Large Dimensional Change Point Detection with FWER Control as Automatic Stopping

Jiacheng Zou [1] **Yang Fan** [2] **Markus Pelger** [1]

Abstract

We propose a statistical inference method for detecting change points in time-series of large panel data. The change points can have a general impact on different subsets of the panel. Our novel statistical perspective for high-dimensional change point detection combines selective inference and multiple testing. Our easy-to-use and computationally efficient procedure has two stages: First, LASSO regressions for each time-series screen a candidate set of change points. Second, we apply post-selection inference with a novel multiple testing adjustment to select the change points. Our method controls for the panel family-wise error rate with theoretical guarantees; hence guarding against p-hacking without the need for tuning parameters. In extensive simulations, our method outperforms leading benchmarks in terms of correct selections and false discovery. We have higher detection and make fewer Type I errors, leading to over 20% higher F1 classification scores.

1 Introduction

Time series change point detection (CPD) is increasingly faced with high dimensions, owing to the growing availability of repeated observations from numerous series. In a variety of domains, such as industrial automation systems, personal health tracking, and financial markets, we consistently monitor a wide range of metrics over time to pinpoint unusual events that could potentially induce structural breaks in the outcome variables of interest. To effectively detect an unknown number of change points across many time series, CPD algorithms must avoid spurious detections and "p-hacking" documented in (Brodeur et al., 2020), while trying to identify as many change points as possible. In this paper, we present a theoretical framework that balances this trade-off.

Our approach introduces theoretical novelty by integrating multiple testing with post-selection inference, creating a two-stage method that, to the best of our knowledge, represents a new approach to CPD problem. In the first stage, we employ LASSO screening to reduce the set of change points under consideration. In the second stage, we design **Panel Multiple Testing** to provide a theoretical control on the Family-Wise Error Rate (FWER), which acts as a stopping criterion and avoids false discovery. This attribute also lends full interpretability to our stopping criterion. The Panel Multiple Testing yields different thresholds for different potential change points by combining evidence cross-sectionally in a data-driven fashion.

Our approach outperforms leading benchmarks by accurately identifying a larger number of change points while making fewer false discoveries in simulations. In terms of performance, our method achieves a higher F1 score, demonstrating its superior balance between Type I and Type II errors. This advantage is evident even in scenarios with low Signal-to-Noise Ratio (SNR) and when structural breaks

Table 1: Performance comparison

Method	# Selected	# Correct	F_1
Panel Multiple Testing	**11.4**	**10.0**	**0.94**
Union rDP	50.5	9.9	0.33
Majority Voting rDP	4.9	4.8	0.64
Panel rDP	1.5	1.4	0.24
SBS	3.8	3.4	0.49
DCBS	6.3	5.6	0.69

This table compares the accuracy of our method with leading benchmarks. Row 1: our method, Panel Multiple Testing. Rows 2 to 6: leading benchmark methods in the multiple CPD literature. The setting involves 10 true shocks dispersed among 300 time steps. The metrics are averaged over 100 simulations with the FWER set to 0.05. Breaks are drawn from a piece-wise constant jump process of similar magnitude, with an expected SNR of approximately 0.5 and all units might observe breaks ($q_0 = 1$). More comprehensive details about the methods can be found in Section 3 and Table 3. Further specifics of the simulation are available in Section 4.

[1]Department of Management Science and Engineering, Stanford University, Stanford CA, USA [2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford CA, USA. Correspondence to: Jiacheng Zou <jiachengzou@stanford.edu>.

Accepted to ICML workshop on Spurious Correlations, Invariance, and Stability, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

affect random subsets of the cross-section. Table 1 provides a preview of these results.

The outline of the paper is as follows. Section 2 describes the first stage of our procedure that uses LASSO to propose a heterogeneous set of change points across the panel. Section 3 completes our procedure by introducing the second stage test that takes LASSO screened periods as input and selects a subset of change points to jointly explain the entire panel as output. In section 4, we examine our method's performance through an extensive series of simulations. Section 5 concludes. Proof of the multiple testing procedure is included in Appendix. The complete two-stage procedure is described in Procedure 1. In addition, Appendix B.5 discusses the computing efficiency of our method.

Related literature CPD is a very rich body of literature that is also growing rapidly. For a comprehensive review, (Aminikhanghahi & Cook, 2017) surveys the literature, and (Truong et al., 2020)'s survey focuses on the offline setting. Here we provide a limited review of works that are either closest to our setting, or provide a contrast that clarifies our scope. We consider that in many practical use cases, CPD is to select structural breaks as argued by (Carlstein et al., 1994). Our work is most closely related to (Levy-Leduc & Harchaoui, 2007) and (Harchaoui & Lévy-Leduc, 2010), as we both use LASSO to propose a sparse candidate set of change points, but our work is optimized for multiple series. Furthermore, our method as an inference procedure with a stopping rule tied to the Type I error budget focuses on interpretation, while (Levy-Leduc & Harchaoui, 2007) follows a long line of literature dating back to (Bellman, 1961) that uses dynamic programming (DP) with a reduced error rate to stop admission and focuses on goodness of fit. (Xie et al., 2012) and (Cho & Fryzlewicz, 2015) propose multiple sparse CPD in high-dimensional time series with consistency, while we focus on structural breaks in the mean. We focus on offline CPD, as opposed to online streaming setting (Chen et al., 2022). In our simulation studies, we consider three distinct heuristics that adapt (Levy-Leduc & Harchaoui, 2007) to the panel setting to draw a performance comparison, as well as directly comparing our method with (Cho & Fryzlewicz, 2015) and (Cho, 2016).

2 Regression model and LASSO screening

2.1 Outcome dynamics

Fundamentally, CPD is challenging because there are as many candidates for the change points as there are number of time periods. We consider it an even more challenging setup where structural breaks can occur for many units in a general pattern. We begin by describing the high-dimensional data-generating process (DGP).

Outcome Y_{it} is observed by ith unit at t time. We denote a change point at time t as β_{it}. In other words, if $\beta_{it} \neq 0$, there is a break in the mean at time t. There are N units and T periods in total. When written as a vector, $\beta_i \in \mathbb{R}^T$ has many entries that are zero and only structural breaks are nonzero. Similar to (Levy-Leduc & Harchaoui, 2007), we construct the accumulator matrix $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{T\times T}, \quad \text{with} \quad X_t = [\underbrace{1,\cdots,1}_{t},\underbrace{0,\cdots,0}_{T-t}]^\top. \tag{1}$$

With the notations above, we assume a DGP that admits an intuitive decomposition of a parametric dynamic and a jump dynamic:

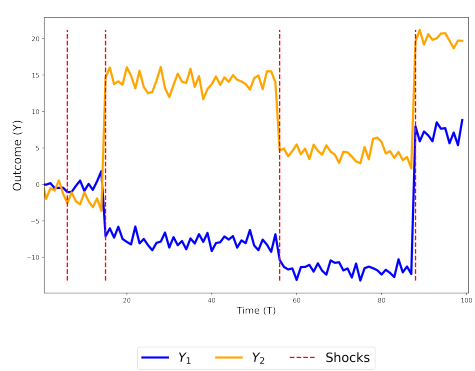
ASSUMPTION 1 (Data generating process). ***(a) Decomposition:*** *The observed outcomes Y_{it} is composed of a mean-shifting process driven by Z_t, mean structural breaks β_i, and a noise process ε_{it}:*

$$Y_{it} = Z_t^\top \gamma_i + X_t^\top \beta_i + \varepsilon_{it} \tag{2}$$

(b) Normal noise: *The noise $\varepsilon_{it} \sim N(0, \sigma^2)$.*
(c) Independence: *Exogenous Z_t, breaks β_i and noise ε_{it} are independent.*

In other words, the parametric mean-shifting process can be modeled by observed variables $\{Z_t\}_t$, and different units are driven by different weights $\{\gamma_i\}_i$. The mean structural breaks explain all other changes of the first moment, and the residuals are white noise. It is worth noting that Assumption 1(b) does not require independence across i or t. The sample paths are visualized in Figure 1.

Figure 1: Example of multiple time series with multiple change points



As an illustration, we show 2 series. Red vertical lines: actual change points.

We focus on the case where shocks are rare. Other than a sparsity assumption, the change points can be extremely

general. For example, any shock could randomly impact $q_0 \cdot N$ units – when $q_0 = 1$, all units are shocked; when $q_0 = 0.5$, only 50% of the units are shocked. We also allow the shock magnitude to differ across the units, i,e. $\beta_{it} \neq \beta_{jt}$ when $i \neq j$ and $\beta_{it} \neq 0, \beta_{jt} \neq 0$. The condition of rare change points is formalized in the following assumption:

ASSUMPTION 2 (Sparse CPD). *Let* $s = \{t : \beta_{it} \neq 0, \forall i\}$. *Then* $|s| = O(1)$ *and* $|s| \ll \min\{T, N\}$.

We have no knowledge of the number of change points s a priori and aim at optimal recovery of s multiple change points. This concludes the introduction of our setting, and we next discuss how to use LASSO first to reduce the dimensionality of the problem.

2.2 LASSO screening

Naively, there are T possible change points, so naturally we are interested in reducing the number of candidate change points in our consideration set. First, by Assumption 1, we run OLS for ith unit on $\boldsymbol{Z}$ to project out the mean-shifting process explained by auxiliary variables. Then to reduce the dimensionality of change points under consideration, we fit unit-level LASSO on the residual for each time-series. Specifically, we apply ℓ_1 penalty to estimate breaks β:

$$\hat{\beta}_i = \arg\min_{\beta} \frac{1}{2T}\|Y_i - \boldsymbol{Z}\hat{\gamma}_i - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1, \quad \text{where} \quad \hat{\gamma}_i = \arg\min_{\gamma} \|Y_i - \boldsymbol{Z}\gamma\|_2^2. \tag{3}$$

LASSO is instrumental in reducing the number of candidates by leveraging sparsity. After the LASSO fit, the active variables of $\hat{\beta}_i$ correspond to atypical means time periods, serving as a lower-dimensional set of time periods to consider in the CPD. This screening process is also utilized in (Levy-Leduc & Harchaoui, 2007) for a single series, but to manage the problem across all series, we need to introduce a new formalism for cross-sectional analysis.

We provide the formalism by writing down a joint hypothesis. The most intuitive hypothesis would be to consider all LASSO active times for any unit, which is exactly the "data-driven hypothesis" H_D that is the combination of LASSO active sets, as in $H_D = \bigcap_{i,t:\hat{\beta}_{it} \neq 0} \{\beta_{it} = 0\}$. We provide more details and examples on H_D in Appendix B.1.

We conceptualize the joint hypothesis H_D as follows: If we can reject any $\beta_{it} = 0$ within H_D, we then regard t as a chosen change point. This strategy readily accounts for scenarios where some of the N units may not record the shocks. If we can convincingly reject any of these $\beta_{it} = 0$ instances while they are part of H_D, we have effectively selected t as a change point.

3 Panel joint selection

We now discuss the key multiple testing steps of our method. First, let us review the formalism of data-driven tests and the appropriate metric of false discovery that we are controlling for under the data-driven hypothesis. Specifically, we are only discussing test procedures that test on H_D, which is a data-driven hypothesis. Since we are only conducting data-driven tests, we also require definitions 1 and 2, which deviate from the classical concepts that date back to (Bonferroni, 1935). Suppose a data-driven testing procedure rejects the hypothesis $\hat{H}$, and the true non-null hypotheses within H_D are $\bar{H}$, then the number of false selections under H_D is defined below.

DEFINITION 1 (False selections). *Number of false selections of a data-driven test is* $V(\hat{H}, \bar{H}, H_D)$, *which is the cardinality of the set of falsely-selected nulls* $\{H_{ij} : H_{ij} \in \hat{H}, H_{ij} \notin \bar{H}, H_{ij} \in H_D\}$.

Note that by their respective definitions, we must have $\hat{H} \subseteq H_D$ since we never test hypotheses that are not in H_D. This formalizes the screening procedure carried out by (Levy-Leduc & Harchaoui, 2007). We write $V(\hat{H}, \bar{H}, H_D)$ as V for brevity when there is no ambiguity. We also define the classical false-discovery metric of FWER, but in the context of data-driven hypotheses:

DEFINITION 2 (FWER). *Family-wise Error Rate (FWER) under the* H_D *is* $\mathbb{P}_{H_D}(V \geq 1)$.

For instance, when we state the selection procedure stops at FWER control rate α, which is commonly set to 5%, we are referring to the probability of making any false discovery within H_D as less than 5%.

In our methodology, we utilize p-values derived from LASSO. We establish the assumption of validity of the p value, a concept widely used in the literature on multiple testing, such as in (Ramdas et al., 2019) and (Vovk et al., 2022). Only p-values that meet this validity criterion should be given as input for our procedure. For our CPD purpose, it is sufficient to use valid post-LASSO p-values, calculated using the methods illustrated in (Tian & Taylor, 2018). We will delve into more detail after introducing Assumption 3. Our primary interest lies in selecting change points via panel multiple testing, hence we encapsulate the detailed discussions on post-LASSO inference by outlining the assumed characteristics of valid p-values below.

ASSUMPTION 3 (Valid p-values). *Individual p-value* p_{it} *satisfies* $\mathbb{P}_{H_D}(p_{it} \leq x) \leq x$, *for* $x \in [0, 1]$.

In the most straightforward case of a single outcome and low-dimensional Ordinary Least Square regression, the traditional p-value derived from t-statistic would adhere to Assumption 3 and be a valid p-value. This represents the exact satisfaction of the inequality in the assumption, as it

means $\mathbb{P}_{H_0}(p \leq x) = x$ because p follows a Uniform[0,1] under the point null hypothesis, H_0. In the realm of high-dimensional statistics, if there is finite-sample bound on the convergence of point-wise pivotal statistic, such as the post-selection p-value for LASSO as demonstrated in Theorem 9 in (Tian & Taylor, 2018), it would likewise fulfill Assumption 3. We directly employ (Tian & Taylor, 2018) to generate post-LASSO p-values after executing Equation (3), hence our p_{it} values comply with Assumption 3.

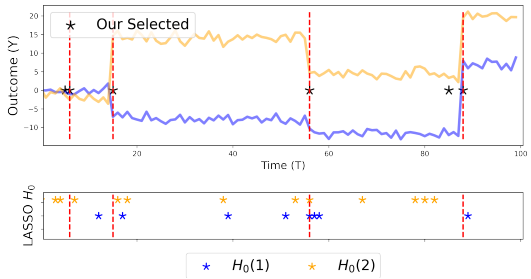
We propose using the following two-stage procedure to select change points in $\boldsymbol{Y}$ controlling for $\boldsymbol{Z}$:

PROCEDURE 1. Change point selection with FWER

1: ***Input:*** *Series $\boldsymbol{Y}$, auxiliary variables $\boldsymbol{Z}$, FWER target α.*
2: ***for*** $i = 1$ ***to*** N ***do***
3: *Run LASSO regression (3) with Y_i and $\boldsymbol{Z}$ to acquire $\hat{\beta}_i$.*
4: *For $\hat{\beta}_{it} \neq 0$, use (Tian & Taylor, 2018) to retrieve p-values: p_{it}.*
5: ***end for***
6: ***for*** $t = 1$ ***to*** T ***do***
7: ***if*** $\hat{\beta}_{it} \neq 0$ *for any i* ***then***
8: *Set $p_t :=$ smallest p_{it} for $\forall i$.*
9: *Calculate simultaneity count N_t.*
10: ***else***
11: *Move onto next t.*
12: ***end if***
13: ***end for***
14: *Calculate panel cohesion coefficient ρ.*
15: *Sort $\rho^{-1}\frac{p_t}{N_t}$ and select $\hat{H} = \{t : \rho^{-1}\frac{p_t}{N_t} \leq \alpha\}$ as change points.*
16: ***Output:*** *Selected times $\hat{H}$ as change points.*

In Figure 2, we provide a visualization of the second stage. In addition, we provide a line-by-line explanation of our procedure: lines 2∼5 outline the first stage of the process, where the LASSO screening is applied to each of the N series. Lines 6∼15 constitute the second stage and implement our selection method by scanning across time periods. Several quantities come into play within our selection method: N_t acts as a local parameter and a simultaneity count that captures how active tth time period is active across all LASSOs, subject to a novel panel localization procedure to reduce overcounting and enhance power (see more in Appendix B.2). ρ is a panel parameter and is interpreted as the panel cohesion coefficient, which is a scalar normalization term between [0,1] that reflects the potential disparity in the active and nonactive sections of the panel. More detailed discussions and interpretations of these relevant quantities are deferred to Appendix B.2.

Figure 2: Second Stage Multiple Testing



Red vertical lines: actual change points. Top subplot: Black stars: Change points selected by our Panel Multiple Testing method. Bottom subplot – LASSO screened change point candidates; different colors correspond to different units.

Our selection step, as described in line 15, is a panel-modified Bonferroni argument with its proof also an algebraic union bound, paralleling the original Bonferroni idea. Thanks to the union bound argument, we allow the p_{it}'s to be arbitrarily dependent, i.e. we allow complex cross-unit dependency.

The rejection procedure picks as many change points as it can, while FWER is controlled, as formalized in the following theorem. The proof is detailed in the Appendix A.

THEOREM 1. *With Assumptions 1, 2 and 3, Procedure 1 selects the most likely change points with FWER$\leq \alpha$ under null hypothesis H_D.*

4 Simulation

In this section, we perform extensive simulations in challenging settings of many units and weak signals as measured by the partial impact rate q_0 and the signal-to-noise ratio (SNR). We then compare our method against leading benchmarks.

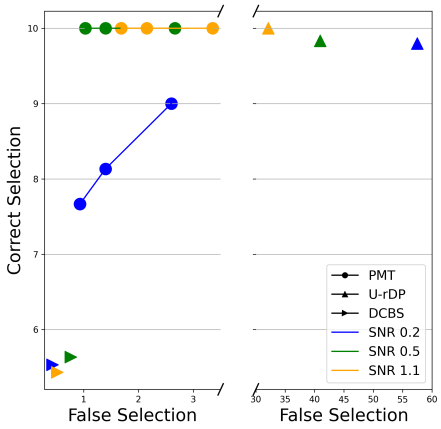
4.1 Setup

Basics Throughout the experiments, we fix $T = 300$ time periods, $N = 200$ series and $s = 10$ true structural breaks. The simulations are repeated 100 times to produce averaged metrics. We adopt an approach similar to the one outlined in (Cho & Fryzlewicz, 2015)'s M4 to generate breaks via a piece-wise constant mean jump process. This method is described in detail in Appendix B.4.

Settings We consider both $q_0 = 1$ and $q_0 = 0.5$. The latter case is that for each change point, only 50% of the N series were impacted, as reported in Table 1. Note that $q_0 = 0.5$ means $\beta_{it} \neq 0$ only appears in half of the cross section. We also consider expected SNR varying from 0.3 to 1.5, as later reported in Figure 3.

Reported Metrics We report the number of selections, the

Figure 3: False vs Correct Selections



This figure compares the correct and falce rejection of our method with the benchmarks. Dots represent our method (Panel Multiple Tetsting). Other shapes represent benchmarks (Union RDP and DCBS). Colors indicate SNR. As greater FWER is allowed, our model has more correct selections and a small increase of false selections. Breaks generated from Equation (15), where SNR 0.3, 0.7 and 1.1 correspond to $a = 2, 5, 10$, with $q_0 = 1$.

Table 2: Performance comparison: Half units impacted ($q_0 = 0.5$)

Method	# Selected	# Correct	F_1
Panel Multiple Testing	**12.0**	**10.0**	**0.91**
Union rDP	78.7	10.0	0.23
Majority Voting rDP	4.9	4.7	0.63
Panel rDP	2.4	2.3	0.36
SBS	3.8	3.0	0.43
DCBS	7.3	6.0	0.69

This table compares the accuracy of our method with leading benchmarks. Row 1: our method, Panel Multiple Testing. Rows 2 to 6: Leading benchmark methods in the multi-CPD literature. The setting involves 10 true shocks dispersed among 300 time steps. The metrics are averaged over 100 simulations with the FWER set to 0.05. Breaks are drawn from a piece-wise constant jump process of similar magnitude, with an expected SNR of approximately 0.7.

number of correct selections and F1 score[1] on selections of change points, averaged over 100 simulations.

4.2 Results

We highlight that our method has the best results measured in F_1 scores, and selects most change points while making fewer Type I errors, as can be seen in both Table 1 where $q_0 = 1$ and Table 2 for $q_0 = 0.5$.

Our method provides better trade-offs in various SNR regimes, as presented in Figure 3. That is, our method has fewer False Selections when compared to methods that make a similarly high number of Correct Selections such as Union RDP, and at the same time our method has more Correct Selections when compared to methods that make a similarly small number of False Selections such as DCBS. Each SNR setting is represented as a different color in Figure 3, and our method maintains the performance advantages across SNRs.

5 Conclusion

We have introduced a novel statistical perspective to the field of high-dimensional CDP problems, which combines post-selection inference and multiple testing, as opposed to existing algorithmic methods. In addition, we provide a formal theory for controlling panel family-wise error rate (FWER) with a theoretical guarantee, guarding against p-hacking. Our method shows superior performance in comprehensive simulations, in terms of higher detection and fewer Type I errors.

Reproducibility statement

Our code for the method and simulations are available on GitHub.

[1]The F1 score is F1 $= \frac{2}{\text{Precision}^{-1}+\text{Recall}^{-1}}$ where Precision $= \frac{\text{TP}}{\text{TP+FP}}$; Recall $= \frac{\text{TP}}{\text{TP+FN}}$.

References

Aminikhanghahi, S. and Cook, D. J. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Bellman, R. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961.

Bonferroni, C. E. Il calcolo delle assicurazioni su gruppi di teste. *In Studi in Onore del Professore Salvatore Ortu Carbon*, 1935.

Brodeur, A., Cook, N., and Heyes, A. Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–3660, 2020.

Carlstein, E. G., Siegmund, D., et al. Change-point problems. 1994.

Chen, Y., Wang, T., and Samworth, R. J. High-dimensional, multiscale online changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):234–266, 2022.

Cho, H. Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, 10 (2):2000 – 2038, 2016. doi: 10.1214/16-EJS1155. URL https://doi.org/10.1214/16-EJS1155.

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

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

Kotecha, J. H. and Djuric, P. M. Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 3, pp. 1757–1760. IEEE, 1999.

Levy-Leduc, C. and Harchaoui, Z. Catching change-points with lasso. *Advances in neural information processing systems*, 20, 2007.

Pakman, A. and Paninski, L. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.

Pelger, M. and Zou, J. Inference for large panel data with many covariates. *arXiv preprint arXiv:2301.00292*, 2022.

Ramdas, A. K., Barber, R. F., Wainwright, M. J., and Jordan, M. I. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47 (5):2790 – 2821, 2019. doi: 10.1214/18-AOS1765. URL https://doi.org/10.1214/18-AOS1765.

Tian, X. and Taylor, J. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

Truong, C., Oudre, L., and Vayatis, N. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Vovk, V., Wang, B., and Wang, R. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.

Xie, Y., Huang, J., and Willett, R. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1): 12–27, 2012.

Appendix

A Proof for Theorem 1

Proof: We show how to count false discoveries under the data-driven null hypothesis conditional on the selection and then evaluate the probability of at least one false discovery under H_D.[2] By design of H_D, we only need to consider false selections for the covariates that are active for some units. For $\mathcal{K}_t \neq \emptyset$, we denote by $t \in H_M$ that the tth time period is tested in H_M. We also use $\mathcal{M}^{(i)}$ to denote the LASSO selection event for nth unit and $\mathcal{M}$ to denote the joint LASSO selection event of all N units.

First, we separate the data-driven hypothesis into individual components. Within H_D, we consider all hypotheses associated with the tth covariate, denoted as $H_{D,t}$. Hypothesis $H_{D,t}$ represents an intermediate level of hypotheses between panel level H_D and individual null unit covariate $H_{0,t}^{(i)}|\mathcal{M}^{(i)}$. When written as a set intersection, $H_{D,t}$ equals $\bigcap_{n\in\mathcal{K}_t} H_{0,t}^{(i)}|\bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}$, that is, $H_{D,t}$ is joint over all the units where the tth covariate is active.

Second, we count the number of false discoveries. Under the null hypothesis H_D, we denote the number of false discoveries in $H_{D,t}$ as V_t. Hence, the total number of false discoveries is the sum of the false discoveries of all covariates given by

$$\mathbb{P}_{H_D}\left(V \geq 1|\mathcal{M}\right) = \mathbb{P}_{H_D}\left(\sum_{t\in H_M} V_t \geq 1 | \bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}\right). \tag{4}$$

Each V_t can be further broken down into the false discovery against the unit-covariate null hypotheses. Each individual potential false discovery is a random event, and the sum of false discoveries greater than or equal to 1 corresponds to the union of these random events. The union of these random events has the conditional distribution

$$\begin{aligned}\mathbb{P}_{H_D}\left(\sum_{t\in H_M} V_t \geq 1 | \bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}\right) =& \mathbb{P}_{H_D}\left(\bigcup_{t\in H_M}\left\{\bigcup_{n\in\mathcal{K}_i}\left\{\text{Rejection made based on } p_{it}|\bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}\right\}\right\}\right)\\ =& \mathbb{P}_{H_D}\left(\bigcup_{t\in H_M}\left\{\bigcup_{n\in\mathcal{K}_i}\left\{p_{it}\leq\rho\frac{\gamma}{N_t}|\bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}\right\}\right\}\right).\end{aligned} \tag{5}$$

The second line simply follows from the design of our rejection procedure. Boole's inequality implies the following union bound:

$$\mathbb{P}_{H_D}\left(\bigcup_{t\in H_M}\left\{\bigcup_{n\in\mathcal{K}_i}\left\{p_{it}\leq\rho\frac{\gamma}{N_t}|\bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}\right\}\right\}\right) \leq \sum_{t\in H_M}\sum_{n\in\mathcal{K}_t}\mathbb{P}\left(p_{it}\leq\rho\frac{\gamma}{N_t}|\bigcap_{n\in\mathcal{K}_t}\mathcal{M}^{(i)}\right). \tag{6}$$

Third, we take advantage of Assumption 3. Under Assumption 3, it holds that $\mathbb{P}_{H_D}(p_{it}\leq\rho\frac{\gamma}{N_t})\leq\rho\frac{\gamma}{N_t}$, which appears in the right-hand side of (6). Thus, combining equations (4)$\sim$(6) yields:

$$\mathbb{P}_{H_D}(V\geq 1|\mathcal{M}) \leq \sum_{t\in H_M}\sum_{n\in\mathcal{K}_t}\rho\frac{\gamma}{N_t} = \gamma\cdot\rho\cdot\sum_{t\in H_M}\frac{1}{N_t}\sum_{n\in\mathcal{K}_t}1 = \gamma\cdot\rho\cdot\sum_{t\in H_M}\frac{|\mathcal{K}_t|}{N_t} = \gamma. \tag{7}$$

The second-to-last equation uses and also explains the definition of ρ. This completes the proof for FWER control.

In addition, let the time periods be sorted ascendingly by $\rho^{-1}\frac{p_t}{N_t}$ into $K = [T_1', T_2', ..., T]$ such that $\frac{p_{K_l}}{N_{K_l}} \leq \frac{p_{K_l'}}{N_{K_l'}}$ if $l \leq l'$. Observe that for every α, our Procedure 1 selects all t up to $\rho^{-1}\frac{p_t}{N_t}\leq\alpha$. Thus, for each FWER threshold α, we uniquely select the top subset of K as change points. **[QED]**

B Technical details

B.1 Data-driven hypotheses H_D

We illustrate the concept of a data-driven hypothesis with a simple example, which we will use throughout this section. For simplicity, we assume that we have $T = 4$ periods and want to explain $N = 6$ cross-sectional units. In the first stage, we

[2] This proof is closely related to the arguments in proof of Theorem 2 and Corollary 1 of (Pelger & Zou, 2022) for unordered testing.

have estimated a sparse model and have obtained the post-selection valid p-values for each of the N units. We collect the fitted sparse estimator $\hat{\beta}_i$ for the nth unit in the matrix $\hat{\boldsymbol{\beta}}$. Note, that this matrix has "holes" due to the sparsity for each $\hat{\beta}_i$. Figure 4(a) illustrates $\hat{\boldsymbol{\beta}}$ for this example.

Figure 4: Illustrative example of data-driven selection

	1	2	3	4
1		-5.43	2.15	
2	-1.10		4.78	-0.08
3		0.19	4.59	
4			4.44	2.10
5		1.44	4.53	2.10
6		-0.46	4.70	

Matrix $\hat{\boldsymbol{\beta}}$

	1	2	3	4
1		0.127	0.587	
2	0.005		<0.001	0.871
3		0.526	<0.001	
4			<0.001	<0.001
5		<0.001	<0.001	<0.001
6		0.102	0.010	

Matrix $\boldsymbol{P}$ of p-values

This figure illustrates in a simple example the data-driven selection of a linear sparse model. In the first stage, we have estimated a regularized sparse linear model for each of the $N = 6$ units with $T = 4$ time periods. Each row represents the selected time periods with their estimated coefficients and p-values. The columns represent the $T = 4$ time periods. The grey shaded boxes represent the LASSO screened set, while the white boxes indicate the inactive time periods. The numbers are purely for demonstrative purposes.

Similarly, we collect the respective p values in the $\boldsymbol{P}$ matrix. For the nth unit, we have only p values for the periods that are active in the nth sparse model. Thus, in Figure 4(b), the same pattern of p-values are unavailable due to the sparse model output. Also, there are white boxes that show the same pattern of p-values that are not available. These holes can appear in different positions for each unit, making this problem non-trivial. This nontrivial shape of the (a) or (b) subplots is fully data driven and is the result of the feature screening of sparse models. We show that a hypothesis can be formed around these non-trivial shapes, which we refer to as the family of data-driven hypotheses.

We want to test which potential change points are jointly insignificant in the full panel. The data-driven hypothesis only tests the significance of the potential change points that were included in the selection and, hence, can drastically reduce the number of hypotheses. However, given the non-trivial shape of the active set, the multiple testing adjustment for the data-driven hypothesis is more challenging.

Before formally defining the families of the hypothesis, we illustrate them in our running example. The data-agnostic hypothesis H_A for explaining the full panel takes the following form:

$$\begin{aligned} H_A &= \{H_{A_{0,1}}, H_{A_{0,2}}, H_{A_{0,3}}, H_{A_{0,4}}\} \\ &= \{\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = 0, \\ &\quad \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = \beta_{26} = 0, \\ &\quad \beta_{31} = \beta_{32} = \beta_{33} = \beta_{34} = \beta_{35} = \beta_{36} = 0, \\ &\quad \beta_{41} = \beta_{42} = \beta_{43} = \beta_{44} = \beta_{45} = \beta_{46} = 0\} \end{aligned} \tag{8}$$

The data-driven hypothesis H_D only includes the active set and hence equals

$$\begin{aligned} H_D = \{&\beta_{12} = 0, \\ &\beta_{21} = \beta_{23} = \beta_{25} = \beta_{26} = 0, \\ &\beta_{31} = \beta_{32} = \beta_{33} = \beta_{34} = \beta_{35} = \beta_{36} = 0, \\ &\beta_{42} = \beta_{44} = \beta_{45}\} \end{aligned} \tag{9}$$

Clearly, H_A has a larger cardinality of $|H_A| = 24 > |H_D| = 14$. This holds in general, unless the first stage selects all periods for each unit, in which case the two hypotheses coincide.

Formally, the data-agnostic family of hypotheses is defined as follows:

DEFINITION 3. Data-agnostic family
The data-agnostic family of hypotheses is

$$H_A = \{H_{A_{0,t}} | t \in [T]\}, \quad \textit{where } H_{A_{0,t}} = \bigcap_{i \in [N]} \{\beta_{it} = 0\}. \tag{10}$$

It is evident that H_A does not need any model output or exploratory analysis, so it is indeed data-agnostic.

As soon as we use a sparsity-constrained model that has censoring capabilities, we no longer observe $(\boldsymbol{Y}, \boldsymbol{X})$ from its data-generating process. Consequently, unless our hypotheses depend on how we built the model, or equivalently on how the data was censored, the data-agnostic hypotheses forgo power without any benefit in false discovery control. Therefore, we formulate the hypothesis on the t time period if it is in the active set of the nth unit. Conditional on observing the model output, there is no inference statement to be made about LASSO inactive time periods, because its estimator is censored by the model.

We denote as $\mathcal{K}_t$ the set of units for which the t time period is active. We define the cross-sectional hypothesis for the t time period as:

$$H_{0,t} = \bigcap_{i \in \mathcal{K}_t} \{\beta_{it} = 0\} \Big| \mathcal{M}, \quad \forall t : \mathcal{K}_t \neq \emptyset. \tag{11}$$

By combining all periods $\{t : \mathcal{K}_t \neq \emptyset\}$ that show up at least once in one of the active sets of our sparse linear estimators, we arrive at a data-driven hypothesis associated with our panel. This is defined as follows:

DEFINITION 4. Data-driven family
The data-driven family of hypotheses conditional on $\mathcal{M}$ is

$$H_D = \{H_{0,t} | t : \mathcal{K}_t \neq \emptyset\}. \tag{12}$$

This demonstrates the non-trivial nature of writing down a hypothesis in the high-dimensional panel: we can only collect $\mathcal{K}_t$ - the set of units for which the t time period is active - after seeing the sparse selection estimation result.

B.2 The panel localization count and other quantities in Algorithm 1

Similar to the conventional definition, we simply count the number of Type I false rejections V, and define FWER as the probability of making at least one false rejection. Importantly, the FWER accounts for the fact that we might repeatedly test a specific covariate for multiple cross-sectional units rather than just for one unit. Our contribution to FWER control in the panel setting is thus to take into consideration both the multiplicities in units and periods when we deal with the "matrix" of p-values $\boldsymbol{P}$. To achieve this goal, we propose a new simultaneity account for the t time period, calculated as

$$N_t = \sum_{i \in \mathcal{K}_t} |M_i| \tag{13}$$

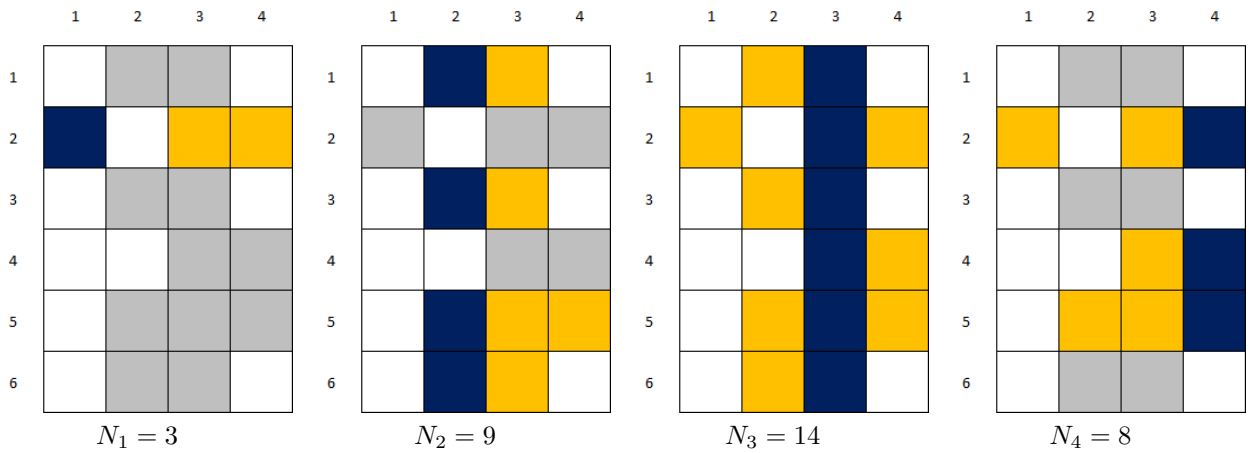
Figures 5∼5 illustrate the simultaneity counting for our running example with $N = 6$ units and $T = 4$ periods. The blue boxes represent the active set for a specific covariate. The yellow boxes indicate the "co-active" periods, which have to be accounted for in a multiple testing adjustment. In the case of the first time period as potential change point $t = 1$, only the second unit has selected this time period. This second unit has also selected time periods $t = 3$ and $t = 4$ as potential change points, which are jointly tested with the first period. Hence, they are "co-active", and the simultaneity count equals $N_1 = 3$. Intuitively, N_t represents all relevant comparisons for the t time period because it counts how many periods are active with the t time period in the regressions. Hence, N_t quantifies the number of "multiple tests" for each covariate.

In Figure 5, we see that $\mathcal{K}_1 = \{2\}$ for the 1st covariate, indicated by the blue box, because it is only active in the second unit's regression. The multiple testing adjustment needs to consider all yellow boxes, and $N_1 = 3$ is thus the total count of 1 blue and 2 yellow boxes. Similarly, for the second covariate, $\mathcal{K}_2 = \{1, 3, 5, 6\}$, so we shade boxes yellow for the

2nd, 3rd and 5th units and obtain $N_2 = 9$. We can already see that our design of simultaneity counts takes all relevant pairwise comparisons into consideration, but avoids counting the white boxes - which would cause overcounting and result in over-conservatism.

Our multiplicity counting is a generalization of the classical Bonferroni adjustment for multiple testing. A conventional Bonferroni method for the data-agnostic hypothesis H_A has a simultaneity count of $|H_A| = N \cdot T = 24$ for testing each covariate. A direct application of a vanilla Bonferroni method to the panel of all selected units and the data-driven hypothesis H_D, would use a simultaneity count of $|H_D| = 14$ for testing each covariate. Our proposed multiplicity counting is a refinement that leverages the structure of the problem, and takes the heterogeneity of the active sets for each covariate into account. Our count has only $N_1 = 3$, $N_2 = 9$ and $N_4 = 8$ for the periods $t = 1, 2$ and 4. Only for time period as potential change point $t = 3$ is the simultaneity count the same as a vanilla Bonferroni count applied to H_D, i.e. $N_3 = 14$.

Figure 5: Simultaneity counts N_t in the illustrative example



This figure shows the simultaneity counts N_t as an illustrative example. The subplots represent the simultaneity counts for the $T = 4$ time periods. The blue boxes indicate the active set, while the yellow boxes indicate the "co-active" time periods. The simultaneity counts are the sum of yellow and blue boxes.

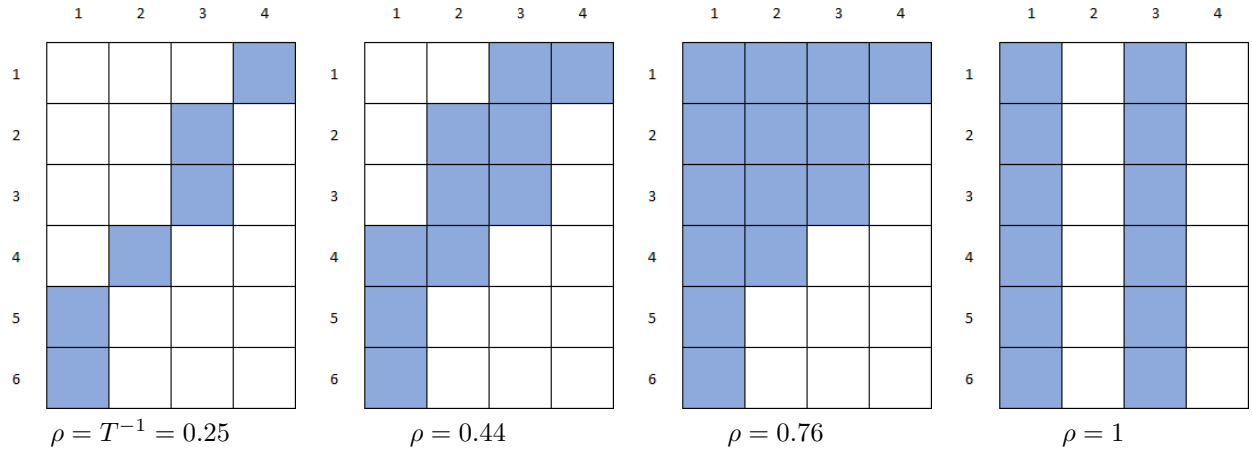
In addition to the simultaneity count of each covariate, we need an additional "global" metric for our testing procedure. We define a panel cohesion coefficient ρ as a scalar that measures how sparse or de-centralized the proposed hypotheses family is:

$$\rho = \left(\sum_{1 \leq t \leq T : \mathcal{K}_t \neq \emptyset} \frac{|\mathcal{K}_t|}{N_t} \right)^{-1} \tag{14}$$

The panel cohesion coefficient ρ is conditional on the data-driven selection of the overall panel. It is straightforward to compute once we observe the sparse selection of the panel. This coefficient takes values between T^{-1} and 1, where larger values of ρ imply that the active set is more dependent on the cross-section. This can be interpreted as the panel Y having a stronger dependency due to the covariates X. Intuitively, in the extreme case when $\rho = T^{-1}$, the panel can be separated into T smaller problems, each containing a subset of response units explained by only one period. Thus the panel would be very incohesive, and could be studied with T separate tests. In the other extreme, if ρ approaches 1, the first-stage models include the same active covariates for all units. We consider this a very cohesive panel. If ρ is between these bounds, the panel is cohesive in a non-trivial way such that some units can be explained by some covariates, and there is no clear separation of the panel into independent subproblems.

Figure 6 illustrates the panel cohesion coefficient with examples. The subplots show four active sets that are different from our running example. The left subplot shows the extreme case of $\rho = T^{-1}$, where the panel is the least cohesive. The right subplot illustrates the other extreme for $\rho = 1$, where the panel is the most cohesive. The middle subplots correspond to the complex cases of a medium cohesion coefficient.

Figure 6: Illustration of the cohesion coefficient



This figure illustrate the cohesion coefficient ρ in four examples. It shows the smallest, largest, and in-between cases of ρ. The columns represent the $T = 4$ time periods.The blue boxes indicate the active sets for each panel.

B.3 Benchmarks methods

We evaluate our method in comparison to 5 established benchmark techniques widely employed for CPD. These benchmarks encapsulate two principal categories of CPD algorithms: Dynamic Programming (DP) and sparse binary segmentation/splitting. For a tabulated summary of all the benchmark methods assessed in our study, please refer to Table 3.

Table 3: Summary of selection methods

Name	Abbreviation	Selection	Multiple Testing	Stopping rule
Panel Multiple Testing	PMT	MT	FWER control	$p^{\text{PoSI}} < \frac{\rho\gamma}{N_i}$
Panel RDP	P-RDP	RDP	None	Depends on hyper-param. J_i's
Majority Voting RDP	MV-RDP	RDP	None	Depends on $\max_i \lvert M_i \rvert$
Union RDP	U-RDP	RDP	None	Depends on hyper-param. J_i's
DCBS	DCBS	Tree	None	Depends on hyper-param.
SBS	SBS	Tree	None	Depends on hyper-param.

This table compares the different methods to estimate a set of covariates from a large dimensional panel. For each method, we list the name and abbreviation. The selection refers to the regression approach for each univariate time-series. The hypothesis is either agnostic or data-driven given the selected subset of covariates. The multiple testing adjustment includes no adjustment, a conventional Bonferroni adjustment and our novel simultaneity count for a data-driven hypothesis. The rejection rules combine the valid p-values and multiple testing adjustment. p^{PoSI} is the debiased post-selection adjusted p-value based on (Tian & Taylor, 2018).

Firstly, we use simple heuristics to adapt (Levy-Leduc & Harchaoui, 2007) to the panel setting. There are three such heuristics that we propose, as enumerated below:

1. Panel RDP (P-RDP): We choose the change points based on panel level MSRE decrease.

2. Majority voting RDP (MV RDP): Here we consider a heuristic that is also a two-stage procedure that has a LASSO first stage, and adapts RDP to the panel setting by allowing an adaptive comparison of unit-level DP results in the second stage.
   Stage 1: Proposed change points are the LASSO active points from individual series. For ith series, the proposing set $Q_i = \{Q_{i1}, ..., Q_{i,|Q_i|}\}$ is potential change points ranked by rDP from most likely to least likely with decreasing reduction of MSRE, i.e. Q_{i1} decreases MSRE more than Q_{i2} for Y_i, so on and so forth.

Stage 2: On the panel, max number of change points is $k_{max} = \max_i |Q_i|$. For $k \leq k_{max}$, all series use simple majority voting to decide which t is the selected change point[3]. The selected change points t^{MVRDP} are thus the voting results of all possible k positions. In math, this is written as: $t^{\text{MVRDP}} = \{t_k^{\text{MVRDP}} : k \leq k_{max}\}$ where $t_k^{\text{MVRDP}} = \arg\max_t \sum_{1 \leq i \leq N} \mathbf{I}(t = Q_{ik})$.

3. Union RDP: We simply union the individual series selection results, which are acquired from running rDP as described in (Levy-Leduc & Harchaoui, 2007).

It's worth noting that MV RDP and Union RDP both employ LASSO screening, making their comparison to our methodology a more direct assessment of the choice between using DP in the second stage versus our multiple testing approach for CPD.

Secondly, we include two tree-structured methods as benchmarks: Sparsified Binary Segmentation (SBS) as presented by (Cho & Fryzlewicz, 2015), and Double CUSUM Binary Segmentation (DCBS) from (Cho, 2016). Both of these are greedy search strategies intended for the detection of an unknown quantity of change points across multiple series.

B.4 Sampling of structural breaks in simulation

We sample the 10 structural breaks randomly from the 300 periods to be $\bar{H}_{Simulation}$, and randomly choose $q_0 \cdot N$ units at each change point $t \in \bar{H}_{Simulation}$ to simulate the breaks.

We adopt an approach similar to the one outlined in (Cho & Fryzlewicz, 2015)'s M4 to generate breaks via a piece-wise constant mean jump process. This method is described in details in Appendix. This method creates bounded variance when T is large, and is compatible with our DPG as delineated in 1. In particular, let the true change points sampled be denoted as $\bar{H}_{Simulation} = t_1, ..., t_s$. The breaks are generated by initially constructing a piece-wise constant mean jump process:

$$b_{it} = \begin{cases} 0, & 0 \leq t < t_1 \\ u_{i1}, & t_1 \leq t < t_2 \\ \vdots & \\ u_{i,s-1}, & t_{s-1} \leq t < t_s \\ u_{is}, & t_s \leq t < T \end{cases} \tag{15}$$

Each u_{it} is drawn from a uniform distribution Unif$[-a, a]$, where we can adjust a to create different SNR settings. We then define the breaks as $\beta_{it} = b_{it} - b_{i,t-1}$, and the magnitude of breaks are $|\beta_{it}|$ when t is a change point. Otherwise, $\beta_{it} = 0$ indicates no break. We define the expected SNR $= \mathbb{E}[\frac{|\beta_i|_2}{|\varepsilon_i|_2}]$.

B.5 Computational cost of the selection method

We now turn to the analysis of the computational efficiency of our selection method (lines 6∼15) as laid out in Procedure 1. The first calculations relevant to selection are the counting processes of N_t in lines 8 and 9, which involves $O(\sum_n |M_n|)$ calculations in total. Suppose the true number of change points is $s = O(1)$, then by LASSO's selection property in large T as discussed in (Harchaoui & Lévy-Leduc, 2010), we have $O(N)$ for lines 8 and 9 as total for all t. Secondly, the panel cohesion coefficient ρ calculation of line 14 is $O(T)$. Lastly, the sorting and rejection in line 15 is $O(T \log T)$. In conclusion, our selection takes $O(\max\{T \log T, N\})$.

However, it is worth pointing out that the above calculation treats the acquisition of p-values as outside of the selection method since they are calculated in the first-stage of the procedure. To calculate p-value post-LASSO via methods of (Tian & Taylor, 2018) is fundamentally about calculating the cumulative distribution function of truncated Gaussians, for which there are many efficient implementations such as (Kotecha & Djuric, 1999) and (Pakman & Paninski, 2014). We consider the discussion of its computational efficiency outside the scope of our algorithm.

[3]Uniform random tie-breaking is used. If ith series does not have Q_{ik}, it does not participate in the voting.