Statistica Sinica Preprint No: SS-2023-0212	
Title	Simultaneous Change Point Detection and Identification for High Dimensional Linear Models
Manuscript ID	SS-2023-0212
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0212
Complete List of Authors	Bin Liu, Xinsheng Zhang and Yufeng Liu
Corresponding Authors	Yufeng Liu
E-mails	yfliu@email.unc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Statistica Sinica

Simultaneous Change Point Detection and Identification for High Dimensional Linear Models

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

[1,2]*Fudan University, China;* [3]*University of North Carolina at Chapel Hill, U.S.A*

Abstract: In this article, we consider change point inference for high dimensional linear models. For change point detection, given any subgroup of variables, we propose a new method for testing the homogeneity of corresponding regression coefficients across the observations. Under some regularity conditions, the proposed new testing procedure controls the type I error asymptotically and is powerful against sparse alternatives and enjoys certain optimality. For change point identification, an "argmax" based change point estimator is proposed which is shown to be consistent for the true change point location. Moreover, combining with the binary segmentation technique, we further extend our new method for detecting and identifying multiple change points. Extensive numerical studies justify the validity of our new method and demonstrate its competitive performance.

Key words and phrases: Change point inference; High dimensions; Linear regression; Multiplier bootstrap; Subgroups.

1. Introduction

Driven by the great improvement of data collection and storage ca-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

pacity, high dimensional linear regression models have attracted a lot of attentions because of its simplicity for interpreting the effect of different variables in predicting the response. Specifically, we are interested in the following model:

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \epsilon,$$

where $Y \in \mathbb{R}$ is the response variable, $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ is the covariate vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional unknown vector of coefficients, and $\epsilon \in \mathbb{R}$ is the error term.

For high dimensional linear regression, the L_1-penalized technique lasso (Tibshirani (1996)) is a popular method for estimating $\boldsymbol{\beta}$. In the past decades, lots of research attentions both in machine learning and statistics have been focused on studying theoretical properties of lasso and other penalized methods. Most of the existing literature on high dimensional linear regression focuses on the case with a homogeneous linear model, where the regression coefficients are assumed invariant across the observations. With many modern complex datasets for analysis in practice, data heterogeneity is a common challenge in many real applications such as economy and genetics. In some applications, the regression coefficients may have a sudden change at some unknown time point, which is called a change point. Typical examples include racial segregation and crime prediction in sociology,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

and financial contagion in economy. For these problems, methods and theories designed for independently and identically ($i.i.d.$) distributed settings are no longer applicable. As a result, ignoring these structural breaks in machine learning applications may lead to misleading results and wrong decision making. For the regression change point problem, a fundamental question is whether the underlying regression model remains homogenous across the observations. To address this issue, in this article, we investigate change point inference for high dimensional linear models. Specifically, let $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ be n ordered independent realizations of $(Y, \boldsymbol{X})$. We aim to detect whether the regression coefficients have a change point during the observations. In particular, let $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ be two p-dimensional vectors of coefficients with $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \ldots, \beta_p^{(1)})^\top$ and $\boldsymbol{\beta}^{(2)} = (\beta_1^{(2)}, \ldots, \beta_p^{(2)})^\top$. We consider the following linear regression model with a possible change point:

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(1)} \mathbf{1}\{1 \le i \le k_*\} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(2)} \mathbf{1}\{k_* + 1 \le i \le n\} + \epsilon_i, \tag{1.1}$$

where k_* is the possible but unknown change point location and $(\epsilon_i)_{i=1}^n$ are the error terms. In this paper, we assume $k_* = \lfloor n t_0 \rfloor$ for some $t_0 \in (0, 1)$. For any given subgroup $\mathcal{G} \subset \{1, \ldots, p\}$, the first goal is to test

$$\begin{aligned} &\mathbf{H}_{0,\mathcal{G}} : \beta_s^{(1)} = \beta_s^{(2)} \ \text{ for all } s \in \mathcal{G} \ \text{ v.s.} \\ &\mathbf{H}_{1,\mathcal{G}} : \text{There exist } s \in \mathcal{G} \text{ and } k_* \in \{1, \ldots, n-1\} \text{ s.t. } \beta_s^{(1)} \neq \beta_s^{(2)}. \end{aligned} \tag{1.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

In other words, under $\mathbf{H}_{0,\mathcal{G}}$, the regression coefficients in each subgroup $\mathcal{G}$ are homogeneous across the observations, and under $\mathbf{H}_{1,\mathcal{G}}$ there is a change point at an unknown time point k_* such that the regression coefficients have a sudden change after k_*. Our second goal of the paper is to identify the change point location once we reject $\mathbf{H}_{0,\mathcal{G}}$ in (1.2). In this paper, we assume that the number of coefficients can be much larger than the number of observations, i.e., $p \succeq n$, which is known as a high dimensional problem.

For the low dimensional setting with a fixed p and $p < n$, change point inference for linear regression models has been well-studied. For example, Quandt (1960) considered testing (1.2) for a simple regression model with $p = 2$. Based on that, several techniques were proposed in the literature. Among them are maximum likelihood ratio tests (Horváth, 1995), partial sums of regression residuals (Bai and Perron, 1998), and the union intersection test (Horváth and Shao, 1995). Moreover, as a special case of linear regression models, Chan et al. (2014) considered change point detection for the autoregressive model. As compared to the broad literature in the low dimensional setting, methods and theory for high dimensional change point inference of (1.1) have not been investigated much until recently. For instance, Lee et al. (2016) considered a high dimensional regression model with a possible change point due to a covariate threshold. Based on the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

L_1/L_0 regularization, Kaul et al. (2019) proposed a two-step algorithm for the detection and estimation of parameters in a high-dimensional change point regression model. As extensions to multiple structural breaks in high dimensional linear models, Leonardi and Bühlmann (2016) proposed fast algorithms for multiple change point estimation based on dynamic programming and binary search algorithms. In addition, Zhang et al. (2015) developed an approach for estimating multiple change points based on sparse group lasso. Wang et al. (2021) proposed a projection-based algorithm for estimating multiple change points. Recently, Cho and Owens (2022); Bai and Safikhani (2023) constructed estimates for the multiple change points in high-dimensional regression models based on methods of moving window and blocked fused lasso. Kaul et al. (2021); Xu et al. (2022) respectively considered the problem of constructing confidence interval for the change point in the context of high dimensional mean vector-based models and linear regression models. Chen et al. (2023) proposed a new method for determing the number of change points with false discovery rate controls. Other related papers include He et al. (2023); Wang et al. (2022).

It is worth noting that the majority of above mentioned papers mainly focus on the estimation of regression coefficients as well as the change point locations by assuming a pre-existing change point in the model. To our

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

best knowledge, the testing problem of (1.2) has not been considered yet. How to make effective change point detection remains to be an urgent but challenging task. To fill this gap, in this article, we consider change point inference in the context of high dimensional linear models.

The main contributions of this paper are as follows. For any pre-specified subgroup $\mathcal{G} \subset \{1, \ldots, p\}$, we propose a new method for testing the homogeneity of corresponding regression coefficients across the observations. For change point detection, the proposed test statistic $T_{\mathcal{G}}$ is constructed based on a weighted L_∞ aggregation, both temporally and spatially, of the process $\{Z_j(\lfloor nt \rfloor)\}_{j\in\mathcal{G}, t\in[\tau_0, 1-\tau_0]}$, where $Z_j(\lfloor nt \rfloor) = \breve{\beta}_j^{(0,t)} - \breve{\beta}_j^{(t,1)}$ with $\breve{\beta}_j^{(0,t)}$ and $\breve{\beta}_j^{(t,1)}$ denoting the de-biased lasso estimators for coordinate j before and after time point $\lfloor nt \rfloor$, respectively. It is shown that $T_{\mathcal{G}}$ is powerful against sparse alternatives with only a few entries in $\mathcal{G}$ having a change point. To approximate its limiting null distribution, a multiplier bootstrap procedure is introduced. The proposed bootstrap can automatically account for the dependence structures of $\{Z_j(\lfloor nt \rfloor)\}_{j\in\mathcal{G}, t\in[\tau_0, 1-\tau_0]}$ and allow the group size $|\mathcal{G}|$ to grow exponentially with the sample size n. Furthermore, to identify the change point location, for each time point $\lfloor nt \rfloor$, we first aggregate the coordinates with the L_∞-norm, then a change point estimator $\widehat{t}_{0,\mathcal{G}}$ is obtained by taking "argmax" with respect to t of the above aggregated

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

process with some proper weights. In addition to single change point detection, by combining with the binary segmentation technique (Vostrikova, 1981), we extend our new algorithm for detecting multiple change points which enjoys better performance than the existing methods.

In terms of theoretical investigation, with mild moment conditions on the covariates and errors in the regression model, we justify the validity of our proposed method in terms of change point detection and identification. In particular, our bootstrap procedure consistently approximates the limiting null distribution of $T_{\mathcal{G}}$, which implies that the proposed new test preserves the pre-specified significance level asymptotically. Furthermore, under $\mathbf{H}_{1,\mathcal{G}}$, our new method is sensitive to sparse alternatives and can reject the null hypothesis with probability tending to one. It is worth mentioning that Xia et al. (2018) considered two sample tests for high dimensional linear regression models. They derived some conditions for consistently distinguishing two sample regression models, which are shown to be minimax optimal. Our requirement for detecting a change point under $\mathbf{H}_{1,\mathcal{G}}$ has the same order as the condition derived in Xia et al. (2018). As for the change point estimation, we prove that our proposed argmax-based change point estimator is consistent for t_0 with an estimation error rate of $\big|\widehat{t}_{0,\mathcal{G}} - t_0\big| = O_p\big(\frac{\log(|\mathcal{G}|n)}{n\|\boldsymbol{\delta}\|^2_{\mathcal{G},\infty}}\big)$, where $\boldsymbol{\delta} := \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

with $\|\boldsymbol{\delta}\|_{\mathcal{G},\infty} =: \max_{j\in\mathcal{G}} |\beta_j^{(1)} - \beta_j^{(2)}|$. Hence, the above estimation result shows that our proposed change point estimator is consistent as long as $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathcal{G},\infty} \gg \sqrt{\log(|\mathcal{G}|n)/n}$ and allows the overall sparsity of regression coefficients and the group's magnitude $|\mathcal{G}|$ to grow simultaneously with the sample size n. We demonstrate that our new testing procedure is relatively simple to implement and extensive numerical studies provide strong support to our theory. Moreover, an R package called "RegCpt" is developed to implement our proposed new algorithms.

The rest of this paper is organized as follows. In Section 2, we introduce our new methodology for Problem (1.2). In Section 3, some theoretical results are derived in terms of change point detection and identification. In Section 4, extensive numerical studies are investigated. The detailed proofs of the main theorems, additional numerical studies and an interesting real data application are given in the Appendix.

For $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, 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 a subset $\mathcal{G} \subset \{1, \ldots, p\}$, denote $\|\boldsymbol{v}\|_{\mathcal{G},\infty}$ by $\max_{j\in\mathcal{G}} |v_j|$. 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

constants c and C such that $c|b_n| \leq |a_n| \leq C|b_n|$ for a sufficiently large n. For a sequence of random variables $\{\xi_1, \xi_2, \ldots\}$, we denote $\xi_n = o_p(1)$ if $\xi_n \xrightarrow{\mathbb{P}} 0$. Define $\lfloor x \rfloor$ as the largest integer less than or equal to x for $x \geq 0$.

2. Methodology

2.1 New test statistic

We present our methodology for testing the existence of a change point in Model (1.1). To this end, we first introduce some basic model settings. Recall the regression model

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(1)} \mathbf{1}\{1 \leq i \leq \lfloor nt_0 \rfloor\} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(2)} \mathbf{1}\{\lfloor nt_0 \rfloor + 1 \leq i \leq n\} + \epsilon_i. \quad (2.1)$$

Denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ as a $n \times 1$ response vector, $\mathbf{X}$ is a $n \times p$ design matrix with $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^\top$ being its i-th row for $1 \leq i \leq n$, and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^\top$ is the error vector. For the unknown $p \times 1$ regression vectors $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \ldots, \beta_p^{(1)})^\top$ and $\boldsymbol{\beta}^{(2)} = (\beta_1^{(2)}, \ldots, \beta_p^{(2)})^\top$, define $\mathcal{S}^{(1)} = \{1 \leq j \leq p : \beta_j^{(1)} \neq 0\}$ and $\mathcal{S}^{(2)} = \{1 \leq j \leq p : \beta_j^{(2)} \neq 0\}$ as the active sets of variables. Denote $s^{(1)} = |\mathcal{S}^{(1)}|$ and $s^{(2)} = |\mathcal{S}^{(2)}|$ as the cardinalities of $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$, respectively. Define $\boldsymbol{\Sigma} = (\Sigma_{i,j}) = \mathrm{Cov}(\boldsymbol{X}_1)$ as the covariance matrix of $\boldsymbol{X}_1$ and $\boldsymbol{\Theta} = (\theta_{i,j})$ as the inverse of $\boldsymbol{\Sigma}$. For $\boldsymbol{\Theta}$, let $s_j = |\{1 \leq k \leq p : \theta_{j,k} \neq 0, k \neq j\}|$. In addition to the above notations, we assume that the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

change point does not happen at the beginning or end of data observations. In other words, there exists some $\tau_0 \in (0, 0.5)$ such that $t_0 \in [\tau_0, 1-\tau_0]$ holds. Note that the search boundary scales with n by allowing $\tau_0 \to 0$.

To propose our method, we first introduce the de-sparsified (de-biased) lasso estimator, which was proposed in Van de Geer et al. (2014) and Zhang and Zhang (2014). Specifically, for Model (2.1), let $\widehat{\boldsymbol{\beta}}_n$ be a lasso estimator from $\widehat{\boldsymbol{\beta}}^n = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/n + 2\lambda_n\|\boldsymbol{\beta}\|_1$, where λ_n is the non-negative regularization parameter. Then for a homogeneous model with no change points, the de-biased lasso estimator is defined:

$$\breve{\boldsymbol{\beta}}^n = \widehat{\boldsymbol{\beta}}^n + \widehat{\boldsymbol{\Theta}}\mathbf{X}^\top\big(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^n\big)/n, \tag{2.2}$$

where $\widehat{\boldsymbol{\Theta}}$ is some appropriate estimator for $\boldsymbol{\Theta}$. Essentially, the de-biased lasso estimator $\breve{\boldsymbol{\beta}}_n$ is a lasso solution by plugging in a Karush-Kuhn-Tucker (KKT) condition. It has been widely used for constructing confidence intervals and statistical tests for high dimensional parameters, and proven to be asymptotically optimal in terms of semiparametric efficiency.

Remark 1. In this paper, we adopt the node-wise estimation for obtaining $\widehat{\boldsymbol{\Theta}}$, as proposed in Meinshausen and Bühlmann (2006). The main idea is to perform regression on each variable using the remaining ones. In particular, denote $\mathbf{X}^j$ as the j-th column of $\mathbf{X}$ and $\mathbf{X}^{-j}$ as the remaining columns. For

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

each $j = 1 \ldots, p$, define

$$\widehat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}}{\arg\min} \Big(\|\mathbf{X}^j - \mathbf{X}^{-j}\boldsymbol{\gamma}\|_2^2/n + 2\lambda_{(j)}\|\boldsymbol{\gamma}\|_1 \Big), \tag{2.3}$$

with $\widehat{\boldsymbol{\gamma}}_j = \{\widehat{\gamma}_{j,k} : k = 1 \ldots, p, k \neq j\}$. Denote by $\widehat{\mathbf{C}} = (\widehat{c}_{i,j})_{i,j}^p$ with $\widehat{c}_{i,i} = 1$ and $\widehat{c}_{i,j} = -\gamma_{i,j}$ for $i \neq j$. Let $\widehat{\tau}_j^2 = \|\mathbf{X}^j - \mathbf{X}^{-j}\widehat{\boldsymbol{\gamma}}_j\|_2^2/n + \lambda_{(j)}\|\widehat{\boldsymbol{\gamma}}_j\|_1$ and $\widehat{\mathbf{T}}^2 = \text{diag}\{\widehat{\tau}_1^2, \ldots, \widehat{\tau}_p^2\}$. The node-wise lasso estimator for $\boldsymbol{\Theta}$ is defined as

$$\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}. \tag{2.4}$$

It is shown that $\widehat{\boldsymbol{\Theta}}$ enjoys good properties in estimation accuracy. More importantly, it is possible to use parallel computation for calculating $\widehat{\boldsymbol{\Theta}}$, which is more appropriate for modern statistical applications with large scale datasets.

Since there is a possible but unknown change point in Model (2.1), we can not use (2.2) directly to make statistical inferences on $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$. The main challenge comes from the unknown change point t_0. To overcome this difficulty, instead of only calculating a single de-biased lasso estimator $\breve{\boldsymbol{\beta}}^n$, we need to construct the de-biased lasso-based process. To that end, we need some notations. For any $0 \leq s < t \leq 1$, define

$$\boldsymbol{Y}_{(s,t)} = (Y_{\lfloor ns \rfloor + 1}, \ldots, Y_{\lfloor nt \rfloor})^\top, \quad \boldsymbol{\epsilon}_{(s,t)} = (\epsilon_{\lfloor ns \rfloor + 1}, \ldots, \epsilon_{\lfloor nt \rfloor})^\top,$$
$$\mathbf{X}_{(s,t)} = (\boldsymbol{X}_{\lfloor ns \rfloor + 1}, \ldots, \boldsymbol{X}_{\lfloor nt \rfloor})^\top, \widehat{\boldsymbol{\Sigma}}_{(s,t)} = \frac{1}{\lfloor nt \rfloor - \lfloor ns \rfloor + 1} \sum_{i = \lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \boldsymbol{X}_i \boldsymbol{X}_i^\top.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

To motivate our testing statistic, for each fixed $t \in [\tau_0, 1-\tau_0]$, we define

$$\boldsymbol{\beta}^{(0,t)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}\|\boldsymbol{Y}_{(0,t)} - \mathbf{X}_{(0,t)}\boldsymbol{\beta}\|_2^2, \boldsymbol{\beta}^{(t,1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}\|\boldsymbol{Y}_{(t,1)} - \mathbf{X}_{(t,1)}\boldsymbol{\beta}\|_2^2. \tag{2.5}$$

By definition, $\boldsymbol{\beta}^{(0,t)}$ and $\boldsymbol{\beta}^{(t,1)}$ are the best regression coefficients for predicting $\boldsymbol{Y}_{(0,t)}$ and $\boldsymbol{Y}_{(t,1)}$ under the squared error loss, respectively. More importantly, suppose there is a change point t_0 in the linear model (2.1). According to the search location t and the true change point location t_0, the underlying true parameters can have the following explicit form:

$$\boldsymbol{\beta}^{(0,t)} = \boldsymbol{\beta}^{(1)}\mathbf{1}\{t \in [\tau_0, t_0]\} + \big(\frac{\lfloor nt_0\rfloor}{\lfloor nt\rfloor}\boldsymbol{\beta}^{(1)} + \frac{\lfloor nt\rfloor - \lfloor nt_0\rfloor}{\lfloor nt\rfloor}\boldsymbol{\beta}^{(2)}\big)\mathbf{1}\{t \in [t_0, 1-\tau_0]\},$$

and

$$\boldsymbol{\beta}^{(t,1)} = \big(\frac{\lfloor nt_0\rfloor - \lfloor nt\rfloor}{n - \lfloor nt\rfloor}\boldsymbol{\beta}^{(1)} + \frac{n - \lfloor nt_0\rfloor}{n - \lfloor nt\rfloor}\boldsymbol{\beta}^{(2)}\big)\mathbf{1}\{t \in [\tau_0, t_0]\} + \boldsymbol{\beta}^{(2)}\mathbf{1}\{t \in [t_0, 1-\tau_0]\}.$$

From the population level, we can define the theoretical signal jump process:

$$\begin{aligned}\boldsymbol{\delta}_n(t) &:= \sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}(\boldsymbol{\beta}^{(0,t)} - \boldsymbol{\beta}^{(t,1)}) \\ &= \sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt_0\rfloor^*}{n}\big(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\big)\mathbf{1}\{t \in [\tau_0, t_0]\} \\ &+ \sqrt{n}\frac{\lfloor nt_0\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}\big(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\big)\mathbf{1}\{t \in [t_0, 1-\tau_0]\},\end{aligned} \tag{2.6}$$

where $\lfloor nt\rfloor^* := n - \lfloor nt\rfloor$.

The signal function in (2.6) has some interesting properties. First, under $\mathbf{H}_{0,\mathcal{G}}$ of no change points, it reduces to a vector of zeros at each

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

time point $\lfloor nt \rfloor$. Second, under $\mathbf{H}_{1,\mathcal{G}}$, $\boldsymbol{\delta}_n(t)$ is at most $(s^{(1)} + s^{(2)})$-sparse since we require sparse regression coefficients in the model. Third, we can see that $\|\boldsymbol{\delta}_n(t)\|_{\mathcal{G},\infty}$ with $t \in [\tau_0, 1-\tau_0]$ obtains its maximum value at the true change point location t_0. Hence, to make change point inference for high dimensional linear models, the key point is how to propose a test statistic that can estimate $\boldsymbol{\delta}_n(t)$ well under $\mathbf{H}_{1,\mathcal{G}}$, and has some theoretically tractable limiting null distributions under $\mathbf{H}_{0,\mathcal{G}}$. A natural idea is to use the lasso estimators directly. Specifically, for each time point, we obtain the lasso estimators $\widehat{\boldsymbol{\beta}}^{(0,t)} = (\widehat{\beta}_1^{(0,t)}, \ldots, \widehat{\beta}_p^{(0,t)})^\top$ and $\widehat{\boldsymbol{\beta}}^{(t,1)} = (\widehat{\beta}_1^{(t,1)}, \ldots, \widehat{\beta}_p^{(t,1)})^\top$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(0,t)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2\lfloor nt \rfloor} \left\| \mathbf{Y}_{(0,t)} - \mathbf{X}_{(0,t)}\boldsymbol{\beta} \right\|_2^2 + \lambda_1(t)\|\boldsymbol{\beta}\|_1, \\ \widehat{\boldsymbol{\beta}}^{(t,1)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2\lfloor nt \rfloor_*} \left\| \mathbf{Y}_{(t,1)} - \mathbf{X}_{(t,1)}\boldsymbol{\beta} \right\|_2^2 + \lambda_2(t)\|\boldsymbol{\beta}\|_1, \end{aligned} \tag{2.7}$$

where $\lambda_1(t)$ and $\lambda_2(t)$ are some regularity parameters to account for the data heterogeneity. It is well known that due to the ℓ_1 regularized penalization in (2.7), the lasso estimators are typically biased and do not have a tractable limiting null distribution. As a result, some"de-biased" process is needed. The main idea is to plug into the KKT conditions under both $\mathbf{H}_{0,\mathcal{G}}$ and $\mathbf{H}_{1,\mathcal{G}}$ for the change point model. To give an insight into the de-biased process for change point detection, in what follows, we assume $\mathbf{H}_{1,\mathcal{G}}$ holds.

Firstly, we consider the case that the search location satisfies $t \in [\tau_0, t_0]$. Let $\widehat{\boldsymbol{\kappa}}_1(t) \in \mathbb{R}^p$ and $\widehat{\boldsymbol{\kappa}}_2(t) \in \mathbb{R}^p$ be the subdifferentials of $\|\boldsymbol{\beta}\|_1$ for the first

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

and second optimization problems in (2.7), respectively. Then, by the KKT condition, we have:

$$\begin{aligned} &-\mathbf{X}_{(0,t)}^{\top}(\boldsymbol{Y}_{(0,t)}-\mathbf{X}_{(0,t)}\widehat{\boldsymbol{\beta}}^{(0,t)})/\lfloor nt\rfloor+\lambda_1(t)\widehat{\boldsymbol{\kappa}}_1(t)=\mathbf{0},\\ &-\mathbf{X}_{(t,1)}^{\top}(\boldsymbol{Y}_{(t,1)}-\mathbf{X}_{(t,1)}\widehat{\boldsymbol{\beta}}^{(t,1)})/\lfloor nt\rfloor^{*}+\lambda_2(t)\widehat{\boldsymbol{\kappa}}_2(t)=\mathbf{0}. \end{aligned} \tag{2.8}$$

Note that for $t\in[\tau_0,t_0]$, the samples $\{\boldsymbol{Y}_{(0,t)},\mathbf{X}_{(0,t)}\}$ are homogeneous with regression coefficients being $\boldsymbol{\beta}^{(0,t)}=\boldsymbol{\beta}^{(1)}$. Hence, similar to the analysis in Van de Geer et al. (2014), for the first term in (2.8), for $t\in[\tau_0,t_0]$, we have the following decomposition:

$$\widehat{\boldsymbol{\beta}}^{(0,t)}+\widehat{\boldsymbol{\Theta}}\lambda_1(t)\widehat{\boldsymbol{\kappa}}_1(t)-\boldsymbol{\beta}^{(1)}=\mathbf{X}_{(0,t)}^{\top}\boldsymbol{\epsilon}_{(0,t)}/\lfloor nt\rfloor\overbrace{-(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{(0,t)}-\mathbf{I})(\widehat{\boldsymbol{\beta}}^{(0,t)}-\boldsymbol{\beta}^{(1)})}^{\boldsymbol{\Delta}_{(0,t)}^{I}}. \tag{2.9}$$

For the second term in (2.8), we note that the samples $\{\boldsymbol{Y}_{(t,1)},\mathbf{X}_{(t,1)}\}$ with $t\in[\tau_0,t_0]$ are heterogeneous due to the change point at t_0. Observe that

$$\mathbf{X}_{(t,1)}^{\top}=(\mathbf{X}_{(t,t_0)}^{\top},\mathbf{X}_{(t_0,1)}^{\top}),\ \text{and}\ \mathbf{Y}_{(t,1)}=((\mathbf{X}_{(t,t_0)}\boldsymbol{\beta}^{(1)})^{\top}+\boldsymbol{\epsilon}_{(t,t_0)}^{\top},(\mathbf{X}_{(t_0,1)}\boldsymbol{\beta}^{(2)})^{\top}+\boldsymbol{\epsilon}_{(t_0,1)}^{\top})^{\top}.$$

Then, the KKT condition for the second equation in (2.8) becomes:

$$\begin{aligned} &\lambda_2(t)\widehat{\boldsymbol{\kappa}}_2(t)\\ &=\mathbf{X}_{(t,t_0)}^{\top}\mathbf{X}_{(t,t_0)}(\boldsymbol{\beta}^{(1)}-\widehat{\boldsymbol{\beta}}^{(t,1)})/\lfloor nt\rfloor^{*}+\mathbf{X}_{(t_0,1)}^{\top}\mathbf{X}_{(t_0,1)}(\boldsymbol{\beta}^{(2)}-\widehat{\boldsymbol{\beta}}^{(t,1)})/\lfloor nt\rfloor^{*}+\mathbf{X}_{(t,1)}^{\top}\boldsymbol{\epsilon}_{(t,1)}/\lfloor nt\rfloor^{*}\\ &=\widehat{\boldsymbol{\Sigma}}_{(t,1)}(\boldsymbol{\beta}^{(2)}-\widehat{\boldsymbol{\beta}}^{(t,1)})+\frac{\lfloor nt_0\rfloor-\lfloor nt\rfloor}{\lfloor nt\rfloor^{*}}\widehat{\boldsymbol{\Sigma}}_{(t,t_0)}(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)})+\mathbf{X}_{(t,1)}^{\top}\boldsymbol{\epsilon}_{(t,1)}/\lfloor nt\rfloor^{*}. \end{aligned} \tag{2.10}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Multiplying $\widehat{\boldsymbol{\Theta}}$ on both sides of (2.10), for the case of $t \in [\tau_0, t_0]$, we have:

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}}^{(t,1)} + \widehat{\boldsymbol{\Theta}}\lambda_2(t)\widehat{\boldsymbol{\kappa}}_2(t) - \underbrace{\Big(\frac{\lfloor nt_0\rfloor - \lfloor nt\rfloor}{\lfloor nt\rfloor^*}\boldsymbol{\beta}^{(1)} + \frac{n - \lfloor nt_0\rfloor}{\lfloor nt\rfloor^*}\boldsymbol{\beta}^{(2)}\Big)}_{\boldsymbol{\beta}^{(t,1)}} \\
&= \underbrace{-(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{(t,1)} - \mathbf{I})(\widehat{\boldsymbol{\beta}}^{(t,1)} - \boldsymbol{\beta}^{(2)}) - \frac{\lfloor nt_0\rfloor - \lfloor nt\rfloor}{\lfloor nt\rfloor^*}(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{(t,t_0)} - \mathbf{I})(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})}_{\boldsymbol{\Delta}^{I}_{(t,1)}} + \mathbf{X}_{(t,1)}^{\top}\boldsymbol{\epsilon}_{(t,1)}/\lfloor nt\rfloor^*.
\end{aligned} \tag{2.11}$$

Secondly, for the case of $t \in [t_0, 1-\tau_0]$, using a very similar analysis, we can prove that:

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}}^{(0,t)} + \widehat{\boldsymbol{\Theta}}\lambda_1(t)\widehat{\boldsymbol{\kappa}}_1(t) - \underbrace{\Big(\frac{\lfloor nt_0\rfloor}{\lfloor nt\rfloor}\boldsymbol{\beta}^{(1)} + \frac{\lfloor nt\rfloor - \lfloor nt_0\rfloor}{\lfloor nt\rfloor}\boldsymbol{\beta}^{(2)}\Big)}_{\boldsymbol{\beta}^{(0,t)}} = \mathbf{X}_{(0,t)}^{\top}\boldsymbol{\epsilon}_{(0,t)}/\lfloor nt\rfloor + \boldsymbol{\Delta}^{II}_{(0,t)}, \\
&\widehat{\boldsymbol{\beta}}^{(t,1)} + \widehat{\boldsymbol{\Theta}}\lambda_2(t)\widehat{\boldsymbol{\kappa}}_2(t) - \underbrace{\boldsymbol{\beta}^{(2)}}_{\boldsymbol{\beta}^{(t,1)}} = \mathbf{X}_{(t,1)}^{\top}\boldsymbol{\epsilon}_{(t,1)}/\lfloor nt\rfloor^* + \boldsymbol{\Delta}^{II}_{(t,1)},
\end{aligned} \tag{2.12}$$

where the two terms $\boldsymbol{\Delta}^{II}_{(0,t)}$ are $\boldsymbol{\Delta}^{II}_{(t,1)}$ are defined as

$$\begin{aligned}
\boldsymbol{\Delta}^{II}_{(0,t)} &:= -\frac{\lfloor nt\rfloor - \lfloor nt_0\rfloor}{\lfloor nt\rfloor}\big(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{(t_0,t)} - \mathbf{I}\big)\big(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\big) - \big(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{(0,t)} - \mathbf{I}\big)\big(\widehat{\boldsymbol{\beta}}^{(0,t)} - \boldsymbol{\beta}^{(1)}\big), \\
\boldsymbol{\Delta}^{II}_{(t,1)} &:= -\big(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{(t,1)} - \mathbf{I}\big)\big(\widehat{\boldsymbol{\beta}}^{(t,1)} - \boldsymbol{\beta}^{(2)}\big).
\end{aligned}$$

Combining the results in (2.8)-(2.12), for each $t \in [\tau_0, 1-\tau_0]$, we then construct the de-biased lasso estimators $\breve{\boldsymbol{\beta}}^{(0,t)} = (\breve{\beta}_1^{(0,t)}, \ldots, \breve{\beta}_p^{(0,t)})^{\top}$ and $\breve{\boldsymbol{\beta}}^{(t,1)} = (\breve{\beta}_1^{(t,1)}, \ldots, \breve{\beta}_p^{(t,1)})^{\top}$ as follows:

$$\begin{aligned}
\breve{\boldsymbol{\beta}}^{(0,t)} &= \widehat{\boldsymbol{\beta}}^{(0,t)} + \widehat{\boldsymbol{\Theta}}\mathbf{X}_{(0,t)}^{\top}\big(\mathbf{Y}_{(0,t)} - \mathbf{X}_{(0,t)}\widehat{\boldsymbol{\beta}}^{(0,t)}\big)/\lfloor nt\rfloor, \\
\breve{\boldsymbol{\beta}}^{(t,1)} &= \widehat{\boldsymbol{\beta}}^{(t,1)} + \widehat{\boldsymbol{\Theta}}\mathbf{X}_{(t,1)}^{\top}\big(\mathbf{Y}_{(t,1)} - \mathbf{X}_{(t,1)}\widehat{\boldsymbol{\beta}}^{(t,1)}\big)/\lfloor nt\rfloor^*.
\end{aligned} \tag{2.13}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

The construction of our new test statistic comes from our important new derivation (2.13). In particular, under some regularity conditions, the difference between $\breve{\boldsymbol{\beta}}^{(0,t)}$ and $\breve{\boldsymbol{\beta}}^{(t,1)}$ has the following decomposition:

$$\sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}\big(\breve{\boldsymbol{\beta}}^{(0,t)}-\breve{\boldsymbol{\beta}}^{(t,1)}\big) = \underbrace{\boldsymbol{\delta}_n(t)}_{\textbf{Signal function}} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor}\widehat{\boldsymbol{\Theta}}\boldsymbol{X}_i\epsilon_i}_{\textbf{Random noise}} + \underbrace{\sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}(\boldsymbol{R}^{(0,t)}-\boldsymbol{R}^{(t,1)})}_{\textbf{Random bias}}, \tag{2.14}$$

where $\boldsymbol{\delta}_n(t)$ is defined in (2.6), and $\boldsymbol{R}^{(0,t)}$ and $\boldsymbol{R}^{(t,1)}$ are the residuals:

$$\boldsymbol{R}_{(0,t)} = \boldsymbol{\Delta}^{I}_{(0,t)}\mathbf{1}\{t\in[\tau_0,t_0]\} + \boldsymbol{\Delta}^{II}_{(0,t)}\mathbf{1}\{t\in[t_0,1-\tau_0]\},$$
$$\boldsymbol{R}_{(t,1)} = \boldsymbol{\Delta}^{I}_{(t,1)}\mathbf{1}\{t\in[\tau_0,t_0]\} + \boldsymbol{\Delta}^{II}_{(t,1)}\mathbf{1}\{t\in[t_0,1-\tau_0]\}.$$

The above de-biased lasso-based process enjoys several advantages for making change point inference. Firstly, under $\mathbf{H}_{0,\mathcal{G}}$ of no change points, it is the combination of a partial sum-based process plus a random bias term. The latter one can be shown to be negligible. Moreover, under $\mathbf{H}_{1,\mathcal{G}}$, we can see that the de-biased lasso-based process is an asymptotically unbiased estimator for the signal function defined in (2.6), allowing us to make change point detection and identification. The derivation of (2.14) is different from the original de-biased lasso estimator in (2.2) and requires a fundamental modification of Bickel et al. (2009)) to account for data heterogeneity. More details can be found in the Appendix.

Motived by the above observation, for any given subgroup $\mathcal{G}\subset\{1,\ldots,p\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

a natural test statistic for the hypothesis (1.2) is defined as

$$\widetilde{T}_{\mathcal{G}} = \max_{t\in[\tau_0,1-\tau_0]} \max_{j\in\mathcal{G}} \sqrt{n}\frac{\lfloor nt\rfloor}{n}\Big(1-\frac{\lfloor nt\rfloor}{n}\Big)\Big|\breve{\beta}_j^{(0,t)} - \breve{\beta}_j^{(t,1)}\Big|.$$

For any given subgroup $\mathcal{G}$, the proposed new statistic $\widetilde{T}_{\mathcal{G}}$ searches all possible locations of time points. It is demonstrated that $\widetilde{T}_{\mathcal{G}}$ is powerful against sparse alternatives with only a few entries in $\mathcal{G}$ having a change point, and a large value of $\widetilde{T}_{\mathcal{G}}$ leads to a rejection of $\mathbf{H}_{0,\mathcal{G}}$.

2.2 Weighted variance estimation

In Section 2.1, we introduced $\widetilde{T}_{\mathcal{G}}$ for the hypothesis (1.2). Considering the variability of the design matrix $\mathbf{X}$ and the error term $\boldsymbol{\epsilon}$, the test statistic $\widetilde{T}_G$ is heterogeneous. Hence, we need to take its variance into account and standardize it. In this paper, we adopt a weighted variance estimator. Specifically, let $\widehat{\boldsymbol{\Omega}} = (\widehat{\omega}_{i,j})_{i,j}^p = \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\Theta}}^\top$ with $\widehat{\boldsymbol{\Sigma}}_n := \mathbf{X}^\top\mathbf{X}/n$. For each $t \in [\tau_0, 1-\tau_0]$, denote

$$\widehat{\sigma}_\epsilon^2(t) = \frac{1}{n}\big(\big\|\boldsymbol{Y}_{(0,t)} - \mathbf{X}_{(0,t)}\widehat{\boldsymbol{\beta}}^{(0,t)}\big\|_2^2 + \big\|\boldsymbol{Y}_{(t,1)} - \mathbf{X}_{(t,1)}\widehat{\boldsymbol{\beta}}^{(t,1)}\big\|_2^2\big). \tag{2.15}$$

Under $\mathbf{H}_{0,\mathcal{G}}$ of no change points in the model, we can prove that

$$\max_{\tau_0\le t\le 1-\tau_0} \max_{1\le j\le p} |\widehat{\sigma}_\epsilon^2(t)\widehat{\omega}_{j,j} - \sigma_\epsilon^2\omega_{j,j}| = o_p(1).$$

Under $\mathbf{H}_{1,\mathcal{G}}$, however, $\widehat{\sigma}_\epsilon^2(t)$ is not a consistent estimator for σ_ϵ^2 because of the unknown change point t_0. Furthermore, as discussed in Shao and Zhang

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

(2010), an inappropriate variance estimator may lead to non-monotonic power performance. In order to form a powerful test statistic, it is necessary to construct consistent variance estimation for $\mathbf{H}_{0,\mathcal{G}}$ and $\mathbf{H}_{1,\mathcal{G}}$. To address this issue, we need to deal with the unknown change point first. In particular, for a given subgroup $\mathcal{G}$, define

$$H_{\mathcal{G}}(t) = \max_{j\in\mathcal{G}} \frac{\lfloor nt \rfloor}{n}\Big(1 - \frac{\lfloor nt \rfloor}{n}\Big)\Big|\breve{\beta}_j^{(0,t)} - \breve{\beta}_j^{(t,1)}\Big|, \text{ with } t \in [\tau_0, 1-\tau_0].$$

By maximizing $H_{\mathcal{G}}(t)$, we obtain the argmax-based change point estimator:

$$\widehat{t}_{0,\mathcal{G}} = \underset{t\in[\tau_0,1-\tau_0]}{\arg\max}\, H_{\mathcal{G}}(t). \tag{2.16}$$

Based on (2.16), let $\widehat{t}_0 = \widehat{t}_{0,\mathcal{G}}$ with $\mathcal{G} = \{1,\ldots,p\}$. We put $\widehat{t}_0$ into $\widehat{\sigma}_\epsilon^2(t)$ and get a weighted variance estimator for σ_ϵ^2 as

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{n}\big(\big\|\boldsymbol{Y}_{(0,\widehat{t}_0)} - \mathbf{X}_{(0,\widehat{t}_0)}\widehat{\boldsymbol{\beta}}^{(0,\widehat{t}_0)}\big\|_2^2 + \big\|\boldsymbol{Y}_{(\widehat{t}_0,1)} - \mathbf{X}_{(\widehat{t}_0,1)}\widehat{\boldsymbol{\beta}}^{(\widehat{t}_0,1)}\big\|_2^2\big). \tag{2.17}$$

As shown in our theoretical analysis, the new variance estimation in (2.17) is consistent under both $\mathbf{H}_{0,\mathcal{G}}$ and $\mathbf{H}_{1,\mathcal{G}}$. The proof is nontrivial since we need to justify the consistency of $\widehat{t}_{0,\mathcal{G}}$ for t_0, which is known to be an important but difficult task for high dimensional linear models (Lee et al. (2016)).

Using the new variance estimator in (2.17), for any given subgroup $\mathcal{G} \subset \{1,\ldots,p\}$, our new test statistic for the hypothesis (1.2) is finally defined as follows:

$$T_{\mathcal{G}} = \max_{t\in[\tau_0,1-\tau_0]} \max_{j\in\mathcal{G}} \sqrt{n}\frac{\lfloor nt \rfloor}{n}\Big(1 - \frac{\lfloor nt \rfloor}{n}\Big)\Big|\frac{\breve{\beta}_j^{(0,t)} - \breve{\beta}_j^{(t,1)}}{\sqrt{\widehat{\sigma}_\epsilon^2\widehat{\omega}_{j,j}}}\Big|. \tag{2.18}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

2.3 Multiplier bootstrap for approximating the null distribution

In Section 2.2, we have proposed the new test statistic $T_{\mathcal{G}}$ for the hypothesis (1.2). It is challenging to directly obtain its limiting null distribution in high dimensions. Bootstrap has been widely used for making statistical inference on high dimensional linear models since the seminal work of Chernozhukov et al. (2013). For high dimensional linear models with change points, however, existing bootstrap techniques are not applicable and it is desirable to design a new method. To overcome this problem, we investigate two types of multiplier bootstrap.

2.3.1 Bootstrap-I

Recall the decomposition in (2.14). Under $\mathbf{H}_{0,\mathcal{G}}$, we have

$$\sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}(\breve{\boldsymbol{\beta}}^{(0,t)}-\breve{\boldsymbol{\beta}}^{(t,1)})=\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor}\widehat{\boldsymbol{\Theta}}\boldsymbol{X}_i\epsilon_i+(\boldsymbol{R}^{(0,t)}-\boldsymbol{R}^{(t,1)}).$$

It is shown that under $\mathbf{H}_{0,\mathcal{G}}$, the residual-based process $\{\boldsymbol{R}^{(0,t)}-\boldsymbol{R}^{(t,1)},t\in[\tau_0,1-\tau_0]\}$ is asymptotically negligible and the partial sum-based process $\{n^{-1/2}\sum_{i=1}^{\lfloor nt\rfloor}\widehat{\boldsymbol{\Theta}}\boldsymbol{X}_i\epsilon_i,t\in[\tau_0,1-\tau_0]\}$ determines the limiting null distribution of $T_{\mathcal{G}}$, which is known as the leading term. This motivates us to first consider the following bootstrap method:

Step 1: For the b-th bootstrap, generate $i.i.d.$ random variables $\epsilon_1^b,\ldots,\epsilon_n^b$ with $\epsilon_i^b\sim N(0,1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Step 2: Calculate the testing statistic for the b-th bootstrap by

$$W_{\mathcal{G}}^b = \max_{t\in[\tau_0,1-\tau_0]} \max_{j\in\mathcal{G}} \sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}\widehat{\omega}_{j,j}^{-1/2}\Big|\frac{1}{\lfloor nt\rfloor}\sum_{i=1}^{\lfloor nt\rfloor}\widehat{\boldsymbol{\Theta}}_j^\top \mathbf{X}_i\epsilon_i^b - \frac{1}{\lfloor nt\rfloor^*}\sum_{i=\lfloor nt\rfloor+1}^{n}\widehat{\boldsymbol{\Theta}}_j^\top \mathbf{X}_i\epsilon_i^b\Big|,$$

where $\widehat{\boldsymbol{\Theta}}_j^\top$ is the j-th row of $\widehat{\boldsymbol{\Theta}}$.

Step 3: Repeat the above process for B times.

Step 4: Based on the bootstrap samples $\{W_{\mathcal{G}}^1,\ldots,W_{\mathcal{G}}^B\}$, calculate the bootstrap sample-based critical value

$$\widehat{w}_{\mathcal{G},\alpha} = \inf\Big\{t : (B+1)^{-1}\sum_{b=1}^{B}\mathbf{1}\{W_{\mathcal{G}}^b \le t|\mathbf{X},\boldsymbol{Y}\} \ge 1-\alpha\Big\}.$$

Step 5: Reject $\mathbf{H}_{0,\mathcal{G}}$ if and only if $T_{\mathcal{G}} \ge \widehat{w}_{\mathcal{G},1-\alpha}$.

Note that the above bootstrap method essentially bootstraps the partial sum-based process, which has been recently used for change point detection of high dimensional mean vectors in Jirak (2015); Yu and Chen (2021). As shown in our numerical studies, Bootstrap-I suffers from serious size distortions. This phenomenon is due to large biases arising from the residual-based process $\{\boldsymbol{R}^{(0,t)} - \boldsymbol{R}^{(t,1)}, t \in [\tau_0, 1-\tau_0]\}$, which can not be ignored in finite sample performance although it is asymptotically negligible. Hence, for change point detection in high dimensional linear models, substantial modifications are needed and it is desirable to consider a new candidate bootstrap method. To overcome this problem, different from the existing methods, we choose to bootstrap the entire de-biased lasso-based

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

process as shown in the following Bootstrap-II.

2.3.2 Bootstrap-II

The key idea of this bootstrap procedure is to approximate the null limiting distribution under both $\mathbf{H}_{0,\mathcal{G}}$ and $\mathbf{H}_{1,\mathcal{G}}$. It proceeds as follows:

Step 1: Given $\widehat{\sigma}_\epsilon^2$ in (2.17), for the b-th bootstrap, let $\epsilon_1^b, \ldots, \epsilon_n^b$ be *i.i.d.* random variables following $N(0, \widehat{\sigma}_\epsilon^2)$. Define the b-th bootstrap of response vectors $\boldsymbol{Y}^b = (Y_1^b, \ldots, Y_n^b)^\top$:

$$Y_i^b = \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{(0,\widehat{t}_0)} \mathbf{1}\{1 \le i \le \lfloor n\widehat{t}_0 \rfloor\} + \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{(\widehat{t}_0,1)} \mathbf{1}\{\lfloor n\widehat{t}_0 \rfloor < i \le n\} + \epsilon_i^b, \quad (2.19)$$

where $\widehat{\boldsymbol{\beta}}^{(0,\widehat{t}_0)}$ and $\widehat{\boldsymbol{\beta}}^{(\widehat{t}_0,1)}$ are the lasso estimators before and after $\widehat{t}_0$.

Step 2: Denote $\boldsymbol{Y}_{(0,t)}^b = (Y_1^b, \ldots, Y_{\lfloor nt \rfloor}^b)^\top$, and $\boldsymbol{Y}_{(t,1)}^b = (Y_{\lfloor nt \rfloor + 1}^b, \ldots, Y_n^b)^\top$. We then define the b-th bootstrap version of the de-biased lasso estimators before and after $\lfloor nt \rfloor$ as $\breve{\boldsymbol{\beta}}^{b,(0,t)} = (\breve{\beta}_1^{b,(0,t)}, \ldots, \breve{\beta}_p^{b,(0,t)})^\top$ and $\breve{\boldsymbol{\beta}}^{b,(t,1)} = (\breve{\beta}_1^{b,(t,1)}, \ldots, \breve{\beta}_p^{b,(t,1)})^\top$, where

$$\begin{aligned} \breve{\boldsymbol{\beta}}^{b,(0,t)} &:= \widehat{\boldsymbol{\beta}}^{b,(0,t)} + \widehat{\boldsymbol{\Theta}} \mathbf{X}_{(0,t)}^\top \Big(\boldsymbol{Y}_{(0,t)}^b - \mathbf{X}_{(0,t)} \widehat{\boldsymbol{\beta}}^{b,(0,t)} \Big) / \lfloor nt \rfloor, \\ \breve{\boldsymbol{\beta}}^{b,(t,1)} &:= \widehat{\boldsymbol{\beta}}^{b,(t,1)} + \widehat{\boldsymbol{\Theta}} \mathbf{X}_{(t,1)}^\top \Big(\boldsymbol{Y}_{(t,1)}^b - \mathbf{X}_{(t,1)} \widehat{\boldsymbol{\beta}}^{b,(t,1)} \Big) / \lfloor nt \rfloor^*, \end{aligned} \quad (2.20)$$

and $\widehat{\boldsymbol{\beta}}^{b,(0,t)}$ and $\widehat{\boldsymbol{\beta}}^{b,(t,1)}$ are the lasso estimators before and after t using the bootstrap samples $\{\boldsymbol{Y}_{(0,t)}^b, \mathbf{X}_{(0,t)}\}$ and $\{\boldsymbol{Y}_{(t,1)}^b, \mathbf{X}_{(t,1)}\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Step 3: Define the bootstrap sample-based signal function $\widehat{\boldsymbol{\delta}}(t) = (\widehat{\delta}_1(t), \ldots, \widehat{\delta}_p(t))^\top$:

$$\widehat{\boldsymbol{\delta}}(t) = \frac{n - \lfloor n\widehat{t}_0 \rfloor}{n - \lfloor nt \rfloor}\big(\widehat{\boldsymbol{\beta}}^{(0,\widehat{t}_0)} - \widehat{\boldsymbol{\beta}}^{(\widehat{t}_0,1)}\big)\mathbf{1}\{t \in [\tau_0, \widehat{t}_0]\} + \frac{\lfloor n\widehat{t}_0 \rfloor}{\lfloor nt \rfloor}\big(\widehat{\boldsymbol{\beta}}^{(0,\widehat{t}_0)} - \widehat{\boldsymbol{\beta}}^{(\widehat{t}_0,1)}\big)\mathbf{1}\{t \in [\widehat{t}_0, 1-\tau_0]\}.$$

Step 4: Calculate the b-th bootstrap version for the test statistic $T_{\mathcal{G}}$ by

$$T_{\mathcal{G}}^b = \max_{t\in[\tau_0,1-\tau_0]} \max_{j\in\mathcal{G}} \sqrt{n}\frac{\lfloor nt \rfloor}{n}\Big(1 - \frac{\lfloor nt \rfloor}{n}\Big)\Big|\frac{\breve{\beta}_j^{b,(0,t)} - \breve{\beta}_j^{b,(t,1)} - \widehat{\delta}_j(t)}{\sqrt{\widehat{\sigma}_\epsilon^2 \widehat{\omega}_{j,j}}}\Big|. \quad (2.21)$$

Step 5: Repeat the above procedures (2.19)-(2.21) for B times and obtain the bootstrap samples $\{T_{\mathcal{G}}^1, \ldots, T_{\mathcal{G}}^B\}$. Let $c_{\mathcal{G},\alpha} := \inf\{t : \mathbb{P}(T_{\mathcal{G}} \leq t) \geq 1 - \alpha\}$ be the theoretical critical value of $T_{\mathcal{G}}$. Using the bootstrap samples $\{T_{\mathcal{G}}^1, \ldots, T_{\mathcal{G}}^B\}$, we estimate $c_{\mathcal{G},\alpha}$ by

$$\widehat{c}_{\mathcal{G},\alpha} = \inf\Big\{t : (B+1)^{-1}\sum_{b=1}^{B} \mathbf{1}\{T_{\mathcal{G}}^b \leq t | \mathbf{X}, \boldsymbol{Y}\} \geq 1 - \alpha\Big\}. \quad (2.22)$$

Step 6: Define the new test for the hypothesis (1.2) as follows:

$$\Phi_{\mathcal{G},\alpha} = \mathbf{1}\{T_{\mathcal{G}} \geq \widehat{c}_{\mathcal{G},\alpha}\}. \quad (2.23)$$

Given a significance level $\alpha \in (0, 1)$ and a prespecified subgroup $\mathcal{G}$, for the hypothesis (1.2), we reject $\mathbf{H}_{0,\mathcal{G}}$ if $\Phi_{\mathcal{G},\alpha} = 1$.

It is shown in theory that the Bootsrap-II-based test statistic $T_{\mathcal{G}}^b$ approximates the limiting null distribution of $T_{\mathcal{G}}$. More importantly, by bootstrapping the whole de-biased lasso-based process, Bootstrap-II enjoys better test size performance than Bootstrap-I under various candidate subgroups. This is supported by our extensive numerical studies in Section 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

3. Theoretical properties

In this section, we examine some theoretical properties of our proposed method including the size, power and the change point estimation results.

3.1 Basic assumptions

To save space, we provide brief descriptions of our assumptions below. More details on basic assumptions for making change point inference on high dimensional linear models can be found in the Supplementary Materials. Assumptions (**A**.**1**) – (**A**.**3**) impose some regular conditions on the design matrix as well as the error terms. Assumption (**A**.**4**) contains basic requirements on model parameters. Assumption (**A**.**5**) is a technical condition on the regularity parameters in (2.3) and (2.7).

3.2 Main results

We derive some theoretical results of our proposed new test. In Section 3.2.1, we consider the control of Type I error. In Section 3.2.2, we examine the power performance as well as the accuracy of change point estimation.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

3.2.1 The validity of test size

Before giving the test size results, we first consider the variance estimation. Theorem 1 shows that the pooled weighted variance estimator is uniformly consistent under the null hypothesis. It is crucial for deriving the Gaussian approximation results as in Theorem 2.

Theorem 1. *Suppose Assumptions* (**A**.1) – (**A**.5) *hold. Under* $\mathbf{H}_{0,\mathcal{G}}$*, for the variance estimator, with probability at least* $1 - C_1(np)^{-C_2}$*, we have*

$$\max_{1\le j,k\le p} |\widehat{\sigma}_\epsilon^2\widehat{\omega}_{j,k} - \sigma_\epsilon^2\omega_{j,k}| \le C_3\big(\sqrt{\frac{\log(n)}{n}} + \max_j \lambda_{(j)}\sqrt{s_j}\big),$$

where $C_1, \ldots, C_3$ *are universal positive constants not depending on* n *or* p*.*

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

Theorem 2. *Suppose Assumptions* (**A**.1) – (**A**.5) *hold. Under* $\mathbf{H}_{0,\mathcal{G}}$*, for any given subgroup* $\mathcal{G} \subset \{1, \ldots, p\}$*, we have*

$$\sup_{z\in(0,\infty)} \big|\mathbb{P}(T_{\mathcal{G}} \le z) - \mathbb{P}(T_{\mathcal{G}}^b \le z|\{\mathbf{X}, \boldsymbol{Y}\})\big| = o_p(1), \ \text{as } n, p \to \infty.$$

Theorem 2 shows that we can uniformly approximate the distribution of $T_{\mathcal{G}}$ using that of $T_{\mathcal{G}}^b$. As a corollary, the following Corollary 1 shows that our proposed new test can control the Type I error asymptotically for any given pre-specified significance level α.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Corollary 1. *Assume Assumptions* (**A**.**1**)–(**A**.**5**) *hold. Under* $\mathbf{H}_{0,\mathcal{G}}$*, for any given subgroup* $\mathcal{G} \subset \{1, \ldots, p\}$*, we have* $\mathbb{P}(\Phi_{\mathcal{G},\alpha} = 1) \to \alpha$*, as* $n, p, B \to \infty$.

3.2.2 Analysis under $\mathbf{H}_{1,\mathcal{G}}$

After analyzing the theoretical results under the null hypothesis, we next consider the performance under $\mathbf{H}_{1,\mathcal{G}}$. To this end, some additional assumptions are needed.

Assumption (**A**.**6**). Let $\boldsymbol{\delta} = \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$. For the signal jump, we require there exists a constant $c^* \in [0, \infty)$ such that $\lim_{n,p\to\infty} s\|\boldsymbol{\delta}\|_\infty \to c^*$.

Note that Assumption (**A**.**6**) is a signal strength requirement for identifying the change point location t_0 with high accuracy. It allows weak signals that can scale to zero as $(n, p) \to \infty$. With the additional assumption as well as those of (**A**.**1**) – (**A**.**5**), the following Theorem 3 provides a non-asymptotic estimation error bound of $\widehat{t}_{0,\mathcal{G}}$ for t_0.

Theorem 3. *Suppose Assumptions* (**A**.**1**) - (**A**.**6**) *hold. Assume additionally* $\|\boldsymbol{\delta}\|_{\mathcal{G},\infty} \gg \sqrt{\log(|\mathcal{G}|n)/n}$ *holds. For any given subgroup* $\mathcal{G} \subset \{1, \ldots, p\}$*, under* $\mathbf{H}_{1,\mathcal{G}}$*, with probability at least* $1 - C_1(np)^{-C_2}$*, we have*

$$\left|\widehat{t}_{0,\mathcal{G}} - t_0\right| \leq C^* \frac{\log(|\mathcal{G}|n)}{n\|\boldsymbol{\delta}\|^2_{\mathcal{G},\infty}}, \tag{3.1}$$

where C^* *is a universal positive constant not depending on* n *or* p.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Theorem 3 shows that our subgroup-based change point estimator is asymptotically consistent, which allows the group size $|\mathcal{G}|$ to grow with the sample size n as long as $\|\boldsymbol{\delta}\|_{\mathcal{G},\infty} \gg \sqrt{\log(|\mathcal{G}|n)/n}$ holds.

Remark 2. Note that Jirak (2015); Yu and Chen (2021) considered the change point estimation for high dimensional mean vectors. They obtained the change point estimators by taking "argmax" of the corresponding partial sum processes with an estimation error rate of $O_p\big(\log(p)/(n\|\boldsymbol{\Delta}\|^2_{\min})\big)$, where $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_p)^\top$ is the signal jump of mean vectors before and after the change point and $\|\boldsymbol{\Delta}\|_{\min}$ is the minimum signal jump for the coordinates with a change point. Different from Jirak (2015); Yu and Chen (2021), we adopt a different proof technique and derive an estimation error bound of $O_p\big(\log(p)/(n\|\boldsymbol{\Delta}\|^2_{\infty})\big)$. Considering $\|\boldsymbol{\Delta}\|_{\infty}$ can be much larger than $\|\boldsymbol{\Delta}\|_{\min}$, our result is sharper than Jirak (2015); Yu and Chen (2021). More proof details can be found in the Appendix.

After analyzing the change point identification, we next consider the change point detection. 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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

two-sample tests where homogenous data are used to construct consistent variance estimation.

Theorem 4 shows that the pooled weighted variance estimation is uniformly consistent under $\mathbf{H}_{1,\mathcal{G}}$. This guarantees that our new testing method has reasonable power performance.

Theorem 4. *Suppose Assumptions* (**A**.**1**) - (**A**.**6**) *hold. Then, for the weighted variance estimation, under* $\mathbf{H}_{1,\mathcal{G}}$*, we have*

$$\max_{1\le j,k\le p} |\widehat{\sigma}_\epsilon^2\widehat{\omega}_{j,k} - \sigma_\epsilon^2\omega_{j,k}| = o_p(1), \ \ as \ \ n,p\to\infty. \tag{3.2}$$

From the proof of Theorem 4, some interesting observations can be found. On one hand, if the signal strength is too weak such that $\|\boldsymbol{\delta}\|_{\mathcal{G},\infty} = O(\sqrt{\log(pn)/n})$ holds, then the pooled weighted variance estimator $\widehat{\sigma}_\epsilon^2$ is a consistent estimator for σ_ϵ^2 even though we can not guarantee a consistent change point estimator in this case. On the other hand, if the signal strength is big enough such that $\|\boldsymbol{\delta}\|_{\mathcal{G},\infty} \gg \sqrt{\log(pn)/n}$ holds, then a consistent change point estimator $\widehat{t}_{0,\mathcal{G}}$ is needed to guarantee (3.2) holds. These are insightful findings for variance estimation in change point analysis, which is different from the *i.i.d.* case.

Lastly, we discuss the power properties. To this end, we need some additional notations. Recall $\Pi = \{j : \beta_j^{(1)} \neq \beta_j^{(2)}\}$ as the set of coordinates

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

having a change point. Define the oracle signal to noise ratio vector $\boldsymbol{D} = (D_1, \ldots, D_p)^\top$ with

$$D_j := \begin{cases} 0, & \text{for } \ j \in \Pi^c \\ \Big| \dfrac{t_0(1-t_0)(\beta_j^{(2)} - \beta_j^{(1)})}{(\sigma_\epsilon^2 \omega_{j,j})^{1/2}} \Big|, & \text{for } \ j \in \Pi. \end{cases} \tag{3.3}$$

With the above notations and some regularity conditions, the following Theorem 5 shows that we can reject the null hypothesis of no change points with overwhelming probability.

Theorem 5. *Suppose Assumptions* (**A**.**1**) – (**A**.**6**) *hold. Let* $\epsilon_n = o(1)$. *For any given subgroup* $\mathcal{G} \subset \{1, \ldots, p\}$, *if* $\boldsymbol{D}$ *satisfies*

$$\sqrt{n} \big\| \boldsymbol{D} \big\|_{\mathcal{G},\infty} \geq \frac{C_0}{1-\epsilon_n} \Big(\sqrt{2\log(|\mathcal{G}|n)} + \sqrt{2\log(\alpha^{-1})} \Big), \tag{3.4}$$

under $\mathbf{H}_{1,\mathcal{G}}$, *we have* $\mathbb{P}(\Phi_{\mathcal{G},\alpha} = 1) \to 1$, *as* $n, p, B \to \infty$, *where* C_0 *is a large enough universal positive constant not depending on* n *or* p.

Theorem 5 demonstrates that with probability tending to one, our proposed new test can detect the existence of a change point for any given subgroup as long as the corresponding signal to noise ratio satisfies (3.4). Combining (3.3) and (3.4), we note that with a larger signal jump, a smaller noise level, and a closer change point location to the middle of data observations, it is more likely to trigger a rejection of the null hypothesis.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Lastly, we would like to point out that the requirements for identifying and detecting a change point are different. More specifically, from Theorem 3, to correctly identify the location of a change point with desirable accuracy, the signal strength should at least satisfy $\|\boldsymbol{\delta}\|_{\mathcal{G},\infty} \gg \sqrt{\log(|\mathcal{G}|n)/n}$. In contrast, Theorem 5 shows that it is sufficient to detect a change point if $\|\boldsymbol{D}\|_{\mathcal{G},\infty} \geq C\sqrt{\log(|\mathcal{G}|n)/n}$ holds. Hence, we need more stringent conditions for locating a change point than detecting its existence.

4. Numerical studies

We examine the numerical performance of our proposed method and compare it with several existing state-of-art techniques.

We first consider single change point detection. For the design matrix $\mathbf{X}$, we generate $\boldsymbol{X}_i$ $(i.i.d.)$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$, where the following two types of covariance structures are investigated: $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$ (**Model 1**) and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$ with $\boldsymbol{\Sigma}^* = (\sigma^*)_{i,j=1}^p$, where $\sigma_{i,j}^* = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$ (**Model 2**).

To show the bootstrap performance, for each model, the error terms $(\epsilon_i)_{i=1}^n$ are $i.i.d.$ generated from standard normal distributions, standardized Gamma$(4, 1)$ distributions as well as Student's t_5 distributions.

For the regression coefficient $\boldsymbol{\beta}^{(1)}$, for each replication, we generate s non-zero covariates randomly selected from $\mathcal{S} = \{1, \ldots, 50\}$. The corre-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

sponding selected coefficients are *i.i.d.* from $\mathrm{U}(0,2)$, and the remaining $p-s$ covariates are 0's. Note that we generate regression coefficients out of $\mathcal{S}$, which is denoted as the active set. Under $\mathbf{H}_{0,\mathcal{G}}$, we set $\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)}$. Throughout the simulations, we consider various combinations of the sample sizes n, data dimensions p, and overall sparsities s by setting $n \in \{200, 300\}$, $p \in \{100, 200, 300, 400\}$ and $s \in \{5, 10\}$. The number of bootstrap replications is $B = 100$. Without additional specifications, all numerical results are based on 2000 replications.

4.1 Empirical sizes

We investigate the empirical sizes. We set the significance levels $\alpha = 1\%, 5\%$. Furthermore, three different types of subgroups are investigated: $\mathcal{G} = \mathcal{S}$, $\mathcal{G} = \mathcal{S}^c$, and $\mathcal{G} = \mathcal{S} \cup \mathcal{S}^c = \{1, \ldots, p\}$. To evaluate the numerical performance, in addition to our proposed methods, we consider four existing well-known techniques for change point detection of high dimensional linear models: the high dimensional lasso-based method in Lee et al. (2016) (Lee2016), the sparse group lasso-based method in Zhang et al. (2015) (SGL), the binary segmentation-based method in Leonardi and Bühlmann (2016) (L&B), and the Variance Projected Wild Binary Segmentation in Wang et al. (2021) (VPWBS).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

It is worth noting that under $\mathbf{H}_{0,\mathcal{G}}$ with $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$, SGL and L&B can potentially select the true homogeneous model by identifying the change points at $\{1, n\}$. Hence, we record their rates of false selections as their "empirical sizes". As for Lee2016, their main purpose is to simultaneously estimate the potential single change point as well as the regression coefficients. Therefore, we do not report their empirical sizes and powers here.

Table 1 summarizes the empirical sizes for Models 1 and 2 with different combinations of (n, p, s) under $N(0, 1)$ distributions. We can see that both SGL and L&B are only applicable for the case of the overall subset with $\mathcal{G} = \{1, \ldots, p\}$. In those cases, SGL suffers from serious size distortions with too many false selections. One reasonable explanation is that SGL builds their algorithms on the sparse group lasso which tends to overestimate the number of change points. Moreover, we observe that L&B seems to be conservative although it can select the homogenous model with no false selections. As for our proposed methods, the empirical sizes of Boot-I are out of control (especially for the active set $\mathcal{S}$). This suggests that for change point detection of high dimensional linear models, the residual term of the de-biased lasso-based process can not be ignored, even though it is asymptotically negligible in theory. As compared to Boot-I, Boot-II benefits from bootstrapping the whole de-biased lasso-based process. In

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Table 1: Empirical sizes for Models 1 and 2. The errors are generated from $N(0,1)$. The results are based on 2000 replications.

Empirical sizes (%) with $(n,s)=(200,5)$

Model	$\mathcal{G}$	p	Boot-I ($\alpha=1\%$)	Boot-II ($\alpha=1\%$)	Boot-I ($\alpha=5\%$)	Boot-II ($\alpha=5\%$)	SGL	L&B
$\mathbf{\Sigma}=\mathbf{I}$	$\mathcal{S}$	200	7.61	1.70	18.52	3.86	NA	NA
		400	10.70	1.80	23.05	5.30	NA	NA
	$\mathcal{S}^c$	200	8.23	1.44	15.43	4.06	NA	NA
		400	11.93	0.93	21.60	3.40	NA	NA
	$\mathcal{S}\cup\mathcal{S}^c$	200	7.41	1.03	14.20	2.93	38.89	0.00
		400	12.55	1.39	27.37	3.86	46.67	0.00
$\mathbf{\Sigma}=\mathbf{\Sigma}^*$	$\mathcal{S}$	200	7.61	1.49	14.40	4.73	NA	NA
		400	8.64	1.65	16.26	4.68	NA	NA
	$\mathcal{S}^c$	200	3.50	0.82	12.14	3.09	NA	NA
		400	5.76	0.67	12.76	3.03	NA	NA
	$\mathcal{S}\cup\mathcal{S}^c$	200	4.73	0.82	13.37	3.29	77.78	0.00
		400	7.82	1.23	17.08	3.19	80.00	0.00

Empirical sizes (%) with $(n,s)=(300,10)$

Model	$\mathcal{G}$	p	Boot-I ($\alpha=1\%$)	Boot-II ($\alpha=1\%$)	Boot-I ($\alpha=5\%$)	Boot-II ($\alpha=5\%$)	SGL	L&B
$\mathbf{\Sigma}=\mathbf{I}$	$\mathcal{S}$	200	12.76	1.83	23.66	3.25	NA	NA
		400	19.55	1.88	33.74	7.35	NA	NA
	$\mathcal{S}^c$	200	8.33	1.02	16.67	3.25	NA	NA
		400	13.79	1.63	26.95	3.06	NA	NA
	$\mathcal{S}\cup\mathcal{S}^c$	200	11.52	0.82	22.43	3.27	56.67	0.00
		400	17.49	2.45	32.30	5.71	62.30	0.00
$\mathbf{\Sigma}=\mathbf{\Sigma}^*$	$\mathcal{S}$	200	10.91	0.62	22.63	2.67	NA	NA
		400	17.07	2.26	28.86	5.56	NA	NA
	$\mathcal{S}^c$	200	4.32	0.41	11.32	1.65	NA	NA
		400	3.66	0.81	10.77	2.44	NA	NA
	$\mathcal{S}\cup\mathcal{S}^c$	200	6.50	1.85	16.06	4.32	56.67	0.00
		400	7.06	0.61	17.57	3.25	55.30	0.00

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

most cases, the empirical sizes for Boot-II are close to the nominal level across various dimensions and subgroups. Interestingly, it shows that the empirical performance of Boot-II is affected by the candidate subgroups. More specifically, empirical sizes for the active set $\mathcal{S}$ are sometimes larger than the nominal level and the size performance of the non-active set $\mathcal{S}^c$ performs the best among all candidate subgroups. Note that similar findings are also observed in constructing simultaneous confidence intervals in Zhang and Cheng (2017) for the given subgroup $\mathcal{G}$. In addition, we can see that Boot-II can still have satisfactory size performance as the non-zero elements increase slowly from $s = 5$ to $s = 10$.

In the supplemental materials, we report the size performance under standardized Gamma$(4, 1)$ and Student's t_5 distributions in Tables S5.1 and S5.2. In both cases, our proposed method can control the size under the nominal level. This suggests that the bootstrap null distribution is correctly calibrated even for non-normal underlying errors.

4.2 Empirical powers

We next analyze the empirical powers. Denote the signal jump

$$\boldsymbol{\delta} = C\sqrt{\log(p)/n} \times \left(2^3, 2^2, 2^1, 2^0, 2^{-1}\right)^\top.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Table 2: Empirical powers (%) under Model 1. The numerical results are based on 2000 replications.

Empirical powers (%) with $\boldsymbol{\delta} = 0.5\sqrt{\log(p)/n} \times (2^3, 2^2, 2^1, 2^0, 2^{-1})$.

			Change point at $k^* = 0.5n$		Change point at $k^* = 0.3n$	
Model	$\mathcal{G}$	p	Boot-II	L&B	Boot-II	L&B
$\boldsymbol{\Sigma} = \mathbf{I}$	$\mathcal{S}$	200	58.33	NA	36.46	NA
		400	64.93	NA	42.71	NA
	$\mathcal{S}^c$	200	2.08	NA	4.17	NA
		400	3.47	NA	3.82	NA
	$\mathcal{S} \cup \mathcal{S}^c$	200	43.75	0.00	29.17	0.00
		400	40.97	0.00	27.17	0.00

Empirical powers (%) with $\boldsymbol{\delta} = \sqrt{\log(p)/n} \times (2^3, 2^2, 2^1, 2^0, 2^{-1})$.

			Change point at $k^* = 0.5n$		Change point at $k^* = 0.3n$	
Model	$\mathcal{G}$	p	Boot-II	L&B	Boot-II	L&B
$\boldsymbol{\Sigma} = \mathbf{I}$	$\mathcal{S}$	200	100.00	NA	99.38	NA
		400	99.59	NA	99.38	NA
	$\mathcal{S}^c$	1 200	3.50	NA	3.91	NA
		400	3.09	NA	2.06	NA
	$\mathcal{S} \cup \mathcal{S}^c$	200	100.00	36.87	99.18	29.29
		400	99.38	38.38	99.38	28.28

Empirical powers (%) with $\boldsymbol{\delta} = 2\sqrt{\log(p)/n} \times (2^3, 2^2, 2^1, 2^0, 2^{-1})$.

			Change point at $k^* = 0.5n$		Change point at $k^* = 0.3n$	
Model	$\mathcal{G}$	p	Boot-II	L&B	Boot-II	L&B
$\boldsymbol{\Sigma} = \mathbf{I}$	$\mathcal{S}$	200	100.00	NA	100.00	NA
		400	100.00	NA	100.00	NA
	$\mathcal{S}^c$	200	2.47	NA	1.65	NA
		400	3.50	NA	2.88	NA
	$\mathcal{S} \cup \mathcal{S}^c$	200	100.00	99.49	100.00	98.48
		400	100.00	100.00	100.00	97.98

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

We set $n = 200$. We first generate $\boldsymbol{\beta}^{(1)}$ with $s = 5$ non-zero elements following $\mathrm{U}(0, 2)$ distributions out of $\mathcal{S} = \{1, \ldots, 50\}$. Then, we add $\boldsymbol{\delta}$ with $C \in \{0.5, 1, 2\}$ on the corresponding 5 non-zero covariates of $\boldsymbol{\beta}^{(1)}$ to generate $\boldsymbol{\beta}^{(2)}$. Note that in this setting, $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ have a common support.

Table 2 shows the power results with $n = 200$, where various data dimensions, change point locations, candidate subgroups, and signal strength are considered. Note that we do not report the results of SGL and Boot-I because of their serious size distortions. According to Table 2, we see that our proposed method can detect a change point with a very high probability across various data dimensions when the candidate subgroup has a change point ($\mathcal{G} = \mathcal{S}$ and $\mathcal{G} = \mathcal{S} \cup \mathcal{S}^c$). Interestingly, it is shown that the powers in $\mathcal{S}^c$ are close to the nominal level since the coefficients in $\mathcal{S}^c$ are zeros before and after the change point. As for L&B, we see that it can successfully detect a change point when the signal jump is relatively strong ($C = 2$). However, L&B is not very sensitive to weak signals with $C = 0.5$ and $C = 1$. The above analysis suggests that our proposed method is very powerful to sparse alternatives and is more efficient and flexible than the existing methods for change point detection of high dimensional linear models. Moreover, Table S5.3 in the supplemental materials shows the power performance similar to Table 2 for Model 2 with banded covariance

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

structures.

In addition to single change point detection, we also combine the proposed new testing method with the binary segmentation technique for detecting multiple change points. The detailed algorithms and numerical performance are provided in the supplementary materials. The results further demonstrate the superior performance of our method over its competitors.

5. Conclusions

In this paper, we propose a new method for change point inference in the context of high dimensional linear models. For any given subgroup $\mathcal{G} \subset \{1, \ldots, p\}$, a L_∞-norm-based test statistic $T_\mathcal{G}$ is constructed for testing the homogeneity of regression coefficients across the observations. To approximate its limiting null distribution, a novel multiplier bootstrap procedure is introduced. Our new method is powerful against sparse alternatives with only a few entries in $\mathcal{G}$ having a change point, and allows the group size $|\mathcal{G}|$ to grow exponentially with the sample size n. As for the change point identification, a new change point estimator is obtained by taking "argmax" of the L_∞-aggregated process $H_\mathcal{G}(t)$. Theoretically, the change point estimator is shown to be consistent, allowing the overall sparsity s of regression coefficients and the group size $|\mathcal{G}|$ to grow simultaneously

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

with the sample size n. In addition to single change point detection, we further combine our proposed method with the binary segmentation-based technique for detecting and identifying multiple change points. Our new testing method is relatively easy to implement and is justified via extensive numerical studies.

Acknowledgement

The authors would like to thank the editor, the associate editor, and the reviewers for their helpful comments and suggestions.

Supplementary Material

The online supplementary materials provide detailed basic assumptions and proofs of the main theory, and additional numerical results including size, power, multiple change point detection. In addition, an interesting real data application is also provided.

References

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

Bai, Y. and A. Safikhani (2023). A unified framework for change point detection in high-dimensional linear models. *Statistica Sinica 33*, 1–28.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Chan, N. H., C. Y. Yau, and R. Zhang (2014). Group lasso for structural break time series. *Journal of the American Statistical Association 109*(506), 590–599.

Chen, H., H. Ren, F. Yao, and C. Zou (2023). Data-driven selection of the number of change-points via error rate control. *Journal of the American Statistical Association 118*(542), 1415–1428.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics 41*(6), 2786–2819.

Cho, H. and D. Owens (2022). High-dimensional data segmentation in regression settings permitting heavy tails and temporal dependence. *arXiv preprint arXiv:2209.08892*.

He, Z., D. Cheng, and Y. Zhao (2023). Multiple testing of local extrema for detection of structural breaks in piecewise linear models. *arXiv preprint arXiv:2308.04368*.

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

Horváth, L. and Q.-M. Shao (1995). Limit theorems for the union-intersection test. *Journal of Statistical Planning and Inference 44*(2), 133–148.

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

Kaul, A., S. B. Fotopoulos, V. K. Jandhyala, and A. Safikhani (2021). Inference on the change point under a high dimensional sparse mean shift. *Electronic Journal of Statistics 15*, 71–134.

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. *The Journal of Machine Learning Research 20*, 1–40.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

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* .

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association 55*(290), 324–330.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

Van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Vostrikova, L. Y. (1981). Detecting disorder in multidimensional random process. *Soviet Math. Dokl 24*, 55–59.

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

Wang, F., O. Madrid, Y. Yu, and A. Rinaldo (2022). Denoising and change point localisation in piecewise-constant high-dimensional regression coefficients. In *International Conference on Artificial Intelligence and Statistics*, Volume 151, pp. 4309–4338. PMLR.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0212

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*.

Yu, M. and X. Chen (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83*(2), 247–270.

Zhang, B., J. Geng, and L. Lai (2015). Change-point estimation in high dimensional linear regression models via sparse group lasso. In *In: 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 815–821. IEEE.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhang, X. and G. Cheng (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association 112*(518), 757–768.

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

E-mail: liubin0145@gmail.com

E-mail: xszhang@fudan.edu.cn

Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Linberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, U.S.A

E-mail: yfliu@email.unc.edu