;\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha), \tag{2.10}$$

where $\widetilde{\boldsymbol{C}}_\alpha(t;\widetilde{\boldsymbol{\tau}},\boldsymbol{b}^{(0)},\boldsymbol{\beta}^{(0)})$ is the random noise based CUSUM process defined in (2.7), $\boldsymbol{R}_\alpha(t;\widehat{\boldsymbol{b}}_\alpha,\widehat{\boldsymbol{\beta}}_\alpha)$ is some random bias which has a very complicated form but can be controlled properly under $\mathbf{H}_1$, and $\boldsymbol{\delta}_\alpha(t)$ is the signal jump function. More specifically, let

$$SNR(\alpha,\widetilde{\boldsymbol{\tau}}) := \frac{(1-\alpha)\big(\frac{1}{K}\sum_{k=1}^{K} f_\epsilon(b_k^{(0)})\big)+\alpha}{\sigma(\alpha,\widetilde{\boldsymbol{\tau}})}, \tag{2.11}$$

where $f_\epsilon(t)$ is the probability density function of $\epsilon$, and define the signal jump function

$$\boldsymbol{\Delta}(t;\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}) := \begin{cases} \dfrac{\lfloor nt\rfloor(n-\lfloor nt_1\rfloor)}{n^{3/2}}\boldsymbol{\Sigma}\big(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}\big), & \text{if } t\le t_1,\\ \dfrac{\lfloor nt_1\rfloor(n-\lfloor nt\rfloor)}{n^{3/2}}\boldsymbol{\Sigma}\big(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}\big), & \text{if } t> t_1. \end{cases} \tag{2.12}$$

Then, the signal jump $\boldsymbol{\delta}_\alpha(t)$ can be explicitly represented as the products of $SNR(\alpha,\widetilde{\boldsymbol{\tau}})$ and $\boldsymbol{\Delta}(t;\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)})$, which has the following explicit form:

$$\boldsymbol{\delta}_\alpha(t) := SNR(\alpha,\widetilde{\boldsymbol{\tau}})\times\boldsymbol{\Delta}(t;\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}). \tag{2.13}$$

By (2.13), we can see that $\boldsymbol{\delta}_\alpha(t)$ can be decomposed into a loss-function-dependent part $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ and a change-point-model-dependent part $\boldsymbol{\Delta}(t; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$. More specifically, the first term $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ (short for the signal-to-noise-ratio) is only related to the parameters $\alpha, K, \boldsymbol{b}^{(0)}$ as well as $\sigma(\alpha, \widetilde{\boldsymbol{\tau}})$, resulting from a user specified weighted loss function defined in (2.4). In contrast, the second term $\{\boldsymbol{\Delta}(t; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}), t \in [0,1]\}$ is only related to Model (2.1), which is based on parameters $t_1$, $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}^{(1)}$, and $\boldsymbol{\beta}^{(2)}$ and is independent of the loss function. Moreover, for any weighted composite loss function, the process $\{\boldsymbol{\Delta}(t; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}), t \in [0,1]\}$ has the following properties: First, under $\mathbf{H}_1$, the non-zero elements of $\boldsymbol{\Delta}(t; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ are at most $(s^{(1)} + s^{(2)})$-sparse since we require sparse regression coefficients in the model; Second, we can see that $\|\boldsymbol{\Delta}(t; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})\|$ with $t \in [q_0, 1-q_0]$ obtains its maximum value at the true change point location $t_1$, where $\|\cdot\|$ denotes some proper norm such as $\|\cdot\|_\infty$. Hence, in theory, the signal jump function $\boldsymbol{\delta}_\alpha(t)$ also achieves its maximum value at $t_1$ under some proper norm. This is the key reason why using the score based CUSUM can correctly localize the true change point if $\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$ is big enough. More importantly, for a given underlying error distribution $\epsilon$ in Model (2.1), we can use $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ to further amplify the magnitude of $\boldsymbol{\delta}_\alpha(t)$ via choosing a proper combination of $\alpha$ and $\widetilde{\boldsymbol{\tau}}$. In particular, recall $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}}) := \mathrm{Var}[(1-\alpha)e_i(\widetilde{\boldsymbol{\tau}}) - \alpha\epsilon_i)]$.

Then, we have

$$\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}}) = (1-\alpha)^2 \mathrm{Var}[e_i(\widetilde{\boldsymbol{\tau}})] + \alpha^2\sigma^2 - 2\alpha(1-\alpha)\mathrm{Cov}(e_i(\widetilde{\boldsymbol{\tau}}), \epsilon_i), \quad (2.14)$$

where $\sigma^2 := \mathrm{Var}(\epsilon)$. Using (2.11) and (2.14), $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ can be further simplified under some specific $\alpha$. For example, if $\alpha = 1$, then $SNR(\alpha, \widetilde{\boldsymbol{\tau}}) = 1/\sigma$; If $\alpha = 0$, then

$$SNR(\alpha, \widetilde{\boldsymbol{\tau}}) = \frac{\sum_{k=1}^{K} f_\epsilon(b_k^{(0)})}{\sqrt{\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \gamma_{k_1 k_2}}}$$

with $\gamma_{k_1 k_2} := \min(\tau_{k_1}, \tau_{k_2}) - \tau_{k_1}\tau_{k_2}$; If we choose $\alpha \in (0,1)$, $K = 1$ and $\widetilde{\boldsymbol{\tau}} = \tau$ for some $\tau \in (0,1)$, then we have

$$SNR(\alpha, \widetilde{\boldsymbol{\tau}}) = \frac{(1-\alpha) f_\epsilon(b_\tau^{(0)}) + \alpha}{[(1-\alpha)^2\tau(1-\tau) + \alpha^2\sigma^2 - 2\alpha(1-\alpha)\mathrm{Cov}(e(\tau), \epsilon)]^{1/2}}. \quad (2.15)$$

Hence, for any underlying error distribution $\epsilon$ in Model (2.1), it is possible for us to choose a proper $\alpha$ and $\widetilde{\boldsymbol{\tau}}$ that makes $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ as large as possible for that distribution. See Figure 1 for a direct illustration.

For change point detection, a natural question is how to aggregate the $\mathbb{R}^p$–dimensional CUSUM process $\widetilde{\boldsymbol{C}}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha)$. Note that for high-dimensional sparse linear models, there are at most $s = s^{(1)} + s^{(2)}$ coordinates in $\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$ that can have a change point, which can be much smaller than the data dimension $p$, although we allow $s$ to grow with the sample size $n$. Motivated by this, in this paper, we aggregate the first $s_0$ largest statistics of $\widetilde{\boldsymbol{C}}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha)$. To that end, we introduce the $(s_0, 2)$-norm as follows. Let $\boldsymbol{v} = (v_1, \ldots, v_p) \in \mathbb{R}^p$. For any $1 \le s_0 \le p$, define

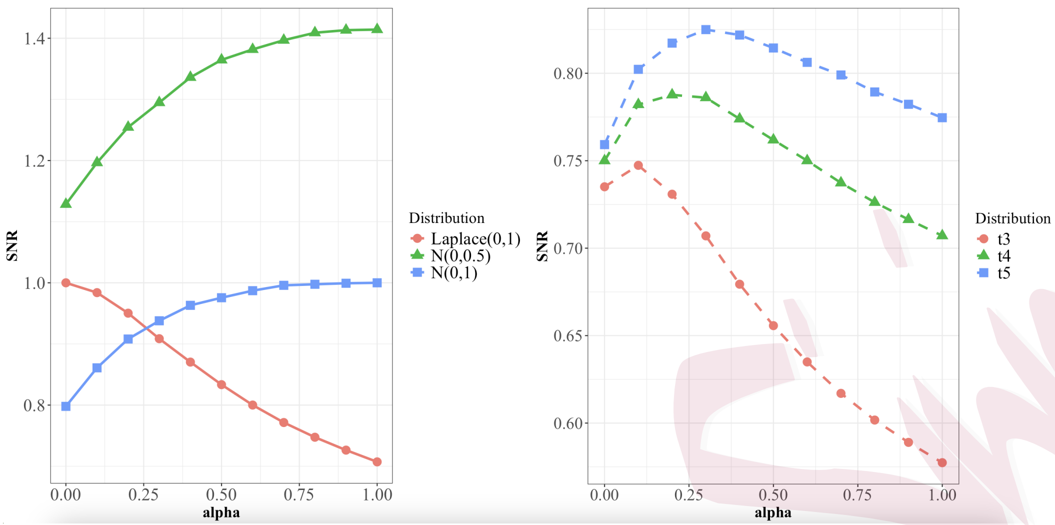


Figure 1: $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ under various errors with different weights $\alpha \in \{0, 0.1, \ldots, 0.9, 1\}$ for the weighted loss with $\widetilde{\tau} = 0.5$ and $K = 1$.

$\|\boldsymbol{v}\|_{(s_0,2)} = (\sum_{j=1}^{s_0} |v_{(j)}|^2)^{1/2}$, where $|v_{(1)} \geq |v_{(2)}| \cdots \geq |v_{(p)}|$ are the order statistics of $\boldsymbol{v}$. By definition, we can see that $\|\boldsymbol{v}\|_{(s_0,2)}$ is the $L_2$-norm for the first $s_0$ largest order statistics of $(|v_1|, \ldots, |v_p|)^\top$, which can be regarded as an adjusted $L_2$-norm in high dimensions. Note that the $(s_0, 2)$-norm is a special case of the $(s_0, \widetilde{p})$-norm proposed in Zhou et al. (2018) by setting $\widetilde{p} = 2$. We also remark that taking the first $s_0$ largest order statistics can account for the sparsity structure of $\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$. Using the $(s_0, 2)$-norm with a user-specified $s_0$ and known variance $\sigma^2(\alpha, \widetilde{\tau})$, we introduce the oracle individual testing statistic with respect to a given $\alpha \in [0,1]$ as

$$\widetilde{T}_{\alpha} = \max_{q_0 \le t \le 1-q_0} \left\| \widetilde{\boldsymbol{C}}_{\alpha}(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_{\alpha}, \widehat{\boldsymbol{\beta}}_{a}) \right\|_{(s_0,2)}, \quad \text{with} \quad \alpha \in [0,1].$$

By construction, $\widetilde{T}_{\alpha}$ can capture the tail structure of the underlying regression errors by choosing a special $\alpha$ and $\widetilde{\boldsymbol{\tau}}$. Specifically, for $\alpha = 1$, it equals to the least squared loss-based method. In this case, since $\widetilde{T}_{\alpha}$ only uses the moment information of the errors, it is powerful for detecting a change point with light-tailed errors such as Gaussian or sub-Gaussian distributions. For $\alpha = 0$, $\widetilde{T}_{\alpha}$ reduces to the composite quantile loss-based method, which only uses the information of ranks or quantiles. In this case, $\widetilde{T}_{\alpha}$ is more robust to data with heavy tails such as Cauchy distributions. As a special case of $\alpha = 0$, if we further choose $\widetilde{\tau} = 0.5$ and $K = 1$, our testing statistic reduces to the median regression-based method. Moreover, if we choose a proper non-trivial weight $\alpha$, $\widetilde{T}_{\alpha}$ enjoys satisfactory power performance for data with a moderate magnitude of tails such as the Student's $t_v$ or Laplace distributions. Hence, our proposed individual testing statistics can adequately capture the tail structures of the data.

Another distinguishing feature for using $\widetilde{T}_{\alpha}$ is that, we can establish a general and flexible framework for aggregating the score based CUSUM for high-dimensional sparse linear models. Instead of taking the $\ell_{\infty}$-norm as most papers adopted for making statistical inference of high-dimensional linear models (e.g., Xia et al., 2018), we choose to aggregate them via using the $\ell_2$-norm of the first $s_0$ largest order statistics. Under this framework, the $\ell_\infty$-norm is a special case by taking $s_0 = 1$.

### 2.2 Variance estimation under $\mathbf{H}_0$ and $\mathbf{H}_1$

Note that $\widetilde{T}_\alpha$ is constructed using a known variance $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})$ which is defined in (2.14). In practice, however, $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})$ is typically unknown. Hence, to yield a powerful testing method, a consistent variance estimation is needed especially under the alternative hypothesis. For high-dimensional change point analysis, the main difficulty comes from the unknown change point $t_1$. To overcome this problem, we propose a weighted variance estimation. In particular, for each fixed $\alpha \in [0, 1]$ and $t \in (0, 1)$, define the score based CUSUM statistic without standardization as follows:

$$\breve{\boldsymbol{C}}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha) = \frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{Z}_\alpha(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha) - \frac{\lfloor nt \rfloor}{n} \sum_{i=1}^{n} \boldsymbol{Z}_\alpha(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha)\Big). \tag{2.16}$$

For each $\alpha \in [0, 1]$, we obtain the individual based change point estimator:

$$\widehat{t}_\alpha = \mathop{\arg\max}_{q_0 \le t \le 1 - q_0} \left\| \breve{\boldsymbol{C}}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha) \right\|_{(s_0, 2)}. \tag{2.17}$$

In Theorem 3, we prove that under some regular conditions, if $\mathbf{H}_1$ holds, $\widehat{t}_\alpha$ is a consistent estimator for $t_1$, e.g. $|n\widehat{t}_\alpha - nt_1| = o_p(n)$. This result enables us to propose a variance estimator which is consistent under both $\mathbf{H}_0$ and $\mathbf{H}_1$. Specifically, let $h \in (0, 1)$ be a user specified constant, and define the samples $n_- = \{i : i \le nh\widehat{t}_\alpha\}$ and $n_+ = \{i : \widehat{t}_\alpha n + (1-h)(1-\widehat{t}_\alpha)n \le i \le n\}$. Let $((\widehat{\boldsymbol{b}}_\alpha^{(1)})^\top, (\widehat{\boldsymbol{\beta}}_\alpha^{(1)})^\top)$ and $((\widehat{\boldsymbol{b}}_\alpha^{(2)})^\top, (\widehat{\boldsymbol{\beta}}_\alpha^{(2)})^\top)$ be the estimators using the samples in $n_-$ and $n_+$. For each $\alpha$, we calculate the regression residuals:

$$\widehat{\epsilon}_i = (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_\alpha^{(1)})\mathbf{1}\{i \in n_-\} + (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_\alpha^{(2)})\mathbf{1}\{i \in n_+\}, \text{ for } i \in n_- \cup n_+. \quad (2.18)$$

Moreover, compute $\widehat{e}_i(\widetilde{\boldsymbol{\tau}}) = K^{-1}\sum_{k=1}^{K} \widehat{e}_i(\tau_k)$ with $\widehat{e}_i(\tau_k)$ defined as

$$\widehat{e}_i(\tau_k) := (\mathbf{1}\{\widehat{\epsilon}_i \le \widehat{b}_{\alpha,k}^{(1)}\} - \tau_k)\mathbf{1}\{i \in n_-\} + (\mathbf{1}\{\widehat{\epsilon}_i \le \widehat{b}_{\alpha,k}^{(2)}\} - \tau_k)\mathbf{1}\{i \in n_+\}. \quad (2.19)$$

Let $\widehat{s}^{(1)}$ and $\widehat{s}^{(2)}$ be the sparsity levels of $\widehat{\boldsymbol{\beta}}_\alpha^{(1)}$ and $\widehat{\boldsymbol{\beta}}_\alpha^{(2)}$. Lastly, based on $\widehat{\epsilon}_i$ and $\widehat{e}_i(\widetilde{\boldsymbol{\tau}})$, we construct a weighted estimator for $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})$:

$$\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}}) = \widehat{t}_\alpha \times \widehat{\sigma}_-^2(\alpha, \widetilde{\boldsymbol{\tau}}) + (1 - \widehat{t}_\alpha) \times \widehat{\sigma}_+^2(\alpha, \widetilde{\boldsymbol{\tau}}), \quad (2.20)$$

where:

$$\widehat{\sigma}_-^2(\alpha, \widetilde{\boldsymbol{\tau}}) := \frac{1}{|n_- - \widehat{s}^{(1)}|}\sum_{i \in n_-}\left[(1-\alpha)\widehat{e}_i(\widetilde{\boldsymbol{\tau}}) - \alpha\widehat{\epsilon}_i\right]^2,$$
$$\widehat{\sigma}_+^2(\alpha, \widetilde{\boldsymbol{\tau}}) := \frac{1}{|n_+ - \widehat{s}^{(2)}|}\sum_{i \in n_+}\left[(1-\alpha)\widehat{e}_i(\widetilde{\boldsymbol{\tau}}) - \alpha\widehat{\epsilon}_i\right]^2.$$

For the above variance estimation, we can prove that $|\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}}) - \sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})| = o_p(1)$ under either $\mathbf{H}_0$ or $\mathbf{H}_1$. As a result, the proposed variance estimator $\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}})$ avoids the problem of non-monotonic power performance as discussed in Shao and Zhang (2010), which is a serious issue in change point analysis. Hence, we replace $\sigma(\alpha, \widetilde{\boldsymbol{\tau}})$ in (2.9) by $\widehat{\sigma}(\alpha, \widetilde{\boldsymbol{\tau}})$ and define the data-driven score-based CUSUM process

$$\boldsymbol{C}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha) = \frac{1}{\sqrt{n}\widehat{\sigma}(\alpha, \widetilde{\boldsymbol{\tau}})}\Big(\sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{Z}_\alpha(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha) - \frac{\lfloor nt \rfloor}{n}\sum_{i=1}^{n} \boldsymbol{Z}_\alpha(\boldsymbol{X}_i, Y_i; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_\alpha)\Big). \quad (2.21)$$

For a user-specified $s_0 \in \{1, \ldots, p\}$ and any $\alpha \in [0,1]$, we define the final individual-based testing statistic as follows:

$$T_\alpha = \max_{q_0 \le t \le 1-q_0} \left\| \boldsymbol{C}_\alpha(t; \widetilde{\boldsymbol{\tau}}, \widehat{\boldsymbol{b}}_\alpha, \widehat{\boldsymbol{\beta}}_a) \right\|_{(s_0,2)}, \quad \text{with} \quad \alpha \in [0,1]. \tag{2.22}$$

In what follows, we use $\{T_\alpha, \alpha \in [0,1]\}$ as our individual testing statistics.

## 2.3 Bootstrap approximation for the individual testing statistic

In high dimensions, it is very difficult to obtain the limiting null distribution of $T_\alpha$. To overcome this problem, we propose a novel bootstrap procedure. In particular, suppose we implement the bootstrap procedure for $B$ times. Then, for each $b$-th bootstrap with $b = 1, \ldots, B$, we generate i.i.d. random variables $e_1^b, \ldots, e_n^b$ with $e_i^b \sim N(0,1)$. Let $e_i^b(\widetilde{\boldsymbol{\tau}}) = K^{-1} \sum_{k=1}^{K} e_i^b(\tau_k)$ with $e_i^b(\tau_k) := \mathbf{1}\{\epsilon_i^b \le \Phi^{-1}(\tau_k)\} - \tau_k$, where $\Phi(x)$ is the CDF for the standard normal distribution. Then, for each individual-based testing statistic $T_\alpha$, with a user specified $s_0$, we define its $b$-th bootstrap sample-based score CUSUM process as:

$$\boldsymbol{C}_\alpha^b(t; \widetilde{\boldsymbol{\tau}}) = \frac{1}{\sqrt{n} v(\alpha, \widetilde{\boldsymbol{\tau}})} \Big( \sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{X}_i ((1-\alpha) e_i^b(\widetilde{\boldsymbol{\tau}}) - \alpha e_i^b) - \frac{\lfloor nt \rfloor}{n} \sum_{i=1}^{n} \boldsymbol{X}_i ((1-\alpha) e_i^b(\widetilde{\boldsymbol{\tau}}) - \alpha e_i^b) \Big), \tag{2.23}$$

where $v^2(\alpha, \widetilde{\boldsymbol{\tau}})$ is the corresponding variance for the bootstrap samples with

$$v^2(\alpha, \widetilde{\boldsymbol{\tau}}) := (1-\alpha)^2 \mathrm{Var}[e_i^b(\widetilde{\boldsymbol{\tau}})] + \alpha^2 - 2\alpha(1-\alpha)\mathrm{Cov}(e_i^b(\widetilde{\boldsymbol{\tau}}), e_i^b). \tag{2.24}$$

Note that for bootstrap, the calculation or estimation of $v^2(\alpha, \widetilde{\boldsymbol{\tau}})$ is not a difficult task since we use $N(0,1)$ as the error term. For example, when

$\widetilde{\tau} = 0.5$, it has an explicit form of

$$v^2(\alpha, \widetilde{\boldsymbol{\tau}}) = (1-\alpha)^2 \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \gamma_{k_1 k_2} + \alpha^2 - \alpha(1-\alpha)\sqrt{\frac{2}{\pi}}.$$

Hence, for simplicity, we directly use the oracle variance $v^2(\alpha, \widetilde{\boldsymbol{\tau}})$ in (2.23).

Using $\boldsymbol{C}_\alpha^b(t; \widetilde{\boldsymbol{\tau}})$ and for a user specified $s_0$, we define the $b$-th bootstrap version of the individual-based testing statistic $T_\alpha$ as

$$T_\alpha^b = \max_{q_0 \le t \le 1-q_0} \left\| \boldsymbol{C}_\alpha^b(t; \widetilde{\boldsymbol{\tau}}) \right\|_{(s_0,2)}, \quad \text{with} \quad \alpha \in [0,1]. \tag{2.25}$$

Let $\gamma \in (0, 0.5)$ be the significance level. For each individual-based testing statistic $T_\alpha$, let $F_\alpha = \mathbb{P}(T_\alpha \le t)$ be its theoretical CDF and $P_\alpha = 1 - F_\alpha(T_\alpha)$ be its theoretical $p$-value. Using the bootstrap samples $\{T_\alpha^1, \ldots, T_\alpha^B\}$, we estimate $P_\alpha$ by

$$\widehat{P}_\alpha = \frac{\sum_{b=1}^B \mathbf{1}\{T_\alpha^b > T_\alpha | \mathcal{X}, \mathcal{Y}\}}{B+1}, \quad \text{with} \quad \alpha \in [0,1]. \tag{2.26}$$

Given the significance level $\gamma$, we can construct the individual test as

$$\Psi_{\gamma,\alpha} = \mathbf{1}\{\widehat{P}_\alpha \le \gamma\}, \quad \text{with} \quad \alpha \in [0,1]. \tag{2.27}$$

For each $T_\alpha$, we reject $\mathbf{H}_0$ if and only if $\Psi_{\gamma,\alpha} = 1$. Note that the above bootstrap procedure is easy to implement since it does not require any model fitting such as obtaining the LASSO estimators which is required by the data-based testing statistic $T_\alpha$.

## 2.4 Constructing the tail-adaptive testing procedure

In Sections 2.1 – 2.3, we propose a family of individual-based testing statistics $\{T_\alpha, \alpha \in [0,1]\}$ and introduce a bootstrap-based procedure for approximating their theoretical $p$-values. As discussed in Sections 2.1 and seen from Figure 1, $T_\alpha$ with different $\alpha$'s can have varying power performance for a given underlying error distribution. For example, $T_\alpha$ with a larger $\alpha$ (e.g. $\alpha = 1$) is more sensitive to change points with light-tailed error distributions by using more moment information. In contrast, $T_\alpha$ with a smaller $\alpha$ (e.g. $\alpha = 0, 0.1$) is more powerful for data with heavy tails such as Student's $t_v$ or even Cauchy distribution. In general, as shown in Figure 1, a properly chosen $\alpha$ can give the most satisfactory power performance for data with a particular magnitude of tails. In practice, however, the tail structures of data are typically unknown. Hence, it is desirable to construct a tail-adaptive method which is simultaneously powerful under various tail structures of data. One candidate method is to find $\alpha^*$ which maximizes the theoretical $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$, i.e. $\alpha^* = \arg\max_\alpha SNR(\alpha, \widetilde{\boldsymbol{\tau}})$, and constructs a corresponding individual testing statistic $T_{\alpha^*}$. Note that in theory, calculating $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ needs to know $\sigma(\alpha, \widetilde{\boldsymbol{\tau}})$ and $\{f_\epsilon(b_k^{(0)}), k = 1, \ldots, K\}$, which could be difficult to estimate especially under the high-dimensional change point model. Instead, we construct our tail-adaptive method by combining all candidate individual tests for yielding a powerful one. In particular, as a small $p$-value leads to rejection of $\mathbf{H}_0$, for the individual tests $T_\alpha$ with $\alpha \in [0, 1]$, we construct the tail-adaptive testing statistic as their minimum

$p$-value, which is defined as follows:

$$T_{\text{ad}} = \min_{\alpha \in \mathcal{A}} \widehat{P}_{\alpha}, \tag{2.28}$$

where $\widehat{P}_{\alpha}$ is defined in (2.26), and $\mathcal{A}$ is a candidate subset of $\alpha$.

**Remark 1.** It's worth noting that for the construction of $T_{\text{ad}}$, the issue of selecting $K$ in the composite quantile regression is typically left to the user. Based on our extensive numerical studies, choosing $K = 1$ and $\tau = 0.5$ has satisfactory performance across data with various tail structures.

In this paper, we require $|\mathcal{A}|$ to be finite, which is a theoretical requirement. Note that our tail-adaptive method is flexible and user-friendly. In practice, if the users have some prior knowledge about the tails of errors, we can choose $\mathcal{A}$ accordingly. For example, we can choose $\mathcal{A} = \{0.9, 1\}$ for light-tailed errors, and $\mathcal{A} = \{0\}$ for extreme heavy tails such as Cauchy distributions. However, if the tail structure is unknown, we can choose $\mathcal{A}$ consisting both small and large values of $\alpha \in [0, 1]$. For example, according to our theoretical analysis of $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$, we find that $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ tends to be maximized near the boundary of $[0, 1]$. Hence, we recommend to use $\mathcal{A} = \{0, 0.1, 0.5, 0.9, 1\}$ in real applications, which is shown by our numerical studies to enjoy stable size performance as well as high powers across various error distributions. Let $F_{\text{ad}}(x)$ be its theoretical distribution function. Note that $F_{\text{ad}}(x)$ is unknown. Hence, we can not use $T_{\text{ad}}$ directly for

Problem (1.2). To approximate its theoretical $p$-value, we adopt the low-cost bootstrap method proposed by Zhou et al. (2018), which is also used in Liu et al. (2020). Let $\widehat{P}_{\text{ad}}$ be an estimation for the theoretical $p$-value of $T_{\text{ad}}$ using the low-cost bootstrap. Given the significance level $\gamma \in (0, 0.5)$, define the final tail-adaptive test:

$$\Psi_{\gamma,\text{ad}} = \mathbf{1}\{\widehat{P}_{\text{ad}} \leq \gamma\}. \tag{2.29}$$

For the tail-adaptive testing procedure, given $\gamma$, we reject $\mathbf{H}_0$ if $\Psi_{\gamma,\text{ad}} = 1$.

## 3. Theoretical results

In this section, we give some theoretical results. In Section 3.1, we provide some basic model assumptions. In Sections 3.2 and 3.3, we discuss the theoretical properties of the individual and tail-adaptive methods.

### 3.1 Basic assumptions

To save space, we offer concise descriptions of our assumptions below. Assumption A gives some conditions for the design matrix such as the non-degeneracy of the covariance matrix $\mathbf{\Sigma}$ in terms of its eigenvalues. Assumptions B-C impose some restrictions on the moments of the error terms as well as the design matrix. Assumptions D are some regular conditions for the underlying distribution of the errors. Assumption E imposes some conditions for the parameter spaces in terms of $(s_0, n, p, s, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$. For more details, please refer to Section 5 of the supplementary materials.

## 3.2 Theoretical results of the individual-based testing statistics

### 3.2.1 Validity of controlling the testing size

Before giving the size results, we first consider the variance estimation. Recall $\sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})$ in (2.14) and $\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}})$ in (2.20). Theorem 1 shows that the pooled weighted variance estimator $\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}})$ is consistent under $\mathbf{H}_0$, which is crucial for deriving the Gaussian approximation results as shown in Theorem 2 and shows that our testing method has satisfactory size performance.

**Theorem 1.** *For $\alpha = 1$, suppose Assumptions A, B, C, E hold. For $\alpha = 0$, suppose Assumptions A, C.1, D, E hold. For $\alpha \in (0, 1)$, suppose Assumptions A - E hold. Let $r_\alpha(n) = \sqrt{s \log(pn)/n}$ if $\alpha = 1$ and $r_\alpha(n) = s\sqrt{\frac{\log(pn)}{n}} \vee s^{\frac{1}{2}}(\frac{\log(pn)}{n})^{\frac{3}{8}}$ if $\alpha \in [0, 1)$. Under $\mathbf{H}_0$, for $\alpha \in [0, 1]$, we have*

$$|\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}}) - \sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})| = O_p(r_\alpha(n)).$$

Based on Theorem 1 as well as some other regularity conditions, the following Theorem 2 justifies the validity of our bootstrap-based procedure.

**Theorem 2.** *Suppose the assumptions in Theorem 1 hold. Then, under $\mathbf{H}_0$, for the individual test with $\alpha \in [0, 1]$, we have*

$$\sup_{z \in (0, \infty)} \left|\mathbb{P}(T_\alpha \leq z) - \mathbb{P}(T_\alpha^b \leq z | \mathcal{X}, \mathcal{Y})\right| = o_p(1), \text{ as } n, p \to \infty. \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337



Theorem 2 demonstrates that we can uniformly approximate the distribution of $T_\alpha$ by that of $T_\alpha^b$. The following Corollary further shows that our proposed new test $\Psi_{\gamma,\alpha}$ can control the Type I error asymptotically for any given significant level $\gamma \in (0,1)$.

**Corollary 1.** *Suppose assumptions in Theorem 2 hold. Under* $\mathbf{H}_0$, *we have*

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

### 3.2.2 Change point estimation

We next consider the performance of the individual test under $\mathbf{H}_1$. We first give some theoretical results on the change point estimation. To that end, some additional assumptions are needed. Recall $\Pi = \{j : \beta_j^{(1)} \neq \beta_j^{(2)}\}$ as the set with change points. For $j \in \{1, \dots, p\}$, define the signal jump $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_p)^\top$ with $\Delta_j := \beta_j^{(1)} - \beta_j^{(2)}$. Let $\Delta_{\min} = \min_{j\in\Pi} |\Delta_j|$ and $\Delta_{\max} = \max_{j\in\Pi} |\Delta_j|$. We now introduce the following Assumption F.

**Assumption F**. There exist constants $\underline{c} > 0$ and $\overline{C} > 0$ such that

$$0 < \underline{c} \leq \liminf_{p\to\infty} \frac{\Delta_{\min}}{\Delta_{\max}} \leq \limsup_{p\to\infty} \frac{\Delta_{\max}}{\Delta_{\min}} \leq \overline{C} < \infty. \tag{3.2}$$

Note that Assumption F is only a technical condition requiring that $\Delta_{\min}$ and $\Delta_{\max}$ are of the same order. Theorem 3 provides a non-asymptotic estimation error bound of the argmax-based individual estimators.

**Theorem 3.** *Suppose* $\|\boldsymbol{\Delta}\|_{(s_0,2)} \gg \sqrt{\log(pn)/n}$ *and Assumption F hold. For* $\alpha = 1$, *suppose Assumptions A, B, C, E.2 - E.4 as well as* $n^{1/4} =$

*$o(s)$ hold; For $\alpha = 0$, suppose Assumptions A, C.1, D, E.2 - E.4 as well as $\lim_{n,p\to\infty} s_0^{1/2}s^2\sqrt{\log(p)/n}\|\mathbf{\Delta}\|_{(s_0,2)} = 0$ hold; For $\alpha \in (0,1)$, suppose Assumptions A, B, C, D, E.2 - E.4 as well as $n^{1/4} = o(s)$ and $\lim_{n,p\to\infty} s_0^{1/2}s^2\sqrt{\log(pn)/n}\|\mathbf{\Delta}\|_{(s_0,2)} = 0$ hold. Then, under $\mathbf{H}_1$, for each $\alpha \in [0,1]$, with probability tending to one, we have*

$$\left|\widehat{t}_\alpha - t_1\right| \le C^*(s_0, \widetilde{\boldsymbol{\tau}}, \alpha)\frac{\log(pn)}{nSNR^2(\alpha, \widetilde{\boldsymbol{\tau}})\|\mathbf{\Sigma\Delta}\|^2_{(s_0,2)}}, \tag{3.3}$$

*where $C^*(s_0, \widetilde{\boldsymbol{\tau}}, \alpha) > 0$ is some constant only depending on $s_0, \widetilde{\boldsymbol{\tau}}$ and $\alpha$.*

**Remark 2.** Theorem 3 shows that our individual estimators are consistent under the condition $\|\mathbf{\Delta}\|_{(s_0,2)} \gg \sqrt{\log(pn)/n}$. Moreover, according to Rinaldo et al. (2021), for high-dimensional linear models, under Assumption F, if $\|\mathbf{\Delta}\|_\infty \gg 1/\sqrt{n}$, any change point estimator $\widehat{t}$ has an estimation lower bound $|\widehat{t} - t_1| \ge c_*\frac{1}{n\|\mathbf{\Delta}\|^2_\infty}$, for some constant $c_* > 0$. Hence, considering (3.2) and (3.3), with a fixed $s_0$, Theorem 3 demonstrates that our individual-based estimators are rate optimal up to a $\log(pn)$ factor.

### 3.2.3 Power performance

We discuss the power properties of the individual tests. Note that for the change point problem, variance estimation under the alternative is a difficult but important task. As pointed out in Shao and Zhang (2010), due to the unknown change point, any improper estimation may lead to non-monotonic power performance. This distinguishes the change point problem substantially from one-sample or two-sample tests where homogenous data are used to construct consistent variance estimation. Hence, for yielding a powerful change point test, we need to guarantee a consistent variance estimation. Theorem 4 shows that the pooled weighted variance estimation is consistent under $\mathbf{H}_1$. This guarantees that our proposed testing method has reasonable power performance.

**Theorem 4.** *Suppose the assumptions in Theorem 3 hold. Let* $r_\alpha(n) = \sqrt{s\log(pn)/n}$ *if* $\alpha = 1$ *and* $r_\alpha(n) = s\sqrt{\frac{\log(pn)}{n}} \vee s^{\frac{1}{2}}(\frac{\log(pn)}{n})^{\frac{3}{8}}$ *if* $\alpha \in [0, 1)$. *Under* $\mathbf{H}_1$*, for each* $\alpha \in [0, 1]$*, we have*

$$|\widehat{\sigma}^2(\alpha, \widetilde{\boldsymbol{\tau}}) - \sigma^2(\alpha, \widetilde{\boldsymbol{\tau}})| = O_p(r_\alpha(n)).$$

Using the consistent variance estimation, we are able to discuss the power properties of the individual tests. Define the oracle signal to noise ratio vector $\boldsymbol{D} = (D_1, \ldots, D_p)^\top$ with

$$D_j := \begin{cases} 0, & \text{for } \ j \in \Pi^c \\ SNR(\alpha, \widetilde{\boldsymbol{\tau}}) \times \Big| t_1(1-t_1)\big(\boldsymbol{\Sigma}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)})\big)_j \Big|, & \text{for } \ j \in \Pi, \end{cases} \tag{3.4}$$

where $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$ is defined in (2.11). Theorem 5 stated below shows that we can reject the null hypothesis with probability tending to 1.

**Theorem 5.** *Let* $\epsilon_n := O(s_0^{1/2} s\sqrt{\log(pn)/n}) \vee O(s_0^{1/2} s^2 \sqrt{\log(pn)/n}\|\boldsymbol{\Delta}\|_{(s_0,2)})$. *For each* $\alpha \in [0, 1]$*, assume the following conditions hold: When* $\alpha = 1$*, suppose that Assumptions A, B, C, E.2 - E.4 hold; When* $\alpha = 0$*, suppose that*

*Assumptions A, C.1, D, E.2 - E.4 as well as*

$$\lim_{n,p\to\infty} s_0^{1/2} s^2 \sqrt{\log(p)/n}\|\boldsymbol{\Delta}\|_{(s_0,2)} = 0 \tag{3.5}$$

*hold; When* $\alpha \in (0,1)$, *suppose that Assumptions A - D, E.2 - E.4 as well as (3.5) hold. Under* $\mathbf{H}_1$, *if* $\boldsymbol{D}$ *in (3.4) satisfies*

$$\sqrt{n} \times \|\boldsymbol{D}\|_{(s_0,2)} \geq \frac{C(\widetilde{\boldsymbol{\tau}}, \alpha)}{1-\epsilon_n} s_0^{1/2}\big(\sqrt{\log(pn)} + \sqrt{\log(1/\gamma)}\big), \tag{3.6}$$

*then we have*

$$\mathbb{P}(\Phi_{\gamma,\alpha} = 1) \to 1, \textit{ as } n, p, B \to \infty,$$

*where* $C(\widetilde{\boldsymbol{\tau}}, \alpha)$ *is some positive constant only depending on* $\widetilde{\boldsymbol{\tau}}$ *and* $\alpha$.

Theorem 5 demonstrates that with probability tending to one, our proposed individual test with $\alpha \in [0,1]$ can detect the existence of a change point for high-dimensional linear models as long as the corresponding signal to noise ratio satisfies (3.6). Combining (3.4) and (3.6), for each individual test, we note that with a larger signal jump and a closer change point location $t_1$ to the middle of data observations, it is more likely to trigger a rejection of the null hypothesis. More importantly, considering $\epsilon_n = o(1)$, Theorem 5 illustrates that for consistently detecting a change point, we require the signal to noise ratio vector to be at least an order of $\|\boldsymbol{D}\|_{(s_0,2)} \asymp s_0^{1/2}\sqrt{\log(pn)/n}$, which is particularly interesting to further discuss under several special cases. For example, if we choose $s_0 = 1$ and $\alpha = 1$, our proposed individual test reduces to the least squared loss based testing statistic with the $\ell_\infty$-norm aggregation. In this case, we require $\|\boldsymbol{D}\|_\infty \asymp \sqrt{\log(pn)/n}$ for detecting a change point. If we choose

$\alpha = 0$ with the composite quantile loss, the test is still consistent as long as $\|\boldsymbol{D}\|_\infty \asymp \sqrt{\log(pn)/n}$. Note that the latter one is of special interest for the robust change point detection. Hence, our theorem provides the unified condition for detecting a change point under a general framework, which may be of independent interest. Moreover, Theorem 5 reveals that for detecting a change point, our individual-based method with $\alpha \in [0, 1]$ can account for the tails of the data. For Model (2.1) with a fixed signal jump $\boldsymbol{\Delta}$ and a change point location $t_1$, considering (3.4) and (3.6), the individual test $T_\alpha$ is more powerful with a larger $SNR(\alpha, \widetilde{\boldsymbol{\tau}})$.

### 3.3 Theoretical results of the tail-adaptive testing statistics

In this section, we discuss the size and power properties of the tail-adaptive test $\Psi_{\gamma,\mathrm{ad}}$ defined in (2.29). To present the theorems, we need additional notations. Let $F_{T_\alpha}(x) := \mathbb{P}(T_\alpha \le x)$ be the CDF of $T_\alpha$. Then $\widehat{P}_\alpha$ in (2.26) approximates the following individual tests' theoretical $P$-values defined as $P_\alpha := 1 - F_{T_\alpha}(T_\alpha)$. Hence, based on the above theoretical $P$-values, we can define the oracle tail-adaptive testing statistic $\widetilde{T}_{\mathrm{ad}} = \min_{\alpha \in \mathcal{A}} P_\alpha$. Let $\widetilde{F}_{T,\mathrm{ad}}(x) := \mathbb{P}(\widetilde{T}_{\mathrm{ad}} \le x)$ be the CDF of $\widetilde{T}_{\mathrm{ad}}$. Then we can also define the theoretical tail-adaptive test's $P$-value as $\widetilde{P}_{\mathrm{ad}} := \widetilde{F}_{T,\mathrm{ad}}(\widetilde{T}_{\mathrm{ad}})$. Recall $\widehat{P}_{\mathrm{ad}}$ be the low cost bootstrap $P$-value for $\widetilde{P}_{\mathrm{ad}}$. In what follows, we show that $\widehat{P}_{\mathrm{ad}}$

converges to $\widetilde{P}_{\text{ad}}$ in probability as $n, p, B \to \infty$.

We introduce Assumption **E.1**$'$ to describe the scaling relationships among $n$, $p$, and $s_0$. Let $\boldsymbol{G}_i = (G_{i1}, \ldots, G_{ip})^\top$ with $\boldsymbol{G}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ being i.i.d. Gaussian random vectors, where $\boldsymbol{\Sigma} := \text{Cov}(\boldsymbol{X}_1)$. Define

$$\boldsymbol{C}^{\boldsymbol{G}}(t) = \frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{G}_i - \frac{\lfloor nt \rfloor}{n} \sum_{i=1}^{n} \boldsymbol{G}_i\Big) \quad \text{and} \quad T^{\boldsymbol{G}} = \max_{q_0 \le t \le 1-q_0} \|\boldsymbol{C}^{\boldsymbol{G}}(t)\|_{(s_0,2)}.$$

As shown in the proof of Theorem 2, we use $T^{\boldsymbol{G}} = \max_{q_0 \le t \le 1-q_0} \left\|\boldsymbol{C}^{\boldsymbol{G}}(t)\right\|_{(s_0,2)}$ to approximate $T_\alpha$. For $T^{\boldsymbol{G}}$, let $f_{T^{\boldsymbol{G}}}(x)$ and $c_{T^{\boldsymbol{G}}}(\gamma)$ be the probability density function (pdf), and the $\gamma$-quantile of $T^{\boldsymbol{G}}$, respectively. We then define $h(\epsilon)$ as $h(\epsilon) = \max_{x \in I(\epsilon)} f^{-1}_{T^{\boldsymbol{G}}}(x)$, where $I(\epsilon) := [c_{T^{\boldsymbol{G}}}(\epsilon), c_{T^{\boldsymbol{G}}}(1-\epsilon)]$.

With the above definitions, we now introduce Assumption **E.1**$'$:

**(E.1)**$'$ For any $0 < \epsilon < 1$, we require $h^{0.6}(\epsilon) s_0^3 \log(pn) = o(n^{1/10})$.

Note that Assumption **E.1**$'$ is more stringent than Assumption **E.1**. The intuition of Assumption **E.1**$'$ is that, we construct our tail-adaptive testing statistic by taking the minimum $P$-values of the individual tests. For analyzing the combinational tests, we need not only the uniform convergence of the distribution functions, but also the uniform convergence of their quantiles on $[\epsilon, 1-\epsilon]$ for any $0 < \epsilon < 1$.

The following Theorem 6 justifies the validity of the low-cost bootstrap procedure in Section 2.4.

**Theorem 6.** *For $T_{\text{ad}}$, suppose Assumptions A - D, E.1′, E.2 - E.4 hold. Under $\mathbf{H}_0$, we have*

$$\mathbb{P}(\Psi_{\gamma,\text{ad}} = 1) \to \gamma, \quad \textit{and} \quad \widehat{P}_{\text{ad}} - \widetilde{P}_{\text{ad}} \xrightarrow{\mathbb{P}} 0, \ \textit{as } n, p, B \to \infty.$$

We now discuss the power. Theorem 7 shows that under some regularity conditions, our tail-adaptive test has its power converging to one.

**Theorem 7.** *Let $\epsilon_n := O(s_0^{1/2} s\sqrt{\log(pn)/n}) \vee O(s_0^{1/2} s^2 \sqrt{\log(pn)/n}\|\mathbf{\Delta}\|_{(s_0,2)})$. Suppose Assumptions A - D, E.2 - E.4 as well as $\lim_{n,p\to\infty} s_0^{1/2} s^2 \sqrt{\log(p)/n}\|\mathbf{\Delta}\|_{(s_0,2)} = 0$ hold. If $\mathbf{H}_1$ holds with*

$$\sqrt{n} \times \|\boldsymbol{D}\|_{(s_0,2)} \geq \frac{C(\widetilde{\boldsymbol{\tau}}, \alpha)}{1 - \epsilon_n} s_0^{1/2} \big(\sqrt{\log(pn)} + \sqrt{\log(|\mathcal{A}|/\gamma)}\big), \tag{3.7}$$

*then for $T_{\text{ad}}$, we have*

$$\mathbb{P}(\Psi_{\gamma,\text{ad}} = 1) \to 1 \ \textit{as } n, p, B \to \infty,$$

*where $C(\widetilde{\boldsymbol{\tau}}, \alpha)$ is some positive constant only depending on $\widetilde{\boldsymbol{\tau}}$ and $\alpha$.*

Note that based on the theoretical results obtained in Section 3.2, Theorems 6 and 7 can be proved using some modifications of the proofs of Theorems 3.5 and 3.7 in Zhou et al. (2018). Hence, we omit the detailed proofs for brevity. Lastly, recall the tail-adaptive based change point estimator $\widehat{t}_{\text{ad}} = \widehat{t}_{\widehat{\alpha}}$ with $\widehat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} \widehat{P}_\alpha$. According to Theorem 3, the tail-adaptive estimator is also consistent.

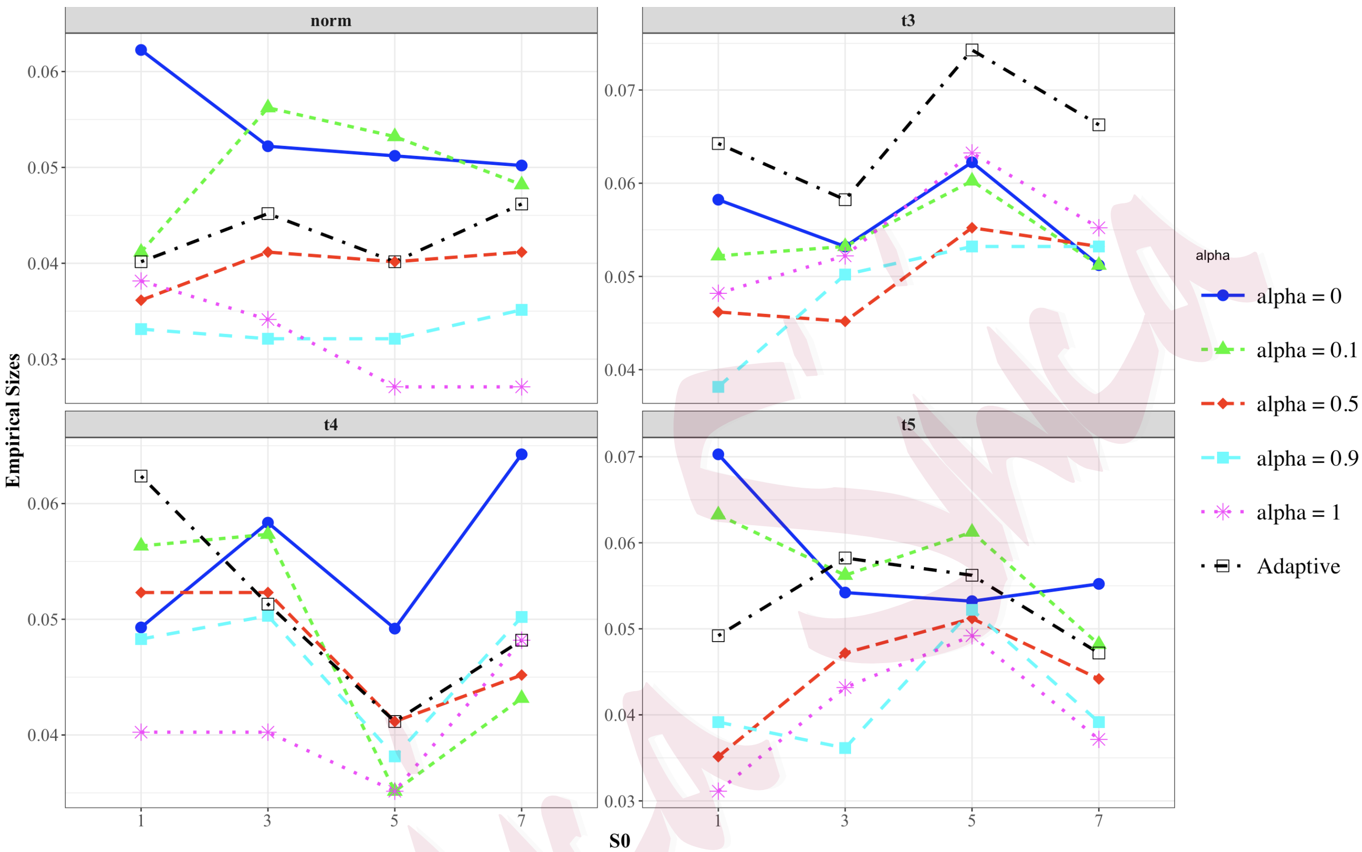


Figure 2: Empirical sizes of the individual and tail-adaptive tests for models with banded covariance matrix under the setting of $(n, p) = (200, 400)$. The results are based on 1000 replications.

## 4. Simulation Studies

We have carried out extensive numerical studies to examine the finite sample performance of our proposed new methods. To save space, we put the detailed model settings and results in Appendix S2 of the supplementary materials. The simulation results, including size, power, and single and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337

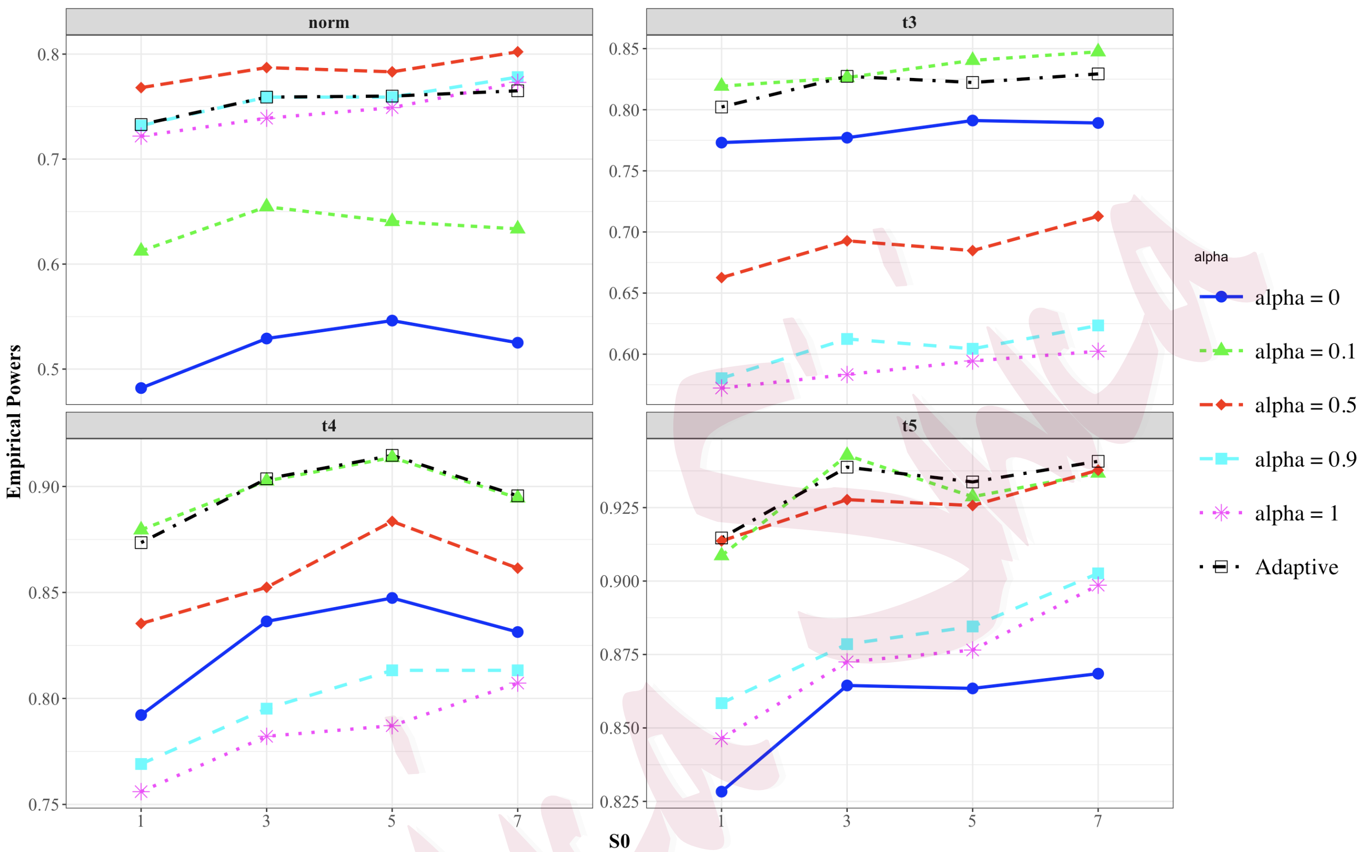


Figure 3: Empirical powers of the individual and tail-adaptive tests for models with banded covariance matrix under the setting of $(n, p) = (200, 400)$. The change point is at $t_1 = 0.5$. The results are based on 1000 replications with $B = 200$ for each replication.

multiple change point estimation, can be summarized as follows:

**(i)** Figure 2 shows that the proposed individual and tail-adaptive tests can control the size very well under various model settings with different tail structures including both lighted and heavy tails. The individual test

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337

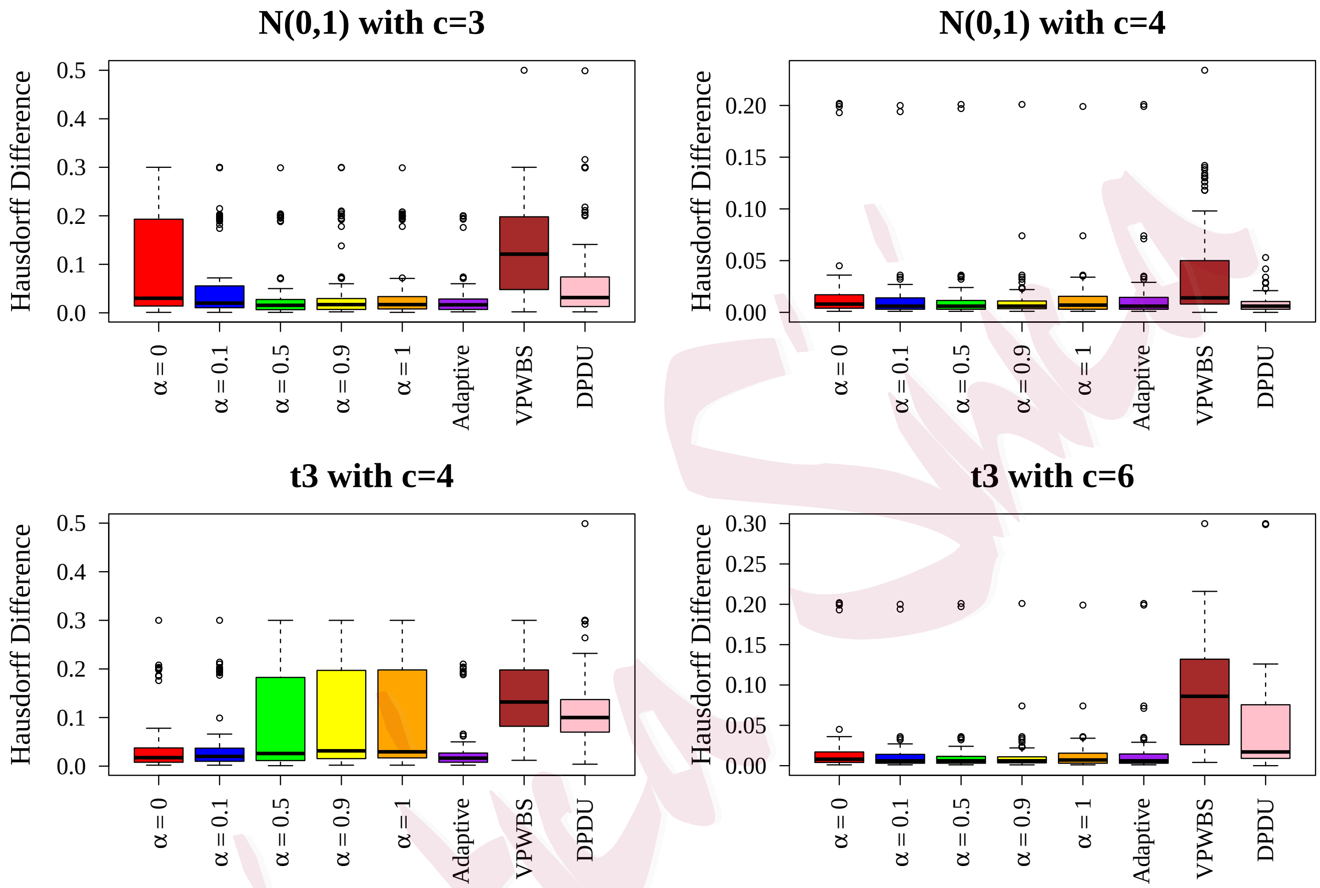


Figure 4: Boxplots of the scaled Hausdorff distance of different methods for detecting multiple change points based on 100 replications. The three change points are at $(0.25, 0.5, 0.75)$. VPWBS and DPDU are techniques in Wang et al. (2021) and Xu et al. (2022). The constant $c$ represents the signal strength and a larger $c$ denotes stronger signal jump.

with $\alpha = 0$ can even control the size well for Student's $t_2$ and Cauchy distributions. **(ii)** In terms of power performance, as shown in Figure 3, the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337

individual tests perform differently under various tail structures. However, the tail-adaptive method can have powers close to its best individual one whenever the errors are lighted or heavy-tailed. **(iii)** For multiple change point estimation, similar to the power analysis, the performance of the individual estimators depends on the underlying error distributions. Figure 4 indicates that the tail-adaptive estimator can perform close to its best individual estimator. Moreover, compared with the existing techniques, the tail-adaptive method enjoys better performance for multiple change point detection. **(iv)** For the choice of $s_0$, the size performance is stable across different choices of $s_0$ under $\mathbf{H}_0$. Moreover, under $\mathbf{H}_1$, choosing $s_0 > 1$ can have high powers and accuracies than using $s_0 = 1$ for change point testing and estimation. In practice, we recommend taking $s_0 = \lfloor \log(p) \rfloor$. In summary, the numerical results are consistent with our theory developed in Section 3 and demonstrate the advantages of our tail-adaptive method over the existing methods.

## 5. Summary

In this article, we propose a general tail-adaptive approach for simultaneous change point testing and estimation for high-dimensional linear regression models. The method is based on the observation that both the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0337

conditional mean and quantile change if the regression coefficients have a change point. Built on a weighted composite loss, we propose a family of individual testing statistics with different weights to account for the unknown tail structures. Then, we combine the individual tests to construct a tail-adaptive method, which is powerful against sparse alternatives under various tail structures. In theory, with mild conditions on the regression covariates and errors, we show the optimality of our methods theoretically in terms of size, power, and change point estimation. In the presence of multiple change points, we combine our tail-adaptive approach with the WBS technique to detect multiple change points. With extensive numerical studies, our proposed methods have better performance than the existing methods under various model setups.

**Acknowledgments**

The authors are indebted to the editor, the associate editor, and two referees, whose helpful comments and suggestions led to a much improved presentation. Bin Liu was supported in part by the National Natural Science Foundation of China 12101132, 12331009.

## Supplementary Materials

The online Supplementary Material provides detailed basic assumptions and proofs of the main theory, and additional numerical results including size, power, multiple change point detection. In addition, an interesting application to the S&P100 data analysis is also provided.

## References

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

Dette, H., J. Gösmann, et al. (2018). Relevant change points in high dimensional time series. Electronic Journal of Statistics 12(2), 2578–2636.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. The Annals of Statistics 42(6), 2243–2281.

Horváth, L. (1995). Detecting changes in linear regressions. Statistics 26(3), 189–208.

Jiang, F., Z. Zhao, and X. Shao (2022). Modelling the covid-19 infection trajectory: A piecewise linear quantile trend model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84(5), 1589–1607.

Jirak, M. (2015). Uniform change point tests in high dimension. The Annals of Statistics 43(6), 2451–2483.

Kaul, A., V. K. Jandhyala, and S. B. Fotopoulos (2019). An efficient two step algorithm for high dimensional change point regression models without grid search. Journal of Machine Learning Research 20(111), 1–40.

Koenker, R. and G. Bassett (1978). Regression quantiles. Econometrica 46, 33–50.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2018). Oracle estimation of a change point in high dimensional quantile regression. Journal of the American Statistical Association 113(523), 1184–1194.

Lee, S., M. H. Seo, and Y. Shin (2016). The lasso for high dimensional regression with a possible change point. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78(1), 193–210.

Leonardi, F. and P. Bühlmann (2016). Computationally efficient change point detection for high-dimensional regression. arXiv preprint: 1601.03704.

Liu, B., C. Zhou, , X.-S. Zhang, and Y. Liu (2020). A unified data-adaptive framework for high dimensional change point detection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(4), 933–963.

Oka, T. and Z. Qu (2011). Estimating structural changes in regression quantiles. Journal of Econometrics 162(2), 248–267.

Page, E. (1955). Control charts with warning lines. Biometrika 42(1-2), 243–257.

Qu, Z. (2008). Testing for structural change in regression quantiles. Journal of Econometrics 146(1), 170–184.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. Journal of the American Statistical Association 53(284), 873–880.

Rinaldo, A., D. Wang, Q. Wen, R. Willett, and Y. Yu (2021). Localizing changes in high-dimensional regression models. In International Conference on Artificial Intelligence and Statistics, pp. 2089–2097. PMLR.

Roy, S., Y. Atchadé, and G. Michailidis (2017). Change point estimation in high dimensional Markov random-field models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(4), 1187–1206.

Shao, X. and X. Zhang (2010). Testing for change points in time series. Journal of the American Statistical Association 105(491), 1228–1240.

Wang, D., Z. Zhao, K. Z. Lin, and R. Willett (2021). Statistically and computationally efficient change point localization in regression settings. Journal of Machine Learning Research 22(248), 1–46.

Wang, X., B. Liu, Y. Liu, and X. Zhang (2023). Efficient multiple change point detection for high-dimensional generalized linear models. The Canadian Journal of Statistics 51(2), 596–629.

Xia, Y., T. Cai, and T. T. Cai (2018). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. Statistica Sinica 28(1), 63–92.

Xu, H., D. Wang, Z. Zhao, and Y. Yu (2022). Change point inference in high-dimensional regression models under temporal dependence. arXiv preprint arXiv:2207.12453.

Zhang, L., H. J. Wang, and Z. Zhu (2014). Testing for change points due to a covariate threshold in quantile regression. Statistica Sinica 24, 1859–1877.

Zhao, T., M. Kolar, and H. Liu (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. arXiv preprint: 1601.03704.

Zhou, C., W.-X. Zhou, X.-S. Zhang, and H. Liu (2018). A unified framework for testing high dimensional parameters: a data-adaptive approach. arXiv preprint: 1601.03704.

Zou, H. and M. Yuan (2008). Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36(3), 1108–1126.

Department of Statistics and Data Science, School of Management at Fudan University

E-mail: liubin0145@gmail.com; xszhang@fudan.edu.cn

Department of Decision Sciences, George Washington University, U.S.A

E-mail: qizhengling321@gmail.com

Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A

E-mail: yfliu@email.unc.edu