Detecting sparse change in regression coefficients in the presence of dense nuisance parameters

FENGNAN GAO

School of Mathematics and Statistics, University College Dublin, O'Brien Centre for Science, D04 V1W8, Dublin, Ireland

AND

TENGYAO WANG*

Department of Statistics, London School of Economics and Political Science, Houghton Street, WC2A 2AE, London, United Kingdom
*Corresponding author: t.wang59@lse.ac.uk

We introduce a new methodology '**charcoal**' for estimating the location of sparse changes in high-dimensional linear regression coefficients, without assuming that those coefficients are individually sparse. The procedure works by constructing different sketches (projections) of the design matrix at each time point so as to eliminate the possible dense nuisance parameters. The sequence of sketched design matrices is then compared against a single sketched response vector to form a sequence of test statistics whose behavior shows a surprising link to the well-known CUSUM statistics of univariate changepoint analysis. The procedure is computationally attractive, and strong theoretical guarantees are derived for its estimation accuracy. Simulations confirm that our methods perform well in extensive settings, and a real-world application to a large single-cell RNA sequencing dataset showcases the practical relevance.

Keywords: high-dimensional regression; changepoint; nuisance parameter; sparsity

1. Introduction

The past twenty years have witnessed rapid development of statistical methodologies for high-dimensional data sets, where the number of variables of interest is often of comparable or even larger order of magnitude than the number of observations. The most prominent example, perhaps, is the line of work on sparse linear regression, which started from the seminal work of Tibshirani [44], and was developed and generalized subsequently by many others (see, e.g. Bühlmann and van de Geer [8], Fan and Lv [16] for a general overview of this area). In many of these works, the primary focus was on how to exploit the sparsity of the regression coefficients for their successful estimation, and to achieve this, a homogeneous data generating mechanism was often assumed for simplicity of analysis.

However, it is usually unrealistic in large, high-dimensional data sets to assume that the data generating mechanism holds true throughout. In fact, heterogeneity is the norm rather than exception in Big Data applications. Several attempts have been made to handle data heterogeneity in high-dimensional linear models. For instance, Krishnamurthy et al. [31], Städler et al. [42], Yin et al. [50] considered the problem of learning mixture of sparse linear regression, where the regression coefficient vector is sampled from a small set of sparse regression coefficients. When observations have a temporal structure, one common way to handle heterogeneity is to break the sequence of observations into shorter time segments on which the data are more homogeneous. This line of thinking is the driving force behind the recent revival of interest in changepoint analysis, which dates back to the early work of Page [38], but

has seen growing interest recently in high-dimensional settings, see e.g., Bai [2], Chen et al. [9], Cho [10], Cho and Fryzlewicz [11], Enikeeva and Klopp [15], Follain et al. [17], Horváth and Hušková [24], Jirak [25], Lévy-Leduc and Roueff [34], Liu et al. [35], Wang and Samworth [48], Zhang et al. [52].

In our linear regression setting, such a changepoint setup means that the sequence of regression coefficient vectors has a piecewise constant structure. More precisely, for an unknown sequence of changepoints $0 < z_1 < \cdots < z_\nu < n$ (for notational convenience, we also define $z_0 := 0$ and $z_{\nu+1} := n$) and an unknown sequence of a regression coefficients $(\beta^{(r)} : 1 \leq r \leq \nu + 1)$, we assume that the data $(x_t, y_t) \in \mathbb{R}^p \times \mathbb{R}$, $1 \leq t \leq n$ are generated according to the following model

$$y_t = x_t^\top \beta_t + \varepsilon_t, \quad \text{where } \beta_t = \beta^{(r)} \text{ for } z_{r-1} < t \leq z_r, \ 1 \leq r \leq \nu + 1, \tag{1.1}$$

and $(\varepsilon_t)_{1 \leq t \leq n}$ are the observational errors distributed as $N(0, \sigma^2 I_n)$ conditionally independent of $(x_t)_{1 \leq t \leq n}$. The goal is to locate the changepoints $z_1, \ldots, z_\nu$ upon observing the response vector $Y = (y_1, \ldots, y_n)^\top$ and the design matrix $X = (x_1, \ldots, x_n)^\top$.

Classically, when the dimension p is far smaller than n, Bai [1], Bai and Perron [3], Julious [26] showed that a least-square-based approach works well in the above changepoint problem, which is equivalent to maximum-likelihood estimations under Gaussianity assumptions. Specifically, for a given ν, the maximum likelihood estimator finds the optimal partition of $\{1, \ldots, n\}$ into $\nu + 1$ segments such that residual sum of squares within each segment is minimized. The least-square (maximum-likelihood) fit from different choices of ν can then be compared using for instance the Bayesian Information Criterion (BIC) to choose the best ν, which is often solved algorithmically via dynamic programming.

In the high-dimensional setting, the above least-square approach no longer works. Several works have appeared to analyze such regression changepoint problems in the high-dimensional context, see for instance Kaul et al. [28], Lee et al. [32], Rinaldo et al. [40], Wang et al. [47] and references therein. However, in addition to the modeling assumption in (1.1), these works also impose the additional assumption that all regression coefficients $(\beta^{(r)} : 1 \leq r \leq \nu + 1)$ are individually sparse. Given a hypothesized set of changepoints, this additional assumption allows them to form estimators of $\beta^{(r)}$, $1 \leq r \leq \nu + 1$, which are in turn used to form goodness-of-fit statistics for the set of hypothesized changepoints.

A major difference between this work and the aforementioned existing line of works is that we do not assume that the regression coefficients within different stationary segments are individually sparse. Instead, we make the less stringent assumption that the difference in the regression coefficient vectors before and after each change, i.e., $\theta^{(r)} := (\beta^{(r+1)} - \beta^{(r)})/2$, are sparse in the sense that $\|\theta^{(r)}\|_0 \leq k$, for $r = 1, \ldots, \nu$. We would argue that this is a more natural assumption, since it is the change in the regression coefficients, rather than the pre- and post-change coefficients themselves, that is the quantity of interest in this statistical problem. Practically, the assumption that all regression coefficients are sparse can be violated in applications. For instance, Kraft and Hunter [30] argued that in genetic studies, "many, rather than few, variant risk alleles are responsible for the majority of the inherited risk of each common disease", leading to non-sparse regression coefficients. However, in such examples, the task of detecting sparse changes in these regression coefficients over time can still be of interest in, e.g., identifying different development stages in gene regulatory networks in species [23]. Furthermore, our 'sparsity-in-change' assumption is also more in line with the assumptions made in the high-dimensional change-in-mean problem [see, e.g. 11, 25, 48], where the pre- and post-change mean vectors are regarded as nuisance parameters and sparsity assumptions only need to be placed on vectors of changes for successful detection and estimation of the changepoints. In other words, our paper falls into the broader

category of statistical inference in the presence of nuisance parameter, where the goal is to extract information of interest, e.g., the differences in regression coefficients $\theta^{(r)}$'s and associated changepoints z_r's, which is often challenging to be disentangled from complex nuisance parameters, e.g., the exact regression coefficients $\beta^{(r)}$'s. The literature of statistical inference in the presence of nuisance parameter has a long history [37], but has received renewed interest in recent years, especially in high-dimensional settings [5, 6]. We aim to solve our problem without the simplifying assumption that the nuisance parameters are estimable in some way. Since the initial draft of this paper, we have also become aware of a recent work [13] that addresses the setting with dense regression coefficients.

Allowing for dense pre- and post-change regression coefficients makes the changepoint estimation problem considerably more challenging. In particular, the general strategy employed by existing works on high-dimensional regression changepoints that relies on forming good estimators of $(\beta^{(r)} : r \in \{1,\dots,\nu\})$ will unlikely be successful here. Our first contribution in this paper is to propose a novel methodology, which we call **charcoal** (<u>cha</u>ngepoint in <u>r</u>egression via a <u>co</u>mplementary-sketching <u>al</u>gorithm), and works by forming a projected response vector and a sequence of projected design matrices to eliminate the dense nuisance parameter. For simplicity of exposition, we consider the single changepoint scenario, where $\beta_i = \beta^{(1)}\mathbb{1}_{\{i \le z_1\}} + \beta^{(2)}\mathbb{1}_{\{i > z_1\}}$. Under the hypothesis that the true change takes place at time t, we have

$$\begin{cases} Y_{(0,t]} = X_{(0,t]}\beta^{(1)} + \varepsilon_{(0,t]}, \\ Y_{(t,n]} = X_{(t,n]}\beta^{(2)} + \varepsilon_{(t,n]}, \end{cases} \tag{1.2}$$

where the subscript $(0,t]$ indicates the concatenation of relating quantities on $(1,\dots,t)$ and $(t,n]$ that of $(t+1,\dots,n)$. We may think of (1.2) as a two-sample problem with different regression coefficients before and after t. We assume throughout the paper that $n > p$ for otherwise it is impossible to estimate the change when both pre- and post-change parameters are dense (see further discussion at the beginning of Section 3). By invoking the complementary sketching method of Gao and Wang [21], we can find matrices $A_{(0,t]} \in \mathbb{R}^{t \times (n-p)}$ and $A_{(t,n]} \in \mathbb{R}^{(n-t) \times (n-p)}$ such that $(A_{(0,t]}^\top, A_{(t,n]}^\top)^\top$ has orthogonal columns spanning the orthogonal complement of the range of X. By forming the projected design matrix $W_t := A_{(0,t]}^\top X_{(0,t]} - A_{(t,n]}^\top X_{(t,n]}$ and the projected response $Z := A_{(0,t]}^\top Y_{(0,t]} + A_{(t,n]}^\top Y_{(t,n]}$, we can eliminate the possibly dense nuisance parameter $\zeta := (\beta^{(2)} + \beta^{(1)})/2$ and conduct tests on $\theta^{(1)} = (\beta^{(2)} - \beta^{(1)})/2$ is zero against that it is non-zero and sparse.

In light of the true changepoint at time z_1, the hypothesized model (1.2) is only correctly specified when $t = z_1$. The further t is away from z_1, the less different the two samples $(X_{(0,t]}, Y_{(0,t]})$ and $(X_{(t,n]}, Y_{(t,n]})$ are, since one of the samples will be further contaminated by the data points assigned to the wrong segment by the hypothesized changepoint. Intuitively, we would expect the aforementioned two-sample test statistics to peak around $t = z_1$, which can thus be used to estimate the location of the single changepoint. Unfortunately, while good for testing, these two-sample test statistics have variances too large for accurate changepoint estimation. Nevertheless, the general idea of using complementary sketching to eliminate nuisance parameters is valid. We introduce in Section 2 several alternative statistics based on the sketched design W_t and response Z that do lead to good changepoint estimation performance. In particular, we will show in Section 3 that a variant of the **charcoal** procedure achieves a rate of convergence of order $k^{1/2}/(n^{1/2}\|\theta^{(1)}\|_2)$, up to logarithmic factors. In the course of theoretical investigations, we have developed new crucial results and proofs in understanding the asymptotic behavior of the sketched design matrices by generalizing existing matrix-variate Beta distribution to rank-deficient cases (Lemma 10 and Corollary 11) and extended sub-Gaussian bounds of Beta random variables to the matrix variate case (Lemma 13). Both rely on novel proof techniques and may be of independent interest. Particularly, the conditioning argument used in the proof of Lemma 13 seems rarely

seen in the literature and may be an alternative way to establish concentration in similar scenarios. Last but not least, for any generic multiple-changepoint estimation procedures obtained by combining the narrowest-over-threshold methodology and a viable single-changepoint method, we provide a general theorem 5 to establish rates of convergence, which may be useful in many other applications.

1.1. *Outline of the paper*

We present the methodology in detail in Section 2, including several algorithms that all use the complementary sketching idea. Section 3 provides theoretical performance guarantees to the slight variants of those proposed in Section 2. In the first part of Section 4, we conduct numerical experiments on the **charcoal** methodology over a comprehensive range of settings for both single and multiple changepoint estimation tasks and compare our methods with other changepoint estimation methods in the high-dimensional linear regression context. In the second part of Section 4, we study a real data example to identify changes for each gene in terms of its interaction with other genes in the gene regulatory network across various development stages of T cells. The proofs of the main results and of the ancillary results are collected in the supplementary materials.

1.2. *Notation*

For a positive integer p, $[p] = \{1, \ldots, p\}$ consists of all positive integers not exceeding p. For vector $v = (v_1, \ldots, v_p)^\top$, $\mathrm{diag}(v)$ is a $p \times p$ matrix such that $(\mathrm{diag}(v))_{i,j} = \mathbb{1}_{\{i=j\}} v_i$ for $i, j \in [p]$. We follow the usual definitions of $\|v\|_0 = \sum_{i \in [p]} \mathbb{1}_{\{v_i \neq 0\}}$, $\|v\|_2 = (\sum_{i \in [p]} v_i^2)^{1/2}$, $\|v\|_1 = \sum_{i \in [p]} |v_i|$ and $\|v\|_\infty = \max_{i \in [p]} |v_i|$.

Given a matrix $A \in \mathbb{R}^{n \times m}$, we make it a convention that $A = (A_{i,j})_{i \in [n], j \in [m]} = [A_1 \mid \cdots \mid A_m] = (a_1, \ldots, a_n)^\top$, where a_i is the transpose of the ith row of A and A_j is the jth column of A. Given any set $S \subseteq \mathbb{R}$, we write A_S to be the submatrix of A with row indices in S. For instance, given positive integers s, t such that $1 \leq s < t \leq n$, $A_{(s,t]} := (a_{s+1}, \ldots, a_t)^\top$. We define the usual norms for A as follows $\|A\|_{\mathrm{op}} := \sup_{v \in \mathbb{R}^m : \|v\|_2 = 1} \|Av\|_2$ and $\|A\|_{\max} := \max_{i \in [n], j \in [m]} |A_{i,j}|$. Assuming $n = m$ in A, $\mathrm{diag}(A)$ is an $n \times n$ matrix such that $(\mathrm{diag}(A))_{i,j} := \mathbb{1}_{i=j} A_{i,j}$ for $i \in [n]$ and $\mathrm{tr}(A) := \sum_{i \in [n]} A_{i,i}$.

For $n \geq m$, $\mathbb{O}^{n \times m} := \{O \in \mathbb{R}^{n \times m} : O^\top O = I_m\}$. We define $\mathscr{S}^{p-1} := \{v \in \mathbb{R}^p : \|v\|_2 = 1\}$ and the k-sparse unit ball as $B_0(k) := \{v \in \mathbb{R}^p : \|v\|_2 \leq 1, \|v\|_0 \leq k\}$.

2. Methodology

In this section, we describe in detail our **charcoal** algorithm for identifying the changepoints in the problem setup of (1.1).

2.1. *Single changepoint estimation*

We start by focusing on the setting of a single changepoint estimation, i.e., $v = 1$, which captures the essence of the difficulty of this problem. For simplicity, we denote $z := z_1$ for the location of the only changepoint and write $m := n - p$. The main idea is to use data-driven projections to sketch the design matrix and the response vector to eliminate the effect of the nuisance parameters.

Recall the data generating model (1.1). At each time point $t \in [n-1]$, we perform a two-sample test for the equality of regression coefficients before and after t using data points $(x_i, y_i)_{i=1}^t$ and $(x_i, y_i)_{i=t+1}^n$ respectively. Motivated by [21], this can be achieved by constructing a matrix $A \in \mathbb{O}^{n \times m}$ whose columns

span the orthogonal complement of the column space of X. We then define for any $t \in [n-1]$

$$W_t := A_{(0,t]}^\top X_{(0,t]} - A_{(t,n]}^\top X_{(t,n]} = 2A_{(0,t]}^\top X_{(0,t]} \in \mathbb{R}^{m \times p},$$

$$Z := A_{(0,t]}^\top Y_{(0,t]} + A_{(t,n]}^\top Y_{(t,n]} = A^\top Y \in \mathbb{R}^m.$$

We define $\theta = (\beta^{(1)} - \beta^{(2)})/2$, $\zeta = (\beta^{(1)} + \beta^{(2)})/2$ and $\xi = A^\top \varepsilon \sim N_m(0, \sigma^2 I_m)$. By the model construction, we have

$$\begin{aligned} Z &= A_{(0,z]}^\top Y_{(0,z]} + A_{(z,n]}^\top Y_{(z,n]} = A_{(0,z]}^\top (X_{(0,z]}\beta^{(1)} + \varepsilon_{(0,z]}) + A_{(z,n]}^\top (X_{(z,n]}\beta^{(2)} + \varepsilon_{(z,n]}) \\ &= A_{(0,z]}^\top X_{(0,z]}(\theta + \zeta) - A_{(z,n]}^\top X_{(z,n]}(\theta - \zeta) + \xi = W_z\theta + \xi, \end{aligned} \tag{2.1}$$

whence we have eliminated the nuisance parameter ζ, and obtain the sketched data in the form of $(Z, (W_t)_{t \in [n-1]})$. By (2.1) and the sparsity assumption on θ, Z can be approximated by a sparse linear combination of the columns of W_z. Therefore, the changepoint estimation problem is reduced to finding t such that W_t forms a 'best' sparse linear approximation to Z.

As mentioned in the introduction, a naive way to achieve this would be based on the two-sample test statistics introduced in Gao and Wang [21]. Specifically, let $Q = (Q_1, \ldots, Q_{n-1})^\top$ be defined such that

$$Q_t := \{\operatorname{diag}(W_t^\top W_t)\}^{-1/2} W_t^\top Z.$$

We view Q_t as the vector of the correlations between columns of W_t and Z, where we naturally seek to find the time point t such that such correlations are as large as possible. To take into account of possible observational errors, we first remove small entries of Q_t via an entrywise hard-thresholding operation $\operatorname{hard}(Q_t, \lambda)$ for $\operatorname{hard}(v, \lambda) : (v_i)_{i=1}^p \mapsto (v_i \mathbb{1}_{\{|v_i| \geq \lambda\}})_{i=1}^p$, where the threshold level λ is a tuning parameter. This allows us to estimate the location of the changepoint via $\hat{z}^{\text{hard}} := \arg\max_{t \in [n-1]} \|\operatorname{hard}(Q_t, \lambda)\|_2$. Note that $\|\operatorname{hard}(Q_t, \lambda)\|_2$ is the statistic from Gao and Wang [21] to test whether the two samples $(X_{(0,t]}, Y_{(0,t]})$ and $(X_{(t,n]}, Y_{(t,n]})$ have the same regression coefficient against the alternative that there is a sparse difference. As is argued before, we expect that two-sample testing statistics gives the strongest signal against the null of no change at $t = z$ — the only point where the two-sample problem is correctly specified.

However, the changepoint estimator $\hat{z}^{\text{hard}}$ is less than ideal in practice, as the discontinuity of the hard-thresholding function creates large variabilities in the test statistics. Moreover, the theoretical guarantees given in Gao and Wang [21] becomes increasingly inapplicable for test statistics away from the true changepoint as one of the two samples contains a mixture of data both before and after the change. Coupled with the fact that W_t has large variance when t is close to the boundary, Q_t may have a large number of entries above the hard-thresholding level $\lambda = 2\sqrt{\log p}$ as recommended in Gao and Wang [21]. Empirically, this is evidenced by high variance of the test statistics near the two endpoints of the interval for changepoint detection, as shown in Figure 1. Quite often, this boundary effect may overwhelm the main signal near the true changepoint, leading to a spurious changepoint being estimated near the boundary. One way to alleviate the instability problem of $\hat{z}^{\text{hard}}$ is to replace the hard-thresholding in $\hat{z}^{\text{hard}}$ by a soft-thresholding operation on each entry of Q. The changepoint is then estimated by $\hat{z}^{\text{soft}} = \arg\max_{t \in [n-1]} \|\operatorname{soft}(Q_t, \lambda)\|_2$, where $\operatorname{soft}(v, \lambda) : (v_i)_{i=1}^p \mapsto (\operatorname{sign}(v_i) \max(|v_i| - \lambda, 0))_{i=1}^p$ with a tuning parameter λ. The continuity of the soft-thresholding function reduces the variance in the test statistics, and in the ensuing changepoint estimator. However, as also shown in Figure 1, the sequence
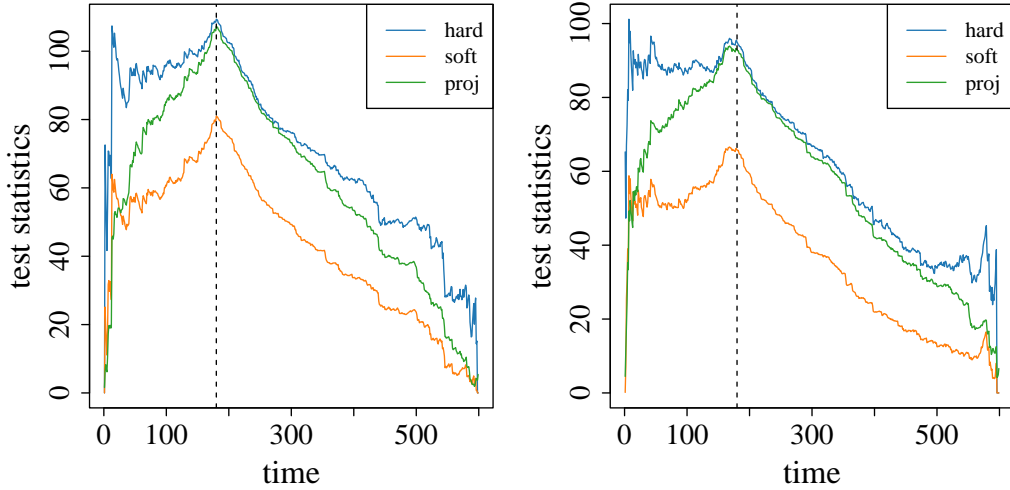
FIG. 1. Visualization of different test statistics for changepoint estimation. We compare the hard-thresholded test statistics $\|\operatorname{hard}(Q_t, \lambda)\|_2$ used to estimate $\hat{z}^{\mathrm{hard}}$ (denoted by `hard`), the soft-thresholded test statistics $\|\operatorname{soft}(Q_t, \lambda)\|_2$ used to estimate $\hat{z}^{\mathrm{soft}}$ (denoted by `soft`) and the projected statistics $|\hat{v}^\top Q_t|$ in Algorithm 1 (denoted by `proj`) over two random realizations. Here, $n = 600$, $p = 200$, $\|\beta_1 - \beta_2\|_0 = 10$, $\|\beta_1 - \beta_2\|_2 = 8$, and the true change takes place at $z = 180$, as indicated by the dashed lines. In both panels, we observe that both $\|\operatorname{hard}(Q_t, \lambda)\|_2$ and $\|\operatorname{soft}(Q_t, \lambda)\|_2$ exhibit relatively strong boundary effect.

of test statistics $(\|\operatorname{soft}(Q_t, \lambda)\|_2)_{t \in [n-1]}$ could still exhibit undesirably large, although less so than $(\|\operatorname{hard}(Q_t, \lambda)\|_2)_{t \in [n-1]}$, variations when t is close to the boundary.

To avoid such boundary effect, we propose instead to aggregate the test statistics $(Q_t)_t$ via a projection-based approach. The key insight here is that, away from the boundary, the matrix $Q = (Q_1, \ldots, Q_{n-1})$ can be well-approximated by a rank-one matrix whose leading left singular vector is proportional to θ. Hence, by first estimating $\theta/\|\theta\|_2$ via the leading left singular vector of $\operatorname{soft}(Q, \lambda)$, we can aggregate each vector of correlation Q_t along the direction of $\hat{v}$ and estimate the changepoint by $\hat{z} := \arg\max_{t \in [\alpha n, (1-\alpha)n]} |\hat{v}^\top Q_t|$. This approach is summarized in Algorithm 1. We allow Algorithm 1 to output both the changepoint estimator $\hat{z}$ and a test statistic $H_{\max}$, which can be used in our multiple changepoint algorithm to determine if an estimated changepoint is spurious.

To compute the sequence $(Q_t)_{t \in [n-1]}$ in Algorithm 1, observe that the same A and Z can be used for all $t \in [n-1]$ and hence only need to be computed once. It is worth noting that we exploit the structures of the sketched designs $(W_t)_t$ to greatly simplify their computations. Recall that $a_t \in \mathbb{R}^{n-p}$ is the tth row vector of A, i.e., $A = (a_1, \ldots, a_n)^\top$. W_t are computed via the simple iterative scheme $W_0 = 0$ and $W_t = W_{t-1} + 2a_t x_t^\top$ for $t \in [n-1]$. As a common measure, we introduce the burn-in parameter α so that we forgo the possibilities of having changepoints in $(0, \alpha n) \cup ((1-\alpha)n, n)$.

While the main focus of our current work is the changepoint estimation problem, we remark that Algorithm 1 can be easily adapted to test the existence of a single changepoint in the sequence of regression coefficients. Specifically, we can construct the test

$$\psi_{\alpha, \lambda, T} := \mathbb{1}\left\{ \max_{t \in [\alpha n, (1-\alpha)n]} \|\operatorname{soft}(Q_t, \lambda)\| \geq T \right\}, \qquad (2.2)$$

where T is some appropriate threshold.

Algorithm 1: Pseudocode for changepoint estimation

Input: $X \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^n$ satisfying $n - p > 0$, a soft threshold level $\lambda \geq 0$, burn-in parameter $\alpha \geq 0$

1   Set $m \leftarrow n - p$
2   Form $A \in \mathbb{O}^{n \times m}$ with columns orthogonal to the column space of X
3   Compute $Z \leftarrow A^\top Y$
4   Set $W_0 = \mathbf{0}_{m \times p}$
5   **for** $1 \leq t \leq n - 1$ **do**
6   Compute $W_t \leftarrow W_{t-1} + 2a_t x_t^\top$
7   Compute $Q_t = \{\mathrm{diag}(W_t^\top W_t)\}^{-1/2} W_t^\top Z$
8   **end**
9   Form $Q := (Q_{\lfloor \alpha n \rfloor}, \ldots, Q_{\lceil (1-\alpha)n \rceil})^\top$
10   Compute $\hat{v} \leftarrow$ the leading left singular vector of $\mathrm{soft}(Q, \lambda)$

Output: $\hat{z} := \arg\max_{\alpha n \leq t \leq (1-\alpha)n} |\hat{v}^\top Q_t|$ and $H_{\max} := \max_{\alpha n \leq t \leq (1-\alpha)n} \|\mathrm{soft}(Q_t, \lambda)\|$.

Finally, we mention that another natural approach to find the W_t whose columns form the best sparse linear approximation of $Z = W_z \theta + \xi$ is to fit a sparse linear model by regressing Z against W_t and compare the goodness-of-fit across t via the Bayesian Information Criterion (BIC). We choose the BIC for the model selection purpose, though it is conceivably straightforward to apply any other model selection criteria. The pseudocode for this procedure is given in Algorithm 2. Specifically, for appropriately chosen $(\lambda_t)_{t \in [n-1]}$, we compute first the Lasso solutions in Step 4 and then the corresponding BICs in Step 5. In practice, the sequence of regularizing parameters $(\lambda_t)_t$ may be chosen via cross-validation for each t.

Algorithm 2: Pseudocode for changepoint estimation with Lasso with BIC

Input: $X \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^n$ satisfying $n > p$, $\alpha > 0$ and a sequence $(\lambda_t)_{t \in [n-1]}$

1   Follow Algorithm 1 until line 4
2   **for** $1 \leq t \leq n - 1$ **do**
3   Compute $W_t \leftarrow W_{t-1} + 2a_t x_t^\top$
4   Compute the Lasso estimator $\hat{\theta}_t \leftarrow \arg\min_{v \in \mathbb{R}^p} \left\{ \frac{1}{2m} \|Z - W_t v\|_2^2 + \lambda_t \|v\|_1 \right\}$
5   Compute $H_t \leftarrow -(\|Z - W_t \hat{\theta}_t\|_2^2 + \|\hat{\theta}_t\|_0 \log m)$
6   **end**

Output: $\hat{z} := \arg\max_{\alpha n \leq t \leq (1-\alpha)n} H_t$

Algorithm 1 has a computational complexity of $O(n^2 p)$, with the most computationally intensive step being its Step 2 to form the sketching matrix A (e.g. via a QR decomposition). For Algorithm 2, each Lasso step has a computational cost of $O(k^2 n)$ [14], leading to an overall computational complexity of $O(n^2(p + k^2))$. It is remarkable that for sparse signals ($k = O(\sqrt{p})$), the changepoint algorithms we proposed here has essentially the same computational complexity as the complementary-sketching-based two-sample test [21] for any hypothesized changepoint location t.

2.2. *Multiple changepoint estimation*

The single changepoint estimation procedure described above can be combined with a generic top-down multiple changepoint estimation method, such as binary segmentation [46], wild binary segmentation [18] and its variants [e.g., 4, 20, 29], to iteratively identify multiple changepoints. For concreteness, we describe an approach combining Algorithm 1 with the narrowest-over-threshold method of Baranowski et al. [4]. Algorithm 3 is a slight generalization of Baranowski et al. [4, Algorithm 1]. It takes as input a single changepoint estimation procedure $\hat{z}$ and a testing procedure ψ. When the data $D_1, \ldots, D_n$ are the covariate-response pair $(X_i, Y_i)_{i \in [n]}$, we may apply Algorithm 1 or 2 to obtain $\hat{z}$ and define $\psi(X, Y) := \mathbb{1}_{\{H_{\max} > T\}}$ for some T using the output $H_{\max}$ of Algorithm 1. However, note that both Algorithms 1 and 2 require the number of observations to be larger than the dimension for the complementary sketching to work. If this is not satisfied, we simply define $\hat{z}(X, Y) := 0$ and $\psi(X, Y) := 0$.

Essentially, in Algorithm 3, we generate multiple intervals and run the single changepoint algorithm on each interval to obtain candidate changepoint estimates and test results. We choose the candidate changepoint associated with the narrowest interval for which the test rejects the null, and add that to the set of estimated changepoints. We then segment the data at this estimated changepoint, and repeat the above process recursively on the data to the left and right segments, using only intervals lying completely within each segment. The process terminates when none of the tests reject the null. Furthermore, for practical reasons, we recommend combining Algorithm 3 with some second-stage refinements, for which we discuss in more details in Section 4.4.

Algorithm 3: Pseudocode for multiple changepoint estimation

Input: Data $D_1, \ldots, D_n$, number of intervals M, burn-in parameter $\varpi > 0$, single changepoint estimation procedure $\hat{z}$ and a single changepoint testing procedure ψ

1 Set $\hat{Z} \leftarrow \emptyset$ and generate a set of M intervals $\mathscr{M} := \{(s_1, e_1], \ldots, (s_M, e_M]\}$ independently and uniformly from $\{(a, b] : 0 \le a < b \le n\}$.

2 Run $\mathrm{NOT}(0, n)$ where NOT is defined below.

3 Let $\hat{v} \leftarrow |\hat{Z}|$ and sort elements of $\hat{Z}$ in increasing order to yield $\hat{z}_1 < \cdots < \hat{z}_{\hat{v}}$.

Output: $\hat{z}_1, \ldots, \hat{z}_{\hat{v}}$

4 **Function** NOT (s, e)

5   Set $\mathscr{R}^{(s,e)} \leftarrow \{m : (s_m, e_m] \subseteq (s, e], \psi(D_{(s_m + n\varpi, e_m - n\varpi]}) = 1\}$

6   **if** $\mathscr{R}^{(s,e)} \neq \emptyset$ **then**

7          $m_0 \leftarrow \arg\min_{m \in \mathscr{R}^{(s,e)}} (e_m - s_m)$

8          $b \leftarrow s_{m_0} + \hat{z}(D_{(s_m, e_m]})$

9          $\hat{Z} \leftarrow \hat{Z} \cup \{b\}$

10   NOT (s, b)

11   NOT (b, e)

12   **end**

13 **end**

---

3. Theoretical guarantees

In this section, we establish theoretical guarantees for the changepoint procedures proposed in Section 2. We start by focusing on the single changepoint estimation problem. For simplicity of analysis, we will assume that the noise variance σ^2 is known in this section, which by scale invariance can be further assumed to be equal to 1. We discuss practical aspects of estimating σ^2 in Section 4.1. We first present two conditions, which we will need to establish the results in this section.

Condition 1 *All entries of X are independent standard normals.*

Condition 2 *n, z, p satisfy that $n > p$, $z/n \to \tau \in (0, 1)$ and $(n - p)/n \to \eta \in (0, 1)$ as $\min(z, n, p) \to \infty$.*

The design Condition 1 requires that the rows of the design matrix $X = (x_1, \ldots, x_n)^\top$ follow the isotropic Gaussian distribution. Condition 2 specifies the asymptotic regime we work in. Note that the assumption $n > p$ is necessary, since otherwise, even if z is known, it is impossible to test $\theta = 0$ against the alternative $\theta = \theta^*$ for any fixed $\theta^* \in \mathbb{R}^p$ (see Lemma 8). The key ingredient of our theoretical analysis is the following proposition, which shows that $W_t^\top W_z$ is close to a multiple of identity in terms of their actions on sparse vectors. We impose both conditions 1 and 2 only to enable the application of the existing random matrix theory on the limiting spectral measure of matrix-variate Beta distributions in the proof of Proposition 1. In principle, even if the above conditions are violated, the theoretical results in the rest of the section hold for any data (X, Y) such that (3.1) is satisfied. In particular, we remark that the empirical study in Section 4.3 has demonstrated that our methodology exhibits good finite-sample performance even when the above two conditions do not hold.

Proposition 1 *Suppose that Conditions 1 and 2 are satisfied and define*

$$g(t; z) := \begin{cases} 4t(n-z)(n-p)/n^2 & \text{if } 1 \leq t \leq z, \\ 4z(n-t)(n-p)/n^2 & \text{if } z < t \leq n-1. \end{cases}$$

There exists a constant $C_{\tau, \eta} > 0$, depending only on τ and η such that with probability 1, for any fixed $v \in \mathscr{S}^{p-1}$ and $\ell \in [p]$, we have for all but finitely many p's that

$$\sup_{t \in [n-1]} \sup_{u \in B_0(\ell)} u^\top \left\{ W_t^\top W_z - g(t; z) I_p \right\} v \leq C_{\tau, \eta} \sqrt{\ell n \log p}. \tag{3.1}$$

Note that we suppress the dependence on n and t in the notation of $g(t; z)$. Taking $\ell = 1$ in the above proposition, we would expect $\text{diag}(W_t^\top W_t)$ to concentrate around $g(t; t) I_p = 4t(n-t)(n-p)n^{-2} I_p$ for each t. This would allow us to approximate the test statistics $Q_t = \{\text{diag}(W_t^\top W_t)\}^{-1/2}(W_t^\top W_z \theta + W_t^\top \xi)$ for a fixed t. However, due to a lack of non-asymptotic probabilistic bounds in random matrix theory on the convergence of the spectral measures of matrix-variate Beta random matrices, we are unable to establish the said convergence of $\text{diag}(W_t^\top W_t)$ uniformly over $t \in [n-1]$. As such, we instead show the theoretical results for a slightly modified variant of Algorithm 1, where we replace the definition of Q_t by

$$Q_t := \sqrt{\frac{n}{t(n-t)}} W_t^\top Z. \tag{3.2}$$

We will henceforth refer to the above variant of Algorithm 1 as Algorithm 1'. It is worth noting that the latter is merely a proof device, and in practice we always recommend applying Algorithm 1. Empirically,

the primed variant has a slightly worse but comparable estimation accuracies than Algorithm 1, which can be seen in Table 1 in Section 4.1.

With this alternative choice of Q_t, after removing the perturbation introduced by observational errors ξ's with an appropriate soft-thresholding tuning parameter λ, we expect

$$H_t := \|\text{soft}(Q_t, \lambda)\|_2 \approx h_t := \|\theta\|_2 \gamma_t,$$

where γ_t is defined by

$$\gamma_t := g(t; z) \sqrt{\frac{n}{t(n-t)}} = \begin{cases} \frac{4(n-p)}{n} \sqrt{\frac{t}{n(n-t)}} (n-z) & \text{if } 1 \leq t \leq z, \\ \frac{4(n-p)}{n} \sqrt{\frac{n-t}{nt}} z & \text{if } z < t \leq n-1. \end{cases} \tag{3.3}$$

Interestingly, h_t (or γ_t) is proportional to the CUSUM statistic in the univariate change-in-mean problem, whence h_t attains its maximum at $t = z$ [cf. 48, Equation (10)]. By exploiting the above observation, we establish in Theorem 2 that the testing procedure (2.2) is capable of determining whether a (single) changepoint is present in the regression data, as mentioned in Section 2.

Recall $\zeta := (\beta^{(1)} + \beta^{(2)})/2$ and $\theta := (\beta^{(1)} - \beta^{(2)})/2$, where we regard ζ as a possibly dense nuisance parameter and wish to estimate the changepoint only assuming the difference parameter θ is sparse, i.e., $\|\theta\|_0 \leq k$ for some unknown but fixed k typically much smaller than p.

Theorem 2 *Assume Conditions 1 and 2 and that data (X, Y) are generated according to (1.1) with $\nu = 1$. Suppose that $\|\theta\|_0 \leq k$ satisfies $(k \log p)/n \to 0$ and that $\min(\tau, 1 - \tau) \geq \alpha$ for some known α. There exists $c_{\tau,\eta,\alpha}, C_{\tau,\eta,\alpha}, c'_{\tau,\eta,\alpha} > 0$, depending only on τ, η, α, such that for $\lambda = c_{\tau,\eta,\alpha} \log p$, $T = C_{\tau,\eta,\alpha} \sqrt{k} \log p$, the following holds.*

1. *If $\theta = 0$, then $\psi_{\alpha,\lambda,T}(X, Y) \xrightarrow{\text{a.s.}} 0$.*
2. *If $\|\theta\|_2 > c'_{\tau,\eta,\alpha} \frac{\sqrt{k} \log p}{\sqrt{n}}$, then $\psi_{\alpha,\lambda,T}(X, Y) \xrightarrow{\text{a.s.}} 1$.*

We now turn our attention to the estimation in Algorithm 1′. The key of understanding the performance of Algorithm 1′ lies in an analysis of the estimated projection vector $\hat{v}$ in Step 10 of the algorithm. By Proposition 1 in the appendix, we expect $Q = (Q_1, \ldots, Q_{n-1})$ to be well-approximated by the rank-one matrix $\theta \gamma^\top$, where $\gamma = (\gamma_t)_{t \in [n-1]}$ is defined in (3.3). Thus, the oracle projection direction to aggregate Q is along $\theta / \|\theta\|_2$. We are now in a position to state the convergence rate of the changepoint estimator from Algorithm 1′.

Theorem 3 *Assume Conditions 1 and 2 and that data (X, Y) are generated according to (1.1) with $\nu = 1$. Suppose that $\|\theta\|_2 \leq 1$, $k \leq p/2$ and that $\min(\tau, 1 - \tau) \geq \alpha > 0$ for some known α. There exists $c_{\tau,\eta,\alpha} > 0$, depending only on τ, η, α, such that if $\lambda > c_{\tau,\eta,\alpha} \log p$, then the output $\hat{z}$ of Algorithm 1′ (defined below Proposition 1) with input (X, Y), λ and α satisfies with probability 1 for all but finitely many p's that*

$$\frac{|\hat{z} - z|}{n} \lesssim_{\tau,\eta,\alpha} \frac{\lambda^2 \sqrt{k}}{\sqrt{n} \|\theta\|_2^2}.$$

Theorem 3 shows that with a tuning parameter choice of order $\log p$, and when $\|\theta\|_2$ is bounded (which is the more difficult regime for estimation), Algorithm 1′ produces a consistent changepoint

estimator with a rate of convergence of order $k^{1/2}n^{-1/2}\|\theta\|_2^{-2}$ up to logarithmic factors. However, in light of the testing viewpoint of Theorem 2, in which it is possible to test apart the null of no change against a sparse alternative if $\sqrt{n}\|\theta\|_2/\sqrt{k}$, up to logarithmic factors, is sufficiently large, the rate in the above theorem appears to have an extra factor of $\|\theta\|_2^{-1}$. This additional factor is likely to arise from the technical difficulty of controlling the weak, though complex, dependence between the estimated projection direction $\hat{v}$ and the sketched Gaussian noises ($W_t^\top \xi : t \in [n-1]$). Indeed, the following theorem shows that if $\hat{v}$ is estimated from an independent sample, then the estimator from Algorithm 1 has a rate of convergence that agrees with what is prescribed in Theorem 2, up to logarithmic factors. Recall that the definition of Q_t in Algorithm 1' is replaced by (3.2).

Theorem 4   *Assume the same conditions as in Theorem 3. Let $(\tilde{X},\tilde{Y})$ be an independent copy of (X,Y). Let Q be the matrix constructed in Step 7 of Algorithm 1' (defined below Proposition 1) with input (X,Y), λ and α. Suppose $\hat{v}$ is computed in Step 9 of Algorithm 1' with input $(\tilde{X},\tilde{Y})$, λ and α. Then $\hat{z} := \arg\max_{t\in[n-1]} |\hat{v}^\top Q_t|$ satisfies with probability 1 for all but finitely many p's that*

$$\frac{|\hat{z}-z|}{n} \lesssim_{\tau,\eta,\alpha} \frac{\lambda\sqrt{k}\log p}{\sqrt{n}\|\theta\|_2}.$$

This additional independent sample $(\tilde{X},\tilde{Y})$ may be obtained in reality via a sample-splitting scheme. For example, we may take all odd time points to construct the Q matrix, and then use the even time points to estimate the projection direction $\hat{v}$. However, such sample splitting is necessary only from a technical viewpoint, and the algorithm typically performs better without sample splitting in practice.

We remark that the rate in the above Theorem 4 is slower compared to the usual results from change-in-mean problems as well as under regression settings, where rates of order $n^{-1}\|\theta\|_2^{-2}$ are achievable under appropriate conditions [see e.g. 12, 45, 47, 48, 49]. Unlike the classical change-in-mean setting, where the perturbation term (the CUSUM of the noise series) can be controlled both globally and locally, with the local fluctuation admitting a modulus of continuity bound scaling as $\sqrt{\delta}$ for time increment δ, our proof provides only a global perturbation bound. The local modulus of continuity bound seems difficult to obtain with the current toolbox of random matrix theory we have. We believe that this lack of local control of the perturbation is behind the apparent suboptimal rate. It remains to be seen if the estimation rate can be improved via alternative and possibly more refined analysis routes.

We now turn our attention to theoretical guarantees in the multiple changepoint setting. The following theorem shows that provided that we have a good single changepoint estimation and testing procedure in any changepoint problem, combining the narrowest-over-threshold with the single change procedures yields a multiple changepoint estimation procedure of similar accuracy with theoretical guarantees.

Theorem 5   *Let $D_1,\ldots,D_n$ be a data sequence with changepoints $0 = z_0 < z_1 < \cdots < z_v < z_{v+1} = n$. We assume that $z_i - z_{i-1} \geq n\Delta_\tau$ for all $i \in [v+1]$. Let $\mathcal{M}$ be defined as in Algorithm 3. Write $\mathcal{I}_0 := \{(s,e) \in \mathcal{M} : (s+n\varpi, e-n\varpi) \cap \{z_1,\ldots,z_v\} = \emptyset\}$ and for $i \in [v]$, define $\mathcal{I}_i := \{(s,e) \in \mathcal{M} : s \in [z_i - n\Delta_\tau/2, z_i - n\Delta_\tau/3], e \in [z_i + n\Delta_\tau/3, z_i + n\Delta_\tau/2]\}$ and $\tilde{\mathcal{I}}_i := \{(s,e) \in \mathcal{M} : \min\{z_i - s, e - z_i\} \geq n\Delta_\tau/6 \text{ and } e - s \leq n\Delta_\tau\}$. Let $\hat{z}$ and ψ be the single changepoint estimation and testing procedure used in Algorithm 3. Define the events $\Omega_0 := \{\forall i \in [v], \exists m \in [M], \text{ s.t. } (s_m,e_m) \in \mathcal{I}_i\}$, $\Omega_1 := \{\psi(D_{(s+n\varpi,e-n\varpi)}) = 0 \text{ for all } (s,e) \in \mathcal{I}_0\}$, $\Omega_2 := \bigcap_{i\in[v]}\{\psi(D_{(s+n\varpi,e-n\varpi)}) = 1 \text{ for all } (s,e) \in \mathcal{I}_i\}$ and for some $\phi_1,\ldots,\phi_v > 0$, $\Omega_3 := \bigcap_{i\in[v]}\{|\hat{z}(D_{(s,e)}) - (z_i - s)| \leq n\phi_i \text{ for all } (s,e) \in \tilde{\mathcal{I}}_i\}$. Let $\hat{z}_1,\ldots,\hat{z}_{\hat{v}}$ be the output of Algorithm 3 with inputs $D_1,\ldots,D_n$, $M > 0$, $\varpi = \Delta_\tau/6$, $\hat{z}$ and ψ. Assume further*

$\phi := \max_{i \in [v]} \phi_i < \varpi$. We have on $\Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3$ that $\hat{v} = v$ and $n^{-1}|\hat{z}_i - z_i| \leq \phi_i$ for all $i \in [v]$. In particular, we have

$$\mathbb{P}\big(\hat{v} = v \text{ and } n^{-1}|\hat{z}_i - z_i| \leq \phi_i \text{ for all } i \in [v]\big) \geq 1 - \mathbb{P}(\Omega_1^c) - \mathbb{P}(\Omega_2^c) - \mathbb{P}(\Omega_3^c) - v e^{-\Delta_\tau^2 M/36}.$$

Note that the theorem is valid for any generic multiple changepoint estimation that combines valid single changepoint estimation and testing procedures and the top-down narrowest-over-threshold multiple changepoint estimation paradigm. Hence, it can be applied in contexts other than the linear regression setting here. The statement of Theorem 5 is slightly stronger than the usual results on narrowest-over-threshold procedures, where ϕ_i are taken identical.

Applying the above theorem to our specific problem, we extend the single changepoint estimation result in Theorem 4 to the multiple changepoint setting and establish the estimation accuracy of Algorithm 3. We first give the following condition, which is the equivalent of Condition 2 in the multiple changepoint setting.

Condition 3 n, p satisfy $n > p$ and that $(n - p)/n \to \eta$ as $\min(n, p) \to \infty$. Assume further the changepoints satisfy $0 = z_0 < z_1 < \cdots < z_v < z_{v+1} = n$ and $z_i - z_{i-1} > n\Delta_\tau$ for all $i \in [v + 1]$ and $z_i/n \to \tau_i$ for $0 \leq i \leq v + 1$.

Due to the asymptotic nature of our theoretical results in the above, we consider a sequence of Algorithm 3. To facilitate proof, we study a specific coupling of the random intervals $\{(s_m, e_m) : m \in [M]\}$ generated across this sequence as follows:

$$\begin{aligned}
(\tilde{s}_m, \tilde{e}_m) &\overset{\text{iid}}{\sim} \text{Unif}\big(\{(\tilde{a}, \tilde{b}) : 0 \leq a < b \leq 1\}\big), \ \forall \, m \in [M], \\
s_m &= \lfloor n\tilde{s}_m \rfloor, e_m = \lceil n\tilde{e}_m \rceil \text{ for } m \in [M] \text{ and } n \in \mathbb{N}.
\end{aligned} \tag{3.4}$$

Note that the intervals generated by (3.4) have the same law as those generated in Algorithm 3.

Corollary 6   Let X and Y be generated by (1.1) and write $D_i := (X_i, Y_i)$ for $i \in [n]$. Assume Conditions 1 and 3 hold. There exist $c, C, c', C' > 0$, depending only on $\alpha, \Delta_\tau, \eta$, such that the following holds. For $\alpha < 1/6$, $\lambda = c \log p$, $T = C\sqrt{k} \log p$, let $\hat{z} = \hat{z}_{\alpha, \lambda}$ be the sample-splitted version of the single changepoint estimator defined in Algorithm 1' and $\psi = \psi_{\alpha, \lambda, T}$ be the testing procedure defined in (2.2). If $\Delta_\tau > 3(1 - \eta)$, $\frac{c'\sqrt{k}\log p}{\sqrt{n}} \leq \|\theta^{(i)}\|_2 \leq 1$ and $\|\theta^{(i)}\|_0 \leq k$ such that $\frac{k \log p}{n} \to 0$, then the output $\hat{z}_1, \ldots, \hat{z}_{\hat{v}}$ of Algorithm 3 with intervals $\{(s_m, e_m) : m \in [M]\}$ generated according to (3.4), inputs $(D_i)_{i \in [n]}$, $M > 0$, $\varpi = \tau/6$, $\hat{z}_{\alpha, \lambda}$ and $\psi_{\alpha, \lambda, T}$ satisfies with probability $1 - v e^{-\Delta_\tau^2 M/36}$ that for all but finitely many p's,

$$\hat{v} = v \text{ and } \frac{|\hat{z}_i - z_i|}{n} \leq \frac{C'\lambda\sqrt{k}\log p}{\sqrt{n}\|\theta^{(i)}\|_2} \text{ for all } i \in [v].$$

4. Numerical study

4.1. *Tuning parameter choice and comparison of variants*

Theoretical analysis in Section 3 have assumed that the noise variance σ^2 is known. In practice, we may obtain an upward-biased estimator $\tilde{\sigma}$ as the median absolute deviation of entries of the Q matrix. We

note

$$\text{Var}(Q_t) = \mathbb{E}[\text{Var}\{\tilde{W}_t^\top (W_z \theta + \xi) \mid X\}] + \text{Var}[\mathbb{E}\{\tilde{W}_t^\top (W_z \theta + \xi) \mid X\}]$$
$$= \sigma^2 \mathbb{E}(\tilde{W}_t^\top \tilde{W}_t) + \text{Var}(\tilde{W}_t^\top W_z \theta).$$

Since $\mathbb{E}(\tilde{W}_t^\top \tilde{W}_t)$ has all diagonal entries equal to 1, every entry of Q has a marginal variance of at least σ^2.
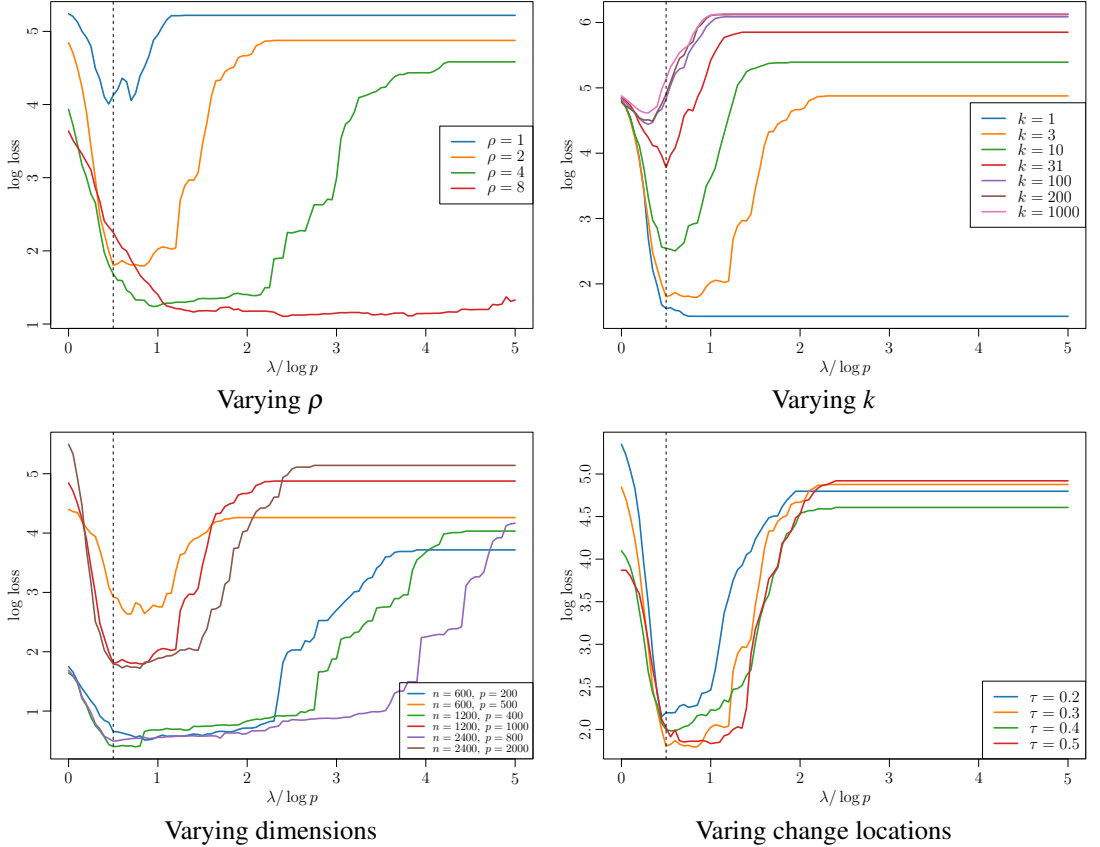

FIG. 2. The effect of choosing different λ in a series of studies. Our recommendation of $\lambda := 0.5 \log p$ is marked in dashed vertical line in each panel. Unless specified otherwise in each one, the panels share the parameters $n = 1200, p = 1000, \tau = 0.3, k = 3, \rho = 2$.

Algorithm 1 requires a soft-thresholding tuning parameter $\lambda > 0$ as an input. The theoretical results in Section 3 suggests using $\lambda = c\sigma \log p$ for some $c > 0$. We investigate here the performance of our algorithm at different soft-thresholding levels λ. Specifically, we computed the logarithmic average loss $|\hat{z} - z|$ of Algorithm 1 over 100 Monte Carlo repetitions for parameter settings of $n \in \{600, 1200, 2400\}$, $p \in \{n/3, 5n/6\}$, $\tau \in \{0.2, 0.3, 0.4, 0.5\}$, $k \in \{1, 3, 10, \lfloor \sqrt{p} \rfloor, \lfloor 0.1p \rfloor, \lfloor 0.2p \rfloor, p\}$, $\rho := \|\theta\|_2 \in \{1, 2, 4, 8\}$, $\sigma = 1$ and various choices of $c \in [0.1, 5]$. In all our simulations here and below, we sample the vector of change in the regression coefficients θ uniformly from the set

$\{v : \|v\|_0 = k, \|v\|_2 = \rho\}$, and generate dense pre-change vector from $N_p(0, \max\{1, \rho^2\}I_p)$. Figure 2 illustrates part of the simulation results where we vary one aspect of the parameters at a time. From the figure, we see that a choice of $c = 0.5$ provides good statistical performance across the parameter settings considered, and we will henceforth adopt this choice of $\lambda = 0.5\tilde{\sigma}\log(p)$ in our subsequent numerical studies.

We now compare the statistical performance of various versions of complementary-sketching-based approaches proposed in the paper, including Algorithms 1 and 2 from Section 2 and the slight variant Algorithm 1′ mentioned in Section 3 to facilitate theoretical analysis. For a demonstrative purpose, we have also included the naive hard- and soft-thresholded changepoint estimators $\hat{z}^{\text{hard}}$ and $\hat{z}^{\text{soft}}$ mentioned just above Algorithm 1. We use the λ choice suggested in the previous paragraph for $\hat{z}^{\text{hard}}$ and $\hat{z}^{\text{soft}}$, Algorithm 1 and its variant, and choose λ_t in Algorithm 2 via a five-fold cross-validation for each $t \in [n-1]$. Empirical observations suggest that Algorithms 1, 1′ and 2 work well without any burn-in (i.e. $\alpha = 0$). However, both $\hat{z}^{\text{soft}}$ and $\hat{z}^{\text{hard}}$ do suffer from more serious boundary effects, as seen in the large root mean squared errors in Table 1. In addition, Algorithm 1 has roughly the same but slightly better estimation accuracy compared to its primed variant. This justifies our recommendation of Algorithm 1 over its primed variant, and the similarity in performance further consolidates the relevance of our theoretical analysis on the primed variant as a proof device.

n	p	k	ρ	$\hat{z}^{\text{soft}}$	$\hat{z}^{\text{hard}}$	Alg1	Alg1′	Alg2
600	200	3	1	50.1	178.35	12.6	40.79	22.34
			2	3.98	88.76	3.1	5.01	4.58
			4	25.37	92.3	2.14	4.33	2.29
		14	1	31.7	122.11	37.36	83.59	129.71
			2	6.4	88.04	6.14	7.32	9.83
			4	3.88	58.92	3.82	4.67	2.95
1200	400	3	1	11.24	196.81	10.95	13.34	19.54
			2	3.15	136.02	3.69	4.58	4.96
			4	35.65	101.17	1.67	4.36	2.09
		20	1	39.17	146.98	18.14	37.63	168.56
			2	5.06	144.5	7.29	7.59	15.51
			4	4.72	106.84	2.72	2.93	4.47

TABLE 1 *Comparisons of performances of $\hat{z}^{\text{soft}}$, $\hat{z}^{\text{hard}}$, Algorithms 1, 1′ and 2 in various settings in terms of root mean squared error. No burn-in is applied anywhere ($\alpha = 0$). The true change takes place at $0.3n$.*

4.2. *Comparisons with other methods*

From the discussion above, we recommend using Algorithms 1 and 2 for their robustness against the choice of the burn-in parameter α, and for the more accurate estimation of Algorithm 1 over its primed variant. We will henceforth focus on Algorithms 1 and 2, which we call **charcoal$_{\text{proj}}$** and **charcoal$_{\text{lasso}}$**, respectively. In this section, we compare the performance of **charcoal$_{\text{proj}}$** (**CP**) and **charcoal$_{\text{lasso}}$** (**CL**)

with existing approaches in the literature. Specifically, we will compare against the VPBS algorithm of Rinaldo et al. [40], two-sided Lasso-based approaches of Lee et al. [32] (LSS), Leonardi and Bühlmann [33] (LB) and Cho and Owens [12], a two-stage refinement approach of Kaul et al. [28] (KJF), a dynamic programming algorithm of Xu et al. [49]. We have used the authors' own implementation for VPBS, KJF, CO and XWZY, and Kaul et al. [28]'s implementation for LSS. We have implemented LB ourselves using the recommended tuning parameter choices as in Leonardi and Bühlmann [33]. It is worth noting that none of the six existing methods in the literature were designed to estimate changes in the regression coefficients when both the pre- and post-change coefficients are dense. As such, we also include simulation settings where the pre-change regression coefficient is s-sparse for $s < p$.

We compare the performance in terms of mean absolute loss of various methods in a single changepoint estimation task for $n = 1200$, $p = 400$, $\tau = 0.3$, $k \in \{3, \lfloor \sqrt{p} \rfloor, p\}$, $s \in \{k, p\}$, $\rho \in \{1, 2, 4, 8\}$ in Table 2 (we also include additional results for $p = 1000$ in Table S.1 in the supplement). We see that when the pre-change coefficients are sparse, most existing methods works well and in particular, KJF tends to have the best average loss, but our **charcoal$_{\text{proj}}$** and **charcoal$_{\text{lasso}}$** remains comparable. In contrast, if the pre-change regression coefficients are dense, our methods can significantly outperform other approaches. Specifically, we see that none of VPBS, LB, KJF, LSS, CO or XWZY show any sign of consistent estimation as their average loss do not decrease as the signal strength increases. On the other hand, both **charcoal$_{\text{proj}}$** and **charcoal$_{\text{lasso}}$** have shown highly promising performance in various settings. It is surprising that our methods also seem to work even when the vector of change is dense. We notice that **charcoal$_{\text{proj}}$** does not perform as well as **charcoal$_{\text{lasso}}$** when the change is dense. This may be related to the fact that the marginal correlation based test statistics Q_t may suffer from larger bias due to covariate collinearity in the transformed design W when the vector of change θ is dense. We further note that **charcoal$_{\text{lasso}}$** shows better estimation accuracy when either the signal strength ρ is high or the vector of change θ is dense.

4.3. *Model misspecification*

While we have focused on the Gaussian Orthogonal Ensemble (GOE) design (i.e. X has independent $N(0, 1)$ entries) and Gaussian noise in the theoretical analysis, our methodology can be applied in more general settings. In this subsection, we investigate the robustness of the estimation accuracy of our method to deviations from this Gaussian distributional assumptions. Specifically for $n = 1,200$, $p = 400$, $\tau = 0.3$, $k = 20$ and $\rho \in \{1.5^0, \ldots, 1.5^8\}$, we varied the design matrix X to have either $N_p(0, \Sigma)$ rows, where $\Sigma = (0.7^{|i-j|})_{i,j \in [p]}$ has an autoregressive Toeplitz structure, or independent Rademacher entries. We also vary the noise distribution to take t_4, t_6, centred $\text{Exp}(1)$ or Rademacher distributions. Overall, we see from Figure 3 that the performance of **charcoal$_{\text{lasso}}$** is robust to both non-GOE design matrices and discrete, heavy-tailed or skewed noise distributions. Similar results hold for the **charcoal$_{\text{proj}}$** method.

We further investigate the robustness of our procedure to more severe departure of the design distribution from Condition 1, by allowing the covariate distributions to change over time. Specifically, we consider the following three scenarios:

(a) For $z' = n - z$, we draw $x_t \sim N(\mathbf{0}, I_p)$ for $1 \le t \le z'$, and $x_t \sim N(-\mathbf{100}, I_p)$ for $z' < t \le n$.

(b) We first draw $\tilde{x}_t \overset{\text{iid}}{\sim} N(0, I_p)$, and then find a permutation π such that $\|\tilde{x}_{\pi(1)}\|_2 \le \cdots \le \|\tilde{x}_{\pi(n)}\|_2$. Finally, we set $x_t = \tilde{x}_{\pi(t)}$.

(c) For $z' = n - z$, we draw $x_t \sim N(\mathbf{0}, \gamma_t' I_p)$, where $\gamma_t' = \sqrt{\frac{t}{n(n-t)}}(n - z')$ for $t \le z'$ and $\gamma_t' = \sqrt{\frac{n-t}{nt}} z'$ for $t > z'$.

k	s	ρ	**CP**	**CL**	VPBS	LB	KJF	LSS	CO	XWZY
3	3	1	6.2	12.6	50.3	549.5	**3.8**	17.5	87.6	14.1
		2	2.0	4.2	7.6	53.5	**1.6**	4.3	12.5	7.3
		4	0.9	1.3	2.6	0.7	1.1	2.2	11.5	**0.5**
		8	1.0	0.8	1.1	**0.3**	1.0	0.8	10.5	3.2
3	400	1	**7.2**	13.2	452.4	556.1	238.8	472.2	247.3	254.6
		2	**2.2**	3.5	476.3	569.2	239.3	364.1	86.5	315.0
		4	**1.1**	1.5	434.2	532.8	239.1	272.1	29.4	325.1
		8	**0.7**	0.8	326.3	496.8	239.1	310.8	62.2	327.0
20	20	1	13.5	96.4	55.2	716.7	**5.8**	33.1	373.3	18.2
		2	4.2	12.2	10.2	675.5	**1.8**	6.9	315.1	2.7
		4	1.6	2.6	2.9	69.8	**1.4**	2.6	174.3	7.7
		8	1.4	1.1	3.1	**0.6**	1.2	1.0	111.3	2.0
20	400	1	**12.4**	85.4	422.7	528.8	238.9	479.5	387.0	273.5
		2	**3.0**	9.2	494.9	546.8	238.9	284.5	364.8	316.2
		4	**2.0**	2.6	431.9	553.1	239.1	268.5	367.0	321.8
		8	1.9	**0.8**	356.2	513.3	239.3	261.5	397.1	322.0
400	400	1	**162.2**	344.2	477.8	569.8	238.8	429.9	406.9	263.7
		2	**46.3**	338.4	504.0	583.2	238.8	252.4	408.5	313.9
		4	25.3	**13.3**	446.3	554.1	238.9	285.6	392.3	323.4
		8	20.7	**3.0**	355.6	487.6	239.1	250.1	374.2	324.8

TABLE 2 *Average loss of various changepoint methods under different settings. Other parameters: $n = 1200$, $p = 400$ and $z = 360$. The method with the least average loss in each line is marked in bold.*

As we see in Figure 4, our **charcoal**$_{\text{lasso}}$ procedure shows remarkable robustness against possibly strong design heterogeneity.

4.4. *Multiple changepoints*

As mentioned in Section 2, our **charcoal** algorithms can be easily combined with generic multiple changepoint methods to deal with multiple changepoints, and we proposed a specific version in Algorithm 3 of such a multiple changepoint estimation procedure. We run Algorithm 3 with $\varpi = 0$ and $M = 200$. For the single changepoint estimator input $\hat{z}$ in Algorithm 3, we employ Algorithm 1 with the recommended value of λ in Section 4.1 and the burn-in parameter $\alpha = 0.05$. For the testing procedure input ψ of Algorithm 3, we run Algorithm 1 to obtain output $H_{\max}$ and define $\psi(X,Y) = \mathbb{1}_{\{H_{\max} > T\}}$, where the testing threshold T is chosen by a Monte Carlo simulation as follows. We generate $B = 1000$ pairs of (X,Y) under model (1.1) with $\nu = 0$ (i.e. no changepoint), and run Algorithm 1 with these synthetic (X,Y) pairs and the same λ and α choices as above. This would return B test statistics $(H_{\max}^b)_{b \in [B]}$, which are used to estimate an upper $0.01/M$ quantile by fitting a generalized extreme value distribution [41].
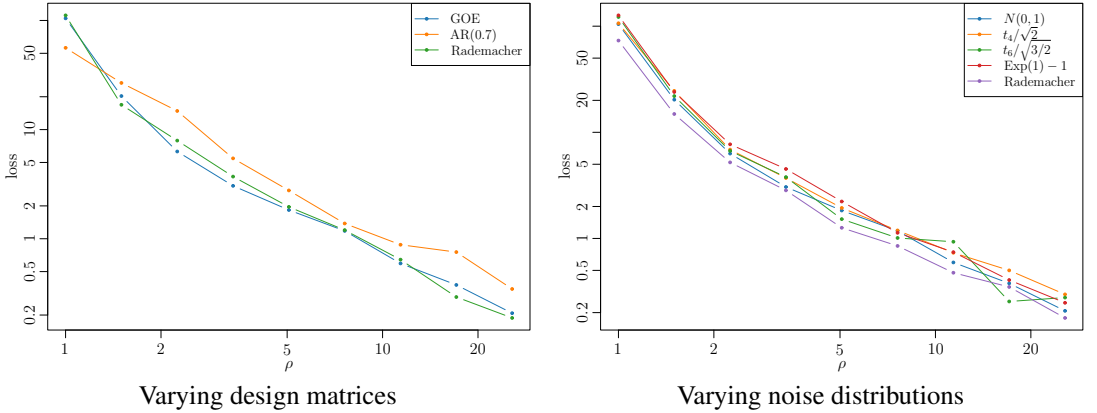
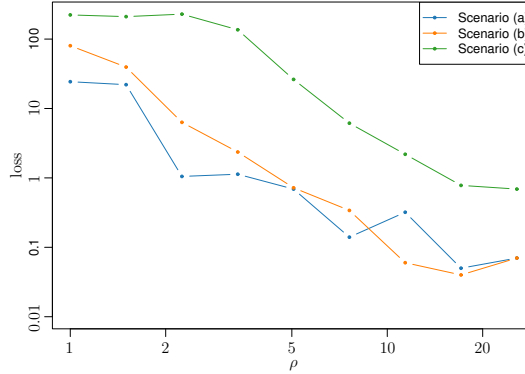FIG. 3. *Robustness to varying design matrices and noise distributions.*

FIG. 4. *Robustness to time-varying design distribution. See Section 4.3 for detailed setup of the three scenarios.*

While Algorithm 3 already produces good estimators of multiple changepoints, its performance can be further improved by the following post-processing procedures. Such post-processing has previously been described in e.g., Fryzlewicz [19, Section 2 of online supplement].

Specifically, after we obtain an initial candidate set of changepoints using Algorithm 3, we iteratively run the test prescribed by (2.2) on the largest time interval containing each candidate changepoint as the only estimated changepoint, and remove that candidate changepoint if the test is non-significant. For the remaining candidate changepoints $\hat{z}_1, \ldots, \hat{z}_{\hat{v}}$, we refine their estimated locations in two steps. We first perform a 'midpoint' refinement, where we use (for instance) Algorithm 2 to output a refined estimator $\tilde{z}_i$ based on data $\{(x_t, y_t) : t \in ((\hat{z}_{i-1} + \hat{z}_i)/2, (\hat{z}_i + \hat{z}_{i+1})/2]\}$ for each $i \in [\hat{v}]$. Here, we use the convention that $\hat{z}_0 = 0$ and $\hat{z}_{\hat{v}+1} = n$. Using midpoints between successive estimated changepoints ensures that each $((\hat{z}_{i-1} + \hat{z}_i)/2, (\hat{z}_i + \hat{z}_{i+1})/2]$ contains with high probability at most one true changepoint. However, it does not use the full data available around each true changepoint. As such, we perform a second refinement step after this, where we use Algorithm 2 to output a further refined estimator $\tilde{z}_i^{\text{refined}}$ based on data $\{(x_t, y_t) : t \in (\tilde{z}_{i-1} + \alpha n, \tilde{z}_{i+1} - \alpha n]\}$ for each $i \in [\hat{v}]$ with α being the burn-in parameter as in Algorithm 3. Again, we use the convention that $\tilde{z}_0 = 0$ and $\tilde{z}_{\hat{v}+1} = n$. For both refinement steps, we may also use Algorithm 1 in place of Algorithm 2, and they have very similar performances in our numerical

n	p	k	$\rho_{\min}$	$\hat{v} - v$ value counts					Haus	ARI
				-3	-2	-1	0	1		
1200	200	3	0.8	0	0	96	4	0	292.8	0.742
			1.2	0	0	22	78	0	75.4	0.918
			1.6	0	0	0	98	2	8.8	0.978
		10	0.8	0	2	97	1	0	304.9	0.710
			1.2	0	0	42	55	3	141.1	0.856
			1.6	0	0	1	96	3	18.0	0.960
		100	0.8	3	67	30	0	0	591.7	0.303
			1.2	0	4	88	8	0	319.3	0.611
			1.6	0	0	52	46	2	217.1	0.759
2400	400	3	0.8	0	0	25	75	0	155.3	0.881
			1.2	0	0	0	100	0	14.3	0.975
			1.6	0	0	0	100	0	10.1	0.983
		10	0.8	0	15	53	32	0	376.9	0.720
			1.2	0	0	2	98	0	37.3	0.945
			1.6	0	0	1	99	0	21.0	0.970
		100	0.8	42	57	1	0	0	1154.9	0.184
			1.2	0	32	54	14	0	647.0	0.457
			1.6	0	0	14	84	2	376.9	0.658

TABLE 3   *Summary of results of multiple changepoint estimations under (M1) and (M2) described in Section 4.4 with $\rho \in \{0.8, 1.2, 1.6\}$ and $k \in \{3, 10, 100\}$. The first nine rows of the table corresponds to setting (M1) and the last nine rows corresponds to (M2). The columns of $\hat{v} - v$ tabulates the difference in number of estimated and true changepoints over 100 Monte Carlo repetitions. The 'Haus' and 'ARI' columns measure the average Hausdorff distance and the average adjusted rand index between the discovered partition and the true partition over 100 repetitions.*

experiment. For definiteness and simpler presentation, we employ Algorithm 2 for both refinement steps in the following numerical experiments.

We assume that the regression noise level σ is known and consider the following two multiple changepoint specifications in our simulations: (M1) $n = 1200$, $p = 200$, $v = 3$ and three changepoints are located at $z = (z_1, z_2, z_3) = (240, 540, 900)$ with signal sizes $(\|\theta^{(1)}\|_2, \|\theta^{(2)}\|_2, \|\theta^{(3)}\|_2) = \rho_{\min} \times (1, 1.5, 2)$ and sparsity $\|\theta^{(1)}\|_0 = \|\theta^{(2)}\|_0 = \|\theta^{(3)}\|_0 = k$ respectively for various $\rho_{\min}$ and k; (M2) $n = 2400$, $p = 400$, $v = 4$ and four changepoints are located at $z = (z_1, z_2, z_3, z_4) = (720, 1320, 1800, 2160)$ with signal sizes $(\|\theta^{(1)}\|_2, \|\theta^{(2)}\|_2, \|\theta^{(3)}\|_2, \|\theta^{(4)}\|_2) = \rho_{\min} \times (1, 1.15, 1.45, 2.18)$ and sparsity $\|\theta^{(1)}\|_0 = \|\theta^{(2)}\|_0 = \|\theta^{(3)}\|_0 = \|\theta^{(4)}\|_0 = k$ respectively for various $\rho_{\min}$ and k.
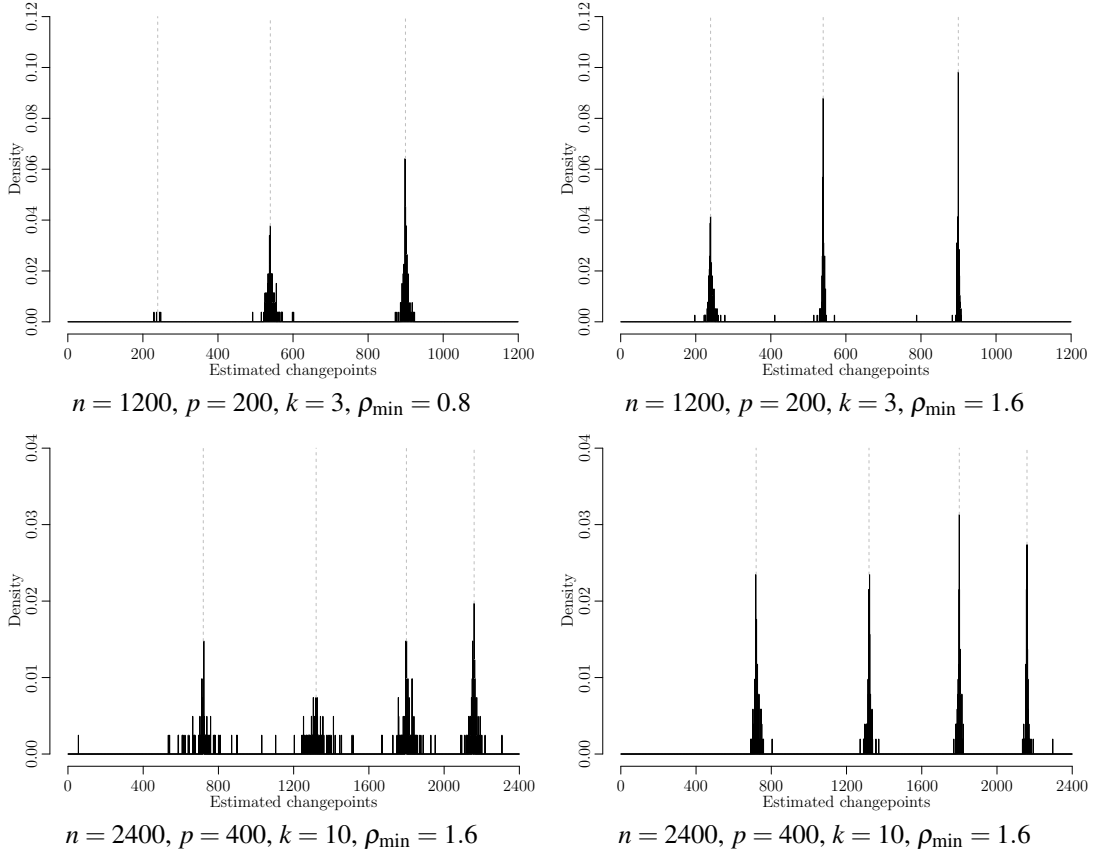
FIG. 5. *Histogram of estimated changepoint locations in four settings. The true changes take place at $z = (240, 540, 900)$ for the $(n, p) = (1200, 200)$ specifications in the above two panels with the signal strengths at respective changes being $\rho_{\min} \times (1, 1.5, 2)$. For the two panels below with $(n, p) = (2400, 400)$, $z = (720, 1320, 1800, 2160)$ with $\rho_{\min} \times (1, 1.15, 1.45, 2.175)$. The locations of true changes are marked in lightly-colored dashed vertical lines in each plot.*

Note that for (M2), the signal sizes are chosen such that

$$\|\theta^{(i)}\|_2^2 \frac{(z_i - z_{i-1})(z_{i+1} - z_i)(z_{i+1} - z_{i-1} - p)}{(z_{i+1} - z_{i-1})^2}$$

is approximately constant for each $i \in [\nu]$, which according to Gao and Wang [21] means that the effective signal-to-noise ratio of testing for each changepoint z_i within the interval $(z_{i-1}, z_{i+1}]$ is almost constant. Table 3 reports the multiple changepoint estimation performances for both (M1) and (M2) with $\rho_{\min} \in \{0.8, 1.2, 1.6\}$ and $k \in \{3, 10, 100\}$. The multiple changepoint estimation accuracy is measured in terms of the difference between the number of estimated and true changepoints, the average Hausdorff distance between the sets $\{z_i : i \in [n]\}$ and $\{\hat{z}_i^{\text{refined}} : i \in [\hat{n}]\}$ and finally the average adjusted Rand index (ARI) [39] of the estimated segments against the truth, over 100 Monte Carlo repetitions. We see from Table 3 that the promising single changepoint estimation performance of our methodology carries over to the multiple changepoint settings.

Figure 5 visualizes the simulation results by showing the histograms of estimated changepoints in four of the parameter settings shown in Table 3. It is worth noting that in the bottom two panels of the figure, where the effective signal-to-noise ratios are chosen to be approximately constant for all the four changepoints, we indeed see a similar number of times in identifications of each changepoint.

4.5. *Real data example*

In this subsection, we showcase how the **charcoal** algorithm can be applied to a single-cell gene expression dataset from Suo et al. [43]. The original data consists of the logarithmic normalized gene expression levels of 3211 genes measured in 11853 cells along the developmental trajectory from proliferating double positive cells (DP(P) cells) to quiescent double positive cells (DP(Q) cells), $\alpha\beta$T entry cells (ABT cells) and finally to CD4+ T cells. These cells have been ordered in pseudotime according to their development stage in Suo et al. [43], which we use as our timeline (see Figure 6). We are interested in understanding the change in the gene regulatory networks along this time trajectory. We can estimate the changepoints by modeling the logarithmic normalized expressions using Gaussian graphical models and seek changes in the nodewise regression coefficients of each gene against the remaining genes. To speed up the computation, we preprocess the data by subsampling $1/3$ of the original cells and only using genes that have non-zero expression in at least 5% of the cells. The changepoints are estimated using Algorithm 1 with tuning parameters chosen as suggested in Section 4.1. In Table 4, we list the genes that reported most significant test statistics in their nodewise regression coefficients along this pseudotime trajectory. From Figure 6, we see that most of the changes are identified immediately before the boundary between the DP(P) and DP(Q) boundary, and most of the associated genes TK1, CKAP2L, TTK, ARHGEF39, DEPDC1, SPC25, GTSE1, HMMR, CENPA are well-known regulators for cell proliferation in biology [7, 22, 36, 51, 53, see, e.g.]. The change in nodewise regression coefficient of the RAG2 gene occurred immediately before the DP(Q) and ABT boundary, which agrees with the existing literature that RAG2 is a regulator for T cell development [27].
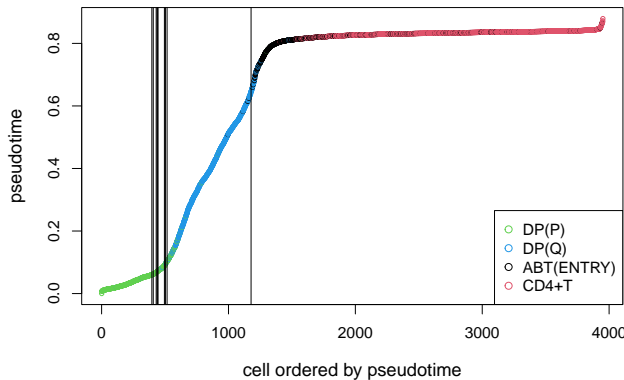

FIG. 6. Ordered pseudotime of cells in the real data example of Section 4.5. Each plot point represents a cell, colored by its annotated cell type. Vertical lines corresponds to estimated changepoint locations of the most significant changes in the nodewise regression coefficients as described in Section 4.5.

Gene	Changepoint	Top interacting partners
TK1	495	PTP4A1, KIF20B, CENPF, SPRY1, ZWINT
CKAP2L	430	TOGARAM1, DEPDC1B, AP001816.1, TUBA1C, FANCI
RAG2	1178	SMPD3, AL365440.2, LZTFL1, AEBP1, HIST1H2BJ
TTK	407	UBE2S, KNL1, CDC20, TRAV19, DDIT3
ARHGEF39	396	HJURP, CD1A, SLC25A25, CCDC152, MBTD1
DEPDC1	444	AL138899.1, USPL1, RIPK4, SERPINF1, EPHB6
SPC25	442	ATF3, ITGAE, CDC42EP3, AC136475.5, EPS8
GTSE1	437	MID1IP1, HIST1H2AG, GADD45G, PSRC1, FBLN5
HMMR	516	TAX1BP3, LAIR1, SERP2, LANCL2, MANEA-DT
CENPA	503	TRBV7-3, SOCS1, FRMD4B, CDKN1A, FXYD2

TABLE 4 *List of genes with most significant changes in the nodewise regression coefficients, together with their changepoint locations and top 5 interacting partners.*

REFERENCES

1. J. Bai. Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4): 551–563, 1997.
2. J. Bai. Common breaks in means and variances for panel data. *J. Econom.*, 157(1):78–92, 2010.
3. J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78, 1998.
4. R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. Roy. Statist. Soc., Ser. B*, 81(3):649–672, 2019.
5. H. Battey and D. Cox. High dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, 3(2):119–148, 2020.
6. H. S. Battey and N. Reid. On inference in high-dimensional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):149–175, 2023.
7. E. E. Bitter, M. H. Townsend, R. Erickson, C. Allen, and K. L. O'Neill. Thymidine kinase 1 through the ages: a comprehensive review. *Cell & Bioscience*, 10(1):1–16, 2020.
8. P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.
9. Y. Chen, T. Wang, and R. J. Samworth. High-dimensional, multiscale online changepoint detection. *J. Roy. Statist. Soc., Ser. B*, 84(1):234–266, 2022.
10. H. Cho. Change-point detection in panel data via double CUSUM statistic. *Electron. J. of Statist.*, 10(2): 2000–2038, 2016.
11. H. Cho and P. Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. Roy. Statist. Soc., Ser. B*, 77(2):475–507, Jul 2014. ISSN 1369-7412. doi: 10.1111/rssb.12079.
12. H. Cho and D. Owens. High-dimensional data segmentation in regression settings permitting temporal dependence and non-gaussianity. *Electronic Journal of Statistics*, 18(1):2620–2664, 2024.
13. H. Cho, T. Kley, and H. Li. Detection and inference of changes in high-dimensional linear regression with nonsparse structures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf029, 2025.
14. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.
15. F. Enikeeva and O. Klopp. Change-point detection in dynamic networks with missing links. *arXiv preprint*, page arXiv:2106.14470, 2021.

16. J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statist. Sinica*, 20(1):101, 2010.

17. B. Follain, T. Wang, and R. J. Samworth. High-dimensional changepoint estimation with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B*, page to appear, 2022.

18. P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281, 2014.

19. P. Fryzlewicz. Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Ann. Statist.*, 46(6B):3390–3421, 2018.

20. P. Fryzlewicz. Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *J. Korean Statist. Soc.*, 49(4):1027–1070, 2020.

21. F. Gao and T. Wang. Two-sample testing of high-dimensional linear regression coefficients via complementary sketching. *Ann. Statist.*, 50(5):2950 – 2972, 2022. doi: 10.1214/22-AOS2216. URL https://doi.org/10.1214/22-AOS2216.

22. L. Guo, S. Zhang, B. Zhang, W. Chen, X. Li, W. Zhang, C. Zhou, J. Zhang, N. Ren, and Q. Ye. Silencing gtse-1 expression inhibits proliferation and invasion of hepatocellular carcinoma cells. *Cell Biology and Toxicology*, 32:263–274, 2016.

23. W. L. Hatleberg and V. F. Hinman. Modularity and hierarchy in biological systems: Using gene regulatory networks to understand evolutionary change. In S. F. Gilbert, editor, *Evolutionary Developmental Biology*, volume 141 of *Current Topics in Developmental Biology*, pages 39–73. Academic Press, 2021. doi: https://doi.org/10.1016/bs.ctdb.2020.11.004.

24. L. Horváth and M. Hušková. Change-point detection in panel data. *J. Time Series Anal.*, 33(4):631–648, 2012.

25. M. Jirak. Uniform change point tests in high dimension. *Ann. Statist.*, 43(6):2451–2483, 2015.

26. S. A. Julious. Inference and estimation in a changepoint regression problem. *J. Roy. Statist. Soc., Ser. D*, 50(1): 51–61, 2001.

27. L. Kalman, M. L. Lindegren, L. Kobrynski, R. Vogt, H. Hannon, J. T. Howard, and R. Buckley. Mutations in genes required for t-cell development: IL7R, CD45, IL2RG, JAK3, RAG1, RAG2, ARTEMIS, and ADA and severe combined immunodeficiency: HuGE review. *Genetics in Medicine*, 6(1):16–26, 2004.

28. A. Kaul, V. K. Jandhyala, and S. B. Fotopoulos. An efficient two step algorithm for high dimensional change point regression models without grid search. *J. Mach. Learn. Res.*, 20:(111), 1–40, 2019.

29. S. Kovács, H. Li, P. Bühlmann, and A. Munk. Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633*, 2020.

30. P. Kraft and D. J. Hunter. Genetic risk prediction—are we there yet? *N. Engl. J. Medcine*, 360(17):1701–1703, 2009.

31. A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal. Sample complexity of learning mixture of sparse linear regressions. *Adv. Neur. Inform. Proc. Sys.*, 32, 2019.

32. S. Lee, M. H. Seo, and Y. Shin. The lasso for high dimensional regression with a possible change point. *J. Roy. Statist. Soc., Ser. B*, 78(1):193–210, 2016.

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

34. C. Lévy-Leduc and F. Roueff. Detection and localization of change-points in high-dimensional network traffic data. *Ann. Appl. Statist.*, pages 637–662, 2009.

35. H. Liu, C. Gao, and R. J. Samworth. Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.*, 49(2):1081–1112, 2021.

36. G. B. Mills, R. Schmandt, M. McGill, A. Amendola, M. Hill, K. Jacobs, C. May, A.-M. Rodricks, S. Campbell, and D. Hogg. Expression of TTK, a novel human protein kinase, is associated with cell proliferation. *Journal of Biological Chemistry*, 267(22):16000–16006, 1992.

37. J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

38. E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527, 1955.

39. W. M. Rand. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.*, 66(336): 846–850, 1971.

40. A. Rinaldo, D. Wang, Q. Wen, R. Willett, and Y. Yu. Localizing changes in high-dimensional regression models. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2089–2097. PMLR, 13–15 Apr 2021.

41. R. L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90, 1985.

42. N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2): 209–256, 2010.

43. C. Suo, K. Polanski, E. Dann, R. G. Lindeboom, R. Vilarrasa-Blasi, R. Vento-Tormo, M. Haniffa, K. B. Meyer, L. M. Dratva, Z. K. Tuong, et al. Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. *Nature Biotechnology*, pages 1–12, 2023.

44. R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, 58(1):267–288, 1996.

45. N. Verzelen, M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Optimal change-point detection and localization. *arXiv preprint arXiv:2010.11470*, 2020.

46. L. Y. Vostrikova. Detecting "disorder" in multidimensional random processes. In *Doklady Akademii Nauk*, volume 259(2), pages 270–274. Russian Academy of Sciences, 1981.

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

48. T. Wang and R. J. Samworth. High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B*, 80(1):57–83, 2018.

49. H. Xu, D. Wang, Z. Zhao, and Y. Yu. Change-point inference in high-dimensional regression models under temporal dependence. *The Annals of Statistics*, 52(3):999–1026, 2024.

50. D. Yin, R. Pedarsani, Y. Chen, and K. Ramchandran. Learning mixtures of sparse linear regressions using sparse graph codes. *IEEE Trans. Inform. Theory*, 65(3):1430–1451, 2018.

51. L. Zhang, Y. Du, S. Xu, Y. Jiang, C. Yuan, L. Zhou, X. Ma, Y. Bai, J. Lu, and J. Ma. DEPDC1, negatively regulated by mir-26b, facilitates cell proliferation via the up-regulation of FOXM1 expression in TNBC. *Cancer Letters*, 442:242–251, 2019.

52. N. R. Zhang, D. O. Siegmund, H. Ji, and J. Z. Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645, 2010.

53. H. Zhou, L. Cai, X. Zhang, A. Li, Y. Miao, Q. Li, X. Qiu, and E. Wang. ARHGEF39 promotes tumor progression via activation of Rac1/P38 MAPK/ATF2 signaling and predicts poor prognosis in non-small cell lung cancer patients. *Laboratory Investigation*, 98(5):670–681, 2018.

Supplementary materials of 'Sparse change detection in high-dimensional linear regression'

14th January 2026

The supplementary materials collect the proofs of the main results of Section A in the main text and the proofs of the ancillary results in Section B.

A   Proof of main results

Proof of Proposition 1. Define $\kappa_1 = \kappa_1(n, z, p) := z(n-z)(n-p)/n^3$, which under Condition 2 is $O(1)$. We decompose

$$
W_t^\top W_z - \frac{4tn\kappa_1}{z} I_p = W_t^\top W_z - \frac{t}{z} W_z^\top W_z + \frac{t}{z} W_z^\top W_z - \frac{4tn\kappa_1}{z} I_p
$$
$$
= 4t \left\{ \left(\frac{S_{0,t}}{t} - \frac{S_{0,z}}{z} \right) S_{0,n}^{-1} S_{z,n} + \frac{n}{z} \left(\frac{W_z^\top W_z}{4n} - \kappa_1 I_p \right) \right\}.
$$

Define $\Omega_\nu := \{ \|S_{0,n}^{-1} S_{z,n}\|_{\mathrm{op}} \|S_{0,z}/z\|_{\mathrm{op}} \le \nu \}$. By Lemma 15, we have a $(1/2)$-net $\mathcal{N}$ of $B_0(\ell)$ with cardinality at most $(5ep/\ell)^\ell$ such that we have

$$
\mathbb{P}(\sup_{u \in B_0(\ell)} |u^\top (S_{0,t}/t - S_{0,z}/z) S_{0,n}^{-1} S_{z,n} v| \ge x)
$$
$$
\le \mathbb{E}[\mathbb{P}(2 \sup_{u \in \mathcal{N}} |u^\top (S_{0,t}/t - S_{0,z}/z) S_{0,n}^{-1} v| \ge x \mid S_{0,z}, S_{z,n}) \mathbb{1}_{\Omega_\nu}] + \mathbb{P}(\Omega_\nu^c)
$$
$$
\le 52 \left(\frac{5ep}{\ell} \right)^\ell \exp\{-t^2 x^2 / (32 z \nu^2)\}) + \mathbb{P}(\Omega_\nu^c),
$$

where the first inequality holds by that Ω_ν is measurable with respect to the σ-algebra generated by $(S_{0,z}, S_{z,n})$ and the second by Lemma 13 and a union bound. Define $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ for any generic symmetric matrix A to

be the smallest and largest eigenvalues of A, respectively. By Wainwright (2019, Theorem 6.1), we have

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{S_{z,n}}{n-z}\right) \geq 2\left(1+\sqrt{p/(n-z)}\right)^2\right) \leq \exp\left(-\frac{(n-z)\left(1+\sqrt{p/(n-z)}\right)^2}{8}\right),$$

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{S_{0,z}}{z}\right) \geq 2\left(1+\sqrt{p/z}\right)^2\right) \leq \exp\left(-\frac{z\left(1+\sqrt{p/z}\right)^2}{8}\right),$$

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{S_{0,n}}{n}\right) \leq \frac{(1-\sqrt{p/n})^2}{4}\right) \leq \exp\left(-\frac{n(1-\sqrt{p/n})^2}{8}\right).$$

We define $\nu := 16[(n-z)/n](1+\sqrt{p/z})^2(1+\sqrt{p/(n-z)})^2(1-\sqrt{p/n})^{-2}$. By a union bound, we arrive at

$$\mathbb{P}(\Omega_\nu^c) = \mathbb{P}(\|S_{0,n}^{-1}S_{z,n}\|_{\mathrm{op}}\|S_{0,z}/z\|_{\mathrm{op}} \leq \nu) \leq \exp\left(-\frac{n\left(1-\sqrt{p/n}\right)^2}{8}\right) + 2\exp(-p/8).$$

Combining the above displays and setting $x := 16\nu\sqrt{\ell z \log(5ep/\ell)}/t$, we have by a union bound that

$$\mathbb{P}\left(\sup_{u\in B_0(\ell)} |u^\top(S_{0,t}/t - S_{0,z}/z)S_{0,n}^{-1}v| \geq x\right) \leq p^{-7} + e^{-n(1-\sqrt{p/n})^2/8} + 2e^{-p/8}.$$

Taking another union bound over $t \in [z]$, and by the Borel–Cantelli lemma, we have with probability 1, for all but finitely many p's that

$$\sup_{t\in[z]}\sup_{u\in B_0(\ell)} u^\top\left(W_t^\top W_z - (t/z)W_z^\top W_z\right)v \leq C'_{\tau,\eta}\sqrt{\ell p \log(ep/\ell)}, \qquad \text{(S.1)}$$

for some constant $C'_{\tau,\eta} > 0$ that depends only on τ and η. By Gao and Wang (2022, (22)), we have for all but finitely many p's that

$$\sup_{u\in B_0(\ell)} u^\top\left(\frac{W_z^\top W_z}{4n}-\kappa_1 I_p\right)u \leq (4+o(1))\sqrt{\frac{(\ell+4)\log(10ep/\ell)}{n}}\left\{(\kappa_1+\kappa_2)\sqrt{n/p}+\kappa_1\right\},$$

where $\kappa_2 > 0$ is again a constant depending only on τ and η. This, together with the first claim of Lemma 9, implies that with probability 1, for all but finitely many p's, we have

$$\sup_{t\in[z]}\sup_{u\in B_0(\ell)} \frac{4tn}{z}u^\top\left(\frac{W_z^\top W_z}{4n} - \kappa_1 I_p\right)v \leq C''_{\tau,\eta}\sqrt{\ell p \log(ep/\ell)}. \qquad \text{(S.2)}$$

2

The conclusion follows by combining (S.1) and (S.2), and the corresponding inequality for $t \in [n-1] \setminus [z]$. $\qquad\square$

Proof of Theorem 2. Applying Proposition 1, we have with probability 1 that for all but finitely many p's that

$$\sup_{t \in [z]} \sup_{u \in B_0(\ell)} u^\top \left(W_t^\top W_z - \frac{4t(n-z)(n-p)}{n^2} I_p \right) \frac{\theta}{\|\theta\|_2} \lesssim_{\tau,\eta} \sqrt{p\ell \log p}. \quad \text{(S.3)}$$

Let $S := \text{supp}(\theta)$. Taking $\ell = k$ in (S.3), we have

$$\sup_{\alpha n \le t \le z} \left\| \sqrt{\frac{n}{t(n-t)}} \left(\left(W_t^\top W_z - \frac{4t(n-z)(n-p)}{n^2} I_p \right) \theta \right)_S \right\|_2$$

$$\le \sup_{\alpha n \le t \le z} \sup_{u \in B_0(k)} u^\top \left(\sqrt{\frac{n}{t(n-t)}} W_t^\top W_z - \frac{4t^{1/2}(n-z)(n-p)}{n^{3/2}(n-t)^{1/2}} I_p \right) \theta$$

$$\lesssim_{\tau,\eta,\alpha} \sqrt{k \log p} \|\theta\|_2. \quad \text{(S.4)}$$

By (S.3) and the second claim of Lemma 9, with probability 1 that for all but finitely many p's,

$$\sup_{t \in [z]} \|(W_t^\top W_z \theta)_{S^c}\|_2 = \sup_{t \in [z]} \left\| \left(\left\{ W_t^\top W_z - \frac{4t(n-z)(n-p)}{n^2} I_p \right\} \theta \right)_{S^c} \right\|_2$$

$$\lesssim_{\tau,\eta} p\sqrt{\log p} \|\theta\|_2.$$

For any $Q \in \mathbb{O}^{p \times p}$, we have $X \stackrel{\mathrm{d}}{=} XQ^\top =: \tilde{X}$, and the latter has the corresponding sketching matrix $\tilde{A} = AQ^\top$ because $\tilde{A}^\top \tilde{A} = I_{n-p}$ and $\tilde{A}^\top \tilde{X} = QA^\top XQ^\top = \mathbf{0}_{(n-p) \times p}$. As such, for any $Q \in \mathbb{O}^{p \times p}$ such that $Q\theta = \theta$, $Q(W_t^\top W_z \theta) = (QW_t Q^\top)^\top (QW_z Q^\top)\theta \stackrel{\mathrm{d}}{=} W_t^\top W_z \theta$. In particular, $(W_t^\top W_z \theta)_{S^c}$ is spherically symmetric on $\mathbb{R}^{p-k}$. Hence, by Lemma 14 (with a choice of $\delta = 2p^{-4}$), with probability 1 we have for all but finitely many p's that

$$\sup_{\alpha n \le t \le z} \left\| \sqrt{\frac{n}{t(n-t)}} (W_t^\top W_z \theta)_{S^c} \right\|_\infty \le \sup_{\alpha n \le t \le z} \|(W_t^\top W_z \theta)_{S^c}\|_2 \sqrt{\frac{n}{t(n-t)}} \frac{4\sqrt{\log p}}{\sqrt{p-k}}$$

$$\lesssim_{\tau,\eta,\alpha} \|\theta\|_2 \log p \quad \text{(S.5)}$$

Let $X = QT$ be the QR decomposition of X and define $B_t := Q_{(0,t]}^\top Q_{(0,t]}$. By Equation (16) in the proof of Gao and Wang (2022, Proposition 8), there

3

exists $C_\eta > 0$, depending only on η that for any fixed p and t and $j \in [p]$, with probability $1 - p^{-4}$, we have

$$\frac{1}{n}(W_t^\top W_t)_{j,j} \leq \frac{4\mathrm{tr}(B_t(I_p - B_t))}{p} + C_\eta \sqrt{\frac{\log p}{n}} \leq 1 + C_\eta \sqrt{\frac{\log p}{n}},$$

where the final inequality follows from the fact that $\|B_t\|_{\mathrm{op}} \leq 1$. Taking union bounds over $j \in [p]$ and $t \in [z]$, and applying the Borel–Cantelli lemma, we have with probability 1 that for all but finitely many p's,

$$\sup_{\alpha n \leq t \leq z} \sup_{j \in [p]} \frac{n}{t(n-t)}(W_t^\top W_t)_{j,j} \lesssim_{\tau,\alpha} 1.$$

Furthermore, applying the Gaussian tail bound followed by a union bound, we have with probability 1 for all but finitely many p's that

$$\sup_{\alpha n \leq t \leq z} \left\| \sqrt{\frac{n}{t(n-t)}} W_t^\top \xi \right\|_\infty \leq 4\sqrt{\log p} \sup_{\alpha n \leq t \leq z} \sup_{j \in [p]} \frac{n}{t(n-t)}(W_t^\top W_t)_{j,j} \lesssim_{\tau,\alpha} \sqrt{\log p}.$$
$$\text{(S.6)}$$

We now work on the probability 1 event Ω, such that (S.4), (S.5), (S.6) all hold for all but finitely many p's.

For sufficiently large $c_{\tau,\eta,\alpha}$, we have the right-hand side of (S.5) and (S.6) are both dominated by $\lambda/2$. Hence, on Ω, we have for all $t \in [\alpha n, z]$ that

$$H_t = \left\| \mathrm{soft}\left(\sqrt{\frac{n}{t(n-t)}} W_t^\top (W_z \theta + \xi), \lambda \right) \right\|_2$$

$$= \left\| \mathrm{soft}\left(\sqrt{\frac{n}{t(n-t)}} (W_t^\top W_z)_{S,S} \theta_S + \sqrt{\frac{n}{t(n-t)}} (W_t^\top \xi)_S, \lambda \right) \right\|_2$$

Writing $\tilde{H}_t := \left\| \sqrt{\frac{n}{t(n-t)}} (W_t^\top W_z)_{S,S} \theta_S \right\|_2$, we have by the triangle inequality and (S.6) that on Ω,

$$\sup_{t \in [\alpha n, z]} \left| H_t - \tilde{H}_t \right| \lesssim_{\tau,\alpha} \sqrt{k}\lambda + \sqrt{k \log p}. \qquad \text{(S.7)}$$

Recall the definition of γ_t in (3.3) and write

$$h_t := \gamma_t \|\theta\|_2 = \frac{4t^{1/2}(n-z)(n-p)}{n^{3/2}(n-t)^{1/2}} \|\theta\|_2 = \frac{4(n-p)}{n} \sqrt{\frac{t}{n(n-t)}} (n-z)\|\theta\|_2.$$

4

Then by (S.4), we also have on Ω that

$$\sup_{t \in [\alpha n, z]} |\tilde{H}_t - h_t| \lesssim_{\tau, \eta, \alpha} \sqrt{k \log p} \|\theta\|_2. \tag{S.8}$$

By a symmetric argument, both (S.7) and (S.8) hold for $t \in [z, (1 - \alpha)n]$ and consequently for $t \in [\alpha n, (1 - \alpha)n]$ (with perhaps a slightly different constant). Consequently, we have that with probability 1 for all but finitely many p,

$$\sup_{t \in [\alpha n, (1-\alpha)n]} |H_t - h_t| \leq C_1 \left(\sqrt{k} \lambda + \sqrt{k \log p} \|\theta\|_2 \right),$$

where C_1 depends only on τ, η and α.

If $\theta = 0$, then $h_t = 0$ for all t, and hence for sufficiently large $C_{\tau, \eta, \alpha}$, we have with probability 1 for all but finitely many p's that $\sup_{t \in [\alpha n, (1-\alpha)n]} |H_t| \leq T$ and thus the first conclusion holds.

For the second conclusion, we have for some C_2 depending only on τ, η and α that

$$|H_z| \geq |h_z| - \sup_{t \in [\alpha n, (1-\alpha)n]} |H_t - h_t| \geq C_2 \sqrt{n} \|\theta\|_2 - C_1 (\sqrt{k} \lambda + \sqrt{k \log p} \|\theta\|_2)$$

$$\geq C_2 (1 - o(1)) \sqrt{n} \|\theta\|_2 - C_1 c_{\tau, \eta, \alpha} \sqrt{k} \log p.$$

The signal size condition on $\|\theta\|_2$ then ensures that $|H_z| \geq (C_2 c'_{\tau, \eta, \alpha}/2 - C_1 c_{\tau, \eta, \alpha}) \sqrt{k} \log p$. Hence, for sufficiently large $c'_{\tau, \eta, \alpha}$, we can ensure that with probability 1 for all but finitely many p's, we have $\max_{t \in [\alpha n, (1-\alpha)n]} |H_t| \geq |H_z| \geq T$, completing the proof. $\qquad \square$

The following proposition shows that the estimated projection direction $\hat{v}$ is well-aligned with this oracle direction.

Proposition 7. *Assume Conditions 1 and 2 and that data (X, Y) are generated according to (1.1) with $\nu = 1$. Suppose that $k \leq p/2$ and that $\min(\tau, 1 - \tau) \geq \alpha$ for some known α. There exists $c_{\tau, \eta, \alpha}, C_{\tau, \eta, \alpha} > 0$, depending only on τ, η, α, such that if $\lambda > c_{\tau, \eta, \alpha} \max(1, \|\theta\|_2) \log p$, then the projection direction estimator $\hat{v}$ in Algorithm 1' (defined below Proposition 1) satisfies with probability 1 for all but finitely many p's that*

$$\sin \angle(\hat{v}, \theta) \leq C_{\tau, \eta, \alpha} \frac{\lambda \sqrt{k}}{\sqrt{n} \|\theta\|_2}.$$

5

Proof. Define for each $t \in [\alpha n, (1-\alpha)n]$ a vector $\tilde{Q}_t \in \mathbb{R}^p$ such that $(\tilde{Q}_t)_{S^c} := 0$ and

$$(\tilde{Q}_t)_S := \sqrt{\frac{n}{t(n-t)}}(W_t^\top W_z)_{S,S}\theta_S.$$

By (S.5) and (S.6) and their symmetric results for $t \in [z, (1-\alpha)n]$, we have with probability 1 for all but finitely many p's that

$$\sup_{\alpha n \leq t \leq (1-\alpha)n} \|Q_t - \tilde{Q}_t\|_\infty \lesssim_{\tau,\eta,\alpha} \max(1, \|\theta\|_2) \log p \qquad (\text{S.9})$$

Recall the definition of $\gamma = (\gamma_t)_{t \in [n-1]}$ in (3.3). Applying (S.3) and its symmetric result for $t \geq z$ with $\ell = 1$, we have with probability 1 for all but finitely many p's that

$$\sup_{t \in [n-1]} \|\tilde{Q}_t - \theta\gamma_t\|_\infty \lesssim_{\tau,\eta,\alpha} \|\theta\|_2 \sqrt{\log p}. \qquad (\text{S.10})$$

By symmetry, both (S.9) and (S.10) are still valid when we replace the supremum over $t \in [z, (1-\alpha)n]$ instead. Define $\tilde{Q} := (\tilde{Q}_{\lceil \alpha n \rceil}, \ldots, \tilde{Q}_{\lfloor (1-\alpha)n \rfloor})^\top$ and $\gamma := (\gamma_t)_{t \in [\alpha n, (1-\alpha)n]}$. We then have

$$\|Q - \theta\gamma^\top\|_\infty \leq \|Q - \tilde{Q}\|_\infty + \|\tilde{Q} - \theta\gamma^\top\|_\infty \lesssim_{\tau,\eta,\alpha} \max(1, \|\theta\|_2) \log p. \quad (\text{S.11})$$

By Wang and Samworth (2018, Propositions 2 and 4 in the online supplement), for $c_{\tau,\eta,\alpha}$ large enough such that $\lambda \geq \|Q - \theta\gamma^\top\|_\infty$, we have

$$\sin \angle(\hat{v}, \theta) \lesssim_{\tau,\eta,\alpha} \frac{\lambda\sqrt{kn}}{\|\theta\|_2\|\gamma\|_2},$$

whence the desired form follows by noting $\|\gamma\|_2 \asymp_{\tau,\eta,\alpha} n$ by Condition 2. $\square$

Proof of Theorem 3. To simplify exposition, all statements should be interpreted as valid with probability 1 for all but finitely many p's. Write $v := \theta/\|\theta\|_2$ for simplicity, and note that $\|v\|_0 = \|\theta\|_0 \leq k$. Since the estimator $\hat{z}$ is unchanged if we replace $\hat{v}$ by $-\hat{v}$ in Algorithm 1, we may assume without loss of generality that $\rho := \hat{v}^\top v \geq 0$. Our strategy is to view $(\hat{v}^\top Q_t : t \in [\alpha n, (1 - \alpha)n])$ as a perturbation of a multiple of $(\gamma_t : t \in [\alpha n, (1 - \alpha)n])$, which is maximized at z. By (S.11), we may

6

choose $c_{\tau,\eta,\alpha}$ large enough such that $\|Q - \theta\gamma^\top\|_\infty \le \lambda$. By Proposition 7, we then have

$$\max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(Q_t - \theta\gamma_t)\right| \le \|\hat{v}\|_1\|Q - \theta\gamma^\top\|_{\max} \le (\|v\|_1 + \|\hat{v} - v\|_1)\lambda$$

$$\lesssim_{\tau,\eta,\alpha} (\sqrt{k} + \sqrt{p}\|\hat{v} - v\|_2)\lambda.$$

From Proposition 7, there exists $C'_{\tau,\eta,\alpha} > 0$, depending only on τ, η, α, such that

$$\|\hat{v} - v\|_2 \le 2\sin\angle(\hat{v}, \theta) \le C'_{\tau,\eta,\alpha}\frac{\lambda\sqrt{k}}{\sqrt{n}\|\theta\|_2}, \tag{S.12}$$

which implies that

$$\max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(Q_t - \theta\gamma_t)\right| \lesssim_{\tau,\eta,\alpha} \frac{\lambda^2\sqrt{k}}{\|\theta\|_2}. \tag{S.13}$$

We may further assume that

$$\frac{\lambda\sqrt{k}}{\sqrt{n}\|\theta\|_2} \le \frac{\lambda^2\sqrt{k}}{\sqrt{n}\|\theta\|_2^2} \le \frac{1}{C'_{\tau,\eta,\alpha}} \tag{S.14}$$

for all p's, since for p where this is not satisfied the result is trivially true. Then, $\sin\angle(\hat{v}, \theta) \le 1/2$ and thus $\rho = \{1 - \sin^2\angle(\hat{v}, \theta)\}^{1/2} \ge 1/2$. Consequently, from (S.13) and (S.14), increasing $C'_{\tau,\eta,\alpha}$ if necessary, we have

$$\hat{v}^\top\theta\gamma_z = \rho\|\theta\|_2\frac{4z^{1/2}(n - z)^{1/2}(n - p)}{n^{3/2}} \ge 2\max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(Q_t - \theta\gamma_t)\right|,$$

which implies in particular that $\hat{v}^\top Q_{\hat{z}} > 0$. Now, since $z = \arg\max_{t\in[n-1]}\gamma_t$ and $\hat{z} = \arg\max_{t\in[\alpha n,(1-\alpha)n]}\hat{v}^\top Q_t$, we have from (S.13) that

$$\hat{v}^\top\theta\gamma_z - \hat{v}^\top\theta\gamma_{\hat{z}} \le \hat{v}^\top Q_z - \hat{v}^\top Q_{\hat{z}} + 2\max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(Q_t - \theta\gamma_t)\right| \lesssim_{\tau,\eta,\alpha} \frac{\lambda^2\sqrt{k}}{\|\theta\|_2}. \tag{S.15}$$

On the other hand, by Wang and Samworth (2018, Lemma 7), we have

$$\inf_{t\in[z-\min\{z,n-z\}/2,z+\min\{z,n-z\}/2]}\frac{\hat{v}^\top\theta\gamma_z - \hat{v}^\top\theta\gamma_t}{|z - t|} \gtrsim_{\tau,\eta} \|\theta\|_2\sqrt{n}. \tag{S.16}$$

We arrive at the conclusion by combining (S.15) and (S.16). □

7

Proof of Theorem 4. As in the proof of Theorem 3, all statements are valid with probability 1 for all but finitely many p's, and we may assume without loss of generality that $\hat{v}^\top v \geq 0$. Let $\tilde{Q}_t$ be as in the proof of Proposition 7. The main difference to the proof of Theorem 3 will be an improvement of (S.13) using the independence between $\hat{v}$ and $Q_t - \theta\gamma_t$. Specifically, since

$$\{t(n-t)/n\}^{1/2}\hat{v}^\top(Q_t - \tilde{Q}_t) = \hat{v}^\top W_t^\top(W_z\theta + \xi) - \hat{v}_S^\top(W_t^\top W_z)_{S,S}\theta_S$$
$$= \hat{v}_{S^c}^\top(W_t^\top W_z\theta)_{S^c} + \hat{v}^\top W_t^\top\xi,$$

we have that

$$\max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(Q_t - \theta\gamma_t)\right| \leq \max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(\tilde{Q}_t - \theta\gamma_t) + \sqrt{\frac{n}{t(n-t)}}\hat{v}^\top(W_t^\top W_z\theta)_{S^c}\right|$$
$$+ \max_{t\in[\alpha n,(1-\alpha)n]}\left|\sqrt{\frac{n}{t(n-t)}}\hat{v}^\top W_t^\top\xi\right| \quad\quad \text{(S.17)}$$

We control the two terms on the right-hand side of (S.17) separately. By (S.10), (S.5), the Cauchy–Schwarz inequality and finally (S.12), the first term from the above display (S.17) is bounded by

$$\|\hat{v}\|_1\left(\|\tilde{Q} - \theta\gamma^\top\|_{\max} + \max_{t\in[\alpha n,(1-\alpha)n]}\left\|\sqrt{\frac{n}{t(n-t)}}\|\hat{v}\|_2\|(W_t^\top W_z\theta)_{S^c}\|_2\right\|_\infty\right)$$
$$\lesssim_{\tau,\eta,\alpha} (\sqrt{k}\|v\|_2 + \sqrt{p}\|\hat{v} - v\|_2)\|\theta\|_2\log p \lesssim_{\tau,\eta,\alpha} \lambda\sqrt{k}\log p. \quad\quad \text{(S.18)}$$

On the other hand, since $\hat{v}$, W_t and ξ are mutually independent, we have $\hat{v}^\top W_t^\top\xi \mid (\hat{v}, W_t) \sim N(0, \|W_t\hat{v}^\top\|_2^2)$. By Gao and Wang (2022, (S.8)), we have

$$\sup_{t\in[\alpha n,(1-\alpha)n]}\left\|\frac{1}{n}W_t^\top W_t\right\|_{\mathrm{op}} \lesssim_{\tau,\eta,\alpha} 1.$$

Hence, by Gaussian tail bounds followed by a union bound, we have

$$\max_{t\in[\alpha n,(1-\alpha)n]}\left|\sqrt{\frac{n}{t(n-t)}}\hat{v}^\top W_t^\top\xi\right| \lesssim_{\tau,\eta,\alpha} \sqrt{\log p}. \quad\quad \text{(S.19)}$$

Substituting (S.18) and (S.19) into (S.17), we have

$$\max_{t\in[\alpha n,(1-\alpha)n]}\left|\hat{v}^\top(Q_t - \theta\gamma_t)\right| \lesssim_{\tau,\eta,\alpha} \lambda\sqrt{k}\log p. \quad\quad \text{(S.20)}$$

8

Following the same argument as in the proof of Theorem 3, with (S.20) replacing (S.13), we arrive at the following counterpart to (S.15):

$$\hat{v}^\top \theta \gamma_z - \hat{v}^\top \theta \gamma_{\hat{z}} \leq \hat{v}^\top Q_z - \hat{v}^\top Q_{\hat{z}} + 2 \max_{t \in [\alpha n, (1-\alpha)n]} \left| \hat{v}^\top (Q_t - \theta \gamma_t) \right| \lesssim_{\tau, \eta, \alpha} \lambda \sqrt{k} \log p.$$
(S.21)

Combining (S.21) with (S.16), the proof is complete. $\square$

Proof of Theorem 5. First observe that since

$$\mathbb{P}(\Omega_0^c) \leq \sum_{i=1}^{\nu} \prod_{m=1}^{M} \left(1 - \mathbb{P}\left((s_m, e_m] \in \mathcal{I}_i \right) \right) \leq \nu \left(1 - \Delta_\tau^2 / 36 \right)^M \leq \nu \exp(-\Delta_\tau^2 M / 36),$$

the second conclusion follows immediately from the first one. To establish the first conclusion, we henceforth work on the event $\Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3$.

For $0 \leq s < e \leq n$, we define the following sets

$$\begin{aligned}
\mathcal{M}^{(s,e]} &:= \{ m \in [M] : s \leq s_m < e_m \leq e \}, \\
\mathcal{R}^{(s,e]} &:= \{ m \in \mathcal{M}^{(s,e]} : \psi(D_{(s_m + n\varpi, e_m - n\varpi]}) = 1 \}, \\
\mathcal{Z}^{(s,e]} &:= \{ i \in [\nu] : z_i \in (s, e] \}, \\
\mathcal{Z}_{\text{good}}^{(s,e]} &:= \{ i \in [\nu] : z_i \in (s, e], \min\{ z_i - s, e - z_i \} \geq n\Delta_\tau / 2 \}, \\
\mathcal{Z}_{\text{bad}}^{(s,e]} &:= \{ i \in [\nu] : z_i \in (s, e], \min\{ z_i - s, e - z_i \} < n\phi_i \}.
\end{aligned}$$

Note that on the event Ω_0, we can associate each true changepoint z_i with an $m_i \in [M]$ such that $(s_{m_i}, e_{m_i}] \in \mathcal{I}_i$. On Ω_2, we have

$$\left\{ m_i : i \in \mathcal{Z}_{\text{good}}^{(s,e]} \right\} \subseteq \mathcal{R}^{(s,e]}. \tag{S.22}$$

Recall the assumption $\phi < \varpi$. For any $(s_0, e_0] \subset (s, e]$ such that $(s_0, e_0] \cap \{ z_i : i \in [\nu], z_i \in (s, e] \} \subseteq \mathcal{Z}_{\text{bad}}^{(s,e]}$, we have $(s_0, e_0] \in \mathcal{I}_0$ and hence $\psi(D_{(s_0 + n\varpi, e_0 - n\varpi]}) = 0$ on Ω_1.

For any set of changepoints $\hat{Z}$, we can partition the original timeline $(0, n]$ into $|\hat{Z}| + 1$ segments, which we call segments induced by $\hat{Z}$. We now prove by induction that as we update $\hat{Z}$ throughout the recursion of Algorithm 3, for any $(s, e]$ induced by $\hat{Z}$, we have $\mathcal{Z}^{(s,e]} = \mathcal{Z}_{\text{good}}^{(s,e]} \cup \mathcal{Z}_{\text{bad}}^{(s,e]}$. The base case is trivially true as at the beginning of the algorithm, $\hat{Z} = \emptyset$, so the only segment induced is $(0, n]$ so $\mathcal{Z}^{(s,e]} = \mathcal{Z}_{\text{good}}^{(s,e]}$ and $\mathcal{Z}_{\text{bad}}^{(s,e]} = \emptyset$ by our assumption that $z_i - z_{i-1} \geq n\Delta_\tau$ for all $i \in [\nu + 1]$.

9

Now assuming that the inductive hypothesis holds at some iteration of the recursion in Algorithm 3. We show that the claimed statement still holds if a new changepoint is estimated. Let $\hat{Z}$ be the set of changepoints identified before this new changepoint, say $\hat{z}_*$, is added to it. We necessarily have $\hat{z}_* = \hat{z}(D_{(s',e']}) + s'$ for some $(s', e'] \in \hat{\mathcal{Z}}^{(s,e]}$ where $(s, e]$ is induced by $\hat{Z}$. From the inductive hypothesis, we have $\mathcal{Z}^{(s,e]} = \mathcal{Z}^{(s,e]}_{\text{good}} \cup \mathcal{Z}^{(s,e]}_{\text{bad}}$ and $\mathcal{Z}^{(s,e]}_{\text{good}}$ is necessarily non-empty, for otherwise all changepoints in $(s, e]$ are within a distance of $n\phi$ to the boundary of the interval, which together with the fact that $\phi < \varpi$, implies that $\mathcal{M}^{(s,e]} \cap \mathcal{R}^{(s,e]} = \emptyset$, contradicting the fact that a new changepoint is identified. Thus, there exists some $i' \in \mathcal{Z}^{(s,e]}_{\text{good}}$, which by (S.22) means that $m_{i'} \in \mathcal{R}^{(s,e]}$. By the definition of m_0 in Line 6 of Algorithm 3, we have $e_{m_0} - s_{m_0} \leq e_{m_{i'}} - s_{m_{i'}} \leq n\Delta_\tau$. Thus, from the condition of the theorem, we have that $(s_{m_0}, e_{m_0}]$ contains at most one changepoint. If $(s_{m_0}, e_{m_0}] \cap \{z_i : i \in \mathcal{Z}^{(s,e]}\} = \emptyset$, then on Ω_1, $\psi(D_{(s_{m_0}+n\varpi, e_{m_0}-n\varpi]}) = 0$, contradicting $m_0 \in \mathcal{R}^{(s,e]}$. If $(s_{m_0}, e_{m_0}]$ contains a single changepoint z_i for $i \in \mathcal{Z}^{(s,e]}_{\text{bad}}$, then since $\phi < \varpi$, we again have on Ω_1 that $\psi(D_{(s_{m_0}+n\varpi, e_{m_0}-n\varpi]}) = 0$, a contradiction. By the inductive hypothesis, this implies that $(s_{m_0}, e_{m_0}]$ contains exactly one true change-point z_{i_0} for some $i_0 \in \mathcal{Z}^{(s,e]}_{\text{good}}$ and that $\min\{e_{m_0} - z_{i_0}, z_{i_0} - s_{m_0}\} \geq n\varpi$. Hence, $(s_{m_0}, e_{m_0}] \in \tilde{\mathcal{I}}_{i_0}$, and thus on Ω_3, we have $|\hat{z}_* - z_{i_0}| \leq n\phi_{i_0}$.

We finally check that the two new segments induced by $\hat{Z} \cup \{\hat{z}_*\}$, say $(\hat{z}_{\text{left}}, \hat{z}_*]$ and $(\hat{z}_*, \hat{z}_{\text{right}}]$ for $\hat{z}_{\text{left}} < \hat{z}_* < \hat{z}_{\text{right}}$, still satisfy the inductive hypothesis. By symmetry, we may assume without loss of generality that $\hat{z}_* \leq z_{i_0}$. Since $|z_{i_0} - \hat{z}_*| \leq n\phi$, we have $i_0 \in \mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]}_{\text{bad}}$. For any $i \in \mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]}$ such that $i < i_0$, we have $\hat{z}_* - z_i \geq z_{i_0} - z_i \geq n\Delta_\tau$, and thus $z_i \in \mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]}_{\text{good}} \cup \mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]}_{\text{bad}}$ by the inductive hypothesis, consequently, $\mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]} = \mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]}_{\text{good}} \cup \mathcal{Z}^{(\hat{z}_{\text{left}}, \hat{z}_*]}_{\text{bad}}$. Similarly, for $i \in \mathcal{Z}^{(\hat{z}_*, \hat{z}_{\text{right}}]}$, we have $i > i_0$ and $z_i - \hat{z}_* \geq z_i - z_{i_0} - (\hat{z}_* - z_{i_0}) \geq n\Delta_\tau - \phi \geq n\Delta_\tau/2$. Again by the inductive hypothesis, $\mathcal{Z}^{(\hat{z}_*, \hat{z}_{\text{right}}]} = \mathcal{Z}^{(\hat{z}_*, \hat{z}_{\text{right}}]}_{\text{good}} \cup \mathcal{Z}^{(\hat{z}_*, \hat{z}_{\text{right}}]}_{\text{bad}}$. This completes the induction.

As a consequence of the above inductive argument, we have shown that a new changepoint will be identified in Algorithm 3 if and only if $(s, e] \cap \{z_i : i \in \mathcal{Z}^{(s,e]}_{\text{good}}\} \neq \emptyset$. Thus, from the inductive claim, at the end of the recursion, each changepoint, say z_i, must be less than $n\phi_i$ away from one of the end points of the segments induced by $\hat{Z}$. This, as well as the assumption that $z_i - z_{i-1} \geq n\Delta_\tau$ for all $i \in [\nu+1]$, means that $|\hat{Z}| = \nu$ and that $|\hat{z}_i - z_i| \leq n\phi_i$ as desired. $\qquad\square$

Proof of Corollary 6. First we write $\phi_i = \frac{C'\lambda\sqrt{k}\log p}{\sqrt{n}\|\theta^{(i)}\|_2}$ and define Ω_0, Ω_1, Ω_2 and Ω_3 as in Theorem 5. Observe that Ω_1, Ω_2 and Ω_3 has implicit dependence on p, whereas in the specific coupling (3.4) considered in this theorem, Ω_0 does not vary with p. We have from the proof of Theorem 5 that $\mathbb{P}(\Omega_0) \geq 1 - \nu e^{-\Delta_\tau^2 M/36}$. Hence, it suffices show that on Ω_0, we have for all but finitely many p's that Ω_1, Ω_2 and Ω_3 hold simultaneously. We keep in mind that M is fixed and finite, and for the rest of the proof, we condition on a realization of $(\tilde{s}_m, \tilde{e}_m)_{m=1}^M$ as in (3.4) such that Ω_0 holds.

Let $\mathcal{I}_0$, $\mathcal{I}_i$ and $\tilde{\mathcal{I}}_i$ be defined as in Theorem 5. We first establish Ω_1 and Ω_2. For any interval $(s, e]$ with $e - s \leq p$, $\psi = 0$ by definition. For every $(s, e] \in \cup_{0 \leq i \leq \nu} \mathcal{I}_i$ whose length is longer than p, the fixed-ratio regime Condition 2 is true by the generating mechanism of the intervals in (3.4), and it is straightforward to verify that Theorem 2 applies. As a result, there exist c, c', C, which may depend on $(s, e]$, such that the conclusion of Theorem 2 holds for each $(s, e] \in \cup_{0 \leq i \leq \nu} \mathcal{I}_i$ with $e - s > p$. Inspecting the proof of Theorem 2 shows that we can take the maximum of all such c, c', C's so that the conclusion of Theorem 2 holds for all intervals in $\cup_{0 \leq i \leq \nu} \mathcal{I}_i$ with length longer than p. As such, we have, for all but finitely many p's, Ω_1 and Ω_2 holds.

Now we turn to Ω_3. Again, by reasoning similar to the above, we see the conditions of Theorem 4 hold for each $(s, e] \in \tilde{\mathcal{I}}_i$, and for the above-mentioned specific choices of c, which may depend on $(s, e]$, the conclusion of Theorem 4 holds for each $(s, e] \in \tilde{\mathcal{I}}_i$. We can again take the maximum of all such c's so that for all intervals in $\tilde{\mathcal{I}}_i$ for all $i \in [\nu]$, the conclusion of Theorem 4 holds, i.e.,

$$\frac{\hat{z}(D_{(s,e]}) - (z_i - s)}{n} \leq C'_{(s,e]}\frac{\lambda\sqrt{k}\log p}{\sqrt{n}\|\theta\|_2}.$$

Setting, e.g., $C' = \max_{i \in [\nu]} \max_{(s,e] \in \tilde{\mathcal{I}}_i} C'_{(s,e]}$, we have, for all but finitely many p's, Ω_3 holds. Invoking Theorem 5 completes the proof. $\qquad\square$

B   Ancillary results

We collect here the ancillary results and their proofs.

The following lemma shows our assumption $p < n$ is necessary in the sense that otherwise it is impossible to test the null $\theta = 0$ against the alternative $\theta = \theta^*$ for any $\theta^* \in \mathbb{R}^p$.

Lemma 8. *Fix $n, p \in \mathbb{N}$, $z \in \{1, \ldots, n-1\}$, $\sigma > 0$ and a distribution P_X on $\mathbb{R}^{n \times p}$. For $\beta, \theta \in \mathbb{R}^p$, define $P_{\beta, \theta}$ to be the joint distribution of $x_1, \ldots, x_n, y_1, \ldots, y_n$ generated by*

$$(x_1, \ldots, x_n)^\top \sim P_X$$
$$y_i = x_i^\top (\beta + \theta \mathbb{1}_{\{i > z\}}) + \epsilon_i, \quad \epsilon_i \mid (x_1, \ldots, x_n) \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Now, for a fixed $\theta^* \neq 0 \in \mathbb{R}^p$, set $\mathcal{P} := \{P_{\beta, 0} : \beta \in \mathbb{R}^p\}$ be the space of null hypotheses and $\mathcal{Q} = \{P_{\beta, \theta^*} : \beta \in \mathbb{R}^p\}$ the space of alternative hypotheses. If $n \leq p$, then

$$\inf_\psi \Big\{ \sup_{P \in \mathcal{P}} P(\psi = 1) + \sup_{Q \in \mathcal{Q}} Q(\psi = 0) \Big\} = 1,$$

where the infimum is taken over all tests $\psi : \mathbb{R}^{pn} \times \mathbb{R}^n \to \{0, 1\}$.

Proof. The left-hand side is trivially upper-bounded by 1 (e.g. taking ψ to be constantly zero), hence it suffices to show that 1 is also a lower bound. Let $\pi := N_p(0, \tau^2 I_p)$ be a prior on β, for some $\tau > 0$, and define

$$P_\pi := \int P_{\beta, 0} \, d\pi(\beta), \quad Q_\pi := \int P_{\beta, \theta^*} \, d\pi(\beta).$$

We then have by Le Cam's two-point test lemma Samworth and Shah (2025, Lemma 8.4)

$$\inf_\psi \Big\{ \sup_{P \in \mathcal{P}} P(\psi = 1) + \sup_{Q \in \mathcal{Q}} Q(\psi = 0) \Big\} \geq \inf_\psi \Big\{ P_\pi(\psi = 1) + Q_\pi(\psi = 0) \Big\}$$
$$= 1 - d_{\text{TV}}(P_\pi, Q_\pi),$$

where d_{TV} denotes the total variation distance. It remains to show that $d_{\text{TV}}(P_\pi, Q_\pi)$ can be made arbitrarily close to 0 by taking appropriate values of τ. Under P_π, $X = (x_1, \ldots, x_n)^\top$ has marginal distribution P_X and $Y = (y_1, \ldots, y_n)^\top$ has conditional distribution

$$Y \mid X \sim N(0, K), \quad \text{where } K = \sigma^2 I_n + \tau^2 X X^\top.$$

Similarly, under Q_π, $X \sim P_X$ and

$$Y \mid X \sim N(\mu, K), \quad \text{where } \mu = (\mathbb{1}_{\{i > z\}} x_i^\top \theta^*)_{i=1}^n.$$

Hence,

$$d_{\mathrm{TV}}(P_\pi, Q_\pi) = \int_X d_{TV}\big(N(0, K), N(\mu, K)\big) \, dP_X(X)$$
$$= \int_X \left\{ 1 - 2\Phi\left(-\frac{1}{2}\sqrt{\mu^\top K^{-1}\mu}\right) \right\} dP_X(X).$$

Since $n \leq p$, $XX^\top$ has full rank and is positive definite P_X-almost surely, which implies that the minimum eigenvalue of K diverges to infinity as $\tau \to \infty$. Thus, $\lim_{\tau \to \infty} \mu^\top K^{-1}\mu \to 0$ P_X-almost surely, and by dominated convergence, we have $\lim_{\tau \to \infty} d_{\mathrm{TV}}(P_\pi, Q_\pi) = 0$. $\qquad\square$

Lemma 9. *Fix $A \in \mathbb{R}^{p \times p}$ and $k \in [p]$. The following are true.*

1. *If A is symmetric, then $\sup_{u,v \in B_0(k)} u^\top Av \leq \sup_{v \in B_0(2k)} v^\top Av$.*

2. *$\sup_{v \in B_0(k)} \|Av\|_2 \leq \sqrt{p/k} \sup_{u,w \in B_0(k)} u^\top Aw$.*

Proof. For the first claim, fix $u, v \in B_0(k)$ and let S and T be their respective support. Then by the symmetry of A, we have

$$u^\top Av \leq \|A_{S,T}\|_{\mathrm{op}} \leq \|A_{S \cup T, S \cup T}\|_{\mathrm{op}} = \sup_{w \in \mathcal{S}^{2k-1}} w^\top A_{S \cup T, S \cup T} w \leq \sup_{v \in B_0(2k)} v^\top Av.$$

The first claim then follows by taking supremum on the left-hand side.

For the second claim, define $\psi := \sup_{u,w \in B_0(k)} u^\top Aw$. Write $\binom{[p]}{k} := \{S \subseteq B : |S| = k\}$. For any $v \in B_0(k)$, let $\tilde{v} := Av/\|Av\|_2$ and $T := \mathrm{supp}(v)$. Then by the Cauchy–Schwarz inequality, we have

$$\|Av\|_2 = \tilde{v}^\top Av = \frac{1}{\binom{p-1}{k-1}} \sum_{S \in \binom{[p]}{k}} \tilde{v}_S^\top A_{S,T} v_T \leq \frac{1}{\binom{p-1}{k-1}} \sum_{S \in \binom{[p]}{k}} \|\tilde{v}_S\|_2 \psi$$
$$\leq \left\{ \frac{1}{\binom{p-1}{k-1}} \sum_{S \in \binom{[p]}{k}} \|\tilde{v}_S\|_2^2 \right\}^{1/2} \left\{ \frac{1}{\binom{p-1}{k-1}} \sum_{S \in \binom{[p]}{k}} \psi^2 \right\}^{1/2} \leq \|\tilde{v}\|_2 \cdot \sqrt{\frac{p}{k}} \psi = \sqrt{\frac{p}{k}} \psi.$$

Taking supremum over v on the left-hand side, we arrive at the conclusion. $\qquad\square$

Suppose that $X = (x_1, \ldots, x_n)^\top$ is generated by independent $x_i \sim N_p(0, \Sigma)$ with $n \geq p$ and some positive definite Σ. Write $S = X^\top X$ and $S_1 = X_{(0,t]}^\top X_{(0,t]}$. By Mitra (1970), for any well-defined function $\phi : S \mapsto \phi(S)$ such that $\phi(S)S\phi(S)^\top = I_p$, $U = \phi(S)S_1\phi(S)^\top$ is said to have a matrix-variate Beta

13

distribution, i.e. $U \sim \text{Beta}_p(t/2, (n-t)/2)$. To the best of our knowledge, it is unclear what happens when $n < p$ in the literature. The following Lemma 10 and Corollary 11 effectively generalize the existing matrix-variate Beta distribution to the rank-deficient case of $n < p$.

Lemma 10. *Suppose $X \in \mathbb{R}^{n \times p}$ has independent $N(0,1)$ entries and write $r := \min\{n, p\}$. There exists an almost surely unique way of writing $X = QR$ as its QR decomposition, where $Q \in \mathbb{O}^{n \times r}$ and $R \in \mathbb{R}^{r \times p}$ such that $R_{i,j} = 0$ for all $i > j$ and $R_{i,i} \geq 0$ for all $i \in [n]$. We have that Q and R are independent and $Q \sim \text{Unif}(\mathbb{O}^{n \times n})$, $R_{i,i}^2 \sim \chi^2(n - i + 1)$ and $R_{i,j} \sim N(0,1)$ for $i \in [n]$ and $j \in [p]$ with $i < j$. Furthermore, $B := Q_{(0,t]}^\top Q_{(0,t]} \sim \text{Beta}_r(t/2, (n-t)/2)$ and is independent of $X^\top X$.*

Proof. First we consider the case of $n \geq p$. Write the (almost surely) unique QR decomposition of X by $X = QR$ with $Q \in \mathbb{O}^{n \times p}$ and $R \in \mathbb{R}^{p \times p}$ being an upper triangular matrix with $R_{i,i} \geq 0$ for all $i \in [p]$. For any fixed $H \in \mathbb{O}^{n \times n}$, $HX \overset{\text{d}}{=} X$, whence $HQR \overset{\text{d}}{=} QR$. As such, the joint density of Q and R is constant for every possible value of $Q \in \mathbb{O}^{p \times n}$, whence Q and R are independent and $Q \sim \text{Unif}(\mathbb{O}^{p \times n})$. By (Muirhead, 2009, Theorem 3.2.14), we have $R_{i,i}^2 \sim \chi^2(n - i + 1)$ and that $R_{i,j} \sim N(0,1)$ and $R_{i,j}$ are independent for all $i \leq j$. We define $S := X^\top X$, $S_1 := X_{(0,t]}^\top X_{(0,t]}$ and $S_2 := X_{(t,n]}^\top X_{(t,n]}$. Define $S^{1/2} := R^\top$, and by Mitra (1970),

$$B := Q_{(0,t]}^\top Q_{(0,t]} = S^{-1/2} S_1 (S^{-1/2})^\top \sim \text{Beta}_p(t/2, (n-t)/2).$$

We note that B as a function of Q is independent of $X^\top X = R^\top R$, by the independence of Q and R.

Now we consider the case $n < p$. Write X_j as the jth column of X. Write $X = [X_\text{L} \mid X_\text{R}]$ where $X_\text{L} = [X_1 \mid \cdots \mid X_n]$ and $X_\text{R} := [X_{n+1} \mid \cdots \mid X_p]$. For X_L whose rank is almost surely n, there exists a unique QR decomposition such that $X_\text{L} = QR_\text{L}$. Take $R_\text{R} := Q^\top X_\text{R}$ and $R := [R_\text{L} \mid R_\text{R}]$, and we have $X = QR$, where both Q and R are almost surely unique. By the same argument as the case of $n \geq p$, we have Q and R are independent and $Q \sim \text{Unif}(\mathbb{O}^{n \times n})$. Applying the conclusion from the case of $n \geq p$ on $X_\text{L} = QR_\text{L}$, we have $(R_L)_{i,i} \sim \chi^2(n - i + 1)$ and $(R_L)_{i,j} \sim N(0,1)$. Furthermore, since $R_\text{R} = Q^\top X_\text{R}$ where both Q and X_R is independent of X_L and Q is independent of X_R, all entries of R_R are standard normals independent of R_L.

Applying the case of $n \geq p$ on $X_{\mathrm{L}} = QR_{\mathrm{L}}$, we have $B := Q_{(0,t]}^\top Q_{(0,t]} \sim$ $\mathrm{Beta}_n(t/2, (n-t)/2)$. By the same argument as before, B is independent of $X^\top X$. $\qquad \square$

Corollary 11. *Let $X = (x_1, \ldots, x_n)^\top$ where $x_i \sim N_p(0, \Sigma)$ where Σ is a positive definite matrix. Write $S = X^\top X$ and $S_1 := X_{(0,t]}^\top X_{(0,t]}$. Let $r := \min\{n,p\}$ and $\phi : \mathbb{R}^{p \times p} \to \mathbb{R}^{p \times r}$ be a function such that $\phi(S)S\phi(S)^\top = I_r$ for any positive semi-definite matrix $S \in \mathbb{R}^{p \times p}$ of rank r. Then $\phi(S)S_1\phi(S)^\top \sim \mathrm{Beta}_r(t/2, (n-t)/2)$ and is independent of $S = X^\top X$.*

Proof. By the positive definiteness of Σ, we find $\Sigma^{1/2} \in \mathbb{R}^{p \times p}$ such that $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^\top$. Find $\Sigma^{-1/2}$ such that $\Sigma^{-1/2}\Sigma^{1/2} = I_p$. Define $y_i := \Sigma^{-1/2}x_i \sim N_p(0, I_p)$ and $Y = X(\Sigma^{-1/2})^\top$ has independent $N(0,1)$ entries. Let $\tilde{S} := Y^\top Y = \Sigma^{-1/2}S(\Sigma^{-1/2})^\top$ and $\tilde{S}_1 := Y_{(0,t]}^\top Y_{(0,t]} = \Sigma^{-1/2}S_1(\Sigma^{-1/2})^\top$, whence by the definition of ϕ we have $\phi(S)S\phi(S)^\top = \phi(S)\Sigma^{1/2}\tilde{S}(\Sigma^{1/2})^\top\phi(S)^\top = I_r$. As such, we define $\tilde{\phi}(\tilde{S}) := \phi(\Sigma^{1/2}\tilde{S}(\Sigma^{-1/2})^\top)\Sigma^{1/2}$, which is a well-defined function, and have $\tilde{\phi}(\tilde{S})\tilde{S}\tilde{\phi}(\tilde{S})^\top = I_r$. Since $\tilde{\phi}(\tilde{S})\tilde{S}_1\tilde{\phi}(\tilde{S})^\top$ can be defined by Y with no dependence on Σ, it suffices to work on the case $\Sigma = I_p$, which we assume for the rest of the proof.

By Lemma 10, write the unique QR decomposition of X by $X = QR$ with $Q \in \mathbb{O}^{n \times r}$ and $R \in \mathbb{R}^{r \times p}$, whence $S = R^\top R$. Write $B = Q_{(0,t]}^\top Q_{(0,t]}$. Since $\phi(S)^\top R^\top R\phi(S) = I_r$, we have $H := R\phi(S) \in \mathbb{O}^{r \times r}$, whence $\phi(S)^\top S_1\phi(S) = H^\top Q_{(0,t]}^\top Q_{(0,t]}H = H^\top BH$. Since B as a function of Q is independent of S and R, it is independent of H. By noting $H^\top BH = \phi(S)S_1\phi(S)^\top$ and $B \overset{\mathrm{d}}{=} H^\top BH$ because $QH \overset{\mathrm{d}}{=} Q$ for any $H \in \mathbb{O}^{r \times r}$. The independence between $\phi(S)S_1\phi(S)$ and S follows from the fact that the distribution of $\phi(S)S_1\phi(S)$ is invariant conditionally on S. $\qquad \square$

Recall that $J_{(a_1, a_2]} := (J_{a_1+1}, \cdots, J_{a_2})^\top$ is the submatrix of J by taking only the $(a_1 + 1)$-th to a_2-th rows for any matrix J. For the rest of the paper, we define shorthand

$$S_{a_1, a_2} := \sum_{i=a_1+1}^{a_2} x_i x_i^\top = X_{(a_1, a_2]}^\top X_{(a_1, a_2]}$$

Define the scalar quantity $\eta(n, p) := (\mathbb{E}[x_1 x_1^\top (n^{-1}S_{0,n})^{-1} x_n x_n^\top])_{1,1}$.

Lemma 12. *For all $t \leq z$, $W_t^\top W_z = 4S_{0,t}S_{0,n}^{-1}S_{z,n}$, whence for $z \in [n]$ and $t \in [z]$ we have*

$$\mathbb{E}[W_t^\top W_z] = 4t(n-z)n\eta(n,p)I_p.$$

15

Furthermore, under Condition 2, $\frac{\eta(n,p)}{(n-p)n^{-1}} \to 1$, i.e., $\eta(n,p) \to \eta$ as $n,p \to \infty$.

Proof. By the construction of A, we have $AA^\top = I_n - X(X^\top X)^{-1}X^\top$. We have

$$
\begin{aligned}
W_t^\top W_z &= \begin{pmatrix} X_{(0,t]}^\top & -X_{(t,n]}^\top \end{pmatrix} \begin{pmatrix} A_{(0,t]} \\ A_{(t,n]} \end{pmatrix} \begin{pmatrix} A_{(0,z]}^\top & A_{(z,n]}^\top \end{pmatrix} \begin{pmatrix} X_{(0,z]} \\ -X_{(z,n]} \end{pmatrix} \\
&= \begin{pmatrix} X_{(0,t]}^\top & -X_{(t,n]}^\top \end{pmatrix} (I_n - X(X^\top X)^{-1}X^\top) \begin{pmatrix} X_{(0,z]} \\ -X_{(z,n]} \end{pmatrix} \\
&= (S_{0,t} - S_{t,z} + S_{z,n}) - (S_{0,t} - S_{t,z} - S_{z,n})S_{0,n}^{-1}(S_{0,t} + S_{t,z} - S_{z,n}) \\
&= 2(S_{0,t} - S_{t,z} - S_{z,n})S_{0,n}^{-1}S_{z,n} + 2S_{z,n} \\
&= 4S_{0,t}S_{0,n}^{-1}S_{z,n}.
\end{aligned}
$$

In particular, we have for all $z \in [n]$ and $t \in [z]$

$$
\mathbb{E}W_t^\top W_z = 4\sum_{i=1}^{t} \sum_{j=z+1}^{n} \mathbb{E}[x_i x_i^\top S_{0,n}^{-1} x_j x_j^\top] = 4t(n-z)\mathbb{E}[x_1 x_1^\top S_{0,n}^{-1} x_n x_n^\top],
$$

where we invoke the exchangeability of $x_i x_i^\top S_{0,n}^{-1} x_j x_j^\top$ for all $1 \le i < j \le n$ in the second equality.

We first note $\mathbb{E}[W_z^\top W_z] = 4z(n-z)\mathbb{E}[x_1 x_1^\top S_{0,n}^{-1} x_n x_n^\top]$, and then show that $\mathbb{E}[W_z^\top W_z]$ must be a scale of I_p. Since for any $U \in \mathbb{O}^{p \times p}$, $x_i^\top U \overset{\mathrm{d}}{=} x_i^\top$ for $i \in [n]$, whence we have $U^\top W_z^\top W_z U = 4U^\top S_{0,t} U (U^\top S_{0,n} U)^{-1} U^\top S_{z,n} U \overset{\mathrm{d}}{=} W_z^\top W_z$. In particular, $W_z^\top W_z$ have identically distributed diagonals and identically distributed off-diagonals. It suffices to verify that its off-diagonals have zero mean.

Let $X = QT$ be the almost surely unique QR decomposition of X, where we only take non-negative diagonal entries in T. By Equation (15) of Gao and Wang (2022), we have $W_z^\top W_z = 4T^\top V\Lambda(I_p - \Lambda)V^\top T$, where $V\Lambda V^\top = Q_{(0,z]}^\top Q_{(0,z]}$ is the eigendecomposition of $B := Q_{(0,t]}^\top Q_{(0,t]}$. Note that $V \sim$ Unif($\mathbb{O}^{p \times p}$), Λ and T are mutually independent and T has independent entries with $T_{j,j} = t_j > 0$ such that $t_j^2 \sim \chi_{n-j+1}^2$ and $T_{j,k} = z_{jk} \sim N(0,1)$ for $j \ne k$. For off-diagonals, it suffices to have

$$
\begin{aligned}
(\mathbb{E}[W_z^\top W_z])_{1,2} &= 4\mathbb{E}\left[\sum_{j=1}^{p} t_1 V_{j,1}\lambda_j(1-\lambda_j)(t_2 V_{j,2} + z_{12}V_{j,1}) \right] \\
&= 4\sum_{j=1}^{p} \left[\mathbb{E}[t_1]\mathbb{E}[\lambda_j(1-\lambda_j)](\mathbb{E}[t_2]\mathbb{E}[V_{j,1}V_{j,2}] + \mathbb{E}[z_{12}]\mathbb{E}[V_{j,1}^2]) \right] = 0,
\end{aligned}
$$

16

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are diagonal elements of Λ and $\mathbb{E}[V_{j,1}V_{j,2}] = (1/p)\sum_{j=1}^p V_{j,1}V_{j,2} = 0$ since $V \in \mathbb{O}^{p \times p}$.

Lastly, under Condition 2, by Proposition 8 of Gao and Wang (2022), $(W_z^\top W_z)_{1,1} \xrightarrow{\text{a.s.}} (4z(n-z)(n-p))/n^3$. Noting $(\mathbb{E}[W_z^\top W_z])_{1,1} = 4z(n-z)n\eta(n,p)$, we conclude the convergence of $\eta(n,p) \to \eta$. $\qquad\square$

Lemma 13. *Let $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N_p(0, I_p)$. Fix $z \in [n]$ and $t \in [z]$. For any nonrandom $u, w \in \mathcal{S}^{p-1}$, we have*

$$\mathbb{P}\left[\frac{1}{t}\left|u^\top\left\{S_{0,t} - \frac{t}{z}S_{0,z}\right\}S_{0,n}^{-1}S_{z,n}w\right| \geq x \,\Big|\, S_{0,z}, S_{z,n}\right] \leq 52\exp\left\{-\frac{t^2x^2}{8z\|S_{0,n}^{-1}S_{z,n}\|_{\text{op}}^2\|S_{0,z}/z\|_{\text{op}}^2}\right\}.$$

Proof. For notational simplicity, we use $\mathbb{P}^z$ and $\mathbb{E}^z$ to denote the conditional probability and expectation with respect to the σ-algebra generated by $(S_{0,z}, S_{z,n})$. Note that $S_{0,t} \mid (S_{0,z}, S_{z,n}) \overset{\text{d}}{=} S_{0,t} \mid S_{0,z}$ and $\mathbb{E}^z[S_{0,t}] = (t/z)S_{0,z}$.

Let $r := \min\{z, p\}$. We note $S_{0,z}$ has (almost surely) rank r, and write $S_{0,z} = R^\top R$ for the (almost surely unique) Cholesky decomposition of $S_{0,z}$ such that $R \in \mathbb{R}^{r \times p}$ is an upper-triangular matrix with positive diagonal entries. Write $R^\dagger \in \mathbb{R}^{p \times r}$ for the (almost surely unique) Moore–Penrose pseudo-inverse of R such that $RR^\dagger = I_p$. By Corollary 11, the matrix $B := (R^\dagger)^\top S_{0,t}R^\dagger \sim \text{Beta}_r(t/2, (z-t)/2)$ has a matrix-variate Beta distribution, and is independent of $S_{0,z}$ with $(t/z)I_p$ as its (conditional) mean. Observe that $R^\top BR = (R^\dagger R)^\top S_{0,t}(R^\dagger R) = S_{0,t}$, since $R^\dagger R$ is a (symmetric) orthogonal projection matrix onto the row space of $X_{(0,z]}$, which contains the row space of $X_{(0,t]}$.

Define $v := S_{0,n}^{-1}S_{z,n}w/\|S_{0,n}^{-1}S_{z,n}w\|_2$. Writing $\tilde{u} := Ru/\sqrt{z}$ and $\tilde{v} := Rv/\sqrt{z}$, we have

$$\mathbb{P}^z\left[\frac{1}{t}|u^\top\{S_{0,t} - \mathbb{E}^z(S_{0,t})\}S_{0,n}^{-1}S_{z,n}w| \geq x\right] \leq \mathbb{P}^z\left[\frac{1}{t}|u^\top\{S_{0,t} - \mathbb{E}^z(S_{0,t})\}v| \geq \frac{x}{\|S_{0,n}^{-1}S_{0,z}\|_{\text{op}}}\right]$$

$$= \mathbb{P}^z\left\{\left|\tilde{u}^\top\left(\frac{z}{t}B - I_p\right)\tilde{v}\right| \geq \frac{x}{\|S_{0,n}^{-1}S_{0,z}\|_{\text{op}}}\right\}$$

$$\leq \mathbb{P}^z\left\{\left|\left(\frac{\tilde{u}}{\|\tilde{u}\|_2}\right)^\top\left(\frac{z}{t}B - I_p\right)\frac{\tilde{v}}{\|\tilde{v}\|_2}\right| \geq \frac{x}{\|S_{0,n}^{-1}S_{0,z}\|_{\text{op}}\|S_{0,z}/z\|_{\text{op}}}\right\}.$$

Write shorthand $\psi := \|S_{0,n}^{-1}S_{0,z}\|_{\text{op}}\|S_{0,z}/z\|_{\text{op}}$, which is measurable with respect to the σ-algebra generated by $(S_{0,z}, S_{z,n})$. There exists an orthogonal matrix U such that $U\tilde{u}/\|\tilde{u}\|_2 = e_1$ and $U\tilde{v}/\|\tilde{v}\|_2 = \alpha e_1 + \beta e_2$ for real α and β such

17

that $\alpha^2 + \beta^2 = 1$, where e_j denotes the jth standard basis vector in $\mathbb{R}^p$. Using the fact that $B \mid (S_{0,z}, S_{z,n}) \stackrel{\mathrm{d}}{=} U^\top B U \mid (S_{0,z}, S_{z,n})$, we have for $J = \{1, 2\}$ that

$$
\mathbb{P}^z \left\{ \left| \left(\frac{\tilde{u}}{\|\tilde{u}\|_2} \right)^\top \left(\frac{z}{t} B - I_p \right) \frac{\tilde{v}}{\|\tilde{v}\|_2} \right| \geq \frac{x}{\psi} \right\}
$$

$$
\leq \mathbb{P}^z \left\{ \left| \frac{z}{t} B_{1,1} - 1 \right| \geq \frac{x}{\sqrt{2}\psi} \right\} + \mathbb{P}^z \left\{ \frac{z}{t} |B_{1,2}| \geq \frac{x}{\sqrt{2}\psi} \right\}
$$

$$
\leq 2\mathbb{P}^z \left\{ \left\| \left(B - \frac{t}{z} I_p \right)_{J,J} \right\|_{\mathrm{op}} \geq \frac{tx}{\sqrt{2}z\psi} \right\},
$$

where the first inequality holds by noting $|\alpha| + |\beta| \leq (2\alpha^2 + 2\beta^2)^{1/2} = \sqrt{2}$. Note that $\{w \in \mathcal{S}^{p-1} : \mathrm{supp}(w) \subseteq J\}$ is isomorphic to S^1, which contains a $(1/4)$-net $\mathcal{N}$ of cardinality $\lceil \frac{2\pi}{4\arcsin(1/8)} \rceil = 13$. By Gupta and Nagar (1999, Theorem 5.3.12), for each $w \in \mathcal{N}$, we have $w^\top B w \sim \mathrm{Beta}(t/2, (z-t)/2)$. Hence, by Vershynin (2012, Lemma 5.4) and a union bound, we have

$$
\mathbb{P}^z \left\{ \left\| \left(B - \frac{t}{z} I_p \right)_{J,J} \right\|_{\mathrm{op}} \geq \frac{tx}{\sqrt{2}z\psi} \right\} \leq \mathbb{P}^z \left\{ \sup_{w \in \mathcal{N}} \left| w^\top B w - t/z \right| \geq \frac{tx}{2\sqrt{2}z\psi} \right\}
$$

$$
\leq 13\mathbb{P}^z \left\{ \left| B_{1,1} - t/z \right| \geq \frac{tx}{2\sqrt{2}z\psi} \right\}
$$

$$
\leq 26 \exp \left\{ -\frac{t^2 x^2}{8z\psi^2} \right\},
$$

where we have used Marchal and Arbel (2017, Theorem 2.1) in the final inequality. $\qquad\square$

Lemma 14. *Let $X = (X_1, \ldots, X_p)^\top$ be uniformly distributed on the sphere $\mathcal{S}^{p-1}$. Then, for $\delta \geq e^{-p/16}$, we have*

$$
\mathbb{P} \left(\|X\|_\infty > \sqrt{\frac{4 \log(2/\delta)}{p}} \right) \leq p\delta.
$$

Proof. Let $Z_1, \ldots, Z_p$ be independent $N(0,1)$ random variables, then $X_1 \stackrel{\mathrm{d}}{=} Z_1/(Z_1^2 + \cdots + Z_p^2)^{1/2}$. By a standard Gaussian tail bound, we have

$$
\mathbb{P} \left\{ Z_1 > \sqrt{2\log(1/\delta)} \right\} \leq \delta.
$$

18

Moreover, $\sum_{j=1}^p Z_j^2 \sim \chi_p^2$. Since we have $\delta \geq e^{-p/16}$, by Laurent and Massart (2000, Lemma 1),

$$\mathbb{P}\left(\sum_{j=1}^p Z_j^2 < \frac{p}{2}\right) \leq \mathbb{P}\left(\sum_{j=1}^p Z_j^2 < p - 2\sqrt{p\log(1/\delta)}\right) \leq \delta.$$

The result follows by combining the above two bounds and applying a union bound. □

Recall $B_0(k) \subseteq \mathbb{R}^p$ is the set of k-sparse unit vectors.

Lemma 15. *For any $A \in \mathbb{R}^{p \times p}$, and any $\epsilon \in (0,1)$, there exists an ϵ-net $\mathcal{N}_\epsilon$ of $B_0(k)$ of cardinality at most $\{(1 + 2/\epsilon)ep/k\}^k$ such that*

$$\sup_{u \in B_0(k)} u^\top A v \leq (1 - \epsilon)^{-1} \max_{u \in \mathcal{N}_\epsilon} u^\top A v.$$

Proof. By Vershynin (2012, Lemma 5.2), for each subset $S \subseteq [p]$ of cardinality k, there exists an ϵ-net $\mathcal{N}_S$ of $\{v \in B_0(k) : \text{supp}(v) \subseteq S\}$ of cardinality at most $(1 + 2/\epsilon)^k$. Define $\mathcal{N}_\epsilon := \cup_{S \subseteq [p]:|S|=k}\mathcal{N}_S$, then $|\mathcal{N}_\epsilon| \leq \binom{p}{k}(1 + 2\epsilon)^k \leq \{(1 + 2\epsilon)ep/k\}^k$. For any fixed $x \in B_0(k)$, find $\tilde{x} \in \mathcal{N}_\epsilon$ such that $\|x - \tilde{x}\|_2 \leq \epsilon$ and $\|x - \tilde{x}\|_0 \leq k$. Thus,

$$\begin{aligned}
x^\top A v &= (x - \tilde{x})^\top A v + \tilde{x}^\top A v \\
&\leq \|x - \tilde{x}\|_2 \sup_{u \in B_0(k)} u^\top A v + \sup_{u \in \mathcal{N}_\epsilon} u^\top A v \\
&\leq \epsilon \sup_{u \in B_0(k)} u^\top A v + \sup_{u \in \mathcal{N}_\epsilon} u^\top A v.
\end{aligned}$$

The desired result follows by taking supremum over $x \in B_0(k)$ above. □

C   Additional simulation results

k	s	ρ	**CP**	**CL**	VPBS	LB	KJF	LSS	CO	XWZY
3	3	1	79.8	105.3	122.8	665.2	**4.5**	22.0	126.3	39.1
		2	8.8	14.1	33.0	493.1	**2.1**	6.2	17.3	10.4
		4	3.3	3.4	8.3	38.0	**1.2**	2.4	13.1	24.8
		8	1.6	1.6	2.5	0.4	1.2	**0.6**	15.6	37.3
3	1000	1	**60.7**	113.3	241.6	429.5	237.2	227.3	222.8	272.9
		2	**8.3**	11.8	243.4	441.4	239.0	228.2	92.3	298.8
		4	**2.9**	4.0	239.5	366.9	243.9	230.6	96.6	304.9
		8	2.4	**1.4**	235.1	245.1	262.2	230.7	53.9	312.7
31	31	1	**239.1**	333.9	145.1	659.5	**15.0**	33.9	410.5	88.3
		2	**89.6**	128.0	59.1	741.1	**3.0**	11.2	452.1	24.8
		4	25.7	**13.6**	39.0	765.9	**1.3**	4.6	389.1	1.0
		8	11.2	**5.9**	68.5	396.2	1.2	**0.8**	408.7	6.0
31	1000	1	300.3	364.9	233.4	440.1	238.8	**227.4**	352.0	280.0
		2	**71.7**	140.9	242.5	469.5	238.9	228.3	398.1	294.5
		4	16.0	**12.5**	251.3	358.4	238.9	224.5	418.7	308.6
		8	13.7	**4.6**	244.5	249.0	238.2	230.1	372.1	315.5
1000	1000	1	275.5	359.8	232.6	483.0	239.3	**231.8**	371.3	273.5
		2	256.9	320.8	238.4	447.4	238.9	**229.2**	348.2	296.1
		4	224.1	**91.0**	242.7	378.2	239.1	228.0	357.6	309.8
		8	194.5	**39.6**	246.4	253.5	242.4	226.7	393.9	311.4

Table S.1: Average loss of various changepoint methods under different settings, similar to Table 2, except for $p = 1000$. Other parameters: $n = 1200$ and $z = 360$. The method with the least average loss in each line is marked in bold.

References

Gao, F. and T. Wang (2022). Two-sample testing of high-dimensional linear regression coefficients via complementary sketching. *Ann. Statist. 50*(5), 2950 – 2972.

Gupta, A. K. and D. K. Nagar (1999). *Matrix Variate Distributions*, Volume 104. CRC Press.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist. 28*, 1302–1338.

Marchal, O. and J. Arbel (2017). On the sub-gaussianity of the beta and dirichlet distributions. *Electron. Commun. Probab. 22*, 1–14.

Mitra, S. K. (1970). A density-free approach to the matrix variate beta distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, 81–88.

Muirhead, R. J. (2009). *Aspects of Multivariate Statistical Theory.* Hoboken, New Jersey: John Wiley & Sons.

Samworth, R. J. and R. D. Shah (2025+). *Modern Statistical Methods and Theory.* Cambridge University Press, Cambridge.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. in y. eldar and g. kutyniok (eds.). *Compressed Sensing, Theory and Applications*, 210–268.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge: Cambridge University Press.

Wang, T. and R. J. Samworth (2018). High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B 80*(1), 57–83.