Journal of Machine Learning Research 22 (2021) 1-46 Submitted 6/19; Revised 7/21; Published 10/21

Statistically and Computationally Efficient Change Point Localization in Regression Settings

Daren Wang DWANG24@ND.EDU
Department of ACMS
University of Notre Dame
Indiana, USA

Zifeng Zhao ZIFENG.ZHAO@ND.EDU
Mendoza College of Business
University of Notre Dame
Indiana, USA

Kevin Z. Lin KEVINL1@WHARTON.UPENN.EDU
Wharton Department of Statistics and Data Science
University of Pennsylvania
Pennsylvania, USA

Rebecca Willett WILLETT@UCHICAGO.EDU
Department of Statistics
University of Chicago
Illinois, USA

Editor: Zaid Harchaoui

Abstract

Detecting when the underlying distribution changes for the observed time series is a fundamental problem arising in a broad spectrum of applications. In this paper, we study multiple change-point localization in the high-dimensional regression setting, which is particularly challenging as no direct observations of the parameter of interest is available. Specifically, we assume we observe $\{x_t, y_t\}_{t=1}^n$ where $\{x_t\}_{t=1}^n$ are p-dimensional covariates, $\{y_t\}_{t=1}^n$ are the univariate responses satisfying $\mathbb{E}(y_t) = x_t^\top \beta_t^*$ for $1 \leq t \leq n$ and $\{\beta_t^*\}_{t=1}^n$ are the unobserved regression coefficients that change over time in a piecewise constant manner. We propose a novel projection-based algorithm, Variance Projected Wild Binary Segmentation (VPWBS), which transforms the original (difficult) problem of change-point detection in p-dimensional regression to a simpler problem of change-point detection in mean of a one-dimensional time series. VPWBS is shown to achieve sharp localization rate $O_p(1/n)$ up to a log factor, a significant improvement from the best rate $O_p(1/\sqrt{n})$ known in the existing literature for multiple change-point localization in high-dimensional regression. Extensive numerical experiments are conducted to demonstrate the robust and favorable performance of VPWBS over two state-of-the-art algorithms, especially when the size of change in the regression coefficients $\{\beta_t^*\}_{t=1}^n$ is small.

Keywords: Change-point detection; High-dimensional regression; CUSUM statistics; Wild binary segmentation; Time series analysis

©2021 Daren Wang, Zifeng Zhao, Kevin Z. Lin and Rebecca Willett.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v22/19-531.html.

1. Introduction

Change-point detection and localization is a classical problem in time series analysis, in which we record a series of measurements and wish to determine whether and at what time(s) the underlying generative model has changed. Due to its flexibility, the model of a time series with multiple structural changes has a wide range of applications including econometrics [Bai and Perron (1998)], epidemiology [Jiang et al. (2021)], stock price analysis [Chen and Gupta (1997)], Internet security monitoring [Peng et al. (2004)], and genetics [Castro et al. (2018); Zhao and Yau (2021)].

Change-point detection is mostly studied and well understood in the mean change-point model, where we typically assume we observe a time series $\{y_t\}_{t=1}^n \subset \mathbb{R}^p$ such that

$$y_t = \beta_t^* + \varepsilon_t, \text{ for all } 1 \leq t \leq n.$$

Here $\{\varepsilon_t\}_{t=1}^n$ are independently and identically distributed measurement noise with mean zero and $\{\beta_t^*\}_{t=1}^n$ are the population mean vectors that change over time in a piecewise constant manner. The important task is to determine whether and where the structural changes of $\{\beta_t^*\}_{t=1}^n$ take place. There is a vast literature of change-point detection in mean for both low and high dimensions, see for example Frick et al. (2014), Cho and Fryzlewicz (2015), Cho et al. (2016), Yau and Zhao (2016) and Wang and Samworth (2018). More recently, Pein et al. (2017) introduced a method that can handle mean and variance changes simultaneously. Cribben and Yu (2017), Wang et al. (2021) and Zhao et al. (2019), among others, investigated the mean change-point problem for the dynamic Bernoulli network models. Enikeeva and Harchaoui (2019) studied the optimal change point detection boundary in the high-dimensional settings. Xie et al. (2020) considered online monitoring change point detection for streaming data in high dimensions.

However, in some other practical settings , we can only obtain indirect measurements of the (potentially high-dimensional) vectors $\{\beta_t^*\}_{t=1}^n$. Specifically, in this paper, we consider change-point detection in high-dimensional linear regression. We assume we observe the time series $\{x_t, y_t\}_{t=1}^n$, where $\{x_t\}_{t=1}^n$ are p-dimensional covariates, $\{y_t\}_{t=1}^n$ are the univariate responses satisfying $\mathbb{E}(y_t|x_t) = x_t^\top \beta_t^*$ for all $1 \leq t \leq n$ and $\{\beta_t^*\}_{t=1}^n$ are the unobserved regression coefficients that potentially change over time. We formally summarize the model as follows.

Model 1 (Change-point model in the regression setting) *Suppose for $1 \leq t \leq n$, the random covariate $x_t \in \mathbb{R}^p$ and response $y_t \in \mathbb{R}$ satisfy*

$$y_t = x_t^\top \beta_t^* + \varepsilon_t, \tag{1}$$

where the noise $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ and is independent of the covariate $x_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$[1]. In addition, there exist $K \geq 0$ change-points $\{\eta_k\}_{k=1}^K \subset \{1, \ldots, n-1\}$ such that

$$\beta_t^* = \beta_{t'}^* \quad \text{if} \quad \eta_{k-1} + 1 \leq t \leq t' \leq \eta_k, \quad \text{for all} \quad k = 1, \cdots, K+1,$$

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

1. We assume for convenience that ε_t and x_t are normally distributed. However our results remain valid as long as ε_t and x_t are i.i.d. sub-Gaussian random variables.

Notation: Before we proceed, for clarity of presentation, we first introduce necessary notation used throughout the paper. For two positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O(b_n)$ if there exists $C > 0$ such that $\limsup_{n\to\infty} a_n/b_n < C$ and write $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We write $a_n \succeq b_n$ if $\liminf_{n\to\infty} a_n/b_n = \infty$. Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of random variables. We write $x_n = O_p(b_n)$ if $x_n/b_n = O_p(1)$ and write $x_n = o_p(b_n)$ if $x_n/b_n = o_p(1)$, where $O_p(1)$ and $o_p(1)$ follow the standard probability notation of big O (stochastic boundedness) and small o (convergence to zero in probability). For a vector $\beta \in \mathbb{R}^p$, denote $\|\beta\|_2 = \sqrt{\sum_{i=1}^p \beta_i^2}$ as its l_2 norm, denote $\|\beta\|_\infty = \max_{1\le i\le p} |\beta_i|$ as its l_∞ norm, and denote $\|\beta\|_0 = \sum_{i=1}^p \mathbb{I}(\beta_i \neq 0)$ as its l_0 norm, where β_i denotes the ith element of β and $\mathbb{I}$ is the indicator function. Given two natural numbers $s < e$, for simplicity, with a slight abuse of notation, we denote $[s, e] := \{t \in \mathbb{N} | s \le t \le e\}$ and denote $(s, e] := \{t \in \mathbb{N} | s < t \le e\}$. In other words, $[s, e]$ contains all natural numbers from s to e (inclusive) and $(s, e]$ contains all natural numbers from $s+1$ to e (inclusive). Throughout the paper, we use c and C to denote generic absolute constants independent of n and p, and the value of c and C may vary from place to place.

For change-point detection in Model (1), the key task is to estimate the unknown $\{\eta_k\}_{k=1}^K$. For any change-point estimator $\{\widehat{\eta}_k\}_{k=1}^{K'}$, we say it is consistent if, with probability approaching 1, $K' = K$ and the sup-norm error satisfies

$$\epsilon := \max_{1\le k\le K} \frac{|\widehat{\eta}_k - \eta_k|}{n} = o_p(1). \tag{2}$$

for all sufficiently large n.

In the literature, change-point detection for low-dimensional ($p \ll n$) linear regression models has been extensively studied by many authors including Bai and Perron (1998), Qu and Perron (2007), and more recently Zhang et al. (2015b). Most of the existing works in this setting focus on the case where the number of change-points, K, is a fixed constant.

Change-point detection for the high-dimensional linear regression model where $p \gg n$, has also received recent attention. In particular, Lee et al. (2016) extended Lasso to the high-dimensional single change-point setting and showed that both the change-point and the regression parameters $\{\beta_t^*\}_{t=1}^n$ can be consistently estimated. Later, Lee et al. (2018) extended their results to the high-dimensional quantile regression model. Kaul et al. (2019) proposed a highly efficient algorithm for the setting of exactly one change-point. Both Lee et al. (2016) and Kaul et al. (2019) showed that in the *single* change-point setting, the change-point can be estimated with sup-norm error satisfying $\epsilon = O_p(1/n)$. Zhang et al. (2015a) studied the Sparse Group Lasso (SGL) algorithm for the multiple change-points setting. The authors showed that SGL returns consistent change-point estimators with $\epsilon = o_p(1)$ when the number of change-points K is bounded. Leonardi and Bühlmann (2016) showed that, by using a binary search algorithm, consistent estimation can be achieved with $\epsilon = O_p(1/\sqrt{n})$ even when the number of change-points K diverges as $n \to \infty$.

In this paper, we focus on the high-dimensional regime ($p \gg n$) and propose a computationally efficient algorithm that can consistently estimate the unknown multiple change-points at the minimax optimal localization rate $O_p(1/n)$ up to a log factor. To the best of our knowledge, no other method in the literature can achieve this rate for multiple change-points estimation under such setting. We refer to more detailed discussion of our contribution at the end of this section.

We proceed by imposing some mild general assumptions on the high-dimensional regression setting in Model (1) and define key quantities that are used to quantify the localization error rate and requirements on the signal-to-noise ratio (SNR) of various change-point estimation methods for Model (1).

Assumption 1
a. *[Design matrix] There exist absolute positive constants c_x and C_x such that the minimal and maximal eigenvalues of the covariance matrix Σ satisfy $\Lambda_{\min}(\Sigma) \geq c_x$ and $\Lambda_{\max}(\Sigma) \leq C_x$.*
b. *[Sparse support] There exist a collection of subsets $\{S_k\}_{k=1}^{K+1} \subset \mathbb{R}^p$ such that, for all $k = 1, \cdots, K+1$,*

$$\beta^*_{t,j} = 0 \text{ if } \eta_{k-1} + 1 \leq t \leq \eta_{k+1} \text{ and } j \notin S_k.$$

In addition, the size of the support satisfies $\max_{1\leq k\leq K+1} |S_k| \leq \mathfrak{s}$ and there exists an absolute constant C_β such that $\max_{1\leq t\leq n} \|\beta^*_t\|_\infty \leq C_\beta < \infty$.

Key quantities: Define $\mathfrak{N} := \max_{1\leq t\leq n} \|\beta^*_t\|_2^2$. By Assumption 1**b**, we have $\mathfrak{N} \leq C_\beta^2 \mathfrak{s}$. Moreover, by Assumption 1**a**, we have $\mathrm{Var}(y_t) = \beta^{*\top}_t \Sigma \beta^*_t + \sigma^2_\varepsilon > c_x \|\beta^*_t\|_2^2$, and thus $\mathfrak{N} < \max_{1\leq t\leq n} \mathrm{Var}(y_t)/c_x$. For $k = 1, \cdots, K+1$, denote $\Delta_k = \eta_k - \eta_{k-1}$ as the spacing between two consecutive change-points and define $\Delta = \min_{1\leq k\leq K+1} \Delta_k$ as the minimum spacing. In addition, for $k = 1, \cdots K$, denote $\kappa_k = \left\|\beta^*_{\eta_k+1} - \beta^*_{\eta_k}\right\|_2$ as the l_2-norm of the change in regression coefficients and define $\kappa = \min_{1\leq k\leq K} \kappa_k$ as the minimum change size. Intuitively, the difficulty of change-point detection for Model 1 depends on the interplay among K, κ, Δ, $\mathfrak{s}$, the dimension p and the sample size n.

We remark that our later theoretical analysis allows the number of change-points K, the minimum change size κ, the minimum spacing Δ, the sparsity $\mathfrak{s}$ and the dimension p to vary with the sample size n. To our best knowledge, this is among the most flexible frameworks in the literature.

Our contributions: For change-point estimation in the high-dimensional regression model, we propose a novel two-stage detection procedure named Variance Projected Wild Binary Segmentation (VPWBS). Given the observations $\{x_t, y_t\}_{t=1}^n$, in Stage 1, VPWBS estimates the regression coefficients $\{\beta^*_t\}_{t=1}^n$ using a group Lasso based local screening algorithm carefully tailored for the high-dimensional regression change-point setting. In Stage 2, via a novel projection step, VPWBS projects the high-dimensional regression data $\{x_t, y_t\}_{t=1}^n$ into a one-dimensional time series, where the optimal projection direction is derived from the estimated $\{\widehat{\beta}_t\}_{t=1}^n$ in Stage 1. Subsequently, VPWBS achieves change-point estimation by performing mean change detection via CUSUM statistics on the resulting one-dimensional time series.

Our theoretical analysis shows that VPWBS can achieve consistent estimation even when the number of change-points K diverges as $n \to \infty$. Furthermore, the sup-norm error ϵ (defined in (2)) of the VPWBS change-point estimator is, up to a log factor, of order $O_p(1/n)$, which is the known minimax optimal rate. To the best of our knowledge, this is a significant improvement for *multiple* change-point estimation in the high-dimensional regression setting, as the aforementioned existing literature can only achieve $\epsilon = O_p(1/\sqrt{n})$ at best. A key step of VPWBS is the estimation of an optimal projection direction. In the theoretical analysis, we establish error bounds on the estimated high-dimensional projection

direction, which may be of independent interest. VPWBS admits a reasonable computational cost of order $O(n(\log(n))^2 \cdot \text{GroupLasso}(n,p))$, which enables its implementation in the high-dimensional regression setting. Here $\text{GroupLasso}(n,p)$ denotes the computational cost of the group Lasso for a p-dimensional regression with n samples. Similar definition applies to $\text{Lasso}(n,p)$. We summarize the localization error bound and computational cost of VPWBS and two other state-of-the-art methods in Table 1 and refer more detailed discussions to Sections 3 and 4.

	Localization Error Bound ϵ	SNR Condition	Computational Complexity
VPWBS	$O_p\big(\mathfrak{N}\log(n)n^{-1}\big)$	$\Delta\kappa^2 \succeq \mathfrak{s}\log(pn)$	$O(n(\log(n))^2 \cdot \text{GroupLasso }(n,p))$
EBSA	$O_p\big(\mathfrak{s}\log(p)n^{-1/2}\big)$	$\Delta\kappa^2 \succeq \mathfrak{N}\mathfrak{s}\log(p)$	$O(n\log(n) \cdot \text{Lasso }(n,p))$
SGL	$o_p(1)$	$\Delta\kappa^2 \asymp n$	$O(\text{Lasso}(n,np))$

Table 1: SGL (Zhang et al., 2015b) and EBSA (Leonardi and Bühlmann, 2016) are two state-of-the-art methods developed for change-point estimation in high-dimensional regression. Recall $\mathfrak{N} \leq C_\beta^2\mathfrak{s}$ and we refer the detailed definition of notation $n, p, \Delta, \kappa, \mathfrak{s}, \mathfrak{N}$ to Assumption 1 and the discussion on key quantities.

The rest of the paper is organized as follows. In Section 2, we introduce the projection based change-point estimation framework and a group Lasso based local screening algorithm for the estimation of the optimal projection direction. Building upon wild binary segmentation, Section 3 proposes the VPWBS for multiple change-point estimation in high-dimensional regression and further establishes its optimal theoretical properties. Extensive numerical experiments are conducted in Section 4 to demonstrate the promising performance of VPWBS when compared with state-of-the-art methods in the literature. Section 5 concludes with a discussion. Technical proofs can be found in the supplementary material.

2. A General Framework and Group Lasso Based Screening

In this section, we introduce the general framework of the proposed change-point estimation procedure for the high-dimensional regression problem in Model (1). Specifically, Section 2.1 discusses the essential idea of a projection based change-point detection framework and Section 2.2 proposes a group Lasso based screening algorithm for estimating the unknown projection direction.

2.1 A projection based change-point estimation framework

To ease presentation, we start the discussion with the problem of single change-point estimation. Specifically, given a sample of high-dimensional regression $\{x_t, y_t\}_{t=1}^n$ with $y_t = x_t^\top \beta_t^* + \varepsilon_t$, assume there is a single change-point at an unknown time point η such that

$$\beta_t^* = \beta^{(1)} \text{ for } 1 \leq t \leq \eta \text{ and } \beta_t^* = \beta^{(2)} \text{ for } \eta + 1 \leq t \leq n.$$

To detect the existence of η and further estimate its location, we need to measure and test the difference between the unknown regression coefficients $\beta^{(1)}$ and $\beta^{(2)}$.

For two regression coefficients $\beta^{(1)}$ and $\beta^{(2)}$, it is natural to directly measure their difference via the l_2-norm $\|\beta^{(1)}-\beta^{(2)}\|_2^2$. However, under the regression context, an arguably more relevant alternative is $(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})$, which equals to $\mathrm{Var}(x_t^\top(\beta^{(1)}-\beta^{(2)}))$ as $\mathrm{Var}(x_t)=\Sigma$. Note that under Assumption 1**a**, we have that

$$c_x\|\beta^{(1)}-\beta^{(2)}\|_2^2 \le (\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)}) \le C_x\|\beta^{(1)}-\beta^{(2)}\|_2^2.$$

Thus, in terms of theoretical magnitude, $\|\beta^{(1)}-\beta^{(2)}\|_2^2$ and $(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})$ are the same and both can capture the change in the regression coefficient. However, compared to $\|\beta^{(1)}-\beta^{(2)}\|_2^2$, the quantity $(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})$ further incorporates the covariance structure Σ of the covariates and thus can better reflect the difference between two regression models $y=x^\top\beta^{(1)}+\varepsilon$ and $y=x^\top\beta^{(2)}+\varepsilon$. We therefore prefer $(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})$ for change-point estimation. We remark that $(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})$ is closely related to the explained variance in the regression literature, see for example Cai and Guo (2020).

For any $1\le m\le n-1$, define $\beta_m^{(1)}=\sum_{t=1}^m\beta_t^*/m$ and $\beta_m^{(2)}=\sum_{t=m+1}^n\beta_t^*/(n-m)$. Note that $\beta_m^{(1)}$ and $\beta_m^{(2)}$ are the unique minimizer of the population squared loss function $\mathbb{E}(\sum_{t=1}^m(y_t-x_t^\top\beta)^2)$ and $\mathbb{E}(\sum_{t=m+1}^n(y_t-x_t^\top\beta)^2)$, respectively. As a function of m, $(\beta_m^{(1)}-\beta_m^{(2)})^\top\Sigma(\beta_m^{(1)}-\beta_m^{(2)})$ achieves its maximum at the true change-point $m=\eta$ due to the fact that $\beta_m^{(1)}-\beta_m^{(2)}=\min(\frac{\eta}{m},\frac{n-\eta}{n-m})(\beta^{(1)}-\beta^{(2)})$. Thus, the sample estimate of $(\beta_m^{(1)}-\beta_m^{(2)})^\top\Sigma(\beta_m^{(1)}-\beta_m^{(2)})$ can be valuable for the detection and estimation of η.

Given a time point m, to estimate $(\beta_m^{(1)}-\beta_m^{(2)})^\top\Sigma(\beta_m^{(1)}-\beta_m^{(2)})$, a natural choice is the plug-in estimator. Specifically, via a penalized M-estimator, we can obtain $\widehat{\beta}_m^{(1)}$ from $\{x_t,y_t\}_{t=1}^m$ and $\widehat{\beta}_m^{(2)}$ from $\{x_t,y_t\}_{t=m+1}^n$. Combined with a covariance matrix estimator $\widehat{\Sigma}$, the plug-in estimator takes the form $(\widehat{\beta}_m^{(1)}-\widehat{\beta}_m^{(2)})^\top\widehat{\Sigma}(\widehat{\beta}_m^{(1)}-\widehat{\beta}_m^{(2)})$. This in some sense resembles the classical Wald-type statistics used in the change-point literature, see for example Richard A. Davis (1995) and Hušková et al. (2007). However, the plug-in estimator requires the estimation of Σ. Without strong structural assumptions on Σ, this is known to be a difficult task in high dimensions.

To bypass this difficulty, we slightly alter the estimation target and propose an alternative estimator via projection. Specifically, given a p-dimensional unit vector u with $\|u\|_2=1$, we define the *one-dimensional* variance-projected time series $\{z_t(u)\}_{t=1}^n$ as

$$z_t(u)=u^\top x_t y_t,\ \text{for } t=1,\cdots,n.$$

A key observation is that $\{z_t(u)\}_{t=1}^n$ has a single change-point in mean at time point η as long as $u^\top\Sigma(\beta^{(1)}-\beta^{(2)})\neq 0$. Importantly, if $u=(\beta^{(1)}-\beta^{(2)})/\|\beta^{(1)}-\beta^{(2)}\|_2$, we have that

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{m}\sum_{t=1}^m z_t(u)-\frac{1}{n-m}\sum_{t=m+1}^n z_t(u)\right)=u^\top\Sigma(\beta_m^{(1)}-\beta_m^{(2)})\\ =&\min(\frac{\eta}{m},\frac{n-\eta}{n-m})\frac{(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})}{\|\beta^{(1)}-\beta^{(2)}\|_2},\end{aligned}\tag{3}$$

which is proportional to the key quantity $(\beta^{(1)}-\beta^{(2)})^\top\Sigma(\beta^{(1)}-\beta^{(2)})$ and also achieves its maximum at $m=\eta$. Note that we further have

$$(3)\geq \min(\frac{\eta}{m},\frac{n-\eta}{n-m})c_x\|\beta^{(1)}-\beta^{(2)}\|_2.$$

Thus, the projection direction $u=(\beta^{(1)}-\beta^{(2)})/\|\beta^{(1)}-\beta^{(2)}\|_2$ is optimal in the sense that it preserves the *original* change size $\|\beta^{(1)}-\beta^{(2)}\|_2$ of the regression coefficients. Therefore, if the projection direction u is reasonably aligned with $\beta^{(1)}-\beta^{(2)}$, we can efficiently detect and estimate the change-point η by performing change-point estimation in mean on the univariate time series $\{z_t(u)\}_{t=1}^n$. To estimate the optimal projection direction, in Section 2.2, we propose a group Lasso based local screening (LGS) algorithm which provides an estimated $\widehat{\beta}^{(1)}-\widehat{\beta}^{(2)}$.

Note that the above projection framework loses its intuition and becomes less effective when $\{x_t,y_t\}_{t=1}^n$ contains multiple change-points. To tackle this issue, in Section 3, we further combine the projection idea with the wild binary segmentation (WBS) in Fryzlewicz (2014) and propose a multiple change-point detection algorithm named variance-projected WBS (VPWBS). Roughly speaking, the strategy is to perform the projection based change-point detection for $\{x_t,y_t\}_{t=1}^n$ on many randomly generated intervals $\{(a_m,b_m]\}_{m=1}^M$ with $1\leq a_m+1<b_m\leq n$, instead of focusing on the whole sample on $(0,n]$. The hope is that for a sufficiently large M, some random intervals will contain only one change-point and the projection based detection method will succeed.

An illustrative example: To facilitate understanding, we provide an illustrative example of how VPWBS works in practice. Specifically, we generate the data $\{x_t,y_t\}_{t=1}^n$ according to simulation setting (i) in Section 4.2, where we have $n=300,p=100$ and there are two change-points of $\{\beta_t^*\}_{t=1}^n$ at $\eta_1=100$ and $\eta_2=200$ with change size $\kappa=1.6\sqrt{40}$. For illustration, we focus on one of the randomly generated intervals $(104,290]$, which contains a single change-point at $\eta_2=200$. Figure 1(a)-(b) plots the subsample observations $\{x_t,y_t\}_{t=105}^{290}$, where no clear pattern of changes can be seen. Based on the above discussion, the optimal projection direction is

$$u^*=(\beta^*_{\eta_2}-\beta^*_{\eta_2+1})/\|\beta^*_{\eta_2}-\beta^*_{\eta_2+1}\|_2=(\underbrace{1,-1,1,-1\ldots,-1}_{10},\underbrace{0,\ldots,0}_{90})/\sqrt{10}.$$

Figure 1(c) plots the projected univariate time series $\{z_t(u^*)=u^{*\top}x_ty_t\}_{t=105}^{290}$ and its one-dimensional CUSUM statistics (see definition in (5) later). Note that there is a clear pattern of mean change for $\{z_t(u^*)\}_{t=105}^{290}$ around the true change-point $\eta_2=200$ and the CUSUM statistics is indeed maximized at $t=200$. Figure 1(d) plots the projected univariate time series $\{z_t(\widehat{u})=\widehat{u}^\top x_ty_t\}_{t=105}^{290}$ and its CUSUM statistics, where $\widehat{u}$ is estimated by the LGS algorithm in Section 2.2 using $\{x_t,y_t\}_{t=105}^{290}$ As can be seen, Figure 1(d) closely resembles Figure 1(c) and thus confirms the success of the proposed projection based change-point estimation framework.

Remark 1 *Given the estimated p-dimensional vector $\widehat{\beta}^{(1)}-\widehat{\beta}^{(2)}$, an intuitive and tempting alternative option is to detect change-points directly based on $\|\widehat{\beta}^{(1)}-\widehat{\beta}^{(2)}\|_2^2$. However, we remark that the extra projection step in our proposed framework helps further turn (and simplify) the p-dimensional problem into one-dimensional change-point detection in mean.*

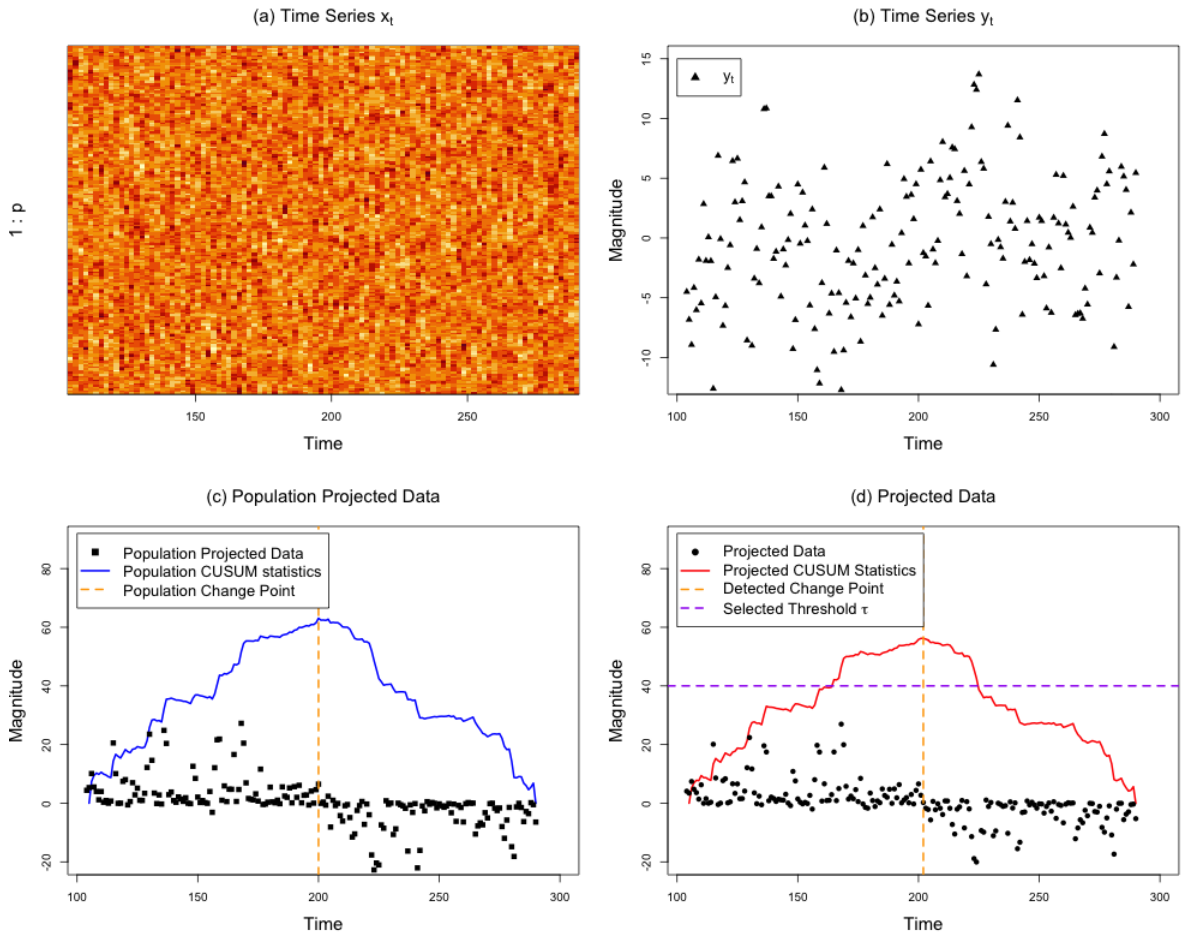

Figure 1: Plots of (a) $\{x_t\}_{t=105}^{290}$ (b) $\{y_t\}_{t=105}^{290}$ (c) $\{z_t(u^*)\}_{t=105}^{290}$ and its CUSUM statistics (d) $\{z_t(\widehat{u})\}_{t=105}^{290}$ and its CUSUM statistics.

This projection step acts as a refinement and is essential for the proposed method to achieve the minimax optimal rate (up to a log factor). In comparison, estimation error may accumulate along the p coordinates for $\|\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\|_2^2$*, making its theoretical analysis much more challenging. See Wang and Samworth (2018) for a similar observation in change-point detection for high-dimensional mean.*

2.2 Local Group Lasso Screening (LGS)

In this section, we propose a local group Lasso based screening (LGS) algorithm for estimating the optimal projection direction given the observed high-dimensional regression $\{x_t, y_t\}_{t=1}^n$.

Specifically, denote $1 \le s+1 < e \le n$ as the subsample index, LGS performs a variant of the group Lasso on the subsample $\{x_t, y_t\}_{t=s+1}^{e}$ and computes

$$(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\nu}) \leftarrow \underset{\substack{\nu \in [s'+1, e'-1], \\ \alpha_1, \alpha_2 \in \mathbb{R}^p}}{\arg\min} \left\{ \sum_{t=s+1}^{\nu} \left(y_t - x_t^\top \alpha_1\right)^2 + \sum_{t=\nu+1}^{e} (y_t - x_t^\top \alpha_2)^2 + \lambda \sum_{i=1}^{p} \sqrt{(\nu - s)(\alpha_{1,i})^2 + (e - \nu)(\alpha_{2,i})^2} \right\}, \tag{4}$$

where s' and e' serve as boundary trimming parameters with $s+1 \le s'+1 < e' \le e$, and λ is the tuning parameter for the group penalty. In the following, for convenience, we set $s' = s + \lfloor (e-s)/10 \rfloor$ and $e' = e - \lfloor (e-s)/10 \rfloor$, and summarize the detailed implementation of LGS in Algorithm 1.

Algorithm 1 Local group Lasso based Screening. LGS $(\{x_t, y_t\}_{t=1}^n, (s, e], \lambda)$.

INPUT: data $\{x_t, y_t\}_{t=1}^n$, subsample index $(s, e]$, tuning parameter λ.

Set $s' = s + \lfloor (e-s)/10 \rfloor$ and $e' = e - \lfloor (e-s)/10 \rfloor$ and compute

$$(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\nu}) \leftarrow \underset{\substack{\nu \in [s'+1, e'-1], \\ \alpha_1, \alpha_2 \in \mathbb{R}^p}}{\arg\min} \left\{ \sum_{t=s+1}^{\nu} \left(y_t - x_t^\top \alpha_1\right)^2 + \sum_{t=\nu+1}^{e} (y_t - x_t^\top \alpha_2)^2 + \lambda \sum_{i=1}^{p} \sqrt{(\nu - s)(\alpha_{1,i})^2 + (e - \nu)(\alpha_{2,i})^2} \right\}.$$

OUTPUT: $\{\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\nu}\}$.

The proposed LGS algorithm in (4) is different from the classical group Lasso or sparse group Lasso, as LGS explicitly targets the single change-point alternative in its formulation by incorporating two separate regression coefficients α_1 and α_2. Intuitively, when the subsample is sufficiently large w.r.t. the signal-to-noise ratio (SNR) and contains only one change-point η, the output $\widehat{\alpha}_1 - \widehat{\alpha}_2$ of LGS can estimate the optimal projection direction $\beta_\eta^* - \beta_{\eta+1}^*$ accurately. Assumption 2 formalizes this intuition and Theorem 2 further establishes the approximation quality of $\widehat{\alpha}_1 - \widehat{\alpha}_2$. Recall the definition in Section 1 that $\kappa = \min_{1 \le k \le K} \left\|\beta_{\eta_k+1}^* - \beta_{\eta_k}^*\right\|_2$ denotes the minimum change size and $\Delta = \min_{1 \le k \le K+1}(\eta_k - \eta_{k-1})$ denotes the minimum spacing between change-points.

Assumption 2

a. *There exists an absolute constant C_κ such that $\left\|\beta_{\eta_k+1}^* - \beta_{\eta_k}^*\right\|_2 \le C_\kappa$ for all $k = 1, \cdots, K$.*

b. *[SNR] We have $\Delta\kappa^2 \ge C_{snr}\mathfrak{s}\log(pn)$ where $C_{snr} = C_{snr}(n)$ is a diverging sequence as $n \to \infty$.*

Assumption 2**a** is a technical condition needed in the proof, which is also used in Lee et al. (2016) and Kaul et al. (2019). Assumption 2**b** implies that $\Delta \ge C_{snr}\kappa^{-2}\mathfrak{s}\log(pn) \ge C_{snr}C_\kappa^{-2}\mathfrak{s}\log(pn)$, which is the standard SNR condition in the Lasso literature. Note that we require $C_{snr} \to \infty$ as $n \to \infty$, but the divergence rate can be arbitrarily slow.

Theorem 2 *Suppose Assumptions 1-2 hold and* $\lambda = C_\lambda \sqrt{\log(pn)}$ *for some sufficiently large constant* C_λ. *Let* $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\nu})$ *be the output of LGS* $(\{x_t, y_t\}_{t=1}^n, (s, e], \lambda)$. *Suppose that* $(s, e]$ *satisfies* $e - s \geq \Delta/2$ *and contains exactly one change-point* η *such that*

$$\min\{\eta - s, e - \eta\} \geq \frac{e - s}{10}.$$

Then with probability at least $1 - 2(pn)^{-4}$, *it holds that*

$$\big\|(\widehat{\alpha}_1 - \widehat{\alpha}_2) - (\beta_\eta^* - \beta_{\eta+1}^*)\big\|_2 \leq \frac{c_x}{32 C_x} \|\beta_\eta^* - \beta_{\eta+1}^*\|_2.$$

Theorem 2 states that when the subsample $\{x_t, y_t\}_{t=s+1}^e$ contains only one change-point η and has sufficient number of observations, the proposed LGS algorithm can accurately estimate the optimal projection direction $\beta_\eta^* - \beta_{\eta+1}^*$, which serves as the foundation for the later theoretical analysis of the projection based framework.

3. Variance-Projected Wild Binary Segmentation

In this section, we formalize the discussion in Section 2 and present the variance-projected wild binary segmentation (VPWBS) algorithm for multiple change-point estimation in high-dimensional linear regression of Model (1).

Note that the LGS algorithm and the projection framework in Section 2 are discussed under the single change-point scenario. To further extend to multiple change-point estimation, VPWBS employs the mechanism of wild binary segmentation in Fryzlewicz (2014), where the essential idea is to perform single change-point estimation on M randomly generated intervals $\{(a_m, b_m]\}_{m=1}^M$ where $1 \leq a_m + 1 < b_m \leq n$. For a sufficiently large M, with high probability, for every true change-point in $\{\eta_k\}_{k=1}^K$, there exists at least one random interval $(a_m, b_m]$ such that η_k is the only change-point contained in $(a_m, b_m]$. More specifically, the good event

$$\mathcal{M} = \bigcap_{k=1}^K \{a_m \in \mathcal{S}_k, b_m \in \mathcal{E}_k, \text{ for some } m \in \{1, 2, \cdots, M\}\}$$

will hold with high probability, where $\mathcal{S}_k = (\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\mathcal{E}_k = (\eta_k + \Delta/2, \eta_k + 3\Delta/4]$, for $k = 1, 2, \cdots, K$. It is easy to see that if $a_m \in \mathcal{S}_k$ and $b_m \in \mathcal{E}_k$, we have $(a_m, b_m]$ only contains a single change-point η_k, as by definition the minimum spacing between two consecutive change-points is Δ. Theorem 3 later provides a rigorous bound for the probability that event $\mathcal{M}$ holds.

Another issue that needs to be addressed is that for the projection idea in (3) to be theoretically valid, the projection direction u is required to be independent from the observations $\{x_t, y_t\}_{t=1}^n$. To tackle this issue, we use sample splitting, a commonly used technique in high-dimensional statistics, see for example Wang and Samworth (2018), Wang et al. (2021) and Zou et al. (2020). Without loss of generality, we assume the original sample $\{x_t, y_t\}_{t=1}^{2n}$ is of length $2n$ (i.e. even) and we estimate the projection direction using LGS on the oddly-indexed observations $\{x_t^{(1)}, y_t^{(1)}\}_{t=1}^n$ and perform change-point estimation on the projected univariate series based on the evenly-indexed observations $\{x_t^{(2)}, y_t^{(2)}\}_{t=1}^n$, where

$$(x_t^{(1)}, y_t^{(1)}) = (x_{2t-1}, y_{2t-1}) \text{ and } (x_t^{(2)}, y_t^{(2)}) = (x_{2t}, y_{2t}) \text{ for } t = 1, \cdots, n.$$

To summarize, VPWBS implements the following two-stage procedure. In the first stage, given M random intervals $\{(a_m, b_m]\}_{m=1}^M$, the LGS in Algorithm 1 is implemented on $\{x_t^{(1)}, y_t^{(1)}\}_{t=1}^n$ for each of the M subsamples indexed by $(a_m, b_m]$ and returns M projection directions $\{u_m\}_{m=1}^M$. In the second stage, based on $\{u_m\}_{m=1}^M$ and $\{x_t^{(2)}, y_t^{(2)}\}_{t=1}^n$, we conduct mean change-point detection on the projected univariate time series $\{z_t(u_m) = u_m^\top x_t^{(2)} y_t^{(2)}\}_{t=a_m+1}^{b_m}$ for $m = 1, \cdots, M$ via the classical CUSUM statistics. For a univariate series $\{z_t(u_m)\}_{t=a_m+1}^{b_m}$ and $a_m \le s_m < \nu < e_m \le b_m$, the CUSUM statistics computed on $\{z_t(u_m)\}_{t=s_m+1}^{e_m}$ is defined as

$$\widetilde{Z}_\nu^{s_m,e_m}(u_m) = \sqrt{\frac{e_m - \nu}{(e_m - s_m)(\nu - s_m)}} \sum_{t=s_m+1}^{\nu} z_t(u_m) - \sqrt{\frac{\nu - s_m}{(e_m - s_m)(e_m - \nu)}} \sum_{t=\nu+1}^{e_m} z_t(u_m). \tag{5}$$

We summarize the detailed description of VPWBS in Algorithm 2. In total, there are four tuning parameters $(M, \lambda, \tau, \zeta)$ of the algorithm, where M is the number of random intervals, λ regulates the group Lasso penalty, τ is the threshold level of the maximum CUSUM statistics and 2ζ is the minimum length required for a subsample $(s_m, e_m]$ to be considered for change-point detection. Theorem 3 establishes the consistency and localization rate of VPWBS and gives the theoretical orders required for the tuning parameters $(M, \lambda, \tau, \zeta)$. We refer more details to the discussion after Theorem 3.

We remark that the sample splitting step of VPWBS in Algorithm 2 is mainly needed for establishing its theoretical validity in Theorem 3. In practice, we find that VPWBS is often more efficient without sample splitting. In other words, we can set both $\{x_t^{(1)}, y_t^{(1)}\}_{t=1}^n$ and $\{x_t^{(2)}, y_t^{(2)}\}_{t=1}^n$ as the original sample in Algorithm 2. See Wang and Samworth (2018) for similar phenomenon in high-dimensional mean change-point estimation. Recall from Section 1 that $\mathfrak{N} := \max_{1\le t\le n} \|\beta_t^*\|_2^2$ and by assumption 1, we have $\mathfrak{N} \le C_\beta^2 \mathfrak{s}$ and $\mathfrak{N} < \max_{1\le t\le n} \mathrm{Var}(y_t)/c_x$.

Theorem 3 *Suppose Assumptions 1-2 hold. Let $\{(a_m, b_m]\}_{m=1}^M$ be a collection of intervals whose end points are drawn independently and uniformly from $\{1, \ldots, n\}$ and that $\max_{1\le m\le M}(b_m - a_m) \le C_R\Delta$ for some absolute constant $C_R > 0$.*

Let $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ be the estimated change-points by VPWBS with data $\{x_t^{(1)}, y_t^{(1)}\}_{t=1}^n$, $\{x_t^{(2)}, y_t^{(2)}\}_{t=1}^n$, random intervals $\{(a_m, b_m]\}_{m=1}^M$, and tuning parameters $\lambda > 0$, $\tau > 0$, $\zeta > 0$, where

$$\lambda = C_\lambda \sqrt{\log(pn)}, \quad \tau = C_\tau \sqrt{(\mathfrak{N}+1)\log(n)}, \quad \text{and} \quad \zeta = C_\zeta(\mathfrak{N}+1)\log(n)$$

for sufficiently large constants C_λ, C_τ and C_ζ. Then there exists an absolute constant C such that

$$\begin{aligned}\mathbb{P}\Big\{\widehat{K} = K;\ |\eta_k - \widehat{\eta}_k| \le & \frac{C(\mathfrak{N}+1)\log(n)}{\kappa_k^2} \text{ for all } 1 \le k \le K\Big\} \\ \ge & 1 - n^{-2} - \exp\left(\log\big(\frac{n}{\Delta}\big) - \frac{M\Delta^2}{16n^2}\right).\end{aligned} \tag{6}$$

Theorem 3 establishes the consistency of VPWBS and further provides the localization error rate. Note that since $\mathfrak{N} \le C_\beta^2 \mathfrak{s}$ and $\mathfrak{N} < \max_{1\le t\le n} \mathrm{Var}(y_t)/c_x$, we have $\mathfrak{N} = O(1)$ if the

Algorithm 2 Variance-Projected Wild Binary Segmentation. VPWBS $(\{(a_m, b_m]\}_{m=1}^M, \lambda, \tau, \zeta)$

INPUT: 1st sample $\{x_t^{(1)}, y_t^{(1)}\}_{t=1}^n$, 2nd sample $\{x_t^{(2)}, y_t^{(2)}\}_{t=1}^n$, random intervals $\{(a_m, b_m]\}_{m=1}^M$, tuning parameters $\lambda > 0$, $\tau > 0$, $\zeta > 0$.

Initialize the set of estimated change-points as $\mathbf{S} = \varnothing$ and set $(s, e] = (0, n]$.

Stage 1: LGS and projection
  for $m = 1, \ldots, M$ **do**
    compute $\{\widehat{\alpha}_1^m, \widehat{\alpha}_2^m\} \leftarrow$ LGS $(\{x_t^{(1)}, y_t^{(1)}\}_{t=1}^n, (a_m, b_m], \lambda)$.
    set the projection direction: $u_m \leftarrow (\widehat{\alpha}_2^m - \widehat{\alpha}_1^m)/\|\widehat{\alpha}_2^m - \widehat{\alpha}_1^m\|_2$.
    set the projected univariate series: $z_t(u_m) \leftarrow u_m^\top x_t^{(2)} y_t^{(2)}$ for $t \in (a_m, b_m]$.
  end for

Stage 2: WBS$((s, e], \{(a_m, b_m]\}_{m=1}^M, \tau, \zeta\ )$
  for $m = 1, \ldots, M$ **do**
    $(s_m, e_m] \leftarrow (s, e] \cap (a_m, b_m]$
    if $e_m - s_m \geq 2\zeta$ **then**
      $D_m \leftarrow \arg\max_{s_m+\zeta \leq t \leq e_m-\zeta} |\widetilde{Z}_t^{s_m, e_m}(u_m)|$ ▷ Recall Equation (5)
      $A_m \leftarrow \max_{s_m+\zeta \leq t \leq e_m-\zeta} |\widetilde{Z}_t^{s_m, e_m}(u_m)|$
    else
      $A_m \leftarrow -1$
    end if
  end for
  $m^* \leftarrow \arg\max_{m=1,\ldots,M} A_m$
  if $A_{m^*} > \tau$ **then**
    add D_{m^*} to the set $\mathbf{S}$
    WBS $((s, D_{m^*}], \{(a_m, b_m]\}_{m=1}^M, \tau, \zeta)$
    WBS $((D_{m^*}, e], \{(a_m, b_m]\}_{m=1}^M, \tau, \zeta)$
  end if

OUTPUT: The set of estimated change-points $\mathbf{S}$.

sparsity level $\mathfrak{s}$ is a constant or the maximum variance of the response y_t is upper bounded, which is a rather mild condition. In such case, the localization error bound in Theorem 3 further implies

$$\epsilon = \max_{1 \leq k \leq K} \frac{|\eta_k - \widehat{\eta}_k|}{n} \leq \max_{1 \leq k \leq K} C \frac{\log(n)}{n\kappa_k^2},$$

where ϵ is the localization error defined in (2) and $\kappa_k = \|\beta_{\eta_k} - \beta_{\eta_k+1}\|_2$ is the change size at η_k. Up to a log factor, this matches the well-known minimax optimal rate for change-point estimation, see Wang et al. (2018) and references therein.

Theorem 3 requires $\max_{1 \leq m \leq M}(b_m - a_m) \leq C_R\Delta$, which essentially implies that the random intervals cannot contain too many change-points. See similar assumptions in Kaul et al. (2019). Note that if $\Delta \asymp n$, the assumption becomes minimal as we can simply set $C_R\Delta = n$. We remark that Theorem 3 still holds without the assumption $\max_{1 \leq m \leq M}(b_m - a_m) \leq C_R\Delta$, however, the localization error rate in (6) will be inflated to $(n/\Delta)^2 \cdot (C(\mathfrak{N} + 1)\log(n)/\kappa_k^2)$

by a factor $(n/\Delta)^2$. This is a phenomenon commonly seen in the high-dimensional change-point literature, see for example Wang and Samworth (2018), Wang et al. (2021) and Li et al. (2021).

Discussion on tuning parameters $(M, \lambda, \tau, \zeta)$: By the probability bound (6) in Theorem 3, for the consistency of VPWBS, it is necessary to choose the number of random intervals $M \succeq n^2\log(n)/\Delta^2$. In particular, suppose that $\Delta \asymp n$ (i.e. there are finite number of change-points), it suffices to choose $M = (\log(n))^2$. The tuning parameter λ is needed in the LGS algorithm and assumes the standard order $C_\lambda\sqrt{\log(pn)}$ of the group Lasso penalty in the literature (see also Theorem 2). The parameter τ is commonly seen in the change-point literature, and is needed to threshold the maximum CUSUM statistics and controls false positive detection. To derive τ, we need to study the order of the maximum CUSUM statistics under the no change-point scenario. As for the tuning parameter ζ, intuitively, for small subsamples, the estimation error of LGS and the CUSUM statistics become difficult to control. The parameter ζ is designed to handle such scenario and regulates the minimum length required for a subsample $(s, e]$ to be considered for change-point detection. See similar tuning parameters in Leonardi and Bühlmann (2016) and Kaul et al. (2019). Note that simple algebra gives that $\Delta \succeq \zeta$. In practice, it suffices to set $\zeta = \log(n)$.

In general, VPWBS is highly robust to the choices of M and ζ, and the key tuning parameters affecting the performance of VPWBS are λ and τ. In Section 4, we propose a cross-validation procedure to select λ and τ in a fully data-driven fashion.

4. Simulations

In this section, we conduct extensive numerical experiments to examine the performance of VPWBS under various simulation settings and further compare it with two other state-of-the-art methods in the literature, specifically, EBSA in Leonardi and Bühlmann (2016) and SGL in Zhang et al. (2015b). Implementations of the numerical experiments can be found at the GitHub link here. We discuss the implementation details such as settings for each algorithm and estimation accuracy metrics in Section 4.1 and present the simulation results in Section 4.2.

4.1 Implementation details

Given estimated change-point estimators $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$, we measure the estimation accuracy via the scaled Hausdorff distance, a popular metric used in the change-point literature. Specifically, denoting the true change-points as $\{\eta_k\}_{k=1}^{K}$, the scaled Hausdorff distance is defined as

$$\mathcal{D}(\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}, \{\eta_k\}_{k=1}^{K}) = d(\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}, \{\eta_k\}_{k=1}^{K})/n,$$

where $d(\cdot, \cdot)$ denotes the Hausdorff distance between two compacts sets A, B in $\mathbb{R}$, given by

$$d(A, B) = \max\left\{\max_{a\in A}\min_{b\in B}|a-b|, \max_{b\in B}\min_{a\in A}|a-b|\right\}.$$

Note that $\mathcal{D}(\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}, \{\eta_k\}_{k=1}^{K}) \leq 1$ when both $K, \widehat{K} \geq 1$. Therefore, following the convention in the change-point literature, we set $\mathcal{D}(\varnothing, \{\eta_k\}_{k=1}^{K}) = 1$.

Implementation of VPWBS: As discussed in Section 3, there are four tuning parameters $(M, \lambda, \tau, \zeta)$ in VPWBS. Throughout the simulation section, we set $M = 40$ and set $\zeta = 5$, which roughly corresponds to $M = (\log(n))^2$ and $\zeta = \log(n)$ across all simulation settings in Section 4.2. We remark that the performance of VPWBS is robust to the choices of (M, ζ), and the key tuning parameters are (λ, τ).

In the following, we provide a sample splitting based cross-validation procedure that selects (λ, τ) in a fully data-driven fashion. Specifically, given the original sample $\{x_t, y_t\}_{t=1}^n$, we set the training data to be the oddly-indexed observations $\{x_{2t-1}, y_{2t-1}\}_{t=1}^{n/2}$ and the test data to be the evenly-indexed observations $\{x_{2t}, y_{2t}\}_{t=1}^{n/2}$, where we assume, without loss of generality, n is even. Note that the training data and test data share the same number and locations of change-points (up to one time point difference).

Denote the candidate sets of λ, τ as $\Lambda, \mathcal{T} \subset \mathbb{R}^+$. For each pair of $(\lambda, \tau) \in \Lambda \times \mathcal{T}$, using the training data, we compute the estimated change-points via VPWBS and further estimate the (piecewise constant) regression coefficients $\{\widehat{\beta}_t\}_{t=1}^{n/2}$ *conditional* on the estimated change-points as in Model (1). We then compute the prediction error of $\{\widehat{\beta}_t\}_{t=1}^{n/2}$ using the test data via

$$e_t = y_{2t} - x_{2t}^\top \widehat{\beta}_t, \quad t = 1, 2, \cdots, n/2.$$

The tuning parameters (λ, τ) are then selected as the pair of $(\lambda, \tau) \in \Lambda \times \mathcal{T}$ that achieves the minimum squared prediction error $\sum_{t=1}^{n/2} e_t^2$ on the test data. Note that in the cross-validation procedure, the random intervals $\{(a_m, b_m]\}_{m=1}^M$ and the minimum length ζ are kept the same across all pairs of (λ, τ). For all simulation experiments in Section 4.2, we set $\Lambda = \{0.5, 1, 1.5, 2\}$ and $\mathcal{T} = \{1, 4, 7, 10, \cdots, 49\}$.

Implementation of competing methods: The EBSA algorithm is proposed in Leonardi and Bühlmann (2016), which performs change-point detection for high-dimensional regression via a model selection point of view. Specifically, a dynamic programming algorithm is proposed to directly estimate the unknown change-points by minimizing an l_0-penalized goodness of fit function. In contrast, VPWBS utilizes a group Lasso based local screening (LGS) algorithm for estimating an optimal projection direction and uses CUSUM statistics for change-point estimation, where the LGS is an l_1-penalized M-estimator. For computational efficiency, a binary segmentation based algorithm is further proposed in Leonardi and Bühlmann (2016) to find an approximate minimizer of the penalized function with strong theoretical guarantees. We choose the tuning parameters of EBSA using its default settings as specified in Leonardi and Bühlmann (2016). We note that EBSA gives slightly worse performance when its tuning parameters are selected via the sample splitting based cross-validation.

The sparse group Lasso (SGL) is first introduced by Simon et al. (2013) and is later used by Zhang et al. (2015b) for change-point detection in the high-dimensional regression setting. See also Harchaoui and Lévy-Leduc (2007, 2010), Bleakley and Vert (2011) and references therein for earlier work along this line of research, where the classical fused Lasso is used for change-point detection in mean for low-dimensional time series.

Given $\{x_t, y_t\}_{t=1}^n$, SGL computes

$$\{\widehat{\beta}_t\}_{t=1}^n = \arg \min_{(\beta_1, \ldots, \beta_n)} \sum_{t=1}^n (y_t - X_t \beta_t)_2^2 + \lambda \sum_{t=1}^{n-1} \|\beta_{t+1} - \beta_t\|_2 + \gamma \sum_{t=1}^{n-1} \|\beta_{t+1} - \beta_t\|_1, \qquad (7)$$

which can be seen as a variant of the classical fused Lasso with an extra group sparsity penalty. Note that SGL is a global method as it estimates $\{\widehat{\beta}_t\}_{t=1}^n$ (and thus multiple change-points) based on the entire sample. In comparison, the local group Lasso based screening (LGS) algorithm in Stage 1 of the proposed VPWBS is a local method and is designed to directly target single change-points.

Define the function $f : \{1, 2, \cdots, n-1\} \to \mathbb{R}$ where $f(t) := \|\widehat{\beta}_t - \widehat{\beta}_{t-1}\|_2$. Note that $\{\widehat{\beta}_t\}_{t=1}^n$ estimated by SGL in (7) may not directly lead to accurate change-point estimation as $\sum_{t=1}^{n-1} \mathbb{I}(f(t) > 0)$ is generally a large number and leads to uncontrollable false positives, where $\mathbb{I}$ denotes the indicator function. In practice, the SGL estimator $\{\widehat{\beta}_t\}_{t=1}^n$ typically exhibits the so-called staircase pattern, a pattern commonly seen in fused Lasso based estimation (e.g. Rojas and Wahlberg, 2014; Owrang et al., 2017), where $\{\widehat{\beta}_t\}_{t=1}^n$ contains large-scale changes accompanied by many small-scale jumps. See Figure 5 of the supplementary material for an illustration of such phenomenon. To avoid false positive estimation, given knowledge of the true number of change-points K, a common practice in the literature, see e.g. Harchaoui and Lévy-Leduc (2010), is to estimate the change-points as the locations where the function f achieves its K largest values.

However, in practice, K is typically unknown. Thus, to further improve the applicability of SGL, in our experiments we consider a variant of the SGL algorithm combined with wild binary segmentation in Fryzlewicz (2014), which we refer to as Wild Binary Segmentation via SGL (WBSSGL). Specifically, WBSSGL further post-processes the estimated $\{\widehat{\beta}_t\}_{t=1}^n$ by SGL on M random intervals $\{(a_m, b_m]\}_{m=1}^M$. For a subsample $\{\widehat{\beta}_t\}_{t=a_m+1}^{b_m}$ and $a_m \leq s_m < \nu < e_m \leq b_m$, WBSSGL computes the ($p$-dimensional) subsample CUSUM statistics for $\{\widehat{\beta}_t\}_{t=s_m+1}^{e_m}$ defined as

$$\widehat{\mathcal{B}}_\nu^{s_m,e_m} = \sqrt{\frac{(e_m-\nu)}{(e_m-s_m)(\nu-s_m)}} \sum_{t=s_m+1}^{\nu} \widehat{\beta}_t - \sqrt{\frac{(\nu-s_m)}{(e_m-s_m)(e_m-\nu)}} \sum_{t=\nu+1}^{e_m} \widehat{\beta}_t, \tag{8}$$

and further compares it with a suitable threshold. The detailed implementation of WBSSGL is given in Algorithm 3. For all simulation experiments in Section 4.2, numerical results indicate that WBSSGL outperforms the original SGL algorithm by a wide margin. Thus, in the following we only present the results for WBSSGL.

There are five tuning parameters $(M, \lambda, \gamma, \tau, \zeta)$ of WBSSGL, which are selected in the same way as VPWBS. Specifically, we set the random intervals $\{(a_m, b_m]\}_{m=1}^M$ and the minimum length ζ of WBSSGL to be the same as VPWBS. The key tuning parameters (λ, γ, τ) of WBSSGL are selected using the same cross-validation procedure as the one implemented for VPWBS. For each combination of (λ, γ), we solve the original SGL in (7) via the R package `SGL`.

4.2 Simulation results

In this section, we conduct extensive numerical experiments to examine the performance of VPWBS, EBSA and WBSSGL in terms of estimation accuracy and computational cost. We design a wide range of simulation settings by varying change size κ, spacing between change-points Δ, number of change-points K, sparsity level $\mathfrak{s}$, sample size n and dimension p. The variance of noise σ_ε^2 is set at 1 for all settings. For each simulation setting, we repeat

Algorithm 3 Wild Binary Segmentation via SGL. WBSSGL($\{(a_m, b_m]\}_{m=1}^M, \lambda, \gamma, \tau, \zeta$)

INPUT: data $\{x_t, y_t\}_{t=1}^n$, random intervals $\{(a_m, b_m]\}_{m=1}^M$,
tuning parameters $\lambda > 0, \gamma > 0, \tau > 0, \zeta > 0$.

Initialize the set of estimated change-points as $\mathbf{S} = \varnothing$ and set $(s, e] = (0, n]$.

Stage 1: SGL
Compute $\{\widehat{\beta}_t\}_{t=1}^n$ via (7) with tuning parameters λ, γ.

Stage 2: WBS($(s, e], \{(a_m, b_m]\}_{m=1}^M, \tau, \zeta$)
for $m = 1, \ldots, M$ **do**
$(s_m, e_m] \leftarrow (s, e] \cap (a_m, b_m]$
if $e_m - s_m \geq 2\zeta$ **then**
$D_m \leftarrow \arg\max_{s_m+\zeta \leq t \leq e_m-\zeta} \|\widehat{\mathcal{B}}_t^{s_m,e_m}\|_2$ ▷ Recall Equation (8)
$A_m \leftarrow \max_{s_m+\zeta \leq t \leq e_m-\zeta} \|\widehat{\mathcal{B}}_t^{s_m,e_m}\|_2$
else
$A_m \leftarrow -1$
end if
end for
$m^* \leftarrow \arg\max_{m=1,\ldots,M} A_m$
if $A_{m^*} > \tau$ **then**
add D_{m^*} to the set $\mathbf{S}$
WBS $((s, D_{m^*}], \{(a_m, b_m]\}_{m=1}^M, \tau, \zeta)$
WBS $((D_{m^*}, e], \{(a_m, b_m]\}_{m=1}^M, \tau, \zeta)$
end if

OUTPUT: The set of estimated change-points $\mathbf{S}$.

the experiments 100 times. The detailed simulation setting is as follows. We further plot typical realizations of $\{y_t\}_{t=1}^n$ for each setting in Figure 2, where it can be seen clearly that information contained in $\{y_t\}_{t=1}^n$ is not sufficient for change-point estimation.

Setting (i): two change-points with varying change size κ. In this setting, we fix $n = 300$, $p = 100$, $K = 2$ and set the covariance matrix Σ of x_t to be the Toeplitz matrix with $\Sigma_{i,j} = 0.6^{|i-j|}$ for $i, j = 1, \cdots, p$. The two change-points occur at $\eta_1 = n/3 = 100$ and $\eta_2 = 2n/3 = 200$. The regression coefficients $\{\beta_t^*\}_{t=1}^n$ take the form

$$\beta_t^* = \begin{cases} \kappa \cdot \alpha/\sqrt{40}, & \text{for } 1 \leq t \leq n/3, \\ -\kappa \cdot \alpha/\sqrt{40}, & \text{for } n/3+1 \leq t \leq 2n/3, \\ \kappa \cdot \alpha/\sqrt{40}, & \text{for } 2n/3+1 \leq t \leq n, \end{cases}$$

where $\alpha = (\underbrace{1, -1, 1, -1 \ldots, -1}_{10}, \underbrace{0, \ldots, 0}_{p-10})$ and we vary $\kappa \in \sqrt{40} \cdot \{1, 1.2, 1.4, 1.6\}$. Simple calculation shows that the change size of β_t^* at both change-points equals κ.

Setting (ii): three change-points with varying sample size n. In this setting, we vary $n \in \{480, 560, 640, 720, 800\}$, fix $p = 100$, $K = 3$ and set the covariance matrix Σ of x_t to be the identity matrix I_p. The three change-points occur evenly at $\eta_i = in/4$ for

$i = 1, 2, 3$. The regression coefficients $\{\beta_t^*\}_{t=1}^n$ take the form

$$\beta_t^* = \begin{cases} 2/5\alpha, & \text{for } 1 \le t \le n/4, \\ -2/5\alpha, & \text{for } n/4+1 \le t \le n/2, \\ 2/5\alpha, & \text{for } n/2+1 \le t \le 3n/4, \\ -2/5\alpha, & \text{for } 3n/4+1 \le t \le n, \end{cases}$$

where $\alpha = (1, -1, 1, -1, 0, \cdots, 0)$. The change size of β_t^* at each change-point equals $8/5$.

Setting (iii): two change-points with varying p and varying support of β_t. In this setting, we fix $n = 320$, $K = 2$, vary $p \in \{90, 100, 110, 120\}$ and set the covariance matrix Σ of x_t to be I_p. The two change-points occur unevenly at $\eta_1 = 120$ and $\eta_2 = 220$. The regression coefficients $\{\beta_t^*\}_{t=1}^n$ take the form

$$\beta_t^* = \begin{cases} 2/3 \cdot (\underbrace{1, \ldots, 1}_{8}, 0, \ldots, 0), & \text{for } 1 \le t \le 120, \\ 2/3 \cdot (\underbrace{0, \ldots, 0}_{8}, \underbrace{1, \ldots, 1}_{8}, 0, \ldots, 0), & \text{for } 121 \le t \le 220, \\ 2/3 \cdot (\underbrace{0, \ldots, 0}_{16}, \underbrace{1, \ldots, 1}_{8}, 0, \ldots, 0), & \text{for } 221 \le t \le 320. \end{cases}$$

Simple calculation shows the change size of β_t^* at each change-point equals $8/3$.

Setting (iv): two change-points with uneven spacing and varying support size s. In this setting, we fix $n = 520$, $K = 2$, $p = 100$ and set the covariance matrix of x_t to be $\Sigma = I_p$. The two change-points occur unevenly at $\eta_1 = 160$ and $\eta_2 = 360$. The regression coefficients $\{\beta_t^*\}_{t=1}^n$ take the form

$$\beta_t^* = \begin{cases} \sqrt{2/s} \cdot (\underbrace{1, \ldots, 1}_{s/2}, \underbrace{3, \ldots, 3}_{s/2}, 0, \ldots, 0), & \text{for } 1 \le t \le 160, \\ \sqrt{2/s} \cdot (\underbrace{2, \ldots, 2}_{s/2}, \underbrace{1, \ldots, 1}_{s/2}, 0, \ldots, 0), & \text{for } 161 \le t \le 360, \\ \sqrt{2/s} \cdot (\underbrace{1, \ldots, 1}_{s/2}, \underbrace{3, \ldots, 3}_{s/2}, 0, \ldots, 0), & \text{for } 361 \le t \le 520, \end{cases}$$

and we vary $s \in \{16, 20, 24, 28\}$. Simple calculation shows that the change sizes of β_t^* at both change-points equal $\kappa = \sqrt{5}$.

Estimation accuracy: Table 2 reports the scaled Hausdorff distance (averaged over 100 repetitions) achieved by VPWBS, EBSA and WBSSGL across all simulation settings. For better visualization, Figure 3 further provides the bar plots based on the results reported in Table 2. First, as expected, the performance of all three algorithms improve with larger sample size n (setting (ii)) and with larger change size κ (setting (i)), and worsen with higher dimension p (setting (iii)) and with higher sparsity (setting (iv)).

Overall, VPWBS offers robust and competitive performance for change-point estimation across all simulation settings, and consistently outperforms its competitors under the low

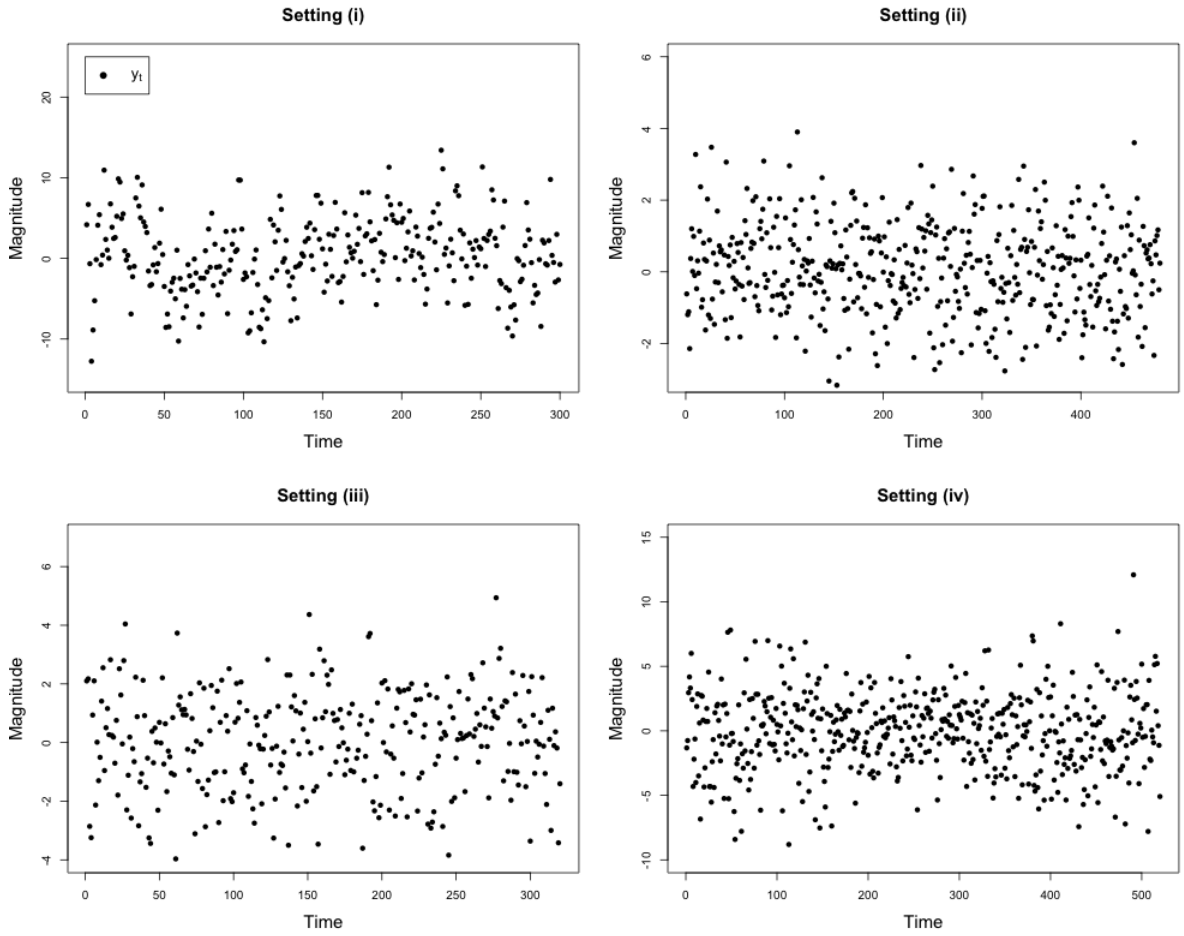

Figure 2: Typical realizations of $\{y_t\}_{t=1}^n$ for Setting (i) with $\kappa = \sqrt{40}$, Setting (ii) with $n = 480$, Setting (iii) with $p = 90$ and Setting (iv) with $\kappa = 0.35\sqrt{40}$.

SNR scenario where the sample size n or the change size κ is small. Compared to WBSSGL, which conducts change-point estimation directly on the estimated $\{\widehat{\beta}_t\}_{t=1}^n$ by a penalized M-estimator (i.e. SGL), VPWBS in general gives more favorable performance, which could be seen as numerical evidence confirming the importance of the projection step in VPWBS. It is worth noting that when the SNR is large, EBSA becomes highly competitive. We conjecture that EBSA is also minimax optimal when the signal strength $\Delta\kappa^2$ is sufficiently large, though rigorous proof of such result seems challenging.

Computational cost: It is straightforward to derive that the computational cost of VPWBS is $O(Mn \cdot \text{GroupLasso}(n, p))$, where M is the number of random intervals used and recall that $\text{GroupLasso}(n, p)$ denotes the computational cost of the group Lasso for a p-dimensional regression with n samples. Since we set $M = (\log(n))^2$, the computational complexity of VPWBS equals $O(n(\log(n))^2 \cdot \text{GroupLasso}(n, p))$. On the other hand, refer-

		VPWBS	EBSA	WBSSGL
Setting (i)	$\kappa = \sqrt{40}$	**0.011** (0.015)	0.090 (0.055)	0.151 (0.094)
	$\kappa = 1.2\sqrt{40}$	**0.009** (0.017)	0.060 (0.052)	0.118 (0.069)
	$\kappa = 1.4\sqrt{40}$	**0.010** (0.024)	0.033 (0.044)	0.101 (0.058)
	$\kappa = 1.6\sqrt{40}$	**0.009** (0.016)	0.025 (0.039)	0.098 (0.061)
Setting (ii)	$n = 480$	**0.062** (0.080)	0.128 (0.177)	0.101 (0.052)
	$n = 560$	**0.044** (0.042)	0.064 (0.112)	0.098 (0.053)
	$n = 640$	**0.034** (0.060)	**0.034** (0.084)	0.094 (0.055)
	$n = 720$	0.026 (0.054)	**0.015** (0.023)	0.093 (0.075)
	$n = 800$	0.022 (0.046)	**0.009** (0.010)	0.091 (0.072)
Setting (iii)	$p = 80$	**0.025** (0.047)	0.052 (0.047)	0.103 (0.065)
	$p = 90$	**0.039** (0.068)	0.068 (0.051)	0.109 (0.068)
	$p = 100$	**0.033** (0.056)	0.056 (0.050)	0.120 (0.072)
	$p = 110$	**0.041** (0.063)	0.055 (0.046)	0.125 (0.072)
	$p = 120$	**0.049** (0.078)	0.061 (0.048)	0.140 (0.079)
Setting (iv)	$s = 16$	**0.027** (0.072)	0.055 (0.023)	0.116 (0.059)
	$s = 20$	**0.024** (0.060)	0.059 (0.024)	0.141 (0.033)
	$s = 24$	**0.043** (0.089)	0.057 (0.020)	0.141 (0.032)
	$s = 28$	**0.057** (0.106)	0.063 (0.022)	0.144 (0.023)

Table 2: Scaled Hausdorff distance for VPWBS, EBSA (Leonardi and Bühlmann (2016)) and WBSSGL (Zhang et al. (2015b)). For each cell, the experiment is repeated 100 times. The numbers in the brackets indicate the sample standard errors of the scaled Hausdorff distance. Each highlighted number indicates the best performance in the corresponding setting.

ring to Table 1, the complexity of EBSA and WBSSGL are $O(n\log(n)\cdot \text{Lasso}(n,p))$ and $O(\text{Lasso}(n,np))$ respectively.

Note that the computational cost of solving Lasso and group Lasso for a p-dimensional linear regression with n observations is both $O(np^2)$, see for example Efron et al. (2004) and Wright et al. (2009). Thus, it is easy to see that in terms of computational efficiency, ESBA is the best, VPWBS comes second, and WBSSGL comes last. In practice, popular

R packages typically implement gradient or coordinate descent to obtain an approximate solution of Lasso and group Lasso and the computation can be much faster than $O(np^2)$.

We conduct further numerical experiments to exam the computational performance of each algorithm in practice. Specifically, given (n, p), we generate the regression coefficients $\{\beta_t^*\}_{t=1}^n$ and observations $\{x_t, y_t\}_{t=1}^n$ using Setting (i) with $\kappa = 1.6\sqrt{40}$. In the first set of experiments, we fix $n = 450$ and vary $p \in \{80, 100, 120, 140, 160, 180, 200, 220\}$; in the second set of experiments, we fix $p = 100$ and vary $n \in \{240, 300, 360, 420, 480, 540, 600, 660\}$. For each simulation setting, we repeat the experiments 100 times and report the average execution time of VPWBS, EBSA and WBSSGL in Figure 4.

As can be seen in Figure 4, the computational costs of VPWBS and EBSA increase linearly with both the dimension p and the sample size n. On the other hand, while the computational cost of WBSSGL grows linearly with p, it does not scale well with n. This is not surprising, as the SGL approach (7) is essentially solving a Lasso with n samples and np covariates.

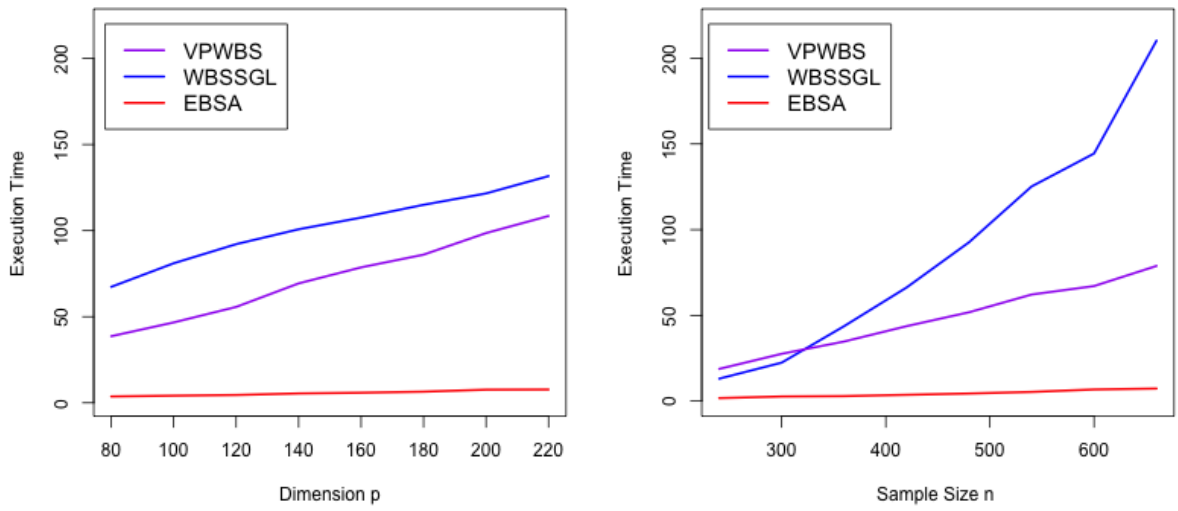

Figure 4: Average execution time of VPWBS, EBSA and WBSSGL across different simple sizes n and dimensions p.

5. Discussion

In this paper, we study the problem of multiple change-point estimation in high-dimensional linear regression model. We propose a novel projection-based algorithm, VPWBS, which performs change-point detection from a dimension reduction angle. Based on an estimated (optimal) projection direction, VPWBS transforms the original (difficult) problem of change-point detection in p-dimensional regression to a simpler problem of change-point detection in mean of a one-dimensional time series. VPWBS is shown to achieve the minimax

optimal localization rate $O_p(1/n)$ up to a log factor, a significant improvement from the best rate $O_p(1/\sqrt{n})$ known in the existing literature. In addition, VPWBS is computationally efficient with a complexity of $O(n(\log(n))^2 \cdot \text{GroupLasso}(n, p))$. Extensive numerical experiments are conducted to demonstrate the robust and favorable performance of VPWBS over two state-of-the-art algorithms in a wide range of simulation settings.

Besides the high-dimensional regression problem, we believe the projection based change-point estimation framework can be useful under other important contexts as well, such as change in covariance matrices and tensors. The key is to design an algorithm that utilizes the structure of the specific problem (such as sparsity or low rank) and provides a provably accurate estimation of the (optimal) projection direction.

Acknowledgments

We would like to thank the editor, Dr. Zaid Harchaoui, as well as the three anonymous reviewers for their thoughtful assessment and constructive comments which helped us to improve the quality and the presentation of our paper. The work of RW was supported in part by AFOSR FA9550-18-1-0166, NSF DMS-1925101, NSF OAC-1934637, and DOE DE-AC02-06CH11357.

References

Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78, 1998.

Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *Technical report HAL-00602121*, 2011.

T. Tony Cai and Zijian Guo. Semi-supervised inference for explained variance in high-dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B*, 82(2):391–419, 2020.

Bruno M Castro, Renan B Lemes, Jonatas Cesar, Tábita Hünemeier, and Florencia Leonardi. A model selection approach for multiple sequence segmentation and dimensionality reduction. *Journal of Multivariate Analysis*, 167:319–330, 2018.

Jie Chen and Arjun K Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical association*, 92(438):739–747, 1997.

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

Haeran Cho et al. Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, 10(2):2000–2038, 2016.

Ivor Cribben and Yi Yu. Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):607–627, 2017.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Farida Enikeeva and Zaid Harchaoui. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079, 2019.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

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

Zaid Harchaoui and Céline Lévy-Leduc. Catching change-points with lasso. *Advances inNeural Information Processing Systems 20*, 2007.

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

Marie Hušková, Zuzana Prášková, and Josef Steinebach. On the detection of changes in autoregressive time series i. asymptotics. *Journal of Statistical Planning and Inference*, 137(4):1243–1259, 2007.

Feiyu Jiang, Zifeng Zhao, and Xiaofeng Shao. Modelling the covid-19 infection trajectory: A piecewise linear quantile trend model. *Journal of the Royal Statistical Society: Series B*, 2021.

Abhishek Kaul, Venkata K Jandhyala, and Stergios B Fotopoulos. An efficient two step algorithm for high dimensional change point regression models without grid search. *Journal of Machine Learning Research*, 80(20):1–40, 2019.

Sokbae Lee, Myung Hwan Seo, and Youngki Shin. 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, 2016.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Oracle estimation of a change point in high-dimensional quantile regression. *Journal of the American Statistical Association*, 113(523):1184–1194, 2018.

Florencia Leonardi and Peter Bühlmann. Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704*, 2016.

Yu-Ning Li, Degui Li, and Piotr Fryzlewicz. Detection of multiple structural breaks in largecovariance matrices. *Working paper*, 2021.

Arash Owrang, Mohammadreza Malek-Mohammadi, Alexandre Proutiere, and Magnus Jansson. Consistent change point detection for piecewise constant signals with normalized fused lasso. *IEEE Signal Processing Letters*, 24(6):799 – 803, 2017.

Florian Pein, Hannes Sieling, and Axel Munk. Heterogeneous change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1207–1227, 2017.

Tao Peng, Christopher Leckie, and Kotagiri Ramamohanarao. Proactively detecting distributed denial of service attacks using source ip address monitoring. In *International conference on research in networking*, pages 771–782. Springer, 2004.

Zhongjun Qu and Pierre Perron. Estimating and testing structural changes in multivariate regressions. *Econometrica*, 75(2):459–502, 2007.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Yi-Ching Yao Richard A. Davis, Dawei Huang. Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, 23(1):282–304, 1995.

Cristian R. Rojas and Bo Wahlberg. On change point detection using the fused lasso method. *arXiv:1401.5408*, 2014.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*, 2017.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Univariate mean change point detection: Penalization, cusum and optimality. *arXiv preprint arXiv:1810.09498*, 2018.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, 49(1):203–232, 2021.

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

Stephen J. Wright, Robert D. Nowak, and MÁrio A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479 – 2493, 2009.

Liyan Xie, Yao Xie, and George V Moustakides. Sequential subspace change point detection. *Sequential Analysis*, 39(3):307–335, 2020.

Chun Yip Yau and Zifeng Zhao. Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B*, 78(4): 895–916, 2016.

Bingwen Zhang, Jun Geng, and Lifeng Lai. Change-point estimation in high dimensional linear regression models via sparse group lasso. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 815–821. IEEE, 2015a.

Bingwen Zhang, Jun Geng, and Lifeng Lai. Multiple change-points estimation in linear regression models via sparse group lasso. *IEEE Trans. Signal Processing*, 63(9):2209–2224, 2015b.

Zifeng Zhao and Chun Yip Yau. Alternating pruned dynamic programming for multiple epidemic change-point estimation. *Journal of Computational and Graphical Statistics*, 30 (3):808–821, 2021.

Zifeng Zhao, Li Chen, and Lizhen Lin. Change-point detection in dynamic networks via graphon estimation. *arXiv:1908.01823*, 2019.

Changliang Zou, Guanghui Wang, and Runze Li. Consistent selection of the number of change-points via sample-splitting. *The Annals of Statistics*, 48(1):413–439, 2020.

Appendix A. Proofs Related to Theorem 2

In this section, we provide all the technical details for the proof of Theorem 2

Proof [Proof of Theorem 2] Denote

$$\kappa' = \|\beta^*_\eta - \beta^*_{\eta+1}\|_2.$$

Since $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\nu})$ be the output of LGS $(\{y_i, x_i\}_{i=1}^n, (s, e], \lambda)$, it holds that

$$(\widehat{\alpha}_1, \widehat{\alpha}_2) = \underset{\alpha_1,\alpha_2\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=s+1}^{\widehat{\nu}} \left(y_i - x_i^\top \alpha_1\right)^2 + \sum_{i=\widehat{\nu}+1}^{e} (y_i - x_i^\top \alpha_2)^2 + \lambda \sum_{j=1}^{p} \sqrt{(\widehat{\nu} - s)(\alpha_{1,j})^2 + (e - \widehat{\nu})(\alpha_{2,j})^2} \right\},$$

and that by Theorem 4,

$$|\eta - \widehat{\nu}| \le C \frac{\mathfrak{s} \log(pn)}{(\kappa')^2}.$$

Without loss of generality, assume that $s < \eta < \widehat{\nu} \le e$. Then by Theorem 5,

$$\left\|\widehat{\alpha}_1 - \beta^*_{(s,\widehat{\nu}]}\right\|_2^2 \le C \frac{\mathfrak{s} \log(pn)}{\Delta} \le \frac{C\mathfrak{s} \log(pn)}{C_{snr} \mathfrak{s} \log(pn) \kappa^{-2}} \le \frac{C\kappa^2}{C_{snr}} \le \frac{C(\kappa')^2}{C_{snr}},$$

For sufficient large constant C_{snr}, it holds that

$$\left\|\widehat{\alpha}_1 - \beta^*_{(s,\widehat{\nu}]}\right\|_2 \le \frac{c_x}{128 C_x} \kappa' \quad \text{and} \quad \left\|\widehat{\alpha}_2 - \beta^*_{(\widehat{\nu},e]}\right\|_2 \le \frac{c_x}{128 C_x} \kappa'.$$

Since $\beta^*_{(\widehat{\nu},e]} = \beta^*_{\eta+1}$, it follows that

$$\left\|\widehat{\alpha}_2 - \beta^*_{\eta+1}\right\|_2 \le \frac{c_x}{128 C_x} \kappa'. \tag{9}$$

Note that

$$\begin{aligned}
\left\|\widehat{\alpha}_1 - \beta^*_{(s,\widehat{\nu}]}\right\|_2 =& \left\|\widehat{\alpha}_1 - \frac{(s-\eta)\beta^*_\eta + (\widehat{\nu}-\eta)\beta^*_{\eta+1}}{\widehat{\nu}-s}\right\|_2 \\
\geq& \|\widehat{\alpha}_1 - \beta^*_\eta\|_2 - \left\|\beta^*_\eta - \frac{(s-\eta)\beta^*_\eta + (\widehat{\nu}-\eta)\beta^*_{\eta+1}}{\widehat{\nu}-s}\right\|_2 \\
\geq& \|\widehat{\alpha}_1 - \beta^*_\eta\|_2 - \frac{\widehat{\nu}-\eta}{\widehat{\nu}-s}\|\beta^*_\eta - \beta^*_{\eta+1}\|_2 \\
\geq& \|\widehat{\alpha}_1 - \beta^*_\eta\|_2 - \frac{C\mathfrak{s}\log(pn)/(\kappa^{'})^2}{\frac{1}{20}C_{snr}\mathfrak{s}\log(pn)/\kappa^2}(\kappa') \\
\geq& \|\widehat{\alpha}_1 - \beta^*_\eta\|_2 - \frac{c_x}{128C_x}\frac{\kappa^2}{\kappa'} \\
\geq& \|\widehat{\alpha}_1 - \beta^*_\eta\|_2 - \frac{c_x}{128C_x}\kappa',
\end{aligned}$$

where the third inequality follows from the fact that $\widehat{\nu} - s \geq \frac{e-s}{10} \geq \frac{\Delta}{20} \geq \frac{1}{20}C_{snr}\kappa^{-2}\mathfrak{s}\log(pn)$ and the fourth inequality holds if C_{snr} is sufficiently large. As a result

$$\|\widehat{\alpha}_1 - \beta^*_\eta\|_2 \leq \frac{c_x}{64C_x}\kappa'. \tag{10}$$

The desired result is an immediate consequence of Equation (9) and Equation (10). ■

Lemma 4 *Suppose that $[s+1, e] \subset [1, n]$ is any interval such that $e - s \geq \frac{\Delta}{2}$ and that $[s+1, e]$ contains exactly one change point η which satisfies*

$$\min\{\eta - s, e - \eta\} \geq \frac{e-s}{10}.$$

Suppose Assumptions 1-2 hold and $\lambda = C_\lambda\sqrt{\log(pn)}$ for a sufficiently large constant C_λ. Let $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\nu})$ be the output of LGS $(\{y_i, x_i\}_{i=1}^n, (s, e], \lambda)$. Denote

$$\kappa' = \|\beta^*_\eta - \beta^*_{\eta+1}\|_2.$$

Then with probability at least $1 - (pn)^{-4}$, it holds that

$$|\eta - \widehat{\nu}| \leq C\frac{\mathfrak{s}\log(pn)}{(\kappa')^2}.$$

Proof [Proof of Theorem 4] Let S_1 be the support of β^*_i when $i \in (s, \eta]$ and S_2 be the support of β^*_i when $i \in (\eta, e]$. Denote

$$S = S_1 \cup S_2.$$

Note that $|S| \leq 2\mathfrak{s}$ and that S is the common support for β^*_i for $i \in (s, e]$. Without loss of generality, we assume that $s < \eta < \widehat{\nu} < e$. Denote

$$\widehat{\beta}_i = \begin{cases} \widehat{\alpha}_1, & i \in (s, \widehat{\nu}], \\ \widehat{\alpha}_2, & i \in (\widehat{\nu}, e]. \end{cases}$$

If $\widehat{\nu} - \eta < C_1 \frac{\mathfrak{s} \log(pn)}{\kappa^2}$ for some sufficiently large constant C_1, then the desired result holds.

So suppose that

$$\widehat{\nu} - \eta \geq C_1 \frac{\mathfrak{s} \log(pn)}{\kappa^2}. \tag{11}$$

Note that since $\kappa < C_\kappa < \infty$, Equation (11) implies that

$$\widehat{\nu} - \eta \geq C_1 C_\kappa^{-2} \mathfrak{s} \log(pn). \tag{12}$$

By assumption,

$$\min\{\eta - s, e - \widehat{\nu}\} \geq \frac{e-s}{10} \geq \frac{\Delta}{20} \geq \frac{1}{20} C_{snr} C_\kappa^{-2} \mathfrak{s} \log(pn). \tag{13}$$

Step 1. In this step, it is shown that with probability at least $1 - C(np)^{-5}$,

$$\sum_{i=s+1}^{e} \|\widehat{\beta}_i - \beta_i^*\|_2^2 \leq C_3 \mathfrak{s} \lambda^2.$$

From Equation (4), it holds that

$$\sum_{i=s+1}^{e} (y_i - x_i^\top \widehat{\beta}_i)^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{i=s+1}^{e} \left(\widehat{\beta}_{i,j}\right)^2} \leq \sum_{i=s+1}^{e} (y_i - x_i^\top \beta_i^*)^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{i=s+1}^{e} \left(\beta_{i,j}^*\right)^2}. \tag{14}$$

Let $\delta_i = \widehat{\beta}_i - \beta_i^*$. It holds that

$$\sum_{i=s+1}^{e-1} \mathbb{1}\{\delta_i \neq \delta_{i+1}\} = 2.$$

Equation (14) implies that

$$\sum_{i=s+1}^{e} \|\delta_i^\top x_i\|_2^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{i=s+1}^{e} \left(\widehat{\beta}_{i,j}\right)^2} \leq 2 \sum_{i=s+1}^{e} (y_i - x_i^\top \beta_i^*) \delta_i^\top x_i + \lambda \sum_{j=1}^{p} \sqrt{\sum_{i=s+1}^{e} \left(\beta_{i,j}^*\right)^2}. \tag{15}$$

Note that since S is the common support for β_i^* for $i \in (s, e]$, it holds that

$$\begin{aligned}
&\sum_{j=1}^{p} \sqrt{\sum_{i=s+1}^{e} \left(\beta_{i,j}^*\right)^2} - \sum_{j=1}^{p} \sqrt{\sum_{i=s+1}^{e} \left(\widehat{\beta}_{i,j}\right)^2} \\
=&\sum_{j \in S} \sqrt{\sum_{i=s+1}^{e} \left(\beta_{i,j}^*\right)^2} - \sum_{j \in S} \sqrt{\sum_{i=s+1}^{e} \left(\widehat{\beta}_{i,j}\right)^2} - \sum_{j \in S^c} \sqrt{\sum_{i=s+1}^{e} \left(\widehat{\beta}_{i,j}\right)^2}
\end{aligned}$$

$$\leq \sum_{j\in S}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}-\sum_{j\in S^c}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}. \tag{16}$$

Note that for any $j\in[1,\ldots,p]$, from Equation (12), Equation (13) and Theorem 8, it holds that

$$\sup_{s<i\leq e}\frac{\delta_{i,j}}{\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}}\leq C_4\frac{1}{\sqrt{\mathfrak{s}\log(pn)}}. \tag{17}$$

As a result, with probability at least $1-(np)^{-5}$

$$\left|\sum_{i=s+1}^{e}(y_i-x_i^\top\beta_i^*)\delta_i^\top x_i\right|=\left|\sum_{i=s+1}^{e}\varepsilon_i\delta_i^\top x_i\right|=\left|\sum_{j=1}^{p}\left\{\left(\frac{\sum_{i=s+1}^{e}\varepsilon_i\delta_{i,j}x_{i,j}}{\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}}\right)\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}\right\}\right|$$

$$\leq \sup_{j=1,\ldots,p}\left|\frac{\sum_{i=s+1}^{e}\varepsilon_i\delta_{i,j}x_{i,j}}{\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}}\right|\sum_{j=1}^{p}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}\leq C_5\sqrt{\log(pn)}\sum_{j=1}^{p}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}$$

$$\leq(\lambda/4)\sum_{j=1}^{p}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}, \tag{18}$$

where the second inequality follows from Theorem 7 and Equation (17) if C_5 is a sufficiently large constant, and the last inequality follows from $\lambda=C_\lambda\sqrt{\log(pn)}$.
Combining (14), (15), (16) and (18) yields

$$\sum_{i=s+1}^{e}(\delta_i^\top x_i)^2+\frac{\lambda}{2}\sum_{j\in S^c}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}\leq\frac{3\lambda}{2}\sum_{j\in S}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2}. \tag{19}$$

Step 2. To apply restricted eigenvalue conditions, let

$$I=(s,e],\quad I_1=(s,\eta],\quad I_2=(\eta,\widehat{\nu}],\quad I_3=(\widehat{\nu},e].$$

Denote

$$\mathcal{E}_I:=\left\{v^\top\left(\frac{1}{|I|}\sum_{i\in I}x_ix_i^\top\right)v\geq\frac{1}{16}v^\top\Sigma v-C_r\frac{\log(p)}{|I|}\|v\|_1^2\right\}.$$

If C_1 in Equation (12) and C_{snr} in Equation (13) are sufficiently large constants, it holds that

$$\min\{|I_1|,\ |I_2|,\ |I_3|\}\geq C_6\mathfrak{s}\log(np) \tag{20}$$

for some sufficiently large C_6. For $h=1,2,3$, denote $\mathcal{E}_{I_h}$ the same way as $\mathcal{E}_I$. So by Theorem 9, $P(\mathcal{E}_{I_h})\geq 1-(np)^{-5}$ for $h=1,2,3$. Denote $\delta_{I_h}=\delta_i$ for any $i\in I_h$. With probability at least $1-3(np)^{-5}$, on the event $\bigcap_{h=1,2,3}\mathcal{E}_{I_h}$,

$$\sum_{i=s+1}^{e}(\delta_i^\top x_i)^2=\sum_{h=1,2,3}\sum_{i\in I_h}(\delta_{I_h}^\top x_i)^2$$

$$
\begin{aligned}
&\geq \sum_{h=1,2,3} \frac{|I_h|}{16} \delta_{I_h}^\top \Sigma \delta_{I_h} - C_r \log(p) \|\delta_{I_h}\|_1^2 \\
&= \sum_{h=1,2,3} \frac{|I_h|}{16} \delta_{I_h}^\top \Sigma \delta_{I_h} - C_r \log(p) \big(\|\delta_{I_h}(S)\|_1 + \|\delta_{I_h}(S^c)\|_1 \big)^2 \\
&\geq \sum_{h=1,2,3} \frac{c_x |I_h|}{16} \|\delta_{I_h}\|_2^2 - 2C_r |S| \log(p) \|\delta_{I_h}\|_2^2 - 2C_r \log(p) \|\delta_{I_h}(S^c)\|_1^2 \\
&\geq \sum_{h=1,2,3} \frac{c_x |I_h|}{20} \|\delta_{I_h}\|_2^2 - 2C_r \log(p) \|\delta_{I_h}(S^c)\|_1^2, \qquad (21)
\end{aligned}
$$

where the last inequality follows from (20) and $|S| \leq 2\mathfrak{s}$ and $\|\delta_{I_h}(S)\|_1 = \sum_{j \in S} |\delta_{I_h,j}|$.

Step 3. Note that

$$
\begin{aligned}
\sqrt{\sum_{h=1,2,3} \|\delta_{I_h}(S^c)\|_1^2} &= \sqrt{\sum_{h=1,2,3} \left(\sum_{j \in S^c} |\delta_{I_h,j}| \right)^2} = \sqrt{\sum_{h=1,2,3} \left(\sqrt{\frac{|I_h|}{|I_h|}} \sum_{j \in S^c} |\delta_{I_h,j}| \right)^2} \\
&\leq \min\{|I_1|, |I_2|, |I_3|\}^{-1/2} \sqrt{\sum_{i=s+1}^{e} \left(\sum_{j \in S^c} |\delta_{i,j}| \right)^2} \\
&\leq \min\{|I_1|, |I_2|, |I_3|\}^{-1/2} \sum_{j \in S^c} \sqrt{\sum_{i=s+1}^{e} (\delta_{i,j})^2} \\
&\leq 3 \min\{|I_1|, |I_2|, |I_3|\}^{-1/2} \sum_{j \in S} \sqrt{\sum_{i=s+1}^{e} (\delta_{i,j})^2} \\
&\leq 3 \min\{|I_1|, |I_2|, |I_3|\}^{-1/2} \sqrt{|S| \sum_{j \in S} \sum_{i=s+1}^{e} (\delta_{i,j})^2} \\
&\leq \sqrt{\frac{c_x}{200 C_r \log(p)}} \sqrt{\sum_{i=s+1}^{e} \|\delta_i\|_2^2} \quad ,
\end{aligned}
$$

where the second inequality follows from the generalized Minkowski's inequality, the third inequality follows from Equation (19), which implies that

$$
\sum_{j \in S^c} \sqrt{\sum_{i=s+1}^{e} \big(\delta_{i,j}\big)^2} \leq 3 \sum_{j \in S} \sqrt{\sum_{i=s+1}^{e} \big(\delta_{i,j}\big)^2},
$$

and the last inequality follows from Equation (20) for sufficiently large C_6. So

$$
\sqrt{\sum_{i=1,2,3} \|\delta_{I_h}(S^c)\|_1^2} \leq \sqrt{\frac{c_x}{200 C_r \log(p)}} \sqrt{\sum_{i=s+1}^{e} \|\delta_i\|_2^2} \, . \qquad (22)
$$

Therefore, for any

$$\begin{aligned}(21) &= \sum_{h=1,2,3} \frac{c_x|I_h|}{20}\|\delta_{I_h}\|_2^2 - 2C_r\log(p)\|\delta_{I_h}(S^c)\|_1^2 \\ &= \sum_{i=s+1}^{e} \frac{c_x}{20}\|\delta_i\|_2^2 - 2C_r\log(p)\|\delta_{I_h}(S^c)\|_1^2 \\ &\geq \sum_{i=s+1}^{e} \frac{c_x}{20}\|\delta_i\|_2^2 - 2C_r\log(p)\frac{c_x}{200C_r\log(p)}\sum_{i=s+1}^{e}\|\delta_i\|_2^2 \\ &\geq \sum_{i=s+1}^{e} \frac{c_x}{25}\|\delta_i\|_2^2 \ ,\end{aligned}$$

where the second last inequality follows from Equation (22).

Step 4. Putting the previous steps together, it holds that

$$\begin{aligned}&\sum_{i=s+1}^{e} \frac{c_x}{25}\|\delta_i\|_2^2 + \frac{\lambda}{2}\sum_{j\in S^c}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2} \\ \leq &\sum_{i=s+1}^{e}(\delta_i^\top x_i)^2 + \frac{\lambda}{2}\sum_{j\in S^c}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2} \\ \leq &\frac{3\lambda}{2}\sum_{j\in S}\sqrt{\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2} \\ \leq &\frac{3\lambda}{2}\sqrt{|S|\sum_{j\in S}\sum_{i=s+1}^{e}\left(\delta_{i,j}\right)^2} \\ \leq &3\lambda\sqrt{\mathfrak{s}}\sqrt{\sum_{i=s+1}^{e}\|\delta_i\|_2^2} \ .\end{aligned}$$

Therefore,

$$\sum_{i=s+1}^{e}\|\widehat{\beta}_i - \beta_i^*\|_2^2 = \sum_{i=s+1}^{e}\|\delta_i\|_2^2 \leq C_7\lambda^2\mathfrak{s} \ . \tag{23}$$

Step 5. From Equation (23) and the definition of $\widehat{\beta}_i$, it holds that

$$\sum_{i=s+1}^{e}\|\widehat{\beta}_i - \beta_i^*\|_2^2 = |I_1|\|\beta_\eta^* - \widehat{\alpha}_1\|_2^2 + |I_2|\|\beta_{\eta+1}^* - \widehat{\alpha}_1\|_2^2 + |I_3|\|\beta_{\eta+1}^* - \widehat{\alpha}_2\|_2^2 \leq C_7\lambda^2\mathfrak{s}. \tag{24}$$

Let

$$\kappa' = \|\beta_\eta^* - \beta_{\eta+1}^*\|_2.$$

By Equation (13)

$$\min\{|I_1|,\ |I_3|\} \geq \frac{\Delta}{20}.$$

Therefore it holds that

$$\|\beta^*_\eta - \widehat{\alpha}_1\|_2^2 \leq \frac{C_7\lambda^2\mathfrak{s}}{|I_1|} \leq \frac{C_7\lambda^2\mathfrak{s}}{\frac{1}{20}\Delta} \leq \frac{C_7C_\lambda^2\log(pn)\mathfrak{s}}{\frac{1}{20}C_{snr}\log(pn)\mathfrak{s}/\kappa^2} \leq \frac{1}{16}\kappa^2,$$

where the first inequality follows from Equation (24), the third inequality follows from Assumption 2 **b** and the last inequality holds for sufficiently large C_{snr}. Similarly

$$\|\beta^*_{\eta+1} - \widehat{\alpha}_2\|_2^2 \leq \frac{1}{16}\kappa^2.$$

So

$$\|\beta^*_{\eta+1} - \widehat{\alpha}_1\|_2 \geq \|\beta^*_{\eta+1} - \beta^*_\eta\|_2 - \|\beta^*_\eta - \widehat{\alpha}_1\|_2 \geq \kappa' - \frac{1}{4}\kappa \geq \kappa'/2.$$

Equation (24) further implies that

$$(\widehat{\nu} - \eta)(\kappa')^2/4 \leq |I_2|\|\beta^*_\eta - \widehat{\alpha}_1\|_2^2 \leq C_7\lambda^2\mathfrak{s},$$

which implies that

$$\widehat{\nu} - \eta \leq \frac{4C_7\lambda^2\mathfrak{s}}{(\kappa')^2},$$

as desired. ■

Lemma 5 *Suppose that $[s+1, e] \subset [1, n]$ is any interval such that $e - s \geq \frac{\Delta}{2}$ and that $[s+1, e]$ contains exactly one change point η which satisfies*

$$\min\{\eta - s, e - \eta\} \geq \frac{e-s}{10}.$$

Suppose

$$(\widehat{\alpha}_1, \widehat{\alpha}_2) = \underset{\alpha_1,\alpha_2\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=s+1}^{\nu} \left(y_i - x_i^\top\alpha_1\right)^2 + \sum_{i=\nu+1}^{e} (y_i - x_i^\top\alpha_2)^2 + \lambda\sum_{j=1}^{p}\sqrt{(\nu-s)(\alpha_{1,j})^2 + (e-\nu)(\alpha_{2,j})^2} \right\}.$$

If in addition, Assumptions 1-2 hold and that $\lambda = C_\lambda\sqrt{\log(pn)}$ for sufficiently large C_λ, then with probability at least $1 - (pn)^{-4}$, it holds that

$$\left\|\widehat{\alpha}_1 - \beta^*_{(s,\nu]}\right\|_2^2 \leq C\frac{\mathfrak{s}\log(pn)}{\Delta} \quad \textit{and} \quad \left\|\widehat{\alpha}_2 - \beta^*_{(\nu,e]}\right\|_2^2 \leq C\frac{\mathfrak{s}\log(pn)}{\Delta},$$

where

$$\beta^*_{(a,b]} = \frac{1}{b-a}\sum_{i=a+1}^{b}\beta^*_i.$$

Proof [Proof of Theorem 5] Let S_1 be the support of β_η^* and S_2 be the support of $\beta_{\eta+1}^*$. Denote

$$S = S_1 \cup S_2.$$

Note that $|S| \leq 2\mathfrak{s}$ and that S is the common support for β_i^* for $i \in (s, e]$. Without loss of generality, assume that $s < \eta < \nu < e$. Note that

$$\min\{\nu - s, e - \nu\} > \frac{\Delta}{20} \geq \frac{1}{20} C_{snr} C_\kappa^{-2} \mathfrak{s} \log(pn). \tag{25}$$

For brevity, denote

$$I_1 = (s, \nu], \quad I_2 = (\nu, e], \quad \alpha_1^* = \beta_{(s,\nu]}^* \quad \text{and} \quad \alpha_2^* = \beta_{(\nu,e]}^* .$$

Step 1. From Equation (4), it holds that

$$\begin{aligned} &\sum_{i \in I_1} (y_i - x_i^\top \widehat{\alpha}_1)^2 + \sum_{i \in I_2} (y_i - x_i^\top \widehat{\alpha}_2)^2 + \lambda \sum_{j=1}^p \sqrt{|I_1|(\widehat{\alpha}_{1,j})^2 + |I_2|(\widehat{\alpha}_{2,j})^2} \\ \leq &\sum_{i \in I_1} (y_i - x_i^\top \alpha_1^*)^2 + \sum_{i \in I_2} (y_i - x_i^\top \alpha_2^*)^2 + \lambda \sum_{j=1}^p \sqrt{|I_1|(\alpha_{1,j}^*)^2 + |I_2|(\alpha_{2,j}^*)^2}. \end{aligned} \tag{26}$$

Let $\phi_i = \widehat{\alpha}_i - \alpha_i^*$. Equation (26) implies that

$$\begin{aligned} &\sum_{i \in I_1} (x_i^\top \phi_1)^2 + \sum_{i \in I_2} (x_i^\top \phi_2)^2 + \lambda \sum_{j=1}^p \sqrt{|I_1|(\widehat{\alpha}_{1,j})^2 + |I_2|(\widehat{\alpha}_{2,j})^2} \\ \leq &2 \sum_{i \in I_1} (y_i - x_i^\top \alpha_1^*) x_i^\top \phi_1 + 2 \sum_{i \in I_2} (y_i - x_i^\top \alpha_2^*) x_i^\top \phi_2 + \lambda \sum_{j=1}^p \sqrt{|I_1|(\alpha_{1,j}^*)^2 + |I_2|(\alpha_{2,j}^*)^2} \\ = &2 \sum_{i \in I_1} (\beta_i^* - \alpha_1^*) x_i x_i^\top \phi_1 + 2 \sum_{i \in I_2} (\beta_i^* - \alpha_2^*) x_i x_i^\top \phi_2 \qquad (27) \\ &+ 2 \sum_{i \in I_1} \varepsilon_i x_i^\top \phi_1 + 2 \sum_{i \in I_2} \varepsilon_i x_i^\top \phi_2 \qquad (28) \\ &+ \lambda \sum_{j=1}^p \sqrt{|I_1|(\alpha_{1,j}^*)^2 + |I_2|(\alpha_{2,j}^*)^2}. \end{aligned}$$

Note that S is the common support for α_1^* and α_2^* for $i \in (s, e]$. So

$$\begin{aligned} &\sum_{j=1}^p \sqrt{|I_1|(\alpha_{1,j}^*)^2 + |I_2|(\alpha_{2,j}^*)^2} - \sum_{j=1}^p \sqrt{|I_1|(\widehat{\alpha}_{1,j})^2 + |I_2|(\widehat{\alpha}_{2,j})^2} \\ = &\sum_{j \in S} \sqrt{|I_1|(\alpha_{1,j}^*)^2 + |I_2|(\alpha_{2,j}^*)^2} - \sum_{j \in S} \sqrt{|I_1|(\widehat{\alpha}_{1,j})^2 + |I_2|(\widehat{\alpha}_{2,j})^2} - \sum_{j \in S^c} \sqrt{|I_1|(\widehat{\alpha}_{1,j})^2 + |I_2|(\widehat{\alpha}_{2,j})^2} \\ \leq &\sum_{j \in S} \sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2} - \sum_{j \in S^c} \sqrt{|I_1|(\widehat{\alpha}_{1,j})^2 + |I_2|(\widehat{\alpha}_{2,j})^2} \end{aligned}$$

$$=\sum_{j\in S}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}-\sum_{j\in S^c}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}.$$

Step 2. Note that for any $h=1,2$ and $j\in[1,p]$, from Equation (25), it holds that

$$\frac{\phi_{h,j}}{\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}}\leq\frac{1}{\sqrt{|I_h|}}\leq C_4\frac{1}{\sqrt{\mathfrak{s}\log(pn)}}. \tag{29}$$

As a result, with probability at least $1-C(np)^{-5}$

$$\begin{aligned}
|(28)|=&\left|\sum_{j=1}^{p}\left\{\frac{\sum_{i\in I_1}\varepsilon_i\phi_{1,j}x_{i,j}+\sum_{i\in I_2}\varepsilon_i\phi_{2,j}x_{i,j}}{\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}}\right\}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}\right|\\
\leq&\sup_{j=1,\ldots,p}\left|\frac{\sum_{i\in I_1}\varepsilon_i\phi_{1,j}x_{i,j}+\sum_{i\in I_2}\varepsilon_i\phi_{2,j}x_{i,j}}{\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}}\right|\sum_{j=1}^{p}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}\\
\leq&C_5\sqrt{\log(pn)}\sum_{j=1}^{p}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}\leq(\lambda/8)\sum_{j=1}^{p}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2},
\end{aligned}$$

where the second inequality follows from Theorem 7 and Equation (29), and the last inequality follows from $\lambda=C_\lambda\sqrt{\log(pn)}$.
In addition, since $s<\eta<\nu<e$, it holds that

$$\alpha_2^*=\frac{1}{(e-\nu)}\sum_{i=\nu+1}^{e}\beta_i^*=\beta_{\eta+1}^*.$$

As a result

$$\begin{aligned}
(27)=&\sum_{i\in I_1}(\beta_i^*-\alpha_1^*)^\top x_ix_i^\top\phi_1+\sum_{i\in I_2}(\beta_i^*-\alpha_2^*)^\top x_ix_i^\top\phi_2\\
=&\sum_{i\in I_1}\left(\beta_i^*-\frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_ix_i^\top\phi_1\\
\leq&\sup_{1\leq j\leq p}\left|\sum_{i\in I_1}\left(\beta_i^*-\frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_ix_{i,j}\right|\|\phi_1\|_1\\
=&\sup_{1\leq j\leq p}\left|\frac{1}{\sqrt{|I_1|}}\sum_{i\in I_1}\left(\beta_i^*-\frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_ix_{i,j}\right|\sqrt{|I_1|}\|\phi_1\|_1\\
\leq&\sup_{1\leq j\leq p}\left|\frac{1}{\sqrt{|I_1|}}\sum_{i\in I_1}\left(\beta_i^*-\frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_ix_{i,j}\right|\sum_{j=1}^{p}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}\quad.
\end{aligned}\tag{30}$$

For any $j\in[1,p]$, denote $\Sigma[,j]$ to be the j-th column of Σ. Then

$$\sum_{i\in I_1}\left(\beta_i^*-\frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top\Sigma[,j]=0.$$

In addition, it is straightforward to see that

$$\sup_{i\in I_1}\left\|\beta_i^* - \frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right\| \le C_\kappa.$$

So $\left(\beta_i^* - \frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_i x_{i,j}$ is a sub-exponential random variable with parameter $C_x^2 C_\kappa$ for any $i \in I_1$. As a result

$$\begin{aligned}
&P\left(\left|\frac{1}{\sqrt{|I_1|}}\sum_{i\in I_1}\left(\beta_i^* - \frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_i x_{i,j}\right| \ge \delta \text{ for all } 1\le j\le p\right)\\
=&P\left(\left|\sum_{i\in I_1}\left(\beta_i^* - \frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top \{x_i x_{i,j} - \Sigma[,j]\}\right| \ge \delta\sqrt{|I_1|} \text{ for all } 1\le j\le p\right)\\
\le& p\exp\left(-c\min\{\frac{\delta^2}{4C_x^2 C_\kappa}, \frac{\delta\sqrt{|I_1|}}{2C_x\sqrt{C_\kappa}}\}\right) \le p\exp\left(-c'\min\{\delta^2, \delta\sqrt{|I_1|}\}\right),
\end{aligned}$$

where the second to the last inequality follows from standard sub-exponential tail bounds. Since $|I_1| \ge \frac{\Delta}{20} \ge \frac{1}{20}C_{snr}C_\kappa^{-2}\mathfrak{s}\log(pn)$, letting $\delta = C_\delta\sqrt{\log(p)}$ for sufficiently large constant C_δ, it holds that with probability at least $1-(pn)^{-5}$,

$$\sup_{1\le i\le p}\left|\frac{1}{\sqrt{|I_1|}}\sum_{i\in I_1}\left(\beta_i^* - \frac{1}{|I_1|}\sum_{i'\in I_1}\beta_{i'}^*\right)^\top x_i x_{i,j}\right| \le C_\delta\sqrt{\log(p)} \le \frac{1}{8}\lambda.$$

Therefore (30) gives

$$(27) \le (\lambda/8)\sum_{j=1}^p \sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2}.$$

Step 3. Combing the previous two steps gives

$$\begin{aligned}
&\sum_{i\in I_1}(x_i^\top\phi_1)^2 + \sum_{i\in I_2}(x_i^\top\phi_2)^2 + \lambda\sum_{j\in S^c}\sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2}\\
\le&\lambda\sum_{j\in S}\sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2} + \frac{\lambda}{4}\sum_{j=1}^p\sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2}.
\end{aligned}$$

This gives

$$\sum_{i\in I_1}(x_i^\top\phi_1)^2 + \sum_{i\in I_2}(x_i^\top\phi_2)^2 + \frac{\lambda}{2}\sum_{j\in S^c}\sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2} \le \frac{3\lambda}{2}\sum_{j\in S}\sqrt{|I_1|(\phi_{1,j})^2 + |I_2|(\phi_{2,j})^2}. \tag{31}$$

Using exactly the same argument as in **Step 4** and **Step 5** in the proof of Theorem 4, it can be shown that

$$\sum_{i\in I_1}(x_i^\top\phi_1)^2 + \sum_{i\in I_2}(x_i^\top\phi_2)^2 \ge \frac{c_x}{25}\big(|I_1|\|\phi_1\|_2^2 + |I_2|\|\phi_2\|_2^2\big). \tag{32}$$

Therefore

$$\begin{aligned}|I_1|\|\phi_1\|_2^2+|I_2|\|\phi_2\|_2^2 \leq & C_6\lambda\sum_{j\in S}\sqrt{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2}\\ \leq & C_6\lambda\sqrt{|S|\sum_{j\in S}\left\{|I_1|(\phi_{1,j})^2+|I_2|(\phi_{2,j})^2\right\}}\\ \leq & C_6\lambda\sqrt{|S||I_1|\|\phi_1\|_2^2+|I_2|\|\phi_2\|_2^2}\\ \leq & 2C_6\lambda\sqrt{\mathfrak{s}}\sqrt{|I_1|\|\phi_1\|_2^2+|I_2|\|\phi_2\|_2^2} \quad ,\end{aligned}$$

where the first inequality follows from (31) and (32). This directly gives

$$|I_1|\|\phi_1\|_2^2+|I_2|\|\phi_2\|_2^2 \leq 8C_6^2\lambda^2\mathfrak{s}.$$

The desired result follows from the assumption that

$$\min\{|I_1|,|I_2|\} > \frac{1}{20}\Delta.$$

■

A.1 Additional Technical Lemmas

Lemma 6 *Let $\mathcal{R}$ be any linear subspace in $\mathbb{R}^n$ and $\mathcal{N}_{1/4}$ be a 1/4-net of $\mathcal{R}\cap B(0,1)$, where $B(0,1)$ is the unit ball in $\mathbb{R}^n$. For any $u\in\mathbb{R}^n$, it holds that*

$$\sup_{v\in\mathcal{R}\cap B(0,1)}\langle v,u\rangle \leq 2\sup_{v\in\mathcal{N}_{1/4}}\langle v,u\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{R}^n$.

Proof Due to the definition of $\mathcal{N}_{1/4}$, it holds that for any $v\in\mathcal{R}\cap B(0,1)$, there exists a $v_k\in\mathcal{N}_{1/4}$, such that $\|v-v_k\|_2<1/4$. Therefore,

$$\langle v,u\rangle=\langle v-v_k+v_k,u\rangle=\langle x_k,u\rangle+\langle v_k,u\rangle\leq\frac{1}{4}\langle v,u\rangle+\frac{1}{4}\langle v^\perp,u\rangle+\langle v_k,u\rangle,$$

where the inequality follows from $x_k=v-v_k=\langle x_k,v\rangle v+\langle x_k,v^\perp\rangle v^\perp$. Then we have

$$\frac{3}{4}\langle v,u\rangle\leq\frac{1}{4}\langle v^\perp,u\rangle+\langle v_k,u\rangle.$$

It follows from the same argument that

$$\frac{3}{4}\langle v^\perp,u\rangle\leq\frac{1}{4}\langle v,u\rangle+\langle v_l,u\rangle,$$

where $v_l\in\mathcal{N}_{1/4}$ satisfies $\|v^\perp-v_l\|_2<1/4$. Combining the previous two equation displays yields

$$\langle v,u\rangle\leq 2\sup_{v\in\mathcal{N}_{1/4}}\langle v,u\rangle.$$

■

For any vector $v \in \mathbb{R}^m$, denote $\mathcal{D}(v)$ to be the number of change point of v. That is,

$$\mathcal{D}(v) = \sum_{i=1}^{m} \mathbb{1}\{v_i \neq v_{i+1}\}.$$

Lemma 7 *For data generated according to Model 1, for any interval $I = (s, e] \subset \{1, \ldots, n\}$, it holds that for any $\delta > 0$ and any $j \in \{1, \ldots, p\}$,*

$$\mathbb{P}\left\{ \sup_{\substack{v \in \mathbb{R}^{(e-s)} \\ \|v\|_2 = 1, \mathcal{D}(v) = 2}} \left| \sum_{i=s+1}^{e} v_i \varepsilon_i x_i[j] \right| > \delta \right\} \leq Cn \exp\left\{ -c \min\left\{ \delta^2, \frac{\delta}{\|v\|_\infty} \right\} \right\}.$$

Proof This is a standard covering lemma. We provide a proof for completeness. For any $v \in \mathbb{R}^{(e-s)}$ satisfying $\mathcal{D}(v) = 2$ and $\|v\|_2 = 1$, let $\eta < \eta'$ be the change points of v. Then there are $(e-s)(e-s-1)/2$ possible choice of η, η'. For any η, η', denote

$$\mathcal{R}(\eta, \eta') = \{w \in \mathbb{R}^{(e-s)}, w_1 = \ldots = w_\eta \neq w_{\eta+1} = \ldots = w_{\eta'} \neq w_{\eta'+1} = \ldots = w_{(e-s)}\}$$

Then $\mathcal{R}(\eta, \eta')$ is a 3-dimensional subspace. Denote $\mathcal{N}_{1/4}(\eta, \eta')$ to be the covering number a the unit ball in $\mathcal{R}(\eta, \eta')$. Then $\mathcal{N}_{1/4}(\eta, \eta') \leq 9^2$. Therefore we have,

$$\begin{aligned}
&\mathbb{P}\left\{ \sup_{\substack{v \in \mathbb{R}^{(e-s)} \\ \|v\|_2 = 1, \mathcal{D}(v) = 2}} \left| \sum_{i=s+1}^{e} v_i \varepsilon_i x_i[j] \right| > \delta \right\} \\
\leq & \frac{(e-s)(e-s-1)}{2} 9^2 \sup_{\eta \in (s,e],\ v \in \mathcal{N}_{1/4}(\eta, \eta')} \mathbb{P}\left\{ \left| \sum_{i=s+1}^{e} v_i \varepsilon_i x_i[j] \right| > \delta/2 \right\} \\
\leq & Cn^2 \exp\left\{ -c' \min\left\{ \frac{\delta^2}{4C_x^2}, \frac{\delta}{2C_x \|v\|_\infty} \right\} \right\} \\
\leq & Cn^2 \exp\left\{ -c \min\left\{ \delta^2, \frac{\delta}{\|v\|_\infty} \right\} \right\},
\end{aligned}$$

where the first inequality follows from Theorem 6 and union bounds, the second last inequality holds because for any fixed v, $v_i \varepsilon_i x_{i,j}$ is a sub-Exponential random variable with parameter bounded by C_x. ■

Lemma 8 *Suppose*

$$v = (\underbrace{a, \ldots, a}_{K_1}, \underbrace{b, \ldots, b}_{K_2}, \underbrace{c, \ldots, c}_{K_3})$$

and that $v \neq 0$. Then

$$\left\| \frac{v}{\|v\|_2} \right\|_\infty \leq \frac{1}{\sqrt{\min\{K_1, K_2, K_3\}}}.$$

Proof It suffices to show $a/\|v\|_2 \leq \frac{1}{\sqrt{\min\{K_1,K_2,K_3\}}}$. If $a = 0$, then this trivially holds. Otherwise

$$\frac{a}{\|v\|_2} = \frac{a}{\sqrt{a^2K_1 + b^2K_2 + c^2K_3}} \leq \frac{1}{\sqrt{K_1}}.$$

■

Theorem 9 *Suppose* $\{x_i\}_{1\leq i\leq n} \overset{i.i.d.}{\sim} N_p(0,\Sigma)$. *Let* $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. *Then there exists constants* c *and* C *such that for all* $v \in \mathbb{R}^p$,

$$v^\top \widehat{\Sigma} v \geq \frac{1}{16} v^\top \Sigma v - C_r \frac{\log(p)}{n}\|v\|_1^2$$

with probability at least $1 - \exp(-cn)$.

Proof This is the well known restricted eigenvalue condition. The proof cam be found in Raskutti et al. (2010). ■

Appendix B. Proofs Related to Theorem 3

For any univariate time series $\{z_i\}_{i=1}^n$ and any $1 \leq s < t < e < n$, denote the CUSUM statistics as

$$\widetilde{Z}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}}\sum_{i=s+1}^{t} z_i - \sqrt{\frac{t-s}{(e-s)(e-t)}}\sum_{i=t+1}^{e} z_i.$$

Proof [Proof of Theorem 3] Throughout the proof, assume that event $\mathcal{A}(\{y_i^{(2)}, x_i^{(2)}\}_{i=1}^n, \{u_m\}_{m=1}^M, \xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)})$ in Equation (44), event $\mathcal{B}(\{y_i^{(2)}, x_i^{(2)}\}_{i=1}^n, \{u_m\}_{m=1}^M, \xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)})$ in Equation (45), event $\mathcal{M}$ in Equation (39) and the good event in Theorem 2 (with data $\{y_i^{(1)}, x_i^{(1)}\}_{i=1}^n$) hold. Denote

$$\delta_k = \frac{C(\mathfrak{N}+1)\log(n)}{\kappa_k^2} \quad \text{and} \quad \delta_{\max} = \frac{C(\mathfrak{N}+1)\log(n)}{\kappa^2}.$$

Since δ_k is the desired localization rate, by induction, it suffices to consider any generic $(s,e] \subset (0,n]$ that satisfies the following three conditions:

$$\begin{aligned}
&\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq -1;\\
&\text{either } \eta_r - s \leq \delta_r \quad \text{or} \quad s - \eta_{r-1} \leq \delta_{r-1};\\
&\text{either } \eta_{r+q+1} - e \leq \delta_{r+q+1} \quad \text{or} \quad e - \eta_{r+q} \leq \delta_{r+q}.
\end{aligned}$$

Here $q = -1$ indicates that there is no change point contained in $(s,e]$.

Observe that under Assumption 2, for sufficiently large constant C_{snr}, it holds that $\delta_{\max} < \Delta/4$. Therefore, it has to be the case that for any true change point $\eta_r \in (0,n]$, either $|\eta_r - s| \leq \delta_r$ or $|\eta_r - s| \geq \Delta - \delta_{\max} \geq \Delta/4$. This means that $\min\{|\eta_r - e|, |\eta_r - s|\} \leq \delta_r$

indicates that η_r is a detected change point in the previous induction step, even if $\eta_r \in (s, e]$. We refer to $\eta_r \in (s, e]$ as an undetected change point if $\min\{\eta_r - s, \eta_r - e\} \geq \Delta/4$.

To complete the induction step, it suffices to show that VPWBS$((s, e], \{(a_m, b_m)\}_{m=1}^M, \lambda, \tau, \zeta)$ **(i)** will not detect any new change point in $(s, e]$ if all the change points in that interval have been previous detected, and
(ii) will find a point D_{m*} in $(s, e]$ such that $|\eta_r - D_{m*}| \leq \delta_r$ if there exists at least one undetected change point in (s, e).

Let $\{\widehat{\alpha}_1^m, \widehat{\alpha}_2^m, \widehat{\nu}^m\}$ be the output of LGS $(\{y_i^{(1)}, x_i^{(1)}\}_{i=1}^n, (a_m, b_m), \lambda)$, and

$$u_m = \frac{\widehat{\alpha}_2^m - \widehat{\alpha}_1^m}{\|\widehat{\alpha}_2^m - \widehat{\alpha}_1^m\|_2} \in \mathbb{R}^p$$

for all $1 \leq m \leq M$. Since the intervals $\{(a_m, b_m)\}_{m=1}^M$ are sampled independently from the data, the rest of the argument is made on the event $\mathcal{M}$, which is defined in Equation (39) and this event has no effects on the distribution of the data.

Step 1. Denote

$$f_i(u_m) = E\{z_i(u_m)\} \quad \text{and} \quad \widetilde{f}_t^{s_m, e_m}(u_m) = E\{\widetilde{Z}_t^{a_m, b_m}(u_m)\}.$$

Note that

$$f_i(u_m) = u_m \Sigma \beta_i.$$

On the event $\mathcal{M}$, for any $\eta_k \in (0, n]$, without loss of generality, there exists

$$a_k \in [\eta_k - 3\Delta/4, \eta_k - \Delta/2] \quad \text{and} \quad b_k \in [\eta_k + \Delta/2, \eta_k + 3\Delta/4]. \tag{33}$$

In this step, it is shown that for each $k \in [1, \ldots, K]$, it holds that

$$\max_{a_k + \zeta \leq t \leq b_k - \zeta} |\widetilde{f}_t^{a_k, b_k}(u_k)| \geq \frac{7c_x}{32} \sqrt{\Delta} \kappa_k, \tag{34}$$

where

$$u_k = \frac{\widehat{\alpha}_2^k - \widehat{\alpha}_1^k}{\|\widehat{\alpha}_2^k - \widehat{\alpha}_1^k\|_2},$$

and

$$\{\widehat{\alpha}_1^k, \widehat{\alpha}_2^k, \widehat{\nu}^k\} = \text{ LGS } (\{y_i^{(1)}, x_i^{(1)}\}_{i=1}^n, (a_k, b_k), \lambda).$$

By Equation (33), $[a_k, b_k]$ contains exactly one change point η_k. Since $f_i(u_k)$ is a one dimensional population time series, it holds that

$$\widetilde{f}_t^{a_k, b_k}(u_k) = \begin{cases} \sqrt{\frac{t - a_k}{(b_k - a_k)(b_k - t)}} (b_k - \eta_k) u_k^\top \Sigma (\beta_{\eta_k}^* - \beta_{\eta_{k-1}}^*), & a_m < t \leq \eta_k, \\ \sqrt{\frac{b_k - t}{(b_k - a_k)(t - a_k)}} (\eta_k - a_k) u_k^\top \Sigma (\beta_{\eta_k}^* - \beta_{\eta_{k-1}}^*), & \eta_k < t \leq b_m. \end{cases} \tag{35}$$

Let

$$\{\widehat{\alpha}_1^k, \widehat{\alpha}_2^k, \widehat{\nu}^k\} \leftarrow \text{ LGS } (\{y_i^{(1)}, x_i^{(1)}\}_{i=1}^n, (a_k, b_k), \lambda).$$

From Theorem 2, with probability at least $1 - n^{-5}$, it holds that

$$\left\|(\widehat{\alpha}_1^k - \widehat{\alpha}_2^k) - (\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}})\right\|_2 \leq \frac{c_x}{32C_x}\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2.$$

Since $c_x < C_x$ by definition, it holds that

$$\frac{32}{33} \leq \frac{\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2}{\|\widehat{\alpha}_1 - \widehat{\alpha}_2\|_2} \leq \frac{32}{31}.$$

As a result,

$$\begin{aligned}
&u_k^\top \Sigma(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}) \\
=&\frac{(\widehat{\alpha}_1 - \widehat{\alpha}_2)^\top}{\|\widehat{\alpha}_1 - \widehat{\alpha}_2\|_2}\Sigma(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}) \\
=&\frac{(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}})^\top}{\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2}\Sigma(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}) + \left(\frac{(\widehat{\alpha}_1 - \widehat{\alpha}_2)^\top}{\|\widehat{\alpha}_1 - \widehat{\alpha}_2\|_2} - \frac{(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}})^\top}{\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2}\right)\Sigma(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}) \\
\geq&\frac{(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}})^\top}{\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2}\Sigma(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}) - 2\frac{\|\widehat{\alpha}_1 - \widehat{\alpha}_2 - (\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}})\|\|\Sigma\|_{\mathrm{op}}\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2^2}{\|\widehat{\alpha}_1 - \widehat{\alpha}_2\|_2\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2} \\
\geq&c_x\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2 - 3C_x\frac{c_x}{32C_x}\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2 \geq \frac{7}{8}c_x\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2. \qquad (36)
\end{aligned}$$

Therefore Equation (35) gives

$$|\widetilde{f}^{a_k,b_k}_{\eta_k}(u_k)| = \sqrt{\frac{(\eta_k - a_k)(b_k - \eta_k)}{b_k - a_k}}\left|u_k^\top\Sigma(\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}})\right| \geq \frac{1}{4}\sqrt{\Delta}\frac{7}{8}c_x\|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2 = \frac{7c_x}{32}\sqrt{\Delta}\kappa_k,$$

where the last inequality follows from Equation (33) and Equation (36). Under Assumption 1 and Assumption 2, for sufficiently large constant C_{snr}, it holds that $\zeta \leq \Delta/4$. Therefore Equation (33) also implies that $\eta_k \in [a_k + \zeta, b_k - \zeta]$

$$\max_{a_k+\zeta\leq t\leq b_k-\zeta}|\widetilde{f}^{a_k,b_k}_t(u_k)| \geq |\widetilde{f}^{a_k,b_k}_{\eta_k}(u_k)| \geq \frac{7c_x}{32}\sqrt{\Delta}\kappa_k.$$

This directly gives Equation (34).

Step 2. In this step, we will show that VPWBS$((s,e],\{(a_m,b_m)\}_{m=1}^M,\lambda,\tau,\zeta)$ consistently detect or reject the existence of undetected change points within $(s,e]$.

Let a_m, b_m and m^* be defined as in VPWBS$((s,e],\{(a_m,b_m)\}_{m=1}^M,\lambda,\tau,\zeta)$. Denote $z_i(u_m) = (u_m)^\top x_i^{(2)} y_i^{(2)}$ and $f_i(u_m) = E\{z_i(u_m)\} = u_m\Sigma\beta_i^*$. Let $\widetilde{Z}_t^{s,e}(u_m)$ and $\widetilde{f}_t^{s,e}(u_m)$ be the CUSUM statistics of the time series $z_i(u_m)$ and $f_i(u_m)$, respectively.

Suppose there exists a change point $\eta_r \in (s,e]$ such that $\min\{\eta_r - s, e - \eta_r\} \geq 3\Delta/4$. Then, on the event $\mathcal{M}$, there exists an interval $(a_k,b_k]$ selected by VPWBS such that $a_k \in [\eta_r - 3\Delta/4, \eta_r - \Delta/2]$ and $b_k \in [\eta_r + \Delta/2, \eta_r + 3\Delta/4]$. Then $[a_k,b_k] \subset [s,e]$ and so

$$(s_k,e_k] = (a_k,b_k] \cap (s,e] = (a_k,b_k].$$

Since Equation (34) in **Step 1** holds for $(a_k, b_k]$, we have that

$$\begin{aligned} A_k &= \max_{a_k+\zeta\le t\le b_k-\zeta} |\widetilde{Z}_t^{a_k,b_k}(u_k)| \\ &\ge \max_{a_k+\zeta\le t\le b_k-\zeta} |\widetilde{f}_t^{a_k,b_k}(u_k)| - C_1\sqrt{(\mathfrak{N}+1)\log(n)} \\ &\ge \frac{7c_x}{32}\sqrt{\Delta}\kappa_r - C_1\sqrt{(\mathfrak{N}+1)\log(n)} \\ &\ge \frac{7c_x}{64}\sqrt{\Delta}\kappa_r \ , \end{aligned}$$

where the first inequality holds on the event $\mathcal{A}(\{y_i^{(2)}, x_i^{(2)}\}_{i=1}^n, \{u_m\}_{m=1}^M, \xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)})$, the second inequality follows from Equation (34), and the last inequality follows from Assumption 2 with sufficiently large constant C_{snr}. Thus for any undetected change point η_r within $(s, e]$, it holds that

$$A_{m^*} = \sup_{1\le m\le M} A_m \ge A_k \ge c'\sqrt{\Delta}\kappa_r. \tag{37}$$

By Assumption 2 with sufficiently large constant C_{snr}, Equation (37) gives

$$A_{m^*} \ge c'\sqrt{\Delta}\kappa_r > C_\tau\sqrt{(\mathfrak{N}+1)\log(n)} = \tau.$$

As a result, VPWBS$((s, e], \{(a_m, b_m)\}_{m=1}^M, \lambda, \tau, \zeta)$ correctly accepts the existence of undetected change points.

Suppose there does not exist any undetected change points in $(s, e]$. Then for any $(s_m, e_m] = (a_m, b_m] \cap (s, e]$, one of the following situations must hold.

(a) There is no change point within $(s_m, e_m]$;

(b) there exists only one change point η_r within $(s_m, e_m]$ and $\min\{\eta_r - s_m, e_m - \eta_r\} \le \delta_r$;

(c) there exist two change points η_r, η_{r+1} within $(s_m, e_m]$ and

$$\eta_r - s_m \le \delta_r \quad \text{and} \quad e_m - \eta_{r+1} \le \delta_{r+1}.$$

The calculations of (c) is provided as the other two cases are similar and simpler. Note that for any $\|u_m\|_2 = 1$, it holds that

$$|f_{\eta_{r+1}}(u_m) - f_{\eta_{r+1}+1}(u_m)| = |u_m^\top\Sigma(\beta^*_{\eta_{r+1}} - \beta^*_{\eta_{r+1}+1})| \le \|u_m\|_2\|\Sigma\|_{\text{op}}\|\beta^*_{\eta_{r+1}} - \beta^*_{\eta_{r+1}+1}\| \le C_x\kappa_{r+1}$$

and similarly

$$|f_{\eta_r}(u_m) - f_{\eta_r+1}(u_m)| \le C_x\kappa_r.$$

By Theorem 12 and the assumption that $(s_m, e_m]$ contains only two change points, it holds that

$$\max_{s_m\le t\le e_m} |\widetilde{f}_t^{s_m,e_m}(u_m)| \le \sqrt{e_m - \eta_{r+1}}|f_{\eta_{r+1}}(u_m) - f_{\eta_{r+1}+1}(u_m)| + \sqrt{\eta_r - s_m}|f_{\eta_r}(u_m) - f_{\eta_r+1}(u_m)|$$

$$\leq C_2\sqrt{\delta_{r+1}}\kappa_{r+1} + C_2\sqrt{\delta_r}\kappa_r \leq C_3\sqrt{(\mathfrak{N}+1)\log(n)}.$$

Therefore under event $\mathcal{A}(\{y_i, x_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)})$,

$$A_m := \max_{s_m+\zeta\leq t\leq e_m-\zeta} |\widetilde{Z}_t^{s_m,e_m}(u_m)| \leq \max_{s_m+\zeta\leq t\leq e_m-\zeta} |\widetilde{f}_t^{s_m,e_m}(u_m)| + C_1\sqrt{(\mathfrak{N}+1)\log(n)} \leq C_4\sqrt{(\mathfrak{N}+1)\log(n)}.$$

So if $\tau = C_\tau\sqrt{(\mathfrak{N}+1)\log(n)}$ for sufficiently large C_τ, it holds that

$$A_m \leq \tau \quad \text{for all } 1 \leq m \leq M.$$

As a result, VPWBS$((s,e], \{(a_m, b_m)\}_{m=1}^M, \lambda, \tau, \zeta)$ correctly reject if $(s,e]$ contains no undetected change points.

Step 3. Assume that there exists a change point $\eta_r \in (s,e]$ such that

$$\min\{\eta_r - s, \eta_r - e\} \geq 3\Delta/4.$$

Let a_m, b_m and m^* be defined as in VPWBS$((s,e], \{(a_m, b_m)\}_{m=1}^M, \lambda, \tau, \zeta)$.

To complete the induction, it suffices to show that, there exists a change point $\eta_k \in (s_{m*}, e_{m*}]$ such that $\min\{\eta_k - s, \eta_k - e\} \geq 3\Delta/4$ and $|D_{m*} - \eta_k| \leq \delta_k$.

Consider the univariate time series

$$z_i(u_{m*}) = (u_m^\top x_i^{(2)})y_i^{(2)}, \quad f_i(u_{m*}) = E\{z_i(u_{m*})\} \quad \text{for all } 1 \leq i \leq n.$$

Since the collection of the change points of the time series $\{f_i(u_{m*})\}_{i=s_{m*}+1}^{e_{m*}}$ is a subset of that of $\{\eta_k\}_{k=0}^{K+1} \cap [s,e]$, we may apply Theorem 11 to the time series $\{z_i(u_{m*})\}_{i=s_{m*}+1}^{e_{m*}}$ and $\{f_i(u_{m*})\}_{i=s_{m*}+1}^{e_{m*}}$. Therefore, it suffices to justify that all the assumptions of Theorem 11 hold.

Let $\zeta = C_\zeta(\mathfrak{N}+1)\log(n)$ and $\xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)}$. Observe that from **Step 2** Equation (37), it holds that

$$A_{m^*} \geq c'\sqrt{\Delta}\kappa_r.$$

for all r such that $\min\{\eta_r - s, e - \eta_r\} \geq 3\Delta/4$. So Equation (40) holds. Equation (41) and Equation (42) are direct consequences of $\mathcal{A}(\{y_i^{(2)}, x_i^{(2)}\}_{i=1}^n, \{u_m\}_{m=1}^M, \xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)})$ and $\mathcal{B}(\{y_i^{(2)}, x_i^{(2)}\}_{i=1}^n, \{u_m\}_{m=1}^M, \xi = C_1\sqrt{(\mathfrak{N}+1)\log(n)})$. Equation (43) is a direct consequence of Assumptions 1 and 2.

Thus, all the conditions in Theorem 11 are met, and we therefore conclude that there exists a change point η_k of $\{f_i(u_{m*})\}_{i=s_{m^*}+1}^{e_{m^*}}$, satisfying

$$\min\{e_{m^*} - \eta_k, \eta_k - s_{m^*}\} > \Delta/4, \tag{38}$$

and

$$|D_{m*} - \eta_k| \leq \max\{C_3\xi^2\kappa_k^{-2}, \zeta\} \leq C(\mathfrak{N}+1)\log(n)\kappa_k^{-2},$$

where the last inequality holds because

$$C(\mathfrak{N}+1)\log(n)\kappa_k^{-2} \geq C(\mathfrak{N}+1)\log(n)C_\kappa^{-2} \geq C_\zeta(\mathfrak{N}+1)\log(n) = \zeta$$

for sufficiently large C. Observe that
i) The change points of $\{f_i(u_{m^*})\}_{i=s+1}^{e}$ belong to $(s,e] \cap \{\eta_k\}_{k=1}^{K}$; and
ii) Equation (38) and $(s_{m^*}, e_{m^*}] \subset (s,e]$ imply that

$$\min\{e-\eta_k, \eta_k - s\} > \Delta/4 > \frac{C(\mathfrak{N}+1)\log(n)}{\kappa^2} = \delta_{\max}.$$

As discussed in the argument before **Step 1**, this implies that η_k must be an undetected change point of $\{\beta_i^*\}_{i=1}^n$. ■

B.1 Additional Technical Lemmas

Let $\{a_m\}_{m=1}^M, \{b_m\}_{m=1}^M$ be two sequences independently selected at random from $\{1,\ldots,n\}$, and

$$\mathcal{M} = \big\{\text{For each } k \in \{1,\ldots,K\}, \text{there exist one } m \in \{1,\ldots,M\} \text{ such that } a_m \in \mathcal{S}_k, b_m \in \mathcal{E}_k\big\}, \tag{39}$$

where $\mathcal{S}_k = [\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\mathcal{E}_k = [\eta_k + \Delta/2, \eta_k + 3\Delta/4]$. In the following lemma below, we give a lower bound on the probability of $\mathcal{M}$.

Lemma 10 *For the event $\mathcal{M}$ defined in* (39)*, we have*

$$\mathbb{P}(\mathcal{M}) \geq 1 - \exp\left\{\log\left(\frac{n}{\Delta}\right) - \frac{M\Delta^2}{16n^2}\right\}.$$

Proof Since the number of change points are bounded by n/Δ,

$$\begin{aligned}
\mathbb{P}(\mathcal{M}^c) \leq& \sum_{k=1}^{K}\prod_{m=1}^{M}\big\{1 - \mathbb{P}\big(a_m \in \mathcal{S}_k, b_m \in \mathcal{E}_k\big)\big\} \\
\leq& K(1-\Delta^2/(16n^2))^M \leq (n/\Delta)(1-\Delta^2/(16n^2))^M \\
\leq& \exp\left\{\log\left(\frac{n}{\Delta}\right) - \frac{M\Delta^2}{16n^2}\right\}.
\end{aligned}$$

■

B.1.1 Univariate CUSUM Statistics

We introduce some notations for one dimensional change point detection and the corresponding CUSUM statistics. Let $\{z_i\}_{i=1}^n, \{f_i\}_{i=1}^n \subset \mathbb{R}$ be two univariate sequences. We will make the following assumptions.

Assumption 3 (Univariate mean change points) *Let* $\{\eta_k\}_{k=0}^{K+1} \subset \{0, \ldots, n\}$, *where* $\eta_0 = 0$ *and* $\eta_{K+1} = n$, *and*

$$f_{\eta_{k-1}+1} = f_{\eta_{k-1}+2} = \ldots = f_{\eta_k} \quad \textit{for all} \quad 1 \le k \le K+1,$$

Assume

$$\min_{k=1,\ldots,K+1} (\eta_k - \eta_{k-1}) \ge \Delta > 0,$$
$$0 < |f_{\eta_{k+1}} - f_{\eta_k}| := \kappa_k \textit{ for all } k = 1, \ldots, K.$$

We also have the corresponding CUSUM statistics over any generic interval $[s, e] \subset [1, T]$ defined as

$$\widetilde{Z}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} z_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} z_i,$$
$$\widetilde{f}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} f_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} f_i.$$

Throughout this section, all of our results are proven by regarding $\{Z_i\}_{i=1}^T$ and $\{f_i\}_{i=1}^T$ as two deterministic sequences. We will frequently assume that $\widetilde{f}_t^{s,e}$ is a good approximation of $\widetilde{Z}_t^{s,e}$ in ways that we will specify through appropriate assumptions.

Lemma 11 *Suppose Assumption 3 holds. Let* $[s_0, e_0]$ *be an interval with* $e_0 - s_0 \le C_R \Delta$ *and contain at lest one change point* η_r *such that*

$$\eta_{r-1} \le s_0 \le \eta_r \le \ldots \le \eta_{r+q} \le e_0 \le \eta_{r+q+1}, \quad q \ge 0.$$

Suppose that $\min\{\eta_{p'} - s_0, e_0 - \eta_{p'}\} \ge \Delta/16$ *for some* p' *and let* $\kappa_{\max}^{s,e} = \max\{\kappa_p : \min\{\eta_p - s_0, e_0 - \eta_p\} \ge \Delta/16\}$. *Let* $[s, e] \subset [s_0, e_0]$ *be any generic intervals. and*

$$b \in \arg\max_{s<t<e} |\widetilde{Z}_t^{s,e}|.$$

For some $c_1 > 0$, $\lambda > 0$ *and* $\delta > 0$, *suppose that*

$$|\widetilde{Z}_b^{s,e}| \ge c_1 \kappa_{\max}^{s,e} \sqrt{\Delta}, \tag{40}$$
$$\sup_{s+\zeta \le t \le e-\zeta} |\widetilde{Z}_t^{s,e} - \widetilde{f}_t^{s,e}| \le \xi, \quad \textit{and} \tag{41}$$
$$\sup_{s_1<t<e_1} \frac{1}{\sqrt{e_1 - s_1}} \left| \sum_{t=s_1+1}^{e_1} (z_t - f_t) \right| \le \xi \quad \textit{for every} \quad e_1 - s_1 \ge \zeta. \tag{42}$$

If there exists a sufficiently small $c_2 > 0$ *such that*

$$\xi \le c_2 \kappa_{\max}^{s,e} \sqrt{\Delta} \quad \textit{and} \quad \zeta \le c_2 \Delta, \tag{43}$$

then there exists a change point $\eta_k \in (s, e)$ *such that*

$$\min\{e - \eta_k, \eta_k - s\} > \Delta/4 \quad \textit{and} \quad |\eta_k - b| \le \min\{C_3 \xi^2 \kappa_k^{-2}, \zeta\}.$$

Proof This is Lemma 22 in Wang et al. (2017). ■

Lemma 12 *If $[s, e]$ contain two and only two change points η_r and η_{r+1}, then*

$$\sup_{s\le t\le e}\left|\widetilde{f}_t^{s,e}\right| \le \sqrt{e-\eta_{r+1}}\kappa_{r+1} + \sqrt{\eta_r - s}\kappa_r.$$

Proof This is Lemma 24 in Wang et al. (2021). ■

B.1.2 Projected CUSUM Statistics

Given a collection of deterministic vectors $\{u_m\}_{m=1}^M \in \mathbb{R}^p$, denote

$$z_i(u_m) = u_m^\top x_i y_i \in \mathbb{R} \quad \text{for } 1 \le m \le M \text{ and } 1 \le i \le n.$$

Let $\widetilde{Z}_t^{s,e}(u_m)$ denote the corresponding one-dimensional CUSUM statistics. That is

$$\widetilde{Z}_t^{s,e}(u_m) = \sqrt{\frac{e-t}{(e-s)(t-s)}}\sum_{i=s+1}^{t} z_i(u_m) - \sqrt{\frac{t-s}{(e-s)(e-t)}}\sum_{i=t+1}^{e} z_i(u_m).$$

Consider the following events

$$\begin{aligned}
&\mathcal{A}(\{y_i,x_i\}_{i=1}^n,\{u_m\}_{m=1}^M,\xi) \qquad (44)\\
&\quad=\left\{\sup_{1\le m\le M}\sup_{0\le s<t<e\le n}|\widetilde{Z}_t^{s,e}(u_m) - E(\widetilde{Z}_t^{s,e}(u_m))| \ge \xi, \min\{t-s, e-t\} \ge (\mathfrak{N}+1)\log(n)\right\};\\
&\mathcal{B}(\{y_i,x_i\}_{i=1}^n,\{u_m\}_{m=1}^M,\xi) \qquad (45)\\
&\quad=\left\{\sup_{1\le m\le M}\sup_{0\le s<t<e\le n}\left|\frac{\sum_{i=s+1}^{e}\{z_i(u_m) - E(z_i(u_m))\}}{\sqrt{e-s}}\right| \ge \xi, \min\{e-s\} \ge (\mathfrak{N}+1)\log(n)\right\}.
\end{aligned}$$

Lemma 13 (Deviation Bounds for Variance-Projected CUSUM statistics) *Suppose Assumption 1* **a** *holds. Let $\{u_m\}_{m=1}^M$ be a collection of vectors in $\mathbb{R}^p$ such that $\|u_m\|_2 = 1$ for all m. Then there exists an absolute constants C_1 and C_2 such that*

$$\begin{aligned}
&P\big[\mathcal{A}(\{y_i,x_i\}_{i=1}^n,\{u_m\}_{m=1}^M, C_1\sqrt{(\mathfrak{N}+1)\log(n)})\big] \ge 1 - CMn^{-3}, \text{ and}\\
&P\big[\mathcal{B}(\{y_i,x_i\}_{i=1}^n,\{u_m\}_{m=1}^M, C_1\sqrt{(\mathfrak{N}+1)\log(n)})\big] \ge 1 - CMn^{-3}.
\end{aligned}$$

Proof The deviation bounds can be established by standard sub-Exponential tail bounds. The analysis for the event $\mathcal{A}$ will be provided, as the analysis for event $\mathcal{B}$ is exactly the same.

Step 1. Note that

$$\widetilde{Z}_t^{s,e}(u_m) - E(\widetilde{Z}_t^{s,e}(u_m) = \sum_{i=s+1}^{e} b_i\Big[u_m^\top x_i y_i - E\{u_m^\top x_i y_i\}\Big],$$

where

$$b_i = \begin{cases} \sqrt{\frac{e-t}{(e-s)(t-s)}} & \text{when} s+1 \le i \le t, \\ -\sqrt{\frac{t-s}{(e-s)(e-t)}} & \text{when} t+1 \le i \le e. \end{cases}$$

Note that $u_m^\top x_i y_i = u_m^\top x_i (x_i^\top \beta_i^* + \varepsilon_i)$ where $x_i^\top \beta_i^* + \varepsilon_i$ is centered Gaussian with

$$Var(x_i^\top \beta_i^* + \varepsilon_i) = (\beta_i^*)^\top \Sigma \beta_i^* + \sigma_\varepsilon^2 \le \mathfrak{N} C_x + \sigma_\varepsilon^2,$$

and $u_m x_i$ is centered Gaussian with with

$$Var(u_m^\top x_i) = u_m^\top \Sigma u_m \le C_x,$$

where $\|u_m\|_2^2 = 1$ is used in the last inequality. So $u_m^\top x_i y_i$ is sub-Exponential with parameter $\mathfrak{N} C_x^2 + \sigma_\varepsilon^2 C_x$. In addition, note that

$$\sum_{i=s+1}^{e} b_i^2 = 1 \quad \text{and} \quad |b_i| \le (\mathfrak{N} \log(n))^{-1/2}.$$

So by sub-Exponential tail bound, it holds that

$$P\bigg(\bigg| \sum_{i=s+1}^{e} b_i \Big[u_m^\top x_i y_i - E\{u_m^\top x_i y_i\} \Big] \bigg| \ge \delta \bigg) \le 2 \exp\left(-c \min\left\{ \frac{\delta^2}{\mathfrak{N} C_x^2 + C_x \sigma_\varepsilon^2}, \frac{\delta \sqrt{(\mathfrak{N}+1)\log(n)}}{\sqrt{\mathfrak{N} C_x^2 + C_x \sigma_\varepsilon^2}} \right\} \right).$$

So by picking $\delta = C_\delta \sqrt{(\mathfrak{N}+1)\log(n)}$ for sufficiently large constant C_δ, it holds that with probability at most $1 - n^{-6}$,

$$\left| \widetilde{Z}_t^{s,e}(u_m) - E(\widetilde{Z}_t^{s,e}(u_m) \right| = \bigg| \sum_{i=s+1}^{e} b_i \Big[u_m^\top x_i y_i - E\{u_m^\top x_i y_i\} \Big] \bigg| \ge C_\delta \sqrt{(\mathfrak{N}+1)\log(n)}.$$

Since there are at most n^2 possible choice for $(s, e] \subset (0, n]$, a straightforward union bound argument shows that

$$P\big[\mathcal{A}(\{y_i, x_i\}_{i=1}^n, \{u_m\}_{m=1}^M, C_1 \sqrt{(\mathfrak{N}+1)\log(n)}) \big] \ge 1 - CMn^{-3}.$$

■

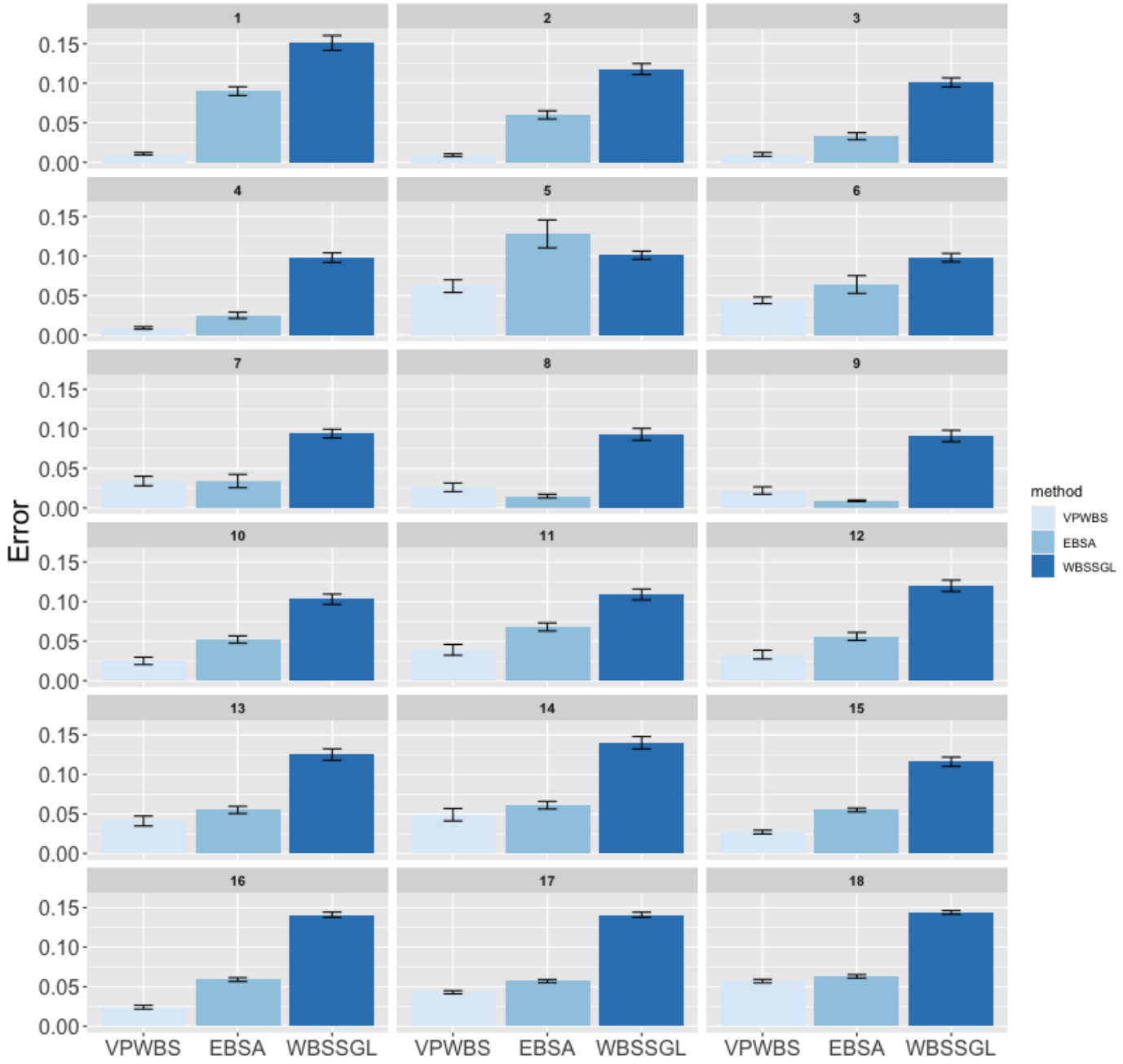

Figure 3: Bar plots for estimation results reported in Table 2. Plots 1-4 correspond to Setting (i) with $\kappa \in \sqrt{40} \cdot \{1, 1.2, 1.4, 1.6\}$. Plots 5-9 correspond to Setting (ii) with $n \in \{480, 560, 640, 720, 800\}$. Plots 10-14 correspond to Setting (iii) with $p \in \{80, 90, 100, 110, 120\}$. Plots 15-18 correspond to Setting (iv) with $s \in \{16, 20, 24, 28\}$.

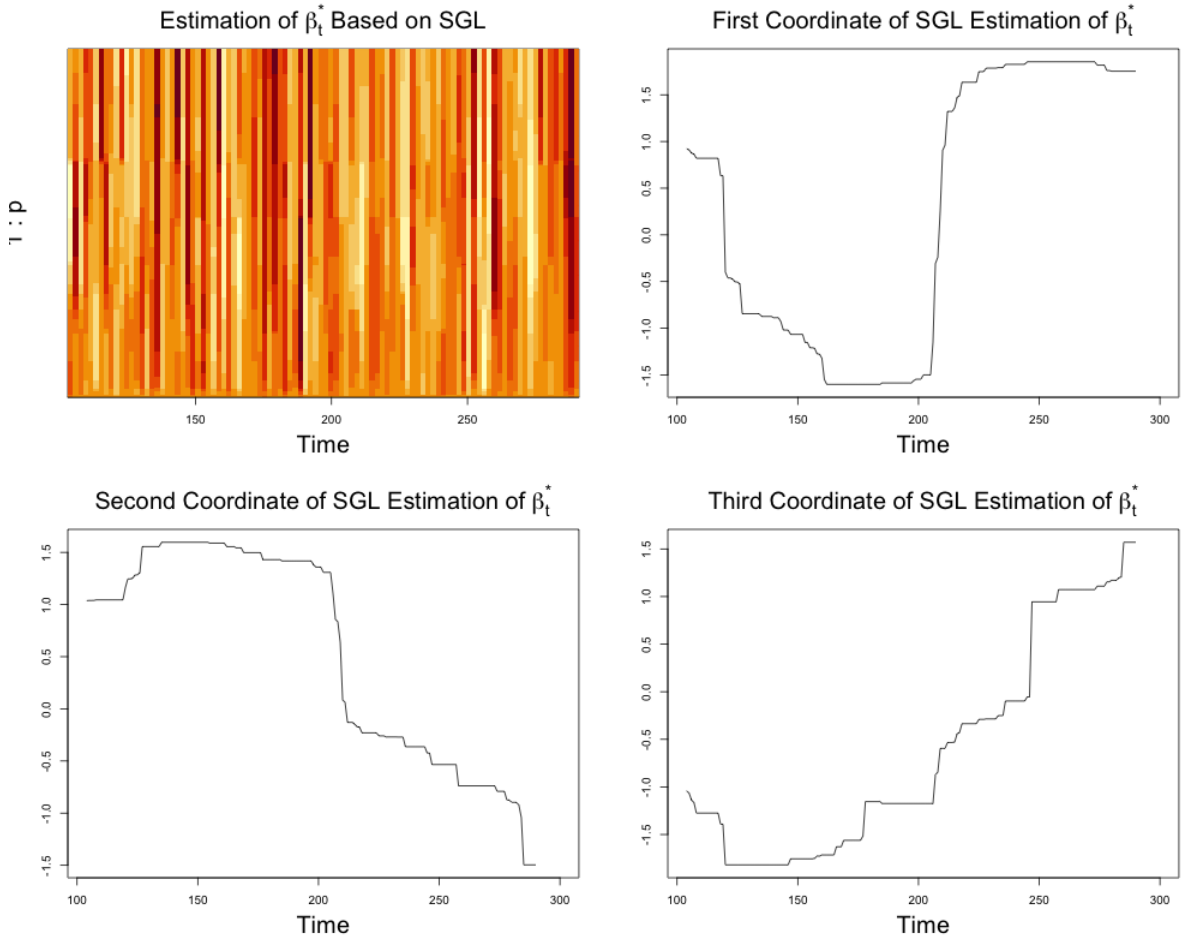

Figure 5: Plots of SGL estimator $\{\widehat{\beta}_t\}_{t=1}^n$ in (7). The data $\{x_t, y_t\}_{t=1}^n$ is the same as the one used to generate the illustration of VPWBS in Figure 1, where we have $n = 300, p = 100$ and that two change-points are at $\eta_1 = 100$ and $\eta_2 = 200$ with change size $\kappa = 1.6\sqrt{40}$. For better comparison with Figure 1, we plot the estimated $\widehat{\beta}_t$ for $t = 105, \cdots, 290$. The true coefficient $\{\beta_t^*\}_{t=105}^{290}$ contains a single change-point at $\eta_2 = 200$.