Efficient multiple change point detection for high-dimensional generalized linear models

Xianru WANG[1], Bin LIU[1*], Xinsheng ZHANG[1], and Yufeng LIU[2], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]*Department of Statistics and Data Science, School of Management at Fudan University, Shanghai, China*
[2]*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, Chapel Hill, North Carolina, U.S.A.*

Abstract: Change point detection for high-dimensional data is an important yet challenging problem for many applications. In this article, we consider multiple change point detection in the context of high-dimensional generalized linear models, allowing the covariate dimension p to grow exponentially with the sample size n. The model considered is general and flexible in the sense that it covers various specific models as special cases. It can automatically account for the underlying data generation mechanism without specifying any prior knowledge about the number of change points. Based on dynamic programming and binary segmentation techniques, two algorithms are proposed to detect multiple change points, allowing the number of change points to grow with n. To further improve the computational efficiency, a more efficient algorithm designed for the case of a single change point is proposed. We present theoretical properties of our proposed algorithms, including estimation consistency for the number and locations of change points as well as consistency and asymptotic distributions for the underlying regression coefficients. Finally, extensive simulation studies and application to the Alzheimer's Disease Neuroimaging Initiative data further demonstrate the competitive performance of our proposed methods.

Résumé: La détection de points de rupture dans des données en hautes dimensions est un problème important mais comporte des défis majeurs pour de nombreuses applications. Dans cet article, nous considérons la détection de points de changement multiples dans le contexte de modèles linéaires généralisés (GLM) de grande dimension et dans lesquels la dimension des covariables p croît de façon exponentielle avec la taille de l'échantillon n. Le modèle étudié est assez général et flexible pour permettre de couvrir différents modèles particuliers. Il peut tenir compte du mécanisme de génération de données sous-jacent de façon automatique et sans connaissance préalable du nombre de points de changement. En utilisant des techniques de programmation dynamique et de segmentation binaire, nous proposons deux algorithmes de détection de points de rupture multiples dont le nombre croît avec n. Pour une efficacité computationnelle accrue, un algorithme plus efficace conçu pour le cas d'un seul point de changement est proposé. Nous établissons les propriétés théoriques des algorithmes proposés, y compris la convergence de l'estimation du nombre

* *Corresponding author: liubin0145@gmail.com*

et de la localisation des points de changement, ainsi que la convergence des coefficients du modèle de régression sous-jacent. Enfin, nous établissons la performance des méthodes proposées sur des échantillons finis par une vaste étude de simulation et les utilisons pour analyser un jeu de données réelles provenant de l'initiative d'imagerie médicale pour la maladie d'Alzheimer (ADNI).

1. INTRODUCTION

With the advance of technology, complex large-scale data are prevalent in various scientific fields. Data heterogeneity creates great challenges for the analysis of complex data that may not be well approximated by a common distribution. Change point detection is a powerful tool to deal with data heterogeneity. Since the seminal work by Page (1955), there is a growing literature on change point detection with a wide range of applications, including genomics (Braun, Braun & Müller, 2000), finance (Pesaran & Pick, 2007; Fan, Lv & Qi, 2011), and social networks (Raginsky et al., 2012).

In this article, we consider multiple change point detection for a general framework of high-dimensional generalized linear models (GLMs). Suppose we have n independent observations $\{Y_i, X_i\}_{i=1}^n$ with

$$g(\mu_i) = X_i^T \beta^{(i)} \quad \text{for} \ i = 1, \dots, n, \tag{1}$$

where $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the real-valued response for the ith observation, $X_i = (X_{i1}, \dots, X_{ip})^\top$ is the corresponding covariate vector in $\mathcal{X} \subset \mathbb{R}^p$, $\mu_i = \mathbb{E}(Y_i | X_i)$, $g(\cdot)$ is the link function, and $\beta^{(i)} = \left(\beta_1^{(i)}, \dots, \beta_p^{(i)}\right)^\top \in \mathbb{R}^p$ is the unknown regression coefficient vector for the ith observation. Then we consider estimating multiple change points with piecewise constant coefficients for model (1). More specifically, let $\tilde{k} \geq 0$ be the true number of unknown change points along with the true location vector $\tilde{\tau} = (\tilde{\tau}_0, \tilde{\tau}_1 \dots, \tilde{\tau}_{\tilde{k}}, \tilde{\tau}_{\tilde{k}+1})^\top$ with $0 = \tilde{\tau}_0 < \tilde{\tau}_1 < \tilde{\tau}_2 < \cdots < \tilde{\tau}_{\tilde{k}} < \tilde{\tau}_{\tilde{k}+1} = 1$. Then, the unknown $\tilde{k}$ change points divide the n time-ordered observations into $\tilde{k} + 1$ intervals and the underlying regression coefficients $\beta^{(i)}$ have the following form:

$$\beta^{(i)} = \begin{cases} \beta^0, & \text{if} \ \tilde{k} = 0, \\ \beta^0(j), & \text{if} \ \tilde{\tau}_{j-1} < i/n \leqslant \tilde{\tau}_j, j = 1, \dots, \tilde{k}+1, \end{cases} \tag{2}$$

where $\beta^0(j) = \left(\beta_1^0(j), \dots, \beta_p^0(j)\right)^\top \in \mathbb{R}^p$ denotes the underlying true regression coefficients in the jth interval. We focus on change point detection, which consists of estimating: (a) the number of change points ($\tilde{k}$); (b) the locations of change points ($\tilde{\tau}$); (c) the regression coefficients $\beta^0(j)$ in each segmentation, where $j = 1, \dots, \tilde{k}+1$.

There is a growing literature on change point detection. Most existing papers focus on change point problems in the mean, variance, or covariance matrix either for a fixed p (Kirch, Muhsal & Ombao, 2015; Zhang & Lavitas, 2018) or for a growing p (Frick, Munk & Sieling, 2014; Jirak, 2015; Barigozzi, Cho & Fryzlewicz, 2018; Wang & Samworth, 2018; Wang, Yu & Rinaldo, 2021). Progress has been made in the literature for detection of multiple change points as well (Lavielle & Teyssière, 2006; Aue et al., 2009; Harchaoui & Lévy-Leduc, 2010; Cho & Fryzlewicz, 2015). Despite progress on change point detection, many fewer papers appear in the literature on regression change point problems, especially for high-dimensional models. The main difficulty comes from the complexity of both calculation and theoretical analysis arising from the growing dimension.

For regression problems, penalized techniques such as Lasso (Tibshirani, 1996) are popular in dealing with high-dimensional data. Some theoretical properties of the Lasso and various extensions can be found in Fan & Peng (2004), Candès & Tao (2007), and van de Geer et al. (2014).

For a general overview and recent developments, we refer to Fan & Lv (2010) and Tibshirani (2011). In terms of change point detection based on Lasso, some methods exist for solving regression change point problems both in low and high dimensions. For example, designed for a fixed p, Ciuperca (2014) considered multiple change point estimation based on the Lasso. Qian & Su (2016) and Li, Qian & Su (2016) proposed a systematic change point estimation framework based on the adaptive fused Lasso. When the data dimension p grows to infinity, Lee, Seo & Shin (2016) considered high-dimensional linear models with a possible change point and proposed a method for estimating regression coefficients as well as the unknown threshold parameter. As an extension, Leonardi & Bühlmann (2016) proposed computationally efficient algorithms for the number and locations of multiple change points in the context of high-dimensional linear models. Recently, Liu, Zhang & Liu (2021) investigated simultaneous change point detection and identification based on a de-biased Lasso process. Wang, Lin & Willett (2021) developed variance projected wild binary segmentation (VPWBS) for multiple change point detection.

Note that the above-mentioned papers focused on change point detection based on linear models with a continuous response, and thus are not directly applicable to the analysis of categorical or count response variables in practice. GLM can be very useful in this situation because it covers the exponential family distributions for the response variable. Because of its generality, GLM is widely used in various applications such as genetics, economics, and epidemiology. Several papers studied low-dimensional, single change point problems in the context of GLM (Lee & Seo, 2008; Lee, Seo & Shin, 2011; Fong, Di & Permar, 2015). To the best of our knowledge, change point detection for high-dimensional GLMs has not been studied in the literature. Hence, it is desirable to consider a flexible and general framework for analyzing high-dimensional data with heterogeneity. Motivated by this, in this article, we consider computationally efficient multiple change point detection in the context of high-dimensional GLMs. Our main contributions are summarized as follows:

- We consider change point problems in a more flexible and general framework of high-dimensional GLMs, allowing the data dimension p to grow exponentially with the sample size n. It covers various model settings including linear models, logistic, and probit models as special cases. As far as we know, change point detection for high-dimensional logistic and probit models has not been considered in the literature.

- Under the above framework, we propose a three-step procedure to estimate the number and locations of change points based on the Lasso estimator of the regression coefficients. The basic idea is to choose a useful contrast function $J(\tau(k))$, which satisfies $J(\hat{\tau}(\hat{k})) < J(\tau(k))$ for any $\tau(k)$. To solve this optimization problem, we propose two algorithms based on dynamic programming and binary segmentation techniques, which have computational costs of $O(n^2\text{GLMLasso}(n))$ and $O(n\log(n)\text{GLMLasso}(n))$, respectively, where $\text{GLMLasso}(n)$ is the cost to compute the Lasso estimator for the GLM with the sample size n. We also propose a much more efficient approach for the single change point case, with a computational cost of $O(\log(n)\text{GLMLasso}(n))$. To the best of our knowledge, this is the most computationally efficient algorithm available for detecting a single change point in GLMs.

- We examine some theoretical properties of our proposed change point estimators computed by the three algorithms. To be specific, under some mild conditions, both the dynamic programming and binary segmentation techniques can obtain a consistent estimator for the number and locations of the true change points with a rate of $O_p(\sqrt{\log(p)/n})$, which covers the case with an asymptotically growing number of change points. Moreover, the estimation error of the Lasso estimator of underlying regression coefficients can be bounded to $o_p(1)$. To achieve further statistical prediction and inference, we introduce the de-biased Lasso estimator of the underlying regression coefficients in each segmentation, which is shown to be asymptotically normal. As for the third efficient approach designed for single change point

cases, we establish that it can identify the location of the change point with high estimation accuracy. Finally, the competitive performance of our proposed methods is demonstrated by extensive numerical results as well as application to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

For a better understanding of our work, we would like to point out its relationship with several related papers. Compared with Lee, Seo & Shin (2011), which considered single change point detection for the binary response variable with low-dimensional covariates, we overcome the challenges of the computational and theoretical complexity arising from the growing dimension and number of change points. Meanwhile, to address the issue of the unknown multiple change points, we construct accurate and effective algorithms based on two techniques, dynamic programming and binary segmentation. These techniques are popular for multiple change point detection and were previously studied by Lavielle & Teyssière (2006), Boysen et al. (2009), Harchaoui & Lévy-Leduc (2010), Cho & Fryzlewicz (2012, 2015), and Leonardi & Bühlmann (2016). Our extension to GLMs involves several technical challenges to overcome. One substantial difficulty comes from the complex form of the contrast functions compared with the least squares for linear models considered in Leonardi & Bühlmann (2016).

The rest of this article is organized as follows. In Section 2, we introduce our methodology and demonstrate how our proposed three algorithms detect change points. In Section 3, the corresponding theoretical results of the change points computed by different algorithms are established. We investigate the performance of our proposed methods by extensive numerical results as well as a real data application in Sections 4 and 5. We summarize the article in Section 6. Detailed proofs of the main theorems and some useful lemmas are given in the Appendix.

2. METHODOLOGY

In this section, we introduce our new methodology for model (1) with multiple unknown change points. In particular, in Section 2.1, some notation is introduced. In Section 2.2, we present a three-step change point estimator including the number and the locations of change points. Meanwhile, the regression coefficients in each segment are estimated based on the Lasso. In Sections 2.2.1 and 2.2.2, based on dynamic programming and binary segmentation techniques, two algorithms are proposed to detect multiple change points. To further improve the computational efficiency, in Section 2.2.3, we present a much more efficient algorithm designed for the case of a single change point.

2.1. Notation

We first introduce some notation. For a vector $\mathbf{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, we denote $\|\mathbf{a}\|_1 = \sum_{i=1}^p |a_i|$, $\|\mathbf{a}\|_2 = (\sum_{i=1}^p a_i^2)^{1/2}$, and $\|\mathbf{a}\|_\infty = \max_{1 \le i \le p} |a_i|$. For two real-valued sequences a_n and b_n, we set $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$, for a sufficiently large n. We set $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. For a sequence of random variables $\{\xi_1, \xi_2, \ldots\}$, we set $\xi_n \xrightarrow{\mathbb{P}} \xi$ if ξ_n converges to ξ in probability as $n \to \infty$. We also denote $\xi_n = o_p(1)$ if $\xi_n \xrightarrow{\mathbb{P}} 0$. Given an interval $(u, v) \subset [0, 1]$ such that $u, v \in L^1 = \{i/n, i = 1, \ldots, n, n \in \mathbb{N}\}$, we denote the vector $(Y_{un+1}, \ldots, Y_{vn})^\top$ by $\mathbf{Y}_{(u,v)}$ and the vector $(\epsilon_{un+1}, \ldots, \epsilon_{vn})^\top$ by $\boldsymbol{\epsilon}_{(u,v)}$. Analogously, we use $\mathbf{X}_{(u,v)}$ to denote the $(v - u)n \times p$ dimensional matrix $(X_{(u,v)}^{(1)}, \ldots, X_{(u,v)}^{(p)})$, where $X_{(u,v)}^{(j)} = (X_{un+1}^{(j)}, \ldots, X_{vn}^{(j)})^\top$ with $j = 1, \ldots, p$, and we use $\hat{\beta}_{(u,v)}$ to denote the Lasso estimator based on the observations $\mathbf{Y}_{(u,v)}$ and $\mathbf{X}_{(u,v)}$. For a set A, we use $\#A$ to denote its cardinality. For any $x \ge 0$, we use $[x]$ to denote the largest integer less than or equal to x. We use $C_1, C_2, \ldots$ to denote generic positive constants that may vary in different places.

2.2. New Estimation and Algorithms

Let $Y = (Y_1, \ldots, Y_n)^\top$ denote the $n \times 1$ response vector, and $\mathbf{X}$ the $n \times p$ design matrix with $X_i = (X_{i1}, \ldots, X_{ip})^\top$ being its ith row for $1 \leq i \leq n$. In this article, we assume $\{X_i\}_{i=1}^n$ are independently and identically distributed (i.i.d.) p-dimensional random vectors with mean zero and covariance matrix $\Sigma = \mathrm{Cov}(X_1)$. Furthermore, for $j = 1, \ldots, \tilde{k} + 1$, we denote by $S^{(j)}$ the set of nonzero elements of the regression coefficients $\beta^0(j)$, i.e., $S^{(j)} = \#\{\ell : \beta_\ell^0(j) \neq 0$ for $\ell = 1, \ldots, p\}$. For any given partition $\tau = (\tau_0, \tau_1, \ldots, \tau_k, \tau_{k+1})^\top$, we denote the jth interval by $I_j(\tau) = (\tau_{j-1}, \tau_j)$, the length of the jth interval by $r_j(\tau) = \tau_j - \tau_{j-1}$, the shortest interval length by $r(\tau) = \min_{1 \leq j \leq k+1} r_j(\tau)$, and the change point number by $l(\tau)$. Moreover, we denote the minimum interval length by δ.

We are now ready to introduce our change point estimator in detail. We consider the Lasso-type ℓ_1-penalized estimator for high-dimensional GLMs. Such estimators have some desirable properties. In particular, van de Geer (2008) derived some theoretical properties including consistency and the oracle inequality, based on which our algorithms are mainly constructed. More specifically, let $\rho_\beta(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be some loss function relative to $g(\cdot)$. For instance, if $g(\cdot)$ is the logit function, $\rho_\beta(x, y)$ will then be the negative-likelihood function in the form $\log\left(1 + e^{x^\top \beta}\right) - yx^\top \beta$. For $\beta \in \mathbb{R}^p$, we define $\dot{\rho}_\beta := \frac{\partial \rho_\beta}{\partial \beta}$ and $\ddot{\rho}_\beta := \frac{\partial \rho_\beta}{\partial \beta \partial \beta^T}$. Note that such complex loss functions lead to substantial difficulty for the estimation of change points as well as regression coefficients. Given data observations $\{Y_i, X_i\}_{i=1}^n$, the Lasso-based GLM method solves the following ℓ_1 penalized problem:

$$\hat{\beta} = \arg \min_\beta \left\{ \frac{1}{n} \sum_{i=1}^n \rho_\beta(X_i, Y_i) + \lambda \|\beta\|_1 \right\}. \tag{3}$$

Because we consider heterogeneous data with possible multiple change points, we cannot use Equation (3) directly to obtain parameter estimation. The main challenge is that both the number ($\tilde{k}$) and the locations ($\tilde{\tau}$) are unknown. To solve this issue, we consider three steps.

Before introducing the change point estimator, we first demonstrate how to estimate the regression coefficients for each segment. To be specific, for any given candidate partition $\tau = (\tau_0, \tau_1, \ldots, \tau_k, \tau_{k+1})^\top$, with $\tau_j \in L^1, j = 1, \ldots, k + 1$, we obtain the estimator with the ℓ_1 penalty in each segment by, for $j = 1, \ldots, k + 1$,

$$\hat{\beta}(\tau, j) = \arg \min_\beta \left\{ \frac{1}{n} \sum_{i=n\tau_{j-1}+1}^{n\tau_j} \rho_\beta(X_i, Y_i) + \lambda_j \sqrt{(\tau_j - \tau_{j-1})} \|\beta\|_1 \right\}, \tag{4}$$

where λ_j is the non-negative regularization parameter.

Based on Equation (4), our new algorithms for estimating both $\tilde{k}$ and $\tilde{\tau}$ are summarized into the following three steps.

Step 1 (Search the "best" partition)**.** Given the candidate number of change points k, we find the "best" partition $\hat{\tau}(k) = (\hat{\tau}_1, \ldots, \hat{\tau}_k)^\top$ that minimizes the total loss function (contrast function):

$$\hat{\tau}(k) = \arg \min_{\tau = (\tau_0, \ldots, \tau_{k+1})^\top} J(\tau, \hat{\beta}(\tau), \mathbf{X}, Y) + \gamma(k + 1), \tag{5}$$

where $\hat{\beta}(\tau) := \left(\hat{\beta}(\tau, 1)^\top, \ldots, \hat{\beta}(\tau, k + 1)^\top\right)^\top$, $J(\tau, \hat{\beta}(\tau), \mathbf{X}, Y) := \sum_{j=1}^{k+1} P_n \rho\left(I_j(\tau), \hat{\beta}(\tau, j)\right)$, and $P_n \rho\left(I_j(\tau), \beta\right) := \frac{1}{n} \sum_{i=n\tau_{j-1}+1}^{n\tau_j} \rho_\beta(X_i, Y_i)$.

Step 2 (Estimate number of change points)**.** We put $\hat{\boldsymbol{\tau}}(k)$ into $J(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}), \mathbf{X}, \boldsymbol{Y})$ and obtain the minimum loss function associated with k as

$$G(k) := J\left(\hat{\boldsymbol{\tau}}(k), \hat{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\tau}}(k)\right), \mathbf{X}, \boldsymbol{Y}\right) + \gamma(k + 1). \tag{6}$$

Then, we find the "best" estimation that minimizes $G(k)$ with a penalty:

$$\hat{k} = \underset{k}{\arg\min}\ G(k). \tag{7}$$

Step 3 (Estimate locations of change points)**.** We put $\hat{k}$ into Step 1 and obtain the final change point estimator $\hat{\boldsymbol{\tau}} := \hat{\boldsymbol{\tau}}(\hat{k}) = (\hat{\tau}_1, \dots, \hat{\tau}_{\hat{k}})^\top$ by

$$\hat{\boldsymbol{\tau}} = \underset{\boldsymbol{\tau}=(\tau_0, \dots, \tau_{\hat{k}+1})^\top}{\arg\min}\ J\left(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}), \mathbf{X}, \boldsymbol{Y}\right). \tag{8}$$

Combining Steps 1–3, our final change point estimators $\hat{k}$ and $\hat{\boldsymbol{\tau}}$ can be obtained equivalently in the following form:

$$\hat{\boldsymbol{\tau}} = \underset{k}{\arg\min}\ \underset{\boldsymbol{\tau}:l(\boldsymbol{\tau})=k}{\min}\ \left\{ J(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}), \mathbf{X}, \boldsymbol{Y}) + \gamma(k + 1) \right\}. \tag{9}$$

After the above three steps, we obtain the change point estimators $\hat{\boldsymbol{\tau}}$. As for $\boldsymbol{\beta}^0$, we recommend two different Lasso estimators of the underlying regression coefficients $\boldsymbol{\beta}^0$, serving different purposes for practitioners. In particular, naturally, we can use the Lasso estimator of $\boldsymbol{\beta}^0$ to select variables and make a prediction, which is defined for $j = 1, \dots, \hat{k} + 1$ as

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=n\hat{\tau}_{j-1}+1}^{n\hat{\tau}_j} \rho_{\boldsymbol{\beta}}\left(X_i, Y_i\right) + \lambda_j \sqrt{(\hat{\tau}_j - \hat{\tau}_{j-1})} \|\boldsymbol{\beta}\|_1 \right\}.$$

For further statistical inference including confidence intervals and hypothesis testing, van de Geer et al. (2014) proposed the de-biased Lasso estimator and analyzed its asymptotic properties for the homogeneous model under high-dimensional set-ups. Similarly, for the heterogeneous observations, we construct a de-biased Lasso estimator for the underlying regression coefficients for each segmentation for $j = 1, \dots, \hat{k} + 1$ as

$$\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) - \hat{\boldsymbol{\Theta}} P_n \dot{\rho}_{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)},$$

where the precision matrix estimator $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Theta}}_{\text{Lasso}}$ can be constructed using nodewise Lasso with $\hat{\boldsymbol{\Sigma}} := P_n \ddot{\rho}_{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)}$ as input (van de Geer et al., 2014).

In what follows, we introduce three specific algorithms for solving Equation (9). Note that $J(\boldsymbol{\tau}, \boldsymbol{\beta}, \mathbf{X}, \boldsymbol{Y})$ and $P_n \rho(I_j(\boldsymbol{\tau}), \boldsymbol{\beta})$ are the loss function for all intervals and the jth interval, respectively. Meanwhile, λ_j in Equation (4) and γ in Equation (5) are positive tuning parameters that encourage coefficient and segment sparsity, respectively. We adopt a cross-validation approach to make a proper choice of these two tuning parameters λ and γ. To compute $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j)$ in Equation (4), we can use, for example, the R package glmnet (https://glmnet.stanford.edu). It is worth mentioning the following two remarks for our proposed estimator: (1) If the number of

change points $\tilde{k}$ is known, we just use Step 1 by plugging in $k = \tilde{k}$ to directly obtain the locations of change points as follows:

$$\hat{\boldsymbol{\tau}}(\tilde{k}) = \underset{\boldsymbol{\tau}=(\tau_0,\ldots,\tau_{\tilde{k}+1})^\top}{\arg\min} J\left(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}), \mathbf{X}, \boldsymbol{Y}\right). \tag{10}$$

In this case, our method covers the setting considered by Lee, Seo & Shin (2016), where at most one change point is assumed. (2) When no change point occurs ($\tilde{k} = 0$), our proposed method can still work. Hence, our proposed method can automatically account for the underlying data generation mechanism ($\tilde{k} = 0$ or $\tilde{k} > 0$) without specifying any prior knowledge about the number of change points $\tilde{k}$. Furthermore, as shown by our extensive numerical studies, our new algorithms can estimate $\tilde{k}$ with high accuracy.

Our main goal is to design efficient algorithms that solve the optimization problem in Equation (9) of the form $J(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}) + \text{Pen}(\boldsymbol{\tau})$. To address this issue, three algorithms are proposed next.

2.2.1. Dynamic programming approach

We introduce a general approach based on the dynamic programming algorithm (DPA), which works well for our change point problem (Eq. 9). It is well known that DPA has excellent accuracy because it considers the global solution of Equation (9). It is widely used in multiple change point detection including the efficient, parallelized approaches introduced recently by Tickle et al. (2020). More details can be found in Boysen et al. (2009) and Leonardi & Bühlmann (2016). Next, we present how to use this technique to solve Equation (9) in detail.

For any given $v \in \{i/n : i = 1, \ldots, n\}$, consider the sample $\left(\boldsymbol{Y}_{(0,v]}, \mathbf{X}_{(0,v]}\right)$. Given a candidate change point number k, denote $F_k(v)$ as the minimum value as follows:

$$F_k(v) := \min_{\boldsymbol{\tau}:\, l(\boldsymbol{\tau})=k} \sum_{j=1}^{k+1} \left(P_n \rho\left(I_j(v\boldsymbol{\tau}), \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j)\right) + \gamma \right). \tag{11}$$

One can see that the optimal $k + 1$ segments $\{(\tau_{j-1}, \tau_j)\}_{j=1}^{k+1}$ corresponding to the change point vector $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{k+1})^\top$ obtained from Equation (11), consist of the optimal first k segments $\{(\tau_{j-1}, \tau_j)\}_{j=1}^{k}$ and a single segment $(\tau_k, \tau_{k+1})^\top$. Recall that $\tau_0 := 0$ and $\tau_{k+1} := 1$. Then τ_k is the rightmost change point estimator. Furthermore, by definition of $F_k(v)$, $\{(\tau_{j-1}, \tau_j)\}_{j=1}^{k}$ obtained from Equation (11) is also a minimizer of $F_{k-1}(\tau_k)$. Hence, the last change point τ_k is the minimizer of $F_{k-1}(u) + P_n \rho\left((u, v), \hat{\boldsymbol{\beta}}_{(u,v)}\right) + \gamma$ with $u < v$.

The above observation motivates us to use the dynamic programming recursion to calculate $F_k(v)$ with $v \in \{i/n : i = 1, \ldots, n\}$. In particular, for any $v \in \{i/n : i = 1, \ldots, n\}$, define

$$F_0(v) = P_n \rho\left((0, v), \hat{\boldsymbol{\beta}}_{(0,v)}\right) + \gamma. \tag{12}$$

Then, the dynamic programming recursion proceeds as follows:

$$F_k(v) = \min_{\substack{u \in \{i/n:\, i=1,\ldots,n\} \\ u < v}} \left\{ F_{k-1}(u) + P_n \rho\left((u, v), \hat{\boldsymbol{\beta}}_{(u,v)}\right) + \gamma \right\}, \quad v \in \{i/n : i = 1, \ldots, n\}. \tag{13}$$

Define $V_n = \{i/n : i = 1, \ldots, n\}$. Based on Equations (12) and (13), we can obtain $\{F_1(v), v \in V_n\}$, $\{F_2(v), v \in V_n\}$, $\ldots$, and $\{F_{k_{\max}}(v), v \in V_n\}$, where $k_{\max}$ (in our case $k_{\max} + 1 = 1/\delta$) is an "upper bound" of the number of change points. See Section 3.2 for more details. By considering

$G(k)$ in (6), we have $F_k(1) = G(k)$ with $k = 1, \ldots, k_{\max}$. Hence, we are ready to estimate the change point number by

$$\hat{k} = \underset{k=1,\ldots,k_{\max}}{\arg \min} F_k(1). \tag{14}$$

The corresponding locations of change points $\hat{\boldsymbol{\tau}} = (0, \hat{\tau}_1, \ldots, \hat{\tau}_{\hat{k}}, 1)^{\top}$ can be obtained by

$$\hat{\tau}_j = \underset{u \in V_n, u < \hat{\tau}_{j+1}}{\arg \min} \left\{ F_{j-1}(u) + P_n \rho \left((u, \hat{\tau}_{j+1}), \hat{\boldsymbol{\beta}}_{(u,\hat{\tau}_{j+1})} \right) + \gamma \right\}, \quad \text{for } j = \hat{k}, \ldots, 1. \tag{15}$$

The following Algorithm 1 describes our procedure for obtaining $\hat{k}$ and $\hat{\boldsymbol{\tau}}$ based on the DPA. Note that DPA solves Equation (9) with globally optimal solutions, which have excellent estimation accuracy. Furthermore, as shown in Leonardi & Bühlmann (2016), it has the computational cost of $O(n^2 \text{GlmLasso}(n))$ operations. This can be computationally expensive, especially when n is very large. Hence, it is desirable to consider a more efficient approach. Next, we introduce an efficient approach based on binary segmentation, which can ensure almost the same estimation accuracy as that of DPA.

Algorithm 1. Dynamic programming procedure for change point detection in high-dimensional GLMs.

Input:	Given the dataset $\{\mathbf{X}, \boldsymbol{Y}\}$, set the value of $k_{\max}$.
Step 1:	Based on Equations (12)–(13), compute $F_k(1)$ for $k = 1, \ldots, k_{\max}$.
Step 2:	Obtain estimate of the number of change points $\hat{k}$ by Equation (14).
Step 3:	Obtain estimate of the change point locations $\hat{\boldsymbol{\tau}}$ by Equation (15).
Output:	Algorithm 1 provides the change point estimator $\hat{\boldsymbol{\tau}}(\hat{k}) = (0, \hat{\tau}_1, \ldots, \hat{\tau}_{\hat{k}}, 1)^{\top}$, including both the number and locations.

2.2.2. Binary segmentation approach

Next we introduce an approach based on the binary segmentation algorithm (BSA) examined in Cho & Fryzlewicz (2012, 2015), and Leonardi & Bühlmann (2016), which is shown to be much more efficient compared with DPA. The main idea of BSA for solving the change point problem for GLMs (Eq. 9) is that for each candidate search interval (u, v), we use the penalized loss function to determine whether a new change point s can be added. If s is identified, then the interval (u, v) is split into two subintervals (u, s) and (s, v) and we conduct the above procedure on (u, s) and (s, v) separately. This algorithm is continued until no new subintervals can be added. In particular, for any given $u, v \in V_n := \{i/n : i = 1, \ldots, n\}$, we define

$$Z(u, v) = \begin{cases} P_n \rho \left((u, v], \hat{\boldsymbol{\beta}}_{(u,v]} \right) + \gamma, & \text{if } (v - u)n \geq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

and

$$h(u, v) = \underset{s \in \{u\} \cup [u+\delta, v-\delta]}{\arg \min} \{ Z(u, s) + Z(s, v) \}. \tag{17}$$

Then, we present our BSA-based algorithm as follows.

Note that this approach searches many fewer candidates for finding a new change point as compared with DPA, which makes it more computationally efficient. More specifically, as shown

by Leonardi & Bühlmann (2016), BSA has a computational cost of $O(n\log(n)\mathrm{GlmLasso}(n))$ operations. Furthermore, in Section 3.2, we prove that the change point estimator computed by Algorithm 2 enjoys almost the same estimation accuracy as that of Algorithm 1.

Algorithm 2. Binary segmentation procedure for change point detection in high-dimensional GLMs.

Input:	Given the dataset $\{\mathbf{X}, Y\}$, initialize the set of change point pairs $T = \{0, 1\}$.
Step 1:	For each pair $\{u, v\}$ in T, compute $s = h(u, v)$ as defined in Equation (17). If $s > u$, add new pairs of nodes $\{u, s\}$ and $\{s, v\}$ to T and update T as $T = T \cup \{u, s\} \cup \{s, v\}$.
Step 2:	Repeat Step 1 until no more new pairs of nodes can be added. Denote the terminal set of change point pairs by $T_{\mathrm{final}} = \cup_{i=1}^{q}\{u_i, v_i\}$.
Output:	Algorithm 2 provides the change point estimator $\hat{\boldsymbol{\tau}}^b = \left(\hat{\tau}_0^b, \ldots, \hat{\tau}_{\hat{k}^b + 1}^b\right)^\top$, where $\hat{k}^b = \#T_{\mathrm{final}}$ and $0 = \hat{\tau}_0^b < \hat{\tau}_1^b < \cdots < \hat{\tau}_{\hat{k}^b}^b < \hat{\tau}_{\hat{k}^b+1}^b = 1 \in T_{\mathrm{final}}$, including both the number and locations.

2.2.3. A fast screening approach for single change point models

So far, we have proposed two efficient algorithms in Sections 2.2.1 and 2.2.2 for solving Equation (9). In this section, we show that under the single change point models, the computational cost can be further reduced. As far as we know, our fast screening approach (FSA) is novel for detecting a single change point in regression models. The main idea is that for detecting a single change point, if we have some prior information about its location, it is not necessary to search all candidate subintervals that have been adopted in the BSA-based algorithm. To see this, we recall $Z(u, v)$ as defined in Equation (16). For $\tau^f \in (0, 1)$, we define the statistics as $W_{\tau^f}((u, v)) = Z\left(u, u + \tau^f(v - u)\right) + Z\left(u + \tau^f(v - u), v\right)$. Based on $W_{\tau^f}((u, v))$, we have the following key observation: Consider any subinterval (u, v) containing a single change point $\tilde{\tau} \in (u, v)$. If $\tilde{\tau}$ lies in the first half interval of (u, v), i.e., $\tilde{\tau} \in \left(u, u + \frac{1}{2}(v - u)\right)$, with a high probability, we can prove that $W_{1/4}((u, v)) < W_{3/4}((u, v))$. If $\tilde{\tau}$ lies in the second half interval of (u, v), i.e., $\tilde{\tau} \in \left(u + \frac{1}{2}(v - u), v\right)$, with a high probability, we have $W_{3/4}((u, v)) < W_{1/4}((u, v))$. The above observation motivates us to design Algorithm 3 for fast change point identification.

Note that Algorithm 3 does not need to search through all data points, and it can quickly identify the half interval where the change point is located by comparing the quarter, half, and three-quarter values of $W(T_i)$ in each iteration. As a result, it only takes $O(\log(n)\mathrm{GLMLasso}(n))$ computational operations to detect the change point. Hence, as compared with Algorithms 1 and 2, the computational cost can be dramatically reduced. Its computational benefits will be validated by our numerical experiments in Section 4.

3. THEORETICAL PROPERTIES

We examine the theoretical properties of our proposed three approaches. In particular, we first show that our estimation of change points and regression coefficients is consistent and has the same rates of convergence as those of linear models. Secondly, for GLMs with change points, we reconstruct our assumptions and lemmas for analyzing high-dimensional data with heterogeneity based on the work of van de Geer (2008). Note that van de Geer (2008) considered the ℓ_1 penalized estimation of the regression coefficients under the setting of all observations from the same GLM. In particular, in Section 3.1, we introduce some assumptions. In Section 3.2, we present theoretical results of the change point estimator computed by the new algorithms.

Algorithm 3. A fast screening approach for single change point detection in high-dimensional GLMs

---

Input: Input the dataset $\{\mathbf{X}, Y\}$.

Step 0: Set $u_0 = 0, v_0 = 1$.

Step 1: For each iteration $i = 0, 1, 2, \ldots$, let $T_i = [u_i, v_i]$. Calculate the values of $W_{\frac{1}{4}}(T_i)$, $W_{\frac{1}{2}}(T_i)$, and $W_{\frac{3}{4}}(T_i)$. Set $M(T_i) = \min\left(W_{\frac{1}{4}}(T_i), W_{\frac{1}{2}}(T_i), W_{\frac{3}{4}}(T_i)\right)$.

For each iteration i, consider the following three cases:

If $M(T_i) = W_{\frac{1}{4}}(T_i)$, set $T_{i+1} = [u_i, u_i + 1/2(v_i - u_i)]$;

if $M(T_i) = W_{\frac{1}{2}}(T_i)$, set $T_{i+1} = [u_i + 1/4(v_i - u_i), u_i + 3/4(v_i - u_i)]$;

if $M(T_i) = W_{\frac{3}{4}}(T_i)$, set $T_{i+1} = [u_i + 1/2(v_i - u_i), v_i]$.

Step 2: Repeat Step 2 until $[n * (v_{i*} - u_{i*})] \leq 4$ holds for some i^*. Denote T_{i*} by $\hat{T}$.

Step 3: Calculate $\hat{\tau}^f = \arg\min_{\tau \in \hat{T}} W_{\tau f}(\hat{T})$.

Output: This algorithm provides a single change point estimator $\hat{\tau}^f$.

Before presenting the theoretical results, we introduce some additional notation. For any $u, v \in V_n := \{i/n : i = 1, \ldots, n\}$ with $u < v$, we denote, for a function $\rho(\mathbf{x}, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the subinterval-based theoretical mean and empirical mean by $P\rho((u, v)) := \frac{1}{n}\sum_{i=un+1}^{vn}\mathbb{E}\rho(X_i, Y_i)$, and $P_n\rho((u, v)) := \frac{1}{n}\sum_{i=un+1}^{vn}\rho(X_i, Y_i)$, respectively. For convenience, we denote $P\rho = P\rho((0, 1))$ and $P_n\rho = P_n\rho((0, 1))$. Consider a linear subspace $\mathcal{F} := \left\{f_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^p\right\}$. For a $f_{\boldsymbol{\beta}} \in \mathcal{F}$, define $\rho_{f_{\boldsymbol{\beta}}}(\mathbf{x}, y) = \rho(f_{\boldsymbol{\beta}}(\mathbf{x}), y)$. Then the empirical risk and theoretical risk at f are defined as $P_n\rho_f$ and $P\rho_f$, respectively. Furthermore, we define the target as the minimizer of the theoretical risk $f^0 := \arg\min_{f \in \mathcal{F}} P\rho_f$ and $\boldsymbol{\beta}^0 := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} P\rho_{f_{\boldsymbol{\beta}}}$, where $\boldsymbol{\beta}^0$ can be regarded as the "truth". By definition, we have $f^0(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\beta}^0$. For $f_{\boldsymbol{\beta}} \in \mathcal{F}$, the excess risk is defined as $\mathcal{E}(f_{\boldsymbol{\beta}}) := P\left(\rho_{f_{\boldsymbol{\beta}}} - \rho_{f^0}\right)$. Lastly, for any subinterval (u, v), we define the oracle $\boldsymbol{\beta}^*_{(u,v)}$ as $\boldsymbol{\beta}^*_{(u,v)} := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p}\left\{\mathcal{E}\left(f_{\boldsymbol{\beta}}\right)\right\}$. The corresponding estimation error is then denoted as $\epsilon^* := \left(P_n - P\right)\rho_{f_{\boldsymbol{\beta}^*}}$.

3.1. Basic Assumptions

We introduce some assumptions as follows.

Assumption A (loss function)**.** The loss function $\rho_f(x, y) := \rho(f(x), y)$ is convex for all $y \in \mathbb{R}$. Moreover, it satisfies the Lipschitz property:

$$|\rho(f_{\boldsymbol{\beta}}(x), y) - \rho(f_{\tilde{\boldsymbol{\beta}}}(x), y)| \leq \left|f_{\boldsymbol{\beta}}(x) - f_{\tilde{\boldsymbol{\beta}}}(x)\right|, \ \forall(x, y) \in \mathcal{X} \times \mathcal{Y}, \ \forall\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}} \in \mathbb{R}^p.$$

Assumption B (design matrix)**.** There exists $K_X < \infty$ such that $\|X_i\|_\infty \leq K_X$ and $\mathbb{E}(X_i) = 0$ hold for all $i = 1, \ldots, n$.

Assumption C (margin condition)**.** There exists an $\eta > 0$ and a strictly convex increasing function $G(x)$, such that for all $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\|f_{\boldsymbol{\beta}} - f^0\|_\infty \leq \eta$, one has

$$\mathcal{E}(f_{\boldsymbol{\beta}}) \geq G\left(\|f_{\boldsymbol{\beta}} - f^0\|\right),$$

where there exists a constant C such that $G(x) \geq Cx^2$ for any positive x.

The Canadian Journal of Statistics / La revue canadienne de statistique

Assumption D (compatibility condition)**.** The compatibility condition is met for the set $S_* = \cup_{j=1}^{\tilde{k}+1} S^{(j)}$ ($S^{(j)}$ defined in Section 2.2) with constant $\phi_* > 0$, if for all $\boldsymbol{\beta} \in \mathbf{R}^p$ satisfying $\left\|\boldsymbol{\beta}_{S_*^c}\right\|_1 \leq 3 \left\|\boldsymbol{\beta}_{S_*}\right\|_1$, it holds that

$$\left\|\boldsymbol{\beta}_{S_*}\right\|_1^2 \leq \left(\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}\right) s_* / \phi_*^2,$$

where $s_* := \#S_*$ is the cardinality of S_*.

Assumption E (parameter space)**.** For $k_0 > 1$, there exist constants $m_* > 0$ and $M_* > 0$ such that

$$\min_{1 \leq i \leq j \leq k \leq \tilde{k}+1} \frac{\left\|\sum_{r=i}^{j} \gamma(i, r, j) \boldsymbol{\beta}^0(r) - \sum_{r=j+1}^{k} \gamma(j+1, r, k) \boldsymbol{\beta}^0(r)\right\|_1}{s_*} \geq m_*,$$

where $\gamma(i, j, k) = \frac{\tilde{\tau}_j - \tilde{\tau}_{j-1}}{\tilde{\tau}_k - \tilde{\tau}_{i-1}}$,

$$s_* = o\left(\sqrt{n}/\log(p)\right), \quad \max_{1 \leq j \leq \tilde{k}+1} \|\boldsymbol{\beta}^0(j)\|_\infty \leq M_*, \quad \text{and} \quad \max_{1 < j \leq \tilde{k}+1} \|\boldsymbol{\beta}^0(j) - \boldsymbol{\beta}^0(j-1)\|_\infty \leq M_*.$$

Note that in the case $\tilde{k} = 1$, the former condition reduces to $\|\boldsymbol{\beta}^0(1) - \boldsymbol{\beta}^0(2)\|_1 \geq m_* s_*$.

We assume in Assumption A that the loss function ρ is Lipschitz in f, which allows us to bound the loss function by the difference between estimated regression parameters and the corresponding true parameters. Many functions can meet this condition, for example, the negative-likelihood function of the logistic regression model. Assumption B imposes relatively weak conditions on the covariates, which covers a wide range of distributional patterns. Assumption C (margin condition) is assumed for a "neighbourhood" of the target linear function $f^0 = X^T \beta^0$ and is a common condition for analyzing the GLM. See Section 6.4 in Bühlmann & van de Geer (2011) for more details. Assumption D (compatibility condition) for the design matrix $\mathbf{X}$ allows us to establish oracle results for Lasso estimation. Note that one can verify that Assumption D is a sufficient condition of Assumption C in van de Geer (2008) by choosing the function $D(\mathcal{K}) = \#\mathcal{K}\beta^2$, where $\#\mathcal{K}$ is the cardinality of the set $\mathcal{K} \subset \{1, \ldots, p\}$ defined in Assumption C of van de Geer (2008). Assumption E presents the minimum and maximum differences between the true regression parameters, which allow us to detect the change points. Furthermore, the sparsity of regression coefficients is required to guarantee the consistency of our proposed estimators. Assumption F introduced in the Appendix imposes some technical conditions on the tuning parameter λ for the Lasso estimation as well as the tuning parameter γ for the change point estimation. Assumption G includes the required condition for the limiting property of the de-biased Lasso estimator.

3.2. Main Results

We are ready to present some theoretical results of our proposed three new algorithms. Before that, we denote $c^* = m_*^2 \phi_*^2 / M^*$ and let d_* be a constant. See more details in Lemma 5. We first present the properties of the estimators computed by DPA in Algorithm 1.

Theorem 1. *Suppose Assumptions A − G hold with* $\log(p) = o(n)$*. Then, for a given* $C_1 > 0$*, with probability of at least* $1 - 7\exp\left(-C_1 \frac{n^2}{\log(p)}\right)$*, we have that*

(1) $l(\hat{\boldsymbol{\tau}}) = \tilde{k}$;

(2) $\|\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}}\|_1 \le c^* \sqrt{\delta}\lambda$;

(3) $\sum_{j=1}^{\tilde{k}+1} \left| P_n\rho\left(I_j(\hat{\boldsymbol{\tau}}), \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)\right) - P_n\rho\left(I_j(\hat{\boldsymbol{\tau}}), \boldsymbol{\beta}^0(j)\right)\right| + \lambda r_j(\hat{\boldsymbol{\tau}})\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) - \boldsymbol{\beta}^0(j)\|_1 \le (\tilde{k}+1)d_* s_* \lambda^2$;

(4) for each $j \in \{1, \dots, \hat{k}\}$,

$$\sqrt{n}\left(\tilde{\beta}_s(\hat{\boldsymbol{\tau}}, j) - \beta_s^0(j)\right)/\hat{\sigma}_{j,s} = V_{j,s} + o_{\mathbf{P}}(1) \text{ for } s \in \{1, \dots, p\},$$

where $\tilde{\beta}_s(\hat{\boldsymbol{\tau}}, j)$ is the sth component of $\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)$, $V_{j,s} \sim \mathcal{N}(0,1)$ and $\hat{\sigma}_{j,s}^2 := \left(\hat{\boldsymbol{\Theta}}_j P_n \dot{\rho}_{\hat{\boldsymbol{\beta}}} \dot{\rho}_{\hat{\boldsymbol{\beta}}}^T \hat{\boldsymbol{\Theta}}_j^T\right)_{s,s}$.

Theorem 1 demonstrates that Algorithm 1 can identify both the number and locations of multiple change points with high estimation accuracy. In particular, the first result shows that we can obtain a consistent estimator $l(\hat{\boldsymbol{\tau}})$ for the true number of change points. As for the locations, the second result indicates that our multiple change point estimator $\hat{\boldsymbol{\tau}}$ converges to the true change point vector $\tilde{\boldsymbol{\tau}}$ with a rate of $O_p(\sqrt{\log(p)/n})$. Furthermore, the third result implies that we can bound the prediction error or the estimation error of the underlying regression parameters within a rate of $O_p(\tilde{k}s_*\lambda^2)$ or $O_p(\tilde{k}s_*\lambda)$. Result (4) implies the asymptotic normality of the de-biased Lasso estimator $\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}})$, which allows the wider statistical inference including confidence intervals and hypothesis testing.

Based on Theorem 1, some other interesting conclusions can be made. To simplify the discussion, we require that all the $\tilde{k}+1$ change point intervals are within the same order of magnitude. Recall δ as the minimum length of change point intervals as defined in Section 2.2. Then we have $\tilde{k}(k_{\max}) = O(1/\delta)$. Furthermore, according to δ, the following two cases are considered: (1) $\delta = O(1)$ and (2) $\delta = o(1)$.

For the first case, we have $\tilde{k} = O(1)$, which means that the number of change points is fixed and does not increase with the sample size n. Furthermore, considering Assumption F1, we have $\lambda = O(\sqrt{\log(p)/n})$. Hence, the three results in Theorem 1 reduce to:

$$\|\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}}\|_1 = O_p\left(\sqrt{\log(p)/n}\right),$$

$$\sum_{j=1}^{\tilde{k}+1} \left| P_n\rho\left(I_j(\hat{\boldsymbol{\tau}}), \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)\right) - P_n\rho\left(I_j(\hat{\boldsymbol{\tau}}), \boldsymbol{\beta}^0(j)\right)\right| = O_p\left(s_* \frac{\log(p)}{n}\right), \text{ and}$$

$$\sum_{j=1}^{\tilde{k}+1} \|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) - \boldsymbol{\beta}^0(j)\|_1 = O_p\left(s_* \sqrt{\frac{\log(p)}{n}}\right). \tag{18}$$

Considering Equation (18), our results are consistent with the Lasso estimation results derived in van de Geer (2008) and estimation consistency is guaranteed as long as $s_* \sqrt{\log(p)/n} = o(1)$ holds.

We next consider the second case with $\delta = o(1)$. In this case, we allow the number of change points to grow with n. Noting that $\lambda \sqrt{\delta} = O\left(\sqrt{\log(p)/n}\right)$, the three results in Theorem 1 reduce to:

$$\|\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}}\|_1 = O_p\left(\sqrt{\log(p)/n}\right),$$

$$\sum_{j=1}^{\tilde{k}+1} \left| P_n\rho\left(I_j(\hat{\boldsymbol{\tau}}), \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)\right) - P_n\rho\left(I_j(\hat{\boldsymbol{\tau}}), \boldsymbol{\beta}^0(j)\right)\right| = O_p\left(s_* \frac{\log(p)}{n\delta^2}\right), \text{ and}$$

$$\sum_{j=1}^{\tilde{k}+1} \|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) - \boldsymbol{\beta}^0(j)\|_1 = O_p\left(s_* \delta^{-3/2} \sqrt{\frac{\log(p)}{n}}\right). \tag{19}$$

Hence, by Equation (19), the estimation consistency can still be obtained as long as $s_* \delta^{-3/2} \sqrt{\frac{\log(p)}{n}} = o(1)$ holds. In other words, the number of change points $\tilde{k}$ cannot grow faster than the order of $\left(\frac{n}{\log(p) s_*^2} \right)^{1/3}$.

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

Theorem 2. *Suppose Assumptions A–G hold with $\log(p) = o(n)$. For a given $C_2 > 0$, with probability of at least $1 - 7 \exp\left(-C_2 \frac{n^2}{\log(p)} \right)$, we have that*

(1) $l(\hat{\boldsymbol{\tau}}^b) = \tilde{k};$

(2) $\|\hat{\boldsymbol{\tau}}^b - \tilde{\boldsymbol{\tau}}\|_1 \leq c^* \sqrt{\delta} \lambda;$

(3) $\sum_{j=1}^{\tilde{k}+1} \left| P_n \rho \left(I_j(\hat{\boldsymbol{\tau}}^b), \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}^b, j) \right) - P_n \rho \left(I_j(\hat{\boldsymbol{\tau}}^b), \boldsymbol{\beta}^0(j) \right) \right| + \lambda r_j(\hat{\boldsymbol{\tau}}^b) \| \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}^b, j) - \boldsymbol{\beta}^0(j) \|_1 \leq (\tilde{k} + 1) d_* s_* \lambda^2;$

(4) *for each $j \in \{1, \dots, \hat{k}\}$,*

$$\sqrt{n} \left(\tilde{\beta}_s(\hat{\boldsymbol{\tau}}, j) - \beta_s^0(j) \right) / \hat{\sigma}_{j,s} = V_{j,s} + o_{\mathbf{P}}(1), \text{ for } s \in \{1, \dots, p\},$$

where $\tilde{\beta}_s(\hat{\boldsymbol{\tau}}, j)$ is the sth component of $\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j)$, $V_{j,s} \sim \mathcal{N}(0, 1)$ and $\hat{\sigma}_{j,s}^2 := \left(\hat{\boldsymbol{\Theta}}_j P_n \dot{\rho}_{\hat{\boldsymbol{\beta}}} \dot{\rho}_{\hat{\boldsymbol{\beta}}}^T \hat{\boldsymbol{\Theta}}_j^T \right)_{s,s}.$

Theorem 2 shows similar results as those of Theorem 1 in terms of consistency of both the number and locations of change points. Furthermore, Theorem 2 allows us to use a much more efficient algorithm to detect multiple change points for GLMs, which enjoys almost the same estimation accuracy as that of the global solutions. The efficiency will be further investigated in our numerical experiments.

Finally, we establish theoretical properties of FSA proposed in Algorithm 3 for single change point models.

Theorem 3. *Suppose Assumptions A–F hold with $\log(p) = o(n)$. Assume that the true single change point $\tilde{\tau} \in (0, 1/2)$. For a given $C_3 > 0$, with probability of at least $1 - 7 \exp\left(-C_3 \frac{n^2}{\log(p)} \right)$, we have that*

$$W_{\frac{1}{4}}((0, 1)) < W_{\frac{3}{4}}((0, 1)). \tag{20}$$

Theorem 3 justifies the validity of Algorithm 3 and demonstrates that the cost of identifying a single change point in GLMs can be reduced to only $O(\log(n)\text{GLMLasso}(n))$ computational operations.

4. SIMULATION STUDIES

In this section, we investigate the numerical performance of our three proposed change point detection procedures in various model settings. For the design matrix $\mathbf{X}$, we generate X_i i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We first consider two types of covariance matrix structures including independent and weakly dependent settings as follows:

Case 1: $\boldsymbol{\Sigma} = \mathbf{I}_{p \times p};$

Case 2: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$ with $\boldsymbol{\Sigma}^* = (\sigma^*)_{i,j=1}^p$, where $\sigma_{i,j}^* = 0.8^{|i-j|}$ for $1 \leq i, j \leq p$.

We consider logistic regression models. For $i = 1, \ldots, n$, we generate $Y_i \in \{0, 1\}$ with $g(\mathbb{P}(Y_i = 1)) = \log \frac{\mathbb{P}(Y_i=1)}{1-\mathbb{P}(Y_i=1)} = X_i^T \beta^{(i)}$. Then the responses $\{Y_i\}_{i=1}^n$ are generated from the following binomial distribution: $Y_i | X_i \sim \text{Bin}\left(1, \frac{\exp\left(X_i^T \beta^{(i)}\right)}{1+\exp\left(X_i^T \beta^{(i)}\right)}\right)$.

For this model set-up, we investigate the performance of our approaches in terms of accuracy and efficiency. For efficiency, we compare our proposed algorithms in terms of the computational cost. Note that BSA and DPA are designed for multiple change point detection. In order to compare efficiency reasonably for the cases with no change point and a single change point, we set these two algorithms to stop after one screening by making $k_{\max} = 1$. To show the accuracy, we record the mean, mean squared error (MSE), and error rate (proportion of false positives) of the change point estimators including the number and locations. We compare the corresponding results with the following existing methods:

- Lee, Seo & Shin (2011) (denoted by Lee2011), which is based on the maximum score estimation.
- Qian & Su (2016) (denoted by SGL), which proposed a systematic estimation framework based on the adaptive fused Lasso in linear regression models. To be specific, they estimate $\{\beta_t\}_{t=1}^n$ by minimizing the ℓ_2-loss with the fused Lasso penalty. In this article, we modify SGL by replacing the ℓ_2-loss with the loss ρ defined in Section 2 for high-dimensional GLMs.
- Wang, Lin & Willett (2021) (denoted by VPWBS), i.e., the variance projected wild binary segmentation based on the sparse group Lasso estimator for linear regression models. In particular, they projected the high-dimensional time series $\{X_i, Y_i\}_{i=1}^n$ onto the univariate time series $\{z_i(u)\}_{i=1}^n$. The optimal projection direction u is obtained by local group Lasso screening (LGS). Then they conducted mean change point detection by wild binary segmentation (WBS) on the univariate time series $\{z_i(u)\}_{i=1}^n$. Note that, for linear models, LGS performs a variant of the group Lasso on any subsample $\{X_i, Y_i\}_{i=s+1}^e$, and computes

$$(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\nu}) \leftarrow \underset{\substack{\nu \in [s'+1, e'-1] \\ \alpha_1, \alpha_2 \in \mathbb{R}^p}}{\arg \min} \left\{ \sum_{i=s+1}^{\nu} \left(Y_i - X_i^\top \alpha_1\right)^2 + \sum_{t=\nu+1}^{e} \left(Y_i - X_i^\top \alpha_2\right)^2 + \lambda_G \sum_{j=1}^{p} \sqrt{(\nu - s)\left(\alpha_{1,j}\right)^2 + (e - \nu)\left(\alpha_{2,j}\right)^2} \right\}, \tag{21}$$

where s' and e' serve as boundary trimming parameters with $s + 1 \le s' + 1 < e' \le e$, and λ_G is the tuning parameter for the group penalty. For a better comparison in the context of high-dimensional GLMs, we modify this ℓ_2-loss based method in Wang, Lin & Willett (2021) by replacing the ℓ_2-loss in Equation (21) with the loss $\rho_\beta (X_i, Y_i)$ defined in Section 2.

For our proposed approaches, the regression coefficients are computed by the R package glmnet (https://glmnet.stanford.edu). All numerical results are based on 100 replications, except for the test by Lee2011, which is based on 500 replications.

4.1. Tuning Parameter Selection

It is essential to properly choose the values of tuning parameters for accurate estimation results. We develop a cross-validation approach for GLMs to choose the parameters λ and γ, which encourage regression coefficient and segmentation sparsity, respectively. To be specific, let the samples with odd indices be the training set $(X_1, X_3, \ldots, X_{n-3}, X_{n-1})$ and the others be

the validation set $(X_2, X_4, \ldots, X_{n-2}, X_n)$. For each of two tuning parameters λ, γ, we conduct our procedure on the training set and obtain the estimated change point $\hat{\tau}(\hat{k})$ and underlying regression coefficients $\hat{\beta}(\hat{\tau}, j), j = 1, \ldots, \hat{k}$. Let $\hat{f}_i = X_i^\top \hat{\beta}(\hat{\tau}, j)$, for $i/n \in I_j(\tau)$ and $i = 1, \ldots, n$. We can calculate the validation loss as:

$$CV(\lambda, \gamma) = \frac{2}{n} \sum_{i:\, i \bmod 2 \equiv 1} \rho(\hat{f}_i, Y_i),$$

where ρ is the loss function of the GLMs and depends on the link function. For the specific regression models such as the linear model and logistic regression model, the corresponding validation losses are defined as:

$$CV_{LM}(\lambda, \gamma) = \frac{2}{n} \sum_{i:\, i \bmod 2 \equiv 1} (\hat{f}_i - Y_i)^2, \text{ and}$$

$$CV_{Logic}(\lambda, \gamma) = \frac{2}{n} \sum_{i:\, i \bmod 2 \equiv 1} \log\left(1 + e^{\hat{f}_i}\right) - y\hat{f}_i.$$

Then we choose (λ, γ) corresponding to the lowest validation loss. Note that it is time-consuming to use the cross-validation procedure to choose the tuning parameters for our various model settings. Based on our extensive numerical simulations, we find that our methods are stable over a certain range of tuning parameters. Hence, we use an empirical choice of the parameters λ and γ to save computational cost. In particular, we set $\lambda = c\left(\sqrt{\log(2p)/n} + \log(2p)/n\right)$, with $c \in (0.15, 0.25)$. As for γ, we set $\gamma = \delta\lambda$. Recall δ is the minimum interval length and δn is the minimum interval size, which controls the maximum number of change points. Note that δ is of key importance for our theoretical guarantee discussed in Section 3.2 and needs to be carefully chosen in simulation studies.

In order to ensure the effective fitting of the regression model, we need to guarantee a sufficient sample size for each interval. According to our numerical studies, setting $\delta \in (0.1, 0.25)$ works well. To investigate how sensitive our proposed methods are to the choice of these tuning parameters, we consider various values of λ and γ by setting the sample size $n \in \{200, 300, 1000\}$ and data dimension $p \in \{200, 300, 400\}$. Note that our proposed methods can automatically account for the underlying data generation mechanism and does not need to know the number of change points. To justify this, in what follows, we present our numerical results under three different cases: (1) $\tilde{k} = 0$, (2) $\tilde{k} = 1$, and (3) $\tilde{k} = 3$, which correspond to data with no change point, one change point, and multiple change points, respectively.

4.2. No Change Point Models

We consider the alternative scenario where no change point occurs. In this case, the underlying regression coefficients satisfy $\beta^{(i)} = \beta^0 := (\beta_1^0, \ldots, \beta_p^0)^\top$ for $i = 1, \ldots, n$. We set the sample size $n = 200$ and the data dimension $p \in \{200, 300, 400\}$. For $s \in S^0$, we generate $\beta_s^0 \overset{\text{iid}}{\sim} U(0, 2)$, where S^0 denotes the set of nonzero elements of β^0 with $\#S^0 = [\log(p)]$.

We implement the corresponding algorithms independently on a 2.50 GHz CPU (Linux) with 6 cores and 4 GB of RAM. As shown in Figure 1 (left), the computational cost of BSA grows moderately (12–737 s) as the data dimension increases from 400 to 2000, while the computational cost of DPA grows exponentially (80–32,000 s). As for the accuracy, the error rates of VPWBS, DPA, and BSA are zero in almost all cases, which suggests these three approaches have almost the same accuracy when no change point occurs. SGL tends to overestimate the number of change points for the homogeneous observations. Note that Lee2011 has relatively large errors,
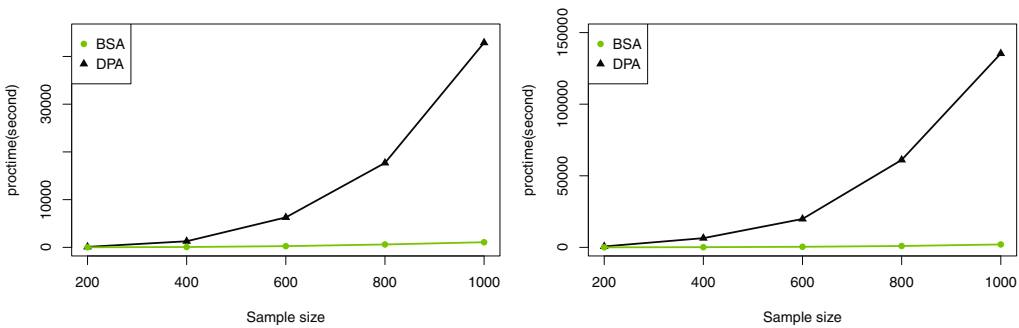
FIGURE 1: Efficiency of change point estimation with $p = 2n$. The left panel shows computational costs of BSA and DPA per replication under the model with no change point. The right panel shows computational costs of BSA and DPA per replication under the model of three change points.

TABLE 1: Change point detection for Cases 1 and 2 under the model with no change point. The numerical results are based on averages of 100 replications.

Case	Measurement	p	Accuracy for the following methods				
			SGL	VPWBS	Lee2011	DPA	BSA
$\mathbf{\Sigma} = \mathbf{I}$	Error rate	200	0.65	0.01	0.73	0.00	0.00
		300	0.85	0.01	0.71	0.00	0.00
		400	0.86	0.01	0.74	0.00	0.00
$\mathbf{\Sigma} = \mathbf{\Sigma}^*$	Error rate	200	0.38	0.00	0.71	0.00	0.00
		300	0.44	0.01	0.71	0.00	0.00
		400	0.47	0.00	0.72	0.00	0.00

which suggests that it may be unreliable in high-dimensional settings. Thus, we do not include it in our comparisons for the single change point models (Table 1).

4.3. Single Change Point Models

Next we consider the alternative scenario where $(\boldsymbol{\beta}^{(i)})_{1 \leq i \leq n}$ have a common change point location $\tilde{\tau}_1$, where $\tilde{\tau}_1 \in \{0.5, 0.7\}$. We set the sample size $n = 300$ and the data dimension $p \in \{200, 300, 400\}$. Furthermore, we assume the regression coefficients have support set $\{S^1, S^2\}$. For $s \in S^1$, we set $\beta_s(1) \overset{\text{iid}}{\sim} \text{U}(0, 2)$. Then, for $s \in S^2$, we set $\beta_s(2) = \beta_s(1) + \delta_s$ with $\delta_s \overset{\text{i.i.d.}}{\sim} \text{U}(0, 10\sqrt{\log(p)/n})$. For each replication, the support sets S^1, S^2 of regression coefficients are randomly selected from the set $\{1, 2, \dots, 0.3p\}$ with $\#S^1 = \#S^2 = [\log(p)]$.

Figure 2 (left) indicates that the computational cost of the BSA grows gradually with the data dimension increasing from 400 to 2000 as compared with the exponential growth of the DPA. Meanwhile, the computational cost of FSA in Algorithm 3 increases slowly as compared with the "exponential" growth of the BSA, as shown in Figure 2 (right). This suggests that FSA is preferable for single change point models. Furthermore, to investigate the computational efficiency for high-dimensional cases, we present the computational cost of our three proposed approaches in Figure 3. It implies that our proposed approaches have stable and good performance as data dimension p grows.
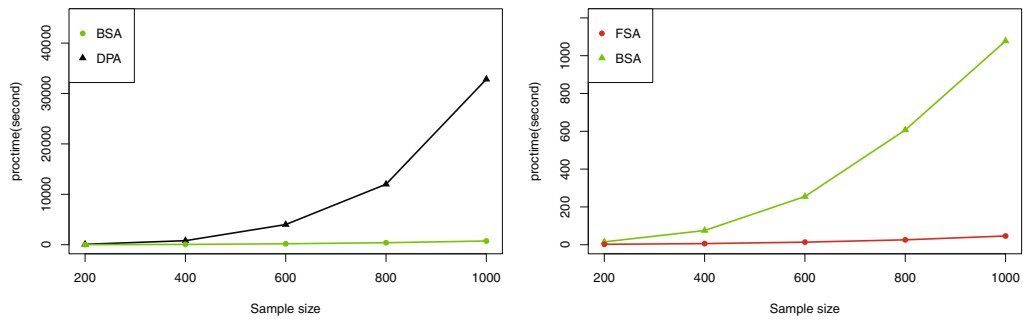
FIGURE 2: Efficiency of change point estimation under the single change point model with $n \in \{200, 400, 600, 800, 1000\}$ and $p = 2n$. The left panel shows the computational costs of BSA and DPA per replication. The right panel shows computational costs per replication of BSA and FSA. The change point is fixed at $\tau_1 = 0.5$.
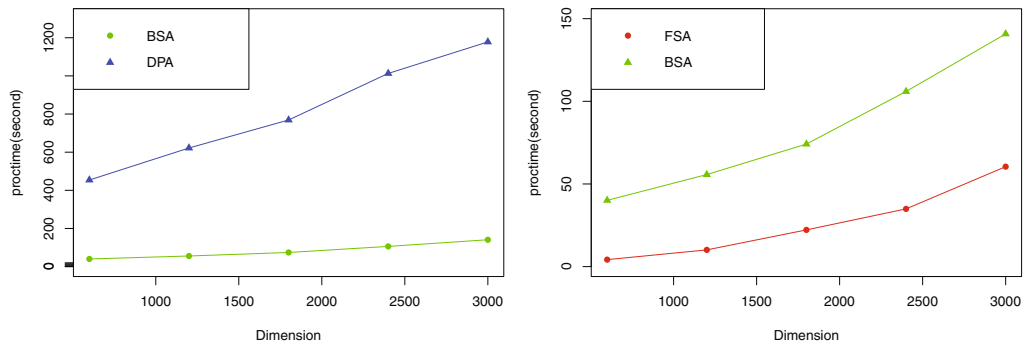

FIGURE 3: Efficiency of change point estimation under the single change point model with $p \in \{600, 1200, 1800, 2400, 3000\}$ and $n = 300$. The left panel shows the computational costs of BSA and DPA per replication. The right panel shows computational costs per replication of BSA and FSA. The change point is fixed at $\tau_1 = 0.5$.

As for the accuracy, we record the percentage of replications (rate (%)) in which DPA and BSA correctly identified a single change point. Note that in Tables 2 and 3, MSEs for the number and location are expressed as factors of 10^{-2} and 10^{-4}, respectively. We can see that DPA, BSA, and VPWBS can identify a single change point with high rates of success. Furthermore, DPA generally has the best performance for estimating the single change point location. VPWBS performs better than BSA, especially when the change occurs near the edge. Both DPA and BSA perform slightly better than FSA. Note that all the proposed algorithms perform better the closer the change point location is to the middle of the data observations, e.g., $\tilde{\tau}_1 = 0.5$.

4.4. Multiple Change Point Models

Finally, we consider the alternative scenario where $(\boldsymbol{\beta}^{(i)})_{1 \leq i \leq n}$ has multiple change point locations $\tilde{\boldsymbol{\tau}}_2$ with $\tilde{\boldsymbol{\tau}}_2 = (0, 0.25, 0.5, 0.75, 1)^\top$. We set the sample size $n = 1000$ and the data dimension $p \in \{200, 300, 400\}$. For $s_1 \in \mathcal{S}^1$, we set $\beta_{s_1}(1) \overset{\text{iid}}{\sim} U(0, 2)$. Then, for $s_j \in \mathcal{S}^j$ $(j = 2, 3, 4)$, we set $\beta_{s_j}(j) = \beta_{s_j}(j - 1) + (j - 1)\delta_{s_j}$ with $\delta_{s_j} \overset{\text{i.i.d.}}{\sim} U(0, 10\sqrt{\log(p)/n})$. For each replication, the support set of regression coefficients $\mathcal{S}$ is randomly selected from the set $\{1, 2, \ldots, 0.3p\}$ with $\#\mathcal{S}^j = [\log(p)]$, $j = 1, 2, 3, 4$.

TABLE 2: Single change point detection for Case 1 with $\Sigma = I_{p \times p}$ under various dimensions and change point locations, based on 100 replications.

	$\tilde{\tau}_1$	Method	Dimension p		
			200	300	400
Number	0.5	SGL	2.49 | 19	2.36 | 17	2.48 | 23
Mean | Rate (%)		VPWBS	1.02 | 96	1.01 | 98	1.02 | 97
		DPA	1.00 | 100	1.01 | 99	1.00 | 100
		BSA	1.00 | 100	1.00 | 100	1.00 | 100
	0.7	SGL	2.48 | 23	2.48 | 23	2.64 | 14
		VPWBS	1.01 | 97	1.04 | 96	0.99 | 95
		DPA	1.04 | 96	1.02 | 98	1.03 | 97
		BSA	1.00 | 100	1.00 | 100	1.00 | 100
Location	0.5	SGL	-	-	-
Mean | MSE (10^{-4})		VPWBS	0.497 | 2.244	0.500 | 3.590	0.496 | 6.322
		DPA	0.499 | 1.836	0.503 | 2.256	0.499 | 3.254
		BSA	0.498 | 2.446	0.501 | 3.779	0.499 | 6.445
		FSA	0.495 | 19.21	0.496 | 9.326	0.493 | 11.18
	0.7	SGL	-	-	-
		VPWBS	0.699 | 2.442	0.701 | 6.963	0.693 | 7.710
		DPA	0.694 | 3.933	0.693 | 4.630	0.691 | 5.086
		BSA	0.691 | 6.884	0.688 | 14.29	0.686 | 11.28
		FSA	0.684 | 25.09	0.678 | 43.94	0.666 | 37.07

We first analyze the efficiency. The results are similar to the other cases. Figure 4 shows that the computational cost of BSA grows gradually with the number of change points. In contrast, the cost of the DPA grows substantially. This suggests that the efficiency of BSA is not sensitive to the number of change points. To compare accuracy, similarly to the previous analysis, we record the percentage of replications in which DPA and BSA can correctly identify the three change points. As shown in Tables 4 and 5, DPA generally has the best performance. VPWBS performs slightly better than BSA. However, there is not much difference in performance among DPA, BSA, and VPWBS in terms of identifying the number and locations of multiple change points.

It is worth mentioning that our proposed methods have good performance in all three models with various data dimensions, sample sizes, and numbers of change points, which suggests that the methods are robust to the suggested choice of tuning parameters λ and γ.

5. REAL DATA ANALYSIS

To illustrate the usefulness of our proposed methods, we consider the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (https://www.loni.ucla.edu/ADNI). The ADNI dataset contains disease state information on different subjects including normal controls (NCs), mild cognitive impairment (MCI), and Alzheimer's disease (AD) as well as some biological markers including features derived from magnetic resonance imaging (MRI) and positron emission

TABLE 3: Single change point detection for Case 2 with $\Sigma = \Sigma^*$ under various dimensions and change point locations. The numerical results are based on 100 replications.

			Dimension p		
	$\tilde{\tau}_1$	Method	200	300	400
Number	0.5	SGL	1.14 | 32	1.30 | 41	1.52 | 40
Mean | Rate (%)		VPWBS	1.02 | 98	1.04 | 96	1.03 | 98
		DPA	1.01 | 99	1.00 | 100	1.00 | 100
		BSA	1.00 | 100	1.00 | 100	1.00 | 100
	0.7	SGL	0.72 | 20	1.16 | 26	1.14 | 36
		VPWBS	0.97 | 97	1.01 | 99	1.05 | 96
		DPA	1.00 | 100	1.02 |98	1.01 | 99
		BSA	0.99 | 99	1.00 |98	1.01 | 99
Location	0.5	SGL	-	-	-
Mean | MSE		VPWBS	0.504 | 5.056	0.495 | 3.146	0.498 | 3.572
		DPA	0.498 | 2.145	0.502 | 1.423	0.499 | 3.670
		BSA	0.493 | 5.897	0.498 | 3.347	0.498 | 4.326
		FSA	0.495 | 12.21	0.489 | 20.10	0.492 | 9.374
	0.7	SGL	-	-	-
		VPWBS	0.694 | 5.959	0.694 | 7.297	0.701 | 6.950
		DPA	0.690 | 5.748	0.690 | 6.411	0.691 | 12.75
		BSA	0.689 | 9.419	0.694 | 3.973	0.688 | 11.44
		FSA	0.689 | 22.41	0.683 | 24.20	0.676 | 27.66

tomography (PET). It is very useful for clinical diagnosis and prevention to study how to measure the progression of AD using the images. For example, in the AD-related literature (see, for example, Reiss & Ogden, 2010), it is popular to use structural MRI or PET to predict the current disease status of the patient (binary response variable), which can be regarded as a classification problem. Usually, they treat the data as homogeneous and ignore the effect of other covariate variables, such as age, gender, and so on. Hence, an interesting question is whether the generalized linear structure between the disease status and biomarkers (MRI or PET) changes due to other covariates. If it changes, how can one estimate (a) the number of change points, (b) the locations of change points, and (c) the regression coefficients (selected variables) in each segmentation? In our study, we address these issues by detecting change points in the generalized linear structure between the disease status and the MRI features together with some covariates. In this application, we choose age as the covariate, which is of particular interest in AD studies.

We use the MRI data of 405 subjects including 220 NCs and 185 AD patients from the ADNI data. For each subject, we obtain the corresponding status (AD/NC), age, and 93 MRI features after using the data processing method proposed in Zhang & Shen (2012). In our model, the predictive variables $X = (X_1, \ldots, X_{93})$ are the 93 MRI features, which are scaled to have mean 0 and variance 1, and the response variable is the binary status obtained by setting $Y_i = \mathbf{1}$ {subject i is an AD patient}, and 0 otherwise. For this dataset $\{X_i, Y_i\}_{i=1}^n$, consider the following logistic regression model: $\log \frac{(\mathbb{P}(Y_i=1))}{(\mathbb{P}(Y_i=0))} = X_i^\top \boldsymbol{\beta}^{(i)}$. Our goal is to detect
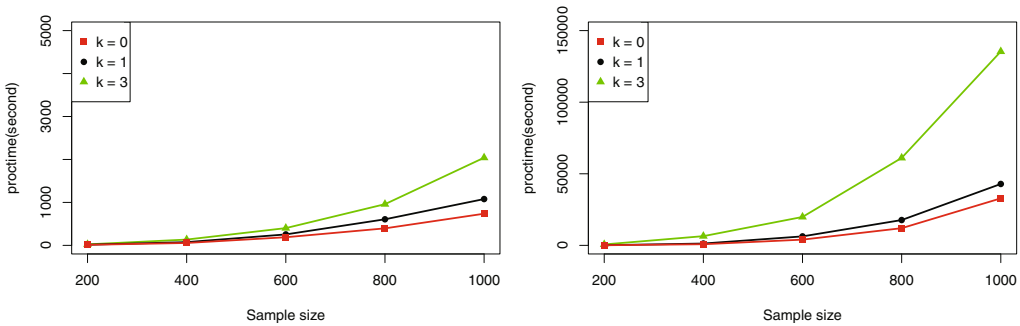
FIGURE 4: Efficiency of change point estimation under the model of multiple change points with $n \in \{200, 400, 600, 800, 1000\}$ and $p = 2n$. The left panel shows the computational costs of BSA per replication in the settings with different numbers of change points. The right panel shows computational cost of DPA per replication in the settings with different numbers of change points.

TABLE 4: Multiple change point detection for Case 1 with $\Sigma = \mathbf{I}_{p \times p}$ under various dimensions. The numerical results are based on 100 replications.

$\tau_2 = (0.25, 0.5, 0.75)^\top$		p	Accuracy for the following methods			
			SGL	VPWBS	DPA	BSA
Number	Mean | Rate (%)	200	3.19 | 55	3.02 | 98	3.00 | 100	2.95 | 95
		300	3.47 | 55	3.03 | 95	3.00 | 100	2.98 | 98
		400	3.50 | 44	3.04 | 93	3.00 | 100	2.94 | 94
Location 1	Mean | MSE (10^{-5})	200	-	0.248 | 1.108	0.249 | 0.665	0.249 | 1.821
		300	-	0.250 | 2.100	0.250 | 0.865	0.247 | 4.168
		400	-	0.249 | 3.913	0.251 | 1.847	0.256 | 5.736
Location 2	Mean | MSE (10^{-5})	200	-	0.499 | 0.894	0.500 | 0.353	0.499 | 2.099
		300	-	0.499 | 1.204	0.501 | 0.508	0.499 | 6.148
		400	-	0.498 | 2.490	0.500 | 0.981	0.500 | 6.736
Location 3	Mean | MSE (10^{-5})	200	-	0.751 | 0.701	0.750 | 1.002	0.749 | 1.689
		300	-	0.750 | 1.370	0.751 | 2.481	0.745 | 7.968
		400	-	0.748 | 2.268	0.749 | 0.688	0.748 | 2.258

potential change points of regression coefficients in $\{\boldsymbol{\beta}^{(i)}\}_{i=1}^{n}$. Taking the effect of different samples into consideration, in our analysis, we divide the data into two parts: training and testing datasets. More specifically, we randomly select 40 subjects from the whole set of 405 subjects according to the empirical distribution of age in Figure 5 (left) as the testing sample and use the remaining 365 subjects as the training data. Then we sort those 365 observations in the training data by age and use BSA to estimate the number and location of change points. We choose the tuning parameters λ and γ as suggested in Section 4. The above process is repeated 100 times. As BSA is more computationally efficient than DPA, we use only BSA to analyze the data.

Figure 5 (right) demonstrates the estimated numbers of change points for 100 replications. To be specific, among the 100 replications, 80% of the estimated numbers of change points are 1

TABLE 5: Multiple change point detection for Case 2 with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$ under various dimensions. The numerical results are based on 100 replications.

| $\boldsymbol{\tau}_2 = (0.25, 0.5, 0.75)^{\top}$ | | p | Accuracy for the following methods | | | |
			SGL	VPWBS	DPA	BSA
Number	Mean | Rate(%)	200	2.65 | 30	3.00 | 100	3.00 | 100	2.99 | 99
		300	2.08 | 17	2.97 | 99	3.00 | 100	2.98 | 98
		400	2.47 | 25	3.00 | 100	3.00 | 100	2.98 | 98
Location 1	Mean | MSE (10^{-5})	200	-	0.247 | 2.435	0.250 | 1.961	0.251 | 3.457
		300	-	0.249 | 3.164	0.251 | 2.844	0.250 | 6.585
		400	-	0.251 | 1.687	0.251 | 2.497	0.249 | 4.994
Location 2	Mean | MSE (10^{-5})	200	-	0.487 | 2.040	0.500 | 1.069	0.499 | 2.667
		300	-	0.502 | 1.403	0.501 | 0.894	0.499 | 4.363
		400	-	0.499 | 2.212	0.501 | 3.613	0.500 | 4.192
Location 3	Mean | MSE (10^{-5})	200	-	0.748 | 1.242	0.750 | 0.711	0.748 | 1.674
		300	-	0.752 | 1.672	0.750 | 0.486	0.749 | 1.398
		400	-	0.747 | 4.511	0.750 | 0.580	0.747 | 4.881
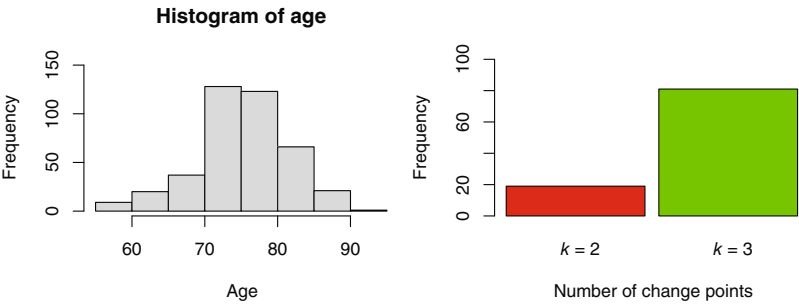
FIGURE 5: The left panel shows distribution of age among 405 subjects. The right panel shows estimated numbers of change points based on 100 replications.

($k = 3$), which suggests that there is a change point in the context of logistic regression between the disease status and the MRI features due to the change of age. Moreover, Table 6 summarizes the change point estimation results computed by BSA. For the change point estimation, 90% of change points were estimated by BSA to be located at 80 years old and the mean is 79.95 years old. This implies that there are significant differences between the regression models of the disease status and the MRI features under 80 and over 80 years old.

In terms of prediction, Table 6 provides the prediction results computed by BSA and the Lasso-based method, where for each replication, we use the training sample to select models and use the testing sample to predict. Note that the Lasso-based method treats the data as homogeneous while we consider a heterogeneous model. For the prediction result, we calculate the predictive MSE on the testing set for these two methods. Our proposed method obtains better prediction performance, which is demonstrated by a 7% lower averaged predictive MSE than that of the Lasso-based method. This suggests that treating the data as heterogeneous and using our method to select models can predict better.

TABLE 6: Change point estimation and prediction for ADNI data. The change point estimator is obtained by using BSA based on 100 replications.

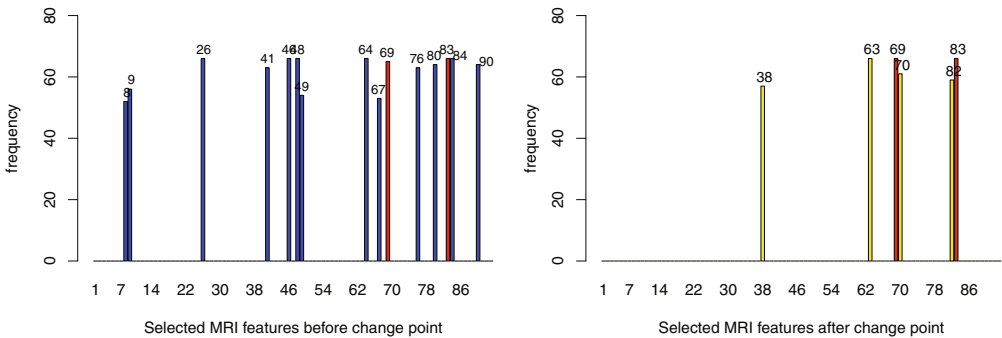| Training/testing sample	Methods	Location of change point	Averaged MSE
365/40	BSA	79.75	0.380
	Lasso-based	-	0.408

FIGURE 6: Frequency of features selected before the change point (left) and after the change point (right) for the ADNI dataset based on 100 replications. Features selected by models both before and after the change point are shown in red.

For variable selection, Figure 6 shows selected features before and after the estimated change point of 80 years old. We see that models both under 80 and over 80 both select the 69th and 83rd features, which correspond to the hippocampal and amygdala regions, respectively. These regions are known to be related to AD by many previous studies (Zhang & Shen, 2012). Moreover, there are a few different features selected separately by these two models under 80 and over 80. We believe these features deserve more scientific attention, and more research studies are needed to study their associations with AD together with age.

6. SUMMARY

In this article, we provide a three-step procedure for change point detection in the context of high-dimensional GLMs, where the dimension p can be much larger than the sample size n. It is worth mentioning that our proposed method can automatically account for the underlying data generation mechanism ($\tilde{k} = 0$ or $\tilde{k} > 0$) without specifying any prior knowledge about the number of change points $\tilde{k}$. Moreover, based on dynamic programming and binary segmentation techniques, two algorithms, DPA and BSA, are proposed to detect multiple change points. To further improve the computational efficiency, we present a much more efficient algorithm designed for the case of a single change point. Furthermore, we investigate the theoretical properties of our proposed change point estimators computed by the three algorithms. Estimation consistency for the number and locations of change points is established. Finally, we demonstrate the efficiency and accuracy of our proposed methods by extensive numerical results under various model settings. A real data application to the ADNI dataset also demonstrates the usefulness of our proposed methods.

ACKNOWLEDGEMENTS

REFERENCES

Aue, A., Hörmann, S., Horváth, L., & Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B), 4046–4087.

Barigozzi, M., Cho, H., & Fryzlewicz, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1), 187–225.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., & Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1), 157–183.

Braun, J. V., Braun, R. K., & Müller, H. G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2), 301–314.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media, Berlin.

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when *p* is much larger than *n*. *The Annals of Statistics*, 35(6), 2313–2351.

Cho, H. & Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22(1), 207–229.

Cho, H. & Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2), 475–507.

Ciuperca, G. (2014). Model selection by LASSO methods in a change-point model. *Statistical Papers*, 55(2), 349–374.

Fan, J. & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1), 101–148.

Fan, J., Lv, J., & Qi, L. (2011). Sparse high-dimensional models in economics. *Annual Review of Economics*, 3(1), 291–317.

Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3), 928–961.

Fong, Y., Di, C., & Permar, S. (2015). Change point testing in logistic regression models with interaction term. *Statistics in Medicine*, 34(9), 1483–1494.

Frick, K., Munk, A., & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3), 495–580.

Harchaoui, Z. & Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492), 1480–1493.

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

Kirch, C., Muhsal, B., & Ombao, H. (2015). Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, 110(511), 1197–1216.

Lavielle, M. & Teyssière, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3), 287–306.

Lee, S. & Seo, M. H. (2008). Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. *Journal of Econometrics*, 144(2), 492–499.

Lee, S., Seo, M. H., & Shin, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association*, 106(493), 220–231.

Lee, S., Seo, M. H., & Shin, Y. (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. & Bühlmann, P. (2016). Computationally efficient change point detection for high-dimensional regression. arXiv preprint, arXiv:1601.03704.

Li, D., Qian, J., & Su, L. (2016). Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association*, 111(516), 1804–1819.

Liu, B., Zhang, X., & Liu, Y. (2021). Simultaneous change point inference and structure recovery for high dimensional Gaussian graphical models. *Journal of Machine Learning Research*, 22, 1–62.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4), 523–527.

Pesaran, M. H. & Pick, A. (2007). Econometric issues in the analysis of contagion. *Journal of Economic Dynamics and Control*, 31(4), 1245–1277.

Qian, J. & Su, L. (2016). Shrinkage estimation of regression models with multiple structural changes. *Econometric Theory*, 32(6), 1376–1433.

Raginsky, M., Willett, R. M., Horn, C., Silva, J., & Marcia, R. F. (2012). Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8), 5544–5562.

Reiss, P. T. & Ogden, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66(1), 61–69.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273–282.

Tickle, S. O., Eckley, I. A., Fearnhead, P., & Haynes, K. (2020). Parallelization of a common changepoint detection method. *Journal of Computational and Graphical Statistics*, 29(1), 149–161.

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

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2), 614–645.

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

Wang, D., Yu, Y., & Rinaldo, A. (2021). Optimal covariance change point localization in high dimensions. *Bernoulli*, 27(1), 554–575.

Wang, T. & Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 57–83.

Zhang, D. & Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2), 895–907.

Zhang, T. & Lavitas, L. (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association*, 113(522), 637–648.

APPENDIX

Notation

We introduce some notation. The random part, empirical process, is defined as $\left\{ v_n(\boldsymbol{\beta}) := (P_n - P) \rho_{f_{\boldsymbol{\beta}}} : \boldsymbol{\beta} \in \mathbb{R}^p \right\}$. We recall the Lasso estimator is, for $j = 1, \ldots, k$, $\hat{\boldsymbol{\beta}}(j) = \arg\min_{\boldsymbol{\beta}} \left\{ P_n \rho_{f_{\boldsymbol{\beta}}} + \lambda r_j(\boldsymbol{\tau}) \|\boldsymbol{\beta}\|_1 \right\}$. We write $\hat{f} = f_{\hat{\boldsymbol{\beta}}}$ and $\mathcal{E}(\boldsymbol{\beta}) := P\left(\rho_{f_{\boldsymbol{\beta}}} - \rho_{f_{\boldsymbol{\beta}^0}}\right)$ for convenience. Recall that for any (u, v), the oracle $\boldsymbol{\beta}^*$ is defined as $\boldsymbol{\beta}^*_{(u,v)} := \arg\min_{\boldsymbol{\beta}} \left\{ \mathcal{E}\left(f_{\boldsymbol{\beta}}\right) \right\}$, which is the best approximation of $\boldsymbol{\beta}^0$ under the compatibility condition, $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0$, if there is no change point between u and v. The estimation error is denoted as $\epsilon^* := v_n(\boldsymbol{\beta}^*) = (P_n - P) \rho_{f_{\boldsymbol{\beta}^*}}$. We define, for some positive constant $L > 0$, $\mathbf{Z}_L := \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq L} \left| v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*) \right|$. We set $L^* := \epsilon^*/\lambda_0$ and require a relatively small L^*, so this indicates that $\epsilon^* \leq \lambda^0$. Based on this, for any $u, v \in V_n := \{i/n : i = 1, \ldots, n\}$ with $u < v$, we define two important sets as follows:

$$\mathcal{T}_0 := \left\{ \mathbf{Z}_{L^*} \leq \epsilon^* \leq \lambda_0 \right\}, \tag{A1}$$

$$\mathcal{T}_1 := \left\{ \max_{(u,v)} \left\| \hat{\Sigma}_{(u,v)} - (v - u)\Sigma \right\|_\infty \leq \lambda_1 \right\}. \tag{A2}$$

We introduce the following assumptions.

Assumption F. We require some technical conditions as follows:

(F1) $\lambda\sqrt{\delta} \geq 8\lambda_0$, where $\lambda_0 = O(\sqrt{\log(p)/n})$.

(F2) $\gamma > 3d_* s_* \lambda^2$, where $d_* = O(1)$ with detailed definition introduced in the Appendix.

(F3) $\frac{\delta m_*^2 \phi_*^2 s_*}{8} > \gamma + 22\epsilon^*$, where $\epsilon^* = O(s_* \log(p)/n)$.

(F4) $\frac{\lambda\delta m_*^2 \phi_*^2 s_*}{8} > 22\epsilon^* - \lambda\delta M_*$.

Assumption G. We require some conditions for achieving limit properties for the de-biased Lasso estimator:

(G1) The derivatives $\dot{\rho}(Y, a) := \frac{d}{da}\rho(Y, a)$, $\ddot{\rho}(Y, a) := \frac{d^2}{da^2}\rho(Y, a)$, exist for all y, a, and for some δ-neighbourhood ($\delta > 0$), $\ddot{\rho}(Y, a)$ is Lipschitz:

$$\max_{a_0 \in \left\{X_i^\top \beta^0\right\}} \sup_{|a-a_0| \vee |\hat{a}-a_0| \leq \delta} \sup_{Y \in \mathcal{Y}} \frac{|\ddot{\rho}(Y, a) - \ddot{\rho}(Y, \hat{a})|}{|a - \hat{a}|} \leq 1$$

Moreover,

$$\max_{a_0 \in \left\{X_i^\top \beta^0\right\}} \sup_{Y \in \mathcal{Y}} \left|\dot{\rho}\left(y, a_0\right)\right| = O(1), \text{ and } \max_{a_0 \in \left\{X_i^\top \beta^0\right\}} \sup_{|a-a_0| \leq \delta} \sup |\ddot{\rho}(Y, a)| = O(1).$$

(G2) It holds that $\left\|P_n \ddot{\rho}_{\hat{\beta}} \hat{\Theta}_j^T - e_j\right\|_\infty = O_P\left(\lambda_*\right)$.

(G3) It holds that $\left\|\mathbf{X}\hat{\Theta}_j^T\right\|_\infty = O_P(K)$ and $\left\|\hat{\Theta}_j\right\|_1 = O_P\left(\sqrt{s_*}\right)$.

(G4) It holds that $\left\|(P_n - P)\dot{\rho}_{\beta^0}\dot{\rho}_{\beta^0}^T\right\|_\infty = O_P\left(K^2\lambda\right)$ and moreover $\max_j 1/\left(\hat{\Theta}P\dot{\rho}_{\beta^0}\dot{\rho}_{\beta^0}^T\hat{\Theta}^T\right)_{j,j} = O(1)$.

(G5) For every j, the random variable $\frac{\sqrt{n}\left(\hat{\Theta}P_n \dot{\rho}_{\beta^0}\right)_j}{\sqrt{\left(\hat{\Theta}P\dot{\rho}_{\beta^0}\dot{\rho}_{\beta^0}^T\hat{\Theta}^T\right)_{j,j}}}$ converges weakly to a

$N(0, 1)$-distribution.

Useful Lemmas

We introduce some useful lemmas that are essential for our main results. More specifically, Lemma 1 presents the upper bound of the difference of the subinterval penalized empirical average of the loss function based on the Lasso and oracle estimators. Corollary 1 shows the equation in Lemma 1 holds with high probability. Lemma 2 provides the results for the subinterval based on the compatibility condition. The margin condition based on the oracle estimator is updated in Lemma 3. Lemma 4 presents the lower bound of the difference of the loss function based on the oracle estimator of adjacent subintervals. Finally, Lemma 5 provides the upper bound of the difference of the subinterval penalized empirical average of the loss function based on the oracle estimator and the truth. Next, we will introduce these useful lemmas in detail.

Lemma 1 (Oracle inequality for the Lasso)**.** *Suppose Assumptions A–F hold for all* $\|\hat{\beta} - \beta^*\|_1 \leq L^*$, *as well as* $\|f_\beta - f_\beta^*\|_\infty \leq L^* K_X$. *Suppose that* λ *satisfies the inequality*

$\lambda\sqrt{\delta} \geq 8\lambda_0$. *Then on the set* $\mathcal{T}_0 \cup \mathcal{T}_1$ *given in (A1)–(A2), we have*

$$\left| P_n\rho\left((u,v),\hat{\beta}_{(u,v)}\right) - P_n\rho\left((u,v),\beta^*_{(u,v)}\right) \right| + \lambda\sqrt{(u-v)}\|\hat{\beta} - \beta^*\|_1 \leq 6\epsilon^*, \qquad (A3)$$

where there exists a constant $C_3 > 0$ *such that* $\epsilon^* \leq C_3 s^* \log(p)/n$.

Proof. Because the assumptions hold, on $\mathcal{T}_0 \cup \mathcal{T}_1$, $8\lambda_0 < \lambda\sqrt{\delta} < \lambda\sqrt{(v-u)}$, we have

$$\left| P\rho\left((u,v),\hat{\beta}_{(u,v)}\right) - P\rho\left((u,v),\beta^*_{(u,v)}\right) \right| + \lambda\sqrt{(u-v)}\|\hat{\beta} - \beta^*\|_1 \leq 4\epsilon^*,$$

by Theorem 6.4 in Bühlmann & van de Geer (2011). By the definition of $P_n\rho, P\rho$, and $v(\beta)$ introduced in Section 3.1, we can obtain

$$\left| P_n\rho\left((u,v),\hat{\beta}_{(u,v)}\right) - P_n\rho\left((u,v),\beta^*_{(u,v)}\right) \right|$$

$$\leq \left| P\rho\left((u,v),\hat{\beta}_{(u,v)}\right) - P\rho\left((u,v),\beta^*_{(u,v)}\right) \right| + \left| v(\beta^*) - v(\hat{\beta}) \right|.$$

If the condition $\|\hat{\beta} - \beta^*\|_1 \leq L^*$ holds, on the set $\mathcal{T}_0 \cup \mathcal{T}_1$, we can have

$$\left| P_n\rho\left((u,v),\hat{\beta}_{(u,v)}\right) - P_n\rho\left((u,v),\beta^*_{(u,v)}\right) \right| + \lambda\sqrt{(u-v)}\|\hat{\beta} - \beta^*\|_1 \leq 4\epsilon^* + 2\epsilon^* = 6\epsilon^*,$$

which completes the proof. ∎

Corollary 1. *Suppose Assumptions A–F hold. Let* $a_n := 4\left(\sqrt{\frac{2\log(2p)}{n}} + \frac{\log(2p)}{n}K_X\right)$ *and*

$$\lambda_0 := \lambda_0(t) := a_n\left(1 + t\sqrt{2\left(1 + 2a_nK_X\right)} + \frac{2t^2a_nK_X}{3}\right).$$

Then we have, with probability of at least $1 - 7exp[-na_n^{-2}t^2] = 1 - \exp\left(-C_1\frac{n^2}{\log(p)}\right)$, *that Equation (A3) holds. We refer to Theorem 2.1 in* van de Geer (2008).

Lemma 2. *Suppose Assumption D and* $s_*\lambda_1 \leq \frac{\phi_*^2}{32}$ *hold. Then on the set* $\mathcal{T}_0 \cup \mathcal{T}_1$, *we have, for all* $(u,v) \in \{i/n, i = 1, \ldots, n\}$ *and all* $\beta \in \mathbb{R}^p$ *with* $\left\|\beta_{S_*}\right\|_1 \leq 3\left\|\beta_{S_*}\right\|_1$,

$$\left\|\beta_{S_*}\right\|_1^2 \leq \frac{(\beta^\top\hat{\Sigma}_{(u,v)}\beta)s_*}{(v-u)\phi_*^2}.$$

Proof. By Assumption D (the compatibility condition), for any $u, v \in \{i/n, n = 1, \ldots, n\}$, we have

$$\left\|\beta_{S_*}\right\|_1^2 \leq \frac{\|f_\beta\|^2(v-u)s_*}{(v-u)\phi^2} = \frac{\left(\beta^T(v-u)\Sigma\beta\right)s_*}{(v-u)\phi_*^2},$$

for all $\beta \in \mathbb{R}^p$ that satisfy $\left\|\beta_{S_*}\right\|_1 \leq 3 \left\|\beta_{S_*}\right\|_1$. Then the matrix $(v - u)\Sigma$ satisfies the compatibility condition for the set S_* with constant $\sqrt{(v - u)\psi_*}$. By Corollary 6.8 in Bühlmann and van de Geer (2011), if $s_* \lambda_1 \leq \frac{\phi_*^2}{32}$, the compatibility condition also holds for the set S_* and the matrix $\hat{\Sigma}_{(u,v)}$, with $\phi_{\hat{\Sigma}_{(u,v)}}^2 \geq (v - u)\phi_*^2/2$. Then we can obtain, for all $\beta \in \mathbb{R}^n$ that satisfy $\left\|\beta_{S_\epsilon}\right\|_1 \leq 3 \left\|\beta_{S.}\right\|_1$,

$$\left\|\beta_{S.}\right\|_1^2 \leq \frac{\left(\beta^T \hat{\Sigma}_{(u,x)}\beta\right) s_*}{\phi_{\hat{\Sigma}_{(0,v)}}^2} \leq \frac{2\left(\beta^T \hat{\Sigma}_{(u,t)}\beta\right) s_*}{(v - u)\phi_*^2}.$$

∎

Lemma 3. *By Assumption B, there exists an $\eta \geq 0$ and strictly convex increasing G, such that for all $\beta_1, \beta_2 \in \mathbb{R}^p$ with $\|f_{\beta_1} - f^0\|_\infty \leq \eta/2$, $\|f_\beta - f^0\|_\infty \leq \eta \|\beta^* - \beta^0\| \leq \eta$, we have*

$$\left|P_n \rho(\beta) - P_n \rho(\beta^*)\right| \geq C\|X((\beta - \beta^*))\|_2^2 - 2\epsilon^*.$$

Proof. According to Assumption C (margin condition), we can directly have

$$\mathcal{E}(f_\beta) \geq G(\|f_\beta - f^0\|) \geq C\|f_\beta - f^0\|^2.$$

By the definition of $\mathcal{E}(f_\beta)$, $v(\beta)$ introduced in Notation and the triangle inequality, we have

$$P_n \rho(\beta) - P_n \rho(\beta^*) \geq C\|f_\beta - f^0\|^2 - 2\epsilon^*.$$

By the definition of β^*, under the compatibility condition, β^* is the best approximation of β^0: $\beta^* = \beta^0$. Then we have

$$P_n \rho(\beta) - P_n \rho(\beta^*) \geq C\|X((\beta - \beta^*))\|_2^2 - 2\epsilon^*.$$

∎

Lemma 4. *Suppose $k_0 > 1$ and that Assumptions A–F and $s_* \lambda_1 \leq \frac{\phi_*^2}{32}$ hold. Then on $\mathcal{T}_0 \cap \mathcal{T}_1$, if $(u, v) \subset \left(\tilde{\tau}_{j-1} - c^*\sqrt{\delta}\lambda, \tilde{\tau}_{j+1} + c^*\sqrt{\delta}\lambda\right)$ and $u < \tilde{\tau}_j < v$ for some $j = 2, \ldots, \tilde{k} - 1$, we have*

$$\left|P_n \rho\left((u, \tilde{\tau}_j), \beta_{(u,v)}^*\right) - P_n \rho\left((u, \tilde{\tau}_j), \beta_{(u,\tilde{\tau}_j)}^*\right)\right| + \left|P_n \rho\left((\tilde{\tau}_j, v), \beta_{(u,v)}^*\right) - P_n \rho\left((\tilde{\tau}_j, v), \beta_{(\tilde{\tau}_j,v)}^*\right)\right|$$

$$\geq \frac{\min\left(\tilde{\tau}_j - u, v - \tilde{\tau}_j\right) m_*^2 \phi_*^2 s_*}{8} - 4\epsilon^*.$$

Proof. By Lemmas 2 and 3, if $(u, v) \subset \left(\tilde{\tau}_{j-1} - c^*\sqrt{\delta}\lambda, \tilde{\tau}_{j+1} + c^*\sqrt{\delta}\lambda\right)$ and $u < \tilde{\tau}_j < v$ for some $j = 2, \ldots, \tilde{k} - 1$, we have

$$\left|P_n \rho\left((u, \tilde{\tau}_j), \beta_{(u,v)}^*\right) - P_n \rho\left((u, \tilde{\tau}_j), \beta_{(u,\tilde{\tau}_j)}^*\right)\right| + \left|P_n \rho\left((\tilde{\tau}_j, v), \beta_{(u,v)}^*\right) - P_n \rho\left((\tilde{\tau}_j, v), \beta_{(\tilde{\tau}_j,v)}^*\right)\right|$$

$$\geq C\left\|X\left(\beta_{(u,v)}^* - \beta_{(u,\tilde{\tau}_j)}^*\right)\right\|_2^2 + C\left\|X\left(\beta_{(u,v)}^* - \beta_{(\tilde{\tau}_j,v)}^*\right)\right\|_2^2 - 4\epsilon^*$$

$$\geq \frac{(\tilde{\tau}_j - u)\left\|\boldsymbol{\beta}^*_{(u,v]} - \boldsymbol{\beta}^*_{(u,\tilde{\tau}_j)}\right\|_1^2 \phi_*^2}{2s_*} + \frac{(v - \tilde{\tau}_j)\left\|\boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^*_{(\tilde{\tau}_j,v]}\right\|_1^2 \phi_*^2}{2s_*} - 4\epsilon^*.$$

Now observe that

$$(v - u)\boldsymbol{\beta}^*_{(u,v]} = (\tilde{\tau}_j - u)\boldsymbol{\beta}^*_{(u,\tilde{\tau}_j)} + (v - \tilde{\tau}_j)\boldsymbol{\beta}^*_{(\tilde{\tau}_j,i)}.$$

Then

$$\boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^*_{(u,\tilde{\tau}_j)} = \left(\frac{v - \tilde{\tau}_j}{v - u}\right)\left(\boldsymbol{\beta}^*_{(\tilde{\tau}_j,v]} - \boldsymbol{\beta}^*_{(u,\tilde{\tau}_j)}\right),$$

by Assumption E. If $(u, v) \subset \left(\tilde{\tau}_{j-1} - c^*\sqrt{\delta}\lambda, \tilde{\tau}_{j+1} + c^*\sqrt{\delta}\lambda\right)$ and $u < \tilde{\tau}_j < v$ for some $j = 2, \dots, \tilde{k} - 1$, we have

$$\left\|\boldsymbol{\beta}^*_{(u,v]} - \boldsymbol{\beta}^*_{(u,\tilde{\tau}_j]}\right\|_1 \geq \frac{(v - \tilde{\tau}_j)m_* s_*}{(v - u)}, \quad \left\|\boldsymbol{\beta}^*_{(u,x]} - \boldsymbol{\beta}^*_{(\tilde{\tau}_j,x)}\right\|_1 \geq \frac{(\tilde{\tau}_j - u)m_* s_*}{(v - u)}. \tag{A4}$$

Then, by the above Equation (A4) and straightforward calculations, we can obtain

$$\frac{(\tilde{\tau}_j - u)\left\|\boldsymbol{\beta}^*_{(u,v]} - \boldsymbol{\beta}^*_{(u,\tilde{\tau}_j)}\right\|_1^2 \phi_*^2}{2s_*} + \frac{(v - u)\left\|\boldsymbol{\beta}^*_{(u,v]} - \boldsymbol{\beta}^*_{(\tilde{\tau}_j,v]}\right\|_1^2 \phi_*^2}{2s_*}$$

$$\geq \frac{(\tilde{\tau}_j - u)^2 + (v - \tilde{\tau}_j)^2}{(v - u)^2} \frac{m_*^2 \phi_*^2 s_*}{2} - 4\epsilon^*.$$

As $\left((\tilde{\tau}_j - u)^2 + (v - \tilde{\tau}_j)^2\right)/(v - u)^2 \geq 1/2$, we can complete the proof. ∎

Lemma 5. *Suppose Assumptions A–F hold and let* $(u, v) \subset \left(\tilde{\tau}_{j-1} - c^*\sqrt{\delta}\lambda, \tilde{\tau}_j + c^*\sqrt{\delta}\lambda\right) \cap (0, 1]$ *for some* $j = 1, \dots, \tilde{k} + 1$, *and* $s_*\lambda_1 \leq \frac{\phi_*^2}{32}, c^*\sqrt{\delta}\lambda < r(\tilde{\boldsymbol{\tau}})$. *Then on the set* $\mathcal{T}_0 \cap \mathcal{T}_1$, *we have*

$$\left|P_n\rho\left((u,v), \hat{\boldsymbol{\beta}}_{(u,v)}\right) - P_n\rho\left((u,v), \boldsymbol{\beta}^0(j)\right)\right| + \lambda\sqrt{(u-v)}\|\hat{\boldsymbol{\beta}}_{(u,v)} - \boldsymbol{\beta}^0(j)\|_1 \leq d_* s_* \lambda^2,$$

where $b = \mathbf{1}\left\{u < \tilde{\tau}_{j-1}\right\} + \mathbf{1}\left\{\tilde{\tau}_j < v\right\}, d_* = \left((b^2 K_X^2 c^* M_* + b)c^* M_* + 6C_4\right).$

Proof. Firstly, by straightforward calculations, we can obtain

$$\left|P_n\rho\left((u,v), \hat{\boldsymbol{\beta}}_{(u,v)}\right) - P_n\rho\left((u,v), \boldsymbol{\beta}^0(j)\right)\right| + \lambda(u - v)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0(j)\|_1$$

$$\leq \left|P_n\rho\left((u,v), \hat{\boldsymbol{\beta}}_{(u,v)}\right) - P_n\rho\left((u,v), \boldsymbol{\beta}^*_{(u,v)}\right)\right| + \left|P_n\rho\left((u,v), \boldsymbol{\beta}^*_{(u,v)}\right) - P_n\rho\left((u,v), \boldsymbol{\beta}^0(j)\right)\right|$$

$$+ \lambda\sqrt{(u-v)}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_{(u,v)}\|_1 + \lambda\sqrt{(u-v)}\|\boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^0(j)\|_1.$$

According to Lemma 1 and on the set $\mathcal{T}_0$, we can obtain

$$\left|P_n\rho\left((u,v), \hat{\boldsymbol{\beta}}_{(u,v)}\right) - P_n\rho\left((u,v), \boldsymbol{\beta}^*_{(u,v)}\right)\right| + \lambda\sqrt{(u-v)}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_{(u,v)}\|_1 \leq 6\epsilon^* \leq 6C_4 s^* \lambda^2. \tag{A5}$$

Now, we will present the bias between $\boldsymbol{\beta}^*_{u,v}$ and $\boldsymbol{\beta}^0_j$, with $(u,v) \subset \left(\tilde{\tau}_{j-1} - c^* \sqrt{\delta} \lambda, \tilde{\tau}_j + c^* \sqrt{\delta} \lambda \right)$. Because $\lambda c^* \sqrt{\delta} < r(\boldsymbol{\tau})$, by Assumption E, we can have that

$$
\begin{aligned}
\left\| \boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^0(j) \right\|_\infty &\leq \frac{\max\left(\tilde{\tau}_{j-1} - u, 0 \right)}{(v-u)} \left\| \boldsymbol{\beta}^0(j) - \boldsymbol{\beta}^0(j-1) \right\|_\infty \\
&\quad + \frac{\max\left(v - \tilde{\tau}_j, 0 \right)}{(v-u)} \left\| \boldsymbol{\beta}^0(j+1) - \boldsymbol{\beta}^0(j) \right\|_\infty \\
&\leq \frac{bM^* c^* \lambda}{\sqrt{v-u}}.
\end{aligned}
\tag{A6}
$$

Furthermore, combining Assumptions A and C, Equation (A6), and the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&\left| P_n \rho \left((u,v), \boldsymbol{\beta}^*_{(u,v)} \right) - P_n \rho \left((u,v), \boldsymbol{\beta}^0(j) \right) \right| + \lambda \sqrt{(u-v)} \left\| \boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^0(j) \right\|_1 \\
&\leq (v-u) \left\| X(\boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^0(j)) \right\|_2^2 + \lambda \sqrt{(u-v)} \left\| \boldsymbol{\beta}^*_{(u,v)} - \boldsymbol{\beta}^0(j) \right\|_1 \\
&\leq \left(b^2 K_X^2 M^* s_* + b \right) c^* M^* s_* \lambda^2.
\end{aligned}
\tag{A7}
$$

Combining Equations (A5) and (A7) can complete the proofs.   ∎

Proof of Theorem 1. To simplify the notation, we denote the value of the penalized total loss function corresponding to the change point vector by

$$
H(\boldsymbol{\tau}) = \sum_{j=1}^{l(\boldsymbol{\tau})+1} P_n \rho \left(I_j(\boldsymbol{\tau}), \widehat{\beta}(\boldsymbol{\tau}, j) \right) + \gamma l(\boldsymbol{\tau}).
\tag{A8}
$$

First, we will show that if the assumptions hold, we must have $l(\hat{\boldsymbol{\tau}}) = \tilde{k}$ and $\|\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}}\|_1 \leq c^* \sqrt{\delta} \lambda$. On the contrary, we assume $\hat{\boldsymbol{\tau}}$ does not satisfy the above two results. We can distinguish three possible cases:

Case 1. Change point number is overestimated, $l(\hat{\boldsymbol{\tau}}) > \tilde{k}$. There exist some i, $1 \leq i \leq \hat{k} - 1$, such that $\{\hat{\tau}_{i-1}, \hat{\tau}_i, \hat{\tau}_{i+1}\} \subset \left(\tilde{\tau}_{j-1} - c^* \sqrt{\delta} \lambda, \tilde{\tau}_j + c^* \sqrt{\delta} \lambda \right)$ for some j, $1 \leq j \leq \tilde{k}$.

Case 2. Change point number is underestimated, $l(\hat{\boldsymbol{\tau}}) < \tilde{k}$. For some $j = 1, \ldots, \tilde{k} - 1$, we have $\hat{\boldsymbol{\tau}} \cap \left(\tilde{\tau}_j - c^* \sqrt{\delta} \lambda, \tilde{\tau}_j - c^* \sqrt{\delta} \lambda \right) = \emptyset$ and $\hat{\boldsymbol{\tau}} \cap \left(\tilde{\tau}_j - c^* \sqrt{\delta} \lambda, \tilde{\tau}_j - c^* \sqrt{\delta} \lambda \right) = \emptyset$.

Case 3. Change point number is correctly estimated, $l(\hat{\boldsymbol{\tau}}) = \tilde{k}$. However, for some $j = 1, \ldots, \tilde{k} - 1$, we have $\hat{\boldsymbol{\tau}} \cap \left(\tilde{\tau}_j - c^* \sqrt{\delta} \lambda, \tilde{\tau}_j - c^* \sqrt{\delta} \lambda \right) = \emptyset$ and $\hat{\boldsymbol{\tau}} \cap \left(\tilde{\tau}_j - c^* \sqrt{\delta} \lambda, \tilde{\tau}_j - c^* \sqrt{\delta} \lambda \right) \neq \emptyset$.

We first consider Case 1, where we have $l(\hat{\boldsymbol{\tau}}) > \tilde{k}$ and there exists some i, such that $\{\hat{\tau}_{i-1}, \hat{\tau}_i, \hat{\tau}_{i+1}\} \subset (\tilde{\tau}_{j-1}, \tilde{\tau}_j)$ for some j, $1 \leq j \leq \tilde{k}$. We define

$$
\boldsymbol{\tau} = \left(\hat{\tau}_1, \ldots, \hat{\tau}_{i-1}, \hat{\tau}_{i+1}, \ldots, \hat{\tau}_{l(\hat{\tau})} \right).
$$

Then we get a new change point vector $\boldsymbol{\tau}$ with $l(\boldsymbol{\tau}) = l(\hat{\boldsymbol{\tau}}) - 1$. Denote the intervals by $S_1 = \left(\hat{\tau}_{i-1}, \hat{\tau}_i \right]$, $S_2 = \left(\hat{\tau}_i, \hat{\tau}_{i+1} \right]$, and $S = (\hat{\tau}_{i-1}, \hat{\tau}_{i+1}]$, and then we obtain

$$
H(\boldsymbol{\tau}) - H(\hat{\boldsymbol{\tau}}) = P_n \rho \left(S, \widehat{\boldsymbol{\beta}}_S \right) - P_n \rho \left(S_1, \widehat{\boldsymbol{\beta}}_{S_1} \right) - P_n \rho \left(S_2, \widehat{\boldsymbol{\beta}}_{S_2} \right) - \gamma.
\tag{A9}
$$

By the definition of the Lasso estimator $\widehat{\beta}$ and the triangle inequality, we can directly have

$$P_n\rho\left(S, \widehat{\beta}_S\right) \le P_n\rho\left(S, \beta^0(j)\right) + \lambda\sqrt{\|S\|}\left\|\beta^0(j) - \widehat{\beta}_J\right\|_1. \tag{A10}$$

Then, combining Equations (A9)–(A10), we can directly have

$$\begin{aligned}
H(\boldsymbol{\tau}) - H(\widehat{\boldsymbol{\tau}}) \le{}& P_n\rho\left(S, \beta^0(j)\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right) \\
&+ \lambda\sqrt{\|S\|}\left\|\beta^0(j) - \widehat{\beta}_J\right\|_1 - \gamma.
\end{aligned} \tag{A11}$$

Using some straightforward calculations and the triangle inequality, we have

$$\begin{aligned}
&P_n\rho\left(S, \beta^0(j)\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right) \\
&= P_n\rho\left(S_1, \beta^0(j)\right) + P_n\rho\left(S_2, \beta^0_j\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right) \\
&\le \left|P_n\rho\left(S_1, \beta^0(j)\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right)\right| + \left|P_n\rho\left(S_2, \beta^0(j)\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right)\right|. \tag{A12}
\end{aligned}$$

By Lemma 5 and the above Equation (A12), with $(u, v) = S_1, S_2$, then we have

$$P_n\rho\left(S, \beta^0(j)\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right) \le 2d_*s^*\lambda^2. \tag{A13}$$

Also by Lemma 5, for the second term in Equation (A10), we can directly have

$$\lambda\sqrt{\|S\|}\left\|\beta^0(j) - \widehat{\beta}_J\right\|_1 \le d_*s^*\lambda^2, \tag{A14}$$

and therefore, by combining Equation (A11) and Equations (A13)–(A14), we can easily obtain

$$H(\boldsymbol{\tau}) - H(\widehat{\boldsymbol{\tau}}) \le 3d_*s^*\lambda^2 - \gamma. \tag{A15}$$

According to Assumption F2, we obtain $M(\boldsymbol{\tau}) < M(\widehat{\boldsymbol{\tau}})$, which is a contradiction.

For Case 2, where we have $l(\widehat{\boldsymbol{\tau}}) < \tilde{k}$, we define a new change points vector $\boldsymbol{\tau} = \widehat{\boldsymbol{\tau}} \cup \{\tilde{\tau}_j\}$, that is,

$$\boldsymbol{\tau} = \left(\widehat{\tau}_1, \dots, \widehat{\tau}_{r_i-1}, \widehat{\tau}_{r_i}, \widehat{\tau}_{r_i+1}, \dots, \widehat{\tau}_{l(\widehat{\tau})+1}\right), \tag{A16}$$

where $\widehat{\tau}_{r_i} = \tilde{\tau}_j$. We obtain a new change point vector $\boldsymbol{\tau}$ with $l(\boldsymbol{\tau}) = l(\widehat{\boldsymbol{\tau}}) + 1$. Also we denote the intervals by $S_1 = \left(\widehat{\tau}_{r_i-1}, \widehat{\tau}_{r_i}\right], S_2 = \left(\widehat{\tau}_{r_i}, \widehat{\tau}_{r_i+1}\right]$ and $S = \left(\widehat{\tau}_{r_i-1}, \widehat{\tau}_{r_i+1}\right]$, then we have

$$H(\widehat{\boldsymbol{\tau}}) - H(\boldsymbol{\tau}) = P_n\rho\left(S, \widehat{\beta}_S\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right) - \gamma. \tag{A17}$$

By Lemma 1, for $u, v \in V_n$, we can obtain $P_n\rho\left((u, v), \widehat{\beta}_{(u,v)}\right) - P_n\rho\left((u, v), \beta^*_{(u,v)}\right) \le 6\epsilon^*$. Thus, by this inequality (with $(u, v) = S_1, S_2$, and S), the triangle inequality and Lemma 4, we can have

$$P_n\rho\left(S, \widehat{\beta}_S\right) - P_n\rho\left(S_1, \widehat{\beta}_{S_1}\right) - P_n\rho\left(S_2, \widehat{\beta}_{S_2}\right)$$

$$\geq P_n \rho\left(S, \boldsymbol{\beta}_S^*\right) - P_n \rho\left(S_1, \boldsymbol{\beta}_{S_1}^*\right) - P_n \rho\left(S_2, \boldsymbol{\beta}_{S_2}^*\right) - 3 * (6\epsilon^*)$$

$$\geq \frac{\delta m_*^2 \phi_*^2 s_*}{8} - 4\epsilon^* - 3 * (6\epsilon^*). \tag{A18}$$

Then, by combining the above Equations (A17)–(A18), we can directly have

$$H(\hat{\boldsymbol{\tau}}) - H(\boldsymbol{\tau}) \geq \frac{\delta m_*^2 \phi_*^2 s_*}{8} - 4\epsilon^* - 3 * (6\epsilon^*) - \gamma, \tag{A19}$$

by Assumption F3, we have $H(\hat{\boldsymbol{\tau}}) > H(\boldsymbol{\tau})$, which is a contradiction.

For Case 3 with $l(\hat{\boldsymbol{\tau}} = \tilde{k})$, we must add some points and remove others to obtain a good change point estimator. Then we define $\boldsymbol{\tau}^{(1)} = \hat{\boldsymbol{\tau}} \bigcup \{\tilde{\tau}_j\}$ with $\tau_{r_i}^{(1)} = \tilde{\tau}_j$. We denote the intervals by $S_1 = \left(\hat{\tau}_{r_i-1}, \hat{\tau}_{r_i}\right], S_2 = \left(\hat{\tau}_{r_i}, \hat{\tau}_{r_i+1}\right],$ and $S = \left(\hat{\tau}_{r_i-1}, \hat{\tau}_{r_i+1}\right],$ and then we can obtain

$$H(\hat{\boldsymbol{\tau}}) - H(\boldsymbol{\tau}^{(1)}) = P_n \rho\left(S, \widehat{\boldsymbol{\beta}}_S\right) - P_n \rho\left(S_1, \widehat{\boldsymbol{\beta}}_{S_1}\right) - P_n \rho\left(S_2, \widehat{\boldsymbol{\beta}}_{S_2}\right) - \gamma. \tag{A20}$$

Without loss of generality, we assume $|S_1| < \delta$ and define a new partition $\boldsymbol{\tau} = \boldsymbol{\tau}^{(1)} \setminus \left\{\tau_{r_i-1}^{(1)}\right\}$. By denoting $K_1 = \left(\tau_{r_i-1}^{(2)}, \tau_{r_i-1}^{(1)}\right)$ and $I = K_1 \bigcup S_1$, then we have that

$$H(\boldsymbol{\tau}^{(1)}) - H(\boldsymbol{\tau}) = P_n \rho\left(K_1, \widehat{\boldsymbol{\beta}}_{K_1}\right) + P_n \rho\left(S_1, \widehat{\boldsymbol{\beta}}_{S_1}\right) - P_n \rho\left(I, \widehat{\boldsymbol{\beta}}_I\right) + \gamma. \tag{A21}$$

Thus, by combining Equations (A20) and (A21), with straightforward calculations, we have

$$H(\hat{\boldsymbol{\tau}}) - H(\boldsymbol{\tau}) = H(\hat{\boldsymbol{\tau}}) - H\left(\boldsymbol{\tau}^{(1)}\right) + H\left(\boldsymbol{\tau}^{(1)}\right) - H(\boldsymbol{\tau})$$

$$= P_n \rho\left(S, \hat{\boldsymbol{\beta}}_S\right) - P_n \rho\left(S_2, \hat{\boldsymbol{\beta}}_{S_2}\right) + P_n \rho\left(K_1, \hat{\boldsymbol{\beta}}_{K_1}\right) - P_n \rho\left(I, \hat{\boldsymbol{\beta}}_I\right),$$

by the definition of the Lasso estimator and the triangle inequality, we can obtain

$$\sqrt{|I|}\left\|\boldsymbol{\beta}_{K_1}^* - \hat{\boldsymbol{\beta}}_I\right\|_1 \leq \sqrt{|I|}\left\|\boldsymbol{\beta}_I^* - \hat{\boldsymbol{\beta}}_I\right\|_1 + \sqrt{|I|}\left\|\boldsymbol{\beta}_{K_1}^* - \boldsymbol{\beta}_I^*\right\|_1$$

$$\leq \sqrt{|I|}\left\|\boldsymbol{\beta}_I^* - \hat{\boldsymbol{\beta}}_I\right\|_1 + \sqrt{|I|}\frac{\delta}{\sqrt{|I|}}M_* s_*$$

$$\leq \sqrt{|I|}\left\|\boldsymbol{\beta}_I^* - \hat{\boldsymbol{\beta}}_I\right\|_1 + \delta M_* s_*. \tag{A22}$$

By Equation (A22), Lemmas 1 and 4, and straightforward calculations, we have

$$P_n \rho\left(S, \hat{\boldsymbol{\beta}}_S\right) - P_n \rho\left(S_2, \hat{\boldsymbol{\beta}}_{S_2}\right) + P_n \rho\left(K_1, \hat{\boldsymbol{\beta}}_{K_1}\right) - P_n \rho\left(I, \hat{\boldsymbol{\beta}}_I\right)$$

$$\geq P_n \rho\left(S, \boldsymbol{\beta}_S^*\right) - P_n \rho\left(S_2, \boldsymbol{\beta}_{S_2}^*\right) + P_n \rho\left(K_1, \boldsymbol{\beta}_{K_1}^*\right)$$

$$- 18\epsilon^* - P_n \rho\left(I, \boldsymbol{\beta}_{K_1}^*\right) - \lambda\sqrt{|I|}\left\|\boldsymbol{\beta}_{K_1}^* - \hat{\boldsymbol{\beta}}_I\right\|_1$$

$$\geq P_n \rho\left(S, \boldsymbol{\beta}_S^*\right) - P_n \rho\left(S_2, \boldsymbol{\beta}_{S_2}^*\right) - P_n \rho\left(S_1, \boldsymbol{\beta}_{S_1}^*\right) - 18\epsilon^* - \lambda|I|\left\|\boldsymbol{\beta}_I^* - \hat{\boldsymbol{\beta}}_I\right\|_1 - \lambda\delta M_* s_*$$

$$\geq \frac{\lambda\delta m_*^2 \phi_s^2 s_*}{8} - 22\epsilon^* - \lambda\delta M_*.$$

According to Assumption F4, we can obtain $H(\hat{\boldsymbol{\tau}}) - H(\boldsymbol{\tau}) > 0$, which is a contradiction. Above all, the first two results (1) and (2) in Theorem 1 have been proved. Result (3) in Theorem 1 can be directly obtained by combining (2) in Theorem 1 with Lemma 5.

Now we prove the fourth result in Theorem 1. Using that by condition $\left|x_i\hat{\boldsymbol{\Theta}}_j^T\right| = \mathcal{O}_{\mathbb{P}}(K)$, we have

$$\hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\boldsymbol{\beta}^0} + \hat{\boldsymbol{\Theta}}P_n\ddot{\rho}_{\hat{\boldsymbol{\beta}}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) + \mathrm{Rem}_1,$$

where

$$\mathrm{Rem}_1 = \mathcal{O}_{\mathbb{P}}(K)\sum_{i=1}^n \left|x_i\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right|^2/n = \mathcal{O}(K)\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right\|_n^2$$

$$= \mathcal{O}_{\mathbb{P}}\left(Ks_0\lambda^2\right) = o_{\mathbb{P}}(1),$$

it follows that

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j) = \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j) - \hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\hat{\boldsymbol{\beta}}}$$

$$= \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j) - \hat{\boldsymbol{\Theta}}_j P_n\dot{\rho}_{\boldsymbol{\beta}^0} - \hat{\boldsymbol{\Theta}}_j P_n\ddot{\rho}_{\hat{\boldsymbol{\beta}}}\left(\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j)\right) - \mathrm{Rem}_1$$

$$= -\hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\boldsymbol{\beta}^0} - \left(\hat{\boldsymbol{\Theta}}P_n\ddot{\rho}_{\hat{\boldsymbol{\beta}}} - I\right)\left(\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j)\right) - \mathrm{Rem}_1$$

$$= -\hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\boldsymbol{\beta}^0} - \mathrm{Rem}_2.$$

By the proof of Theorem 3.1 in van de Geer et al. (2014), we have $\hat{\boldsymbol{\Theta}}P_n\ddot{\rho}_{\hat{\boldsymbol{\beta}}} - I = O(\lambda)$. According to the third result in Theorem 1, we have $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j) = O_P(s_*\lambda)$. Then, it follows that

$$\left|\mathrm{Rem}_2\right| \leq \left|\mathrm{Rem}_1\right| + O\left(\lambda\right)\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}, j) - \boldsymbol{\beta}^0(j)\right\|_1 = o_P\left(n^{-1/2}\right) + O_P\left(s_*\lambda^2\right) = o_P\left(n^{-1/2}\right),$$

since the condition $s_*\lambda^2 = o\left(s_*\frac{\log p}{n}\right) = o\left(n^{-1/2}\right)$ holds. By straight calculations, we have

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\tau}}, j) - \boldsymbol{\beta}^0(j)\right) = -\sqrt{n}\hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\boldsymbol{\beta}^0} - \sqrt{n}\mathrm{Rem}_2$$

$$= -\sqrt{n}\hat{\boldsymbol{\Theta}}P_n\dot{\rho}_{\boldsymbol{\beta}^0} - o_P(1).$$

By the proof of Theorem 3.1 in van de Geer et al. (2014), we can easily conclude that

$$\sqrt{n}\left(\tilde{\beta}_s(\hat{\boldsymbol{\tau}}, j) - \beta_s^0(j)\right)/\hat{\sigma}_{j,s} = V_{j,s} + o_{\mathbf{P}}(1), s \in \{1, \ldots, p\},$$

where $V_{j,s} \sim \mathcal{N}(0, 1)$ and $\hat{\sigma}_s^2 := \left(\hat{\boldsymbol{\Theta}}_j P_n\dot{\rho}_{\hat{\boldsymbol{\beta}}}\dot{\rho}_{\hat{\boldsymbol{\beta}}}^T\hat{\boldsymbol{\Theta}}_j^T\right)_{s,s}$. ∎

Proof of Theorem 2. First we will show that under the conditions of Theorem 1, on $\mathcal{T}_0 \cap \mathcal{T}_1$, we have three cases:

Case A. Change point number is overestimated, $l(\hat{\tau}) > \tilde{k}$. We have $\tilde{k} = 1$ and $l(\hat{\tau}^b) = 1$.
Case B. Change point number is underestimated, $l(\hat{\tau}) < \tilde{k}$. For $\tilde{k} > 1$, we have $h(0,1)$ and $l(\hat{\tau}^b) = 0$.
Case C. Change point number is correctly estimated, $l(\hat{\tau}) = \tilde{k}$. For $\tilde{k} > 1$, we have $h(0,1) \in [\delta, 1-\delta]$.

This fact can be derived straightforwardly from the proof of Theorem 1, as the objective functions coincide for 1 or 2 segments; that is, for all $u \in [0,1]$,

$$H((0,u,1)) = Z(0,u) + Z(u,1).$$

For Case A, suppose $\tilde{k} = 1$ and $\tau = (0,1)$. As in the proof of Case 1 in Theorem 1, we can obtain

$$H(\tau^0) < \min_{u \in (\delta, 1-\delta)} H((0,u,1)),$$

and $h(0,1) = 0$. We suppose $\tilde{k} > 1$ and $h(0,1) \notin \cup_{j=1}^{\tilde{k}-1} \left(\tilde{\tau}_j - c^*\delta\lambda, \tilde{\tau}_j - c^*\delta\lambda \right)$. We define $\tau^{(0)} = (0, h(0,1), 1), \tau^{(1)} = \tau^{(0)} \cup \left\{ \tilde{\tau}_j \right\}, \tau^{(2)} = \tau^{(1)} \setminus \{h(0,1)\}$.

For Case B, $h(0,1) = 0$, as in the proof of Case 2 in Theorem 1, we can obtain $H(\tau^{(0)}) > H(\tau^{(1)})$. For Case C, $h(0,1) \in [\delta, 1-\delta]$, as in the proof of Case 3 in Theorem 1, we can obtain

$$H\left(\tau^{(0)}\right) - H\left(\tau^{(2)}\right) = H\left(\tau^{(0)}\right) - H\left(\tau^{(1)}\right) + H\left(\tau^{(1)}\right) - H\left(\tau^{(2)}\right) > 0.$$

Because $h(0,1)$ minimizes Equation (17), all three cases result in a contradiction. Then, we can replace (0,1) by each subinterval and obtain the same results, which completes the proof. ∎

Proof of Theorem 3. Firstly, we recall our proposed statistics as follows:

$$W_{1/4} - W_{3/4} = P_n\rho\left(\left(0, \frac{1}{4}\right), \hat{\beta}_{\left(\frac{1}{4},1\right)}\right) + P_n\rho\left(\left(\frac{1}{4}, 1\right), \hat{\beta}_{\left(\frac{1}{4},1\right)}\right)$$
$$- P_n\rho\left(\left(0, \frac{3}{4}\right), \hat{\beta}_{\left(0,\frac{3}{4}\right)}\right) - P_n\rho\left(\left(\frac{3}{4}, 1\right), \hat{\beta}_{\left(\frac{3}{4},1\right)}\right).$$

Then, for convenience, we consider two cases for the change point location denoted by τ:

Case D. $0 < \tau \leq 1/4$,
Case E. $1/4 < \tau < 1/2$.

Next, we will discuss each case in detail. Case D: In this case, $0 < \tau \leq 1/4$, so we have

$$W_{1/4} - W_{3/4} = \left(P_n\rho\left(\left(0, \frac{1}{4}\right), \hat{\beta}_{\left(0,\frac{1}{4}\right)}\right) - P_n\rho\left(\left(0, \frac{1}{4}\right), \hat{\beta}_{\left(0,\frac{3}{4}\right)}\right)\right)$$
$$+ \left(P_n\rho\left(\left(\frac{1}{4}, \frac{3}{4}\right), \hat{\beta}_{\left(\frac{1}{4},1\right)}\right) - P_n\rho\left(\left(\frac{1}{4}, \frac{3}{4}\right), \hat{\beta}_{\left(0,\frac{3}{4}\right)}\right)\right)$$
$$+ \left(P_n\rho\left(\left(\frac{1}{4}, 1\right), \hat{\beta}_{\left(\frac{1}{4},1\right)}\right) - P_n\rho\left(\left(\frac{3}{4}, 1\right), \hat{\beta}_{\left(\frac{3}{4},1\right)}\right)\right).$$

We observe there is no change point in $(1/4, 1)$. So the Lasso estimations of any subinterval $(u, v) \subset (1/4, 1)$ make no difference and we can replace $\hat{\boldsymbol{\beta}}_{\left(\frac{1}{4}, 1\right)}$ by $\hat{\boldsymbol{\beta}}_{\left(\frac{1}{4}, \frac{3}{4}\right)}$ or $\hat{\boldsymbol{\beta}}_{\left(\frac{3}{4}, 1\right)}$, and it follows that

$$W_{1/4} - W_{3/4} = \left(P_n \rho \left(\left(0, \frac{1}{4}\right), \hat{\boldsymbol{\beta}}_{\left(0, \frac{1}{4}\right)} \right) - P_n \rho \left(\left(0, \frac{1}{4}\right), \hat{\boldsymbol{\beta}}_{\left(0, \frac{3}{4}\right)} \right) \right)$$
$$+ \left(P_n \rho \left(\left(\frac{1}{4}, \frac{3}{4}\right), \hat{\boldsymbol{\beta}}_{\left(\frac{1}{4}, \frac{3}{4}\right)} \right) - P_n \rho \left(\left(\frac{1}{4}, \frac{3}{4}\right), \hat{\boldsymbol{\beta}}_{\left(0, \frac{3}{4}\right)} \right) \right). \tag{A23}$$

We denote the intervals by $J_1 = \left(0, \frac{1}{4}\right)$, $J_2 = \left(\frac{1}{4}, \frac{3}{4}\right)$, and $J = \left(0, \frac{3}{4}\right)$. Equation (A23) can be organized as follows:

$$W_{1/4} - W_{3/4}$$
$$= \left(P_n \rho(J_1, \hat{\boldsymbol{\beta}}_{J_1}) - P_n \rho(J_1, \hat{\boldsymbol{\beta}}_J) \right) + \left(P_n \rho(J_2, \hat{\boldsymbol{\beta}}_{J_2}) - P_n \rho(J_2, \hat{\boldsymbol{\beta}}_J) \right)$$
$$= P_n \rho(J_1, \hat{\boldsymbol{\beta}}_{J_1}) + P_n \rho(J_2, \hat{\boldsymbol{\beta}}_{J_2}) - P_n \rho(J, \hat{\boldsymbol{\beta}}_J).$$

Then using the same argument as Case 1 in Theorem 1, we can obtain $W_{1/4} < W_{3/4}$.

Case E: In this case, $1/4 < \tau < 1/2$, we observe that there is no change point within these two intervals $(0, 1/4)$ and $(3/4, 1)$. Then, by straightforward calculations, we can obtain

$$W_{1/4} - W_{3/4} = P_n \rho \left(\left(\frac{1}{4}, 1\right), \hat{\boldsymbol{\beta}}_{\left(\frac{1}{4}, 1\right)} \right) - P_n \rho \left(\left(0, \frac{3}{4}\right), \hat{\boldsymbol{\beta}}_{\left(0, \frac{3}{4}\right)} \right)$$
$$= P_n \rho \left(\left(\frac{1}{4}, \frac{3}{4}\right), \hat{\boldsymbol{\beta}}_{\left(\frac{1}{4}, 1\right)} \right) + P_n \rho \left(\left(\frac{3}{4}, 1\right), \hat{\boldsymbol{\beta}}_{\left(\frac{1}{4}, 1\right)} \right)$$
$$- P_n \rho \left(\left(0, \frac{1}{4}\right), \hat{\boldsymbol{\beta}}_{\left(0, \frac{3}{4}\right)} \right) - P_n \rho \left(\left(\frac{1}{4}, \frac{3}{4}\right), \hat{\boldsymbol{\beta}}_{\left(0, \frac{3}{4}\right)} \right).$$

We denote the intervals by $(0, 1/4) = K_1$, $(1/4, 3/4) = J_1$, $(3/4, 1) = J_2$, and it follows that

$$W_{1/4} - W_{3/4}$$
$$= P_n \rho(J_1, \hat{\boldsymbol{\beta}}_J) + P_n \rho(J_2, \hat{\boldsymbol{\beta}}_J) - P_n \rho(K_1, \hat{\boldsymbol{\beta}}_I) - P_n \rho(J_1, \hat{\boldsymbol{\beta}}_I).$$

Then using the same argument as Case 3 in Theorem 1, we can obtain $W_{1/4} < W_{3/4}$. ∎