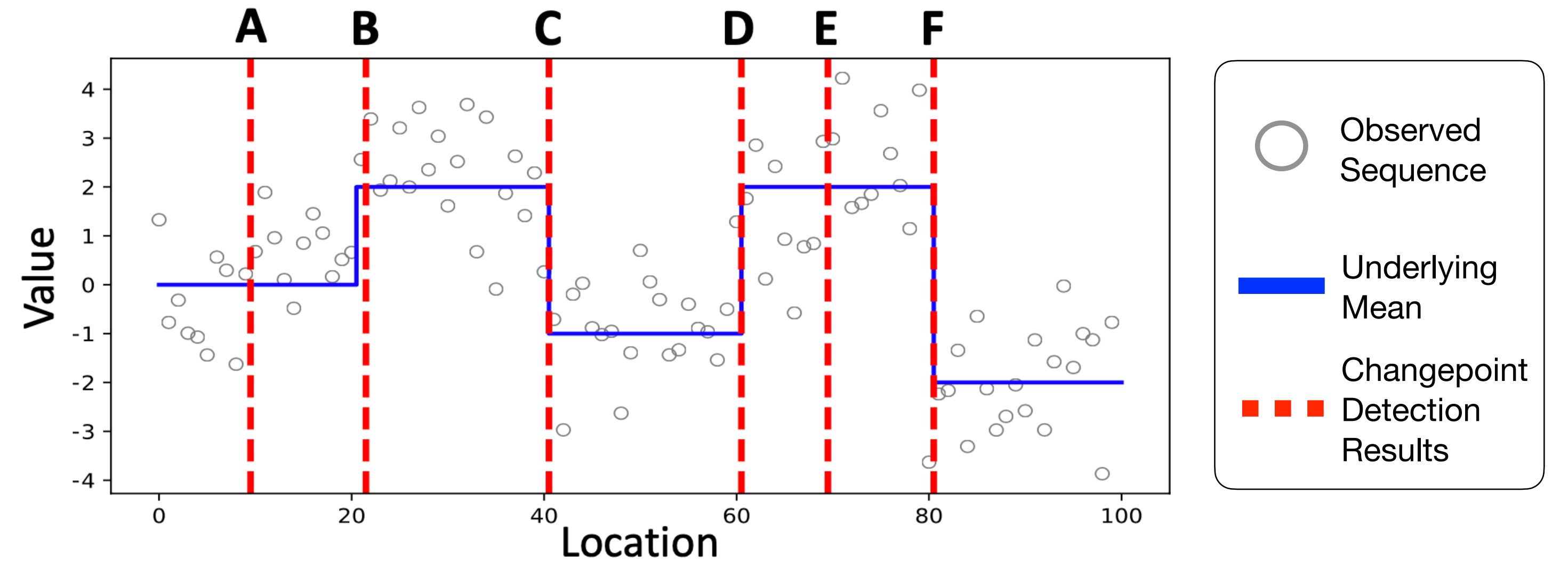
Computing Valid p-value for Optimal Changepoint by Selective Inference using Dynamic Programming

Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, Ichiro Takeuchi

Nagoya Institute of Technology · RIKEN · Neural Information Processing Systems

Introduction and Motivation

- Changepoint (CP) detection: find changes in the underlying mechanism of the observed sequential data.
- CP detection is usually formulated as the problem of minimizing the segmentation cost where Dynamic Programming (DP) is commonly used.
- There are several CP detection methods. However, less attention has been paid to quantify the reliability of the detected CPs.

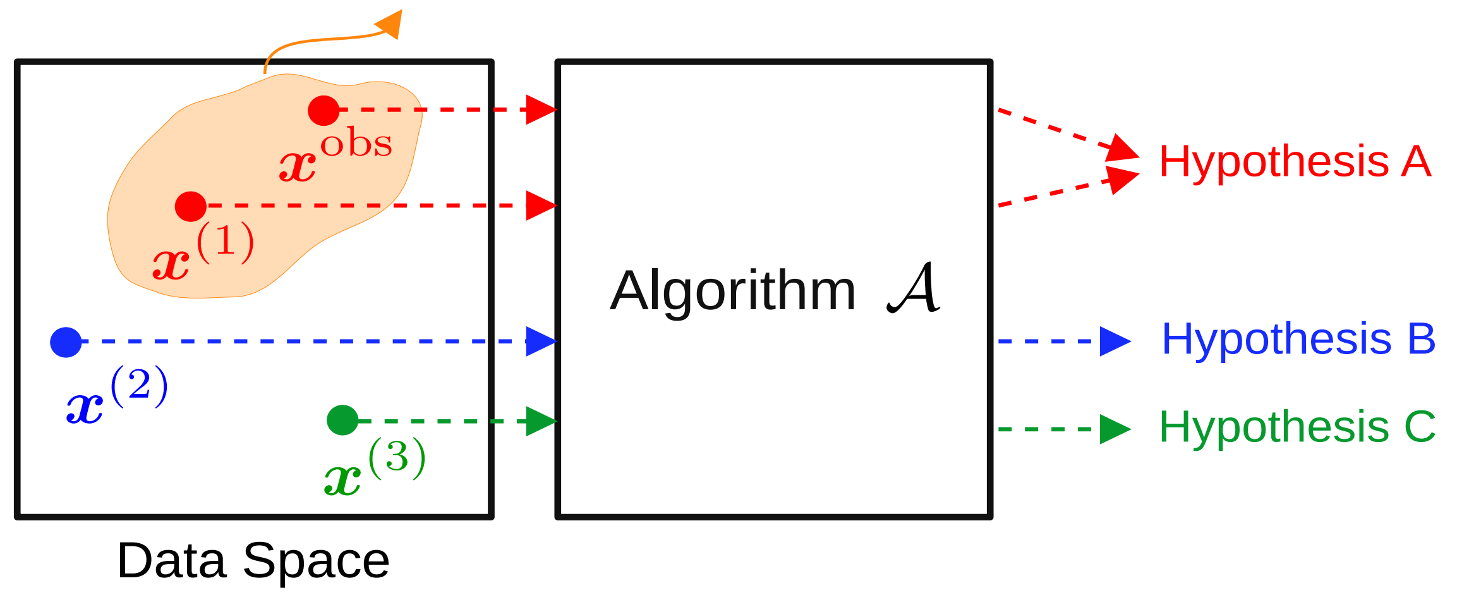


p-value table	A	B	C	D	E	F
OptSeg-SI	0.361	0.0	3×10^{-11}	1×10^{-5}	0.025	1×10^{-15}

- **A** and **E** are falsely detected CPs
  - ➡ Results from CP detection algorithms are unreliable
  - ➡ Harmful for high-stake decision making such as medical diagnosis
- We propose **OptSeg-SI** method to provide **valid** p-value, which is used as a criterion to quantify the reliability of the detected CPs, based on the concept of **Selective Inference (SI)**.
  - ➡ Large p-value indicates false detection (**A** and **E**) and small p-value indicates true detection (**B**, **C**, **D** and **F**)
  - ➡ OptSeg-SI can identify both false and true positive detections

Concept of Selective Inference (SI)

Conditional Data Space $\mathcal{X} = \{\boldsymbol{x} \mid \mathcal{A}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}^{\text{obs}})\}$

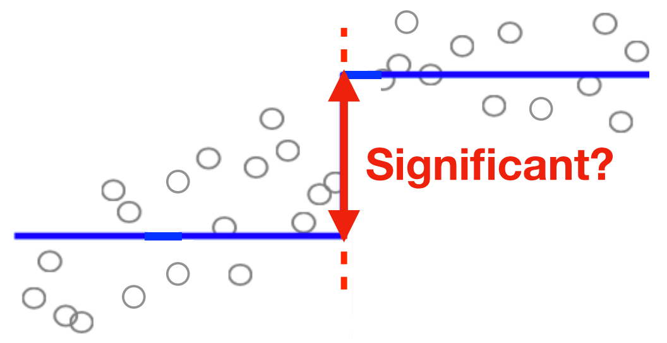


Conditional inference: $\Pr(T(\boldsymbol{x}) \mid \mathscr{A}(\boldsymbol{x}) = \mathscr{A}(\boldsymbol{x}^{\text{obs}}))$, where $T(\boldsymbol{x})$ is the test statistic.

Problem Setting

- We consider the following statistical test

$$\mathrm{H}_0 : \mu_{\text{left}} = \mu_{\text{right}} \quad \text{vs.} \quad \mathrm{H}_1 : \mu_{\text{left}} \neq \mu_{\text{right}}$$



where μ is population mean.

- The conditional p-value (selective p-value) is defined as

$$p_{\text{selective}} = \mathbb{P}_{\mathrm{H}_0}(\,|\,\Delta\,| \geq |\,\Delta^{\text{obs}}\,|\mid \mathscr{X})$$

- Δ^{obs} is is the difference in sample mean between the left segment and right segment in the **observed** sequence
- Δ is the mean difference in **any random sequence**
- $\mathscr{X}$ is the conditional data space defined as

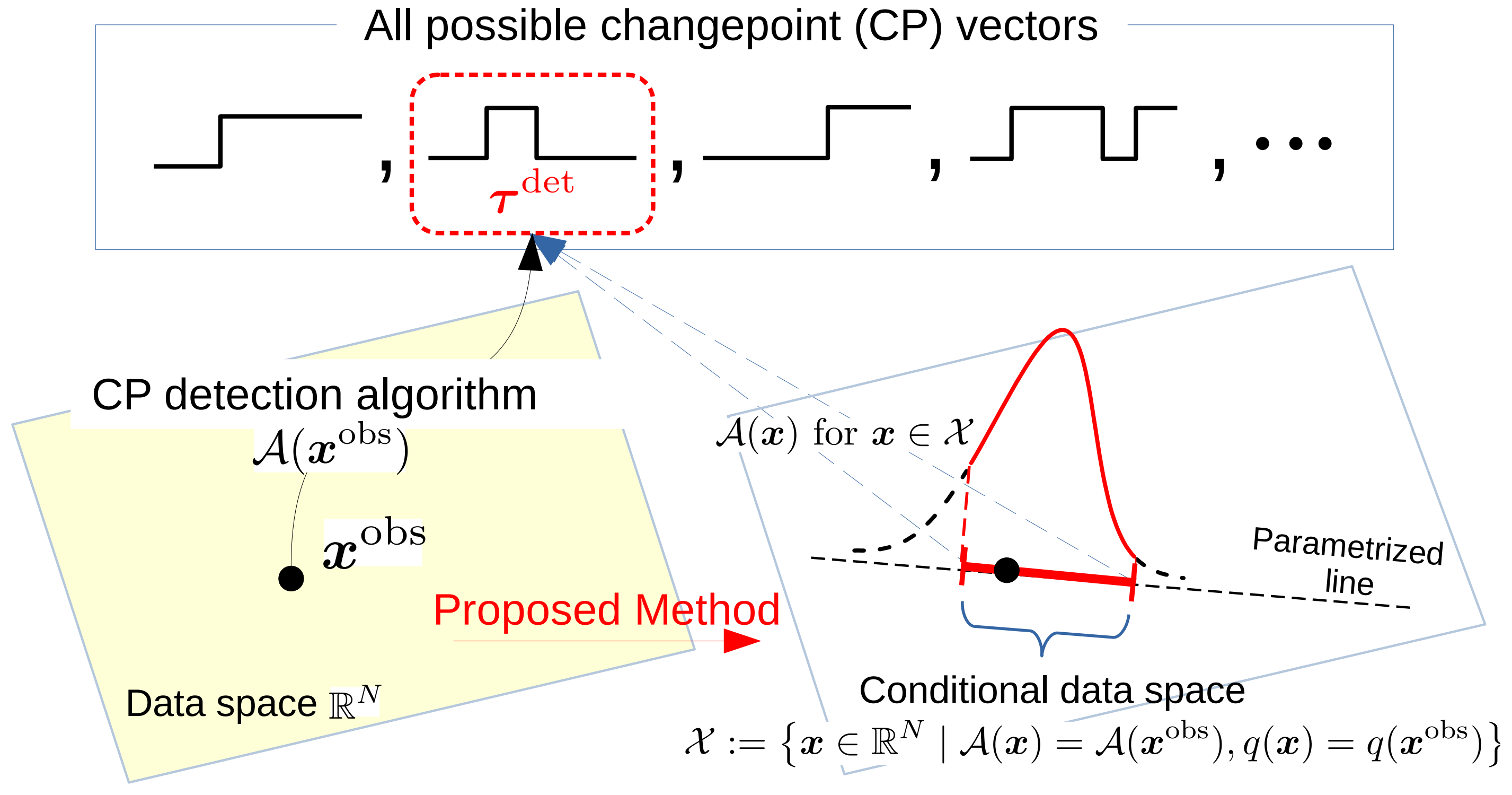
$$\mathscr{X} = \{\boldsymbol{x} : \{\text{left}, \text{right}\} \leftarrow \text{DP algorithm } \mathscr{A}(\boldsymbol{x})\}$$

- In other words, $\mathscr{X}$ is the data space whose data has the same detected CP as the observed sequence.
- The selective p-value is **valid** since

$$\mathbb{P}_{\mathrm{H}_0}(p_{\text{selective}} < \alpha) = \alpha, \quad \forall \alpha \in [0,1]\,.$$

However, characterization of the conditional data space $\mathscr{X}$ is challenging

Proposed Method - Schematic illustration


All possible changepoint (CP) vectors
$\boldsymbol{\tau}^{\text{det}}$
CP detection algorithm
$\mathcal{A}(\boldsymbol{x}^{\text{obs}})$
$\boldsymbol{x}^{\text{obs}}$
Proposed Method
Data space $\mathbb{R}^N$
$\mathcal{A}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{X}$
Parametrized line
Conditional data space
$\mathcal{X} := \{\boldsymbol{x} \in \mathbb{R}^N \mid \mathcal{A}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}^{\text{obs}}), q(\boldsymbol{x}) = q(\boldsymbol{x}^{\text{obs}})\}$

- Step 1: Obtain CP results from the observed data $\boldsymbol{x}^{\text{obs}}$
- Step 2: By restricting data on the line, we perform DP on parametrized data and identify the sub-space whose data has the same CP results as $\boldsymbol{x}^{\text{obs}}$

Proposed Method - Details

- We first restrict the data to the line by using a scalar parameter $z \in \mathbb{R}$

$$\boldsymbol{x}(z) = \boldsymbol{a} + \boldsymbol{b}z,$$

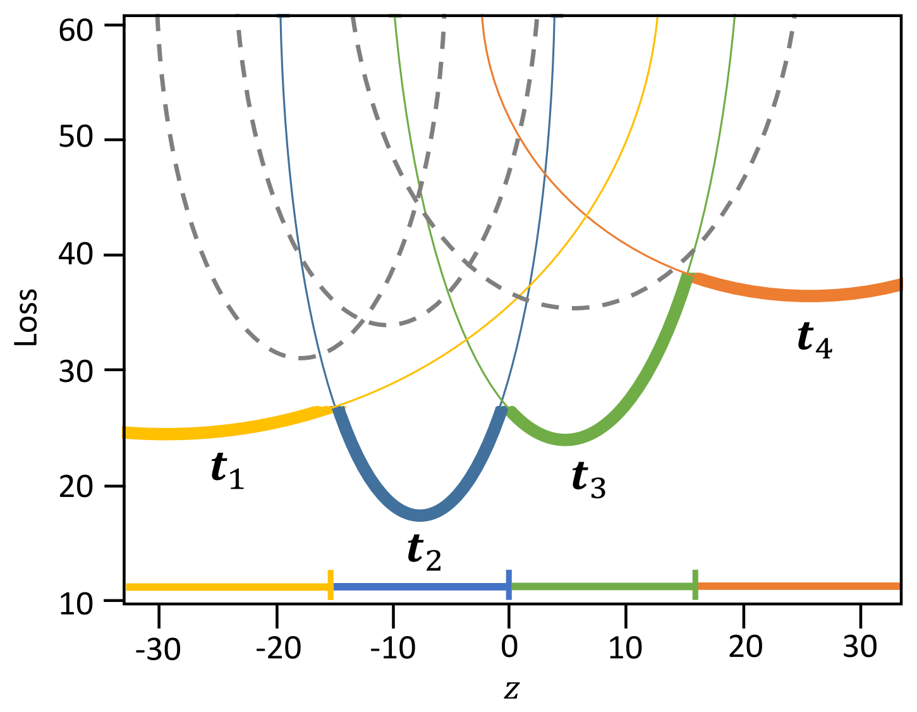
where $\boldsymbol{a}$ and $\boldsymbol{b}$ have specific forms.

- The conditional data space $\mathscr{X}$ is then re-written as

$$\mathscr{X} = \{\boldsymbol{x}(z) = \boldsymbol{a} + \boldsymbol{b}z \mid z \in \mathscr{Z}\},$$

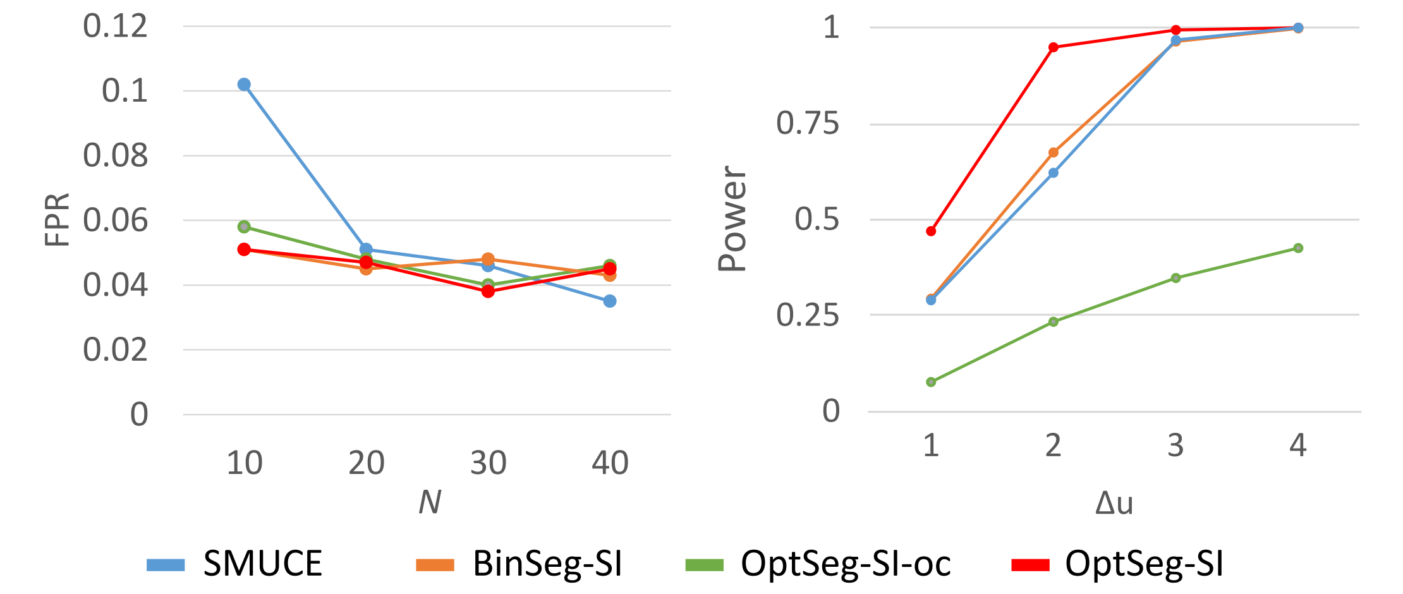
where $\mathscr{Z} = \{z \in \mathbb{R} : \{\text{left}, \text{right}\} \leftarrow \text{DP algorithm } \mathscr{A}(\boldsymbol{x}(z))\}$.

$\Longrightarrow$ **The remaining task is to identify truncation region $\mathscr{Z}$**

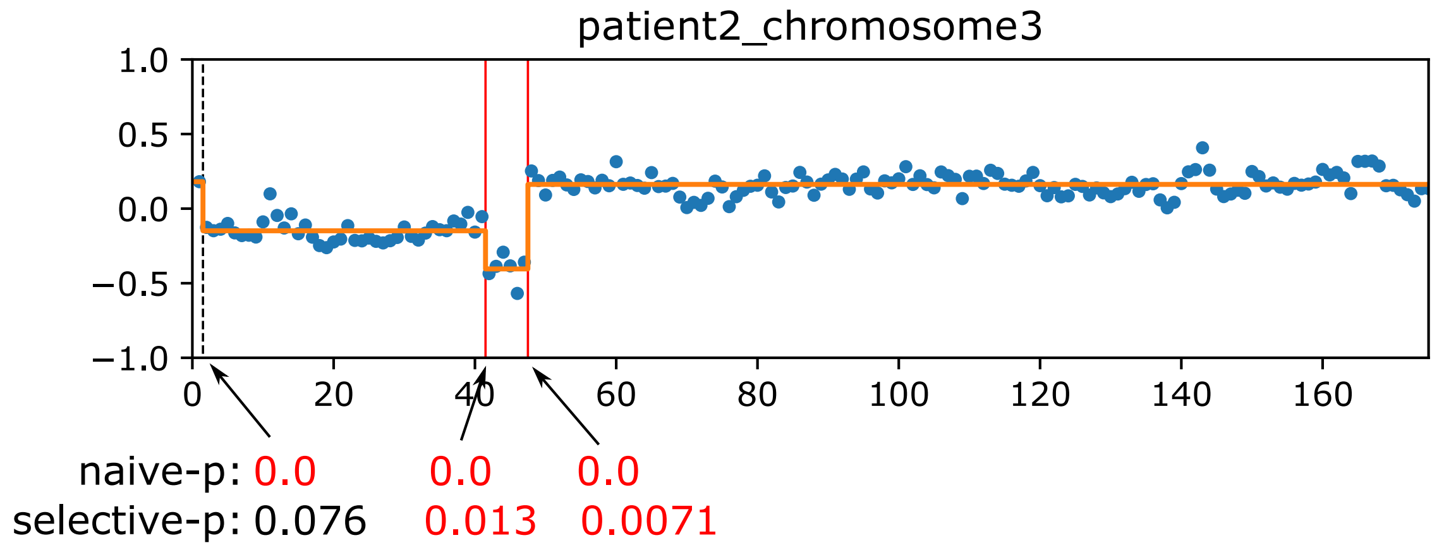


- We propose a **parametric DP approach** to compute CP results of $\boldsymbol{x}(z)$ for all $z \in \mathbb{R}$
- The region $\mathscr{Z}$ is then the union of intervals of z on which we obtain the same CP results as the observed data

Experimental Results



- The OptSeg-SI method (red) is powerful while successfully controlling the false positive rate (FPR)



- Application to real-world bioinformatics dataset

References

Lee et al. (2016). “Exact post-selection inference, with application to the lasso” In: The Annals of Statistics.

Duy et al. (2020). “Parametric programming approach for more powerful and general lasso selective inference”. In: arXiv:2004.09749.

Duy et al. (2020). “Quantifying statistical significance of neural network representation-driven hypotheses by selective inference”. In: arXiv:2010.01823.