Electronic Journal of Statistics
ISSN: 1935-7524

Estimation of high-dimensional change-points under a group sparsity structure

Hanqing Cai and Tengyao Wang*

Department of Statistical Sciences
1–19 Torrington Place
London WC1E 7HB, United Kingdom
e-mail: hanqing.cai.15@ucl.ac.uk

Department of Statistics
London School of Economics
Columbia House, 69 Aldwych
London WC2B 4RR, United Kingdom
e-mail: t.wang59@lse.ac.uk

Abstract: Change-points are a routine feature of 'big data' observed in the form of high-dimensional data streams. In many such data streams, the component series possess group structures and it is natural to assume that changes only occur in a small number of all groups. We propose a new change point procedure, called `groupInspect`, that exploits the group sparsity structure to estimate a projection direction so as to aggregate information across the component series to successfully estimate the change-point in the mean structure of the series. We prove that the estimated projection direction is minimax optimal, up to logarithmic factors, when all group sizes are of comparable order. Moreover, our theory provide strong guarantees on the rate of convergence of the change-point location estimator. Numerical studies demonstrates the competitive performance of `groupInspect` in a wide range of settings and a real data example confirms the practical usefulness of our procedure.

AMS 2000 subject classifications: 62H12, 62M10.
Keywords and phrases: change-point analysis, high-dimensional data, group sparsity.

1. Introduction

Modern applications routinely generate time-ordered high-dimensional datasets, where many covariates are simultaneously measured over time. Examples include wearable technologies recording the health state of individuals from multi-sensor feedbacks (Hanlon and Anderson, 2009), internet traffic data collected by tens of thousands of routers (Peng, Leckie and Ramamohanarao, 2004) and functional Magnetic Resonance Imaging (fMRI) scans that record the time evolution of blood oxygen level dependent (BOLD) chemical contrast in different areas of the brain (Aston and Kirch, 2012). The explosion in number of such

*Research supported by EPSRC grant EP/T02772X/1.

high-dimensional data streams calls for methodological advances for their analysis.

Change-point analysis is an essential statistical technique used in identifying abrupt changes in a time series. Time points at which such abrupt change occurs are called 'change-points'. Through estimating the location of change-points, we can divide the time series into shorter segments that can be analysed using methods designed for stationary time series. Moreover, in many applications, the estimated change-points indicate specific events that are themselves of great interest. In the examples mentioned in the previous paragraph, they can be used to raise alarms about abnormal health events, detect distributed denial of service attacks on the network and pinpoint the onset of certain brain activities.

Classical change-point analysis focuses on univariate time series. The current state-of-art methods including Killick, Fearnhead and Eckley (2012); Frick, Munk and Sieling (2014); Fryzlewicz (2014). However, classical univariate change-point methods are often inadequate for high-dimensional datasets that are routinely encountered in modern applications. When applied componentwise, they are often sub-optimal as signals can spread over many components. As a result, several new methodologies have been proposed to test and estimate change-points in the high-dimensional settings. These include methods that apply a simple ℓ_2 or ℓ_∞ aggregation of test statistics across different components (Horváth and Hušková, 2012; Jirak, 2015), and more complex methods such as a scan-statistics based approach by Enikeeva and Harchaoui (2019), the Sparsified Binary Segmentation algorithm by Cho and Fryzlewicz (2015), the double CUSUM algorithm of Cho (2016) and a projection-based approach by Wang and Samworth (2018).

To overcome the curse of dimensionality, existing high-dimensional change-point methods often assume that the signal of change possesses some form of sparsity. For example, in the high-dimensional mean change setting studied in Jirak (2015); Cho and Fryzlewicz (2015); Wang and Samworth (2018); Enikeeva and Harchaoui (2019), it is assumed that the difference in mean before and after a change-point is nonzero only in a small subset of coordinates. While the sparsity assumption greatly reduces the complexity of the original high-dimensional problem, it often does not capture the the full extent of the structure in the vector of change available in real data applications. For instance, in many applications, the coordinates of the high-dimensional vectors are naturally clustered into groups and coordinates within the same group tend to change together. At each change-point, only a small number of groups will undergo a change. Such a group sparsity change-point structure is useful in modelling many practical applications. Examples include financial data stream where changes are often grouped by industry sectors and a small number of sectors may experience virtually simultaneous market shocks. Also, in functional magnetic resonance imaging data, voxels belonging to the same brain functional regions tend to change simultaneously over time. Similar group sparsity assumptions have been made in other statistical problems including Yuan and Lin (2006); Wang and Leng (2008); Simon et al. (2020).

In this work, we provide a new high-dimensional change-point methodol-

ogy that exploits the group sparsity structure of the changes. More precisely, given pre-specified grouping information of all the coordinates, our algorithm, named `groupInspect` (standing for **group**-based **in**formative **s**parse **p**rojection **e**stimator of **c**hange-poin**t**s), will first estimate a vector of projection that is closely aligned with the true vector of change at each change-point. It will then project the high-dimensional data series along this estimated direction and apply a univariate change-point method on the projected series to identify the location of the change. The above procedure can be combined with the narrowest-over-threshold algorithm of Baranowski et al. (2019) to recursively identify multiple change-points. We show that, in a single change-point setting, the projection direction estimator employed in `groupInspect` has a minimax optimal dependence, up to logarithmic factors, on both the ℓ_0 sparsity parameter and the group-sparsity parameter, representing respectively the number of nonzero elements and the number of nonzero groups in the vector of change. Furthermore, under appropriate conditions, `groupInspect` achieves a minimax optimal $\log\log(n)/(n\vartheta^2)$ rate of convergence for the estimated location of a single change-point and a $\log(n)/(n\vartheta^2)$ rate of convergence for multiple change-points, where ϑ denotes the ℓ_2 norm of the vector of change.

The outline of the paper is as follows. In Section 2, we describe the formal setup of our problem. The `groupInspect` methodology is then introduced in Section 3, with its theoretical performance guarantees provided in Section 4. We illustrate the empirical performance of `groupInspect` via simulatinos and a real-data example in Section 5. Proofs of all theoretical results are deferred to Section 6, and ancillary results and their proofs are given in Appendix A.

1.1. Notation

For $n \in \mathbb{N}$, we write $[n] = \{1, \ldots, n\}$. For a vector $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$, we define $\|v\|_0 = \sum_{i=1}^n \mathbb{1}_{\{v_i \neq 0\}}$, $\|v\|_\infty = \max_{i\in[n]} |v_i|$ and $\|v\|_q = \big\{\sum_{i=1}^n (v_i)^q\big\}^{1/q}$ for any positive integer q, and let $\mathbb{S}^{n-1} = \{v \in \mathbb{R}^n : \|v\|_2 = 1\}$. For a matrix $A \in \mathbb{R}^{p\times n}$, we write $\|A\|_*$ for its nuclear norm and write $\|A\|_{\mathrm{F}}$ for its Frobenius norm.

For any $S \subseteq [n]$, we write v_S for the $|S|$-dimensional vector obtained by extracting coordinates of v in S. For a matrix $A \in \mathbb{R}^{p\times n}$, $J \in [p]$ and $S \in [n]$, we write $A_{J,S}$ for the submatrix obtained by extracting rows and columns of A indexed by J and S respectively. When $S = [n]$, we abbreviate $A_{J,[n]}$ by A_J. When $S = \{t\}$ is a single element set, we slightly abuse notation and write $A_{J,t}$ instead of $A_{J,\{t\}}$.

Given two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ such that $a_n, b_n > 0$ for all n, we write $a_n \lesssim b_n$ (or equivalently $b_n \gtrsim a_n$) if $a_n \leq Cb_n$ for some universal constant C.

2. Problem description

Let $X_1, \ldots, X_n$ be independent random vectors with distribution:

$$X_t \sim N_p(\mu_t, \Sigma), \quad 1 \leq t \leq n, \quad \text{where } \|\Sigma\|_{\text{op}} \leq B \tag{1}$$

for some $B \in (0, \infty)$. We remark that the main focus of the current work is to understand the effect of group sparasity structure on the change-point estimation accuracy, and as such, to simplify exposition, we have assumed here that observations are independent normal random vectors. All our theoretical results can be extended to the case where the observations are sub-Gaussian and have short-ranged temporal dependence (see Appendices B and C for details). We can combine into a single data matrix $X \in \mathbb{R}^{p \times n}$. We assume that the sequence of mean vectors $(\mu_t)_{t=1}^n$ undergoes changes at times $z_i \in \{1, \ldots, n-1\}$ for $i \in \{1, \ldots, \nu\}$, in the sense that

$$\mu_{z_i+1} = \cdots = \mu_{z_{i+1}} =: \mu^{(i)}, \quad \forall\, i \in \{0, \ldots, \nu\}, \tag{2}$$

where we use the convention that $z_0 = 0$ and $z_{\nu+1} = n$. We assume that consecutive change-points are sufficiently separated in the sense that

$$\min\{z_{i+1} - z_i : 0 \leq i \leq \nu\} \geq n\tau.$$

Suppose further that each of the p coordinates belong to (at least) one of G groups. Specifically, let $\mathcal{J}_g$ denotes the set of indices associated with the gth group for $g \in \{1, \ldots, G\}$, we have that

$$\bigcup_{g=1}^{G} \mathcal{J}_g = [p]. \tag{3}$$

We assume that coordinates in the same group will tend to change together. We will consider both the case of overlapping and non-overlapping groups. In the latter scenario, each coordinate belongs to a unique group and $(\mathcal{J}_g)_{g \in [G]}$ forms a partition of $[p]$.

Our goal is to estimate the locations of change $z_1, \ldots, z_\nu$ from the data matrix X and the pre-specified grouping information $(\mathcal{J}_g)_{g \in [G]}$. Motivated by Wang and Samworth (2018), when the coordinates are independent, the best way to aggregate the component series so as to maximise the signal-to-noise ratio around the ith change-point is to project the data along a direction close to the vector of change $\theta^{(i)} = \mu^{(i)} - \mu^{(i-1)}$. Let $v^{(i)}$ be the unit vector parallel to $\theta^{(i)}$:

$$v^{(i)} = \theta^{(i)} / \|\theta^{(i)}\|_2.$$

In our setting, we would like to find the optimiser of $\max \frac{v^\top \theta}{(v^\top \Sigma v)^{1/2}}$, which is $\Sigma^{-1}\theta$. We measure the quality of any estimated projection direction $\hat{v}$ with the Davis–Kahan sin θ loss (Davis and Kahan, 1970)

$$L(\hat{v}, v^{(i)}) = \sqrt{1 - (\hat{v}^\top v^{(i)})^2}$$

and measure the quality of the subsequent location estimator $\hat{z}_i$ by $\mathbb{E}|\hat{z}_i - z_i|$.

The difficulty of the estimation task depends on both the noise level σ and the vector of change $\theta^{(i)} = \mu^{(i)} - \mu^{(i-1)}$. More precisely, we assume that the change is localised in a small number of the G groups as defined in (3). Define $\phi : \mathbb{R}^p \to \mathbb{R}^G$ such that $\phi(x) = (\|x_{\mathcal{J}_1}\|_2, \|x_{\mathcal{J}_2}\|_2, \ldots, \|x_{\mathcal{J}_G}\|_2)^\top$, we assume that

$$\|\phi(\theta^{(i)})\|_0 \leq s, \qquad \sum_{g \in [G]: \theta^{(i)}_{\mathcal{J}_g} \neq 0} |\mathcal{J}_g| \leq k, \quad \text{and} \quad \|\theta^{(i)}\|_2 \geq \vartheta. \tag{4}$$

3. Methodology

3.1. Single change-point estimation

Initially, we will consider estimation of a single change-point, where $\nu = 1$. This can be extended to estimate multiple change-points in conjunction with top-down approaches such as wild binary segmentation and narrowest-over-threshold approach of Baranowski et al. (2019), which we will discuss in Section 3.2.

We define the CUSUM transformation $\mathcal{T} : \mathbb{R}^{p \times n} \to \mathbb{R}^{p \times (n-1)}$ by

$$\mathcal{T}(M)_{j,t} = \sqrt{\frac{t(n-t)}{n}} \left(\frac{1}{n-t} \sum_{r=t+1}^{n} M_{j,r} - \sum_{r=1}^{t} \frac{1}{t} M_{j,r} \right), \tag{5}$$

and compute the CUSUM matrix $T = \mathcal{T}(X)$. As discussed in Section 2, our general strategy is to use the matrix T to estimate a projection direction that is well-aligned with the direction of change, and then project the data along this direction to estimate the change-point location from the univariated projected series. More precisely, we would like to solve for

$$\hat{v} \in \operatorname*{arg\,max}_{u \in \mathbb{S}^{p-1}, \|\phi(u)\|_0 \leq s} \|u^\top T\|_2, \tag{6}$$

where, $\mathbb{S}^{p-1} = \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$. However, the above optimisation problem is non-convex due to the group-sparsity constraint. Consequently, we perform the following convex relaxation of the above problem. We first note that the set of optimisers of (6) is equal to the set of leading left singular vectors of

$$\operatorname*{arg\,max}_{\substack{M \in \mathbb{R}^{p \times (n-1)}: \|M\|_* = 1, \operatorname{rank}(M) = 1 \\ \sum_{g \in [G]} \mathbb{1}_{\{\|M_{\mathcal{J}_g}\|_F \neq 0\}} \leq s}} \langle M, T \rangle,$$

We relax the above matrix-variate optimisation problem by dropping the combinatorial rank constraint, and replacing the nuclear norm constraint set by the larger Frobenius norm set of $\mathcal{S} = \{M \in \mathbb{R}^{p \times (n-1)} : \|M\|_F \leq 1\}$. The constraint that M has at most s groups of non-zero rows can be written as an ℓ_0 constraint on the vector of Frobenius norms of such submatrices, i.e. $\|(\|M_{\mathcal{J}_g}\|_F : g \in \{1, \ldots, G\})\|_0 \leq s$. Motivated by the group lasso penalty (Yuan

and Lin, 2006), we replace this group sparsity constraint with a *group norm* penalty, where the group norm for a matrix $M \in \mathbb{R}^{p\times(n-1)}$ is defined as

$$\|M\|_{\text{grp}} = \sum_{g=1}^{G} p_g^{1/2} \|M_{\mathcal{J}_g}\|_{2,1}, \tag{7}$$

where $\|M_{\mathcal{J}_g}\|_{2,1}$ is the sum of column ℓ_2 norms of the submatrix $M_{\mathcal{J}_g}$ and $p_g = |\mathcal{J}_g|$. Overall, we obtain the following optimisation problem:

$$\hat{M} \in \arg\max_{M\in\mathcal{S}} \big\{\langle T, M\rangle - \lambda\|M\|_{\text{grp}}\big\}, \tag{8}$$

where $\lambda \in [0,\infty)$ is a regularization parameter.

If the groups are non-overlapping, in the sense that $\mathcal{J}_g \cap \mathcal{J}_{g'} = \emptyset$ for all $g \neq g'$, then we see from Proposition 8 that (8) has a closed form solution

$$\hat{M} = \frac{T - R^*}{\|T - R^*\|_{\text{F}}}, \tag{9}$$

where $R^*_{\mathcal{J}_g,t} = T_{\mathcal{J}_g,t} \min\big\{\frac{\lambda p_g^{1/2}}{\|T_{\mathcal{J}_g,t}\|_2}, 1\big\}$.

For overlapping groups, (8) can be optimised using Frank–Wolfe algorithm (Frank and Wolfe, 1956), as described in Algorithm 1. We first compute the gradient of the objective function which is the step 4 in Algorithm 1. We then project the $\hat{M}$ back onto $\mathcal{S}$.

After solving the optimization problem, we can obtain the estimated projection direction $\hat{v}$ by computing the leading left singular vector of $\hat{M}$. Then, we project the data along $\hat{v}$ to obtain a univariate series for which existing one-dimensional change-point estimation methods apply. Specifically, we perform the CUSUM transformation over the projected data series, and locate the change-point by the maximum absolute value of the CUSUM vector. The full procedure is described in Algorithm 2.

3.2. Multiple change-point estimation

When the data matrix possess multiple change-points, we may combine Algorithm 2 with a top-down approach (Fryzlewicz, 2014; Baranowski et al., 2019, e.g), to recursively identify all the change-points. Specifically, in Algorithm 3, we adopt the narrowest-over-threshold approach of Baranowski et al. (2019). We start by drawing a large number of random intervals $[s_1, e_1], \ldots, [s_Q, e_Q]$ and perform a test in each of these intervals to find windows that contain at least one change-point (Line 5 of Algorithm 3, with justification given by Corollary 4 in Section 4). We then select the narrowest interval for which the test rejects the null and apply Algorithm 2 to estimate a change-point within that window. We then partition the data into two submatrices to the left and right of this identified change-point and repeat the above procedures until no windows within the segmented submatrices contain any change-point.

Algorithm 1: Frank–Wolfe algorithm for optimising (8)

Input: $T \in \mathbb{R}^{p\times(n-1)}$, grouping $(\mathcal{J}_g)_{g\in[G]}$, $\lambda > 0$ and $\epsilon > 0$.

Initialise $\hat{M}^{[0]} = T/\|T\|_{\mathrm{F}}$ and $i = 0$.

repeat

$i \leftarrow i + 1$

Compute $G^{[i]} = (G_1^{[i]}, \ldots, G_p^{[i]})^\top \in \mathbb{R}^{p\times(n-1)}$ such that

$$G_{j,t}^{[i]} \leftarrow T_{j,t} - \sum_{g: j\in\mathcal{J}_g} \lambda_g \frac{M_{j,t}^{[i-1]}}{\|M_{\mathcal{J}_{g,t}}^{[i-1]}\|_{\mathrm{F}}},$$

where $\lambda_g = p_g^{1/2}\lambda$

if $G^{[i]} = 0$ **then break**

Compute

$$\tilde{M}^{[i]} = \frac{i}{i+2} M^{[i-1]} + \frac{2}{i+2} \frac{G^{[i]}}{\|G^{[i]}\|_{\mathrm{F}}},$$

Normalise $\hat{M}^{[i]} \leftarrow \tilde{M}^{[i]}/\|\tilde{M}^{[i]}\|_{\mathrm{F}}$

until $\|\hat{M}^{[i+1]} - \hat{M}^{[i]}\|_{\mathrm{F}} \le \epsilon$;

Output: $\hat{M}^{[i]}$

Algorithm 2: Single change-point estimation procedure for data with group structure

Input: $X \in \mathbb{R}^{p\times n}$, $(\mathcal{J}_g)_{g\in[G]}$, and $\lambda > 0$

Compute $T \leftarrow \mathcal{T}(X)$ as in (5).

Solve

$$\hat{M} \in \operatorname*{arg\,max}_{M\in\mathcal{S}} \big\{ \langle T, M\rangle - \lambda\|M\|_{\mathrm{grp}} \big\}$$

using either the closed-form solution in (9) if groups are non-overlapping, or Algorithm 1.

Let $\hat{v}$ be the leading left singular vector of $\hat{M}$.

Estimate z by $\hat{z} = \arg\max_{1\le t\le n-1} |\hat{v}^\top T_t|$, where T_t is the tth column of T.

Output: $\hat{z}$, $\bar{T}_{\max} = \hat{v}^\top T_z$

Algorithm 3: Multiple change-point estimation procedure

Input: $X \in \mathbb{R}^{p\times n}$, $(\mathcal{J}_g)_{g\in[G]}$, $\lambda > 0$, β, $M \in \mathbb{N}$

Set $\hat{\mathcal{Z}} \leftarrow \emptyset$

Draw M pairs of integers $(s_1, e_1), \ldots, (s_M, e_M)$ uniformly at random from the set $\{(\ell, r) \in \mathbb{Z}^2 : 0 \le \ell < r \le n\}$

Function NOT(s, e)

  Set $\mathcal{M}_{s,e} = \{m \in [M] : s \le s_m < e_m \le e\}$

  Set $\mathcal{R}_{s,e} := \{m \in \mathcal{M}_{s,e} : \|\mathcal{T}(X^{(s_m+\beta, e_m-\beta]})\|_{\mathrm{grp}*} > \lambda\}$, where $X^{(a,b]}$ is the submatrix of X obtained using columns indexed in $(a, b]$

  if $\mathcal{R}_{s,e} \neq \emptyset$ **then**

    Find $m^* \in \arg\min_{m\in\mathcal{R}_{s,e}} |e_m - s_m|$

    Set $\hat{z}^{[m^*]}$ as the output from Algorithm 2 with inputs $X^{(s_{m^*}, e_{m^*}]}$ and λ

    $b \leftarrow \hat{z}^{[m^*]} + s_{m^*}$

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

    Run recursively **NOT**(s, b) and **NOT**(b, e)

Output: $\hat{\mathcal{Z}}$

4. Theoretical guarantees

In this section, we provide theoretical guarantees to the performance of the groupInspect algorithm. As we have noted in Section 2, a key to the successful change-point estimation in the current problem is a good estimator of the oracle projection direction $v = \theta/\|\theta\|_2$.

The following theorem controls the sine angle risk of the estimated projection direction $\hat{v}$ in Step 3 of Algorithm 2 when data has a single change. We define $\mathcal{P}^{(\nu)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$ to be the set of data distributions satisfying (1), (2), (3) and (4). For any $P \in \mathcal{P}$, we write $v(P) = \theta/\|\theta\|_2$ where θ is the difference between post-change and pre-change means.

Theorem 1. *For a given grouping $(\mathcal{J}_g)_{g\in[G]}$, let $p_* = \min_{g\in[G]} |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \le C_1$. Let $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$ be a $p \times n$ data matrix, let θ be the vector of change and let $\hat{v}$ be as in Step 3 of Algorithm 2 with input X, $(\mathcal{J}_g)_{g\in[G]}$ and $\lambda \ge B^{1/2}(1+\sqrt{8\log(nG)/p_*})$. Then there exists $C > 0$, depending only on C_1, such that*

$$\sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{P}_P\left\{\sin\angle(\hat{v}, v) > \frac{C\lambda k^{1/2}}{n^{1/2}\tau\vartheta}\right\} \le \frac{1}{(nG)^3}. \tag{10}$$

We remark that the condition $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \le C_1$ is to control the extent of overlapping between different groups. Specifically, it requires that each coordinate can belong to at most C_1 groups. In the special case when all groups $\mathcal{J}_g$ are disjoint, which is often true in practical applications, then it suffices to take $C_1 = 1$.

We note that, when $\lambda = B^{1/2}(1 + \sqrt{8\log(nG)/p_*})$, with high probability, the sine angle loss in (10) has an upper bound that is proportional to

$Bk^{1/2}n^{-1/2}\tau^{-1}\vartheta^{-1}$, similar to what has been previously observed in Wang and Samworth (2018, Proposition 1). However, Theorem 1 reveals an interesting interaction between the ℓ_0 sparsity k and the group sparsity s when all groups are of comparable size. Specifically, for $\lambda = B^{1/2}(1+\sqrt{8\log(nG)/p_*})$ and assuming that $\max_{g\in[G]} p_g \lesssim p_*$, then we can simplify (10) to obtain that

$$\mathbb{E}\{\sin\angle(\hat{v}, v)\} \lesssim \sqrt{\frac{B\{k + s\log(nG)\}}{n\tau^2\vartheta^2}}.$$

In other words, the risk upper bound undergoes a phase transition as the number of coordinates per group increases above a $\log(nG)$ level. Similar phase transitions have been previously observed in the context of high-dimensional linear model where the regression coefficients satisfy a group sparsity assumption (see, e.g. Cai et al., 2019, Theorem 3).

We now turn our attention to a minimax lower bound of the estimation risk of the oracle projection direction. Theorem 2 below shows that the phase transition observed in Theorem 1 is not due to the specific proof techniques employed but rather an intrinsic feature of the problem.

Theorem 2. *Suppose $s > 0$, $k > 0$ and a grouping $(\mathcal{J}_g)_{g\in[G]}$ satisfy that $\mathcal{J}_g \cap \mathcal{J}_{g'} = \emptyset$ for all $g \neq g'$, $\min\{k, (s-1)\log(G/s)\} \geq 20$, and $\sum_{r=1}^{s} p_{(G-r+1)} \geq k/2$, where $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(G)}$ are order statistics of $p_1, \ldots, p_G$. Let $\Sigma = BI_p$. Then for some universal constant $c > 0$, we have*

$$\inf_{\tilde{v}} \sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) \geq c\sqrt{\frac{B\{k + s\log(G/s)\}}{n\tau\vartheta^2}},$$

where the infimum is taken over the set of all measurable functions $\tilde{v}$ of the data X.

The condition that $\sum_{r=1}^{s} p_{(G-r+1)} \geq k/2$ is to ensure that the upper bound k on the ℓ_0-sparsity is not too loose in the sense that k is not too much larger than the cardinality of the union of the largest s groups. If we assume that $\log(G/s) \asymp \log(n)$, $\tau \asymp 1$ and $\max_{g\in[G]} p_g \lesssim p_*$, then the lower bound in Theorem 2 matches the upper bound of Theorem 1 up to universal constants, when all groups are non-overlapping. We remark that the upper and lower bounds in Theorems 1 and 2 do not match in their dependence on the parameter τ. As Proposition 11 shows, this suboptimality is unlikely due to the convex relaxation carried out in (8) since the same τ dependence appears in the risk upper bound of the (computationally infeasible) optimiser of (6).

After obtaining guarantees on the quality of the projection direction estimator, we now provide theoretical guarantees of the overall change-point procedure. We note that the projection direction estimator $\hat{v}$ is dependent on the CUSUM panel T. While this dependence is observed to be very weak in practice, it creates difficulties in analysing the projected CUSUM series $\hat{v}^\top T$ in Step 4 of Algorithm 2. As such, for theoretical convenience, we will instead analyse a sample-splitting version of the algorithm.. Specifically, we split the data into

$X^{(1)}$ and $X^{(2)}$, consisting of odd and even time points respectively, as described in Algorithm 4. We use $X^{(1)}$ to estimate the projected direction $\hat{v}^{(1)}$ and then project $X^{(2)}$ along this direction to locate the change-point. Theorem 3 below provides a performance guarantee for the estimated location of the change-point of this sample-splitting version of our procedure.

Algorithm 4: Change-point estimation procedure: sample splitting version

Input: $X \in \mathbb{R}^{p\times n}$ and $\lambda > 0$

1 Define $X^{(1)}$ as $X^{(1)}_{j,t} = X_{j,2t-1}$ and $X^{(2)}$ as $X^{(2)}_{j,t} = X_{j,2t}$.

2 Compute $T^{(1)} \leftarrow \mathcal{T}(X^{(1)})$ and $T^{(2)} \leftarrow \mathcal{T}(X^{(2)})$ as in (5).

3 Solve

$$\hat{M}^{(1)} \in \operatorname*{arg\,max}_{M\in\mathcal{S}} \big\{ \langle T^{(1)}, M\rangle - \lambda \|M\|_{\mathrm{grp}} \big\}$$

using either the closed-form solution in (9) if groups are non-overlapping, or Algorithm 1.

4 Let $\hat{v}$ be the leading left singular vector of $\hat{M}^{(1)}$.

5 Estimate z by $\hat{z} = 2 \arg\max_{1\le t\le n_1-1} |(\hat{v}^{(1)})^\top T^{(2)}_t|$, where $T^{(2)}_t$ is the tth column of $T^{(2)}$.

Output: $\hat{z}$

Theorem 3. *Given data matrix $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})$, let $\hat{z}$ be the output from the Algorithm 4 with input X and $\lambda \ge B^{1/2}(1+\sqrt{p_*^{-1}8\log(nG)})$. There exist universal constants C, $C' > 0$ such that, if $n \ge 12$ is even, z is even, and*

$$\frac{C\sqrt{k}\lambda}{\vartheta\tau\sqrt{n}} \le 1, \tag{11}$$

then for any $\lambda_1 > \sqrt{B}$, we have

$$\mathbb{P}\left\{ \frac{1}{n}|\hat{z}-z| \le \frac{C'\lambda_1^2}{n\vartheta^2} \right\} \ge 1 - \frac{8}{n^3} - (3\lambda_1+1)e^{-\lambda_1^2/(4B)}\log n.$$

If we choose $\lambda_1 = C\sqrt{B\log\log n}$ for a sufficiently large absolute constant $C > 0$, then Theorem 3 shows that the location estimator $\hat{z}/n$ converges to z/n at a rate of $\frac{B\log\log n}{n\vartheta^2}$ in probability. This rate is minimax optimal even for the problem of estimating a single change in mean in a univariate series; see Proposition 6. While Theorem 3 concerns primarily with the estimation task, we remark that the argument used in its proof can be easily adapted to derive a testing procedure with good theoretical guarantees. Specifically, given data matrix $X \sim P \in \mathcal{P}_{n,p}(s,k,\tau,\vartheta,\Sigma,(\mathcal{J}_g)_{g\in[G]})$, we are interested to test the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta \ne 0$. We can use the We construct a test based on the dual norm to the $\|\cdot\|_{\mathrm{grp}}$ norm defined in (7). More precisely, for any $R \in \mathbb{R}^{p\times n}$ and a grouping $(\mathcal{J}_g)_{g\in[G]}$ of $[p]$, we define

$$\|R\|_{\mathrm{grp}*} = \max_{g\in[G]}\max_{t\in[n]} p_g^{-1/2}\|R_{\mathcal{J}_g,t}\|_2. \tag{12}$$

It can be seen from Lemma 7 that $\|\cdot\|_{\mathrm{grp}*}$ is indeed dual to $\|\cdot\|_{\mathrm{grp}}$. For any $\lambda > 0$, we define a test ψ_λ such that

$$\psi_\lambda(X) = \mathbb{1}_{\{\|\mathcal{T}(X)\|_{\mathrm{grp}^*} \geq \lambda\}}.$$

The following Corollary shows that with an appropriately chosen testing threshold λ, the test ψ_λ define above has good size and power controls.

Corollary 4. *Given data matrix $X \sim P \in \mathcal{P}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$. Let k be the total number of coordinates with change and $\|\theta\|_2$ be the magnitude of the change. Fix $\lambda \geq B^{1/2}\big(1 + \sqrt{4p_*^{-1}\log(nG)}\big)$.*

- *If $s = 0$, then $\mathbb{P}_P(\psi_\lambda(X) = 1) \leq 1/(nG)$.*
- *If $\vartheta \geq \frac{\sqrt{8k}\lambda}{\sqrt{n\tau}}$, then $\mathbb{P}_P(\psi = 1) \geq 1 - 1/(nG)$.*

Our single change-point theory can be applied iteratively to show that the `groupInspect` algorithm in in Algorithm 3 can consistently estimate both the number and the locations of the true change-points. In line with Theorem 3, we consider a sample-splitting version of Algorithm 3, which we call Algorithm Algorithm 3′, where we use Algorithm 4 in place of Algorithm 2 in line 6 of Algorithm 3.

Theorem 5. *Given data matrix $X \sim P \in \mathcal{P}^{(\nu)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$. Let $\hat{Z}$ be the output from the Algorithm 3′ with input X and $\lambda = B^{1/2}(1+\sqrt{8p_*^{-1}\log(nG)})$, Q and $\beta = n\tau/10$. Let $\tau\sqrt{n} \geq C'B\log n/\vartheta^2$. There exist universal constants C, $C' > 0$ such that, if $n \geq 12$ is even, z is even, and*

$$\frac{C\sqrt{Bk}}{\vartheta\tau\sqrt{n\tau}}\left(1 + \sqrt{\frac{8\log(nG)}{p_*}}\right) \leq 1, \tag{13}$$

then,

$$\mathbb{P}\left(\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \leq \frac{C'B\log n}{\vartheta^2}\ \forall\, i \in [\nu]\right) \geq 1 - \nu e^{-\tau^2 M/36} - \frac{1}{nG^3} - \frac{7}{n\tau^3}.$$

5. Numerical studies

In this section, we provide some simulation results to demonstrate the empirical performance of the groupInspect method. In all our numerical studies, unless otherwise specified, we will assume that data are generated according to (1), (2), (3) and (4). In all simulations, we do not assume that the covariance matrix Σ is known. Instead, we estimate the variance in each row using the mean absolute deviation of successive differences of the observations. We then standardise the data by the estimated row standard deviation. The `groupInspect` procedure is then applied to the standardised data assuming that Σ is a well-conditioned matrix with all diagonal entries equal to 1.

5.1. Theory validation

We first show that the practical performance of the `groupInspect` procedure is well captured by the theoretical results in Theorems 1 and 2. There are two related measures of the signal sparsity in our problem, which are the total number of coordinates of change k and the total number of groups with a change s. We conduct two sets of simulation experiments fixing one of these sparsity measures and varying the other. Specifically, for $n = 1000$, $p \in \{600, 1200, 2400\}$ and $\vartheta \in \{1, 2, 4, 8, 16\}$ and $\Sigma = I_p$, we split the p coordinates into disjoint groups of p_* coordinates per group, where p_* is allowed to vary over all divisors of 60. In the first set of experiments, we fix $k = 60$ so that $s = k/p_*$ varies with p_*, whereas in the second set of experiments, we fix $s = 3$ so that $k = sp_*$ varies with p_*. The vector of change is constructed so that the magnitude of change is equal across all coordinates of change. We will use the theoretical choice of tuning parameter λ for both sets of experiments here. Figure 1 shows how the $\sin\theta$ loss, averaged over 100 Monte Carlo repetitions, varies with p_*, for different choices of p and ϑ in both settings.

In the left panel of Figure 1, where the number of signal coordinates k is fixed, we see that the average loss decreases as p_* increases. Furthermore, at a log-log scale, and for relatively large signal sizes of $\vartheta \in \{4, 8, 16\}$, we see the loss curves follow an initial linear decreasing trend as p_* increases before plateauing eventually. This is in agreement with the two terms contributing to the loss described in Theorem 1. Specifically, for small p_*, we expect the second term of (10) to dominate and the loss decreases at a rate approximately proportional to $1/\sqrt{p_*}$ initially. For large p_*, we expect the first term of (10) to dominate and the loss will have minimal dependence on p_*. In the right panel of Figure 1, where the number of signal groups s is fixed, the average loss increases with p_*, as expected from our theory. It appears that for $s = 3$ studied here, the first term of (10) is dominant and the average loss increases linearly at the log-log scale with respect to p_*.

We further remark that in both panels of Figure 1, the average loss for large p_* shows equally spaced separation for the signal size ϑ in the dyadic grid $\{1, 2, 4, 8, 16\}$. This is in good agreement with the $1/\theta$ dependence of expected loss given in Theorem 1. Finally, we note that the ambient dimension p has minimal effect on the loss curves, for all signal strengths studied here. Again, this is predicted by our theory as the dimension p enters the mean loss in (10) only through the $\log(nG) = \log(pn/p_*)$ expression in the second term.

5.2. Practical choice of tuning parameter

The theoretical choice of λ turns out to be conservative in practical use. In this subsection, we will perform numerical simulations to suggest a suitable practical tuning parameter choice. We fix $n = 1000$, $z = 400$, $s = 3$, $G \in \{10, 25\}$ and assume $\Sigma = I_p$. The signal size ϑ is varied in $\{1, 2, 4, 8, 16\}$ and p is chosen from $\{500, 1000\}$. All groups are set to have equal size. We run the `groupInspect`

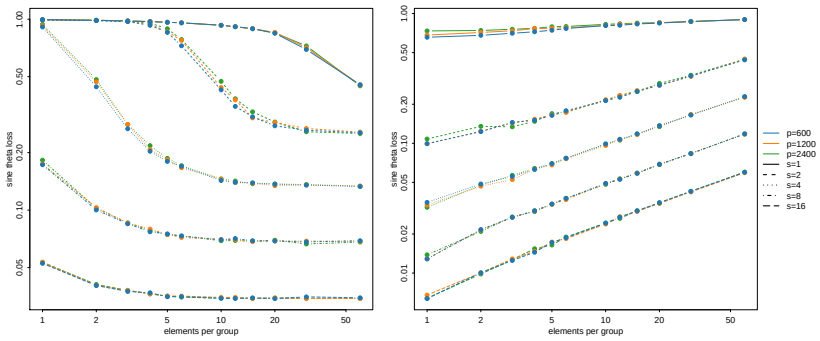

FIG 1. *Average loss (over 100 repetitions) of* `groupInspect` *for varying elements per group* p_**, plotted on a log-log scale. Left panel:* $k = 60$ *and* $s = k/p_*$*. Right panel:* $s = 3$ *and* $k = sp_*$*. Other parameter:* $n = 1000$.

algorithm for tuning parameters $\lambda = a(1 + \sqrt{4p_*^{-1}\log(nG)})$, where a is chosen from a logarithmic sequence of values between 0.1 and 3.

We plot $\sin\theta$ loss against a in Figure 2. In most cases, the loss is minimized when $a \approx 1/2$, i.e. tuning parameter value is half of the theoretical value. This suggests that when $\Sigma = I_p$, the choice $\lambda = 2^{-1}(1+\sqrt{4p_*^{-1}\log(nG)})$ leads to more accurate estimation in practice. Theorems 1 and 3 suggests that for non-identity covariance structure, the tuning parameter choice should scale proportional to the square root of the operator norm of Σ. It is in general a challenging statistical problem to estimate the operator norm of the covariance matrix in a high-dimensional setting. One can in principal use the estimator proposed by Liu, Gao and Samworth (2021), though we observe that this estimator typically incurs a large upward bias when the dimension is high in comparison to the sample size. Moreover, an inspsection of our proof reveals that the presence of the additional factor B is used to capture some worst-case large deviation bound, which is often too conservative for a generic covariance Σ. In view of the above, we recommend that practitioners use the same $\lambda = 2^{-1}(1 + \sqrt{4p_*^{-1}\log(nG)})$ when Σ is unknown.

5.3. Comparison between different methods

Now, we would like to compare our method with other existing change-point estimation procedures. As `groupInspect` is a two-stage procedure that first estimates a projection direction before localising the change-point on the projected series, we will investigate its performance both in terms of its accuracy in estimating the projection direction and the quality of the final change-point location estimator. For the former, we compare the estimated projection direction from `groupInspect` with that from the `inspect` algorithm. We measure the

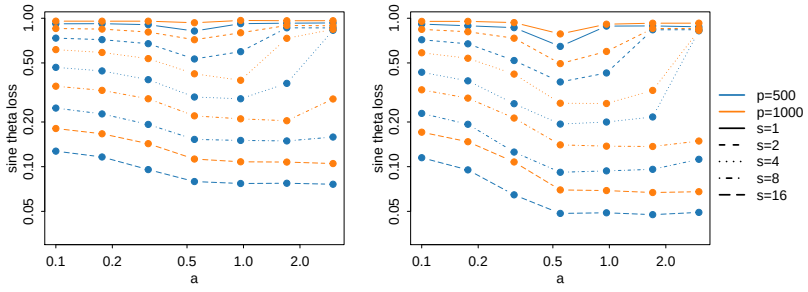

FIG 2. *Average loss (over 100 repetitions) of* `groupInspect` *for tuning parameter* $\lambda = a(1 + \sqrt{4p_*^{-1}\log(nG)})$ *with varying choice of* a. *Left panel:* $G = 10$. *Right panel:* $G = 25$. *Other parameter:* $n = 1000$, $s = 3$.

accuracy in terms of the sine angle loss introduced in Section 2. We use the recommended values for tuning parameters in both methods, i.e., $\sqrt{2^{-1}\log\{p\log n\}}$ in `inspect` as in Wang and Samworth (2018) and $2^{-1}(1 + \sqrt{4p_*^{-1}\log(nG)})$ for `groupInspect` as suggested in Section 5.2.

We fix $n = 1000$, $p = 1000$, vary ϑ in $\{1, 2, 4, 8, 16\}$ and set the covariance matrix to be $\Sigma = I_p$. We consider settings with both non-overlapping groups and overlapping groups. For the non-overlapping setting, we have $G = 10$ groups of equal size $p_* = 100$, whereas for the overlapping setting, we have $G = 19$ groups of size 100 each, where neighbouring groups overlap in exactly 50 coordinates. Both methods have access to exactly the same data sets and the performance is averaged over 100 Monte Carlo repetitions.

Figure 3 shows the comparison of the average sine angle loss between `inspect` and `groupInspect` over all signal sizes on a logarithmic scale, in both the non-overlapping and overlapping settings. In both cases, `groupInspect` outperforms the `inspect` algorithm. From the left panel, we can see that the estimation accuracy of the projection direction using `groupInspect` is substantially better even when the signal is small.

We now turn our attention to the overall change-point localisation accuracy of the `groupInspect` procedure. To this end, we compare the mean absolute deviation of various high-dimensional change-point procedures over 300 Monte Carlo repetitions using the same data sets. In addition to `inspect`, we also compare against the ℓ_2 aggregation procedures of Horváth and Hušková (2012), the ℓ_∞ aggregation procedure of Jirak (2015), the double CUSUM procedure of Cho (2016) and a multiscale testing procedure Pilliat et al. (2020). We set $n = 1000$, $p \in \{500, 1000, 2000\}$, $\vartheta \in \{0.25, 0.5, 1, 2, 4\}$ and $\Sigma = (2^{-|j-k|})_{j,k\in[p]}$. The simulation results are presented in Table 1. For simplicity, we have only shown the results for 10 equal-sized non-overlapping groups here, but qualitatively similar results were obtained in other settings as well. We see that `groupInspect` is very competitive over a wide range of dimensions and signal-to-noise ratio

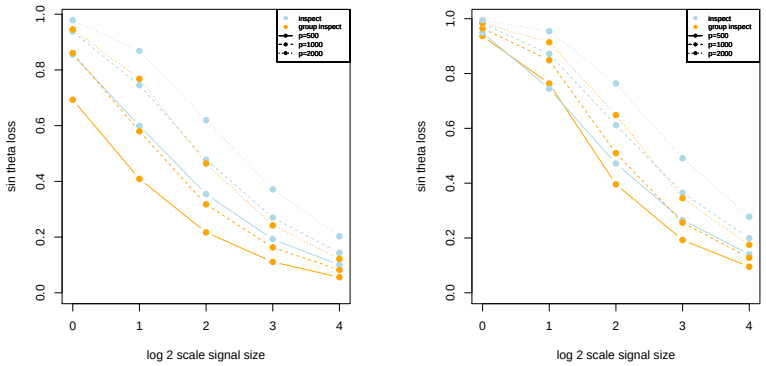

FIG 3. *Average loss (over 100 repetitions) comparison between* `groupInspect` *and* `Inspect`*. Left panel: non-overlap setting. Right panel: overlap setting*

settings, and `groupInspect` dominates the `inspect` procedure in all simulation settings by successfully exploiting the group-sparsity structure.

p	ϑ	`groupInspect`	`inspect`	ℓ_2-aggregate	ℓ_∞-aggregate	double cusum	pilliat
500	0.25	151	158	370	368	364	**113**
500	0.5	**89.6**	98.6	271	332	298	102.6
500	1	**8.7**	14.8	18.5	108	66.8	56.82
500	2	**0.95**	1.30	1.64	15.9	5.42	19.53
500	4	**0.057**	0.063	0.080	3.11	0.51	15
1000	0.25	116	147	368	344	385	**115**
1000	0.5	**85**	120	309	316	335	102
1000	1	**23.4**	32.6	41.0	194	110	67.2
1000	2	**1.31**	1.67	2.04	32.2	7.47	24.36
1000	4	**0.09**	0.14	0.123	6.29	0.850	15
2000	0.25	**106**	128	356	356	374	131
2000	0.5	**89.6**	118	321	344	341	119
2000	1	**47.61**	55.56	106	283	177	92.91
2000	2	**2.91**	3.23	3.39	63.3	10.4	39.141
2000	4	**0.11**	0.160	0.17	9.94	1.32	30.75

TABLE 1

Average mean absolute deviation (over 300 repetitions) comparison between different methods. Other parameters used: $n = 1000$ *with* $G = 10$

5.4. Multiple change-points simulation

The numerical studies so far have focused mainly on the single change-point estimation problem. In this subsection, we investigate the empirical performance of `groupInspect` in multiple change-point estimation tasks. We will compare its performance as implemented in Algorithm 3 to that of the `inspect` algorithms for estimating multiple change-points under different settings. We choose

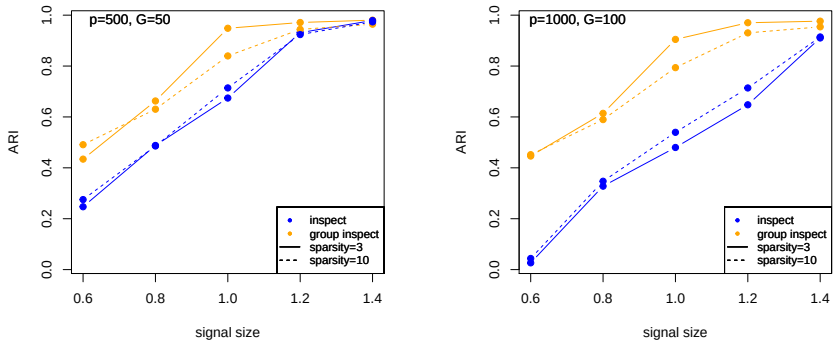

FIG 4. *Average ARI comparsion between* `groupInspect` *and* `inspect`*. Left panel:* $p = 500, G = 50$*. Right panel:* $p = 1000, G = 100$*.*

$n = 1200$, $p \in \{500, 1000\}$, $s \in \{3, 10\}$, $G \in \{50, 100\}$ and $\Sigma = I_p$. Each data series contains three true change-points located at 300, 600 and 900 with the ℓ_2 norm of the change equal to ϑ, 1.5ϑ and 2ϑ respectively. We vary ϑ in $\{0.6, 0.8, 1, 1.2, 1.4\}$. For simplicity, we further assume that the same s coordinates undergo change in all three change-points and that all groups have 10 elements. The total number of coordinates with change k is calculated as $10s$. We use the λ tuning parameter choice suggested in Section 5.2 for the `groupInspect` method and that suggested in Wang and Samworth (2018) for the `inspect` algorithm. For the thresholding parameter ξ of the wild binary segmentation recursion used in both `groupInspect` and `inspect`, we choose via Monte Carlo simulation. More precisely, we randomly generate 1000 data sets from the null model with no change-points and take the maximum absolute CUSUM statistics from Algorithm 3 and Wang and Samworth (2018, Algorithm 4) as ξ_g and ξ_i respectively. We compare the performance of two algorithms using the Adjusted Rand index (ARI) of the estimated segmentation against the truth (Rand, 1971; Hubert and Arabie, 1985).

From Figure 4, we see that the `groupInspect` algorithm generally performs much better than the `inspect` algorithm in the multiple change-point localisation tasks. The advantage of `groupInspect` is more pronounced when the signal is sparser and when the dimension of the data is higher.

To further visualise the output of the two procedures, we plot the estimated change-point locations for one specific setting ($s = 3$ and $\vartheta = 1$) of each of the two panels in Figure 4. The resulting histograms in Figure 5 shows that when $p = 500$, `groupInspect` was better at picking out all three change-points with higher accuracies. When $p = 1000$, `inspect` was only able to pick out the change at $t = 600$ in most of the trials, whereas `groupInspect` was still able to identify even the weakest change signal at $t = 300$ in a substantial fraction of all trials.

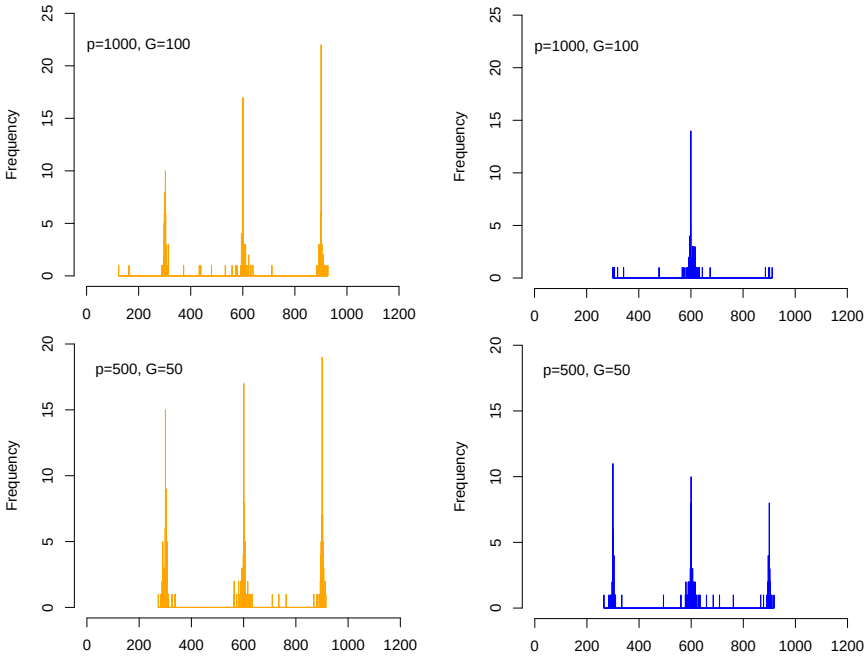

FIG 5. *Histograms of estimated locations by* `groupInspect` *and* `inspect` *under two settings when* $P = 500, G = 50$ *and* $p = 1000, G = 100$. *Other parameter used:* $s = 3$, $\vartheta = 1$ *are fixed in both settings.*

5.5. Real data analysis

In this section, we apply `groupInspect` to an S&P 500 daily stock return dataset. The data consist of the logarithmic daily returns (computed from the adjusted closing prices) of S&P 500 stocks traded during the period of 1 January 2007 to 31 December 2011. We only included the 257 stocks which have continuously traded throughout this this period to construct a multivariate time series of dimension $p = 257$ and length $n = 1259$. We divided the 257 companies into $G = 11$ non-overlapping groups according to their Global Industry Classification Standard sector memberships. For each stock logarithmic returns, we fitted an AR(1) model, and then rescaled the residuals by their estimated standard deviation according to the method described in Section 5.

Figure 6 displays the ten most significant change-points identified by our `groupInspect` algorithm. For each change-point, we derived a sector-weighting vector from the estimated projection direction by `groupInspect`. Specificially, given the projection direction $\hat{v} \in \mathbb{S}^{p-1}$ for each estimated change-point, and the grouping $(\mathcal{J}_g)_{g\in[G]}$, we computed a weight vector $\hat{w} := (\|\hat{v}_{\mathcal{J}_g}\|)_{g\in[G]}$. This vector gives us information about which sectors had driven the change for each change-point estimated. For instance, we see from Figure 6 that the the change-point at 12 Sep 2008 was predominantly driven by price fluctuations in financial stocks, which coincides with the Federal takeover of Fannie Mae and Freddie Mac on 7 Sep 2008 and the bankruptcy of Lehman Brothers on 15 Sep 2008. The change-point identified at 10 Feb 2009, though still heavily weighted on financial stocks, showed a broader impact across other sectors. This is consistent with the passing of the American Reovery and Reinvestment Act of 2009 on 13 Feb 2009 sending a general positive signal to the entire economy.

6. Proofs of main results

In this section, we will give the proof of our results in section 4.

6.1. Proof of Theorem 1

Proof. From the definition of the CUSUM transformation in (5), we can explicitly write the matrix $A := \mathbb{E}(T) = (A_{j,t})_{j\in[p],t\in[n-1]}$ as

$$A_{j,t} = \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n-z)\theta_j & \text{if } 1 \le t \le z, \\ \sqrt{\frac{n-t}{nt}}z\theta_j & \text{if } z < t \le n-1. \end{cases}$$

In particular, we have that A is a rank 1 matrix of the form

$$A = \theta\gamma^\top, \tag{14}$$

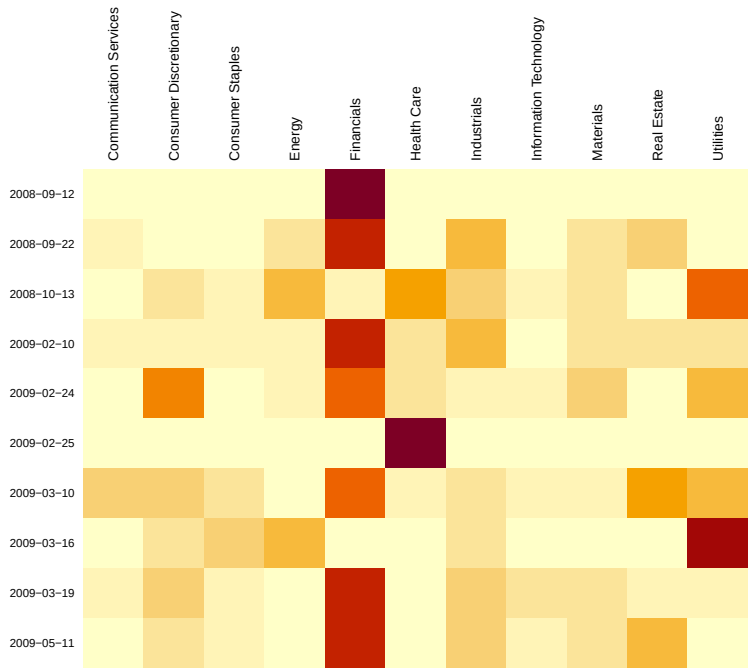

FIG 6. *Estimated change point locations (red dashed lines) by* `groupInspect` *applied to the stock return data. For ease of illustration, we have plotted the* ℓ_2 *norm of the returns of all stocks within each of the 11 groups over time.*

with

$$\gamma = \frac{1}{\sqrt{n}}\Big(\sqrt{\frac{1}{n-1}}(n-z), \sqrt{\frac{2}{n-2}}(n-z), \cdots, \sqrt{z(n-z)}, \sqrt{\frac{n-z-1}{z+1}}z, \cdots, \sqrt{\frac{1}{n-1}}z\Big)^\top.$$

By Wang and Samworth (2018, Lemma 3), we have $\|\gamma\|_2 \geq n\tau/4$, so $\|A\|_{\mathrm{op}} \geq n\tau\vartheta/4$. By Lemma 14 with $\delta = (nG)^{-4}$, we have

$$\mathbb{P}(\|T - A\|_{\mathrm{grp}*} > \lambda) < \frac{1}{(nG)^3}.$$

By Proposition 12, on the event $\{\|T - A\|_{\mathrm{grp}_*} \leq \lambda\}$, we have

$$\max\big\{\sin\angle(v, \hat{v}), \sin\angle(u, \hat{u})\big\} \leq \frac{32\lambda(C_1 k)^{1/2}}{n^{1/2}\tau\vartheta},$$

as desired. □

6.2. *Proof of Theorem 2*

Proof. We will use two different constructions to derive separate lower bounds of order $\sqrt{Bs\log(G/s)/(n\tau\vartheta^2)}$ and $\sqrt{Bk/(n\tau\vartheta^2)}$ respectively. Without loss of generality, we may assume that $z < n/2$.

For the first bound, let $s_0 = s - 1$, $G_0 = G - 1$. By the Gilbert–Varshamov lemma as stated in Massart (2007, Lemma 4.10) (applied with $\alpha = 3/4$ and $\beta = 1/3$), we can construct a set $\mathcal{U}_0$ of s_0-sparse vectors in $\{0,1\}^{G_0}$, with cardinality at least $(G_0/s_0)^{s_0/5}$, such that the pairwise Hamming distance between any pair of vectors in $\mathcal{U}_0$ is at least $s_0/2$. Let $\epsilon \in (0,1)$ to be chosen later, we can define a set

$$\mathcal{U} = \left\{ \begin{pmatrix} \sqrt{1-\epsilon^2} \\ s_0^{-1/2}\epsilon u_0 \end{pmatrix} : u_0 \in \mathcal{U}_0 \right\} \subseteq \mathbb{S}^{G-1}.$$

We remark that for any pair of distinct $u, u' \in \mathcal{U}$, we have by construction that $\epsilon/\sqrt{2} \leq \|u' - u\|_2 \leq \epsilon$. We then define a map $\psi : \mathbb{R}^G \to \mathbb{R}^p$ such that for any $u \in \mathcal{U}$ and $j \in \mathcal{J}_g$, we have $\psi(u)_j = u_g p_g^{-1/2}$. Finally, let $\mathcal{V} = \{\psi(u) : u \in \mathcal{U}\}$. We note that $\|\psi(u') - \psi(u)\|_2 = \|u' - u\|_2$. Therefore, for distinct $v, v' \in \mathcal{V}$, we have

$$L(v', v) = \sqrt{1 - (v^\top v')^2} = \frac{\|v' - v\|_2}{\sqrt{2}} \geq \frac{\epsilon}{2}. \tag{15}$$

Now, for each $v \in \mathcal{V}$, we define a distribution $P_v \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$, such that the pre-change mean is $-\vartheta v$ and the post-change mean is 0 (we check that P_v indeed satisfies the conditions of $\mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$). Then for

any distinct $v, v' \in \mathcal{V}$, we have

$$D(P_v \| P_{v'}) = zD(N_p(-v\vartheta, B) \| N_p(-v'\vartheta, B)) \leq \frac{z\vartheta^2}{2B}\|v - v'\|_2^2 \leq \frac{z\vartheta^2\epsilon^2}{2B}. \tag{16}$$

By (15) and (16), we can apply Fano's lemma (Yu, 1997, Lemma 3) to obtain that

$$\inf_{\tilde{v}} \sup_{P \in \mathcal{P}_{n,p}^{(1)}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) \geq \inf_{\tilde{v}} \sup_{v\in\mathcal{V}} \mathbb{E}_{P_v} L(\tilde{v}(X), v) \geq \frac{\epsilon}{4}\left\{1 - \frac{z\vartheta^2\epsilon^2/(2B) + \log 2}{(s_0/5)\log(G_0/s_0)}\right\}.$$

By the condition $(s-1)\log(G/s) \geq 20$, we have $(s_0/5)\log(G_0/s_0) \geq 2\log 2$. Moreover, the choice of

$$\epsilon = \sqrt{\frac{Bs_0\log(G_0/s_0)}{10z\vartheta^2}}$$

ensures that $(s_0/5)\log(G_0/s_0) \geq 2z\vartheta^2\epsilon^2/B$. Therefore,

$$\inf_{\tilde{v}} \sup_{P \in \mathcal{P}_{n,p}^{(1)}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) \geq \frac{\epsilon}{16} \geq \frac{1}{72}\sqrt{\frac{Bs\log(G/s)}{z\vartheta^2}}. \tag{17}$$

For the second lower bound, let $g_1, \ldots, g_s$ be the indices of the s groups with largest cardinalities. By the given condition of the Theorem, we have that $\tilde{k} = \sum_{r=1}^{s} p_{g_r} = \sum_{r=1}^{s} p_{(G-r+1)} \geq k/2$. Let $S = \cup_{r=1}^{s}\mathcal{J}_{g_r}$, so $|S| = \tilde{k}$. By Massart (2007, Lemma 4.7), we can construct a subset $\mathcal{V}_0$ of $\{-1,1\}^{\tilde{k}_0}$ of cardinality at least $e^{\tilde{k}/8}$, such that any two points in the set are separated in Hamming distance by at least $\tilde{k}/4$. Construct

$$\mathcal{V} = \left\{v : v_S = \begin{pmatrix} \sqrt{1-\epsilon^2} \\ \tilde{k}_0^{-1/2}\epsilon v_0 \end{pmatrix} \text{ for some } v_0 \in \mathcal{V}_0 \text{ and } v_{S^c} = 0\right\}.$$

Therefore, for distinct $v, v' \in \mathcal{V}$, we have $\epsilon \leq \|v' - v\|_2 \leq 2\epsilon$,then,

$$L(v', v) = \sqrt{1 - (v^\top v')^2} = \frac{\|v' - v\|_2}{\sqrt{2}} \geq \frac{\epsilon}{\sqrt{2}}.$$

Following the same derivation as in (16), we have that

$$D(P_v \| P_{v'}) = zD(N_p(-v\vartheta, \Sigma) \| N_p(-v'\vartheta, \Sigma)) \leq \frac{z\vartheta^2}{2B}\|v - v'\|_2^2 \leq \frac{2z\vartheta^2\epsilon^2}{B}.$$

Again, we can use Fano's lemma (Yu, 1997, Lemma 3) to obtain that

$$\inf_{\tilde{v}} \sup_{v\in\mathcal{V}} \mathbb{E}_{P_v} L(\tilde{v}(X), v) \geq \frac{\epsilon}{\sqrt{2}} \left\{ 1 - \frac{2z\vartheta^2\epsilon^2/B + \log 2}{\tilde{k}/8} \right\} \geq \frac{\epsilon}{\sqrt{2}} \left\{ 1 - \frac{2z\vartheta^2\epsilon^2/B + \log 2}{k/16} \right\}.$$

Now, choose $\epsilon = (kB)^{1/2} z^{-1/2} \vartheta^{-1}/4\sqrt{6}$. Since $k \geq 20$, we have $k/16 \geq 9\log(2)/5$, so that

$$\inf_{\tilde{v}} \sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) \geq \inf_{\tilde{v}} \sup_{v\in\mathcal{V}} \mathbb{E}_{P_v} L(\tilde{v}(X), v) \geq \frac{\epsilon}{9\sqrt{2}} \geq \frac{1}{72\sqrt{3}} \sqrt{\frac{kB}{z\theta^2}}. \quad (18)$$

The desired result follows by combining (17) with (18), and noting that $z \geq n\tau$. □

6.3. *Proof of Theorem 3*

Proof. Recall the definition of $X^{(2)}$ and let $T^{(2)} = \mathcal{T}(X^{(2)})$. Define similarly $\mu^{(2)} = (\mu^{(2)}_1, \ldots, \mu^{(2)}_{n_1}) \in \mathbb{R}^{p\times n_1}$ and a random $W^{(2)} = (W^{(2)}_1, \ldots, W^{(2)}_{n_1})$ taking values in $\mathbb{R}^{p\times n_1}$ by $\mu^{(2)}_t = \mu_{2t}$ and $W^{(2)}_t = W_{2t}$. Now, let $A^{(2)} = \mathcal{T}(\mu^{(2)})$ and $E^{(2)} = \mathcal{T}(W^{(2)})$. We also write $\bar{X} = (\hat{v}^{(1)})^\top X^{(2)}$,$\bar{\mu} = (\hat{v}^{(1)})^\top \mu^{(2)}$,$\bar{W} = (\hat{v}^{(1)})^\top W^{(2)}$,$\bar{A} = (\hat{v}^{(1)})^\top A^{(2)}$, $\bar{E} = (\hat{v}^{(1)})^\top E^{(2)}$ and $\bar{T} = (\hat{v}^{(1)})^\top T^{(2)}$ for the one-dimensional projected images. Note that by linearity, we have $\bar{T} = \mathcal{T}(\bar{X})$, $\bar{A} = \mathcal{T}(\bar{\mu})$ and $\bar{E} = \mathcal{T}(\bar{W})$,

Now, conditional on $\hat{v}^{(1)}$, the random variables $\bar{X}_1, \ldots, \bar{X}_{n_1}$ are independent with

$$\bar{X}_t \mid \hat{v}^{(1)} \sim N(\bar{\mu}_t, \sigma^2)$$

and the row vector $\bar{\mu}$ undergoes a single change at $z^{(2)} = z/2$ with magnitude of change

$$\bar{\theta} = \bar{\mu}_{z^{(2)}+1} - \bar{\mu}_{z^{(2)}} = \hat{v}^{(1)\top}\theta.$$

Finally, let $\hat{z}^{(2)} \in \arg\max_{1\leq t\leq n_1-1} |\bar{T}_t|$, so the first component of the output of the algorithm is $\hat{z} = 2\hat{z}^{(2)}$. Consider the set

$$\Upsilon = \{u \in \mathbb{S}^{p-1} : \sin\angle(u, v) \leq 1/2\}.$$

By Condition (11) and Theorem 1, we have that

$$\mathbb{P}(\hat{v}^{(1)} \in \Upsilon) \geq 1 - \frac{1}{(n_1 G)^3}. \quad (19)$$

Moreover, on the event $\{\hat{v}^{(1)} \in \Upsilon\}$, we have that $|\bar{\theta}| \geq \sqrt{3}\vartheta/2$. Noting that we have $\bar{E}_t \mid \hat{v}^{(1)} \sim N(0, \hat{v}^{(1)\top}\Sigma\hat{v}^{(1)})$, we have by Wang and Samworth (2018, Lemma 4) for any $\lambda_1 \geq \sqrt{B}$ that

$$\mathbb{P}(\|\bar{E}\|_\infty \geq \lambda_1) \leq \sqrt{\frac{2}{\pi}}\lceil \log n_1 \rceil \left(\frac{\lambda_1}{\sqrt{B}} + 2\right) e^{-\lambda_1^2/B} \leq 3\lambda_1 e^{-\lambda_1^2/B} \log n. \tag{20}$$

Define $\Omega_0 := \{\hat{v}_1 \in \Upsilon, \|\bar{E}\|_\infty \leq \lambda_1\}$. From (19) and (20), we have $\mathbb{P}(\Omega_0) \geq 1 - n_1^{-3} - 3\lambda_1 e^{-\lambda_1^2/B} \log n$.

Notice that the procedure produces the same output if we replace $\hat{v}^{(1)}$ by $-\hat{v}^{(1)}$, hence we may assume without loss of generality that $\bar{\theta} \geq 0$, which implies that $\bar{A}_t \geq 0$ for all $t \in [n_1 - 1]$. Condition (11) implies that

$$\sqrt{n\tau}\vartheta \geq C\lambda_1, \tag{21}$$

for sufficient large C. Therefore, by Lemma 16 and (21), if we choose $C \geq 8/\sqrt{3}$, then for t satisfying $|z^{(2)} - t| \geq n_1\tau/2$, we have

$$A_{z^{(2)}} = \sqrt{\frac{z^{(2)}(n_1 - z^{(2)})}{n_1}}\bar{\theta} \geq \sqrt{\frac{n_1\tau}{2}}\bar{\theta} \geq \frac{\sqrt{3}}{4}\sqrt{n\tau}\vartheta \geq 2\lambda_1.$$

In particular, we must have on Ω_0 that $T_{\hat{z}^{(2)}} \geq T_{z^{(2)}} \geq A_{z^{(2)}} - \lambda_1 \geq -A_t + \lambda_1 \geq -T_t$ for any $t \in [n-1]$. Hence, $\arg\max_{t\in[n-1]} |\bar{T}_t| = \arg\max_{t\in[n-1]} \bar{T}_t$.

Since $\bar{T} = \bar{A} + \bar{E}$ and $(\bar{A}_t)_t$ and $(\bar{T}_t)_t$ are respectively maximized at $t = z^{(2)}$ and $t = \hat{z}^{(2)}$. We have on the event Ω_0 that

$$\bar{A}_{z^{(2)}} - \bar{A}_{\hat{z}^{(2)}} = (\bar{A}_{z^{(2)}} - \bar{T}_{\hat{z}^{(2)}}) + (\bar{T}_{z^{(2)}} - \bar{T}_{\hat{z}^{(2)}}) + (\bar{T}_{\hat{z}^{(2)}} - \bar{A}_{\hat{z}^{(2)}}) \leq \bar{E}_{\hat{z}^{(2)}} - \bar{E}_{z^{(2)}}. \tag{22}$$

Note that on Ω_0, the right-hand side of (22) is bounded by $2\lambda_1$. Hence, applying Lemma 16 to the left-hand side of (22), and using the unimodality of $\bar{A}$, if $C \geq 24$, on the event Ω_0, we have that

$$\frac{|\hat{z}^{(2)} - z^{(2)}|}{n_1\tau} \leq \frac{3\sqrt{6}\lambda_1}{\bar{\theta}\sqrt{n_1\tau}} \leq \frac{12\lambda_1}{\vartheta\sqrt{n\tau}} \leq \frac{1}{2}.$$

By Lemma 15, there exists an event Ω_1 with probability at least $1 - e^{-\lambda_1^2/(2B)} \log n$ on which

$$|\bar{E}_{z^{(2)}} - \bar{E}_{\hat{z}^{(2)}}| \leq 4\lambda_1\sqrt{\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}} + 16\lambda_1\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}. \tag{23}$$

Substituting the improved bound of (23) into the right-hand side of (22), and again applying Lemma 16 to the left-hand side of (22), we have on $\Omega_0 \cap \Omega_1$ that

$$\frac{\vartheta}{3}\frac{|z^{(2)} - \hat{z}^{(2)}|}{\sqrt{n\tau}} \leq 4\lambda_1\sqrt{\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}} + 16\lambda_1\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}.$$

When $C \geq 96$, from (11), we have $16\lambda_1 \frac{|z^{(2)}-\hat{z}^{(2)}|}{n\tau} \leq \frac{\vartheta}{6} \frac{|z^{(2)}-\hat{z}^{(2)}|}{\sqrt{n\tau}}$. Consequently, on $\Omega_0 \cap \Omega_1$, we have

$$|\hat{z} - z| \leq \frac{C'\lambda_1^2}{\vartheta^2},$$

as desired. Finally, we compute that the desired event occurs with probability

$$\mathbb{P}(\Omega_0 \cap \Omega_1) \geq 1 - \frac{1}{n_1^3} - (3\lambda_1 + 1)e^{-\lambda_1^2/(2B)} \log n.$$

as desired. □

6.4. Proof of Corollary 4

Proof. Define $A := \mathbb{E}(T)$ and $E := T - A$. Under null hypothesis where there is no change in the segment, by Lemma 14, we have that $\mathbb{P}(\|T\|_{\text{grp}^*} \geq \lambda) = \mathbb{P}(\|E\|_{\text{grp}^*} \geq \lambda) < 1/(nG)$.

Under the alternative, we have:

$$\|T\|_{\text{grp}^*} = \|A + E\|_{\text{grp}^*} \geq \|A\|_{\text{grp}^*} - \|E\|_{\text{grp}^*}.$$

By (14), we have

$$\|A\|_{\text{grp}^*} = \|\theta\gamma^\top\|_{\text{grp}^*} = \|\theta\|_\infty \max_{g\in[G]} p_g^{-1/2} \|\theta_{\mathcal{J}_g}\|_2 \geq \frac{\|\gamma\|_\infty \|\theta\|_2}{\sqrt{k}}.$$

Also, by definition of γ, we have that $\|\gamma\|_\infty = \sqrt{\frac{z(n-z)}{n}} \geq \sqrt{n\tau/2}$. Therefore, for $\|\theta\|_2 \geq \frac{2\sqrt{2k}\lambda}{\sqrt{n\tau}}$, combining with Lemma 13, we have that with probability at least $1 - 1/(nG)$ that $\|T\|_{\text{grp}^*} \geq 2\lambda - \lambda = \lambda$. □

6.5. Proof of Theorem 5

Proof. Let $\{z_1, \ldots, z_\nu\}$ be the set of true change points, such that $0 =: z_0 < z_1 < \cdots < z_\nu < n =: z_{\nu+1}$. For each $i \in [\nu]$, define intervals

$$\mathcal{I}_i^L = \big(z_i - n\tau/3, z_i - n\tau/6\big) \quad \text{and} \quad \mathcal{I}_i^R = \big(z_i + n\tau/6, z_i + n\tau/3\big).$$

These intervals contain at least one integer for $n\tau \geq 6$. For simplicity of exposition, we have ignored various rounding issues in this proof. Now, define the following event:

$$\Omega_0 := \{\forall i \in [\nu], \exists m \in [M], \text{ s.t. } (s_m, e_m) \in \mathcal{I}_i^L \times \mathcal{I}_i^R\}.$$

Then, we have

$$\mathbb{P}(\Omega_0^c) \leq \sum_{i=1}^{\nu} \prod_{m=1}^{M} \Big(1 - \mathbb{P}((s_m, e_m) \in \mathcal{I}_i^L \times \mathcal{I}_i^R)\Big) \leq \nu \left(1 - \frac{\tau^2}{36}\right)^M \leq \nu e^{-\tau^2 M/36}.$$

On Ω_0, for each change point z_i, we can find an interval $(s_m, e_m]$ which only captures one change-point, which is at least $n\tau/6$ away from the endpoints s_m and e_m of the interval.

We write $X^{(s,e]}$ for the submatrix of X obtained by extracting columns indexed in $(s, e]$. Let $T^{(s,e]} := \mathcal{T}(X^{(s,e]})$, $A^{(s,e]} := \mathbb{E}T^{(s,e]}$ and $E^{(s,e]} := T^{(s,e]} - A^{(s,e]}$. Set

$$\Omega_1 := \Big\{ \max_{1\le s<e\le n} \|E^{(s,e]}\|_{\mathrm{grp}*} < \lambda \Big\}.$$

By Lemma 14 and a union bound, we have that

$$\mathbb{P}(\Omega_1^{\mathrm{c}}) \le n^2 \frac{(n-1)G}{(nG)^4} \le \frac{1}{nG^3}.$$

Now, for any interval $(s, e]$, we write $\hat{z}^{(s,e]}$ to be the change-point estimate of Algorithm 4 applied to data $X^{(s,e]}$. We define $O := \{(s, e) : 0 \le s < e \le n,\ z_{i-1} \le s < z_i < e < z_{i+1}$ for some $i \in [\nu]$ and $\min\{z_i - s, e - z_i\} \ge n\tau/10\}$ to be the set of intervals $(s, e]$ that captures exactly one true change-point, which is at least $n\tau/10$ away from the boundaries. We then define the event

$$\Omega_2 := \Big\{ |\hat{z}^{(s,e]} + s - z_i| \le \frac{C' B \log n}{\vartheta^2} \text{ for all } (s, e] \in O \Big\}.$$

For a sufficiently large C and C', by Condition (3) and Theorem 3 applied with $\lambda_1 = \sqrt{16B\log(n\tau B)}$, together with union bound, we have that

$$\mathbb{P}(\Omega_2^{\mathrm{c}}) \le \frac{7}{n\tau^3}.$$

We will henceforth work on $\Omega_0 \cap \Omega_1 \cap \Omega_2$.

For any interval $(s, e] \subseteq (0, n]$, we define $\mathcal{Z}^{(s,e]} := \{z_i : i \in [\nu] \text{ and } z_i \in (s, e]\}$ and the following subsets of $\mathcal{Z}^{(s,e]}$:

$$\mathcal{Z}^{(s,e]}_{\mathrm{good}} := \{z \in \mathcal{Z}^{(s,e]} : \min\{z - s, e - z\} \ge n\tau/3\},$$
$$\mathcal{Z}^{(s,e]}_{\mathrm{bad}} := \Big\{ z \in \mathcal{Z}^{(s,e]} : \min\{z - s, e - z\} \le \frac{C' B \log n}{\vartheta^2} \Big\},$$

where C' is chosen to be the same constant as in the definition of Ω_2. We note that $\mathcal{Z}^{(s,e]}_{\mathrm{good}}$ and $\mathcal{Z}^{(s,e]}_{\mathrm{bad}}$ respectively contain change-points within $(s, e]$ that are well-separated from the boundary and close to the boundary. We will informally refer to these change-points as "good" and "bad" change-points in $(s, e]$. On Ω_0, for every $i \in [\nu]$, we can associate it with an $m_i \in [M]$ such that $s_{m_i} \in \mathcal{I}_i^L$ and $e_{m_i} \in \mathcal{I}_i^R$. We claim that

$$\{m_i : z_i \in \mathcal{Z}^{(s,e]}_{\mathrm{good}}\} \subseteq \mathcal{R}_{s,e}. \tag{24}$$

To see this, we first note that from the definition of $\mathcal{I}_i^L$ and $\mathcal{I}_i^R$, and the condition $\min\{z_i - s, e - z_i\} \ge n\tau/3$ that for every i with $z_i \in \mathcal{Z}^{(s,e]}_{\mathrm{good}}$ we have $(s_{m_i}, e_{m_i}] \subseteq$

$(s, e]$. On Ω_1, by Condition (13) with a sufficiently large choice of $C > 0$ and the proof of Corollary 4 we have

$$\|T^{(s_{m_i}+\beta, e_{m_i}-\beta]}\|_{\mathrm{grp}*} \geq \lambda.$$

Hence $m_i \in \mathcal{R}_{s,e}$, establishing the claim. On the other hand, under Condition (13) for sufficiently large C, we have $\frac{C'B\log n}{\vartheta^2} < n\tau/10 = \beta$. Hence on Ω_1, for any $(s_0, e_0] \subseteq (s, e]$ containing only "bad" change-points, i.e. $(s_0, e_0] \cap \mathcal{Z}^{(s,e]} \subseteq \mathcal{Z}^{(s,e]}_{\mathrm{bad}}$, we get:

$$\|T^{(s_0+\beta, e_0-\beta]}\|_{\mathrm{grp}*} < \lambda,$$

as there are no change points within the interval $(s_0 + \beta, e_0 - \beta]$. Thus,

$$\left\{m \in \mathcal{M}_{s,e} : (s_m, e_m] \cap \mathcal{Z}^{(s,e]} \subseteq \mathcal{Z}^{(s,e]}_{\mathrm{bad}}\right\} \cap \mathcal{R}_{s,e} = \emptyset \tag{25}$$

Given a set $\hat{Z}$ of estimated change-points, we can partition $(0, n]$ into $|\hat{Z}| + 1$ segments. We call these the segments induced by $\hat{Z}$. We now prove by induction that throughout the recursion of NOT, the following statement holds:

$$\text{For any } (s, e] \text{ induced by } \hat{Z},\ \mathcal{Z}^{(s,e]} = \mathcal{Z}^{(s,e]}_{\mathrm{good}} \cup \mathcal{Z}^{(s,e]}_{\mathrm{bad}}. \tag{P}$$

For the base case, at the beginning of the algorithm, we have $\hat{Z} = \emptyset$, so the only induced segment by $\hat{Z}$ is $(0, n]$. The statement (P) is true since the cloest change-point from the boundary is at least $n\tau$ away. Now assuming that (P) is true at some stage of the recursion when $\hat{Z}$ is the set of estimated change-points so far, we need to show that (P) still holds when a new change-point is estimated by NOT. This new change-point must be identified from running NOT on some $(s, e]$ where $(s, e]$ is one of the induced segments by $\hat{Z}$. From the inductive hypothesis, we know that $\mathcal{Z}^{(s,e]} = \mathcal{Z}^{(s,e]}_{\mathrm{good}} \cup \mathcal{Z}^{(s,e]}_{\mathrm{bad}}$. We note that $\mathcal{Z}^{(s,e]}_{\mathrm{good}}$ is necessarily nonempty for otherwise by (25) we have $\mathcal{M}_{s,e} \cap \mathcal{R}_{s,e} = \emptyset$ and hence $\mathcal{R}_{s,e} = \emptyset$, so no new change-point will be identified in $(s, e]$. Thus, there exists some i' with $z_{i'} \in \mathcal{Z}^{(s,e]}_{\mathrm{good}}$ and by (24), $m_{i'} \in \mathcal{R}_{s,e}$ and hence $e_{m^*} - s_{m^*} \leq e_{m_{i'}} - s_{m_{i'}} \leq n\tau/3$. In particular, we have that $(s_{m^*}, e_{m^*}]$ must capture exactly one change-point (it has to capture at least one change-point by (25) and cannot capture more than one since two consecutive change-points are spaced at least $n\tau$ away), say z_{i^*}. On the event Ω_2, we know that the change-point output $\hat{z}$ of Algorithm 4 on $X^{(s_{m^*}, e_{m^*}]}$ satisfies

$$|\hat{z} + s_{m^*} - z_{i^*}| \leq \frac{C'B\log n}{\vartheta^2}. \tag{26}$$

We now check that the two new segments induced by $\hat{Z} \cup \{\hat{z} + s_{m^*}\}$ still satisfy (P). For this, it suffices to check that z_{i^*-1}, z_{i^*} and z_{i^*+1} are either within $\frac{C'B\log n}{\vartheta^2}$ of $\hat{z} + s_{m^*}$ or at least $n\tau/3$ away from it. This can be seen by combining (26) with the fact that $\min\{z_{i^*} - z_{i^*-1}, z_{i^*+1} - z_{i^*}\} \geq n\tau$. This completes the induction.

We remark that as a side product of the above inductive argument, we have shown that if $(s, e] \cap \mathcal{Z}_{\text{good}}^{(s,e]} \neq \emptyset$, then $\mathcal{R}_{s,e}$ is non-empty and NOT will estimate a new change-point. Hence, at the end of the recursion, we must have that all segments induced by $\hat{Z}$ contains no change-point at least $n\tau/3$ away from the boundaries. In other words, all change-points $z_1, \ldots, z_\nu$ must be at most $n\tau/10$ away from the endpoints of one of the induced segments. This, together with the fact that consecutive change-points (including z_0 and z_{n+1}) are spaced at least $n\tau$ away, means that there must be exactly ν estimated change-points in $\hat{Z}$ at the end of the algorithm. Let $\hat{z}_1 < \hat{z}_2 < \cdots < \hat{z}_\nu$ be elements of $\hat{Z}$ arranged in an increasing order. Then, since all change-points are "bad" at the end of the NOT recursion, we must have

$$\max_{i \in [\nu]} |\hat{z}_i - z_i| \leq \frac{C' B \log n}{\vartheta^2}$$

as desired. □

Appendix A: Ancillary results

We collect in this section all ancillary propositions and lemmas used in the paper. For all results in this section, we assume that we are given a grouping $(\mathcal{J}_g)_{g \in [G]}$ of $[p]$ and the associated group norm $\|\cdot\|_{\text{grp}}$.

Proposition 6. *Fix $n \in \mathbb{N}$. Let $P_{z,\mu_{\text{L}},\mu_{\text{R}}}$ denote the joint distribution of $(X_i)_{i \in [n]}$ such that $X_i \sim N(\mu_i, \sigma^2)$ are independent random variables with $\mu_i = \mu_{\text{L}} \mathbb{1}_{\{i \leq z\}} + \mu_{\text{R}} \mathbb{1}_{\{i > z\}}$. Then*

$$\inf_{\hat{z}} \sup_{(z,\mu_{\text{L}},\mu_{\text{R}}) \in [n-1] \times \mathbb{R}^2} \mathbb{E}_{P_{z,\mu_{\text{L}},\mu_{\text{R}}}} |\hat{z} - z| (\mu_{\text{L}} - \mu_{\text{R}})^2 \geq c\sigma^2 \log\log n.$$

Proof. Suppose $n = 2^L$ for some $L \in \mathbb{N}$. For $\ell \in [L]$, we define $\boldsymbol{\mu}^{(\ell)} \in \mathbb{R}^{2n}$ to be the vector whose last 2^ℓ entries are equal to $\sqrt{\sigma^2 2^{-\ell} \log\log_2(2n)/60}$ and the remaining entries are 0. Gao et al. (2020, Theroem 2.2 and the argument immediately above its statement) shows that for some universal constant $c_1 > 0$, we have

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\ell \in [L]} \mathbb{E}_{\boldsymbol{\mu}^{(\ell)}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(\ell)}\|_2^2 \geq c_1 \sigma^2 \log\log(16n). \tag{27}$$

Let $c > 0$ be a constant to be chosen later. We assume that the conclusion of the proposition does not hold, which means that there exists an estimator $\hat{z}$ such that for all $z \in [n-1]$ and $\mu_{\text{L}}, \mu_{\text{R}} \in \mathbb{R}$, we have

$$\mathbb{E}_{P_{z,\mu_{\text{L}},\mu_{\text{R}}}} |\hat{z} - z| < \frac{c\sigma^2 \log\log n}{(\mu_{\text{L}} - \mu_{\text{R}})^2}. \tag{28}$$

Let $(Z_i)_{i \in [2n]}$ be a sequence of $2n$ independent random variables such that $Z_i \sim N(\mu_{\text{L}} \mathbb{1}_{\{i \leq 2z\}} + \mu_{\text{R}} \mathbb{1}_{\{i > 2z\}}, \sigma^2)$. We can apply the estimator $\hat{z}$ on data

$\mathcal{Z}_{\text{odd}} := (Z_1, Z_3, \ldots, Z_{2n-1})$ of length n to obtain a changepoint location estimate $\hat{z}(\mathcal{Z}_{\text{odd}})$, which for notational simplicity, we will denote also as $\hat{z}$ henceforth. Now, define

$$\hat{\mu}_{\text{L}} := \frac{1}{\hat{z}} \sum_{i=1}^{\hat{z}} Z_{2i} \quad \text{and} \quad \hat{\mu}_{\text{R}} := \frac{1}{n-\hat{z}} \sum_{i=\hat{z}+1}^{n} Z_{2i}.$$

Then the vector $\hat{\boldsymbol{\mu}} := (\hat{\mu}_{\text{L}} \mathbb{1}_{\{i \le 2\hat{z}\}} + \hat{\mu}_{\text{R}} \mathbb{1}_{\{i > 2\hat{z}\}})_{i \in [2n]}$ is an estimator of $\boldsymbol{\mu} := (\mathbb{E} Z_i)_{i \in [2n]}$. Without loss of generality, we may assume that $\hat{z} \ge z$; the opposite case can be handled symmetrically. This means that

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 = 2z(\hat{\mu}_{\text{L}} - \mu_{\text{L}})^2 + 2(\hat{z} - z)(\hat{\mu}_{\text{L}} - \mu_{\text{R}})^2 + 2(n - \hat{z})(\hat{\mu}_{\text{R}} - \mu_{\text{R}})^2 \quad (29)$$

Using independence between $\hat{z}$ and $(Z_{2i})_{i \in [n]}$, we have $\hat{\mu}_{\text{L}} \mid \mathcal{Z}_{\text{odd}} \sim N\big(\frac{z}{\hat{z}} \mu_{\text{L}} + \frac{\hat{z}-z}{\hat{z}} \mu_{\text{R}}, \sigma^2/\hat{z}\big)$ and $\hat{\mu}_{\text{R}} \mid \mathcal{Z}_{\text{odd}} \sim N\big(\mu_{\text{R}}, \sigma^2/(n-\hat{z})\big)$. Hence, from (29), we have

$$\begin{aligned}
\mathbb{E}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \mid \mathcal{Z}_{\text{odd}}) &= 4\sigma^2 + (\mu_{\text{L}} - \mu_{\text{R}})^2 \left\{ \frac{2z(\hat{z} - z)^2}{\hat{z}^2} + \frac{2z^2(\hat{z} - z)}{\hat{z}^2} \right\} \\
&\le 4\sigma^2 + 4(\mu_{\text{L}} - \mu_{\text{R}})^2 (\hat{z} - z).
\end{aligned}$$

Then, since $\mathbb{E}_{P_{z,\mu_{\text{L}},\mu_{\text{R}}}} |\hat{z} - z| < \frac{c\sigma^2 \log\log n}{(\mu_{\text{L}} - \mu_{\text{R}})^2}$, we have

$$\begin{aligned}
\mathbb{E}_{P_{z,\mu_{\text{L}},\mu_{\text{R}}}} (\|\hat{\mu} - \mu\|_2^2) &\le 4\sigma^2 + 4(\mu_L - \mu_R)^2 \mathbb{E}_{P_{z,\mu_{\text{L}},\mu_{\text{R}}}} (\hat{z} - z) \\
&< 4\sigma^2 + c\sigma^2 \log\log n.
\end{aligned}$$

Now, choosing $c = c_1/2$, then for sufficiently large n, the above inequality contradicts (27), which means that (28) cannot hold, thus establishing the desired conclusion. □

Lemma 7. *The norm $\|\cdot\|_{\text{grp}*}$ is a dual to $\|\cdot\|_{\text{grp}}$ with respect to the inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^{p \times n}$.*

Proof. To prove the lemma, it suffices to show that $\|M\|_{\text{grp}} = \sup_{\|R\|_{\text{grp}*} \le 1} \langle R, M \rangle$ for all $M \in \mathbb{R}^{p \times (n-1)}$. First, for any $M \in \mathbb{R}^{p \times (n-1)}$, let $M_{\mathcal{J}_g, t}$ be the tth column of $M_{\mathcal{J}_g}$. Define $\tilde{R} = \tilde{R}(M)$ such that

$$\tilde{R}_{\mathcal{J}_g, t} = \frac{p_g^{1/2} M_{\mathcal{J}_g, t}}{\max\{\|M_{\mathcal{J}_g, t}\|_2, 1\}}.$$

Then, $\|\tilde{R}\|_{\text{grp}*} \le \max_{g \in [G]} \max_{t \in [n-1]} p_g^{-1/2} p_g^{1/2} \frac{\|M_{\mathcal{J}_g, t}\|_2}{\|M_{\mathcal{J}_g, t}\|_2} = 1$. Hence,

$$\begin{aligned}
\sup_{\|R\|_{\text{grp}*} \le 1} \langle R, M \rangle \ge \langle \tilde{R}, M \rangle &= \sum_{g=1}^{G} \sum_{t=1}^{n-1} p_g^{1/2} \frac{\langle M_{\mathcal{J}_g, t}, M_{\mathcal{J}_g, t} \rangle}{\|M_{\mathcal{J}_g, t}\|_2} \\
&= \sum_{g=1}^{G} \sum_{t=1}^{n-1} p_g^{1/2} \|M_{\mathcal{J}_g, t}\|_2 = \|M\|_{\text{grp}}.
\end{aligned}$$

On the other hand, for any R such that $\|R\|_{\text{grp}*} \le 1$, we have $\|R_{\mathcal{J}_g,t}\|_2 \le p_g^{1/2}$ for all g and t. Consequently, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\langle R, M\rangle &= \sum_{g\in[G]}\sum_{t\in[n-1]}\langle R_{\mathcal{J}_g,t}, M_{\mathcal{J}_g,t}\rangle \le \sum_{g\in[G]}\sum_{t\in[n-1]}\|R_{\mathcal{J}_g,t}\|_2\|M_{\mathcal{J}_g,t}\|_2\\ &\le \sum_{g\in[G]}\sum_{t\in[n-1]} p_g^{1/2}\|M_{\mathcal{J}_g,t}\|_2 = \|M\|_{\text{grp}},\end{aligned}$$

thus establishing the result. □

Proposition 8. *Let* $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_{\text{F}} \le 1\}$. *For* $T \in \mathbb{R}^{p\times(n-1)}$, $\lambda > 0$, *we have*

$$\arg\max_{M\in\mathcal{S}}\Big\{\langle T, M\rangle - \lambda\|M\|_{\text{grp}}\Big\} = \frac{T - R^*}{\|T - R^*\|_{\text{F}}},$$

where R^* *satisfies* $R^*_{\mathcal{J}_g,t} = T_{\mathcal{J}_g,t}\min\big\{\frac{\lambda p_g^{1/2}}{\|T_{\mathcal{J}_g,t}\|_{\text{F}}}, 1\big\}$.

Proof. Define functions $h : \mathbb{R}^{p\times(n-1)} \times \mathbb{R}^{p\times(n-1)} \to \mathbb{R}$ and $f, g : \mathbb{R}^{p\times(n-1)} \to \mathbb{R}$ such that for $M, R \in \mathbb{R}^{p\times(n-1)}$, $h(M, R) = \langle T - \lambda R, M\rangle$ and $f(M) = \inf_{\|R\|_{\text{grp}*}\le 1} h(M, R)$ and $g(R) = \sup_{M\in\mathcal{S}} h(M, R)$. By (12) and Lemma 7, we have that

$$\begin{aligned}\langle T, M\rangle - \lambda\|M\|_{\text{grp}} &= \langle T, M\rangle - \lambda\sup_{\|R\|_{\text{grp}*}\le 1}\langle R, M\rangle\\ &= \inf_{\|R\|_{\text{grp}*}\le 1}\langle T - \lambda R, M\rangle = f(M).\end{aligned}$$

By the minimax equality theorem (Fan, 1953, Theorem 1), we obtain that

$$\sup_{M\in\mathcal{S}} f(M) = \sup_{M\in\mathcal{S}}\inf_{\|R\|_{\text{grp}*}\le 1} h(M, R) = \inf_{\|R\|_{\text{grp}*}\le 1}\sup_{M\in\mathcal{S}} h(M, R) = \inf_{\|R\|_{\text{grp}*}\le 1} g(R).$$

Observe that $g(R) = \|T - \lambda R\|_{\text{F}}$. To find the $R^* \in \arg\min_{\|R\|_{\text{grp}*}\le 1}\|T - \lambda R\|_{\text{F}}$, we consider the G groups individually. For each group g, and in the tth column, if $\|T_{\mathcal{J}_g,t}\|_2 \le \lambda p_g^{1/2}$, then $R^*_{\mathcal{J}_g,t} = T_{\mathcal{J}_g,t}/\lambda$; and if $\|T_{\mathcal{J}_g,t}\|_2 > \lambda p_g^{1/2}$, then $R^*_{\mathcal{J}_g,t} = p_g^{1/2}T_{\mathcal{J}_g,t}/\|T_{\mathcal{J}_g,t}\|_2$. Since the minimizer of $g(R)$ is unique, we have that

$$\arg\max_{M\in\mathcal{S}} f(M) = \arg\max_{M\in\mathcal{S}} h(M, R^*) = \frac{T - \lambda R^*}{\|T - \lambda R^*\|_{\text{F}}},$$

as desired. □

Lemma 9. *For any* $A, B \in \mathbb{R}^{p\times n}$, *we have* $\langle A, B\rangle \le \|A\|_{\text{grp}}\|B\|_{\text{grp}*}$.

Proof. By Cauchy–Schwarz inequality, we have that

$$\begin{aligned}\langle A, B\rangle &= \sum_{g,t}\langle A_{\mathcal{J}_g,t}, B_{\mathcal{J}_g,t}\rangle \le \sum_{g\in[G],t\in[n]}\|A_{\mathcal{J}_g,t}\|_{\text{F}}\|B_{\mathcal{J}_g,t}\|_{\text{F}}\\ &\le \Bigg(\sum_{g\in[G],t\in[n]} p_g^{1/2}\|A_{\mathcal{J}_g,t}\|_{\text{F}}\Bigg)\Bigg(\max_{g\in[G],t\in[n]} p_g^{-1/2}\|B_{\mathcal{J}_g,t}\|_{\text{F}}\Bigg) = \|A\|_{\text{grp}}\|B\|_{\text{grp}*}.\end{aligned}$$

as desired. □

Lemma 10. *Let $p_g = |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \leq C_1$. Then, for any $M \in \mathbb{R}^{p\times n}$, we have $\|M\|_{\mathrm{grp}} \leq (C_1 n \sum_g p_g)^{1/2} \|M\|_{\mathrm{F}}$.*

Proof. Define m with $m_{\mathcal{J}_g,t} = \|M_{\mathcal{J}_g,t}\|_{\mathrm{F}}$. Then by applying the Cauchy–Schwarz inequality twice, we have

$$\begin{aligned}\|M\|_{\mathrm{grp}} &= \sum_{g\in[G]} p_g^{1/2} \sum_{t=1}^{n} \|M_{\mathcal{J}_g,t}\|_2 \leq \sum_{g\in[G]} (np_g)^{1/2} \|M_{\mathcal{J}_g}\|_{\mathrm{F}} \\ &\leq \sqrt{n}\Big(\sum_{g\in[G]} p_g\Big)^{1/2} \Big(\sum_{g\in[G]} \|M_{\mathcal{J}_g}\|_{\mathrm{F}}^2\Big)^{1/2} \leq \Big(C_1 n \sum_{g\in[G]} p_g\Big)^{1/2} \|M\|_{\mathrm{F}},\end{aligned}$$

as desired. □

The following proposition establishes a sine angle loss upper bound for the (computationally infeasible) optimiser of (6). We see that the risk bound has essentially the same form as that given in Theorem 1.

Proposition 11. *For a given grouping $(\mathcal{J}_g)_{g\in[G]}$, let $p_* = \min_{g\in[G]} |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \leq C_1$. Let $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$ be a $p \times n$ data matrix, let θ be the vector of change and let $\hat{v} \in \arg\max_{\tilde{v}\in\mathbb{S}^{p-1}, \|\phi(\tilde{v})\|_0 \leq s} \|\tilde{v}^\top T\|_2$. Let $\lambda \geq B^{1/2}(1 + \sqrt{4\log(nG)/p_*})$. Then, with probability at least $1 - \frac{1}{nG}$ we have that*

$$\sin\angle(v, \hat{v}) \leq \frac{8\sqrt{2C_1}\lambda k^{1/2}}{n^{1/2}\tau\vartheta} \tag{30}$$

Proof. Let A, γ be defined as in the proof of Theorem 1. Let $u := \gamma/\|\gamma\|_2$ and $\hat{u} := T^\top\hat{v}/\|T^\top\hat{v}\|_2$. Then, by the basic inequality, we have that:

$$\langle \hat{v}\hat{u}^\top, T\rangle = \|T^\top\hat{v}\|_2 \geq \|T^\top v\|_2 \geq v^\top T u = \langle vu^\top, T\rangle.$$

Combining with Wang and Samworth (2018, Lemma 2), we have:

$$\begin{aligned}\|vu^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}}^2 &= \frac{2}{\|\theta\|_2\|\gamma\|_2}(\langle A - T, vu^\top - \hat{v}\hat{u}^\top\rangle + \langle T, vu^\top - \hat{v}\hat{u}^\top\rangle) \\ &\leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\langle A - T, vu^\top - \hat{v}\hat{u}^\top\rangle \\ &\leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\|A - T\|_{\mathrm{grp}^*}\|vu^\top - \hat{v}\hat{u}^\top\|_{\mathrm{grp}}\end{aligned}$$

Since $vu^\top - \hat{v}\hat{u}^\top$ has at most $2k$ rows with non-zero entries, By Lemmas 10 and 14, for the choice of λ in the proposition, we have with probability at least $1 - 1/(nG)$ that

$$\|uv^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}}^2 \leq \frac{2\sqrt{2}\lambda(C_1 nk)^{1/2}}{\|\theta\|_2\|\gamma\|_2}\|uv^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}}.$$

Consequently, by the same argument as in the proof pf Proposition 12 we have

$$\sin\angle(v,\hat{v}) \le \|vu^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}} \le \frac{2\sqrt{2}\lambda(C_1 nk)^{1/2}}{\|\theta\|_2\|\gamma\|_2} \le \frac{8\sqrt{2}\lambda(C_1 k)^{1/2}}{n^{1/2}\tau\vartheta},$$

as required. □

Proposition 12. *Let $p_g = |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \le C_1$. Let A be a rank one matrix with $A = \delta vu^\top$ for $\delta > 0$, $\|v\|_2 = \|u\|_2 = 1$ and $\sum_{g:v_{\mathcal{J}_g}\neq 0} p_g \le k$. Suppose $T \in \mathbb{R}^{p\times(n-1)}$ satisfies $\|T - A\|_{\mathrm{grp}*} \le \lambda$ for some $\lambda > 0$, and let $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_{\mathrm{F}} \le 1\}$. Then, for any*

$$\hat{M} \in \arg\max_{M\in\mathcal{S}} \big\{\langle T, M\rangle - \lambda\|M\|_{\mathrm{grp}}\big\},$$

we have

$$\|vu^\top - \hat{M}\|_{\mathrm{F}} \le \frac{4\lambda(C_1 nk)^{1/2}}{\delta},$$

and

$$\max\{\sin\angle(v,\hat{v}), \sin\angle(u,\hat{u})\} \le \frac{8\lambda(C_1 nk)^{1/2}}{\delta}.$$

Proof. Define $\mathcal{G}_0 = \{g : v_{\mathcal{J}_g} \neq 0\}$. Since $vu^\top \in \mathcal{S}$, from the basic inequality, we have

$$\langle T, vu^\top\rangle - \lambda\|vu^\top\|_{\mathrm{grp}} \le \langle T, \hat{M}\rangle - \lambda\|\hat{M}\|_{\mathrm{grp}}. \tag{31}$$

When $\|A-T\|_{\mathrm{grp}*} \le \lambda$, or equivalently, $p_g^{-1/2}\|A_{\mathcal{J}_g,t} - T_{\mathcal{J}_g,t}\|_2 \le \lambda$ for all $g \in [G]$ and $t \in [n-1]$, we have by Wang and Samworth (2018, Lemma 2) and (31) that

$$\begin{aligned}
\|vu^\top - \hat{M}\|_{\mathrm{F}}^2 &\le \frac{2}{\delta}\langle A, vu^\top - \hat{M}\rangle \le \frac{2}{\delta}\big(\langle T, vu^\top - \hat{M}\rangle + \langle A - T, vu^\top - \hat{M}\rangle\big) \\
&\le \frac{2\lambda}{\delta}\big(\|vu^\top\|_{\mathrm{grp}} - \|\hat{M}\|_{\mathrm{grp}} + \|vu^\top - \hat{M}\|_{\mathrm{grp}}\big) \\
&= \frac{4\lambda}{\delta}\sum_{g\in\mathcal{G}_0}\sum_{t\in[n-1]}\|(vu^\top - \hat{M})_{\mathcal{J}_g,t}\|_2 \le \frac{4\lambda(C_1 nk)^{1/2}}{\delta}\|vu^\top - \hat{M}\|_{\mathrm{F}},
\end{aligned}$$

where we used Lemma 9 in the penultimate inequality and Lemma 10 in the final bound. This proves the first claim of the proposition, and the second claim follows from the first by the same argument as used in Wang and Samworth (2018, online supplement (18) and (19)). □

Lemma 13. *Suppose $\Sigma \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix and let $E \sim N(0,\Sigma)$. Then we have for any $\delta > 0$ that*

$$\mathbb{P}\big(\|E\|^2 > \mathrm{tr}(\Sigma) + 2\|\Sigma\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + 2\|\Sigma\|_{\mathrm{op}}\log(1/\delta)\big) \le \delta.$$

Proof. Let $\Sigma = U^\top \Lambda U$ be the eigendecomposition of Σ, such that $U \in \mathbb{R}^{d\times d}$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1(\Sigma), \ldots, \lambda_d(\Sigma))$ is a diagonal matrix with eigenvalues of E on its diagonal. Hence, there exist $Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} N(0,1)$ such that $\|E\|_2^2 = \|UE\|_2^2 = \sum_{j=1}^d \lambda_j(\Sigma) Z_j^2$. Applying Laurent and Massart (2000, Lemma 1), we have with probability at least $1-\delta$ that

$$\begin{aligned}\|E\|_2^2 &\le \sum_{j=1}^d \lambda_j(\Sigma) + 2\Big(\sum_{j=1}^d \lambda_j^2(\Sigma)\Big)^{1/2}\sqrt{\log(1/\delta)} + 2\max_{j=1}^d \lambda_j(\Sigma)\log(1/\delta)\\ &\le \mathrm{tr}(\Sigma) + 2\|\Sigma\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + 2\|\Sigma\|_{\mathrm{op}}\log(1/\delta)\end{aligned}$$

as desired. □

Lemma 14. *Suppose $\Sigma \in \mathbb{R}^{p\times p}$ is a symmetric positive semidefinite matrix with $\|\Sigma\|_{\mathrm{op}} \le B$. Let $W = (W_1, \ldots, W_n)$ be an $p \times n$ random matrix with independent columns $W_t \sim N_p(0, \Sigma)$. Define $E := \mathcal{T}(W)$. Let $p_g = |\mathcal{J}_g|$ with $p_* = \min_{g\in[G]} p_g$. Then for any $\delta \in (0,1)$ and $\lambda = B^{1/2}\big(1 + \sqrt{2p_*^{-1}\log(1/\delta)}\big)$, we have that*

$$\mathbb{P}(\|E\|_{\mathrm{grp}*} > \lambda) \le (n-1)G\delta.$$

Proof. By the definition of the CUSUM transformation $\mathcal{T}$ in (5), we have that $E_{\mathcal{J}_g,t} \sim N(0, \Sigma_{\mathcal{J}_g,\mathcal{J}_g})$. By a union bound, we have

$$\begin{aligned}\mathbb{P}(\|E\|_{\mathrm{grp}*} > \lambda) &\le \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}(\|E_{\mathcal{J}_g,t}\|_2^2 > p_g\lambda^2)\\ &\le \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\bigg(\|E_{\mathcal{J}_g,t}\|_2^2 > Bp_g\bigg(1+\sqrt{\frac{2\log(1/\delta)}{p_g}}\bigg)^2\bigg)\\ &\le \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\bigg(\|E_{\mathcal{J}_g,t}\|_2^2 > B\Big(p_g + 2\sqrt{p_g\log(1/\delta)} + 2\log(1/\delta)\Big)\bigg)\\ &\le \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\big(\|E_{\mathcal{J}_g,t}\|_2^2 > \mathrm{tr}(\Sigma_{\mathcal{J}_g,\mathcal{J}_g}) + 2\|\Sigma_{\mathcal{J}_g,\mathcal{J}_g}\|_{\mathrm{F}}\sqrt{\log(1/\delta)}\\ &\qquad\qquad + 2\|\Sigma_{\mathcal{J}_g,\mathcal{J}_g}\|_{\mathrm{op}}\log(1/\delta)\big)\\ &\le (n-1)G\delta.\end{aligned}$$

as desired, where we used the fact that $\|\Sigma_{\mathcal{J}_g,\mathcal{J}_g}\|_{\mathrm{op}} \le \|\Sigma\|_{\mathrm{op}} \le B$ in the penultimate inequality and Lemma 13 in the final bound. □

Lemma 15. *Let $W = (W_1, \ldots, W_n)$ be a $p \times n$ random matrix with $W_i \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ and $E = \mathcal{T}(W) = (E_1, \ldots, E_{n-1})$. Suppose $\|\Sigma\|_{\mathrm{op}} \le B$ and that $\min(z, n-z) \ge n\tau$ and $|z-t| \le n\tau/2$. For a deterministic vector $v \in \mathbb{R}^p$ and any $\lambda_1 > 0$, there exists an event Ω_1 with probability at least $1 - 16e^{-\lambda_1^2/(4B)}\log n$ such that on this event, we have*

$$|v^\top E_z - v^\top E_t| \le 2\sqrt{2}\lambda_1\sqrt{\frac{z-t}{n\tau}} + 8\lambda_1\frac{z-t}{n\tau}.$$

Proof. Define event

$$\Omega_1 := \left\{ \left| \sum_{r=1}^{s} v^\top W_r - \sum_{r=1}^{t} v^\top W_r \right| \le \lambda_1 \sqrt{|s-t|}, \text{ for } 0 \le t \le n \text{ and } s \in \{0, z, n\} \right\}.$$

Since $v^\top W_1, \ldots, v^\top W_n \overset{\text{iid}}{\sim} N(0, v^\top \Sigma v)$, with $v^\top \Sigma v \le B$, by Wang and Samworth (2018, Lemma 5), for any $u \ge 0$, and $m \in \mathbb{N}$, we have

$$\mathbb{P}\left(\max_{1 \le t \le m} \left| \frac{1}{\sqrt{t}} \sum_{r=1}^{t} v^\top W_r \right| \ge u B^{1/2} \right) \le 4 e^{-u^2/4} \log m. \tag{32}$$

Applying the above bound four times, we have

$$\mathbb{P}(\Omega_1^{\mathrm{c}}) \le 4 e^{-\lambda_1^2/(4B)} \{ 2 \log n + \log z + \log(n-z) \} \le 16 e^{-\lambda_1^2/(4B)} \log n.$$

It hence suffices to show that on Ω_1, the desired inequality holds. By symmetry, we may assume without loss of generality that $t < z$. From the definition of the CUSUM transformation in (5), we have

$$\begin{aligned} v^\top E_z - v^\top E_t &= \sqrt{\frac{n}{z(n-z)}} \left(\frac{z}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{z} v^\top W_r \right) \\ &\qquad - \sqrt{\frac{n}{t(n-t)}} \left(\frac{t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r \right) \\ &= \sqrt{\frac{n}{z(n-z)}} \left(\frac{z-t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=t+1}^{z} v^\top W_r \right) \\ &\qquad + \left(\sqrt{\frac{n}{z(n-z)}} - \sqrt{\frac{n}{t(n-t)}} \right) \left(\frac{t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r \right). \end{aligned} \tag{33}$$

On the event Ω_1,

$$\begin{aligned} \left| \frac{z-t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=t+1}^{z} v^\top W_r \right| &\le \frac{z-t}{n} \left| \sum_{r=1}^{n} v^\top W_r \right| + \left| \sum_{r=t+1}^{z} v^\top W_r \right| \\ &\le \frac{z-t}{n} \lambda_1 \sqrt{n} + \lambda_1 \sqrt{z-t} \le 2 \lambda_1 \sqrt{z-t} \end{aligned} \tag{34}$$

Similarly, we have on Ω_1 that

$$\left| \frac{t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r \right| \le \frac{\lambda_1 t}{\sqrt{n}} + \lambda_1 \sqrt{t} \le 2 \lambda_1 \sqrt{t}.$$

Noticing that $\frac{t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r = \frac{n-t}{n} \sum_{r=1}^{n} v^\top W_r - \sum_{r=t+1}^{n} v^\top W_r$, we can similarly bound the left-hand side above by $2\lambda_1 \sqrt{n-t}$. Therefore, on

Ω_1, we have

$$\left|\frac{t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\right| \le 2\lambda_1 \min\{\sqrt{t}, \sqrt{n-t}\}$$
$$\le 2\lambda_1 \min\left\{\sqrt{z}, \sqrt{n-z+\frac{n\tau}{2}}\right\}. \tag{35}$$

By the mean value theorem, there exists $\xi \in [t, z]$ such that

$$\left|\sqrt{\frac{n}{z(n-z)}} - \sqrt{\frac{n}{t(n-t)}}\right| \le \frac{z-t}{2}\left(\frac{n}{\xi(n-\xi)}\right)^{3/2}$$
$$\le \frac{\sqrt{2}(z-t)}{\min\{(z-n\tau/2)^{3/2}, (n-z)^{3/2}\}}. \tag{36}$$

Combining (33), (34), (35) and (36), we have on Ω_1 that

$$\left|v^\top E_z - v^\top E_t\right| \le 2\lambda_1\sqrt{\frac{n(z-t)}{z(n-z)}} + \frac{2^{3/2}\lambda_1(z-t)\min\left\{z^{1/2}, (n-z+n\tau/2)^{1/2}\right\}}{\min\{(z-n\tau/2)^{3/2}, (n-z)^{3/2}\}}$$
$$\le 2\sqrt{2}\lambda_1\sqrt{\frac{z-t}{n\tau}} + 8\lambda_1\frac{z-t}{n\tau},$$

as desired. □

Lemma 16. *Suppose $\mu = (\mu_1, \ldots, \mu_n)$ has a single change point at z, in the sense that $\mu_1 = \cdots = \mu_z = \mu^{(1)}$ and $\mu_{z+1} = \cdots = \mu_n = \mu^{(2)}$. Let $A = \mathcal{T}(\mu) = (A_1, \ldots, A_n)$. Define $\theta = \mu^{(1)} - \mu^{(2)}$. Then for any $v \in \mathbb{R}^p$, and $|z-t| \le n\tau/2$, we have*

$$\left|v^\top A_z - v^\top A_t\right| \ge \frac{2}{3\sqrt{6}}\frac{|z-t|}{\sqrt{n\tau}}(v^\top\theta).$$

Proof. Observe that A is a rank one matrix given by (14). Hence, $v^\top A = (v^\top\theta)\gamma^\top$. The desired result is then a consequence of Wang and Samworth (2018, Lemma 7). □

Appendix B: Extensions to sub-Gaussian distributions

In the previous sections, we assumed that $X_i = \mu_i + W_i$, for $W_1, \ldots, W_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$. In this session, we discuss how the previous results can be generalised to settings where $W_1, \ldots, W_n$ are independent sub-Gaussian random vectors. Adpoting notation from Zhu, Wang and Samworth (2022), for any random vector U in $\mathbb{R}^p$, we write

$$\|U\|_{\psi_2} := \sup_{w\in\mathcal{S}^{p-1}}\sup_{q\in\mathbb{N}}\frac{\mathbb{E}(|w^\top U|^q)^{1/q}}{\sqrt{q}},$$
$$\|U\|_{\psi_2^*} := \sup_{w\in\mathcal{S}^{p-1}}\frac{\|w^\top U\|_{\psi_2}}{(w^\top\mathrm{Var}(U)w)^{1/2}} = \|\mathrm{Var}^{-1/2}(U)U\|_{\psi_2}.$$

For sub-Gaussian data, Lemma 17 can be used in place of Lemma 14 to derive the equivalent result of Theorem 1 for the sub-Gaussian data.

Lemma 17. *Let $W = (W_1, \ldots, W_n)$ be a $p \times n$ random matrix with independent columns W_t satisfying $\|W_t\|_{\psi_2^*} \leq L$ and $\|\mathrm{Var}(W_t)\|_{\mathrm{op}} \leq B$ for $t \in [n-1]$. Define $E := \mathcal{T}(W)$. Let $p_g = |\mathcal{J}_g|$ with $p_* = \min_{g\in[G]} p_g$. There exists a universal constant $C > 0$ such that for any $\delta \in (0,1)$, we have*

$$\mathbb{P}\left\{\|E\|_{\mathrm{grp}*} > CLB^{1/2}\left(1 + \sqrt{\frac{\log(nG/\delta)}{p_*}}\right)\right\} \leq \delta.$$

Proof. By the definition of the CUSUM transformation $\mathcal{T}$ in (5), we can write E_t as $E_t = \sum_{s\in[n]} a_s W_s$ for a contrast vector $a = (a_1, \ldots, a_n)^\top$ such that $\|a\|_2 = 1$. For each $t \in [n]$, Since $\|W_t\|_{\psi_2^*} \leq L$, we have for any $v \in \mathcal{S}^{p-1}$ that $\|v^\top W_s / \{v^\top \mathrm{Var}(W_t) v\}^{1/2}\|_{\psi_2} \leq L$. Therefore, by Vershynin (2012, Proposition 5.10), there exists a constant $C_1 > 0$ such that for every $t \in [n-1]$ we have

$$\|E_t\|_{\psi_2^*} = \sup_{v\in\mathcal{S}^{p-1}} \frac{\|v^\top E_t\|_{\psi_2}}{(v^\top \mathrm{Var}(W_t) v)^{1/2}} = \sup_{v\in\mathcal{S}^{p-1}} \left\| \frac{\sum_{s=1}^n a_s v^\top W_s}{(v^\top \mathrm{Var}(W_t) v)^{1/2}} \right\|_{\psi_2} \leq C_1 L.$$

Then, we can bound $\|E_t\|_{\psi_2}$ by:

$$\|E_t\|_{\psi_2} \leq \|E_t\|_{\psi_2^*} \|\Sigma\|_{\mathrm{op}}^{1/2} \leq C_1 L B^{1/2}.$$

Define $\mathcal{S}_g^{p-1} := \{v \in \mathcal{S}^{p-1} : \mathrm{supp}(v) \subseteq \mathcal{J}_g\}$ and let $\mathcal{N}_g \subseteq \mathcal{S}_g^{p-1}$ be a 1/2-net of the set $\mathcal{S}_g^{p-1}$. By Vershynin (2012, Lemma 5.2), we can choose $\mathcal{N}_g$ such that $|\mathcal{N}_g| \leq 5^{p_g}$. Obseve that

$$\begin{aligned}\|E_{\mathcal{J}_g,t}\|_2 &= \sup_{v\in\mathcal{S}_g^{p-1}} v^\top E_t \leq \sup_{v\in\mathcal{N}_g} v^\top E_t + \sup_{u:\|u\|_2\leq 1/2, \mathrm{supp}(u)\subseteq\mathcal{J}_g} |u^\top E_t| \\ &= \sup_{v\in\mathcal{N}_g} v^\top E_t + \frac{1}{2}\|E_{\mathcal{J}_g,t}\|_2 \leq 2 \sup_{v\in\mathcal{N}_g} v^\top E_t.\end{aligned}$$

Hence, by a union bound and a tail bound of sub-Gaussian random variables, we have we have for some universal constant $C_2 > 0$ that

$$\mathbb{P}(\|E_{\mathcal{J}_g,t}\|_2 \geq x) \leq \mathbb{P}\left(\sup_{v\in\mathcal{N}_g} v^\top E_t \geq \frac{x}{2}\right) \leq 5^{p_g} e^{-x^2/(C_2^2 L^2 B)}.$$

By another union bound, we have

$$\begin{aligned}&\mathbb{P}\left\{\|E\|_{\mathrm{grp}*} > 2C_2 LB^{1/2}\left(1 + \sqrt{\frac{\log(nG/\delta)}{p_*}}\right)\right\} \\ &\leq \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\left(\|E_{\mathcal{J}_g,t}\|_2 > C_2 LB^{1/2}\sqrt{2p_g + \log(nG/\delta)}\right) \\ &\leq \sum_{g\in[G]} (n-1) 5^{p_g} e^{-2p_g - \log(nG/\delta)} \leq \delta,\end{aligned}$$

as desired. □

Lemma 19 below can be used in place of Lemma 15 to establish the equivalent of Theorem 3 for the sub-Gaussian data. To prove Lemma 19, we first establish Lemma 18.

Lemma 18. *Let $W_1, \ldots, W_n$ be independent centered sub-Gaussian random variables with $\max_t \|W_t\|_{\psi_2} \le K$ for $t \in [n]$. Define $Z_t := t^{-1/2} \sum_{r=1}^t W_r$. Then for $n \ge 5$ and $u \ge 0$, we have for some universal constant $C > 0$ that*

$$\mathbb{P}(\max_{1 \le t \le n} Z_t \ge u) \le 2e^{-u^2/(CK^2)} \log n.$$

Proof. Define $S_t := \sum_{r=1}^t W_r$. Then, $(S_t)_t$ is a martingle and $(e^{S_t})_t$ is a non-negative sub-martingle. Then, by a union bound, we have

$$\mathbb{P}\Big(\max_{1 \le t \le n} Z_t \ge u\Big) \le \sum_{j=1}^{\lceil \log_2(n+1) \rceil} \mathbb{P}\Big(\max_{2^{j-1} \le t < 2^j} Z_t \ge u\Big)$$

Then by Doob's martingle inequality and Vershynin (2012, Lemma 5.9), we have for some universal constant $C_1 > 0$ that

$$\begin{aligned}
\mathbb{P}\Big(\max_{2^{j-1} \le t < 2^j} Z_t \ge u\Big) &\le \sum_{j=1}^{\lceil \log_2(n+1) \rceil} \inf_{\lambda > 0} \mathbb{P}\Big(\max_{2^{j-1} \le t < 2^j} e^{\lambda S_t} \ge e^{2^{(j-1)/2} \lambda u}\Big) \\
&\le \sum_{j=1}^{\lceil \log_2(n+1) \rceil} \inf_{\lambda > 0} \mathbb{E} e^{\lambda S_{2^j}} e^{-2^{(j-1)/2} \lambda u} \\
&\le \sum_{j=1}^{\lceil \log_2(n+1) \rceil} \inf_{\lambda > 0} e^{C_1 \lambda^2 2^{j-1} K^2} e^{-2^{(j-1)/2} \lambda u} \\
&= \sum_{j=1}^{\lceil \log_2(n+1) \rceil} e^{-u^2/(4CK^2)} \le 2e^{-u^2/(4C_1K^2)} \log n,
\end{aligned}$$

where in the final step, we used the fact that $n \ge 5$. The desired result follows by taking $C = 4C_1$. □

Lemma 19. *Let $W = (W_1, \ldots, W_n)$ be a $p \times n$ random matrix with columns satisfying $\max_t \|W_t\|_{\psi_2^*} \le L$ and $E = \mathcal{T}(W) = (E_1, \ldots, E_{n-1})$. Suppose $\min(z, n - z) \ge n\tau$ and $|z - t| \le n\tau/2$. For a deterministic vector v and $\lambda_1 = L\sqrt{CB \log n}$, we have with probability at least $1 - \frac{16 \log n}{n}$ that*

$$|v^\top E_z - v^\top E_t| \le 2\sqrt{2}\lambda_1 \sqrt{\frac{z - t}{n\tau}} + 8\lambda_1 \frac{z - t}{n\tau}$$

Proof. By a similar argument as in the proof of Lemma 17, we have for all $r \in [n]$ and $v \in \mathcal{S}^{p-1}$ that $\|v^\top W_r\|_{\psi_2} \le LB^{1/2}$. Define event

$$\Omega_1 := \Big\{ \Big| \sum_{r=1}^s v^\top W_r - \sum_{r=1}^t v^\top W_r \Big| \le \lambda_1 \sqrt{|s - t|}, \text{ for } 0 \le t \le n \text{ and } s \in \{0, z, n\} \Big\}.$$

Then, by Lemma 18, for any $u \geq 0$, and $m \in \mathbb{N}$, we have

$$\mathbb{P}\left(\max_{1 \leq t \leq m} \left| \frac{1}{\sqrt{t}} \sum_{r=1}^{t} v^\top W_r \right| \geq \lambda_1 \right) \leq 4e^{-\lambda_1^2/(CL^2B)} \log m.$$

Applying the above bound four times, we have

$$\begin{aligned} \mathbb{P}(\Omega_1^c) &\leq 4e^{-\lambda_1^2/(CL^2B)} \{2 \log n + \log z + \log(n-z)\} \\ &\leq 16 e^{-\lambda_1^2/(CL^2B)} \log n \leq \frac{16 \log n}{n}. \end{aligned}$$

It hence suffices to show that on Ω_1, the desired inequality holds. This deterministic calculation follows verbatim from the proof of Lemma 15. □

Appendix C: Extensions to temporal dependence

In this section, we consider the case when the columns of X are not independent. We assume that $W_1, \ldots, W_n$ are stationary and let $K(u) = \mathrm{Cov}(W_t, W_{t+u})$. We further assume that the dependence is short-ranged in the sense that:

$$\left\| \sum_{u=0}^{n-1} K(u) \right\|_{\mathrm{op}} \leq B^*. \tag{37}$$

The oracle projection direction does not change in this case, the following Lemma can be used in place of Lemma 14 to establish the equivalent result of Theorem 1 for data with short-ranged time-dependence.

Lemma 20. *Suppose $\Sigma \in \mathbb{R}^{p \times p}$ is a symmetric positive semidefinite matrix with $\|\Sigma\|_{\mathrm{op}} \leq B$. Let $W = (W_1, \ldots, W_n)$ be an $p \times n$ random matrix with dependent columns $W_t \sim N_p(0, \Sigma)$ satisfying equation (37). Define $E := \mathcal{T}(W)$. Let $p_g = |\mathcal{J}_g|$ with $p_* = \min_{g \in [G]} p_g$. Then for any $\delta \in (0,1)$ and $\lambda = \sqrt{2B^*}\big(1 + \sqrt{2p_*^{-1} \log(1/\delta)}\big)$, we have that*

$$\mathbb{P}(\|E\|_{\mathrm{grp}*} > \lambda) \leq (n-1)G\delta.$$

Proof. Fix $t \in [n-1]$ and define $\kappa = (\kappa_1, \ldots, \kappa_n)^\top \in \mathbb{R}^n$ by $\kappa_r = -\sqrt{\frac{n-t}{nt}} \mathbb{1}_{\{r \leq t\}} + \sqrt{\frac{t}{n(n-t)}} \mathbb{1}_{\{r > t\}}$ (for simplicity, we have suppressed the t dependence in the definition of κ). Then we have $E_t = \sum_{r=1}^n \kappa_r W_r \sim N(0, \Sigma^*)$ for some positive semidefinite matrix $\Sigma^* \in \mathbb{R}^{p \times p}$. For any $v \in \mathcal{S}^{p-1}$, we have

$$\begin{aligned} v^\top \Sigma^* v = \mathrm{Var}(v^\top E_t) &= \sum_{r_1=1}^n \sum_{r_2=1}^n \kappa_{r_1} \kappa_{r_2} v^\top K(|r_2 - r_1|) v \\ &\leq 2 \sum_{u=0}^{n-1} v^\top K(u) v \sum_{r=1}^{n-u} \kappa_r \kappa_{r+u} \\ &\leq 2 \sum_{u=0}^{n-1} v^\top K(u) v \left\{ \frac{(n-t)(t-u)_+}{nt} + \frac{t(n-t-u)_+}{n(n-t)} \right\} \leq 2B^*. \end{aligned}$$

Consequently, we have $\|\Sigma^*\|_{\text{op}} \leq 2B^*$. Then, following the proof of Lemma 14 and B with $2B^*$, we can obtaine the desired result. □

References

Aston, J. A. D. and Kirch, C. (2012) Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.*, **6**, 1906–1948.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2019) Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. Roy. Statist. Soc., Ser. B*, **81**, 649–672.

Cai, T. T., Zhang, A. and Zhou, Y. (2019) Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *arXiv preprint*, arxiv:1909.09851.

Cho, H. (2016) Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.*, **10**, 2000-2038.

Cho, H. and Fryzlewicz, P. (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B*, **77**, 475–507.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, **7**, 1–46.

Enikeeva, F. and Harchaoui, Z. (2019) High-dimensional change-point detection under sparse alternatives. *Ann. Statist.*, **47**, 2051–2079.

Fan, K. (1953) Minimax theorems. *Proc. Natl. Acad. Sci. USA*, **39**, 42–47.

Frank, M. and Wolfe, P. (1956) An algorithm for quadratic programming. *Naval Res. Logist.*, **3**, 95—110.

Frick, K., Munk, A. and Sieling, H. (2014) Multiscale change-point inference. *J. Roy. Statist. Soc., Ser. B*, **76**, 495–580.

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

Gao, C., Han, F., Zhang, C. H. (2020) On estimation of isotonic piecewise constant signals. *Ann. Statist.*, **48**, 629–654.

Hanlon, M. and Anderson, R. (2009) Real-time gait event detection using wearable sensors. *Gait & Posture*, **30**, 523–527.

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

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classification*, **2**, 193—218.

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

Killick, R., Fearnhead, P. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *J. Amer. Stat. Assoc.*, **107**, 1590–1598.

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

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

Massart, P. (2007) *Concentration Inequalities and Model Selection*, Springer, Berlin.

Peng, T., Leckie, C. and Ramamohanarao, K. (2004) Proactively detecting distributed denial ofservice attacks using source IP address monitoring. In Mitrou, N., Kontovasilis, K., Rouskas, G. N., Iliadis, I. and Merakos, L. eds, *Networking 2004*, pp. 771–782. Springer-Verlag, Berlin.

Pilliat, E., Carpentier, A. and Verzelen, N. (2020) Optimal multiple change-point detection for high-dimensional data

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

Simon, N, Friedman, J, Hastie, T and Tibshirani, R (2013) A sparse-group lasso. *J. Comput. Graph. Statist.*, **22**, 231–245.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.) *Compressed Sensing, Theory and Applications*. Cambridge University Press, Cambridge. 210–268.

Wang, H and Leng, C (2008) A note on adaptive group lasso. *Comput. Statist. Data Anal.* **52(12)**, 5277–5286.

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

Yu, B. (1997) Assouad, Fano and Le Cam. In Pollard, D., Torgersen, E. and Yang G. L. (Eds.) *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, 423–435. Springer, New York.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc., Ser. B*, **68**, 49–67.

Zhu, Z., Wang, T. and Samworth, R. J. (2022) High-dimensional principal component analysis with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B*, to appear.