Journal of Machine Learning Research 26 (2025) 1-57 Submitted 7/24; Published 8/25

Reliever: Relieving the Burden of Costly Model Fits for Changepoint Detection

Chengde Qian QIANCHD@GMAIL.COM
School of Mathematical Sciences
Shanghai Jiao Tong University
Shanghai 200240, China

Guanghui Wang GHWANG.NK@GMAIL.COM
School of Statistics and Data Science, LPMC, KLMDASR, and LEBPS
Nankai University
Tianjin 300071, China

Changliang Zou ZOUCL@NANKAI.EDU.CN
NITFID, School of Statistics and Data Science, LPMC, KLMDASR, and LEBPS
Nankai University
Tianjin 300071, China

Editor: Ji Zhu

Abstract

Changepoint detection typically relies on a grid-search strategy for optimal data segmentation. When model fitting itself is expensive, repeatedly fitting a model on every candidate segment dominates the computation. Existing approaches mitigate this by pruning the grid, thus reducing the number of segments (and model fits). We propose Reliever, which instead cuts the number of model fits directly and nests seamlessly within standard grid-search routines. Reliever fits a small, deterministic collection of proxy models and reuses them wherever they apply, making it compatible with a wide range of existing algorithms. For high-dimensional regression with changepoints, coupling Reliever with an optimal grid-search method yields changepoint and coefficient estimators that are rate-optimal up to a logarithmic factor. Extensive numerical experiments demonstrate that Reliever rapidly and accurately detects changepoints across a wide range of high-dimensional and nonparametric models.

Keywords: Binary segmentation; Grid search; High-dimensional regression; Multiple changepoint detection; Optimal partitioning.

1. Introduction

Changepoint detection serves to identify changes in statistical properties such as mean, variance, slope, or distribution within ordered observations. This technique finds applications in diverse domains including time series analysis, signal processing, finance, neuroscience, and environmental monitoring.

To identify the number and locations of changepoints, a common method is to conduct a *grid search* to optimize data segmentation by minimizing (or maximizing) a specific criterion. This criterion often integrates a sum of segment-wise losses (or gains, respectively) with a penalty for excessive segmentation. Grid-search algorithms are broadly categorized into *optimal* and *greedy* strategies. Optimal strategies use dynamic programming (Auger and Lawrence, 1989; Jackson et al., 2005; Killick et al., 2012) to find the global minimum, while greedy strategies, such as binary segmentation (Fryzlewicz, 2014; Baranowski et al., 2019; Kovács et al., 2022) and moving windows (Hao et al., 2013; Eichinger and Kirch, 2018), iteratively approximate this minimum. These algorithms necessitate repeatedly fitting models and evaluating loss functions across numerous data segments. Table 1 outlines the computational complexity of the grid-search step *in isolation*—that is, it treats the required model fits and loss values as if they were already available—so that one can compare how each algorithm scales with the sample size n. These algorithms include segment neighborhood (SN, Auger and Lawrence, 1989), optimal partitioning (OP, Jackson et al., 2005), pruned exact linear time (PELT, Killick et al., 2012), wild binary segmentation (WBS, Fryzlewicz, 2014), and seeded binary segmentation (SeedBS, Kovács et al., 2022). For an extensive review of grid-search algorithms, please refer to Cho and Kirch (2021).

Table 1: Computational complexity of grid-search algorithms in isolation and total model-fitting operations, comparing the original implementations and the proposed Reliever implementations. The notation a_n denotes the complexity of fitting a single model on an interval of length n. The set $\mathcal{R}$ is the pre-specified, deterministic collection of intervals on which Reliever actually fits models, with its cardinality $|\mathcal{R}| = O(n)$; see Definition 1.

	Optimal			Greedy	
Grid-search algorithm	SN	OP	PELT[†]	WBS	SeedBS
Complexity of the grid-search step (in isolation)					
	$O(Kn^2)^{\ddagger}$	$O(n^2)$	$O(n)$	$O(Mn)^{\S}$	$O(n\log n)$
Total model-fitting operations					
Original	$O(n^2 a_n)$	$O(n^2 a_n)$	$O(na_n)$	$O(Mna_n)$	$O\big(n(\log n)a_n\big)$
Reliever	$O(\lvert\mathcal{R}\rvert a_n)$	$O(\lvert\mathcal{R}\rvert a_n)$	$O(\lvert\mathcal{R}\rvert a_n)$	$O(\lvert\mathcal{R}\rvert a_n)$	$O(\lvert\mathcal{R}\rvert a_n)$

[†] For cases when the pruning condition is met (Killick et al., 2012, Eq. (4)); if pruning fails, PELT reduces to OP.

[‡] K: user-specified upper bound on the number of changepoints.

[§] M: number of random intervals in WBS.

To evaluate the loss on any candidate interval $I \subset (0, n]$ with integer endpoints, we must first fit a model $\widehat{\mathcal{M}}_I$ on that segment. Across the set of candidate intervals determined by the grid-search algorithm, the resulting sequence of model fits $\{\widehat{\mathcal{M}}_I\}$ often dominates the runtime in modern changepoint procedures, far outweighing both the associated loss evaluations $\{\mathcal{L}(I; \widehat{\mathcal{M}}_I)\}$ and the modest overhead of iterating through the interval grid. For instance, consider high-dimensional linear models with changepoints, estimated using the lasso (Lee et al., 2016; Leonardi and Bühlmann, 2016; Kaul et al., 2019b; Wang et al., 2021b; Xu et al., 2024). Fitting the lasso on an interval of length n by coordinate descent requires $O(np)$ operations per iteration, so the number of variables p directly drives runtime. If the penalty parameter is selected through cross-validation, the cost of a single model fit increases multiplicatively. Additionally, unlike classical mean-change models—where the sample mean can be updated incrementally (Auger and Lawrence, 1989)—high-dimensional fits cannot be adjusted cheaply when observations are added or removed. Consequently, the sequence of model fits dominates the overall complexity; see Table 1. Similar computational bottlenecks arise in changepoint models that incorporate graphical structures (Londschien et al., 2021), vector autoregressive dynamics (Safikhani and Shojaie, 2022; Bai et al., 2023), network topologies (Wang et al., 2021a), nonparametric frameworks (Zou et al., 2014; Jiang et al., 2022; Chen and Chu, 2023), and mechanisms for handling missing data (Follain et al., 2022).

1.1 Our Idea

Our approach—*Reliever*—operates as follows. For each candidate interval $I \subset (0, n]$, a standard grid-search algorithm $\mathcal{A}$ involves fitting an interval-specific model $\widehat{M}_I$ and evaluating the loss $\mathcal{L}(I; \widehat{M}_I)$. Reliever replaces this costly step with a proxy fit: we pair I with an interval R_I, chosen from a pre-specified deterministic collection $\mathcal{R}$ (see Definition 1). We fit the model $\widehat{M}_{R_I}$ on R_I and evaluate the loss on the target interval I via $\mathcal{L}(I; \widehat{M}_{R_I})$. This substitution continues until the algorithm $\mathcal{A}$ has visited every candidate interval. Figure 1 illustrates the procedure.

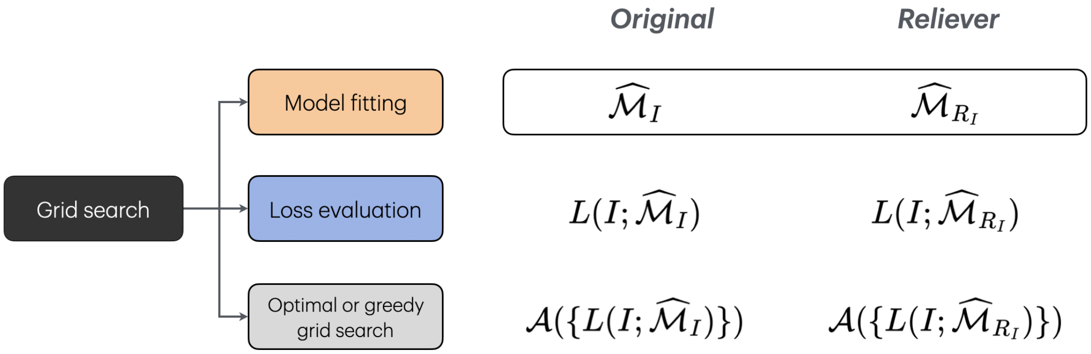

Original
Reliever
Grid search
Model fitting
Loss evaluation
Optimal or greedy grid search
$\widehat{\mathcal{M}}_I$
$\widehat{\mathcal{M}}_{R_I}$
$L(I; \widehat{\mathcal{M}}_I)$
$L(I; \widehat{\mathcal{M}}_{R_I})$
$\mathcal{A}(\{L(I; \widehat{\mathcal{M}}_I)\})$
$\mathcal{A}(\{L(I; \widehat{\mathcal{M}}_{R_I})\})$

Figure 1: Workflow comparison between a standard grid-search algorithm and the same algorithm equipped with Reliever.

The collection $\mathcal{R}$ is intentionally small—$|\mathcal{R}| = O(n)$ in our construction—so only $O(|\mathcal{R}|)$ actual model fits are required overall. Consequently, the total fitting cost drops to at most $O(na_n)$, where a_n denotes the operations needed to fit a single model on an interval of length n; see Table 1. In practice the cost is often lower because the algorithm $\mathcal{A}$ may visit only a subset of intervals in $\mathcal{R}$. Besides controlling the size of $\mathcal{R}$, each interval R_I is selected such that $R_I \subset I$ and the reminder $I \setminus R_I$ is short. With this design, replacing the original loss sequence $\{\mathcal{L}(I; \widehat{\mathcal{M}}_I)\}$ by its proxy counterpart $\{\mathcal{L}(I; \widehat{\mathcal{M}}_{R_I})\}$ in the grid-search algorithm would preserve changepoint-detection accuracy, regardless of how many changepoints lie within any individual interval I.

To illustrate the benefits of Reliever, we consider a high-dimensional linear model with multiple changepoints (see Section 4.1), using $n = 600$ observations and $p = 100$ variables. Figure 2(a) compares the average computational time spent on model fits (including loss evaluations) with and without Reliever, alongside the average time spent solely on the grid search for each algorithm. Clearly, the primary computational burden arises from model fits, and employing Reliever substantially reduces this burden. Figure 2(b) further shows the reduction in the average number of model fits required along the search path. Finally, Figure 2(c) presents a boxplot of changepoint detection error, measured by the Hausdorff distance (see Section 4), confirming that Reliever significantly reduces computational cost without sacrificing detection accuracy.

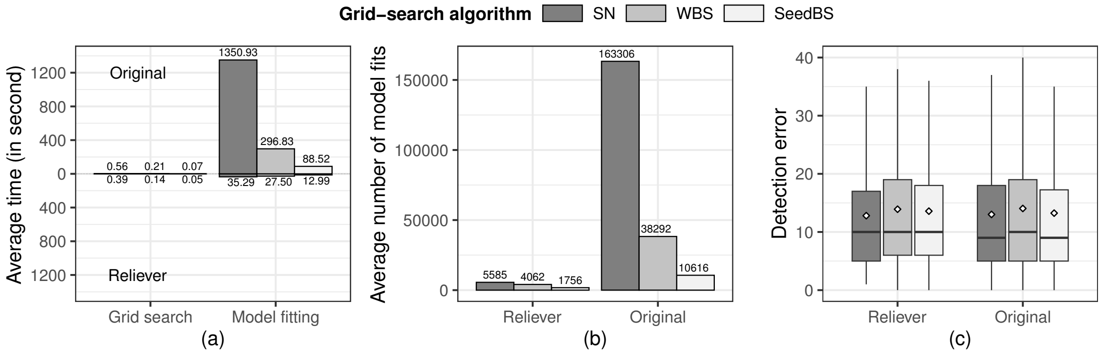


Figure 2: Runtime savings and accuracy retention provided by Reliever in a high-dimensional linear model. (a) Average time spent on grid search in isolation and on model fits for the SN, WBS, and SeedBS algorithms, with and without Reliever. (b) Average number of model fits executed along each search path. (c) Changepoint detection error (Hausdorff distance); circles mark mean values.

1.2 Our Contributions

We introduce Reliever, a highly flexible framework that speeds up changepoint detection whenever model fitting is the main cost. While earlier work cuts runtime by reducing the number of intervals that a grid-search algorithm visits, Reliever follows a complementary route: it fits models on only a pre-specified deterministic set of $O(n)$ intervals and reuses those fits wherever possible. Because the grid-search logic itself is left untouched, Reliever can be dropped straight into common optimal and greedy algorithms like SN, OP, PELT, WBS and SeedBS.

This sharp cut in model fits greatly shortens running time in both high-dimensional and nonparametric settings, while maintaining detection accuracy. In the context of high-dimensional linear models with multiple changepoints—a topic that has garnered significant research interest—we demonstrate that Reliever combined with the OP algorithm (for example, Leonardi and Bühlmann, 2016), produces estimators for both changepoints and corresponding regression coefficients that are rate-optimal, up to a logarithmic factor.

1.3 Related Works

Comparison with grid-pruning methods. The collection of pre-specified and deterministic intervals used by Reliever resembles the *seeded intervals* in SeedBS (Kovács et al., 2022). These two acceleration ideas, however, serve different aims. Kovács et al. (2022) refined WBS by replacing its random intervals with seeded intervals, targeting near-linear scaling of the grid search relative to the sample size (as outlined in Table 1). A similar approach, involving the use of deterministic intervals for constructing scan statistics, is explored in Chan and Walther (2013). In contrast, our approach retains existing grid-search schemes but re-uses a proxy model fitted on one deterministic interval whenever that model is relevant to another data segment. This strategy allows Reliever to adapt to any grid-search algorithm, including SeedBS and WBS. Because the two kinds of deterministic intervals target different goals, their construction principles also differ; see Remark 3.

Comparison with two-step methods. Our method's strategy to reduce intensive model fitting relates to two-step procedures that use a preliminary set of changepoint candidates. In the context of high-dimensional linear models with a single changepoint, Kaul et al. (2019b) proposed an approach involving initial fitting of two regression models, one for data before and another after an initial changepoint estimator, followed by searching for the best split to minimize training error. To achieve nearly-optimal convergence rates for the resulting estimator, the initial estimator must be consistent. For multiple changepoint scenarios, Kaul et al. (2019a) extended this approach by incorporating multiple initial candidates and employing a simulated annealing algorithm to allocate available model fits. This method presupposes proximities of all true changepoints to some of the initial candidates. Cho and Owens (2024) uses moving window on a coarse grid to scan for initial changepoint candidates and then refine their locations, whereas Li et al. (2023) employs dynamic programming on a coarse grid to obtain initial candidates before refinement. In the context of univariate mean change models, Lu et al. (2017) introduced a method that leverages a sparse subsample to derive pilot changepoint estimators; for these pilot estimators to yield optimal changepoint estimators, they must accurately reflect both the number and locations of the changepoints. In contrast, our Reliever framework does not rely on consistent initial estimators. It offers broad applicability and can be integrated as a foundational component in a variety of existing changepoint detection algorithms.

1.4 Notation

The L_q norm of a vector $\mathbf{z} \in \mathbb{R}^p$ is defined by $\|\mathbf{z}\|_q = (\sum_{j=1}^p z_j^q)^{1/q}$. For a p-by-p positive semi-definite matrix $\mathbf{A}$, we denote $\|\mathbf{z}\|_{\mathbf{A}} = (\mathbf{z}^\top \mathbf{A}\mathbf{z})^{1/2}$. The sub-Gaussian norm of a sub-Gaussian random variable X is $\|X\|_{\Psi_2} = \inf\{t > 0 : \mathbb{E}\{\exp(X^2/t^2)\} \leq 2\}$. The sub-Exponential norm of a sub-Exponential random variable X is defined as $\|X\|_{\Psi_1} = \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}$. For a vector $\boldsymbol{X} \in \mathbb{R}^p$, define $\|\boldsymbol{X}\|_{\Psi_j} = \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \|\boldsymbol{v}^\top \boldsymbol{X}\|_{\Psi_j}$, where $\mathbb{S}^{p-1}$ is the unit sphere in $\mathbb{R}^p$ and $j = 1, 2$.

2. Methodology

In this section, we first describe the general multiple changepoint models and algorithms with examples. Then we formally introduce the construction of the Reliever procedure.

2.1 Changepoint Models and Grid-Search Algorithms

Consider a dataset $\{\mathbf{z}_i\}_{i=1}^n$ from a multiple changepoint model

$$\mathbf{z}_i \sim \mathcal{M}_k^*,\ \tau_{k-1}^* < i \leq \tau_k^*,\ k = 1, \ldots, K^* + 1;\ i = 1, \ldots, n, \tag{1}$$

where K^* and $\{\tau_k^*\}$ denote the number and locations of changepoints, respectively, with $\tau_0^* = 0$ and $\tau_{K^*+1}^* = n$. The notations $\{\mathcal{M}_k^*\}$ represent the models governing each data segment, ensuring that $\mathcal{M}_{k-1}^* \neq \mathcal{M}_k^*$. These models may describe nonparametric distributions of $\{\mathbf{z}_i\}$ or specific parametric forms with parameters $\{\boldsymbol{\theta}_k^*\}$, where $\boldsymbol{\theta}_{k-1}^* \neq \boldsymbol{\theta}_k^*$. For specific instances of these models, please refer to Examples 1–3.

Changepoint detection typically proceeds through a grid-search process, involving a model fitting procedure, a loss function to evaluate fit quality, and a grid-search algorithm to determine the optimal segmentation, as illustrated in Figure 1. For a candidate interval $I \subset (0, n]$, a model fitting procedure yields a fitted model $\widehat{\mathcal{M}}_I$ (or $\widehat{\boldsymbol{\theta}}_I$ in parametric scenarios) based on the data segment $\{\mathbf{z}_i : i \in I\}$. The quality of this fit is evaluated using a loss function $\mathcal{L}(I; \widehat{\mathcal{M}}_I)$ (or $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)$ for parametric models).

Example 1 (Parametric models, convex M-estimation) *Consider the general parametric changepoint models within the framework of (1), where each $\mathbf{z}_i \in \mathbb{R}^p$ and*

$$\mathbf{z}_i \textit{ has distribution } P_{\boldsymbol{\theta}_k^*},\ \tau_{k-1}^* < i \leq \tau_k^*,\ k = 1, \ldots, K^* + 1;\ i = 1, \ldots, n.$$

In scenarios with small p and large n, one uses M-estimation for model fitting, which yields $\widehat{\boldsymbol{\theta}}_I = \arg\min_{\boldsymbol{\theta} \in \Theta} \sum_{i \in I} \ell(\mathbf{z}_i, \boldsymbol{\theta})$, where $\ell(\mathbf{z}, \boldsymbol{\theta})$ is a convex function with respect to $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. The loss evaluation is defined as $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I) = \sum_{i \in I} \ell(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_I)$. Employing convex losses, such as the absolute deviation or the Huber loss, is particularly effective in managing heavy-tailed observations or outliers for changepoint detection (Fearnhead and Rigaill, 2019).

Example 2 (High-dimensional linear models, lasso) *In (1), each sample pair $\mathbf{z}_i = (y_i, \mathbf{x}_i)$ consists of a response $y_i \in \mathbb{R}$ and covariates $\mathbf{x}_i \in \mathbb{R}^p$, modeled by*

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}_k^* + \epsilon_i,\ \tau_{k-1}^* < i \leq \tau_k^*,\ k = 1, \ldots, K^* + 1;\ i = 1, \ldots, n, \tag{2}$$

where $\{\boldsymbol{\theta}_k^*\}$ are regression coefficients and $\{\epsilon_i\}$ denote random noises. In high-dimensional settings where both p and n are large, lasso is used for model fitting, that is,

$$\widehat{\boldsymbol{\theta}}_I = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min} \{ \sum_{i \in I} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2 + \lambda_I \|\boldsymbol{\theta}\|_1 \},$$

with λ_I as a tuning parameter. The loss evaluation function is specified as $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I) = \sum_{i \in I} (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\theta}}_I)^2$. Detecting changes in high-dimensional linear models has recently garnered considerable attention; see, for example, Leonardi and Bühlmann (2016), Rinaldo et al. (2021), Wang et al. (2021b) and Xu et al. (2024).

Example 3 (Nonparametric distributions) *Within the framework of (1), consider the samples $\mathbf{z}_i \equiv z_i \in \mathbb{R}$ and*

$$z_i \textit{ has a distribution function } F_k^*,\ \tau_{k-1}^* < i \leq \tau_k^*,\ k = 1, \ldots, K^* + 1;\ i = 1, \ldots, n.$$

Model fitting is performed using the empirical distribution function $\{\widehat{F}_I(t) : t \in \mathbb{R}\}$ for the data subset $\{z_i : i \in I\}$. The loss evaluation function, proposed by Zou et al. (2014), is the negative of an integrated nonparametric maximum log-likelihood.

A grid-search algorithm is then employed to minimize a specific criterion over all possible segmented data sequences. This criterion typically comprises the sum of losses evaluated for each segment, along with a penalty that accounts for the complexity of the segmentation. To be specific, let $\mathcal{T}_K(\delta_{\mathsf{m}}) = \{(\tau_1, \ldots, \tau_K) : 0 \equiv \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} \equiv n, \tau_{k+1} - \tau_k \geq \delta_{\mathsf{m}},\ k = 0, \ldots, K\}$ be the set of K candidate changepoints, where $\delta_{\mathsf{m}} > 0$ is the minimal-spacing parameter; see Remark 4. For $(\tau_1, \ldots, \tau_K) \in \mathcal{T}_K(\delta_{\mathsf{m}})$ that partitions the data into $K + 1$ segments, the criterion is generally formulated as

$$\sum_{k=1}^{K+1} \mathcal{L}((\tau_{k-1}, \tau_k]; \widehat{\mathcal{M}}_k) + \gamma K, \tag{3}$$

where $\gamma \geq 0$ controls the level of penalization to avoid overfitting. Optimal-kind algorithms (for example, SN, OP, or PELT, see Section 1) aim to find the exact minimizer over the entire search space $\mathcal{T}_K(\delta_{\mathsf{m}})$. This involves evaluating a sequence of losses (and fitting the corresponding models) for all $O(n^2)$ intervals $I \subset (0, n]$ satisfying $|I| \geq \delta_{\mathsf{m}}$, sequentially explored using a dynamic programming scheme. Although PELT uses a pruning strategy to skip certain intervals and thus reduces this complexity to $O(n)$, this reduction does not always apply (see Eq. (4) in Killick et al. (2012)). In contrast, greedy-kind algorithms, such as binary segmentation (BS), WBS, narrowest-over-threshold, or SeedBS, consider only a subset of these intervals in a sequential and greedy manner, aiming to reach a local minimizer. To illustrate, consider the BS algorithm. This algorithm begins by solving (3) with $K = 1$, which involves approximately $O(n)$ intervals. The resulting changepoint divides the data sequence into two segments. The algorithm then repeats the same procedure within each segment to identify new changepoints. This iterative process continues until a segment contains fewer observations than δ_{m} or until a stopping rule is triggered. Overall, BS involves approximately $O(n \log n)$ intervals. Throughout, we regard any interval $I = (a, b]$ with integers $0 \leq a < b \leq n$ as a *search interval* and collect all such candidates in the set $\mathcal{I} = \{I : I \subset (0, n]\}$. Given the collection of loss values $\{\mathcal{L}(I; \widehat{\mathcal{M}}_I) : I \in \mathcal{I}\}$, the grid-search algorithm can be regarded as the operator $\mathcal{A} = \mathcal{A}(\{\mathcal{L}(I; \widehat{\mathcal{M}}_I) : I \in \mathcal{I}\})$, which maps these evaluations to a final segmentation. In fact, $\mathcal{A}$ inspects only a subset of $\mathcal{I}$—either because intervals shorter than a minimal-spacing parameter δ_{m} are excluded or because greedy strategies (for example, BS) deliberately restrict the search. Whenever it is necessary to distinguish between the two, we denote this subset actually explored by the algorithm by $\mathcal{I}_{\mathcal{A}} \subset \mathcal{I}$.

The grid-search process becomes computationally demanding when model fits along the search path, $\{\widehat{\mathcal{M}}_I : I \in \mathcal{I}_{\mathcal{A}}\}$, becomes costly. This is particularly evident in scenarios like those described in Examples 1–3, where a single model fit requires substantial computational effort, and updating neighboring model fits by adding or removing observations remains elusive.

2.2 Relief Intervals

Our approach is straightforward yet highly adaptable, and it integrates seamlessly with any grid-search algorithm $\mathcal{A}$. We begin by constructing a set of deterministic intervals $\mathcal{R}$. During the search process, for a search interval $I \in \mathcal{I}$, a proxy or *Relief* model, $\widehat{\mathcal{M}}_{R_I}$, fitted using data from an interval $R_I \in \mathcal{R}$, replaces $\widehat{\mathcal{M}}_I$ when evaluating the loss $\mathcal{L}(I; \widehat{\mathcal{M}}_I)$. Each interval $R_I \in \mathcal{R}$ is referred to as a *relief interval* to distinguish it from a search interval I. It is possible for multiple search intervals to correspond to a single relief interval, and not all relief intervals may be visited during the search. The notation R_I indicates that the selection of a relief interval is dependent on the current search interval I. For simplicity, we will use R interchangeably with R_I, which should not lead to confusion. The key to the construction of $\mathcal{R}$ lies in satisfying two properties. First, it significantly reduces the number of intervals for which a sequence of models needs to be fitted, as opposed to fitting models for every search interval. Second, it ensures that the losses computed using the relief models exhibit behaviors similar to those computed with the original models, thereby allowing for consistent changepoint detection.

Definition 1 (Relief intervals) *Let $\delta_\mathsf{m} > 0$ be the minimal-spacing parameter. Define $0 < w \leq 1$ as the wriggle parameter and $b > 1$ as the growth parameter. For each $0 \leq k \leq \lfloor \log_b\{(1+w)n/\delta_\mathsf{m}\}\rfloor$, construct the kth layer of relief intervals, consisting of n_k intervals of length ℓ_k, evenly shifted by s_k, that is, $\mathcal{R}_k = \{(qs_k, qs_k + \ell_k] + a_k : 0 \leq q \leq n_k\}$, where $\ell_k = b^k\delta_\mathsf{m}/(1+w)$, $s_k = w\ell_k$, $n_k = \lfloor (n-\ell_k)/s_k \rfloor$, and $a_k = n/2 - (\ell_k + n_k s_k)/2$ is an adjustment factor to center the intervals in $\mathcal{R}_k$ around $n/2$. The complete set of relief intervals is $\mathcal{R} = \bigcup_{k=0}^{\lfloor \log_b\{(1+w)n/\delta_\mathsf{m}\}\rfloor} \mathcal{R}_k$.*

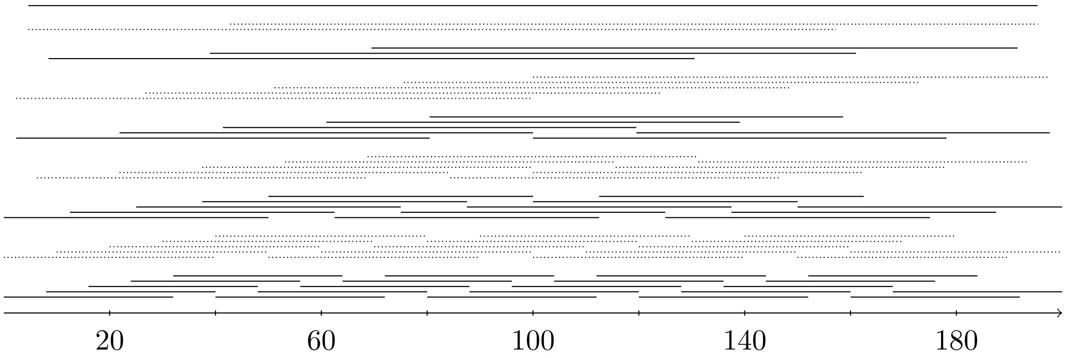


Figure 3: Illustration of relief intervals with $n = 200, \delta_\mathsf{m} = 50, w = 0.25$ and $b = 1.25$.

The rationale behind the construction of relief intervals is to ensure that for any search interval $I \in \mathcal{I}$ with $|I| \geq \delta_\mathsf{m}$, there is always a relief interval $R \in \mathcal{R}$ such that $R \subset I$ and $|R|/|I|$ is maximized. We define the *coverage ratio* as

$$r = \min_{I \in \mathcal{I}: |I| \geq \delta_\mathsf{m}} \max_{R \in \mathcal{R}; R \subset I} \frac{|R|}{|I|}, \qquad 0 < r \leq 1.$$

Proposition 2 *(i) In general, $|\mathcal{R}| \leq c_{w,b} n/\delta_\mathsf{m}$ and $r \geq \{(1+w)b\}^{-1}$, where $c_{w,b} = \{(1+w)b\}/\{w(b-1)\}$. (ii) Setting $(1+w) = b = r^{-\frac{1}{2}}$, $|\mathcal{R}| = \{n(r^{-\frac{1}{2}} - 1)^2\}/(\delta_\mathsf{m} r)$. Additionally, if δ_m and $r \in (0,1)$ are fixed constants, then $|\mathcal{R}| = O(n)$. (iii) If $\delta_\mathsf{m} = C\log n$ for some constant $C > 0$ and $w = b - 1 = \delta_\mathsf{m}^{-\frac{1}{2}}$, then $|\mathcal{R}| \leq n\{1 + (C\log n)^{-\frac{1}{2}}\}^2 = O(n)$ and $r \geq \{1 + (C\log n)^{-\frac{1}{2}}\}^{-2} \approx 1 - 2(C\log n)^{-\frac{1}{2}}$.*

Proposition 2 demonstrates that, by selecting appropriate wriggle and growth parameters, alongside a minimal-spacing parameter, the number of relief intervals approaches linearity with the sample size n while achieving a nearly perfect coverage ratio. To facilitate practical applications, setting a single coverage ratio parameter $r \in (0,1)$ is sufficient, with $1+w=b=r^{-\frac{1}{2}}$. Figure 3 illustrates the construction of relief intervals with $n=200$, $\delta_{\mathsf{m}}=50$, $w=0.25$, and $b=1.25$ (corresponding to $r=0.64$). The parameter r balances computational complexity and estimation accuracy; see Remark 10. Table 2 reports, for $n=1200$ and $\delta_{\mathsf{m}}=30$, the number of search intervals examined by the original SN implementation ($r=1$; that is, $|\mathcal{I}_{\mathcal{A}}|$ for SN), as well as the corresponding number of relief intervals obtained for various coverage ratios of r.

Table 2: Number of model fits required the SN algorithm: Reliever for various coverage ratios r versus the original implementation ($r=1$).

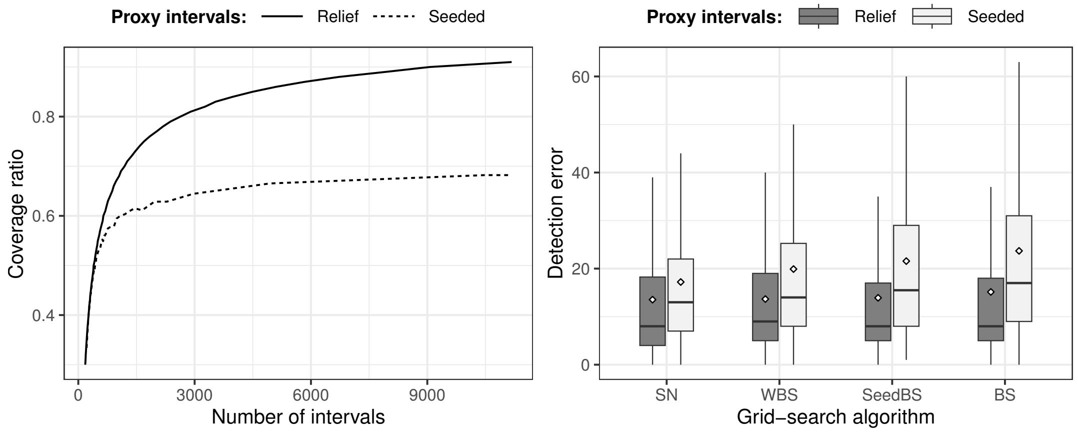
r	0.5	0.6	0.7	0.8	0.9	0.95	0.97	0.99	1
Model fits	440	762	1,298	2,744	12,227	31,699	57,522	196,395	686,206

Remark 3 (Comparison with seeded intervals) *The deterministic nature of our relief intervals is conceptually inspired by the seeded intervals in SeedBS (Kovács et al., 2022), yet they diverge significantly in their design principles due to differing objectives. The seeded intervals were developed to replace random (wild) intervals in WBS, focusing on shorter intervals that typically contain a single changepoint to reduce the occurrence of longer intervals that may encompass multiple changepoints. In contrast, the relief intervals are designed to ensure that each search interval closely matches a relief interval of similar length, thereby producing comparable loss values. Our approach is compatible with various grid-search algorithms, including WBS and SeedBS. A natural question arises: can seeded intervals serve as proxy intervals in place of relief intervals within the proposed Reliever framework? Figure 4(a) displays the coverage ratio r against the number of proxy intervals, revealing that starting at* $r \approx 0.5$, *our construction achieves a higher coverage ratio with the same number of intervals. Notably, using seeded intervals limits the coverage ratio to approximately* 0.68, *even with an increased number of intervals. Figure 4(b) compares the changepoint detection error between the two constructions, both utilizing* $1,000$ *proxy intervals, across various grid-search algorithms, under a high-dimensional linear model with multiple changepoints (as described in Section 4.1). This comparison demonstrates that our construction, with its higher coverage, typically yields better detection accuracy.*

2.3 The Reliever Procedure

(1) Require a gird search algorithm $\mathcal{A}=\mathcal{A}(\{\mathcal{L}(I;\widehat{\mathcal{M}}_I) : I \in \mathcal{I}\})$ with a minimal-spacing parameter $\delta_{\mathsf{m}} \geq 0$ and a model-fitting procedure $\widehat{\mathcal{M}}_I$ for any interval I such that $|I| \geq \delta_{\mathsf{m}}$, and a coverage ratio parameter $r \in (0,1]$;

(2) Create relief intervals $\mathcal{R}$ with the wriggle and growth parameters satisfying $1+w=b=r^{-\frac{1}{2}}$ according to Definition 1;



(a) (b)

Figure 4: Comparison of coverage ratio and changepoint detection error for two proxy interval strategies within the Reliever framework.

(3) Execute the gird search with relief models, that is, $\mathcal{A} = \mathcal{A}(\{\mathcal{L}(I; \widehat{\mathcal{M}}_{R_I}) : I \in \mathcal{I}\})$ with $R_I = \arg\max_{R \in \mathcal{R}, R \subset I} |R|$.

The Reliever procedure can be used in conjunction with both optimal- and greedy-kind grid-search algorithms, as outlined in Section 1, represented by $\mathcal{A} = \mathcal{A}(\{\mathcal{L}(I; \widehat{\mathcal{M}}_I) : I \in \mathcal{I}\})$. The primary distinction when using Reliever compared to the original implementation is the employment of a relief model $\widehat{\mathcal{M}}_{R_I}$ to evaluate the loss function $\mathcal{L}(I; \widehat{\mathcal{M}}_I)$. This modification not only maintains versatility but also significantly reduces the number of model fits required, as each loss evaluation leverages a pre-fitted model from a relief interval, dramatically lowering computational overhead compared to the original implementations (refer to Table 1 for computational comparisons).

Remark 4 (On implementations) *The minimal-spacing parameter δ_{m} is an inherent component of the grid-search algorithm, ensuring that the model/parameter is estimatable on each segment I. For example, in linear models, Bai and Perron (1998) set $\delta_{\mathrm{m}} = p$ (with p the covariate dimension) so that the least-squares estimator is well-posed. Our construction of relief intervals imposes no additional practical restriction on δ_{m}. Theoretically, large δ_{m} guarantees that $\widehat{\mathcal{M}}_I$ behaves well and converges to its population counterpart as $|I|$ grows—a standard requirement in nonparametric and high-dimensional settings. For instance, in multivariate nonparametric changepoint detection via kernel density, Padilla et al. (2023) use $\delta_{\mathrm{m}} = h^{-p} \log(n)$ (with h the bandwidth and p the data dimension). In high-dimensional linear models with temporal dependence, Xu et al. (2024) set $\delta_{\mathrm{m}} = O((s \log(p \vee n))^{2/\zeta - 1})$ (with p the covariate dimension, s the sparsity level, and ζ characterizing dependence and noise tails), which simplifies to $O(s \log(p \vee n))$ in the independent, sub-Gaussian case ($\zeta = 1$). For numerical stability, we use $\delta_{\mathrm{m}} = 20$ in simulation studies (Section 4). In Section E.3, we investigate robustness to small choice of δ_{m}.*

The construction of relief intervals also hinges on the coverage-ratio parameter r, which governs the trade-off between computational cost and estimation accuracy (see Remark 10).

We recommend viewing r as a budget parameter. With ample computing time, choose $r \approx 1$ (which recovers the original algorithm). When runtime is critical, pick the largest r that satisfies the time constraints. Extensive experiments show that $0.8 < r < 0.9$ typically cuts runtime substantially while maintaining satisfactory detection accuracy compared to the original implementation; see Section 4. When the set of grid-search intervals is known (or easily predicted) in advance, the total fitting time for Reliever can be approximated. Practitioners may then tune r to match a target runtime with reasonable confidence. Implementation details are provided in Section E.5.

3. Theoretical Justifications

Despite the broad applicability of Reliever across various changepoint detection algorithms and model settings, establishing a unified theoretical framework for analyzing detection accuracy is challenging without specific assumptions regarding the model, the fitting procedure, and the grid-search algorithm. Here, we first offer an indirect justification by examining the variations in loss values resulting from the application of Reliever. This examination focuses on parametric changepoint models with convex minimization routines for model fitting, as demonstrated in Example 1. Furthermore, in Section 3.2, we present rigorous theoretical results concerning the estimation accuracy of multiple changepoints in the context of high-dimensional linear models employing lasso, as detailed in Example 2. Unless stated otherwise, we assume that the observations $\mathbf{z}_i$ are temporally independent.

3.1 Variations in Loss Values: Convex Minimization

Continue with Example 1. In the original implementation of a grid-search algorithm $\mathcal{A}$, losses are evaluated as $\mathcal{L}(I;\widehat{\boldsymbol{\theta}}_I) = \sum_{i\in I} \ell(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_I)$, where $\widehat{\boldsymbol{\theta}}_I = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i\in I} \ell(\mathbf{z}_i, \boldsymbol{\theta})$. With the Reliever approach, these loss calculations are replaced by $\mathcal{L}(I;\widehat{\boldsymbol{\theta}}_R) = \sum_{i\in I} \ell(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_R)$, where $R \in \mathcal{R}$ is a relief interval corresponding to I. To analyze the impact of this substitution, let $\overline{\mathcal{L}}(I,\boldsymbol{\theta}) = \mathbb{E}\mathcal{L}(I;\boldsymbol{\theta})$ and $\boldsymbol{\theta}_I^\circ = \arg\min_{\boldsymbol{\theta}\in\Theta} \overline{\mathcal{L}}(I,\boldsymbol{\theta})$ denote population loss and its minimizer, respectively. Define $G_I(\boldsymbol{\alpha}) = |I|^{-1}\sum_{i\in I} g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ + \boldsymbol{\alpha}|I|^{-\frac{1}{2}})$ and $\overline{G}_I(\boldsymbol{\alpha}) = \mathbb{E}G_I(\boldsymbol{\alpha})$, where $g(\mathbf{z},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell(\mathbf{z},\boldsymbol{\theta})$.

Under mild regularity conditions for convex M-estimation, Proposition 5 shows that the difference between the losses $\mathcal{L}(I;\widehat{\boldsymbol{\theta}}_I)$ and $\mathcal{L}(I;\widehat{\boldsymbol{\theta}}_R)$ remains uniformly controlled over all search intervals $I \in \mathcal{I}$, regardless of whether I contains changepoints.

(a) $\ell(\cdot,\mathbf{z})$ is convex on the domain Θ for all fixed $\mathbf{z}$ and Θ is a compact and convex subset of $\mathbb{R}^p$.

(b) The expectation $\mathbb{E}\ell(\mathbf{z}_i,\boldsymbol{\theta})$ is finite for all $\mathbf{z}_i$ and fixed $\boldsymbol{\theta}\in\Theta$.

(c) The population minimizer $\boldsymbol{\theta}_I^\circ$ uniquely exists and is interior point of Θ.

(d) $\|g(\mathbf{z}_i,\boldsymbol{\theta})\|_{\Psi_1} \le C_{A.1}$ for each $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_I^\circ$.

(e) $\overline{\mathcal{L}}(I,\boldsymbol{\theta})$ is twice differentiable at $\boldsymbol{\theta}_I^\circ$ and $\mathbf{H}_I \triangleq |I|^{-1}\nabla^2_{\boldsymbol{\theta}}\overline{\mathcal{L}}(I,\boldsymbol{\theta}_I^\circ)$ is positive-define.

(f) $|\overline{G}_I(|I|^{\frac{1}{2}}(\boldsymbol{\theta}-\boldsymbol{\theta}_I^\circ)) - \mathbf{H}_I(\boldsymbol{\theta}-\boldsymbol{\theta}_I^\circ)| = C_{A.2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_I^\circ\|_2^2$.

(g) $\|g(\mathbf{z}_i, \boldsymbol{\theta}) - g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ)\|_{\Psi_1} \le C_{A.3}\|\boldsymbol{\theta} - \boldsymbol{\theta}_I^\circ\|_2$.

(h) $|I|^{-1}\overline{\mathcal{L}}(I, \boldsymbol{\theta})$ is ρ-strongly convex in the compact set Θ.

(i) $\mathbb{E}g(\mathbf{z}_i, \boldsymbol{\theta})$ is ζ-Lipschitz continuous w.r.t. $\boldsymbol{\theta}$.

(j) For $i \in I \setminus R$, $\|\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_i^\circ\|_2 \le \Delta_\infty$ where $\Delta_\infty > 0$ is a fixed constant.

(k) $\|\mathbf{H}_R^{-1} - \mathbf{H}_I^{-1}\|_{\mathsf{op}} \le C_{A.4}\|\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_I^\circ\|_2$ and $\|\mathbf{H}_I^{-1}\|_{\mathsf{op}} \le C_{A.5}$ for any interval I.

Proposition 5 *Given that the conditions (a)–(k) hold, with probability at least* $1 - n^{-C}$ *for some constant* $C > 0$*, the event*

$$0 \le \frac{1}{|I|}\{\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_R) - \mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)\} \le O\left(\left\|\frac{1}{|I|}\sum_{i \in I \setminus R} \mathbb{E}\nabla_{\boldsymbol{\theta}}\ell(\mathbf{z}_i, \boldsymbol{\theta}_R^\circ)\right\|_2^2 + \frac{(1-r)\log n}{r|I|} + \frac{(\log n)^2}{r^2|I|^2}\right) \quad (4)$$

holds uniformly for all intervals $R \subset I \in \mathcal{I}$*, where* $\nabla_{\boldsymbol{\theta}}\ell(\cdot, \boldsymbol{\theta})$ *denotes the gradient or subgradient. In particular, in scenarios where* $I = (s, e]$ *contains no changepoint, or there is only one changepoint* $\tau \in I$ *such that* $\min(\tau - s, e - \tau) = O(\log n)$*, this event simplifies to*

$$0 \le \frac{1}{|I|}\{\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_R) - \mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)\} \le C_1\left(\frac{(1-r)\log n}{r|I|} + \frac{(\log n)^2}{r^2|I|^2}\right), \quad (5)$$

where $C_1 > 0$ *is a constant.*

Remark 6 (On Conditions (a)–(k)) *Conditions (a)–(k) parallel the regularity assumptions of Niemiro (1992), which analyzed M-estimators obtained through convex minimization under independent and identically distributed (i.i.d.) data. These conditions fundamentally concern the smoothness and convexity of the loss function* ℓ *and its expectation. Conditions (a)–(f) are the standard ingredients for deriving estimation error bounds in classical convex M-estimation with i.i.d. samples. The additional conditions (g)–(k) address distributional heterogeneity created by changepoints and ensure uniform control of the difference terms* $\|\boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_R^\circ\|_2$ *and* $\|\widehat{\boldsymbol{\theta}}_I - \widehat{\boldsymbol{\theta}}_R\|_2$ *for all* $I \in \mathcal{I}$*. The proof of Proposition 5 relies on a novel non-asymptotic Bahadur-type representation for* $\widehat{\boldsymbol{\theta}}_I - \widehat{\boldsymbol{\theta}}_R$ *in the presence of changepoints across all* $I \in \mathcal{I}$*, which may be of independent interest.*

In Proposition 5, Eq.(5) indicates that the discrepancy between the Reliever-based loss $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_R)$ and the original loss $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)$ vanishes when the data within I are (nearly) homogeneous and $(\log n)/|I|$ goes to zero. This provides a justification for employing Reliever. Conversely, for heterogeneous I containing changepoints distant from the boundaries, this vanishing property of the discrepancy may not hold. Surprisingly, the inequality $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_R) \ge \mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)$ in Eq.(4) becomes instrumental in excluding inconsistent changepoint estimators in such scenarios. Therefore, we can expect that Reliever effectively tracks the original search path. To gain some intuition, consider the scenario with a single changepoint τ^* such that $\min(\tau^*, n - \tau^*) \ge \delta_{\mathsf{m}}$ or $\tau^* \in \mathcal{T}_1(\delta_{\mathsf{m}})$. And the grid-search algorithm is specified as the first BS step. Define the changepoint estimator as $\widehat{\tau}_{\text{original}} = \arg\min_{\tau \in \mathcal{T}_1(\delta_{\mathsf{m}})} S_I^{(I)}(\tau)$, where $S_I^{(I)}(\tau) = \mathcal{L}(I_{1,\tau}, \widehat{\boldsymbol{\theta}}_{I_{1,\tau}}) + \mathcal{L}(I_{2,\tau}, \widehat{\boldsymbol{\theta}}_{I_{2,\tau}})$, and for any τ, $I_{1,\tau} = (0, \tau]$ and $I_{2,\tau} = (\tau, n]$.

The Reliever-based changepoint estimator is then defined as $\widehat{\tau} = \arg\min_{\tau \in \mathcal{T}_1(\delta_{\mathsf{m}})} S_I^{(R)}(\tau)$, where $S_I^{(R)}(\tau) = \mathcal{L}(I_{1,\tau}, \widehat{\boldsymbol{\theta}}_{R_{1,\tau}}) + \mathcal{L}(I_{2,\tau}, \widehat{\boldsymbol{\theta}}_{R_{2,\tau}})$, and $R_{j,\tau} \subset I_{j,\tau}$ is the corresponding relief interval for $j = 1, 2$. We present the following corollary, which establishes the consistency of $\widehat{\tau}$ in the sense that $|\widehat{\tau} - \tau^*|/n \to 0$.

Corollary 7 *Assume $\delta_{\mathsf{m}} = C_{\mathsf{m}} \log n$ for some constant $C_{\mathsf{m}} > 0$ and the event described in Proposition 5 holds. If there exists a sufficiently large constant $C_2 > 0$ such that for any $\tau \in \mathcal{T}_1(\delta_{\mathsf{m}})$ satisfying $|\tau - \tau^*| > \delta$ for a constant $\delta > 0$,*

$$S_I^{(I)}(\tau) - S_I^{(I)}(\tau^*) > C_2 \log n \tag{6}$$

holds, then $|\widehat{\tau} - \tau^*| \leq \delta$.

Corollary 7 is a direct consequence of Proposition 5. Assume $|\widehat{\tau} - \tau^*| > \delta$. Since $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_R) \leq \mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)$ according to Eq.(4), it follows that $S_I^{(R)}(\widehat{\tau}) \geq S_I^{(I)}(\widehat{\tau})$. Utilizing Eq.(5), we derive

$$S_I^{(R)}(\tau^*) \leq S_I^{(I)}(\tau^*) + 2C_1 \Big\{ \frac{1-r}{r} + \frac{n \log n}{\tau^*(n - \tau^*) r^2} \Big\} \log n.$$

Considering Eq.(6), by setting $C_2 \geq 2C_1\{(1-r)r^{-1} + C_{\mathsf{m}}^{-1} r^{-2}\}$, we have $S_I^{(R)}(\widehat{\tau}) - S_I^{(R)}(\tau^*) > [C_2 - 2C_1\{(1-r)r^{-1} + C_{\mathsf{m}}^{-1} r^{-2}\}] \log n \geq 0$. Therefore, the assumption $|\widehat{\tau} - \tau^*| > \delta$ leads to a contradiction, establishing the validity of Corollary 7. Eq.(6) imposes implicit constraints on the model, ensuring that the original grid-search algorithm produces a consistent changepoint estimator, that is, $|\widehat{\tau}_{\text{original}} - \tau^*| \leq \delta$. Verifying Eq.(6) or establishing a lower bound for $S_I^{(I)}(\tau) - S_I^{(I)}(\tau^*)$ is a well-accepted technique for justifying the consistency of changepoint estimators (Csörgő and Horváth, 1997). Corollary 7 demonstrates that the consistency proof for the original grid-search algorithm can readily be extended to the Reliever estimator. This approach is also applicable to multiple changepoint detection tasks.

3.2 Changepoint Detection Accuracy: Lasso Regression

To deepen our understanding of how variations in loss values impact the accuracy of changepoint detection with the Reliever method, we delve into the detection of multiple changepoints in high-dimensional linear models (cf. Example 2).

We use the OP algorithm (for example, Leonardi and Bühlmann, 2016). The standard implementation minimizes the criterion

$$\sum_{k=1}^{K+1} \mathcal{L}((\tau_{k-1}, \tau_k]; \widehat{\boldsymbol{\theta}}_{(\tau_{k-1}, \tau_k]}) + \gamma K, \tag{7}$$

over all candidate changepoints $(\tau_1, \ldots, \tau_K) \in \mathcal{T}_K(\delta_{\mathsf{m}})$. For any search interval $I \in \mathcal{I}$ with $|I| \geq \delta_{\mathsf{m}}$, model parameters are estimated as $\widehat{\boldsymbol{\theta}}_I = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \{\sum_{i \in I} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2 + \lambda_I \|\boldsymbol{\theta}\|_1\}$, and loss values are computed as $\mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I) = \sum_{i \in I} (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\theta}}_I)^2$. The tuning parameter γ helps avoid overestimating the number of changepoints. Specific values for λ_I and γ are established through a rigorous theoretical analysis discussed later in this section.

To integrate Reliever with OP, we construct a collection of relief intervals $\mathcal{R}$ with a fixed coverage ratio parameter $0 < r < 1$. The optimization criterion within the Reliever framework is thus reformulated as

$$\sum_{k=1}^{K+1} \mathcal{L}((\tau_{k-1}, \tau_k]; \widehat{\boldsymbol{\theta}}_{R_k}) + \gamma K, \text{ with } R_k = \arg\max_{R \in \mathcal{R}, R \subset (\tau_{k-1}, \tau_k]} |R|. \tag{8}$$

This criterion and the original in Eq.(7) are specific instances of a more general optimization problem

$$\min_{(\tau_1, \ldots, \tau_K) \in \mathcal{T}_K(\delta_{\mathsf{m}})} \left\{ \sum_{k=1}^{K+1} \mathcal{L}\Big((\tau_{k-1}, \tau_k]; \widetilde{\boldsymbol{\theta}}\big((\tau_{k-1}, \tau_k]\big)\Big) + \gamma K \right\}. \tag{9}$$

Here, $\widetilde{\boldsymbol{\theta}}(I)$ can be any valid estimator of regression coefficients for $I \in \mathcal{I}$ such that $|I| \geq \delta_{\mathsf{m}}$. Setting $\widetilde{\boldsymbol{\theta}}(I) = \widehat{\boldsymbol{\theta}}_I$ recovers the original criterion (7). Alternatively, choosing $\widetilde{\boldsymbol{\theta}}(I) = \widehat{\boldsymbol{\theta}}_{R_I}$ with $R_I = \arg\max_{R \in \mathcal{R}, R \subset I} |R|$, gives us the reformulated problem (8). OP manages this optimization (9) by integrating a sequence of parameter estimation $\{\widetilde{\boldsymbol{\theta}}(I)\}$ and loss evaluation $\{\mathcal{L}_I \equiv \mathcal{L}(I; \widetilde{\boldsymbol{\theta}}(I))\}$ steps along the search path, with dynamic ordering of intervals I determined by OP itself.

We first establish a deterministic claim about the consistency and near rate-optimality of the resulting changepoint estimators, conditional on an event measuring the quality or *goodness* of the evaluated losses. We introduce some notations and conditions crucial for this analysis. For any search interval $I \in \mathcal{I}$, denote $\boldsymbol{\theta}_I^\circ = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\{\mathcal{L}(I; \boldsymbol{\theta})\}$, and define $\Delta_I = (|I|^{-1} \sum_{i \in I} \|\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ\|_\Sigma^2)^{\frac{1}{2}}$, where $\boldsymbol{\theta}_i^\circ = \boldsymbol{\theta}_{\{i\}}^\circ$ for $i = 1, \ldots, n$. For $k = 1, \ldots, K^*$, let $\Delta_k = \|\boldsymbol{\theta}_{k+1}^* - \boldsymbol{\theta}_k^*\|_\Sigma$ be the magnitude of change at τ_k^*, with $\Delta_0 = \Delta_{K^*+1} = \infty$.

Condition 1 (Change signals) *There exists a sufficiently large constant $C_{\mathsf{snr}} > 0$ such that for $k = 1, \ldots, K^* + 1$, $\tau_k^* - \tau_{k-1}^* \geq C_{\mathsf{snr}} s \log(p \vee n)(\Delta_{k-1}^{-2} + \Delta_k^{-2} + 1)$.*

Condition 2 (Regression coefficients) *(a) Sparsity: $|\mathcal{S}_k| \leq s < p$, where $\mathcal{S}_k = \{1 \leq j \leq p : \theta_{k,j}^* \neq 0\}$ and $\theta_{k,j}^*$ is the jth component of $\boldsymbol{\theta}_k^*$; (b) Boundness: $|\theta_{k,j}^*| \leq C_\theta$ for some constant $C_\theta > 0$.*

Condition 3 (Covariates and noises) *(a) Covariates $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. sub-Gaussian with zero mean and covariance Σ, satisfying $0 < \underline{\kappa} \leq \sigma_x^2 < \infty$, where $\underline{\kappa} = \lambda_{\min}(\Sigma)$ and $\sigma_x^2 = \lambda_{\max}(\Sigma)$ are the minimum and maximum eigenvalues of Σ, respectively. Furthermore, $\|\Sigma^{-\frac{1}{2}} \mathbf{x}_i\|_{\Psi_2} \leq C_x$ for some constant $C_x > 0$; (b) Noises $\{\epsilon_i\}_{i=1}^n$ are i.i.d. sub-Gaussian with zero mean, variance σ_ϵ^2, and a sub-Gaussian norm C_ϵ.*

These conditions are commonly adopted in the literature for changepoint detection in high-dimensional linear models (Leonardi and Bühlmann, 2016; Wang et al., 2021b; Rinaldo et al., 2021; Xu et al., 2024). Specifically, Condition 1 introduces a *local multiscale* signal-to-noise ratio (SNR) requirement for the spacing between neighboring changepoints, providing greater flexibility compared to the global SNR condition in existing works like Leonardi and Bühlmann (2016) and Wang et al. (2021b).

Lemma 8 (Goodness of loss evaluations) *Under Condition 1, for the problem (9) with $\delta_{\mathsf{m}} = C_{\mathsf{m}} s \log(p \vee n)$ for a sufficiently large constant $C_{\mathsf{m}} > 0$, and $\gamma = C_\gamma s \log(p \vee n)$ for a constant $C_\gamma > 0$, the solution $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{K}})$ satisfies:*

$$\widehat{K} = K^* \text{ and } \max_{1 \le k \le K^*} \min_{1 \le j \le \widehat{K}} \frac{1}{2} \Delta_k^2 |\tau_k^* - \widehat{\tau}_j| \le \widetilde{C} s \log(p \vee n),$$

for some constant $\widetilde{C} > 0$, conditional on the event $\mathbb{G} = \mathbb{G}_1 \cap \mathbb{G}_2^- \cap \mathbb{G}_2^+ \cap \mathbb{G}_3$. Here,

$$\mathbb{G}_1 = \left\{ \text{for any } I \in E_1, \left| \mathcal{L}_I - \sum_{i \in I} \epsilon_i^2 \right| < C_{8.1} s \log(p \vee n) \right\},$$

$$\mathbb{G}_2^- = \left\{ \text{for any } I \in E_2^-, \mathcal{L}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_I^2 |I| > -C_{8.2} s \log(p \vee n) \right\},$$

$$\mathbb{G}_2^+ = \left\{ \text{for any } I \in E_2^+, \mathcal{L}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_I^2 |I| < C_{8.2} s \log(p \vee n) \right\},$$

$$\mathbb{G}_3 = \left\{ \text{for any } I \in E_3, \mathcal{L}_I - \sum_{i \in I} \epsilon_i^2 > (1 - C_{8.3}) \Delta_I^2 |I| \right\},$$

where $E_1 = \{I : |I| \ge \delta_{\mathsf{m}}, \Delta_I = 0\}$, $E_2^- = \{I = (a, b] : |I| \ge \delta_{\mathsf{m}}; I \cap \mathcal{T}^* = \{\tau_k^*\}, \min(\tau_k^* - a, b - \tau_k^*) \le 2\widetilde{C} \Delta_k^{-2} s \log(p \vee n)\}$, $E_2^+ = \{I \in E_2^- : I \cap \mathcal{T}^* = \{\tau_k^*\}, |I| \ge (\Delta_k^2 \vee 1) C_{\mathsf{snr}} s \log(p \vee n)\}$, and $E_3 = \{I : |I| \ge \delta_{\mathsf{m}}, \Delta_I^2 |I| \ge \widetilde{C} s \log(p \vee n)\}$. Here $C_{8.1}$, $C_{8.2}$ and $C_{8.3}$ are positive constants. In addition, the constants C_γ and $\widetilde{C}$ only depends on C_{snr}, C_{m}, $C_{8.1}$, $C_{8.2}$, and $C_{8.3}$.

Lemma 8 presents a deterministic result. The probabilistic conditions come into play when certifying that the event $\mathbb{G}$ holds with high probability for both the standard OP implementation with $\mathcal{L}_I = \mathcal{L}(I; \widehat{\boldsymbol{\theta}}_I)$ and the accelerated Reliever version with $\mathcal{L}_I = \mathcal{L}(I; \widehat{\boldsymbol{\theta}}_{R_I})$. This lemma offers new insights into the necessary conditions for the evaluated losses using a general model-fitting procedure $\widehat{\boldsymbol{\theta}}(I)$ along the OP grid-search path to produce consistent and nearly rate-optimal changepoint estimators, which may be of independent interest. Theorem 9 further asserts that this event $\mathbb{G}$ occurs with high probability under additional Conditions 2–3.

Theorem 9 *Suppose that Conditions 1–3 hold. Let C_λ and C_γ be positive constants, and $0 < C_{\mathsf{m}} < C_{\mathsf{snr}}$ be sufficiently large constants. The solution $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{K}})$ of either Problem (7) or Problem (8) with $\delta_{\mathsf{m}} = C_{\mathsf{m}} s \log(p \vee n)$, $\lambda_I = C_\lambda C_x \sigma_x D_I \{|I| \log(p \vee n)\}^{\frac{1}{2}}$, and $\gamma = C_\gamma s \log(p \vee n)$, satisfies that*

$$\mathbb{P}\Big\{ \widehat{K} = K^* \text{ and } \max_{1 \le k \le K^*} \min_{1 \le j \le \widehat{K}} \frac{1}{2} \Delta_k^2 |\tau_k^* - \widehat{\tau}_j| \le \widetilde{C} s \log(p \vee n) \Big\} \ge 1 - (p \vee n)^{-c},$$

where $D_I = (C_x^2 \Delta_I^2 + C_\epsilon^2)^{\frac{1}{2}}$. The constants C_γ, C_λ, $\widetilde{C}$ and c are independent of (n, p, s, K^*). Moreover, under the same probability event, there exists a constant $C > 0$ such that for all $1 \le k \le K^* + 1$,

$$\|\widehat{\boldsymbol{\theta}}_{(\widehat{\tau}_{k-1}, \widehat{\tau}_k]} - \boldsymbol{\theta}_k^*\|_2 \le C \left\{ \frac{s \log(p \vee n)}{\tau_k^* - \tau_{k-1}^*} \right\}^{\frac{1}{2}}.$$

Theorem 9 demonstrates that, under mild conditions and with appropriately chosen tuning parameters γ and λ_I, both the original and Reliever-enhanced implementations of OP consistently estimate the number of changepoints and achieve a state-of-the-art localization rate $n^{-1}|\tau_k^* - \widehat{\tau}_k| \le C\Delta_k^{-2}n^{-1}s\log(p \vee n)$ with high probability. This rate exhibits the phenomenon of *superconsistency* for changepoint estimation in high-dimensional linear regression with multiple changepoints, extending a well-known result for single changepoint scenarios (Lee et al., 2016). Importantly, our theoretical analysis supports scenarios where K^*, the number of changepoints, varies with n and may potentially diverge. When $K^* = O(1)$, our findings are consistent with those reported in Rinaldo et al. (2021) and Xu et al. (2024), which use OP-type algorithms. Wang et al. (2021b) allows for K^* to diverge and derives this rate using a WBS-type algorithm. Additionally, it is noteworthy that the tuning parameter λ_I, which serves as the regularization factor for the lasso model within each interval I, not only scales with $|I|^{\frac{1}{2}}$ but is also modulated by the change magnitude Δ_I^2. In fact, determining the rate of λ_I involves examining the uniform bound of a sequence of mean-zero (sub-)gradients, where the variance is, however, influenced by Δ_I^2. Previous works, such as those by Wang et al. (2021b) and Xu et al. (2024), typically assume $\sup_I \Delta_I^2 = O(1)$, which simplifies the dependency of λ_I to $|I|^{\frac{1}{2}}$ alone. Theorem 9 underscores the nuanced, change-adaptive nature of the regularization parameter λ_I. While the comprehensive exploration of this parameter's dynamics is outside the scope of our current study, it marks a promising avenue for future research and merits further investigation.

Remark 10 (Tradeoff between computational time and estimation accuracy) *At first glance, Reliever might appear to provide an advantage without a corresponding cost, as the localization rate initially appears to be unaffected by the coverage ratio r. However, a deeper analysis of the underlying proofs reveals that r subtly influences the constant $\tilde{C}$ in the localization rate, particularly since r is held constant. More precisely, the magnitude of $\tilde{C}$ is dependent on several constants including C_{snr}, C_{m}, $C_{8.1}$, $C_{8.2}$, and $C_{8.3}$, as detailed in Lemma 8. By setting C_{snr} and C_{m} to sufficiently large values, we determine that $\widetilde{C} = 2(1 - C_{8.3})^{-1}(3C_{8.1} + 10C_{8.2})$. From the proof, it becomes evident that $C_{8.j} \propto r^{-2}$ for $j = 1, 2, 3$. Therefore, as r decreases, the constants $C_{8.j}$ increase, which in turn elevates $\tilde{C}$ and leads to deteriorated localization rates for smaller values of r. Similar relation can also be observed in Proposition 5 for single changepoint detection. The observation illustrates a pivotal tradeoff: lower values of r enhance computational speed at the expense of localization precision. As r approaches 1, the distinction between Reliever and the original grid-search algorithm diminishes, indicating minimal computational gains in exchange for optimal localization accuracy. It is also worth emphasizing that by choosing a fixed $0 < r < 1$, Reliever can always reduce the number of model fits to $O(n)$ as Proposition 2 shows. For detailed derivations and specific values of $C_{8.j}$, $j = 1, 2, 3$, please refer to Corollary 24 in Section C.3.*

3.2.1 Temporal Dependence: Extending the Localization Theory

A closer inspection of the proof of Theorem 9 (independent case) reveals that independence is invoked only inside a Bernstein-type tail bound that establishes oracle inequalities (see Section C). Replacing this bound with a version suited to dependent data therefore suffices to extend the theory.

Xu et al. (2024) establish such a Bernstein bound for functionally dependent sequences. Incorporating their bound yields Reliever's nearly rate-optimal localization guarantee for temporally dependent data. The main ingredients are summarized below; detailed proofs are deferred to Section D.

Definition 11 (Functional dependent sequence (Wu, 2005; Xu et al., 2024)) *For each $t \in \mathbb{Z}$, let*

$$\mathbf{x}_t = \mathbf{g}_t(\mathcal{F}_t^X),$$

where $\mathcal{F}_t^X = \{X_s\}_{s\le t}$ is generated from i.i.d. elements $\{X_s\}_{s\in\mathbb{Z}}$, and $\mathbf{g}_t : \mathcal{F}_t^X \to \mathbb{R}^p$ is measurable. The sequence $\{\mathbf{x}_t\}_{t\in\mathbb{Z}}$ is then called functionally dependent, with dependence functions $\{\mathbf{g}_t\}_{t\in\mathbb{Z}}$ and generating elements $\{X_t\}_{t\in\mathbb{Z}}$. Let $\mathcal{F}_{t,s}^X$ be the same as $\mathcal{F}_t^X$ except that X_s is replaced by an independent copy $\widetilde{X}_s$. Define, for $q \ge 1$, the functional dependence measure and its cumulative version as

$$\delta_{s,q}^{\mathbf{x}} = \sup_{\mathbf{v}\in\mathcal{S}^{p-1},\, t\in\mathbb{Z}} [\mathbb{E}|\mathbf{v}^\top(\mathbf{x}_t - \mathbf{x}_{t-s})|^q]^{\frac{1}{q}} \text{ and } \Delta_{m,q}^{\mathbf{x}} = \sum_{s=m}^{\infty} \delta_{s,q}^{\mathbf{x}},\ m \in \mathbb{Z},$$

respectively.

Within the functional-dependence framework, we impose the following conditions.

Condition 4 (Change signals) *There exists a sufficiently large constant $C_{\mathsf{snr}} > 0$ such that for $k = 1, \ldots, K^* + 1$, $\tau_k^* - \tau_{k-1}^* \ge C_{\mathsf{snr}} s \log(p \vee n)[\Delta_{k-1}^{-2} + \Delta_k^{-2} + \{s\log(p\vee n)\}^{2/\zeta-2}]$ where $\zeta \in (0,1)$. We additionally assume $K = O(1)$ and $\sup_{1\le k\le K^*} \Delta_k \le C_\Delta$ for some universal constant $C_\Delta > 0$.*

Condition 5 (Regression coefficients) $|\mathcal{S}_k| \le s < p$*, where* $\mathcal{S}_k = \{1 \le j \le p : \theta_{k,j}^* \ne 0\}$.

Condition 6 (Covariates and noises) *Let $\zeta_1 > 0$ and $\zeta_2 \in (0, 2]$ be two constants such that $(\zeta_1^{-1}+\zeta_2^{-1})^{-1} = \zeta \in (0,1)$. (a) **Covariates.** The sequence $\{\mathbf{x}_i\}_{i=1}^n$ is a consecutive subsequence of an infinite functionally dependent sequence $\{\mathbf{x}_t\}_{t\in\mathbb{Z}} \subset \mathbb{R}^p$ with a time-invariant dependence function $\mathbf{g}^{\mathbf{x}}$ ($\mathbf{g}_t = \mathbf{g}^{\mathbf{x}}$ for all $t \in \mathbb{Z}$). Moreover, $\sup_{m\ge0} \exp(cm^{\zeta_1})\Delta_{m,4}^{\mathbf{x}} \le D_x$, for some constant $D_x > 0$. Assume each $\mathbf{x}_i$ has mean zero and covariance Σ, satisfying $0 < \underline{\kappa} \le \sigma_x^2 < \infty$, where $\underline{\kappa} = \lambda_{\min}(\Sigma)$ and $\sigma_x^2 = \lambda_{\max}(\Sigma)$, respectively. Furthermore, $\|\Sigma^{-\frac{1}{2}}\mathbf{x}_i\|_{\Psi_{\zeta_2}} \le C_x$ for some constant $C_x > 0$. (b) **Noises**. The sequence $\{\epsilon_i\}_{i=1}^n$ is a consecutive subsequence of an infinite functionally dependent sequence $\{\epsilon_t\}_{t\in\mathbb{Z}} \subset \mathbb{R}$ with a time-invariant dependence function g^ϵ. Moreover, $\sup_{m\ge0}\exp(cm^{\zeta_1})\Delta_{m,4}^{\epsilon} \le D_\epsilon$, for some constant $D_\epsilon > 0$. Assume $\{\epsilon_i\}_{i=1}^n$ are independent of $\{\mathbf{x}_i\}_{i=1}^n$, and each ϵ_i has mean zero and variance σ_ϵ^2. Furthermore, $\|\epsilon_i\|_{\Psi_{\zeta_2}} \le C_\epsilon$.*

Conditions 4–6 mirror their counterparts in the temporally independent setting (Conditions 1–3) in temporal independence scenarios; we additionally assume $K = O(1)$ and $\sup_{1\le k\le K^*} \Delta_k \le C_\Delta$, as in Xu et al. (2024).

Corollary 12 *Suppose that Conditions 4–6 hold. Let C_λ and C_γ be some positive constants, and $0 < C_{\mathsf{m}} < C_{\mathsf{snr}}$ be sufficiently large constants. The solution $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{K}})$ of either*

Problem (7) or Problem (8) with $\delta_{\mathsf{m}} = C_{\mathsf{m}}\{s\log(p\vee n)\}^{2/\zeta-1}$, $\lambda_I = C_\lambda C_x \sigma_x \{|I|\log(p\vee n)\}^{\frac{1}{2}}$, *and* $\gamma = C_\gamma s\log(p\vee n)$, *satisfies that*

$$\mathbb{P}\Big\{\widehat{K} = K^* \text{ and } \max_{1\le k\le K^*}\min_{1\le j\le \widehat{K}} \Delta_k^2|\tau_k^* - \widehat{\tau}_j| \le \widetilde{C}s\log(p\vee n)\Big\} \ge 1-(p\vee n)^{-c}.$$

The constants C_γ, C_λ, $\widetilde{C}$ *and* c *are independent of* (n, p, s, K^*). *Moreover, under the same probability event, there exists a constant* $C > 0$ *such that for all* $1 \le k \le K^* + 1$,

$$\|\widehat{\boldsymbol{\theta}}_{(\widehat{\tau}_{k-1},\widehat{\tau}_k]} - \boldsymbol{\theta}_k^*\|_2 \le C\left\{\frac{s\log(p\vee n)}{\tau_k^* - \tau_{k-1}^*}\right\}^{\frac{1}{2}}.$$

4. Numerical Studies

To evaluate the effectiveness of the Reliever approach compared to the original implementation of various grid-search algorithms, we explore two scenarios: high-dimensional linear changepoint models (cf. Example 2) and nonparametric changepoint models (cf. Example 3). The grid-search algorithms assessed include SN, WBS (with $M = 100$ random intervals), and SeedBS (using a decay parameter $a = 2^{-1/2}$, as recommended by Kovács et al. (2022)), implemented with a known number of changepoints for a fair comparison. For nonparametric models, we also consider OP and PELT, which do not presuppose the number of changepoints. The accuracy of changepoint estimation is quantified using the Hausdorff distance $\max\{\mathrm{OE}, \mathrm{UE}\}$, where $\mathrm{OE} = \max_{1\le j\le \widehat{K}}\min_{1\le k\le K^*}|\tau_k^* - \widehat{\tau}_j|$ is the over-segmentation error and $\mathrm{UE} = \max_{1\le k\le K^*}\min_{1\le j\le \widehat{K}}|\tau_k^* - \widehat{\tau}_j|$ is the under-segmentation error. The following results are based on 500 replications.

4.1 High-Dimensional Linear Models

In this scenario, we examine changepoint detection in high-dimensional linear models as outlined in Example 2, with $n \in \{300, 600, 900, 1200\}$ and $p = 100$. The covariates $\{\mathbf{x}_i\}$ are i.i.d. from the standard multivariate Gaussian distribution, and the noises $\{\epsilon_i\}$ are i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, 1)$. We introduce three changepoints at $\{\tau_k^*\}_{k=1}^3 = \{\lfloor 0.22n\rfloor, \lfloor 0.55n\rfloor, \lfloor 0.77n\rfloor\}$. The regression coefficients $\{\boldsymbol{\theta}_k^*\}$ are generated such that $\theta_{k,j} = 0$ for $j = 3, \ldots, p$, and $\theta_{k,1}$ and $\theta_{k,2}$ are uniformly sampled, satisfying the SNRs $\|\boldsymbol{\theta}_1\|_2^2/\operatorname{Var}(\epsilon_1) = 2$ and $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2^2/\operatorname{Var}(\epsilon_1) = 2^{-1}$ for $k = 2, 3, 4$. Here $\theta_{k,j}$ denotes the jth element of $\boldsymbol{\theta}_k$. We set $\delta_{\mathsf{m}} = 20$ for numerical stability and employ lasso for parameter estimation using the glmnet package (Friedman et al., 2010) in R. We apply three grid-search algorithms, SN, WBS, and SeedBS, assuming a known number of changepoints $\widehat{K} = K^* = 3$. For each algorithm, we scale the regularization parameter $\lambda_I = \lambda|I|^{\frac{1}{2}}$, with λ ranging from a set of 30 values. The changepoint detection error for each algorithm is reported as the minimal Hausdorff distance achieved across all values of λ.

Figures 5–6 display the changepoint detection error and the computational time for each grid-search algorithm across different values of the coverage ratio r. The value $r = 1$ corresponds to the original implementation. The results indicate that as r approaches 1, the performance of Reliever approaches that of the original implementation. For $r = 0.9$, Reliever delivers comparable results to the original implementations but with substantial reductions in computational time. Even when $r = 0.6$, the performance is still acceptable, considering the negligible running time.

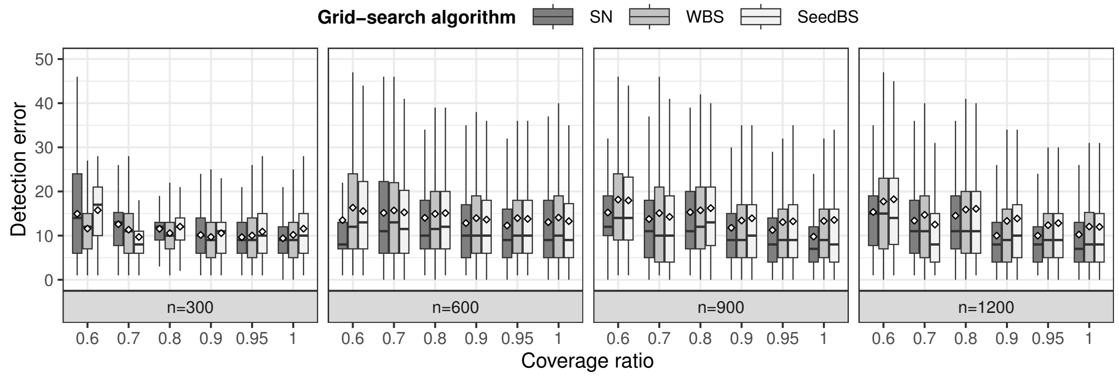


Figure 5: Changepoint detection error for various grid-search algorithms across varying values of the coverage ratio, under the high-dimensional linear model.

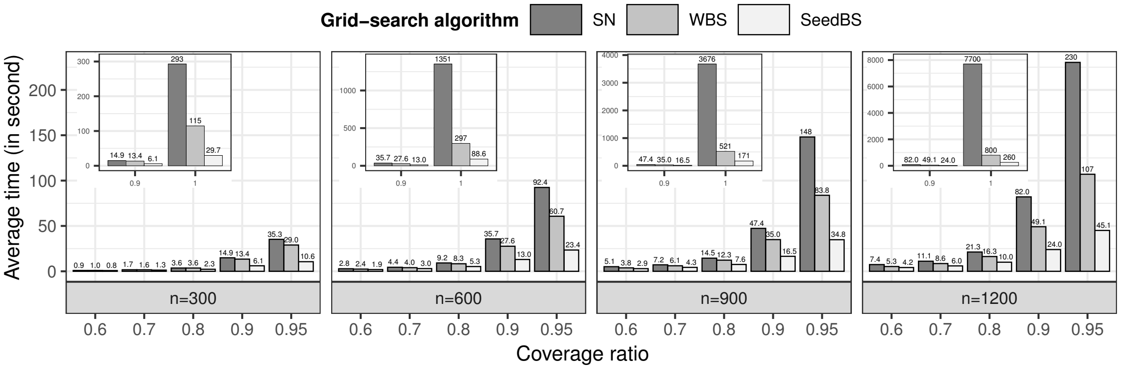


Figure 6: Computational time for various grid-search algorithms across varying values of the coverage ratio, under the high-dimensional linear model.

4.2 Univariate Nonparametric Models

In the second scenario, we explore changepoint detection for univariate nonparametric distributions as described in Example 3. We employ the same three-changepoint structure used in the first scenario. The data within the four segments are generated from different distributions, that is, $\mathcal{N}(0,1)$, $\chi^2_{(3)}$ (standardized to have unit variance), $\chi^2_{(1)}$ (likewise standardized), and $\mathcal{N}(0,1)$, respectively. We implement SN, WBS, and SeedBS to identify changepoints, assessing their effectiveness across varying coverage ratios r. Figures 7–8 summarize the changepoint detection error and computational time for each algorithm. The Reliever method demonstrates robust performance, particularly for r values above 0.7. Notably, SN maintains consistent accuracy and efficiency across a range of r values.

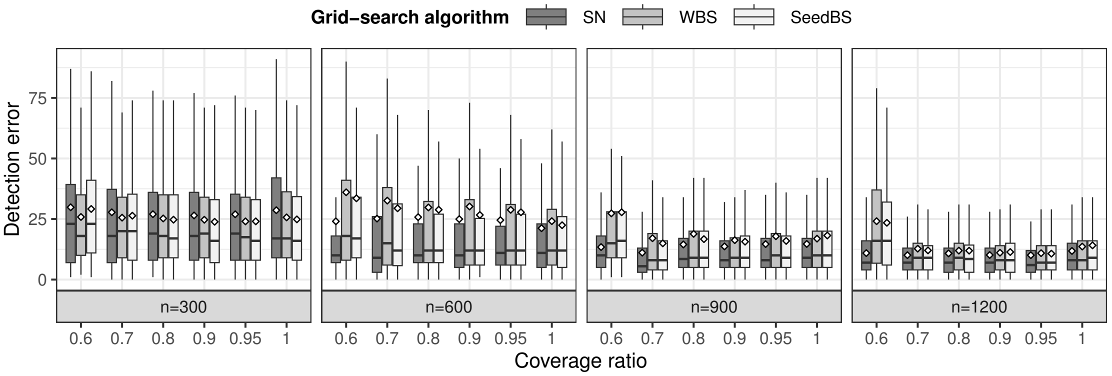


Figure 7: Changepoint detection error for various grid-search algorithms across varying values of the coverage ratio, under the nonparametric model.

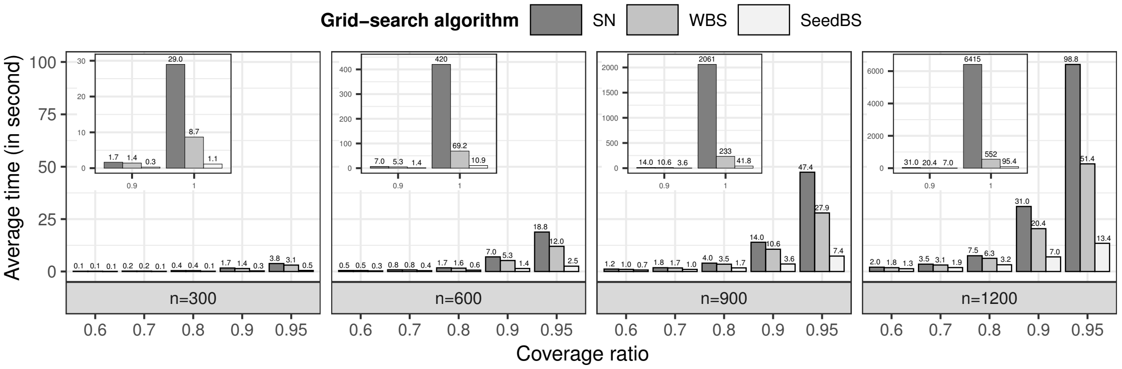


Figure 8: Computational time for various grid-search algorithms across varying values of the coverage ratio, under the nonparametric model.

4.3 Nonparametric Changepoint Detection via Kernel-Density Estimation

This section shows how Reliever integrates naturally with the kernel-density CUSUM framework of Padilla et al. (2023). Consider observations $\{\mathbf{z}_i\}_{i=1}^n \subset \mathbb{R}^p$. For any search interval $I = (a, b] \in \mathcal{I}$ and for a kernel $\mathcal{K}(\cdot)$ with bandwidth $h > 0$, the density is estimated by $\widehat{f}_I(\mathbf{z}) = \{|I|h^p\}^{-1} \sum_{i\in I} \mathcal{K}\Big((\mathbf{z} - \mathbf{z}_i)/h\Big)$, $\forall \mathbf{z} \in \mathbb{R}^p$. Given a split point $t \in I$, define the pointwise CUSUM contrast $\tilde{f}_{I,t}(\mathbf{z}) = \sqrt{(b-t)(t-a)/(b-a)}\{\widehat{f}_{(a,t]}(\mathbf{z}) - \widehat{f}_{(t,b]}(\mathbf{z})\}$ and the overall CUSUM contrast $\|\tilde{f}_{I,t}\|_n^2 = n^{-1}\sum_{j=1}^n \tilde{f}_{I,t}^2(\mathbf{z}_j)$. The changepoint within I is then estimated as $\arg\max_{a+\delta_{\mathrm{m}}<t\le b-\delta_{\mathrm{m}}} \|\tilde{f}_{I,t}\|_n^2$.

Padilla et al. (2023) adopt the SeedBS algorithm; to embed their method in our Reliever framework we translate the contrast-based CUSUM into a loss-based evaluation. Define the loss $\mathcal{L}(I; \widehat{f}_I) = \sum_{j=1}^n \sum_{i\in I} \Big\{ h^{p-1} K\Big((\mathbf{z}_j - \mathbf{z}_i)/h\Big) - \widehat{f}_I(\mathbf{z}_j)\Big\}^2$. Maximizing the overall contrast is equivalent to $\arg\min_{a+\delta_{\mathrm{m}}<t\le b-\delta_{\mathrm{m}}} \mathcal{L}((a,t]; \widehat{f}_{(a,t]}) + \mathcal{L}((t,b]; \widehat{f}_{(t,b]})$.

In the original SeedBS algorithm, $\widehat{f}_I$ and $\mathcal{L}(I; \widehat{f}_I)$ are computed for every search interval $I \in \mathcal{I}$ along the search path. With Reliever, we instead fit $\widehat{f}_R$ on only $O(n)$ deterministic proxy intervals $R \in \mathcal{R}$, reuse these fits, and evaluate $\mathcal{L}(I; \widehat{f}_R)$ for all required I.

We replicate the three-changepoint scenario of Sections 4.1–4.2 (with $n = 1200$, $p = 5$, $\{\tau_k^*\}_{k=1}^3 = \{0.22n, 0.55n, 0.77n\}$, $\delta_{\mathrm{m}} = 2$). Data are generated similar to those in Scenario 3 of Padilla et al. (2023): two independent sequences $\{\mathbf{e}_{z,i}\} \subset \mathbb{R}^p$ and $\{\mathbf{e}'_{z,i}\} \subset \mathbb{R}^p$ with i.i.d. entries $\mathbf{e}_{z,i,j} \sim \text{Pareto}(3, 1)$ and $\mathbf{e}'_{z,i,j} \sim \text{Uniform}(-\sqrt{3}, \sqrt{3})$ drive AR(1) processes $\mathbf{z}_i = 0.3\mathbf{z}_{i-1} + \mathbf{e}_{z,i}$ and $\mathbf{z}'_i = 0.3\mathbf{z}'_{i-1} + \mathbf{e}'_{z,i}$ (starting with $\mathbf{z}_0 = \mathbf{z}'_0 = \mathbf{0}$). Then for $i \in (\tau_1^*, \tau_2^*] \cup (\tau_3^*, n]$, we reset $\mathbf{z}_i$ by $\mathbf{z}_i = \sqrt{0.8}\mathbf{z}_i - \sqrt{0.2}\mathbf{z}'_i$. We use a standard radial basis function (RBF) kernel $\mathcal{K}$ with bandwidth $h = 1$, and set Reliever's coverage ratio to $r = 0.9$.

Table 3 reports changepoint detection error and density-fit time. Although one can update $\widehat{f}_I$ and $\mathcal{L}(I; \widehat{f}_I)$ in $O(p)$ per step—for instance, $\widehat{f}_{(a,b+1]}(\mathbf{z}) = \frac{1}{b-a+1}\{(b-a)\widehat{f}_{(a,b]}(\mathbf{z}) + h^{-p}K((\mathbf{z} - \mathbf{z}_{b+1})/h)\}$, Reliever still cuts total fitting time across WBS, SeedBS, and SN, with almost no change in detection error.

Table 3: Average detection error and density-fit time (centiseconds) for multivariate nonparametric changepoint detection via kernel density estimation, comparing original and Reliever-enabled versions (coverage ratio $r = 0.9$) of WBS, SeedBS, and SN.

Grid search	WBS		SeedBS		SN	
Model fitting	Original	Reliever	Original	Reliever	Original	Reliever
Error	38.9	39.6	41.7	41.6	34.8	34.8
Time	91.9	54.9	19.2	14.9	447.1	147.3

4.4 Integration of Reliever with OP and PELT

To investigate how Reliever integrates with OP and PELT, without presupposing the number of changepoints, we revisit the nonparametric changepoint model discussed in Section 4.2. This model accommodates the applicability of PELT, proposed by Haynes et al. (2017). Additionally, we examine the data-generating process of Model 1 from Zou et al. (2014) and Haynes et al. (2017), with $K^* = 11$ changepoints and Student-$t(3)$-distributed noises. Specifically,

$$z_i = \sum_{k=1}^{K^*} h_k \mathbf{1}_{\{i > \tau_k^*\}} + \sigma\epsilon_i,\ i = 1, \ldots, n,$$

where $\{\tau_k^*\}/n = \{0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81\}$, $\{h_k\} = \{2.01, -2.51, 1.51, -2.01, 2.51, -2.11, 1.05, 2.16, -1.56, 2.56, -2.11\}$, $\{\epsilon_i\} \overset{i.i.d}{\sim} t(3)$, and $\sigma = 0.5$. This configuration is designated as Model (B), in contrast to the three-changepoint setting, which is referred to as Model (A). Table 4 summarises the results for $n = 1000$. The findings highlight that while PELT significantly reduces computational time compared to OP, Reliever can further decrease this burden without compromising the detection accuracy, remaining nearly identical to those obtained via the original OP algorithm.

Table 4: Average absolute errors of changepoint number estimates, detection error, and computational time of OP and PELT, with and without Reliever, under the univariate nonparametric model with $n = 1000$ and $K^* = 11$.

Model	Coverage ratio	$\lvert\widehat{K} - K^*\rvert$		OE		UE		Time (Second)	
		OP	PELT	OP	PELT	OP	PELT	OP	PELT
(A)	(Original) 1.0	0.13	0.13	14.30	14.30	48.18	48.18	2804.17	1098.88
	0.9	0.13	0.13	14.01	14.01	49.16	49.16	22.65	19.22
	0.8	0.14	0.14	14.05	14.10	49.05	49.11	7.71	6.60
	0.7	0.15	0.15	13.95	14.12	54.36	55.83	3.56	3.01
	0.6	0.17	0.17	14.87	15.12	57.39	58.21	2.07	1.72
	0.5	0.21	0.21	18.67	19.72	69.16	72.34	1.34	1.07
(B)	(Original) 1.0	0.01	0.01	2.32	2.32	2.54	2.54	3010.80	65.85
	0.9	0.01	0.01	2.29	2.29	2.51	2.51	24.26	8.56
	0.8	0.01	0.01	2.25	2.25	2.47	2.47	8.22	3.51
	0.7	0.01	0.01	2.37	2.37	2.41	2.41	3.80	1.80
	0.6	0.00	0.00	2.18	2.18	2.18	2.18	2.19	1.13
	0.5	0.01	0.01	2.45	2.45	2.56	2.56	1.41	0.75

4.5 Integration of Reliever with DCDP

Li et al. (2023) propose the two-stage Divide and Conquer Dynamic Programming (DCDP) algorithm. Stage I runs dynamic programming (DP, that is, SN) on a coarse grid of candidate points, while Stage II locally refines each preliminary changepoint to improve accuracy. This design reduces computation by largely eliminating the grid-search space. Because DCDP and Reliever address different bottlenecks, the two can be combined. We therefore evaluate (i) DP (Original versus Reliever); (ii) DCDP Stage I (Original versus Reliever); and (iii) DCDP Stages I–II (Original versus Reliever). Here, “Original” corresponds to Reliever with $r = 1$ (that is, full model fits). We set $r = 0.9$ for Reliever and use a coarse-grid step of 20 for DCDP. Results for the high-dimensional linear setting of Section 4.1 ($n = 1200$) are summarized in Table 5. Across all three schemes, Reliever cuts fitting time while keeping error roughly at the same level—so the combined strategy “Reliever + DCDP” is promising for very large-scale problems.

Table 5: Detection error and runtime for DP and DCDP with/without Reliever in a high-dimensional linear model (coarse step = 20; coverage ratio $r = 0.9$).

Grid search	DP		DCDP I		DCDP I–II	
Model fitting	Original	Reliever	Original	Reliever	Original	Reliever
Error	12.6	13.4	13.9	14.3	13.6	13.9
Time	7,700.0	82.0	32.9	10.7	33.4	11.2

4.6 Comparison with Two-Step Methods

We present a comparative analysis between the Reliever method and the two-step approach proposed by Kaul et al. (2019b). The two-step method is specifically designed to detect a single changepoint in a high-dimensional linear model. It involves an initial guess of the changepoint, which divides the data into two intervals. Proxy models are then fitted within these intervals. Consequently, both methods expedite the process of change detection by reducing extensive model fits. For mitigating the uncertainty in the initialization, multiple guesses are considered, and a changepoint estimator that minimizes the total loss on both segments is reported. In our study, we consider the high-dimensional linear model discussed in Section 5.1 of Kaul et al. (2019b), with $n = 1200$ and $\tau^* = 120$. We consider multiple initial guesses, specifically $0.25n, 0.5n, 0.75n$. The results presented in Table 6 indicate that although the two-step method may offer faster computation due to fewer model fits, it also exhibits larger changepoint detection error. This can be attributed to its performance being heavily reliant on the accuracy of the initial changepoint estimate (or the quality of the corresponding intervals). In contrast, the Reliever method demonstrates stability across a range of choices for the parameter r, varying from 0.9 to 0.3.

Table 6: Comparison of average changepoint detection error and computational time (*in centiseconds*) between the Reliever method and the two-step method under the high-dimensional linear model with single changepoint, and $(n, \delta_{\mathsf{m}}) = (1200, 30)$. The numbers in parentheses represent the corresponding standard errors.

	Two-step	$r = 0.9$	$r = 0.7$	$r = 0.5$	$r = 0.3$
Error	18.6(3.2)	9.2(1.0)	9.4(1.1)	7.6(0.8)	8.7(1.4)
Time (10ms)	60.7(0.6)	480.5(1.1)	141.0(0.4)	89.0(0.3)	64.1(0.3)

Though without theoretical guarantees, the two-step method can be extended for multiple changepoint detection by incorporating BS along with the multiple guess scheme, as suggested by Londschien et al. (2023). This extension can also be applied to WBS and SeedBS in a similar manner. In our study, we examine the examples presented in Sections 4.1 and 4.2 with $n = 1200$. Multiple initial guesses are selected as m-equally spaced quantiles within a search interval, following the recommendation by Londschien et al. (2023). The results depicted in Table 7 reveal that the two-step approach is less efficient for multiple changepoint detection, and increasing the number of multiple initial guesses can even have a detrimental impact on its performance. In contrast, the Reliever method (with $r = 0.9$) exhibits performances that are almost comparable to the original implementation.

We also report the corresponding average computational time in Table 8. It is noteworthy to emphasize that, for multiple changepoint detection tasks, even with $r = 0.9$, the Reliever method shows comparable computational time to the two-step approach. When $r = 0.8$, the Reliever method becomes more efficient. The result is slightly different from the single changepoint case. The reason is that for multiple changepoint detection algorithms like WBS and SeedBS, the two-step method should repeatedly fit the models for every wild/seeded interval. In contrast, the relief models are shared with the global system.

Furthermore, we posit that Reliever serves as a complementary tool rather than a rival to the two-step approach in the domain of multiple changepoint detection. The Reliever method can be combined with the two-step method for multiple changepoint detection tasks. This integration is beneficial because the two-step method still involves a significant number of model fits within the seeded/wild intervals. The Reliever method can further enhance computational efficiency by reducing the time required for these model fits.

Table 7: Comparison of average changepoint detection error between the Reliever method and the two-step method under the multiple changepoint setting in Section 4.1 and Section 4.2. The numbers in parentheses represent the corresponding standard errors

Example	Algorithm	$m = 1$	$m = 3$	$m = 5$	$r = 0.9$	$r = 0.8$	Original
HD	WBS	19.7(0.9)	14.7(0.6)	16.8(0.8)	13.3(0.7)	15.1(0.5)	12.1(0.6)
	SeedBS	21.4(1.0)	17.3(0.8)	17.8(0.8)	13.9(0.7)	15.2(0.5)	12.0(0.6)
NP	WBS	85.5(3.2)	17.4(1.2)	17.5(1.2)	11.1(0.5)	12.3(0.6)	13.6(1.0)
	SeedBS	87.5(3.2)	18.7(1.5)	17.6(1.2)	11.4(0.5)	11.9(0.6)	14.1(1.0)

Table 8: Comparison of average computational time (*in seconds*) between the Reliever method and the two-step method under the multiple changepoint setting in Section 4.1 and Section 4.2.

Example	Algorithm	$m = 1$	$m = 3$	$m = 5$	$r = 0.9$	$r = 0.8$
HD	WBS	20.8(0.2)	37.0(0.6)	52.4(1.2)	49.1(0.8)	16.3(0.3)
	SeedBS	17.2(0.1)	30.2(0.3)	42.6(0.8)	24.0(0.4)	10.0(0.1)
NP	WBS	2.6(0.1)	5.7(0.2)	8.7(0.2)	20.4(0.1)	6.3(0.1)
	SeedBS	1.2(0.1)	2.5(0.1)	3.9(0.1)	7.0(0.1)	3.2(0.1)

5. Concluding Remarks

Searching for multiple changepoints in complex models with large datasets poses significant computational challenges. Current algorithms involve fitting a sequence of models and evaluating losses within numerous intervals during the search process. Existing approaches, such as PELT, WBS, SeedBS, and optimistic search algorithms, aim to reduce the number of (search) intervals. In this paper, we introduce Reliever which specifically relieves the computational burden by reducing the number of fitted models, as they are the primary contributors to computational costs. Our method associates each search interval with a deterministic (relief) interval from a pre-defined pool, enabling the fitting of models only within (or partially within) these selected intervals. The simplicity of the Reliever approach allows for seamless integration with various grid-search algorithms and accommodates different models, providing tremendous potential for leveraging modern machine learning tools (Londschien et al., 2023; Liu et al., 2021; Li et al., 2024).

Reliever incorporates a coverage ratio parameter, which balances computational efficiency and estimation accuracy. For high-dimensional regression models with changepoints, by employing an OP algorithm, we characterize requirements on the search path to ensure consistent and nearly rate-optimal estimators for changepoints; see Lemma 8. Our analysis demonstrates that the Reliever method satisfies these properties for any fixed coverage ratio parameter. Further investigation is warranted to characterize the search path for other algorithms and broader model classes. Additionally, our theoretical analysis highlights the importance of adaptively selecting the nuisance parameter based on the underlying change magnitude. Future research should focus on extending the Reliever to enable data-driven selection of nuisance parameters. While the Reliever focuses on changepoint estimation, it is worth exploring the generalization of these concepts to quantify uncertainty in changepoint detection (Frick et al., 2014; Chen et al., 2023) and perform post-change-estimation inference (Jewell et al., 2022).

Acknowledgments

All authors contributed equally to this work and are listed in alphabetical order. We would like to acknowledge the action editor, Ji Zhu, and anonymous referees for their valuable comments and suggestions, which have improved the manuscript greatly. Qian's research was supported by the National Natural Science Foundation of China (No. 12501410) and the China Postdoctoral Science Foundation (No. 2024M761934, GZB20250716). Wang was supported by the National Natural Science Foundation of China (No. 12471255), the Natural Science Foundation of Shanghai (No. 23ZR1419400), and the Fundamental Research Funds for the Central Universities (No. 63253110). Zou was supported by the National Key R&D Program of China (No. 2022YFA1003703) and the National Natural Science Foundation of China (No. 12231011).

Appendix

The appendix provides proofs of all theoretical results in this article and offers additional numerical analyses.

Appendix A. Proof of Proposition 5

For a fixed $\boldsymbol{\alpha}$, denote the random vectors $\mathbf{x}_i$ by

$$\mathbf{x}_i = g\Big(\mathbf{z}_i, \boldsymbol{\theta}_I^{\circ} + \frac{\boldsymbol{\alpha}}{|I|^{\frac{1}{2}}}\Big) - g(\mathbf{z}_i, \boldsymbol{\theta}_I^{\circ}).$$

Denote $v_I = (\log n)^{\frac{1}{2}}$. By (g), uniformly for all $\|\boldsymbol{\alpha}\|_2 \leq M v_I$ (with some constant $M > 0$), $\|\mathbf{x}_i\|_{\Psi_1} \leq C_{A.3} M v_I |I|^{-\frac{1}{2}}$. Therefore, by applying an exponential inequality,

$$\sup_{\|\boldsymbol{\alpha}\|_2 \leq M v_I} \mathbb{P}\Big[\Big|G_I(\boldsymbol{\alpha}) - G_I(\mathbf{0}) - \overline{G}_I(\boldsymbol{\alpha})\Big| \geq \frac{C_u C_{A.3} M}{c_b} |I|^{-1} v_I (\log n)^{\frac{1}{2}}\Big] \leq 2\exp(-C_u \log n).$$

By (f),

$$\sup_{\|\boldsymbol{\alpha}\|_2 \leq M v_I} \Big|\frac{\mathbf{H}_I \boldsymbol{\alpha}}{|I|^{\frac{1}{2}}} - \overline{G}_I(\boldsymbol{\alpha})\Big| \leq C_{A.2} M^2 v_I^2 |I|^{-1}.$$

The above two inequalities imply that

$$\sup_{\|\boldsymbol{\alpha}\|_2 \leq M v_I} \mathbb{P}\Big[\Big|G_I(\boldsymbol{\alpha}) - G_I(\mathbf{0}) - \frac{\mathbf{H}_I \boldsymbol{\alpha}}{|I|^{\frac{1}{2}}}\Big| \geq \frac{C_u C_{A.3} M}{c_b} |I|^{-1} v_I (\log n)^{\frac{1}{2}}\Big] \leq 2\exp(-C_u \log n).$$

By the chaining technique for convex function, that is, the δ-triangulation argument used in Niemiro (1992),

$$\mathbb{P}\Big[\sup_{\|\boldsymbol{\alpha}\|_2 \leq M v_I} \Big|G_I(\boldsymbol{\alpha}) - G_I(\mathbf{0}) - \frac{\mathbf{H}_I \boldsymbol{\alpha}}{|I|^{\frac{1}{2}}}\Big| \geq C_{A.6} |I|^{-1} v_I (\log n)^{\frac{1}{2}}\Big] \leq 2|I|^{\frac{p}{2}} \exp(-C_u \log n).$$

By the sub-exponential assumption, we choose $M > 0$ such that $\mathbb{P}[\||I|^{\frac{1}{2}} \mathbf{H}_I^{-1} G_I(\mathbf{0})\|_2 \geq (M-1)(\log n)^{\frac{1}{2}}] \leq 2\exp(-C_u \log n)$. It implies that with high probability, $|I|^{\frac{1}{2}} \mathbf{H}_I^{-1} G_I(\mathbf{0})$ is in the ball $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_2 < (M-1)(\log n)^{\frac{1}{2}}\}$. For all $\mathbf{e} \in \mathbb{R}^p$ with $\|\mathbf{e}\|_2 = 1$, let $\boldsymbol{\alpha} = -|I|^{\frac{1}{2}}\{\mathbf{H}_I^{-1} G_I(\mathbf{0}) + (K \log n)|I|^{-1}\mathbf{e}\}$ with $K = 2C_{A.6}/\lambda_{\min}(\mathbf{H}_I)$. With probability at least $1 - 2(1 + |I|^{p/2})\exp(-C_u \log n)$,

$$\begin{aligned} &\mathbf{e}^\top G_I\big(|I|^{\frac{1}{2}} \mathbf{H}_I^{-1} G_I(\mathbf{0}) + (K \log n)|I|^{-\frac{1}{2}} \mathbf{e}\big) \\ \geq &(K|I|^{-1} \log n) \cdot \mathbf{e}^\top \mathbf{H}_I \mathbf{e} - C_{A.6} |I|^{-1} \log n > 0. \end{aligned}$$

It means that $\widehat{\boldsymbol{\theta}}_I$ is in the open ball $\{\boldsymbol{\theta}_I^{\circ} - \mathbf{H}_I^{-1} G_I(\mathbf{0}) + (K \log n)|I|^{-1}\mathbf{e} : \|\mathbf{e}\|_2 < 1\}$. By taking the union bounds over the intervals $I \subset (0, n]$, uniformly with probability at least $1 - \exp(-C_{A.7} \log n)$,

$$(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^{\circ}) = -\mathbf{H}_I^{-1} G_I(\mathbf{0}) + \mathbf{r}_I, \tag{10}$$

where $\max_{|I| \subset (0,n]} \mathbf{r}_I |I| / \log n = O(1)$.

We have now obtained the uniform Bahadur representation, which holds over $I \subset (0, n]$ with high probability. To measure the difference between $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_R$, we first consider the population one. Recall that $R \in \mathcal{R}$ is the relief interval of I. First of all, we study the population minimizers. By the ρ-strong convexity and the definition of $\boldsymbol{\theta}_I^\circ$ and $\boldsymbol{\theta}_R^\circ$,

$$0 \leq \overline{\mathcal{L}}(I, \boldsymbol{\theta}_R^\circ) - \overline{\mathcal{L}}(I, \boldsymbol{\theta}_I^\circ) \leq \nabla_\theta \overline{\mathcal{L}}(I, \boldsymbol{\theta}_R^\circ)^\top (\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_I^\circ) - \frac{\rho |I|}{2} \|\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_I^\circ\|_2^2,$$

which implies that

$$\|\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_I^\circ\|_2 \leq \frac{2}{\rho |I|} \Big\| \sum_{i \in I \setminus R} \mathbb{E} g(\mathbf{z}_i, \boldsymbol{\theta}_R^\circ) \Big\|_2 \leq \frac{2\zeta}{\rho |I|} \sum_{i \in I \setminus R} \|\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_i^\circ\|_2 = O(1 - r).$$

Assume that the Bahadur representation Eq. (10) holds thereafter. We have the following identity of the difference between $\widehat{\boldsymbol{\theta}}_I$ and $\widehat{\boldsymbol{\theta}}_R$,

$$\widehat{\boldsymbol{\theta}}_I - \widehat{\boldsymbol{\theta}}_R = \boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_R^\circ + \mathbf{H}_R^{-1} G_R(\mathbf{0}) - \mathbf{H}_I G_I(\mathbf{0}) + \mathbf{r}_I - \mathbf{r}_R.$$

For $\mathbf{H}_R^{-1} G_R(\mathbf{0}) - \mathbf{H}_I G_I(\mathbf{0})$, further consider the following decomposition,

$$\mathbf{H}_R^{-1} G_R(\mathbf{0}) - \mathbf{H}_I G_I(\mathbf{0}) = (\mathbf{H}_R^{-1} - \mathbf{H}_I^{-1}) G_R(\mathbf{0}) + \mathbf{H}_I^{-1} \{G_R(\mathbf{0}) - G_I(\mathbf{0})\}.$$

For the first part, by the sub-exponential assumption (d), with probability at least $1 - \exp(-C_u \log n)$,

$$\|(\mathbf{H}_R^{-1} - \mathbf{H}_I^{-1}) G_R(\mathbf{0})\|_2 \leq C_{A.8} \|\boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_R^\circ\|_2 \Big[\Big(\frac{\log n}{|R|}\Big)^{\frac{1}{2}} + \frac{\log n}{|R|} \Big].$$

For the second part,

$$G_R(\mathbf{0}) - G_I(\mathbf{0}) = \sum_{i \in R} \Big[\frac{1}{|R|} g(\mathbf{z}_i, \boldsymbol{\theta}_R^\circ) - \frac{1}{|I|} g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ) \Big] - \sum_{i \in I \setminus R} \frac{1}{|I|} g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ) \triangleq \frac{1}{|I|} \sum_{i \in I} \mathbf{x}_i,$$

where $\mathbf{x}_i = [g(\mathbf{z}_i, \boldsymbol{\theta}_R^\circ)|I|/|R|] - g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ)$ for $i \in R$ and $\mathbf{x}_i = -g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ)$ for $i \in I \setminus R$. For any individual $i \in R$, by assumptions (d) and (g),

$$\begin{aligned} \|\mathbf{x}_i\|_{\Psi_1} &= \Big\| \{g(\mathbf{z}_i, \boldsymbol{\theta}_R^\circ) - g(\mathbf{z}_i, \boldsymbol{\theta}_I^\circ)\} + \frac{(1-r)}{r} g(\mathbf{z}_i, \boldsymbol{\theta}_R^\circ) \Big\|_{\Psi_1} \\ &\leq \|\boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_R^\circ\|_2 + \frac{1-r}{r} (C_{A.3} \|\boldsymbol{\theta}_R^\circ - \boldsymbol{\theta}_i^\circ\|_2 + C_{A.1}) \leq \|\boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_R^\circ\|_2 + \frac{1-r}{r} C_{A.9} \end{aligned}$$

For $i \in I \setminus R$,

$$\|\mathbf{x}_i\|_{\Psi_1} \leq (C_{A.3} \|\boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_i^\circ\|_2 + C_{A.1}) \leq C_{A.9}.$$

In the above two inequalities, we make use of Condition (j), the boundness of parameters. By Bernstein's inequality (Lemma 15), with probability at least $1 - \exp(-C_u \log n)$,

$$\|G_R(\mathbf{0}) - G_I(\mathbf{0})\|_2 \leq C_{A.10} \Big[\big(\|\boldsymbol{\theta}_I^\circ - \boldsymbol{\theta}_R^\circ\|_2 + r^{-\frac{1}{2}} (1-r)^{\frac{1}{2}} \big) \Big(\frac{\log n}{|I|} \Big)^{\frac{1}{2}} + \frac{\log n}{r |I|} \Big].$$

Overall we obtain,

$$\|\widehat{\boldsymbol{\theta}}_I - \widehat{\boldsymbol{\theta}}_R\|_2 \le O\bigg(\|\boldsymbol{\theta}_I^{\circ} - \boldsymbol{\theta}_R^{\circ}\|_2 + (1-r)^{\frac{1}{2}}\Big(\frac{\log n}{r|I|}\Big)^{\frac{1}{2}} + \frac{\log n}{r|I|}\bigg).$$

By the definition of $\widehat{\boldsymbol{\theta}}_R$, one obtains $\nabla_{\boldsymbol{\theta}}\mathcal{L}(I, \widehat{\boldsymbol{\theta}}_R) = \sum_{i\in I\setminus R} g(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_R)$. Similarly, by the δ-triangulation argument used in the proof of the Bahadur representation, with probability at least $1-\exp(-C_u \log n)$, uniformly for all intervals I,

$$\Big\|\sum_{i\in I\setminus R}\Big\{g(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_R) - g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ}) - \mathbb{E}\big[g(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_R) - g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ})\big]\Big\}\Big\|_2 = O(\log n),$$

$$\Big\|\sum_{i\in I\setminus R}\mathbb{E}\{g(\mathbf{z}_i, \widehat{\boldsymbol{\theta}}_R) - g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ})\}\Big\|_2 \le \zeta(1-r)|I|\|\widehat{\boldsymbol{\theta}}_R - \boldsymbol{\theta}_R^{\circ}\|_2 = O\bigg((1-r)\Big\{\Big(\frac{|I|\log n}{r}\Big)^{\frac{1}{2}} + \frac{\log n}{r}\Big\}\bigg),$$

$$\Big\|\sum_{i\in I\setminus R} g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ})\Big\|_2 = \Big\|\sum_{i\in I\setminus R}\mathbb{E}g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ})\Big\|_2 + O(\{(1-r)|I|\log n\}^{\frac{1}{2}} + \log n).$$

Combining the above three upper bounds,

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(I, \widehat{\boldsymbol{\theta}}_R) = \Big\|\sum_{i\in I\setminus R}\mathbb{E}g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ})\Big\|_2 + O\bigg((1-r)^{\frac{1}{2}}\Big(\frac{|I|\log n}{r}\Big)^{\frac{1}{2}} + \frac{\log n}{r}\bigg).$$

By the convexity condition (h),

$$\begin{aligned}\frac{1}{|I|}\{\mathcal{L}(I, \widehat{\boldsymbol{\theta}}_R) - \mathcal{L}(I, \widehat{\boldsymbol{\theta}}_I)\} &\le \frac{1}{|I|}\nabla_{\boldsymbol{\theta}}\mathcal{L}(I, \widehat{\boldsymbol{\theta}}_R)^\top(\widehat{\boldsymbol{\theta}}_R - \widehat{\boldsymbol{\theta}}_I) \le \frac{1}{|I|}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(I, \widehat{\boldsymbol{\theta}}_R)\|_2\|\widehat{\boldsymbol{\theta}}_R - \widehat{\boldsymbol{\theta}}_I\|_2 \\ =&O\bigg(\frac{1}{\rho|I|^2}\Big\|\sum_{i\in I\setminus R}\mathbb{E}g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ})\Big\|_2^2 + \frac{(1-r)\log n}{r|I|} + \frac{(\log n)^2}{r^2|I|^2}\bigg). \qquad (11)\end{aligned}$$

When $I = (s, e]$ contains no changepoint, or it is nearly homogeneous such that if a true changepoint $\tau \in I$, then $\min(\tau - s, e - \tau) = O(\log n)$, we have $\sum_{i\in I\setminus R}\mathbb{E}g(\mathbf{z}_i, \boldsymbol{\theta}_R^{\circ}) = O(\min(\tau - s, e - \tau)) = O(\log n)$. Therefore,

$$\frac{1}{|I|}\{\mathcal{L}(I, \widehat{\boldsymbol{\theta}}_R) - \mathcal{L}(I, \widehat{\boldsymbol{\theta}}_I)\} = O\bigg(\frac{(1-r)\log n}{r|I|} + \frac{(\log n)^2}{r^2|I|^2}\bigg).$$

Appendix B. Proof of Lemma 8

We first introduce some notations. For a given changepoint estimation $\tau \in [n]$ and a changepoint set $\mathcal{T} = \{0 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n\}$, denote

$$\mathcal{L}(\mathcal{T}) \triangleq \sum_{k=1}^{K+1}\mathcal{L}((\tau_{k-1}, \tau_k]; \widetilde{\boldsymbol{\theta}}((\tau_{k-1}, \tau_k]))$$

as the loss function in Eq. (9), $\mathcal{K}_+(\tau, \mathcal{T}) \triangleq \min_k\{k : \tau_k > \tau\}$ and $\mathcal{K}_-(\tau, \mathcal{T}) \triangleq \max_k\{k : \tau_k < \tau\}$. For simplicity, further denote $k^*_{\tau,+} = \mathcal{K}_+(\tau, \mathcal{T}^*)$, $\widehat{k}_{\tau,+} = \mathcal{K}_+(\tau, \widehat{\mathcal{T}})$, $k^*_{\tau,-} = \mathcal{K}_-(\tau, \mathcal{T}^*)$

and $\widehat{k}_{\tau,-} = \mathcal{K}_-(\tau, \widehat{\mathcal{T}})$. Let $\widehat{\mathcal{T}} = \{\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{K}}\}$ be the minimizer of Eq. (9). Denote $\delta_{\mathsf{m}} = C_{\mathsf{m}} s \log(p \vee n)$ and $\delta_k = \widetilde{C} s \log(p \vee n) \Delta_k^{-2}$ where $\Delta_k = \|\boldsymbol{\theta}_{k+1}^* - \boldsymbol{\theta}_k^*\|_\Sigma$, and $\mathcal{H} = \{(\widehat{\tau}_a, \widehat{\tau}_{a+1}] : \exists k \in [K^*], \min(\tau_k^* - \widehat{\tau}_a, \widehat{\tau}_{a+1} - \tau_k^*) > \delta_k\}$.

Assume that $\mathcal{H} \neq \varnothing$, that is, $\exists k \in [K^*]$ such that $\widehat{\mathcal{T}} \cap [\tau_k^* - \delta_k, \tau_k^* + \delta_k] = \varnothing$. For such h and a, without loss of generality assume that $\tau_k^* - \widehat{\tau}_a > \delta_k$, it can be observed that $(\tau_k^* - \delta_k, \tau_k^* + \delta_k] \subset (\widehat{\tau}_a, \widehat{\tau}_{a+1}]$ and $\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) \geq 2\delta_k \Delta^2_{(\tau_k^* - \delta_k, \tau_k^* + \delta_k]} = \delta_k \Delta_k^2/2 = 2^{-1}\widetilde{C} s \log(p \vee n)$.

To move further, we need the following definitions to divide $\mathcal{H}$ into four groups.

Definition 13 (Separability of a point) *For a changepoint estimation τ and the true changepoint set $\mathcal{T}^*$, let $u = k^*_{\tau,-}$ and $v = k^*_{\tau,+}$. We say that τ is separable from the left if $\tau - \tau_u^* > \delta_u \vee \delta_{\mathsf{m}}$ and separable from the right if $\tau_v^* - \tau > \delta_v \vee \delta_{\mathsf{m}}$. Otherwise, τ is inseparable from the left (right).*

Definition 14 (Separability of an interval) *For the intervals $(\tau_l, \tau_r] \in \mathcal{H}$, we make the following definitions,*

$\mathcal{H}_1$: $(\tau_l, \tau_r] \in (0, n]$ *is separable if τ_l is separable from the right and τ_r is separable from the left.*

$\mathcal{H}_2$: $(\tau_l, \tau_r] \in (0, n]$ *is left-separable if τ_l is separable from the right and τ_r is inseparable from the left.*

$\mathcal{H}_3$: $(\tau_l, \tau_r] \in (0, n]$ *is right-separable if τ_l is inseparable from the right and τ_r is separable from the left.*

$\mathcal{H}_4$: $(\tau_l, \tau_r] \in (0, n]$ *is inseparable if τ_l is inseparable from the right and τ_r is inseparable from the left.*

Now the sub-intervals in $\mathcal{H}$ have been classified into four groups $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3 \cup \mathcal{H}_4$. We will show that $\mathcal{H} = \varnothing$ by emptying these groups.

Case 1: $\mathcal{H}_1 = \varnothing$

For $(\widehat{\tau}_a, \widehat{\tau}_{a+1}] \in \mathcal{H}_1$, let $h = k^*_{\widehat{\tau}_a,+}$. Denote $\mathcal{T}_a = \{\tau_h^*, \ldots, \tau_{h+t}^*\} = \mathcal{T}^* \cap (\widehat{\tau}_a, \widehat{\tau}_{a+1})$. Let $\widetilde{\mathcal{T}} = \widehat{\mathcal{T}} \cup \mathcal{T}_a$. Since $\gamma = C_\gamma s \log(p \vee n)$,

$$\begin{aligned}
\mathcal{L}(\widehat{\mathcal{T}}) - \mathcal{L}(\widetilde{\mathcal{T}}) = {} & \mathcal{L}_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]} - \left[\mathcal{L}_{(\widehat{\tau}_a, \tau_h^*]} + \mathcal{L}_{(\tau_{h+t}^*, \widehat{\tau}_{a+1}]} + \sum_{j=h}^{h+t-1} \mathcal{L}_{(\tau_j^*, \tau_{j+1}^*]} + (t+1)\gamma\right] \\
> & (1 - C_{8.3})\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) - (t+2)C_{8.1} s \log(p \vee n) - (t+1)\gamma \\
= & (1 - C_{8.3}) \sum_{i \in (\widehat{\tau}_a, \widehat{\tau}_{a+1}]} \|\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}^\circ_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}\|_\Sigma^2 - [(t+2)C_{8.1} + (t+1)C_\gamma] s \log(p \vee n) \\
\geq & \left[(1 - C_{8.3})(t+1)2^{-1}\widetilde{C} - (t+2)C_{8.1} - (t+1)C_\gamma\right] s \log(p \vee n) > 0,
\end{aligned}$$

provided that $\widetilde{C} \geq 2(1 - C_{8.3})^{-1}(2C_{8.1} + C_\gamma)$. Therefore $\mathcal{H}_1 = \varnothing$.

Case 2: $\mathcal{H}_2 = \mathcal{H}_3 = \varnothing$

Without loss of generality, by the symmetry of $\mathcal{H}_2$ and $\mathcal{H}_3$, we only show that $\mathcal{H}_3 = \varnothing$. If the claim does not hold, one can choose $(\widehat{\tau}_a, \widehat{\tau}_{a+1}] \in \mathcal{H}_3$ to be the leftmost one. Hence $\widehat{\tau}_a$ must be separable from the left by Condition 1. Since $\mathcal{H}_1 = \varnothing$ and $(\widehat{\tau}_a, \widehat{\tau}_{a+1}]$ is the leftmost interval in $\mathcal{H}_3$, one obtains $(\widehat{\tau}_{a-1}, \widehat{\tau}_a] \notin \mathcal{H}$. Denote $h = k^*_{\widehat{\tau}_a,+}$ and $\mathcal{T}_a = \mathcal{T}^* \cap (\widehat{\tau}_a + \delta_{\mathsf{m}}, \widehat{\tau}_{a+1} - \delta_{\mathsf{m}}) = \{\tau^*_{h+1}, \ldots, \tau^*_{h+t}\}$ ($t = 0$ if $\mathcal{T}_a = \varnothing$). Let $\widetilde{\mathcal{T}} = (\widehat{\mathcal{T}} \setminus \widehat{\tau}_a) \cup \tau^*_h \cup \mathcal{T}_a = (\widehat{\mathcal{T}} \setminus \widehat{\tau}_a) \cup \{\tau^*_j\}_{j=h}^{h+t}$.

$$\begin{aligned}\mathcal{L}(\widehat{\mathcal{T}}) - \mathcal{L}(\widetilde{\mathcal{T}}) =& \mathcal{L}_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]} + \big(\mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]}\big) - \Big[\sum_{j=h}^{h+t-1} \mathcal{L}_{(\tau^*_j, \tau^*_{j+1}]} + \mathcal{L}_{(\tau^*_{h+t}, \widehat{\tau}_{a+1}]} + t\gamma\Big] \\ >& (1 - C_{8.3})\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) - [(t+1)C_{8.1} + tC_\gamma] s \log(p \vee n) \\ &+ \Big(\sum_{i \in (\widehat{\tau}_a, \tau^*_h]} \epsilon_i^2 + \mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]}\Big). \end{aligned} \tag{12}$$

Since $(\widehat{\tau}_{a-1}, \widehat{\tau}_a] \notin \mathcal{H}$ and $0 < \widehat{\tau}_a - \tau^*_h < \delta_{\mathsf{m}}$, one must obtain that either $(\widehat{\tau}_{a-1}, \widehat{\tau}_a) \cap \mathcal{T}^* = \varnothing$ or $0 < \tau^*_{h-1} - \widehat{\tau}_{a-1} < \delta_{h-1} = \widetilde{C}\Delta^{-2}_{h-1} s \log(p \vee n)$.

For the first scenario, under $\mathbb{G}_1$,

$$\Big|\sum_{i \in (\widehat{\tau}_a, \tau^*_h]} \epsilon_i^2 + \mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]}\Big| \le 2C_{8.1} s \log(p \vee n).$$

Hence,

$$\begin{aligned}\mathcal{L}(\widehat{\mathcal{T}}) - \mathcal{L}(\widetilde{\mathcal{T}}) &> (1 - C_{8.3})\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) - [(t+3)C_{8.1} + tC_\gamma] s \log(p \vee n) \\ &\ge \Big\{(1 - C_{8.3})(t \vee 1)2^{-1}\widetilde{C} - (t+3)C_{8.1} - tC_\gamma\Big\} s \log(p \vee n) > 0,\end{aligned}$$

provided that $\widetilde{C} \ge 2(1 - C_{8.3})^{-1}(4C_{8.1} + C_\gamma)$.

For the second scenario, let $I_1 = (\widehat{\tau}_{a-1}, \widehat{\tau}_a]$ and $I_2 = (\widehat{\tau}_{a-1}, \tau^*_h]$. Firstly, we will bound the gap $\Delta^2_{I_2}|I_2| - \Delta^2_{I_1}|I_1|$. Since $I_1 \subset I_2$, we have $\Delta^2_{I_2}|I_2| - \Delta^2_{I_1}|I_1| \ge 0$.

Denote $d_1 = \tau^*_{h-1} - \widehat{\tau}_{a-1}$, $d_2 = \widehat{\tau}_a - \tau^*_{h-1}$ and $d_3 = \tau^*_h - \widehat{\tau}_a$. Recall that $\Delta_{h-1} = \|\boldsymbol{\theta}^*_h - \boldsymbol{\theta}^*_{h-1}\|_\Sigma$ and the definition of Δ^2_I, we have

$$\Delta^2_{I_2}|I_2| = \frac{d_1(d_2 + d_3)}{d_1 + d_2 + d_3}\Delta^2_{h-1},\ \Delta^2_{I_1}|I_1| = \frac{d_1 d_2}{d_1 + d_2}\Delta^2_{h-1}.$$

It follows that

$$\Delta^2_{I_2}|I_2| - \Delta^2_{I_1}|I_1| = \frac{d_1^2 d_3 \Delta^2_{h-1}}{(d_1 + d_2)(d_1 + d_2 + d_3)} \le \frac{\widetilde{C}^2(\widetilde{C} \vee C_{\mathsf{m}})}{C_{\mathsf{snr}}(C_{\mathsf{snr}} - \widetilde{C} \vee C_{\mathsf{m}})} s \log(p \vee n).$$

where the last inequality is from the conditions $d_1 \le \widetilde{C}\Delta^{-2}_{h-1} s \log(p \vee n)$, $d_3 \le \delta_h \vee \delta_{\mathsf{m}}$ and $d_1 + d_2 + d_3 \ge C_{\mathsf{snr}} s \log(p \vee n)[1 + \Delta^{-2}_{h-1} + \Delta^{-2}_h]$. Denote $C_{m,1} = \widetilde{C}^2(\widetilde{C} \vee C_{\mathsf{m}})/\{C_{\mathsf{snr}}(C_{\mathsf{snr}} - \widetilde{C} \vee C_{\mathsf{m}})\}$.

By $0 < \tau^*_{h-1} - \widehat{\tau}_{a-1} < \delta_{h-1} = \widetilde{C}\Delta^{-2}_{h-1} s\log(p \vee n)$, $\Delta^2_{I_1}|I_1| \le \Delta^2_{I_2}|I_2| \le \widetilde{C}s\log(p \vee n)$. It means that $I_1 \subset I_2 \in E_2^+ \subseteq E_2^-$. Hence by $\mathbb{G}_2^- \cap \mathbb{G}_2^+$,

$$\sum_{i \in (\widehat{\tau}_a, \tau^*_h]} \epsilon_i^2 + \mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]} > -(2C_{8.2} + C_{m,1})s\log(p \vee n). \tag{13}$$

By Eq. (12) and Eq. (13),

$$\begin{aligned}
&\mathcal{L}(\widehat{\mathcal{T}}) - \mathcal{L}(\widetilde{\mathcal{T}}) \\
>&(1 - C_{8.3})\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) - [(t+1)C_{8.1} + tC_\gamma + 2C_{8.2} + C_{m,1}]s\log(p \vee n) \\
\ge&[(1 - C_{8.3})(t \vee 1)2^{-1}\widetilde{C} - (t+1)C_{8.1} - tC_\gamma - 2C_{8.2} - C_{m,1}]s\log(p \vee n) > 0,
\end{aligned}$$

provided that $\widetilde{C} \ge 2(1 - C_{8.3})^{-1}(2C_{8.1} + C_\gamma + 2C_{8.2} + C_{m,1})$. Hence $\mathcal{H}_2 \cup \mathcal{H}_3 = \varnothing$.

Case 3: $\mathcal{H}_4 = \varnothing$

Similar to Case 2, let $(\widehat{\tau}_a, \widehat{\tau}_{a+1}] \in \mathcal{H}_4$, then $\widehat{\tau}_a$ is separable from the left and $\widehat{\tau}_{a+1}$ is separable from the right. By the fact that $\mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3 = \varnothing$, we also obtain $(\widehat{\tau}_{a-1}, \widehat{\tau}_a] \notin \mathcal{H}$ and $(\widehat{\tau}_{a+1}, \widehat{\tau}_{a+2}] \notin \mathcal{H}$. Let $h = k^*_{\widehat{\tau}_a,+}$ and $h + t = k^*_{\widehat{\tau}_{a+1},-}$. Denote $\mathcal{T}_a = \{\tau^*_h, \ldots, \tau^*_{h+t}\}$ and $\widetilde{\mathcal{T}} = (\widehat{\mathcal{T}} \setminus \{\widehat{\tau}_a, \widehat{\tau}_{a+1}\} \cup \mathcal{T}_a$. We have

$$\begin{aligned}
\mathcal{L}(\widehat{\mathcal{T}}) - \mathcal{L}(\widetilde{\mathcal{T}}) =&\mathcal{L}_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]} + [\mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} + \mathcal{L}_{(\widehat{\tau}_{a+1}, \widehat{\tau}_{a+2}]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]} - \mathcal{L}_{(\tau^*_{h+t}, \widehat{\tau}_{a+2}]}] \\
&- \sum_{j=h}^{h+t-1} \mathcal{L}_{(\tau^*_j, \tau^*_{j+1}]} - (t-1)\gamma \\
>&(1 - C_{8.3})\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) - (tC_{8.1} + (t-1)C_\gamma)s\log(p \vee n) \\
&+[\sum_{i \in (\widehat{\tau}_a, \tau^*_h] \cup (\tau^*_{h+1}, \widehat{\tau}_{a+1}]} \epsilon_i^2 + \mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} + \mathcal{L}_{(\widehat{\tau}_{a+1}, \widehat{\tau}_{a+2}]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]} - \mathcal{L}_{(\tau^*_{h+t}, \widehat{\tau}_{a+2}]}].
\end{aligned}$$

Following the same discussion in Case 2, that is, Eq. (13), we have

$$\begin{aligned}
&\sum_{i \in (\widehat{\tau}_a, \tau^*_h] \cup (\tau^*_{h+1}, \widehat{\tau}_{a+1}]} \epsilon_i^2 + \mathcal{L}_{(\widehat{\tau}_{a-1}, \widehat{\tau}_a]} + \mathcal{L}_{(\widehat{\tau}_{a+1}, \widehat{\tau}_{a+2}]} - \mathcal{L}_{(\widehat{\tau}_{a-1}, \tau^*_h]} - \mathcal{L}_{(\tau^*_{h+t}, \widehat{\tau}_{a+2}]} \\
>& - (4C_{8.2} + 2C_{m,1})s\log(p \vee n).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathcal{L}(\widehat{\mathcal{T}}) - \mathcal{L}(\widetilde{\mathcal{T}}) \\
>&(1 - C_{8.3})\Delta^2_{(\widehat{\tau}_a, \widehat{\tau}_{a+1}]}(\widehat{\tau}_{a+1} - \widehat{\tau}_a) - [tC_{8.1} + (t-1)C_\gamma + 4C_{8.2} + 2C_{m,1}]s\log(p \vee n) \\
\ge&\Big\{(1 - C_{8.3})[(t-1) \vee 1]2^{-1}\widetilde{C} - tC_{8.1} - (t-1)C_\gamma - 4C_{8.2} - 2C_{m,1}\Big\}s\log(p \vee n) \ge 0,
\end{aligned}$$

provided that $\widetilde{C} \ge 2(1 - C_{8.3})^{-1}(2C_{8.1} + C_\gamma + 4C_{8.2} + 2C_{m,1})$.

In summary, we obtain $\mathcal{H} = \varnothing$ provided that $\widetilde{C} \ge 2(1 - C_{8.3})^{-1}(2C_{8.1} + C_\gamma + 4C_{8.2} + 2C_{m,1})$. Hence $\max_{1 \le j \le K^*} \min_{1 \le k \le \widehat{K}} \Delta^2_j |\tau^*_j - \widehat{\tau}_k| \le \widetilde{C}s\log(p \vee n)$. It also implies that $\widehat{K} \ge K^*$.

It remains to show that $\widehat{K} \le K^*$. Otherwise, assume that $\widehat{K} > K^*$. Then there must be $j \in [0, K^*]$ and $k \in [1, \widehat{K}]$ such that $\tau_j^* - \delta_j \le \widehat{\tau}_{k-1} < \widehat{\tau}_k < \widehat{\tau}_{k+1} \le \tau_{j+1}^* + \delta_{j+1}$. Similar to the decomposition of $\mathcal{H}$, we can also divide it into four groups.

$\mathcal{G}_1 :\ \tau_j^* \le \widehat{\tau}_{k-1} < \widehat{\tau}_k < \widehat{\tau}_{k+1} \le \tau_{j+1}^*$.

$\mathcal{G}_2 :\ \tau_j^* - \delta_j \le \widehat{\tau}_{k-1} < \tau_j^*$ and $\tau_j^* \le \widehat{\tau}_k < \widehat{\tau}_{k+1} \le \tau_{j+1}^*$.

$\mathcal{G}_3 :\ \tau_j^* \le \widehat{\tau}_{k-1} < \widehat{\tau}_k \le \tau_{j+1}^*$ and $\tau_{j+1}^* < \widehat{\tau}_{k+1} \le \tau_{j+1}^* + \delta_{j+1}$.

$\mathcal{G}_4 :\ \tau_j^* - \delta_j \le \widehat{\tau}_{k-1} < \tau_j^* \le \widehat{\tau}_k \le \tau_{j+1}^* < \widehat{\tau}_{k+1} \le \tau_{j+1}^* + \delta_{j+1}$.

Case 1: $\mathcal{G}_1 = \varnothing$

Let $\widetilde{\mathcal{T}} = \widehat{\mathcal{T}} \setminus \{\widehat{\tau}_k\}$. We have

$$\begin{aligned}\mathcal{L}(\widetilde{\mathcal{T}}) - \mathcal{L}(\widehat{\mathcal{T}}) &= \mathcal{L}_{(\widehat{\tau}_{k-1},\widehat{\tau}_{k+1}]} - \mathcal{L}_{(\widehat{\tau}_{k-1},\widehat{\tau}_{k}]} - \mathcal{L}_{(\widehat{\tau}_{k},\widehat{\tau}_{k+1}]} - \gamma \\ &< (3C_{8.1} - C_\gamma) s \log(p \vee n) \le 0,\end{aligned}$$

provided that $C_\gamma \ge 3C_{8.1}$.

Case 2: $\mathcal{G}_2 \cup \mathcal{G}_3 = \varnothing$

We will show that $\mathcal{G}_2 = \varnothing$ because the proof for $\mathcal{G}_3 = \varnothing$ is the same by symmetry. Assume that the pair (j, k) is the leftmost one that satisfies $\mathcal{G}_2$. It implies that $\widehat{\tau}_{k-2} \in [\tau_{j-1}^* - \delta_{j-1}, \tau_{j-1}^* + \delta_{j-1}]$. Otherwise assume $\widehat{\tau}_{k-2} > \tau_{j-1}^* + \delta_{j-1}$. Since $\max_{1\le j\le K^*} \min_{1\le k\le \widehat{K}} \Delta_j^2 |\tau_j^* - \widehat{\tau}_k| \le \widetilde{C} s \log(p \vee n)$, there must be $\widehat{\tau}_{k-h} \in [\tau_{j-1}^* - \delta_{j-1}, \tau_{j-1}^* + \delta_{j-1}]$ for some $h > 2$. It contradicts the fact that $\mathcal{G}_1 = \varnothing$ and the choice of k.

Let $\widetilde{\mathcal{T}} = \{\tau_j^*\} \cup \widehat{\mathcal{T}} \setminus \{\widehat{\tau}_{k-1}, \widehat{\tau}_k\}$. Note that $(\widehat{\tau}_{k-2}, \widehat{\tau}_{k-1}], (\widehat{\tau}_{k-1}, \widehat{\tau}_k] \in E_2^-$ and $(\widehat{\tau}_{k-2}, \tau_j^*] \in E_2^+$, we have

$$\begin{aligned}\mathcal{L}(\widetilde{\mathcal{T}}) - \mathcal{L}(\widehat{\mathcal{T}}) &= \mathcal{L}_{(\widehat{\tau}_{k-2},\tau_j^*]} + \mathcal{L}_{(\tau_j^*,\widehat{\tau}_{k+1}]} - \Big[\sum_{t=k-2}^{k} \mathcal{L}_{(\widehat{\tau}_t,\widehat{\tau}_{t+1}]} + \gamma\Big] \\ &= [\mathcal{L}_{(\widehat{\tau}_{k-2},\tau_j^*]} - \mathcal{L}_{(\widehat{\tau}_{k-2},\widehat{\tau}_{k-1}]}] + \mathcal{L}_{(\tau_j^*,\widehat{\tau}_{k+1}]} - \Big[\sum_{t=k-1}^{k} \mathcal{L}_{(\widehat{\tau}_t,\widehat{\tau}_{t+1}]} + \gamma\Big] \\ &< \Big[\mathcal{L}_{(\widehat{\tau}_{k-2},\tau_j^*]} - \mathcal{L}_{(\widehat{\tau}_{k-2},\widehat{\tau}_{k-1}]} - \sum_{i\in(\widehat{\tau}_{k-1},\tau_j^*]} \epsilon_i^2\Big] + (2C_{8.1} + C_{8.2} - C_\gamma) s \log(p \vee n) \\ &\le (2C_{8.1} + 3C_{8.2} + C_{m,1} - C_\gamma) s \log(p \vee n) \le 0,\end{aligned}$$

provided $C_\gamma \ge 2C_{8.1} + 3C_{8.2} + C_{m,1}$. The second last inequality is from Eq. (13).

Case 3: $\mathcal{G}_4 = \varnothing$

Now $\mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3 = \varnothing$. Assume that $\{\widehat{\tau}_{k-1}, \widehat{\tau}_k, \widehat{\tau}_{k+1}\}$ satisfies $\mathcal{G}_4$. Similar to the analysis of $\mathcal{G}_2 = \varnothing$, we have $\widehat{\tau}_{k-2} \in [\tau_{j-1}^* - \delta_{j-1}, \tau_{j-1}^* + \delta_{j-1}]$ and $\widehat{\tau}_{k+2} \in [\tau_{j+1}^* + \delta_{j+1}, \tau_{j+1}^* + \delta_{j+1}]$. Follow the same arguments in the proof for $\mathcal{H}_4 = \varnothing$, we can set $\widetilde{\mathcal{T}} = \{\tau_j^*, \tau_{j+1}^*\} \cup \widehat{\mathcal{T}} \backslash \{\widehat{\tau}_{k-1}, \widehat{\tau}_k, \widehat{\tau}_{k+1}\}$.

$$
\begin{aligned}
&\mathcal{L}(\widetilde{\mathcal{T}}) - \mathcal{L}(\widehat{\mathcal{T}}) \\
=&\mathcal{L}_{(\widehat{\tau}_{k-2}, \tau_j^*]} + \mathcal{L}_{(\tau_j^*, \tau_{j+1}^*]} + \mathcal{L}_{(\tau_{j+1}^*, \widehat{\tau}_{k+2}]} - \Big[\sum_{t=k-2}^{k+1} \mathcal{L}_{(\widehat{\tau}_t, \widehat{\tau}_{t+1}]} + \gamma \Big] \\
=&[\mathcal{L}_{(\widehat{\tau}_{k-2}, \tau_j^*]} - \mathcal{L}_{(\widehat{\tau}_{k-2}, \widehat{\tau}_{k-1}]} + \mathcal{L}_{(\tau_{j+1}^*, \widehat{\tau}_{k+2}]} - \mathcal{L}_{(\widehat{\tau}_{k+1}, \widehat{\tau}_{k+2}]}] \\
&+ \mathcal{L}_{(\tau_j^*, \tau_{j+1}^*]} - \Big[\sum_{t=k-1}^{k} \mathcal{L}_{(\widehat{\tau}_t, \widehat{\tau}_{t+1}]} + \gamma \Big] \\
<&\Big[\mathcal{L}_{(\widehat{\tau}_{k-2}, \tau_j^*]} - \mathcal{L}_{(\widehat{\tau}_{k-2}, \widehat{\tau}_{k-1}]} + \mathcal{L}_{(\tau_{j+1}^*, \widehat{\tau}_{k+2}]} - \mathcal{L}_{(\widehat{\tau}_{k+1}, \widehat{\tau}_{k+2}]} - \sum_{i \in (\widehat{\tau}_{k-1}, \tau_j^*] \cup (\tau_{j+1}^*, \widehat{\tau}_{k+1}]} \epsilon_i^2 \Big] \\
&+ (C_{8.1} + 2C_{8.2} - C_\gamma) s \log(p \vee n) \\
\leq&(C_{8.1} + 6C_{8.2} + 2C_{m,1} - C_\gamma) s \log(p \vee n) \leq 0,
\end{aligned}
$$

provided $C_\gamma \geq C_{8.1} + 6C_{8.2} + 2C_{m,1}$. The second last inequality is from Eq. (13).

Combining the proof in the $\mathcal{H}$ and $\mathcal{G}$ parts, we can determine the two constants by solving the following inequalities,

$$
\begin{cases}
C_\gamma \geq C_{8.1} + 6C_{8.2} + 2C_{m,1} \\
\widetilde{C} \geq 2(1 - C_{8.3})^{-1}(2C_{8.1} + C_\gamma + 4C_{8.2} + 2C_{m,1})
\end{cases}
\tag{14}
$$

Since C_{snr} and C_{m} are sufficiently large, we have $C_{m,1} = \widetilde{C}^2(\widetilde{C} \vee C_{\mathsf{m}})/\{C_{\mathsf{snr}}(C_{\mathsf{snr}} - \widetilde{C} \vee C_{\mathsf{m}})\} = \widetilde{C}^2 C_{\mathsf{m}}/\{C_{\mathsf{snr}}(C_{\mathsf{snr}} - C_{\mathsf{m}})\}$. Let $C_\gamma = C_{8.1} + 6C_{8.2} + 2C_{m,1}$, we obtain the following inequality w.r.t. $\widetilde{C}$,

$$
\frac{C_{\mathsf{m}} \widetilde{C}^2}{C_{\mathsf{snr}}(C_{\mathsf{snr}} - C_{\mathsf{m}})} - \frac{(1 - C_{8.3})\widetilde{C}}{2} + 3C_{8.1} + 10C_{8.2} \geq 0. \tag{15}
$$

Treat it as a quadratic inequality w.r.t. $\widetilde{C}$, we can figure out that there exist solutions if and only if $C_{\mathsf{snr}}(C_{\mathsf{snr}} - C_{\mathsf{m}}) \geq 16(1 - C_{8.3})^{-2} C_{\mathsf{m}}(3C_{8.1} + 10C_{8.2})$. And by solving Eq. (15), we have

$$
\widetilde{C} = 2\{a - (a^2 - b)^{\frac{1}{2}}\} \leq \frac{b}{(a^2 - b)^{\frac{1}{2}}} \leq \frac{2b}{a} = 4(1 - C_{8.3})^{-1}(3C_{8.1} + 10C_{8.2}), \tag{16}
$$

satisfies Eq. (14) with $a = (1 - C_{8.3}) C_{\mathsf{snr}}(C_{\mathsf{snr}} - C_{\mathsf{m}})/(8C_{\mathsf{m}})$, $b = (3C_{8.1} + 10C_{8.2}) C_{\mathsf{snr}}(C_{\mathsf{snr}} - C_{\mathsf{m}})/(4C_{\mathsf{m}})$. The last inequality in Eq. (16) holds provided that C_{snr} is sufficiently large so that $b \leq 3a^2/4$.

It follows that $\widehat{\mathcal{T}} = \widetilde{\mathcal{T}}$ provided that Eq. (14) holds. Finally, we obtain

$$
\widehat{K} = K^*; \max_{1 \leq k \leq K^*} \min_{1 \leq j \leq \widehat{K}} \Delta_k^2 |\tau_k^* - \widehat{\tau}_j| \leq \widetilde{C} s \log(p \vee n),
$$

with any $\widetilde{C} \geq 4(1 - C_{8.3})^{-1}(3C_{8.1} + 10C_{8.2})$.

Appendix C. Proof of Theorem 9

For a interval I, denote the sparsity constant $s_I = s \vee |\{1 \le j \le p : \exists i \in I, \boldsymbol{\theta}^\circ_{i,j} \neq 0\}|$. Observe that $s_I \le (1 \vee |\mathcal{T}^* \cap I|) \times s$. Define $\Delta_{I,q,\boldsymbol{\theta}} = (|I|^{-1}\sum_{i\in I}\|\boldsymbol{\theta}^\circ_i - \boldsymbol{\theta}\|^q_\Sigma)^{1/q}$ and let $\Delta_{I,\boldsymbol{\theta}} = \Delta_{I,2,\boldsymbol{\theta}}$ be the root average square variation of I and $\Delta_{I,\infty,\boldsymbol{\theta}} = \max_{i\in I}\|\boldsymbol{\theta}^\circ_i - \boldsymbol{\theta}\|_\Sigma$ be the maximum variation of I. For simplicity, denote $\Delta_{I,q} = \Delta_{I,q,\boldsymbol{\theta}^\circ_I}$, $\Delta_I = \Delta_{I,2}$ and $\Delta_{I,\infty} = \Delta_{I,\infty,\boldsymbol{\theta}^\circ_I}$.

As stated in Lemma 8, to show that the bound of localization error in Theorem 9 holds, we only need to certify that the event $\mathbb{G}$ holds with high probability for both the original full model-fitting approach and the Reliever approach with suitable constants. These two claims are shown in Corollary 22 and Corollary 24, respectively. Finally, the L_2 error bound of the parameter estimation follows the oracle inequality of lasso.

This section is organized as follows. In Section C.1, we introduce several useful non-asymptotic probability bounds, including the oracle inequality of lasso with heterogeneous data. In Section C.2 and C.3, we show that $\mathbb{G}$ holds with high probability for the two approaches correspondingly. All the proofs are relegated to the last part.

C.1 Deviation Bounds via the Bernstein's Inequality

The deviation bounds in this subsection will rely on the following Bernstein's inequality.

Lemma 15 (Bernstein's inequality) *Let $\{X_i\}_{i=1}^n$ be independent, mean-zero random variables with sub-exponential tails. For every $t > 0$, we have*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^n X_i\Big| > t\Big\} \le 2\exp\Big[-c_b\Big(\frac{t^2}{\sum_{i=1}^n \|X_i\|^2_{\Psi_1}} \wedge \frac{t}{\max_i \|X_i\|_{\Psi_1}}\Big)\Big],$$

where $c_b > 0$ is an absolute constant. Choose

$$t = \frac{C_u}{c_b}\Big[\Big(\sum_{i\in[n]}\|X_i\|^2_{\Psi_1}A_{n,p,s}\Big)^{\frac{1}{2}} \vee (\max_i \|X_i\|_{\Psi_1}A_{n,p,s})\Big]$$

with $C_u \ge c_b$, we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^n X_i\Big| > t\Big\} \le 2\exp\{-C_u A_{n,p,s}\}.$$

Here $A_{n,p,s}$ is a diverging sequence. For instance, $A_{n,p,s} = \log(p\vee n)$ and $A_{n,p,s} = s\log(p\vee n)$.

Lemma 16 (Uniform restricted eigenvalue condition) *Assume that Condition 3(a) holds. For any interval $I \subset (0, n]$, denote $\widehat{\Sigma}_I = |I|^{-1}\sum_{i\in I}\mathbf{x}_i\mathbf{x}_i^\top$. Uniformly for all intervals $I \subset (0, n]$ such that $|I| \ge s_I \log(p \vee n)$, with probability at least $1 - \exp\{-C_{u,1}\log(p \vee n)\}$,*

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \ge \|\mathbf{v}\|^2_\Sigma - C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\Big(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2\Big), \forall \mathbf{v} \in \mathbb{R}^p,$$

and

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \le \|\mathbf{v}\|^2_\Sigma + C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\Big(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2\Big), \forall \mathbf{v} \in \mathbb{R}^p,$$

where $C_{u,1}$ and $C_{u,2}$ are two universal constants. Furthermore, let $|I| \geq C_{\mathsf{re}} s_I \log(p \vee n)$ with a sufficiently large constant $C_{\mathsf{re}} \geq 1 \vee (34 C_{u,2} C_x^2 \sigma_x^2 / \underline{\kappa})^2$. For any support set $\mathcal{S} \in [p]$ with $|\mathcal{S}| \leq s_I$ and $\mathbf{v} \in \mathbb{R}^p$ such that $\|\mathbf{v}_{\mathcal{S}^{\complement}}\|_1 \leq 3\|\mathbf{v}_{\mathcal{S}}\|_1$, under the same event above,

$$\frac{1}{2}\|\mathbf{v}\|_\Sigma^2 \leq (1 - C_{16})\|\mathbf{v}\|_\Sigma^2 \leq \mathbf{v}^\top \widehat{\Sigma}_I \mathbf{v} \leq (1 + C_{16})\|\mathbf{v}\|_\Sigma^2 \leq \frac{3}{2}\|\mathbf{v}\|_\Sigma^2,$$

$$\frac{\underline{\kappa}}{2}\|\mathbf{v}\|_2^2 \leq (1 - C_{16})\underline{\kappa}\|\mathbf{v}\|_2^2 \leq \mathbf{v}^\top \widehat{\Sigma}_I \mathbf{v} \leq (\sigma_x^2 + C_{16}\underline{\kappa})\|\mathbf{v}\|_2^2 \leq \Big(\sigma_x^2 + \frac{\underline{\kappa}}{2}\Big)\|\mathbf{v}\|_2^2,$$

where $C_{16} = 17 C_{u,2} C_x^2 \sigma_x^2 \kappa^{-1} C_{\mathsf{re}}^{-\frac{1}{2}}$.

Proof [Proof of Lemma 16.] For sparsity level s, denote $\mathcal{A}(s) = \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v}\|_2 = 1, |\operatorname{supp}(\mathbf{v})| \leq s\}$. We will first show that with high probability, $\sup_{\mathbf{v} \in \mathcal{A}(2s_I)} |\mathbf{v}^\top(\widehat{\Sigma}_I - \Sigma)\mathbf{v}| = O(C_x^2 \{|I|^{-1} s_I \log(p \vee n)\}^{\frac{1}{2}})$ uniformly for all intervals $\{I\}$ such that $\sup_{|I| \geq s_I \log(p \vee n)}$. Then the result follows from Lemma 12 in Loh and Wainwright (2012).

For a fixed interval I, let $\mathbf{D} = \widehat{\Sigma}_I - \Sigma$. For any $\mathcal{U} \subset [p]$ and $|\mathcal{U}| = 2s_I$, let $\mathbf{D}_{\mathcal{U}} \in \mathbb{R}^{2s_I \times 2s_I}$ be the sub-matrix of $\mathbf{D}$ with $\mathcal{U}$ being the set of row and column indices. Let $\mathcal{B}_{\mathcal{U}} = \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v}\|_2 = 1, \operatorname{supp}(\mathbf{v}) = \mathcal{U}\}$. There is a 4^{-1}-net $\mathcal{N}_{\mathcal{U}} \subseteq \mathcal{B}_{\mathcal{U}}$ of $\mathcal{B}_{\mathcal{U}}$ with cardinality $|\mathcal{N}_{\mathcal{U}}| \leq 9^{2s_I}$. For any $\mathbf{v} \in \mathcal{B}_{\mathcal{U}} - \mathcal{N}_{\mathcal{U}}$, there is $\mathbf{u} \in \mathcal{N}_{\mathcal{U}}$ such that $\|\mathbf{v} - \mathbf{u}\|_2 \leq 4^{-1}$ and $\|\mathbf{v} - \mathbf{u}\|_2^{-1}(\mathbf{v} - \mathbf{u}) \in \mathcal{S}_{\mathcal{U}}$. Therefore,

$$|\mathbf{v}^\top \mathbf{D} \mathbf{v} - \mathbf{u}^\top \mathbf{D} \mathbf{u}| = |\mathbf{v}^\top \mathbf{D}(\mathbf{v} - \mathbf{u}) + \mathbf{u}^\top \mathbf{D}(\mathbf{v} - \mathbf{u})| \leq 2\|\mathbf{D}_{\mathcal{U}}\|_{\mathsf{op}}\|\mathbf{v} - \mathbf{u}\|_2 \leq \frac{1}{2}\|\mathbf{D}_{\mathcal{U}}\|_{\mathsf{op}}.$$

By the definition of $\mathbf{D}_{\mathcal{U}}$, we have $\|\mathbf{D}_{\mathcal{U}}\|_{\mathsf{op}} = \sup_{\mathbf{v} \in \mathcal{B}_{\mathcal{U}}} |\mathbf{v}^\top \mathbf{D} \mathbf{v}|$. Hence

$$\sup_{\mathbf{v} \in \mathcal{B}_{\mathcal{U}}} |\mathbf{v}^\top \mathbf{D} \mathbf{v}| \leq 2 \sup_{\mathbf{v} \in \mathcal{N}_{\mathcal{U}}} |\mathbf{v}^\top \mathbf{D} \mathbf{v}|.$$

Let $\mathcal{N} = \cup_{|\mathcal{U}| = 2s_I} \mathcal{N}_{\mathcal{U}}$. We have $|\mathcal{N}| \leq \binom{p}{2s_I} 9^{2s_I} \leq (9p)^{2s_I}$ and $\mathcal{N}$ is the 4^{-1}-net of $\mathcal{A}(2s_I)$ because $\mathcal{A}(2s_I) = \cup_{|\mathcal{U}| = 2s_I} \mathcal{B}_{\mathcal{U}}$. Also,

$$\sup_{\mathbf{v} \in \mathcal{A}(2s_I)} |\mathbf{v}^\top \mathbf{D} \mathbf{v}| \leq 2 \sup_{\mathbf{v} \in \mathcal{N}} |\mathbf{v}^\top \mathbf{D} \mathbf{v}|.$$

For a fixed $\mathbf{v} \in \mathcal{A}(2s_I)$, by the Bernstein's inequality (Lemma 15),

$$\mathbb{P}\bigg[|\mathbf{v}^\top \mathbf{D} \mathbf{v}| > \frac{t}{|I|}\bigg] \leq 2\exp\bigg[-c_b\Big(\frac{t^2}{C_x^4 \sigma_x^4 |I|} \wedge \frac{t}{C_x^2 \sigma_x^2}\Big)\bigg].$$

Set $t = c_b^{-1} C_u C_x^2 \sigma_x^2 \{|I| s_I \log(p \vee n)\}^{\frac{1}{2}}$ with $C_u \geq c_b$ be a sufficiently large constant. With probability at least $1 - \exp\{-C_u s_I \log(p \vee n)\}$,

$$|\mathbf{v}^\top \mathbf{D} \mathbf{v}| \leq c_b^{-1} C_u C_x^2 \sigma_x^2 \Big\{\frac{s_I \log(p \vee n)}{|I|}\Big\}^{\frac{1}{2}}.$$

Note that $s_I \geq s$ by its definition. By taking the union bound over $\mathbf{v} \in \mathcal{N}$ and $\{I : |I| \geq s_I \log(p \vee n)\}$, with probability at least $1 - n^2(9p)^{2s}\exp\{-C_u s \log(p \vee n)\} \geq 1 - \exp\{-C_{u,1}\log(p \vee n)\}$ for some $C_{u,1} > 0$, uniformly for all I such that $|I| \geq s_I \log(p \vee n)$,

$$\sup_{\mathbf{v} \in \mathcal{A}(2s_I)} |\mathbf{v}^\top(\widehat{\Sigma}_I - \Sigma)\mathbf{v}| \leq 2 \sup_{\mathbf{v} \in \mathcal{N}} |\mathbf{v}^\top(\widehat{\Sigma}_I - \Sigma)\mathbf{v}| \leq 2 c_b^{-1} C_u C_x^2 \sigma_x^2 \Big\{\frac{s_I \log(p \vee n)}{|I|}\Big\}^{\frac{1}{2}}.$$

By Lemma 12 in Loh and Wainwright (2012), under the above event,

$$|\mathbf{v}^\top(\widehat{\Sigma}_I - \Sigma)\mathbf{v}| \le 54c_b^{-1}C_uC_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2),$$

for all $\mathbf{v} \in \mathbb{R}^p$ and all intervals in $\{I : |I| \ge s_I\log(p\vee n)\}$. Let $C_{u,2} = 54c_b^{-1}C_u$. With probability at least $1-\exp\{-C_{u,1}\log(p\vee n)\}$, for all $\mathbf{v}\in\mathbb{R}^p$ and all I such that $|I| \ge s_I\log(p\vee n)$,

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \ge \|\mathbf{v}\|_\Sigma^2 - C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\Big(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2\Big), \tag{17}$$

and

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \le \|\mathbf{v}\|_\Sigma^2 + C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\Big(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2\Big). \tag{18}$$

If there exists a support set $\mathcal{S} \in [p]$ with $|\mathcal{S}| \le s_I$ so that $\|\mathbf{v}_{\mathcal{S}^\complement}\|_1 \le 3\|\mathbf{v}_{\mathcal{S}}\|_1$, we have $\|\mathbf{v}\|_1 \le 4\|\mathbf{v}_{\mathcal{S}}\|_1 \le 4s_I^{\frac{1}{2}}\|\mathbf{v}_{\mathcal{S}}\|_2 \le 4s_I^{\frac{1}{2}}\|\mathbf{v}\|_2$. By Eq. (17)–(18) and the inequality that $\frac{1}{s_I}\|\mathbf{v}\|_1^2 \le 16\|\mathbf{v}\|_2^2$, we have

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \ge \|\mathbf{v}\|_\Sigma^2 - 17C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\|\mathbf{v}\|_2^2,$$

and

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \le \|\mathbf{v}\|_\Sigma^2 + 17C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\|\mathbf{v}\|_2^2.$$

Because $|I| \ge C_{\mathsf{re}}s_I\log(p\vee n)$ and $C_{\mathsf{re}} \ge 1 \vee (34\underline{\kappa}^{-1}C_{u,2}C_x^2\sigma_x^2)^2$, we have

$$17C_{u,2}C_x^2\sigma_x^2|I|^{-\frac{1}{2}}\{s_I\log(p\vee n)\}^{\frac{1}{2}} \le \frac{\underline{\kappa}}{2}.$$

The last two results in the lemma are due to the inequality $\|\mathbf{v}\|_2^2 \le \underline{\kappa}^{-1}\|\mathbf{v}\|_\Sigma^2$. ■

Lemma 17 *Assume that Condition 3 holds. For interval $I \subset (0,n]$ and a fixed $\boldsymbol{\theta}$, with probability at least $1-n^{-2}\exp\{-C_{u,1}\log(p\vee n)\}$,*

$$\begin{aligned}&\Big\|\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\mathbf{x}_i\}\Big\|_\infty\\ \le& C_{u,2}C_x\sigma_x\bigg\{(C_x^2\Delta_{I,\boldsymbol{\theta}}^2 + C_\epsilon^2) \vee \frac{(C_x^2\Delta_{I,\infty,\boldsymbol{\theta}}^2 + C_\epsilon^2)\log(p\vee n)}{|I|}\bigg\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\end{aligned}$$

where $C_{u,1}$ and $C_{u,2}$ are two universal constants.

Proof [Proof of Lemma 17.] By condition 3, $\mathbb{E}[\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\mathbf{x}_i\}] = \mathbf{0}$. By Condition 3, $\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta})$ is sub-Gaussian with mean zero and Ψ_2-norm $C_x\|\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}\|_\Sigma$

and ϵ_i is sub-Gaussian with mean zero and Ψ_2-norm $\|\epsilon\|_{\Psi_2} = C_\epsilon$. Hence $\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i$ is sub-Gaussian with mean zero and

$$\|\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\|_{\Psi_2} \le (C_x^2\|\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}\|_\Sigma^2 + C_\epsilon^2)^{\frac{1}{2}}.$$

Then $\mathbf{x}_i\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\} - \Sigma(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta})$ is sub-exponential with Ψ_1-norm:

$$\|\mathbf{x}_i\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\} - \Sigma(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta})\|_{\Psi_1} \le \|\mathbf{x}_i\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\}\|_{\Psi_1} \le C_x\sigma_x(C_x^2\|\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}\|_\Sigma^2 + C_\epsilon^2)^{\frac{1}{2}}.$$

By the Bernstein's inequality (Lemma 15), for any given $\mathbf{v} \in \mathbb{S}^{p-1}$,

$$\begin{aligned}&\mathbb{P}\Big\{\Big|\sum_{i\in I}\mathbf{v}^\top\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\mathbf{x}_i\}\Big| > t\Big\}\\ \le& 2\exp\bigg(-\frac{c_b t^2}{C_x^2\sigma_x^2(C_x^2\Delta_{I,\boldsymbol{\theta}}^2 + C_\epsilon^2)|I|} \wedge \frac{c_b t}{C_x\sigma_x(C_x^2\Delta_{I,\infty,\boldsymbol{\theta}}^2 + C_\epsilon^2)^{\frac{1}{2}}}\bigg).\end{aligned}$$

Set $t = c_b^{-1}(C_{u,1}+3)C_x\sigma_x[\{(C_x^2\Delta_{I,\boldsymbol{\theta}}^2 + C_\epsilon^2)|I|\log(p\vee n)\}^{\frac{1}{2}} \vee \{(C_x^2\Delta_{I,\infty,\boldsymbol{\theta}}^2 + C_\epsilon^2)\log^2(p\vee n)\}^{\frac{1}{2}}]$ with any constant $C_{u,1} \ge c_b$. With probability at least $1 - p\exp\{-(C_{u,1}+3)\log(p\vee n)\} \ge 1 - n^{-2}\exp\{-C_{u,1}\log(p\vee n)\}$,

$$\begin{aligned}&\Big\|\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\mathbf{x}_i\}\Big\|_\infty\\ \le& C_{u,2}C_x\sigma_x\Big[\{(C_x^2\Delta_{I,\boldsymbol{\theta}}^2 + C_\epsilon^2)|I|\log(p\vee n)\}^{\frac{1}{2}} \vee \{(C_x^2\Delta_{I,\infty,\boldsymbol{\theta}}^2 + C_\epsilon^2)\log^2(p\vee n)\}^{\frac{1}{2}}\Big],\end{aligned}$$

where $C_{u,2} = c_b^{-1}(C_{u,1}+3)$. ■

Lemma 18 (Oracle inequalities for the parametric estimates) *Assume that Condition 2(a) and Condition 3 hold. For any interval $I \subset (0,n]$, recall that $D_I = [(C_x^2\Delta_I^2 + C_\epsilon^2) \vee \{|I|^{-1}(C_x^2\Delta_{I,\infty}^2 + C_\epsilon^2)\log(p\vee n)\}]^{\frac{1}{2}}$. We have with probability at least $1 - 2\exp\{-C_{u,1}\log(p\vee n)\}$, uniformly for any interval $I \subset (0,n]$ with $|I| \ge C_{\mathsf{re}}s_I\log(p\vee n)$, provided that $\lambda_I = 4C_{u,2}C_x\sigma_x D_I\{|I|\log(p\vee n)\}^{\frac{1}{2}}$, the solution $\widehat{\boldsymbol{\theta}}_I$ satisfies that*

$$\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_2 \le \underline{\kappa}^{-\frac{1}{2}}\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_\Sigma \le C_{18}D_I\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}},$$

$$\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_1 \le C_{18}D_I s_I\Big\{\frac{\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}},$$

where the model-based constant $C_{18} = 12\underline{\kappa}^{-1}C_{u,2}C_x\sigma_x$. Furthermore, let $\mathcal{S}_I$ be the support set of $\boldsymbol{\theta}_I^\circ$, we have $\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I^\complement} - \boldsymbol{\theta}_{I,\mathcal{S}_I^\complement}^\circ\|_1 \le 3\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I} - \boldsymbol{\theta}_{I,\mathcal{S}_I}^\circ\|_1$.

Proof [Proof of Lemma 18.] (Oracle inequality for the mixture of distributions.)

In the following proof, we assume that the inequalities in Lemmas 16–17 hold for all intervals I such that $|I| \ge C_{\mathsf{re}}s_I\log(p\vee n)$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_I^\circ$. The claim holds with a probability lower bound $1 - 2\exp\{-C_{u,1}\log(p\vee n)\}$.

By the definition of $\widehat{\boldsymbol{\theta}}_I$,

$$\begin{aligned}
&\sum_{i\in I}(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\theta}}_I)^2 + \lambda_I \|\widehat{\boldsymbol{\theta}}_I\|_1 = \sum_{i\in I}\{y_i - \mathbf{x}_i^\top \boldsymbol{\theta}_i^\circ + \mathbf{x}^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ) + \mathbf{x}_i^\top(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)\}^2 + \lambda_I \|\widehat{\boldsymbol{\theta}}_I\|_1 \\
=&\sum_{i\in I} \epsilon_i^2 + \{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\}^2 + \{\mathbf{x}_i^\top(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)\}^2 + \lambda_I \|\widehat{\boldsymbol{\theta}}_I\|_1 \\
+&2\sum_{i\in I}\{\epsilon_i \mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ) + \epsilon_i \mathbf{x}_i^\top(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)\} + 2(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)^\top \sum_{i\in I}\mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ) \\
\leq&\sum_{i\in I}(y_i - \mathbf{x}_i^\top \boldsymbol{\theta}_I^\circ)^2 + \lambda_I \|\boldsymbol{\theta}_I^\circ\|_1 = \sum_{i\in I}\{y_i - \mathbf{x}_i^\top \boldsymbol{\theta}_i^\circ + \mathbf{x}^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\}^2 + \lambda_I \|\boldsymbol{\theta}_I^\circ\|_1 \\
=&\sum_{i\in I} \epsilon_i^2 + \{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\}^2 + 2\sum_{i\in I}\epsilon_i \mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ) + \lambda_I \|\boldsymbol{\theta}_I^\circ\|_1.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i\in I}\{\mathbf{x}_i^\top(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ)\}^2 + \lambda_I \|\widehat{\boldsymbol{\theta}}_I\|_1 \\
\leq&2(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ)^\top \sum_{i\in I}\{\epsilon_i \mathbf{x}_i + \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\} + \lambda_I \|\boldsymbol{\theta}_I^\circ\|_1 \leq \lambda_{I,1}\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_1 + \lambda_I \|\boldsymbol{\theta}_I^\circ\|_1,
\end{aligned}$$

where $\lambda_{I,1} = 2\|\sum_{i\in I}\{\epsilon_i \mathbf{x}_i + \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\}\|_\infty$. By Lemma 17,

$$\lambda_{I,1} \leq 2C_{u,2}C_x\sigma_x D_I\{|I|\log(p\vee n)\}^{\frac{1}{2}}.$$

where $D_I = [(C_x^2\Delta_I^2 + C_\epsilon^2) \vee \{|I|^{-1}(C_x^2\Delta_{I,\infty}^2 + C_\epsilon^2)\log(p\vee n)\}]^{\frac{1}{2}}$ for easing the notation.

Since $\sum_{i\in I}\{\mathbf{x}_i^\top(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ)\}^2 \geq 0$, $(\lambda_I - \lambda_{I,1})\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I^\complement} - \boldsymbol{\theta}^\circ_{I,\mathcal{S}_I^\complement}\|_1 \leq (\lambda_I + \lambda_{I,1})\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I} - \boldsymbol{\theta}^\circ_{I,\mathcal{S}_I}\|_1$. Choosing $\lambda_I = 2\lambda_{I,1}$, we have $\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I^\complement} - \boldsymbol{\theta}^\circ_{I,\mathcal{S}_I^\complement}\|_1 \leq 3\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I} - \boldsymbol{\theta}^\circ_{I,\mathcal{S}_I}\|_1$.

Apply Lemma 16, the uniform restricted eigenvalue condition holds for any interval I with $|I| \geq C_{\mathsf{re}} s_I \log(p\vee n)$. Hence $2^{-1}|I|\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_\Sigma^2 \leq \sum_{i\in I}\{\mathbf{x}_i^\top(\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ)\}^2 \leq \lambda_{I,1}\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_1 + \lambda_I\|\boldsymbol{\theta}_I^\circ\|_1 - \lambda_I\|\widehat{\boldsymbol{\theta}}_I\|_1 \leq (\lambda_I + \lambda_{I,1})\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I} - \boldsymbol{\theta}^\circ_{I,\mathcal{S}_I}\|_1 - \lambda_{I,1}\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I^\complement}\|_1 \leq 3\lambda_{I,1}s_I^{\frac{1}{2}}\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_2 \leq 3\lambda_{I,1}s_I^{\frac{1}{2}}\underline{\kappa}^{-\frac{1}{2}}\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_\Sigma$. Therefore,

$$\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_\Sigma \leq \frac{3\lambda_{I,1}\underline{\kappa}^{-\frac{1}{2}}s_I^{\frac{1}{2}}}{2^{-1}|I|} \leq \frac{12C_{u,2}C_x\sigma_x D_I}{\underline{\kappa}^{\frac{1}{2}}}\left\{\frac{s_I\log(p\vee n)}{|I|}\right\}^{\frac{1}{2}}.$$

Similarly, we have

$$\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_2 \leq \frac{3\lambda_{I,1}s_I^{\frac{1}{2}}}{2^{-1}\underline{\kappa}|I|} \leq \frac{12C_{u,2}C_x\sigma_x D_I}{\underline{\kappa}}\left\{\frac{s_I\log(p\vee n)}{|I|}\right\}^{\frac{1}{2}},$$

and

$$\|\widehat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I^\circ\|_1 \leq \frac{3\lambda_{I,1}s_I}{2^{-1}\underline{\kappa}|I|} \leq \frac{12C_{u,2}C_x\sigma_x D_I s_I}{\underline{\kappa}}\left\{\frac{\log(p\vee n)}{|I|}\right\}^{\frac{1}{2}}.$$

■

Lemma 19 *Assume that Condition 3 holds. For interval $I \subset (0,n]$ and a fixed $\boldsymbol{\theta}$, with probability at least $1 - n^{-2}\exp\{-C_{u,1}\log(p \vee n)\}$, uniformly for any sub-interval $I \subset (0,n]$,*

$$\Big|\sum_{i\in I}\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta})\}^2 - \Delta_{I,\boldsymbol{\theta}}^2|I|\Big| \le C_{u,2}C_x^2\bigg\{\Delta_{I,4,\boldsymbol{\theta}}^4 \vee \frac{\Delta_{I,\infty,\boldsymbol{\theta}}^4\log(p\vee n)}{|I|}\bigg\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}},$$

where $C_{u,1}$ and $C_{u,2}$ are two universal constants.

Lemma 20 *Assume that Condition 3 holds. For interval $I \subset (0,n]$ and a fixed $\boldsymbol{\theta}$, with probability at least $1 - n^{-2}\exp\{-C_{u,1}\log(p \vee n)\}$,*

$$\Big|\sum_{i\in I}\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta})\epsilon_i\Big| \le C_{u,2}C_xC_\epsilon\bigg\{\Delta_{I,\boldsymbol{\theta}}^2 \vee \frac{\Delta_{I,\infty,\boldsymbol{\theta}}^2\log(p\vee n)}{|I|}\bigg\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}},$$

where $C_{u,1}$ and $C_{u,2}$ are two universal constants.

Proof [Proof of Lemma 19–20.] They both follow from Bernstein's inequality with similar arguments as in the proof of Lemma 17. ■

C.2 Certifying $\mathbb{G}$ for the Full Model-fitting

Lemma 21 (In-sample error) *Assume Condition 2 (a) and Condition 3 hold. Under the setting in Lemma 18, with probability at least $1 - 4\exp\{-C_{u,1}\log(p\vee n)\}$, for any interval $I = (\tau_l, \tau_r]$ such that $|I| \ge C_{\mathsf{re}}s_I\log(p\vee n)$,*

$$\begin{aligned}\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| &\le \frac{48C_{u,2}^2C_x^2\sigma_x^2D_I^2s_I\log(p\vee n)}{\underline{\kappa}}\\ &+C_{u,2}C_x(C_x\Delta_{I,\infty} + 2C_\epsilon)[(\Delta_I^2|I|)\vee\{\Delta_{I,\infty}^2\log(p\vee n)\}]^{\frac{1}{2}}\{\log(p\vee n)\}^{\frac{1}{2}}.\end{aligned}$$

Additionally if Condition 2 (b) holds,

$$\begin{aligned}\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| &\le \frac{48C_{u,2}^2C_x^2\sigma_x^2D_I^2s_I\log(p\vee n)}{\underline{\kappa}}\\ &+C_{u,2}C_x(C_x\sigma_xC_\theta + 2s_I^{-\frac{1}{2}}C_\epsilon)[(\Delta_I^2|I|)\vee\{\sigma_x^2C_\theta^2s_I\log(p\vee n)\}]^{\frac{1}{2}}\{s_I\log(p\vee n)\}^{\frac{1}{2}}.\end{aligned}$$

Proof [Proof of Lemma 21.] Assume that the inequalities in Lemmas 16–20 hold for all interval I such that $|I| \ge C_{\mathsf{re}}s_I\log(p\vee n)$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_I^\circ$. It holds with a probability lower bound $1 - 4\exp\{-C_{u,1}\log(p\vee n)\}$.

For any interval $I = (c,d]$, we will analyze the cost $\mathcal{L}_I$. By the definition of the cost $\mathcal{L}_I$,

$$\begin{aligned}\mathcal{L}_I &= \sum_{i\in I}(y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\theta}}_I)^2 = \sum_{i\in I}\{y_i - \mathbf{x}_i^\top\boldsymbol{\theta}_I^\circ + \mathbf{x}_i^\top(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)\}^2\\ &= \sum_{i\in I}\{(y_i - \mathbf{x}_i^\top\boldsymbol{\theta}_I^\circ)^2 + \{\mathbf{x}_i^\top(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)\}^2\} + 2\sum_{i\in I}\{\mathbf{x}_i\epsilon_i + \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\}^\top(\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I)\\ &\ge \sum_{i\in I}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta}_I^\circ)^2 - \lambda_{I,1}\|\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I\|_1 \ge \sum_{i\in I}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta}_I^\circ)^2 - \lambda_I\|\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I\|_1,\end{aligned}$$

where the second last inequality follows from Lemma 17 and the last one is from $\lambda_I = 2\lambda_{I,1} > 0$. By the definition of $\widehat{\boldsymbol{\theta}}_I$,

$$\mathcal{L}_I - \sum_{i\in I}(y_i - \mathbf{x}_i^\top \boldsymbol{\theta}_I^\circ)^2 \le \lambda_I(\|\boldsymbol{\theta}_I^\circ\|_1 - \|\widehat{\boldsymbol{\theta}}_I\|_1) \le \lambda_I\|\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I\|_1.$$

By combining the result in Lemma 18,

$$\Big|\mathcal{L}_I - \sum_{i\in I}(y_i - \mathbf{x}_i^\top \boldsymbol{\theta}_I^\circ)^2\Big| \le \lambda_I\|\boldsymbol{\theta}_I^\circ - \widehat{\boldsymbol{\theta}}_I\|_1 \le \frac{12\lambda_{I,1}^2 s_I}{\underline{\kappa}|I|} \le \frac{48C_{u,2}^2C_x^2\sigma_x^2D_I^2 s_I\log(p\vee n)}{\underline{\kappa}}. \tag{19}$$

By Lemma 19 and Lemma 20,

$$\begin{aligned}
&\Big|\sum_{i\in I}[(y_i - \mathbf{x}_i^\top\boldsymbol{\theta}_I^\circ)^2 - \epsilon_i^2] - \Delta_I^2|I|\Big|\\
=&\Big|\sum_{i\in I}\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\}^2 - \Delta_I^2|I| + \sum_{i\in I}2\epsilon_i\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\Big|\\
\le& C_{u,2}C_x^2\Big\{\Delta_{I,4}^4 \vee \frac{\Delta_{I,\infty}^4\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\\
&+2C_{u,2}C_xC_\epsilon\Big\{\Delta_I^2 \vee \frac{\Delta_{I,\infty}^2\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}.
\end{aligned} \tag{20}$$

From Eq. (19) and Eq. (20),

$$\begin{aligned}
&\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| \le \frac{48C_{u,2}^2C_x^2\sigma_x^2D_I^2 s_I\log(p\vee n)}{\underline{\kappa}}\\
&+C_{u,2}C_x^2\Big\{\Delta_{I,4}^4 \vee \frac{\Delta_{I,\infty}^4\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\\
&+2C_{u,2}C_xC_\epsilon\Big\{\Delta_I^2 \vee \frac{\Delta_{I,\infty}^2\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}.
\end{aligned}$$

By Condition 2 (b), $\Delta_{I,\infty} \le C_\theta\sigma_x s_I^{\frac{1}{2}}$. Hence we have

$$\Big\{\Delta_I^2\vee\frac{\Delta_{I,\infty}^2\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}} \le s_I^{-\frac{1}{2}}\Big[\{C_\theta\sigma_x s_I\log(p\vee n)\}\vee\{\Delta_I^2|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big].$$

Since $\Delta_{I,4}^2 \le \Delta_{I,\infty}\Delta_{I,2}$, it also holds that

$$\begin{aligned}
&\Big\{\Delta_{I,4}^4 \vee \frac{\Delta_{I,\infty}^4\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\\
\le& C_\theta\sigma_x\Big[\{C_\theta\sigma_x s_I\log(p\vee n)\}\vee\{\Delta_I^2|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big].
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| \le \frac{48C_{u,2}^2C_x^2\sigma_x^2D_I^2 s_I\log(p\vee n)}{\underline{\kappa}}\\
&+C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon s_I^{-\frac{1}{2}})[(\Delta_I^2|I|)\vee\{\sigma_x^2C_\theta^2 s_I\log(p\vee n)\}]^{\frac{1}{2}}\{s_I\log(p\vee n)\}^{\frac{1}{2}}.
\end{aligned}$$

■

Corollary 22 *Assume Condition 1, Condition 2, and Condition 3 hold. Under the same probability event in Lemma 21 and with sufficiently large $C_{\mathsf{m}} \ge C_{\mathsf{re}}$, we have the following conclusions.*

(a) For I such that $\Delta_I = 0$ and $|I| \ge C_{\mathsf{m}} s\log(p \vee n)$,

$$\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2\Big| \le \frac{48C_{u,2}^2C_x^2\sigma_x^2C_\epsilon^2 s\log(p\vee n)}{\underline{\kappa}} \triangleq C_{22.1}s\log(p\vee n).$$

(b) For $I = (a,b]$ such that $I \cap \mathcal{T}^* = \{\tau_k^*\}$, $\min(\tau_k^* - a, b - \tau_k^*) \le \widetilde{C}\Delta_k^{-2}s\log(p\vee n)$ and $|I| \ge C_{\mathsf{m}} s\log(p\vee n)$,

$$\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| \le C_{22.2}s\log(p\vee n),$$

where $C_{22.2} = 2C_{22.1} + (C_{\mathsf{m}}\underline{\kappa})^{-1}48C_{u,2}^2C_x^4\sigma_x^2\widetilde{C} + C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon)\{(2C_\theta\sigma_x)\vee(\widetilde{C}^{\frac{1}{2}})\}$.

(c) For I such that $|I| \ge C_{\mathsf{m}} s\log(p\vee n)$ and $\Delta_I^2|I| \ge 2^{-1}\widetilde{C}s\log(p\vee n)$ for some sufficiently large $\widetilde{C} \ge 6C_\theta^2\sigma_x^2$,

$$\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 \ge (1 - C_{22.3})\Delta_I^2|I|,$$

where $C_{22.3} = (C_{\mathsf{m}}\underline{\kappa})^{-1}(96C_{u,2}^2C_x^4\sigma_x^2) + 6^{\frac{1}{2}}C_{u,2}C_x\Big(C_x\sigma_xC_\theta + 2C_\epsilon\Big)\widetilde{C}^{-\frac{1}{2}} + 6\widetilde{C}^{-1}C_{22.1}$.

Proof [Proof of Corollary 22.] All of the results in the three parts of Corollary 22 follow from the proof of Lemma 21 and the conditions about the change signals, that is, the variations $\Delta_I^2|I|$.
Part (a). If $\Delta_I = 0$, we have $D_I^2 = C_\epsilon^2$ and $\Delta_{I,\infty} = 0$. Lemma 21 reduces to

$$\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| \le \frac{48C_{u,2}^2C_x^2\sigma_x^2C_\epsilon^2}{\underline{\kappa}}s\log(p\vee n) = C_{22.1}s\log(p\vee n).$$

Part (b). Since $|I\cap\mathcal{T}^*| = 1$, we have $s_I \le 2s$ and Lemma 21 holds for $|I| \ge C_{\mathsf{m}}s\log(p\vee n) \ge 2^{-1}C_{\mathsf{m}}s_I\log(p\vee n) \ge C_{\mathsf{re}}s_I\log(p\vee n)$ with sufficiently large $C_{\mathsf{m}} \ge 2C_{\mathsf{re}}$. By $|I| \ge C_{\mathsf{m}}s\log(p\vee n)$, we have $|I|^{-1}(C_x^2\Delta_{I,\infty}^2 + C_\epsilon^2)\log(p\vee n) \le 2(C_{\mathsf{m}}s)^{-1}C_x^2\sigma_x^2C_\theta^2 s + C_\epsilon^2 \le C_\epsilon^2$ provided that C_{m} is sufficiently large. Hence $D_I^2 = C_x^2\Delta_I^2 + C_\epsilon^2$. Combining $\Delta_I^2|I| \le 2^{-1}\widetilde{C}s\log(p\vee n)$ and $|I| \ge C_{\mathsf{m}}s\log(p\vee n)$, $\Delta_I^2 \le 2^{-1}C_{\mathsf{m}}^{-1}\widetilde{C}$. Therefore by Lemma 21,

$$\Big|\mathcal{L}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_I^2|I|\Big| \le C_{22.2}s\log(p\vee n),$$

where

$$C_{22.2} = \frac{96C_{u,2}^2C_x^2\sigma_x^2C_\epsilon^2}{\underline{\kappa}} + \frac{48C_{u,2}^2C_x^4\sigma_x^2\widetilde{C}}{C_{\mathsf{m}}\underline{\kappa}} + C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon)\{(2C_\theta\sigma_x)\vee(\widetilde{C}^{\frac{1}{2}})\}.$$

Part (c). When $|I\cap\mathcal{T}^*|\le 1$, the discussion in (b) is still valid. We have $|I|\ge C_{\mathsf{m}}s\log(p\vee n)\ge 2^{-1}C_{\mathsf{m}}s_I\log(p\vee n)$ and $\Delta_I^2|I|\ge 4^{-1}\widetilde{C}s_I\log(p\vee n)$. Otherwise when $|I\cap\mathcal{T}^*|\ge 2$, by Condition 1, we can obtain $|I|\ge 3^{-1}C_{\mathsf{snr}}s_I\log(p\vee n)\ge 2^{-1}C_{\mathsf{m}}s_I\log(p\vee n)$ and $\Delta_I^2|I|\ge 6^{-1}\widetilde{C}s_I\log(p\vee n)$ since C_{snr} is sufficiently large. The above claim becomes trivial when $|I\cap\mathcal{T}^*|\ge 3$. When $|I\cap\mathcal{T}^*|=2$, the result follows from the fact that $s\ge 3^{-1}s_I$. Recall that $\widetilde{C}\ge 6C_\theta^2\sigma_x^2$. We have $\Delta_I^2\ge(6|I|)^{-1}\widetilde{C}s_I\log(p\vee n)\ge|I|^{-1}C_\theta^2\sigma_x^2s_I\log(p\vee n)\ge|I|^{-1}\Delta_{I,\infty}^2\log(p\vee n)$. It implies that $D_I^2=C_x^2\Delta_I^2+C_\epsilon^2$. By Lemma 21,

$$\begin{aligned}\mathcal{L}_I-\sum_{i\in I}\epsilon_i^2 \ge{}& \Delta_I^2|I| - \frac{48C_{u,2}^2C_x^2\sigma_x^2(C_x^2\Delta_I^2+C_\epsilon^2)s_I\log(p\vee n)}{\underline{\kappa}}\\ &-C_{u,2}C_x(C_x\sigma_xC_\theta+2s_I^{-\frac{1}{2}}C_\epsilon)\{\Delta_I^2|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\ge(1-C_{22.3})\Delta_I^2|I|,\end{aligned}\tag{21}$$

where

$$C_{22.3}=\frac{96C_{u,2}^2C_x^4\sigma_x^2}{\underline{\kappa}C_{\mathsf{m}}}+\frac{6C_{22.1}}{\widetilde{C}}+\frac{6^{\frac{1}{2}}C_{u,2}C_x\big(C_x\sigma_xC_\theta+2C_\epsilon\big)}{\widetilde{C}^{\frac{1}{2}}}.$$

■

C.3 Certifying $\mathbb{G}$ for Reliever

To proceed, we first introduce some notations used in the theoretical justification for Reliever. Let R be the surrogate interval w.r.t. I and $J=I\setminus R$ be the complement. Denote oracle change variation for Reliever by $\overline{\Delta}_I^2=\Delta^2_{I,\boldsymbol{\theta}_R^\circ}$. To distinguish the losses for Reliever from these for full model-fitting, we denote the losses of interval I under the relief model estimate $\widehat{\boldsymbol{\theta}}_R$, named relief losses, by $\widetilde{\mathcal{L}}_I=\sum_{i\in I}(y_i-\mathbf{x}_i^\top\widehat{\boldsymbol{\theta}}_R)^2$.

Lemma 23 (Relief error) *Assume Condition 2 and Condition 3 hold. Under the setting in Lemma 18, with probability at least* $1-4\exp\{-C_{u,1}\log(p\vee n)\}$, *for any interval* $I=(\tau_l,\tau_r]$ *such that* $|I|\ge r^{-1}C_{\mathsf{re}}s_I\log(p\vee n)$ *so that its relief interval* R *satisfies* $|R|\ge C_{\mathsf{re}}s_I\log(p\vee n)$, *the relief loss* $\widetilde{\mathcal{L}}_I$ *satisfies that:*

$$\begin{aligned}&\big|\widetilde{\mathcal{L}}_I-\sum_{i\in I}\epsilon_i^2-\Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I|\big|\\ \le{}&(1+C_{16})r^{-1}\underline{\kappa}C_{18}^2D_R^2s_R\log(p\vee n)\\ &+6^{-1}r^{-\frac{1}{2}}\underline{\kappa}C_{18}^2D_{I,\boldsymbol{\theta}_R^\circ}D_Rs_R\log(p\vee n)\\ &+2r^{-\frac{1}{2}}(1-r)^{\frac{1}{2}}\underline{\kappa}^{\frac{1}{2}}C_{18}D_R\{\Delta^2_{J,\boldsymbol{\theta}_R^\circ}|J|\}^{\frac{1}{2}}\{s_R\log(p\vee n)\}^{\frac{1}{2}}.\\ &+C_{u,2}C_x(C_x\sigma_xC_\theta+2C_\epsilon s_I^{-\frac{1}{2}})\Big[\{C_\theta\sigma_xs_I\log(p\vee n)\}\vee\{\Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big].\end{aligned}$$

Proof [Proof of Lemma 23] Similarly, we begin with the control of the difference $\widetilde{\mathcal{L}}_I - \widetilde{\mathcal{L}}_I^\circ$. Observe that,

$$
\begin{aligned}
&\widetilde{\mathcal{L}}_I - \widetilde{\mathcal{L}}_I^\circ\\
=&|I|\|\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R\|_{\widehat{\Sigma}_I}^2 + 2\sum_{i\in I}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta}_R^\circ)\mathbf{x}_i^\top(\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R)\\
=&|I|\|\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R\|_{\widehat{\Sigma}_I}^2 + 2\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ) + \epsilon_i\mathbf{x}_i\}^\top(\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R)\\
&+ 2\sum_{i\in I\setminus R}(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ)^\top\Sigma(\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R).
\end{aligned}
$$

We will derive the upper bounds for the absolute values of the three terms in the above decomposition.

For the first term, assuming the probability events in Lemma 16 and Lemma 18 hold, we have:

$$
|I|\|\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R\|_{\widehat{\Sigma}_I}^2 \le (1 + C_{16})r^{-1}\underline{\kappa}C_{18}^2D_R^2s_R\log(p\vee n). \tag{22}
$$

For the second term, assuming the probability events in Lemma 17 and Lemma 18 hold, we have:

$$
\begin{aligned}
&\Big|\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ) + \epsilon_i\mathbf{x}_i\}^\top(\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R)\Big|\\
\le&\Big\|\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ) + \epsilon_i\mathbf{x}_i\}\Big\|_\infty\big\|\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R\big\|_1\\
\le&C_{u,2}C_x\sigma_xD_{I,\boldsymbol{\theta}_R^\circ}\{|I|\log(p\vee n)\}^{\frac{1}{2}} \times C_{18}D_Rs_R\Big\{\frac{\log(p\vee n)}{|I|r}\Big\}^{\frac{1}{2}}\\
=&12^{-1}r^{-\frac{1}{2}}\underline{\kappa}C_{18}^2D_{I,\boldsymbol{\theta}_R^\circ}D_Rs_R\log(p\vee n).
\end{aligned}
$$

where $D_{I,\boldsymbol{\theta}} = [(C_x^2\Delta_{I,\boldsymbol{\theta}}^2 + C_\epsilon^2) \vee \{|I|^{-1}(C_x^2\Delta_{I,\infty,\boldsymbol{\theta}}^2 + C_\epsilon^2)\log(p\vee n)\}]^{\frac{1}{2}}$.

For the third term, assuming the probability event in Lemma 18 holds, we have:

$$
\begin{aligned}
&\Big|\sum_{i\in J}(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ)^\top\Sigma(\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R)\Big|\\
\le&\sum_{i\in J}\|\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ\|_\Sigma\|\boldsymbol{\theta}_R^\circ - \widehat{\boldsymbol{\theta}}_R\|_\Sigma\\
\le&\Delta_{J,\boldsymbol{\theta}_R^\circ}|J| \times \underline{\kappa}^{\frac{1}{2}}C_{18}D_R\Big\{\frac{s_R\log(p\vee n)}{|R|}\Big\}^{\frac{1}{2}}\\
\le&\underline{\kappa}^{\frac{1}{2}}C_{18}D_R\{\Delta_{J,\boldsymbol{\theta}_R^\circ}^2|J|\}^{\frac{1}{2}}\Big\{\frac{(1-r)s_R\log(p\vee n)}{r}\Big\}^{\frac{1}{2}}.
\end{aligned}
\tag{23}
$$

Therefore,

$$
\begin{aligned}
|\widetilde{\mathcal{L}}_I - \widetilde{\mathcal{L}}_I^\circ| \le&(1 + C_{16})r^{-1}\underline{\kappa}C_{18}^2D_R^2s_R\log(p\vee n)\\
&+ 6^{-1}r^{-\frac{1}{2}}\underline{\kappa}C_{18}^2D_{I,\boldsymbol{\theta}_R^\circ}D_Rs_R\log(p\vee n)\\
&+ 2r^{-\frac{1}{2}}(1-r)^{\frac{1}{2}}\underline{\kappa}^{\frac{1}{2}}C_{18}D_R\{\Delta_{J,\boldsymbol{\theta}_R^\circ}^2|J|\}^{\frac{1}{2}}\{s_R\log(p\vee n)\}^{\frac{1}{2}}.
\end{aligned}
\tag{24}
$$

We now turn to deriving the upper bound of $|\widetilde{\mathcal{L}}_I^\circ - \sum_{i\in I}\epsilon_i^2 - \Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I||$. By Lemma 19 and Lemma 20,

$$\begin{aligned}
&|\widetilde{\mathcal{L}}_I^\circ - \sum_{i\in I}\epsilon_i^2 - \Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I||\\
=&\Big|\Big[\sum_{i\in I}\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_R^\circ)\}^2\Big] - \Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I| + \sum_{i\in I}2\epsilon_i\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}_I^\circ)\Big|\\
\le& C_{u,2}C_x^2\Big\{\Delta^4_{I,4,\boldsymbol{\theta}_R^\circ} \vee \frac{\Delta^4_{I,\infty,\boldsymbol{\theta}_R^\circ}\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\\
&+2C_{u,2}C_xC_\epsilon\Big\{\Delta^2_{I,\boldsymbol{\theta}_R^\circ} \vee \frac{\Delta^2_{I,\infty,\boldsymbol{\theta}_R^\circ}\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}.
\end{aligned}$$

By Condition 2 (b), $\Delta_{I,\infty,\boldsymbol{\theta}_R^\circ} \le C_\theta\sigma_x s_I^{\frac{1}{2}}$. Hence we have

$$\begin{aligned}
&\Big\{\Delta^2_{I,\boldsymbol{\theta}_R^\circ} \vee \frac{\Delta^2_{I,\infty,\boldsymbol{\theta}_R^\circ}\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\\
\le& s_I^{-\frac{1}{2}}\Big[\{C_\theta\sigma_x s_I\log(p\vee n)\} \vee \{\Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big].
\end{aligned}$$

Since $\Delta^2_{I,4,\boldsymbol{\theta}_R^\circ} \le \Delta_{I,\infty,\boldsymbol{\theta}_R^\circ}\Delta_{I,\boldsymbol{\theta}_R^\circ}$, it also holds that

$$\begin{aligned}
&\Big\{\Delta^4_{I,4,\boldsymbol{\theta}_R^\circ} \vee \frac{\Delta^4_{I,\infty,\boldsymbol{\theta}_R^\circ}\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}}\\
\le& C_\theta\sigma_x\Big[\{C_\theta\sigma_x s_I\log(p\vee n)\} \vee \{\Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&|\widetilde{\mathcal{L}}_I^\circ - \sum_{i\in I}\epsilon_i^2 - \Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I||\\
\le& C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon s_I^{-\frac{1}{2}})\Big[\{C_\theta\sigma_x s_I\log(p\vee n)\} \vee \{\Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big]. \qquad (25)
\end{aligned}$$

By Eq.(24) and Eq.(25),

$$\begin{aligned}
&|\widetilde{\mathcal{L}}_I - \sum_{i\in I}\epsilon_i^2 - \Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I||\\
\le&(1 + C_{16})r^{-1}\underline{\kappa}C_{18}^2D_R^2s_R\log(p\vee n)\\
&+ 6^{-1}r^{-\frac{1}{2}}\underline{\kappa}C_{18}^2D_{I,\boldsymbol{\theta}_R^\circ}D_Rs_R\log(p\vee n)\\
&+ 2r^{-\frac{1}{2}}(1-r)^{\frac{1}{2}}\underline{\kappa}^{\frac{1}{2}}C_{18}D_R\{\Delta^2_{J,\boldsymbol{\theta}_R^\circ}|J|\}^{\frac{1}{2}}\{s_R\log(p\vee n)\}^{\frac{1}{2}}.\\
&+ C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon s_I^{-\frac{1}{2}})\Big[\{C_\theta\sigma_x s_I\log(p\vee n)\} \vee \{\Delta^2_{I,\boldsymbol{\theta}_R^\circ}|I|s_I\log(p\vee n)\}^{\frac{1}{2}}\Big].
\end{aligned}$$

■

We now proceed to certify that $\mathbb{G} = \mathbb{G}_1 \cap \mathbb{G}_2^- \cap \mathbb{G}_2^+ \cap \mathbb{G}_3$ holds with high probability.

Corollary 24 *Assume Condition 1, Condition 2, and Condition 3 hold. Under the same probability event in Lemma 23 and with sufficiently large* $C_{\mathsf{m}} \geq C_{\mathsf{re}}$*, we have the following conclusions.*

(a) For I *such that* $\Delta_I = 0$ *and* $|I| \geq C_{\mathsf{m}} s \log(p \vee n)$*,*

$$\left|\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2\right| \leq \frac{48 C_{u,2}^2 C_x^2 \sigma_x^2 C_\epsilon^2 s \log(p \vee n)}{\underline{\kappa}} \triangleq C_{24.1} s \log(p \vee n),$$

where $C_{24.1} = \frac{6+6C_{16}+r^{\frac{1}{2}}}{6r} \underline{\kappa} C_{18}^2 C_\epsilon^2 + C_{u,2} C_x \sigma_x (C_x \sigma_x C_\theta + 2C_\epsilon s^{-\frac{1}{2}}) C_\theta$*.*

(b) For $I = (a, b] \in E_2^-$ *such that* $I \cap \mathcal{T}^* = \{\tau_k^*\}$, $\min(\tau_k^* - a, b - \tau_k^*) \leq \widetilde{C} \Delta_k^{-2} s \log(p \vee n)$ *and* $|I| \geq C_{\mathsf{m}} s \log(p \vee n)$*,*

$$\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_I^2 |I| \geq -C_{24.2} s \log(p \vee n),$$

and for $I \in E_2^+ \subset E_2^-$ *such that* $I \cap \mathcal{T}^* = \{\tau_k^*\}$, $\min(\tau_k^* - a, b - \tau_k^*) \leq \widetilde{C} \Delta_k^{-2} s \log(p \vee n)$ *and* $|I| \geq (\Delta_k^2 \vee 1) C_{\mathsf{snr}} s \log(p \vee n)$*,*

$$\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_I^2 |I| \leq C_{24.2} s \log(p \vee n),$$

where

$$\begin{aligned} C_{24.2} =& 2\{(1 + C_{16})(r^{-1} C_\epsilon^2 + r^{-2} C_x^2 C_m^{-1} \widetilde{C}) + 6^{-1}(r^{-\frac{1}{2}} C_\epsilon^2 + r^{-\frac{3}{2}} C_x^2 C_m^{-1} \widetilde{C})\} \underline{\kappa} C_{18}^2 \\ &+ 2^{\frac{3}{2}}(r^{-\frac{1}{2}} C_\epsilon \widetilde{C}^{\frac{1}{2}} + r^{-1} C_x C_{\mathsf{m}}^{-1} \widetilde{C})(1 - r)^{\frac{1}{2}} \underline{\kappa}^{\frac{1}{2}} C_{18} \\ &+ 2C_{u,2} C_x (C_x \sigma_x C_\theta + 2C_\epsilon s_I^{-\frac{1}{2}})(C_\theta \sigma_x + 2^{-\frac{1}{2}} r^{-\frac{1}{2}} \widetilde{C}^{\frac{1}{2}}) + r^{-1} C_{\mathsf{snr}}^{-1} \widetilde{C}. \end{aligned}$$

(c) For I *such that* $|I| \geq C_{\mathsf{m}} s \log(p \vee n)$ *and* $\Delta_I^2 |I| \geq 2^{-1} \widetilde{C} s \log(p \vee n)$ *for some sufficiently large* $\widetilde{C} \geq 6 C_\theta^2 \sigma_x^2$*,*

$$\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 \geq (1 - C_{22.3}) \Delta_I^2 |I|,$$

where

$$\begin{aligned} C_{24.3} =& (1 + C_{16}) \underline{\kappa} C_{18}^2 \{6 C_\epsilon^2 \widetilde{C}^{-1} r^{-1} + 2(C_x^2 + C_{24.3}' \widetilde{C}^{-1}) C_{\mathsf{m}}^{-1} r^{-2}\} \\ &+ \underline{\kappa} C_{18}^2 \{C_\epsilon^2 \widetilde{C}^{-1} r^{-\frac{1}{2}} + 6^{-1}(C_x^2 + C_{24.3}' \widetilde{C}^{-1}) C_{\mathsf{m}}^{-1}(r^{-\frac{3}{2}} + r^{-\frac{1}{2}})\} \\ &+ (1 - r)^{\frac{1}{2}} \underline{\kappa}^{\frac{1}{2}} C_{18} \{12^{\frac{1}{2}} C_\epsilon \widetilde{C}^{-\frac{1}{2}} r^{-\frac{1}{2}} + 2(C_x^2 + C_{24.3}' \widetilde{C}^{-1})^{\frac{1}{2}} C_{\mathsf{m}}^{-\frac{1}{2}} r^{-1}\} \\ &+ C_{u,2} C_x (C_x \sigma_x C_\theta + 2C_\epsilon)(6 C_\theta \sigma_x \widetilde{C}^{-1} + 6^{\frac{1}{2}} \widetilde{C}^{-\frac{1}{2}}). \end{aligned}$$

Proof [Proof of Corollary 24] All of the results in the three parts of Corollary 24 follow from the proof of Lemma 23 and the conditions about the change signals, that is, the variations $\Delta_I^2 |I|$. In the proof, we will also need to analyze the relationship of $\Delta_I^2 |I|$ and $\Delta_{I, \boldsymbol{\theta}_R^\circ}^2 |I|$.

Part (a). For I such that $\Delta_I = 0$ and $|I| \geq \delta_{\mathsf{m}}$, there is not changepoint in I. So $s_I = s_R = s$, $\boldsymbol{\theta}^\circ_R = \boldsymbol{\theta}^\circ_I$, $\Delta_{I,\boldsymbol{\theta}^\circ_R} = \Delta_{J,\boldsymbol{\theta}^\circ_R} = 0$. Since C_{m} is sufficiently large and $r \in (0, 1]$ is a fixed constant, $D_R = D_{I,\boldsymbol{\theta}^\circ_R} = C_\epsilon$. Therefore,

$$\begin{aligned}
&|\widetilde{\mathcal{L}} - \sum_{i\in I} \epsilon_i^2| \\
\leq&\Big\{\frac{6 + 6C_{16} + r^{\frac{1}{2}}}{6r}\underline{\kappa} C_{18}^2 C_\epsilon^2 + C_{u,2}C_x\sigma_x(C_x\sigma_xC_\theta + 2C_\epsilon s^{-\frac{1}{2}})C_\theta\Big\} s\log(p \vee n) \\
=&C_{24.1}s\log(p \vee n),
\end{aligned}$$

Part (b). We study the interval I from the set

$$E_2^- = \{I = (a, b] : |I| \geq \delta_{\mathsf{m}}, I \cap \mathcal{T}^* = \{\tau_k^*\}, \min(\tau_k^* - a, b - \tau_k^*) \leq \widetilde{C}\Delta_k^{-2}s\log(p \vee n)\}.$$

Since I contains only one changepoint, we have $s_R \leq s_I \leq 2s$, $\Delta_{R,\infty}^2 \leq \Delta_{I,\infty,\boldsymbol{\theta}^\circ_R}^2 \leq 2\sigma_x^2C_\theta^2 s$.

For the average variation terms, by definition, we have $\Delta_R^2|R| \leq \Delta_I^2|I| \leq \widetilde{C}s\log(p \vee n)$. It means that $\Delta_R^2 \leq r^{-1}\Delta_I^2 \leq r^{-1}C_{\mathsf{m}}^{-1}\widetilde{C}$. Similarly, with some calculation, we can also have $\Delta_{J,\boldsymbol{\theta}^\circ_R}^2|J| \leq \Delta_{I,\boldsymbol{\theta}^\circ_R}^2|I| \leq r^{-1}\Delta_I^2|I| \leq r^{-1}\widetilde{C}s\log(p \vee n)$ and $\Delta_{I,\boldsymbol{\theta}^\circ_R}^2 \leq r^{-1}C_{\mathsf{m}}^{-1}\widetilde{C}$.

We now discuss the terms D_R^2 and $D_{I,\boldsymbol{\theta}^\circ_R}^2$. Since $|I| \geq \delta_{\mathsf{m}} \geq C_{\mathsf{m}}s\log(p \vee n)$ and C_{m} is sufficiently large, we have

$$\begin{aligned}
&|I|^{-1}(C_x^2\Delta_{I,\infty}^2 + C_\epsilon^2)\log(p \vee n) \\
\leq&|I|^{-1}(2C_x^2\sigma_x^2C_\theta^2 s + C_\epsilon^2)\log(p \vee n) \\
\leq&(2C_x^2\sigma_x^2C_\theta^2 + C_\epsilon^2 s^{-1})C_{\mathsf{m}}^{-1} \leq C_\epsilon^2.
\end{aligned}$$

Therefore $D_{I,\boldsymbol{\theta}^\circ_R}^2 = C_x^2\Delta_{I,\boldsymbol{\theta}^\circ_R}^2 + C_\epsilon^2$ and $D_R^2 = C_x^2\Delta_R^2 + C_\epsilon^2$.

Combining Lemma 23, we have

$$\begin{aligned}
&\big|\widetilde{\mathcal{L}}_I - \sum_{i\in I}\epsilon_i^2 - \Delta_{I,\boldsymbol{\theta}^\circ_R}^2|I|\big| \\
\leq&(1 + C_{16})r^{-1}\underline{\kappa}C_{18}^2C_\epsilon^2 s_R\log(p \vee n) + (1 + C_{16})r^{-1}\underline{\kappa}C_{18}^2C_x^2\Delta_R^2 s_R\log(p \vee n) \\
&+ 6^{-1}r^{-\frac{1}{2}}\underline{\kappa}C_{18}^2C_\epsilon^2 s_R\log(p \vee n) + 12^{-1}r^{-\frac{1}{2}}\underline{\kappa}C_{18}^2C_x^2(\Delta_R^2 + \Delta_{I,\boldsymbol{\theta}^\circ_R}^2)s_R\log(p \vee n) \\
&+ 2r^{-\frac{1}{2}}(1 - r)^{\frac{1}{2}}\underline{\kappa}^{\frac{1}{2}}C_{18}(C_x\Delta_R + C_\epsilon)\{\Delta_{J,\boldsymbol{\theta}^\circ_R}^2|J|\}^{\frac{1}{2}}\{s_R\log(p \vee n)\}^{\frac{1}{2}}. \\
&+ C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon s_I^{-\frac{1}{2}})\Big[\{C_\theta\sigma_x s_I\log(p \vee n)\} \vee \{\Delta_{I,\boldsymbol{\theta}^\circ_R}^2|I|s_I\log(p \vee n)\}^{\frac{1}{2}}\Big] \\
\leq&2\{(1 + C_{16})(r^{-1}C_\epsilon^2 + r^{-2}C_x^2C_m^{-1}\widetilde{C}) + 6^{-1}(r^{-\frac{1}{2}}C_\epsilon^2 + r^{-\frac{3}{2}}C_x^2C_m^{-1}\widetilde{C})\}\underline{\kappa}C_{18}^2 s\log(p \vee n) \\
&+ 2^{\frac{3}{2}}(r^{-\frac{1}{2}}C_\epsilon\widetilde{C}^{\frac{1}{2}} + r^{-1}C_xC_{\mathsf{m}}^{-1}\widetilde{C})(1 - r)^{\frac{1}{2}}\underline{\kappa}^{\frac{1}{2}}C_{18}s\log(p \vee n) \\
&+ 2C_{u,2}C_x(C_x\sigma_xC_\theta + 2C_\epsilon s_I^{-\frac{1}{2}})(C_\theta\sigma_x + 2^{-\frac{1}{2}}r^{-\frac{1}{2}}\widetilde{C}^{\frac{1}{2}})s\log(p \vee n) \\
\triangleq&C'_{24.2}s\log(p \vee n).
\end{aligned}$$

For $I = (a,b] \in E_2^+ \subseteq E_2^-$, since $|I| \geq (\Delta_k^2 \vee 1) C_{\mathsf{snr}} s \log(p \vee n)$ and C_{snr} is sufficiently large, we have $0 \leq \Delta_{I,\boldsymbol{\theta}}^2 |I| - \Delta_I^2 |I| \leq r^{-1} C_{\mathsf{snr}}^{-1} \widetilde{C} s \log(p \vee n)$. Therefore, for $I \in E_2^+$,

$$\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_I^2 |I| \leq C_{24.2} s \log(p \vee n),$$

with $C_{24.2} = C'_{24.2} + r^{-1} C_{\mathsf{snr}}^{-1} \widetilde{C}$.

For $I \in E_2^-$, similarly,

$$\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_I^2 |I| \geq \widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_{I,\boldsymbol{\theta}_R^\circ}^2 |I| \geq -C'_{24.2} s \log(p \vee n) \geq -C_{24.2} s \log(p \vee n).$$

Part (c). For I such that $|I| \geq C_{\mathsf{m}} s \log(p \vee n)$ and $\Delta_I^2 |I| \geq 2^{-1} \widetilde{C} s \log(p \vee n)$, it also holds that $\Delta_{I,\boldsymbol{\theta}_R^\circ}^2 |I| \geq \Delta_I^2 |I| \geq 2^{-1} \widetilde{C} s \log(p \vee n)$.

If I contains only one changepoint, we have $s_I \leq 2s$ and $s_R \leq 2s$. Therefore it holds that $|I| \geq 2^{-1} C_{\mathsf{m}} s_I \log(p \vee n)$ and $|R| \geq 2^{-1} C_{\mathsf{m}} r s_R \log(p \vee n)$. If I contains more changepoints, by Conditions 1 and 2 and provided that C_{snr} is sufficiently large, it still holds that $|I| \geq 2^{-1} C_{\mathsf{m}} s_I \log(p \vee n)$ and $|R| \geq 2^{-1} C_{\mathsf{m}} r s_R \log(p \vee n)$. Following similar arguments in the proof of Corollary 22, we can obtain that $\Delta_I^2 |I| \geq 6^{-1} \widetilde{C} s_I \log(p \vee n) \geq 6^{-1} \widetilde{C} s_R \log(p \vee n)$. Therefore we have $s_R \log(p \vee n) \leq \frac{2|R|}{C_{\mathsf{m}} r}$ and $s_R \log(p \vee n) \leq \frac{2|R|}{C_{\mathsf{m}} r}$ and $s_R \log(p \vee n) \leq s_I \log(p \vee n) \leq 6 \widetilde{C}^{-1} \Delta_I^2 |I| \leq 6 \widetilde{C}^{-1} \Delta_{I,\boldsymbol{\theta}_R^\circ}^2 |I|$.

Then we consider the two terms $D_{I,\boldsymbol{\theta}_R^\circ}^2$ and D_R^2. By the above justifications, we can obtain that

$$\begin{aligned}
&|I|^{-1} (C_x^2 \Delta_{I,\infty,\boldsymbol{\theta}_R^\circ}^2 + C_\epsilon^2) \log(p \vee n) \\
\leq &|I|^{-1} (C_x^2 \sigma_x^2 C_\theta^2 s_I + C_\epsilon^2) \log(p \vee n) \\
\leq &6 \widetilde{C}^{-1} (C_x^2 \sigma_x^2 C_\theta^2 + C_\epsilon^2 s_I^{-1}) \Delta_{I,\boldsymbol{\theta}_R^\circ}^2.
\end{aligned}$$

It follows that

$$D_{I,\boldsymbol{\theta}_R^\circ}^2 \leq C_\epsilon^2 + \{C_x^2 + 6 \widetilde{C}^{-1} (C_x^2 \sigma_x^2 C_\theta^2 + C_\epsilon^2 s_I^{-1})\} \Delta_{I,\boldsymbol{\theta}_R^\circ}^2.$$

Similarly,

$$D_R^2 \leq C_\epsilon^2 + \{C_x^2 + 6 \widetilde{C}^{-1} (C_x^2 \sigma_x^2 C_\theta^2 + C_\epsilon^2 s_R^{-1})\} \Delta_R^2.$$

Denote $C'_{24.3} = 6(C_x^2 \sigma_x^2 C_\theta^2 + C_\epsilon^2)$. By the above inequalities for the change variation terms, we have:

$$\begin{aligned}
&\Big|\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 - \Delta_{I,\boldsymbol{\theta}_R^\circ}^2 |I|\Big| \\
\leq &\Delta_{I,\boldsymbol{\theta}_R^\circ}^2 |I| \Big[(1 + C_{16}) \underline{\kappa} C_{18}^2 \{6 C_\epsilon^2 \widetilde{C}^{-1} r^{-1} + 2(C_x^2 + C'_{24.3} \widetilde{C}^{-1}) C_{\mathsf{m}}^{-1} r^{-2}\} \\
&\qquad + \underline{\kappa} C_{18}^2 \{C_\epsilon^2 \widetilde{C}^{-1} r^{-\frac{1}{2}} + 6^{-1} (C_x^2 + C'_{24.3} \widetilde{C}^{-1}) C_{\mathsf{m}}^{-1} (r^{-\frac{3}{2}} + r^{-\frac{1}{2}})\} \\
&\qquad + (1 - r)^{\frac{1}{2}} \underline{\kappa}^{\frac{1}{2}} C_{18} \{12^{\frac{1}{2}} C_\epsilon \widetilde{C}^{-\frac{1}{2}} r^{-\frac{1}{2}} + 2(C_x^2 + C'_{24.3} \widetilde{C}^{-1})^{\frac{1}{2}} C_{\mathsf{m}}^{-\frac{1}{2}} r^{-1}\} \\
&\qquad + C_{u,2} C_x (C_x \sigma_x C_\theta + 2 C_\epsilon)(6 C_\theta \sigma_x \widetilde{C}^{-1} + 6^{\frac{1}{2}} \widetilde{C}^{-\frac{1}{2}}) \Big] \\
= &C_{24.3} \Delta_{I,\boldsymbol{\theta}_R^\circ}^2 |I|.
\end{aligned}$$

Note that C_{m} is sufficiently large, we can set $\widetilde{C}$ large enough to make $C_{24.3}$ is sufficiently small such that $C_{24.3} < 1$. Finally, we have

$$\widetilde{\mathcal{L}}_I - \sum_{i \in I} \epsilon_i^2 \geq (1 - C_{24.3}) \Delta^2_{I, \boldsymbol{\theta}^\circ_R} |I| \geq (1 - C_{24.3}) \Delta_I^2 |I|.$$

■

C.4 Final Result

Here, we will finally prove Theorem 9. Recall that in the proof of Lemma 8, we identify the constant $\widetilde{C}$ by solving the inequality (15). It concludes that any $\widetilde{C}$ with $\widetilde{C} \geq 4(1 - C_{8.3})^{-1}(3C_{8.1} + 10C_{8.2})$ suffices for Lemma 8. Since C_{m} can be sufficiently large and $\widetilde{C}C_{\mathsf{m}}^{-1}$ can be sufficiently small, by the results in 22 and Corollary 24, we can verify that the feasible $\widetilde{C}$ exists for the inequality $\widetilde{C} \geq 4(1 - C_{8.3})^{-1}(3C_{8.1} + 10C_{8.2})$, both for the full model-fitting (Corollary 22) and the Reliever (Corollary 24). Furthermore, the $\widetilde{C}$ will only depends on the constants in the expressions of $\{C_{22.j}\}_{j=1}^3$ and $\{C_{24.j}\}_{j=1}^3$. Thus it is independent of (n, p, s, K^*).

In summary, the localization error bound of $\{\widehat{\tau}_k\}$ in Theorem 9 follows from Lemma 8, Corollary 22 and Corollary 24. And provided the localization error bound, the error bound of the parameter estimation follows from Lemma 18.

Appendix D. Proof of Corollary 12

Firstly, note that under Condition 4, Lemma 8 still holds if we replace $\delta_{\mathrm{m}} = C_{\mathrm{m}} s \log(p \vee n)$ by $\delta_{\mathrm{m}} = C_{\mathrm{m}} \{s \log(p \vee n)\}^{2/\gamma - 1}$. The proof will mainly rely on the replacement of Bernstein's inequality for independent data (Lemma 15) by the functionally dependent one. The following lemma is a rearranged result of the functionally dependent Bernstein's inequality (Theorem 33 in Xu et al. (2024)) using the notations of this current manuscript. The proof can be found in the Supplementary Material of Xu et al. (2024).

Lemma 25 (Bernstein's inequality under dependence) *Let $\{X_i\}_{i=1}^n$ be a consecutive subsequence of an infinite functional dependent sequence $\{X_t\}_{t \in \mathbb{Z}}$ with dependence function $\{g_t^X\}$. Assume that the sequence is with zero means, exponentially decayed cumulative function dependence measures:*

$$\sup_{m \geq 0} \exp(c m^{\zeta_1}) \Delta^X_{m,2} \leq C_1,$$

and exponentially decayed tails $\sup_{1 \leq i \leq m} \|X_i\|_{\Psi_{\zeta_2}} \leq C_2$ for some positive constants C_1 and C_2 and $\zeta = (\zeta_1^{-1} + \zeta_2^{-1})^{-1} \in (0, 1]$. Then we have for $t > \sqrt{n}$,

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^n X_i \Big| > t \Big\} \leq 2 \exp\Big\{ -c_b \Big(t^\zeta \wedge \frac{t^2}{n} \Big) \Big\}.$$

By choosing $t = c_b^{-1} C_u \{(n A_{n,p,s})^{\frac{1}{2}} \vee (A_{n,p,s})^{\frac{1}{\zeta}}\}$ with $C_u \geq c_b$, we have

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^n X_i \Big| > t \Big\} \leq 2 \exp\{ -C_u A_{n,p,s} \}.$$

When $n \geq A_{n,p,s}^{\frac{2}{\zeta}-1}$, we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n} X_i\Big| > c_b^{-1} C_u (n A_{n,p,s})^{\frac{1}{2}}\Big\} \leq 2\exp\{-C_u A_{n,p,s}\}.$$

Here $A_{n,p,s}$ is a diverging sequence. For instance, $A_{n,p,s} = \log(p \vee n)$ and $A_{n,p,s} = s\log(p \vee n)$.

By applying Lemma 25, we can obtain parallel results to those in Section C.1. For example, under Condition 6 (a), we have the restricted eigenvalue conditions in Lemma 16 hold uniformly for intervals I such that $|I| \gtrsim \{s_I \log(p \vee n)\}^{\frac{2}{\zeta}-1}$.

Lemma 26 (Uniform restricted eigenvalue condition under dependence) *Assume Condition 6 (a) holds. For any interval $I \subset (0, n]$, denote $\widehat{\Sigma}_I = |I|^{-1}\sum_{i\in I} \mathbf{x}_i\mathbf{x}_i^\top$. Uniformly for all intervals $I \subset (0, n]$ such that $|I| \geq \{s_I \log(p \vee n)\}^{\frac{2}{\zeta}-1}$, with probability at least $1 - \exp\{-C_{u,1}\log(p \vee n)\}$,*

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \geq \|\mathbf{v}\|_\Sigma^2 - C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p \vee n)}{|I|}\Big\}^{\frac{1}{2}}\Big(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2\Big),\ \forall \mathbf{v} \in \mathbb{R}^p,$$

and

$$\mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \leq \|\mathbf{v}\|_\Sigma^2 + C_{u,2}C_x^2\sigma_x^2\Big\{\frac{s_I\log(p \vee n)}{|I|}\Big\}^{\frac{1}{2}}\Big(\|\mathbf{v}\|_2^2 + \frac{1}{s_I}\|\mathbf{v}\|_1^2\Big),\ \forall \mathbf{v} \in \mathbb{R}^p,$$

where $C_{u,1}$ and $C_{u,2}$ are two universal constants. If additionally $|I| \geq C_{\mathsf{re}}s_I\log(p \vee n)$ with a sufficiently large constant $C_{\mathsf{re}} \geq 1 \vee (34C_{u,2}C_x^2\sigma_x^2/\underline{\kappa})^2$, then for any support set $\mathcal{S} \in [p]$ with $|\mathcal{S}| \leq s_I$ and $\mathbf{v} \in \mathbb{R}^p$ such that $\|\mathbf{v}_{\mathcal{S}^\complement}\|_1 \leq 3\|\mathbf{v}_{\mathcal{S}}\|_1$, under the same event above,

$$\frac{1}{2}\|\mathbf{v}\|_\Sigma^2 \leq (1 - C_{26})\|\mathbf{v}\|_\Sigma^2 \leq \mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \leq (1 + C_{26})\|\mathbf{v}\|_\Sigma^2 \leq \frac{3}{2}\|\mathbf{v}\|_\Sigma^2,$$

$$\frac{\underline{\kappa}}{2}\|\mathbf{v}\|_2^2 \leq (1 - C_{26})\underline{\kappa}\|\mathbf{v}\|_2^2 \leq \mathbf{v}^\top\widehat{\Sigma}_I\mathbf{v} \leq (\sigma_x^2 + C_{26}\underline{\kappa})\|\mathbf{v}\|_2^2 \leq \Big(\sigma_x^2 + \frac{\underline{\kappa}}{2}\Big)\|\mathbf{v}\|_2^2,$$

where $C_{26} = 17C_{u,2}C_x^2\sigma_x^2\kappa^{-1}C_{\mathsf{re}}^{-\frac{1}{2}}$.

The proof of Lemma 26 is the same as Lemma 16's by replacing the Bernstein's inequality with the functional dependence version.

Note that under Condition 4 (b), we have for any $I, R \in \mathcal{I}$, $\{\Delta_{I,q}, \Delta_{I,q,\boldsymbol{\theta}_R^\circ}\}_{q=2,4,\infty}$ are all bounded as $O(1)$. The boundness ensures that we can directly apply Lemma 25 in the following lemma.

Lemma 27 *Assume that Conditions 4 (b), 6 hold. For intervals $I, R \subset (0, n]$ with $|I| \geq \{\log(p \vee n)\}^{\frac{2}{\zeta}-1}$ and a fixed $\boldsymbol{\theta} = \boldsymbol{\theta}_R^\circ$, with probability at least $1 - n^{-2}\exp\{-C_{u,1}\log(p \vee n)\}$,*

$$\Big\|\sum_{i\in I}\{(\mathbf{x}_i\mathbf{x}_i^\top - \Sigma)(\boldsymbol{\theta}_i^\circ - \boldsymbol{\theta}) + \epsilon_i\mathbf{x}_i\}\Big\|_\infty \leq C_{27}\{|I|\log(p \vee n)\}^{\frac{1}{2}},$$

for some positive constant $C_{27} > 0$.

By replacing Lemmas 16–17 with Lemmas 26–27, we have the following oracle inequalities for lasso with functionally dependent data.

Lemma 28 (Oracle inequalities for the parametric estimates) *Assume that Conditions 4, 5 and 6 hold. We have with probability at least* $1-2\exp\{-C_{u,1}\log(p\vee n)\}$*, uniformly for any interval* $I\subset(0,n]$ *with* $|I|\geq\{C_{\mathsf{re}}s_I\log(p\vee n)\}\vee[\{s_I\log(p\vee n)\}^{\frac{2}{\zeta}-1}]$*, provided that* $\lambda_I=C_\lambda\{|I|\log(p\vee n)\}^{\frac{1}{2}}$ *for some constant* $C_\lambda>0$*, the solution* $\widehat{\boldsymbol{\theta}}_I$ *satisfies that*

$$\|\widehat{\boldsymbol{\theta}}_I-\boldsymbol{\theta}_I^\circ\|_2\leq\underline{\kappa}^{-\frac{1}{2}}\|\widehat{\boldsymbol{\theta}}_I-\boldsymbol{\theta}_I^\circ\|_\Sigma\leq C_{28}\Big\{\frac{s_I\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}},$$

$$\|\widehat{\boldsymbol{\theta}}_I-\boldsymbol{\theta}_I^\circ\|_1\leq C_{28}s_I\Big\{\frac{\log(p\vee n)}{|I|}\Big\}^{\frac{1}{2}},$$

for some constant $C_{28}>0$*. Furthermore, let* $\mathcal{S}_I$ *be the support set of* $\boldsymbol{\theta}_I^\circ$*, we have* $\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I^\complement}-\boldsymbol{\theta}^\circ_{I,\mathcal{S}_I^\complement}\|_1\leq3\|\widehat{\boldsymbol{\theta}}_{I,\mathcal{S}_I}-\boldsymbol{\theta}^\circ_{I,\mathcal{S}_I}\|_1$*.*

Lastly, we present lemmas that parallel Lemmas 19–20.

Lemma 29 *Assume that Condition 6 and the condition that* $K=O(1)$ *in Condition 4 hold. For interval* $I\subset(0,n]$ *and a fixed* $\boldsymbol{\theta}$*, with probability at least* $1-n^{-2}\exp\{-C_{u,1}\log(p\vee n)\}$*, uniformly for any sub-interval* $I\subset(0,n]$*,*

$$\Big|\sum_{i\in I}\{\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ-\boldsymbol{\theta})\}^2-\Delta_{I,\boldsymbol{\theta}}^2|I|\Big|\leq C_{u,2}C_x^2\bigg[\Delta_{I,4,\boldsymbol{\theta}}^4\vee\frac{\Delta_{I,\infty,\boldsymbol{\theta}}^4\{\log(p\vee n)\}^{\frac{2}{\zeta}-1}}{|I|}\bigg]^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}},$$

where $C_{u,1}$ *and* $C_{u,2}$ *are two universal constants.*

Lemma 30 *Assume that Condition 6 and the condition that* $K=O(1)$ *in Condition 4 hold. For interval* $I\subset(0,n]$ *and a fixed* $\boldsymbol{\theta}$*, with probability at least* $1-n^{-2}\exp\{-C_{u,1}\log(p\vee n)\}$*,*

$$\Big|\sum_{i\in I}\mathbf{x}_i^\top(\boldsymbol{\theta}_i^\circ-\boldsymbol{\theta})\epsilon_i\Big|\leq C_{u,2}C_xC_\epsilon\bigg[\Delta_{I,\boldsymbol{\theta}}^2\vee\frac{\Delta_{I,\infty,\boldsymbol{\theta}}^2\{\log(p\vee n)\}^{\frac{2}{\zeta}-1}}{|I|}\bigg]^{\frac{1}{2}}\{|I|\log(p\vee n)\}^{\frac{1}{2}},$$

where $C_{u,1}$ *and* $C_{u,2}$ *are two universal constants.*

For Lemmas 26–28, only the Bernstein's inequality (Lemma 15) needs to be replaced by Lemma 25 in the proofs. However, for Lemmas 19–20, directly applying Lemma 25 will induce upper bound with order $\{|I|\log(p\vee n)\}^{\frac{1}{2}}$, which drops the multiplier $\Delta_{I,4,\boldsymbol{\theta}}^4$ and $\Delta_{I,\boldsymbol{\theta}}^2$ in the upper bound. It is because in Lemma 25, the tail bound is scaled with n^{-1} in the term $\frac{t^2}{n}$ rather than $\frac{t^2}{\mathrm{Var}(\sum_{i=1}^n X_i)}$ or $\frac{t^2}{\sum_{i=1}^n\|X_i\|_{\Psi_{\zeta_2}}^2}$.

To solve this issue, we need the condition that $K=O(1)$ which is also assumed in Xu et al. (2024) for the functional dependence setting. For an interval I, let $\{I_j\}_{j=1}^a$ be the sub-segments of I divided by the true changepoints $\mathcal{T}^*\cap I$. Then we can apply Lemma 25 to $\{\sum_{i\in I_j}\{\mathbf{x}_i^\top(\boldsymbol{\theta}_{I_j}^\circ-\boldsymbol{\theta})\}^2-\Delta_{I_j,\boldsymbol{\theta}}^2|I_j|\}_{j=1}^a$ and $\{\sum_{i\in I_j}\mathbf{x}_i^\top(\boldsymbol{\theta}_{I_j}^\circ-\boldsymbol{\theta})\epsilon_i\}_{j=1}^a$ and then taking the union bound will lead to the deviation bounds in Lemmas 19–20.

Finally, by replacing the deviation bounds in Section C.1 with the above functionally dependent bounds, we can obtain the results in Corollary 12.

Appendix E. Additional Numerical Results

E.1 Single-Changepoint Scenario (Section 4.6)

The data in the single changepoint scenario in Section 4.6 are generated from the following model,

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}_1 \mathbf{1}\{i \le \tau^*\} + \mathbf{x}_i^\top \boldsymbol{\theta}_2 \mathbf{1}\{i > \tau^*\} + \epsilon_i,\ i = 1, \dots, n,$$

where $\{\epsilon_i\}$ and $\{\mathbf{x}_i\}$ are drawn independently satisfying $\epsilon_i \sim \mathcal{N}(0,1)$ and $\mathbf{x}_i \sim \mathcal{N}_p(0, \Sigma)$. Here Σ is a $p \times p$ matrix with elements $\Sigma_{ij} = 1/2^{|i-j|}$. The regression parameters of the model are set to be

$$\boldsymbol{\theta}_1 = (1/3, 1/3, 1/3, 1/3, 0, \dots, 0)_{p\times 1}^\top$$

and

$$\boldsymbol{\theta}_2 = (\mathbf{0}_{1\times 4}, 1/3, 1/3, 1/3, 1/3, 0, \dots, 0)_{p\times 1}^\top.$$

We set $n = 1200$ and the true changepoint $\tau^* = 120$.

E.2 Extended Comparison with the Two-Step Method

In Figures 9–10, we provide the complementary numerical results of the multiple changepoint scenarios in Section 4.6 with n varying from $n = 300$ to $n = 1200$ under both the high-dimensional linear model and the univariate nonparametric model. In the Reliever method, we set $r = 0.9$ as recommended. The Reliever provides almost comparable performance with the original algorithm in all the cases.

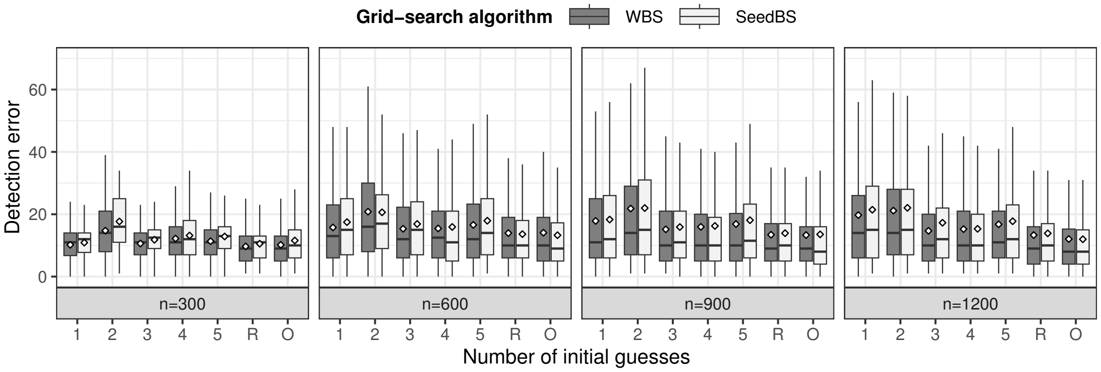


Figure 9: Comparison of the modified two-step approach with multiple initial guesses (1-5), the Reliever method (R) and the original full model-fitting (O), under the setting in Section 4.1.

E.3 Impact of a Small Minimal-Spacing Parameter

For numerical stability, we choose $\delta_{\mathrm{m}} = 20$ in all simulations in the main text. Here we rerun experiments with much smaller values. In high-dimensional linear models, setting $\delta_{\mathrm{m}} = 3$ (the smallest value accepted by glmnet without runtime errors) still works effectively—see Figure 11. The results closely match those with larger δ_{m}, confirming that Reliever is robust to this choice.

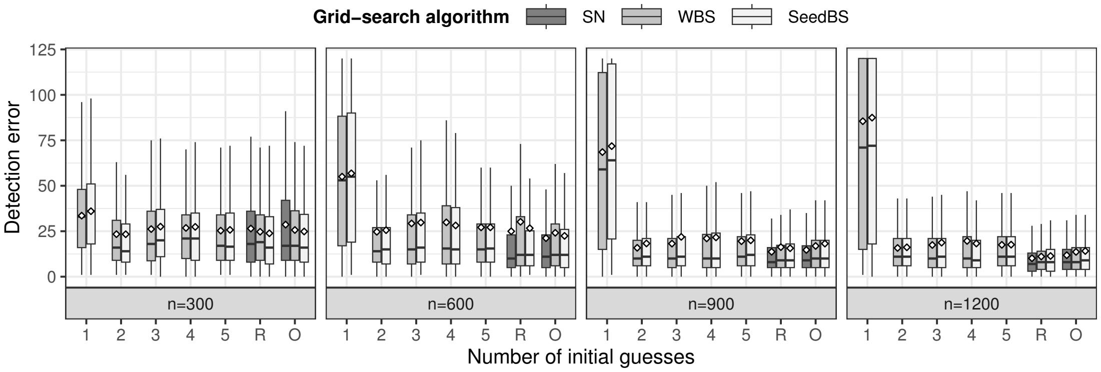


Figure 10: Comparison of the modified two-step approach with multiple initial guesses (1-5), the Reliever method (R) and the original full model-fitting (O), under the setting in Section 4.2.

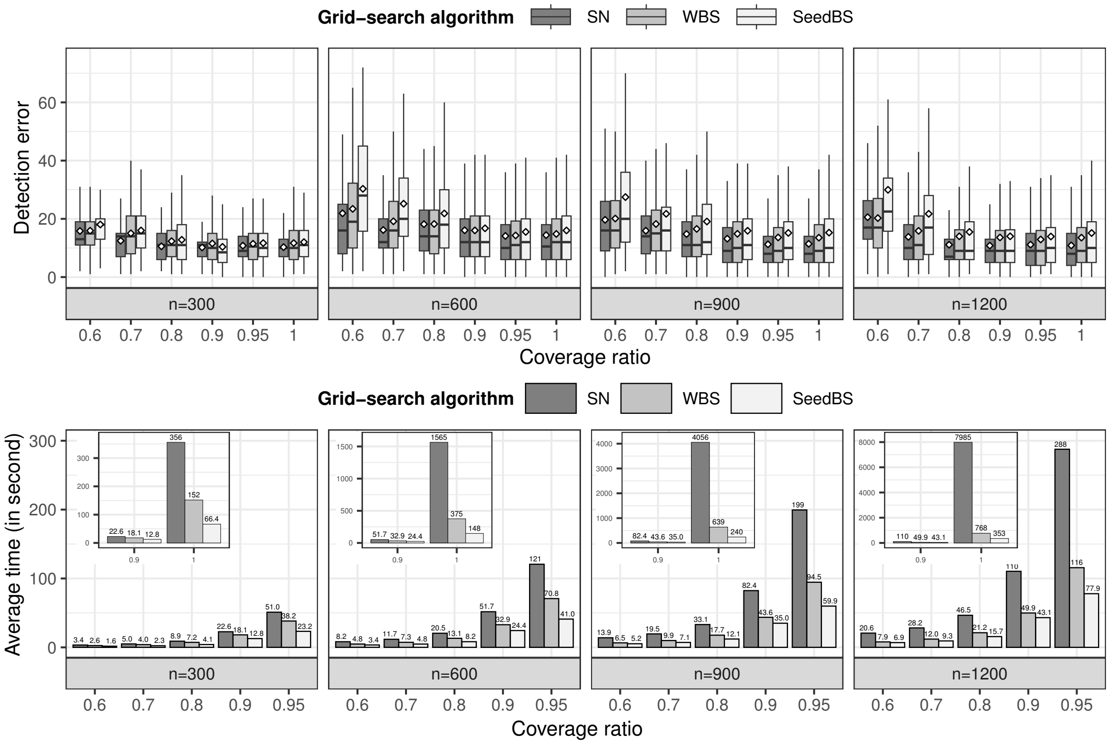


Figure 11: Performance of grid-search algorithms with Reliever under a high-dimensional linear model using a small minimum spacing $\delta_{\mathrm{m}} = 3$.

E.4 High-Dimensional Linear Model with Temporal Dependence

We conduct simulations for the high-dimensional linear setting of Section 4.1 under an AR(1) dependence ($\rho = 0.3$) in both covariates and noise, following Xu et al. (2024):

$$\mathbf{x}_i = \rho \mathbf{x}_{i-1} + \sqrt{1-\rho^2}\widetilde{\mathbf{x}}_i, \quad \epsilon_i = \rho \epsilon_{i-1} + \sqrt{1-\rho^2}\widetilde{\epsilon}_i,$$

with $\widetilde{\mathbf{x}}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ and $\{\widetilde{\epsilon}_i\} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Figure 12 shows that, even under dependence, Reliever matches the accuracy of the original algorithms while greatly reducing runtime.

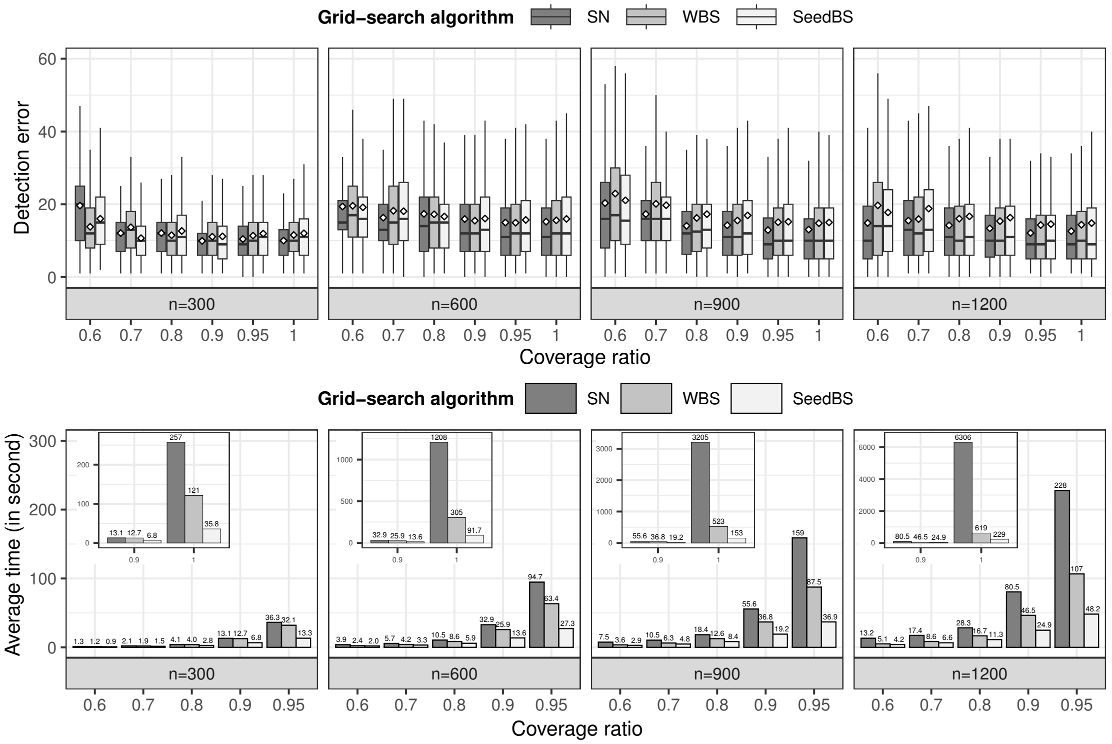


Figure 12: Performance under temporal dependence in high-dimensional linear models.

E.5 A Priori Estimation of Model-Fitting Time

In settings where the grid-search intervals can be anticipated, total fitting time for Reliever can be approximated as follows:

- *Build relief layers.* Follow Definition 1, form the layers $\mathcal{R}_k$ so that $\mathcal{R} = \bigcup_{k=0}^{\lfloor \log_b \{(1+w)n/\delta_{\mathrm{m}}\} \rfloor} \mathcal{R}_k$.
- *Sample a few intervals.* From each layer $\mathcal{R}_k$, draw c intervals (with replacement if $|\mathcal{R}_k| < c$).
- *Measure single-fit times.* Fit the model on those samples to obtain $\widehat{\mathsf{time}}_k$, the average time per fit in layer k.
- *Count expected visits.* Estimate num_k, the number of layer-k intervals the grid-search algorithm will actually visit.
  - For OP or SN, num_k follows directly from their deterministic schedules.
  - For WBS or SeedBS, generate the wild/seeded intervals $\mathcal{W}$ and set num_k as the size of $\{R \in \mathcal{R}_k : \exists (a, b] \in \mathcal{W} \,\&\, t \in (a, b], R \text{ is the relief interval of } (a, t] \text{ or } (t, b]\}$.
  - For PELT, use the OP count as an upper bound when pruning is inactive.
- *Total time estimate.* Compute $\sum_k \widehat{\mathsf{time}}_k \times \mathsf{num}_k$.

Table 9: Estimated and realized model-fitting times (in seconds) for the changepoint detection in the high-dimensional linear model in Section 4.1 with the grid search methods including SN, WBS and SeedBS, under a single run with $n = 1200$.

Method	r	0.5	0.6	0.7	0.8	0.9	0.95	0.97	0.98	1.0
SN	Estimate	3.1	6.8	11.2	25.3	94.1	276.6	591.1	1000.0	7011.1
	Realization	3.0	6.1	9.6	22.3	80.7	267.2	600.9	979.6	7617.3
WBS	Estimate	2.7	5.0	10.3	18.0	55.9	131.4	240.5	313.4	889.7
	Realization	2.9	5.1	9.1	19	56.4	131.5	241.8	299.4	885.8
SeedBS	Estimate	2.0	4.2	6.9	10.2	23.5	48.1	74.6	96.8	221.2
	Realization	2.6	3.9	6.5	11.9	27.6	52.5	76.9	108.3	235.5

Using $c = 1$, we applied this recipe to SN, WBS, and SeedBS on a high-dimensional linear model; Table 9 shows the estimated versus actual fitting times, which match closely. Users can therefore select r to match a target runtime with reasonable confidence.

References

Ivan E. Auger and Charles E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1):39–54, 1989.

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

Peiliang Bai, Abolfazl Safikhani, and George Michailidis. Multiple change point detection in reduced rank high dimensional vector autoregressive models. *J. Amer. Statist. Assoc.*, 118(544):2776–2792, 2023.

Rafal Baranowski, Yining Chen, and Piotr Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 81(3):649–672, 2019.

Hock Peng Chan and Guenther Walther. Detection with the scan and the average likelihood ratio. *Statist. Sinica*, 23(1):409–428, 2013. ISSN 1017-0405,1996-8507.

Hao Chen and Lynna Chu. Graph-based change-point analysis. *Annu. Rev. Stat. Appl.*, 10: 475–499, 2023. ISSN 2326-8298.

Hui Chen, Haojie Ren, Fang Yao, and Changliang Zou. Data-driven selection of the number of change-points via error rate control. *J. Amer. Statist. Assoc.*, 118(542):1415–1428, 2023.

Haeran Cho and Claudia Kirch. Data segmentation algorithms: Univariate mean change and beyond. *Econom. Stat.*, page To appear, 2021.

Haeran Cho and Dom Owens. High-dimensional data segmentation in regression settings permitting temporal dependence and non-gaussianity. *Electronic Journal of Statistics*, 18 (1):2620–2664, 2024.

Miklós Csörgő and Lajos Horváth. *Limit theorems in change-point analysis.* John Wiley & Sons, Ltd., Chichester, 1997.

Birte Eichinger and Claudia Kirch. A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564, 2018.

Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *J. Amer. Statist. Assoc.*, 114(525):169–183, 2019. ISSN 0162-1459. doi: 10.1080/01621459.2017.1385466. URL `https://doi.org/10.1080/01621459.2017.1385466`.

Bertille Follain, Tengyao Wang, and Richard J. Samworth. High-dimensional changepoint estimation with heterogeneous missingness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 84 (3):1023–1055, 2022.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580, 2014. With discussion.

Jerome H. Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

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

Ning Hao, Yue Selena Niu, and Heping Zhang. Multiple change-point detection via a screening and ranking algorithm. *Statistica Sinica*, 2013.

Kaylea Haynes, Paul Fearnhead, and Idris A. Eckley. A computationally efficient nonparametric approach for changepoint detection. *Stat. Comput.*, 27(5):1293–1305, 2017.

Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Proc. Let.*, 12 (2):105–108, 2005.

Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after changepoint detection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 84(4):1082–1104, 2022.

Feiyu Jiang, Zifeng Zhao, and Xiaofeng Shao. Modelling the COVID-19 infection trajectory: a piecewise linear quantile trend model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 84(5): 1589–1607, 2022. ISSN 1369-7412.

Abhishek Kaul, Venkata K. Jandhyala, and Stergios B. Fotopoulos. Detection and estimation of parameters in high dimensional multiple change point regression models via ℓ_1/ℓ_0 regularization and discrete optimization. *arXiv preprint*, art. arXiv:1906.04396, 2019a.

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

Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, 107(500):1590–1598, 2012.

Solt Kovács, Housen Li, Peter Bühlmann, and Axel Munk. Seeded binary segmentation: A general methodology for fast and optimal changepoint detection. *Biometrika*, page To appear, 2022.

Sokbae Lee, Myung Hwan Seo, and Youngki Shin. The lasso for high dimensional regression with a possible change point. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(1):193–210, 2016.

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

Jie Li, Paul Fearnhead, Piotr Fryzlewicz, and Tengyao Wang. Automatic Change-Point Detection in Time Series via Deep Learning. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, page qkae004, 01 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae004. URL https://doi.org/10.1093/jrsssb/qkae004.

Wanshan Li, Daren Wang, and Alessandro Rinaldo. Divide and conquer dynamic programming: An almost linear time change point detection methodology in high dimensions. In *International Conference on Machine Learning*, pages 20065–20148. PMLR, 2023.

Lang Liu, Joseph Salmon, and Zaid Harchaoui. Score-based change detection for gradient-based learning machines. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4990–4994. IEEE, 2021.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 2012.

Malte Londschien, Solt Kovács, and Peter Bühlmann. Change-point detection for graphical models in the presence of missing values. *J. Comput. Graph. Statist.*, 30(3):768–779, 2021.

Malte Londschien, Peter Bühlmann, and Solt Kovács. Random forests for change point detection. *Journal of Machine Learning Research*, 24(216):1–45, 2023. URL http://jmlr.org/papers/v24/22-0512.html.

Zhiyuan Lu, Moulinath Banerjee, and George Michailidis. Intelligent sampling for multiple change-points in exceedingly long time series with rate guarantees. *arXiv preprint*, art. arXiv:1710.07420, 2017.

Wojciech Niemiro. Asymptotics for M-estimators defined by convex minimization. *Ann. Statist.*, 20(3):1514–1533, 1992.

Carlos Misael Madrid Padilla, Haotian Xu, Daren Wang, Oscar Hernan Madrid Padilla, and Yi Yu. Change point detection and inference in multivariate non-parametric models under mixing conditions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 21081–21134, 2023.

Alessandro Rinaldo, Daren Wang, Qin Wen, Rebecca Willett, and Yi Yu. Localizing changes in high-dimensional regression models. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 2089–2097. PMLR, 2021.

Abolfazl Safikhani and Ali Shojaie. Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *J. Amer. Statist. Assoc.*, 117(537): 251–264, 2022.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist.*, 49(1):203–232, 2021a. ISSN 0090-5364.

Daren Wang, Zifeng Zhao, Kevin Z. Lin, and Rebecca Willett. Statistically and computationally efficient change point localization in regression settings. *J. Mach. Learn. Res.*, 22(248):1–46, 2021b.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

Haotian Xu, Daren Wang, Zifeng Zhao, and Yi Yu. Change-point inference in high-dimensional regression models under temporal dependence. *The Annals of Statistics*, 52(3):999–1026, 2024.

Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, 42(3):970–1002, 2014.