Labeled Optimal Partitioning

Toby Dylan Hocking
toby.hocking@nau.edu
joint work with Anuraag Srivastava
arXiv:2006.13967

October 15, 2020

Introduction: supervised changepoint detection for cancer diagnosis with DNA copy number data

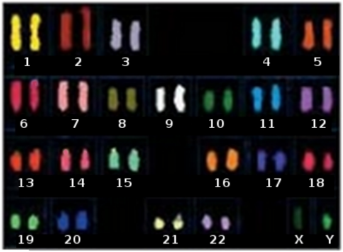
New Labeled Optimal Partitioning (LOPART) Algorithm

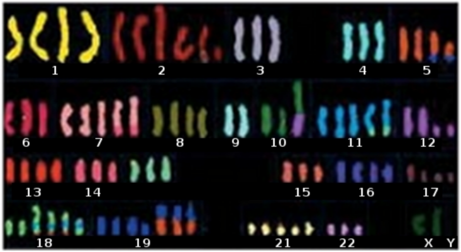
Demonstration of LOPART on example labeled data

Results and Discussion

Cancer cells show chromosomal copy number alterations

Spectral karyotypes show the number of copies of the sex chromosomes (X,Y) and autosomes (1-22).
Source: Alberts *et al.* 2002.

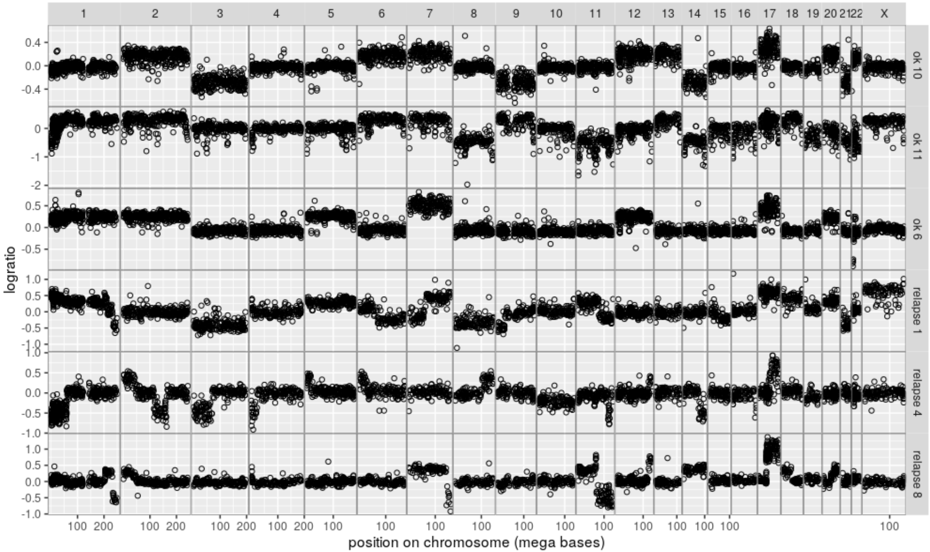




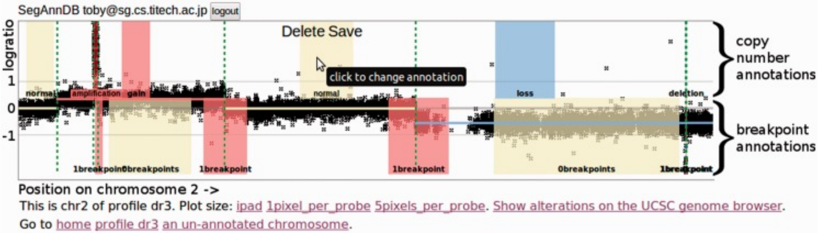
Normal cell with 2 copies of each autosome.

Cancer cell with many copy number alterations.

DNA copy number profiles from neuroblastoma patients with or without relapse



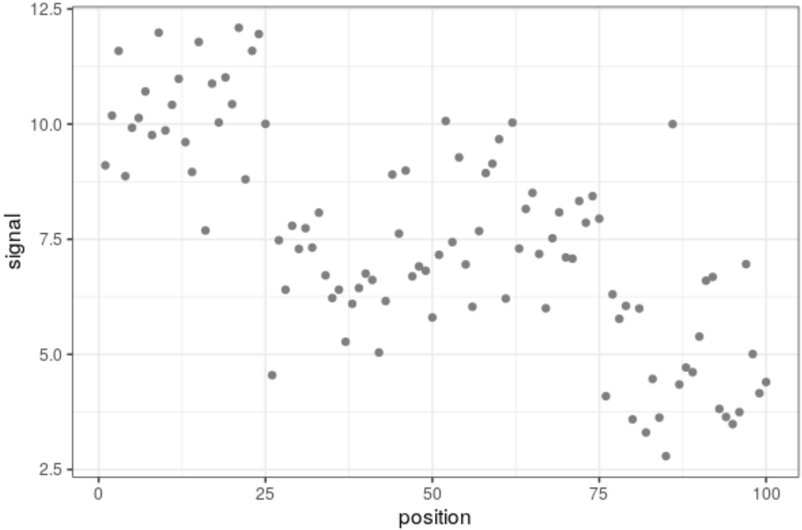
Previous work: SegAnnDB interactive machine learning system



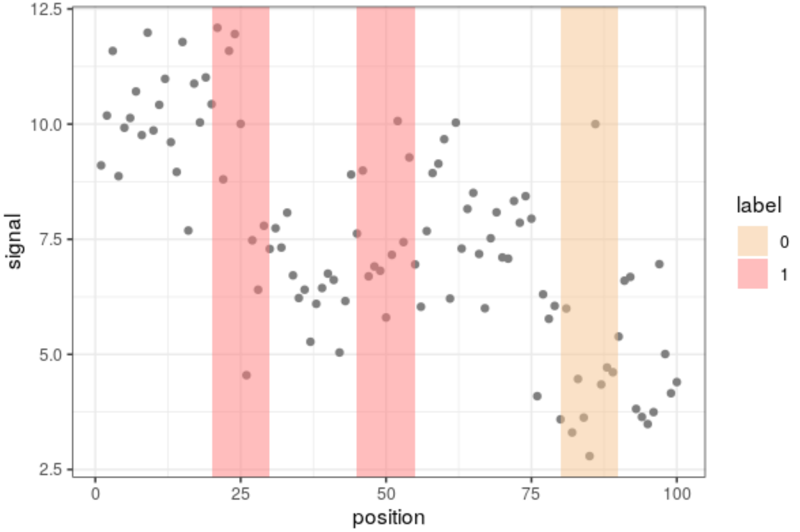
Hocking *et al.*, 2014.

- User uploads noisy data sets for machine learning analysis.
- User can provide labels which indicate presence(1) or absence(0) of changepoints in specific regions of data sets.
- Classic optimal changepoint model (max penalized Gaussian likelihood) used if it has zero label errors. OPART algorithm, Jackson *et al.*, 2005. FPOP algorithm, Maidstone *et al.*, 2016.
- Label-aware SegAnnot algorithm used otherwise (Hocking and Rigaill, 2012). Always zero train label errors, but never predicts any changepoints outside of positive(1) labels.

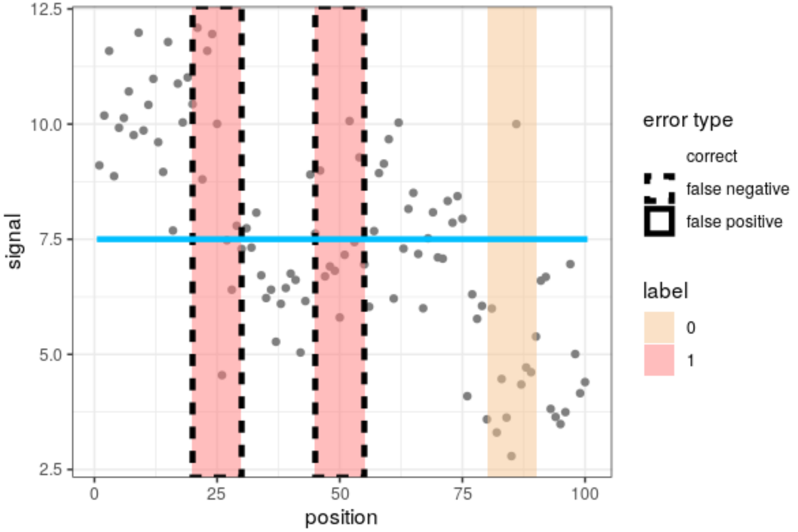
Example noisy data sequence



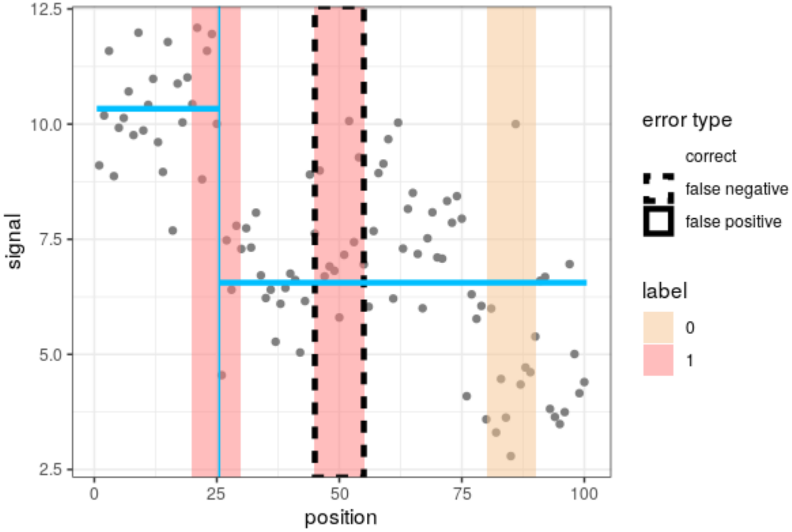
Example noisy data sequence with labels



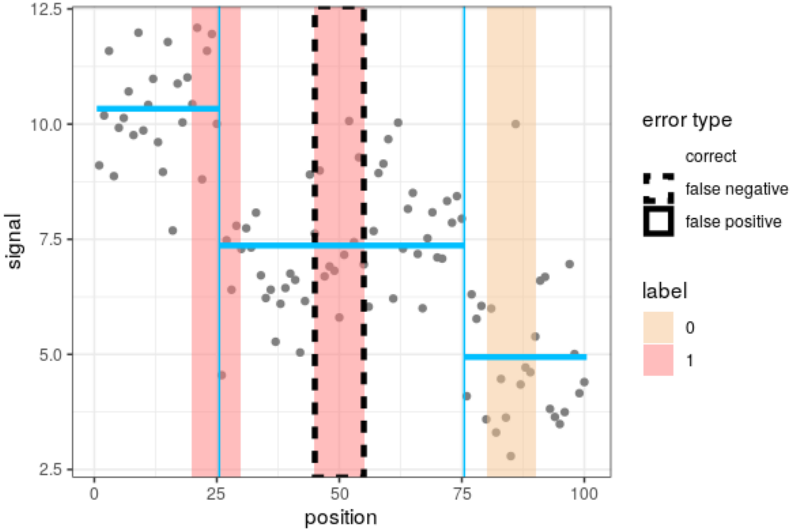
OPART with penalty $\lambda = 1000$ (ignores labels)



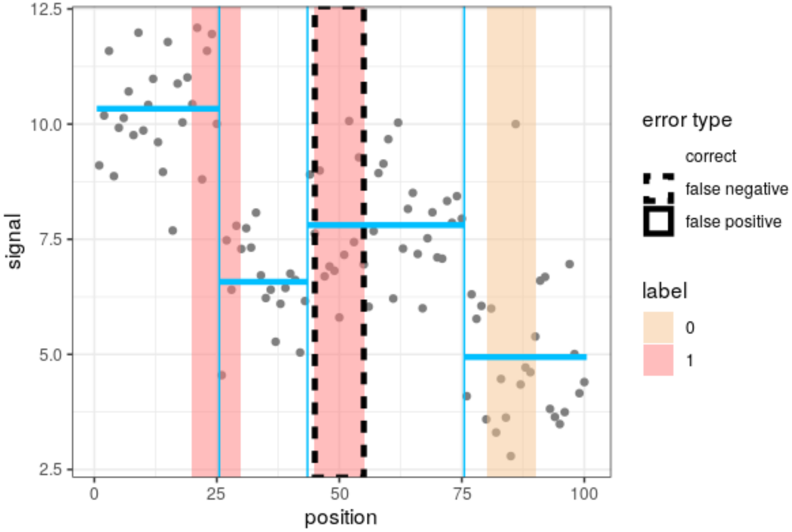
OPART with penalty $\lambda = 100$ (ignores labels)



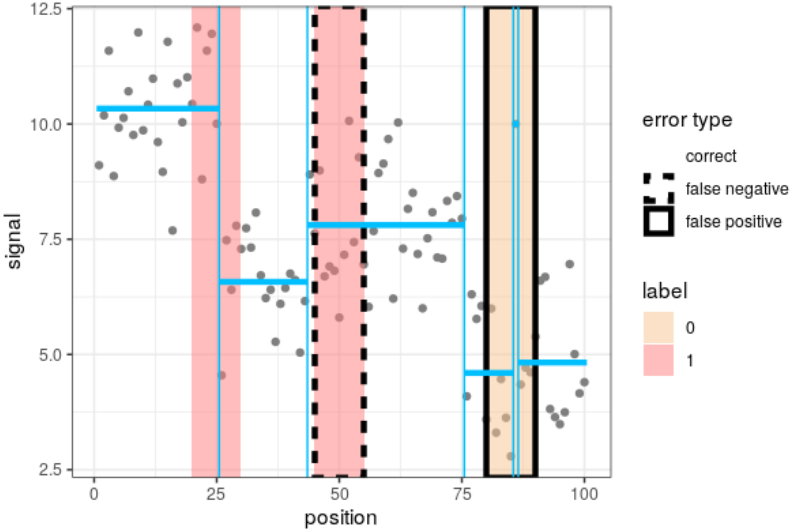
OPART with penalty $\lambda = 20$ (ignores labels)



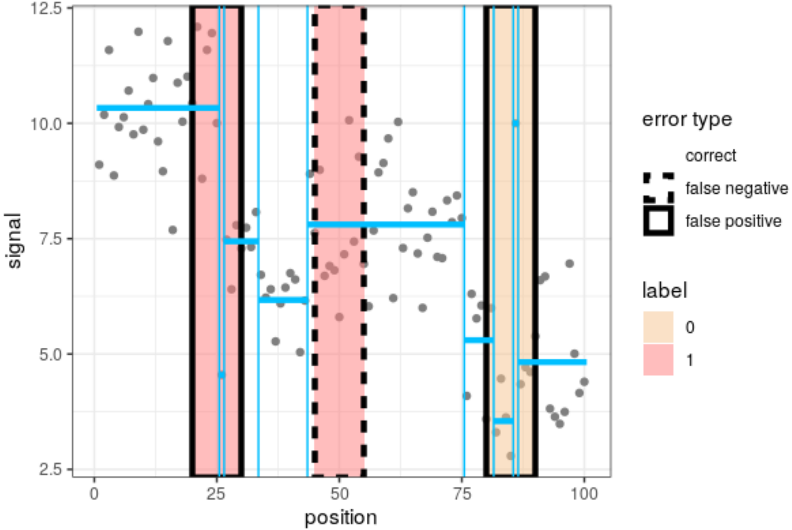
OPART with penalty $\lambda = 15$ (ignores labels)



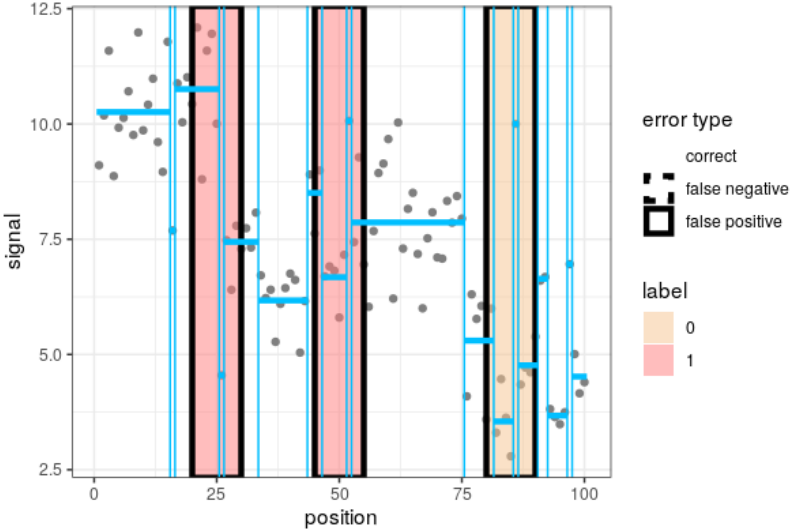
OPART with penalty $\lambda = 10$ (ignores labels)



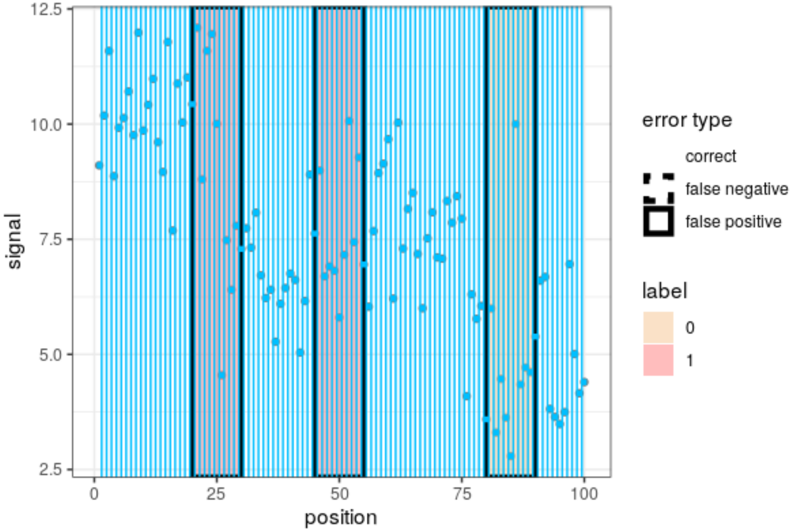
OPART with penalty $\lambda = 5$ (ignores labels)



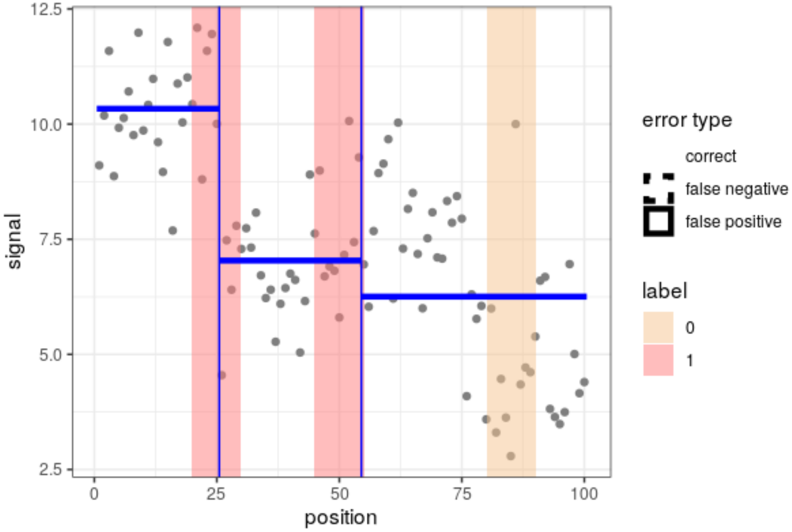
OPART with penalty $\lambda = 4$ (ignores labels)



OPART with penalty $\lambda = 0$ (ignores labels)



SegAnnot (no changes in unlabeled regions)



Introduction: supervised changepoint detection for cancer diagnosis with DNA copy number data

New Labeled Optimal Partitioning (LOPART) Algorithm

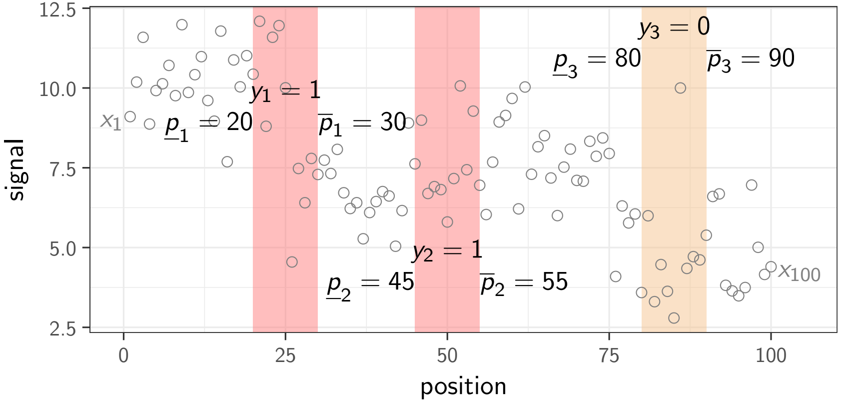
Demonstration of LOPART on example labeled data

Results and Discussion

Geometric interpretation of data and labels

Assume N data and M labels defined by

- $\mathbf{x} = [x_1 \cdots x_N]$ is the sequence of N data,
- $1 \leq \underline{p}_1 < \overline{p}_1 \leq \cdots \leq \underline{p}_M < \overline{p}_M \leq N$ are region start/ends,
- $y_j \in \{0, 1\}$ is the number of changes expected in the region.



Baseline/previous OPART algorithm

Assume

- $\mathbf{x} = [x_1 \cdots x_N]$ is the sequence of N data,
- ℓ is a loss function (e.g. square loss),
- I the indicator function counts the number of changes,
- λ is a non-negative penalty (larger for fewer changes).

Then the problem and algorithm are

$$\begin{aligned} C_N &= \min_{\mathbf{m}\in\mathbb{R}^N} \sum_{i=1}^{N} \ell(m_i, x_i) + \lambda \sum_{i=1}^{N-1} I[m_i \neq m_{i+1}]. \\ &= \min_{\tau\in\{0,1,\dots,N-1\}} C_\tau + \lambda + L(\tau+1, N, \mathbf{x}). \end{aligned}$$

where

- τ is the last changepoint optimization variable,
- C_τ is the optimal cost computed in previous iteration τ,
- L is the cost of the last segment.

Proposed LOPART optimization problem

Assume M labels are defined by

- $1 \leq \underline{p}_1 < \overline{p}_1 \leq \cdots \leq \underline{p}_M < \overline{p}_M \leq N$ are region start/ends,
- $y_j \in \{0, 1\}$ is the number of changes expected in the region.

The problem we want to solve is

$$\min_{\mathbf{m} \in \mathbb{R}^N} \quad \sum_{i=1}^{N} \ell(m_i, x_i) + \lambda \sum_{i=1}^{N-1} I[m_i \neq m_{i+1}].$$

$$\text{subject to} \quad \text{for all } j \in \{1, \dots, M\},\ y_j = \sum_{i=\underline{p}_j}^{\overline{p}_j - 1} I[m_i \neq m_{i+1}].$$

Objective function in black same as previous/unlabeled problem,
constraint in red new to proposed/labeled problem.

Proposed LOPART dynamic programming algorithm

Define the set of possible last changepoints at time t as

$$T_t = \begin{cases} T_{t-1} & \text{if } \exists j : y_j = 0 \text{ and } t \in \{\underline{p}_j + 1, \dots, \overline{p}_j\} \\ T_{t-1} & \text{if } \exists j : y_j = 1 \text{ and } t \in \{\underline{p}_j + 1, \dots, \overline{p}_j - 1\} \\ \{\underline{p}_j, \dots, t-1\} & \text{if } \exists j : y_j = 1 \text{ and } t = \overline{p}_j \\ T_{t-1} \cup \{t-1\} & \text{otherwise. (unlabeled region)} \end{cases}$$

Update rule in black same as previous/unlabeled problem,
rules in red for labeled regions new to proposed problem.
Then the optimal cost can be computed via

$$\tau_t^*, W_t = \arg\min, \min_{\tau \in T_t} W_\tau + \lambda + L(\tau + 1, t, \mathbf{x}).$$

(see paper for proof)

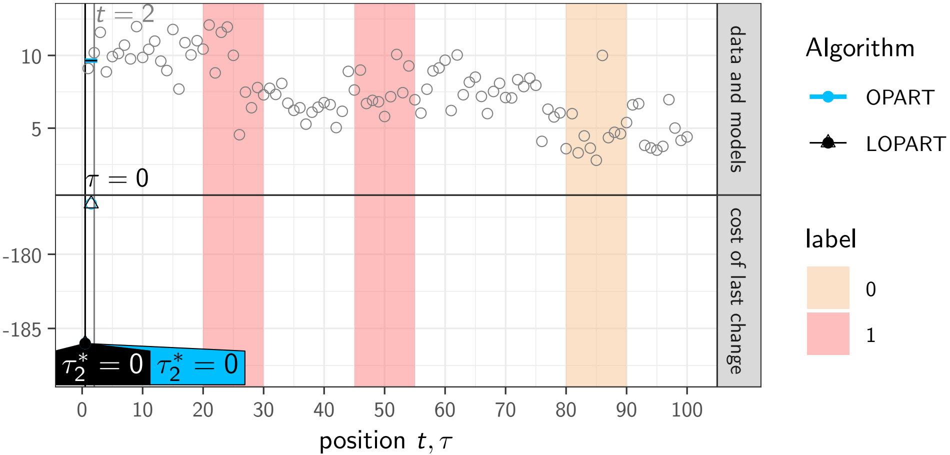
Introduction: supervised changepoint detection for cancer diagnosis with DNA copy number data

New Labeled Optimal Partitioning (LOPART) Algorithm

Demonstration of LOPART on example labeled data

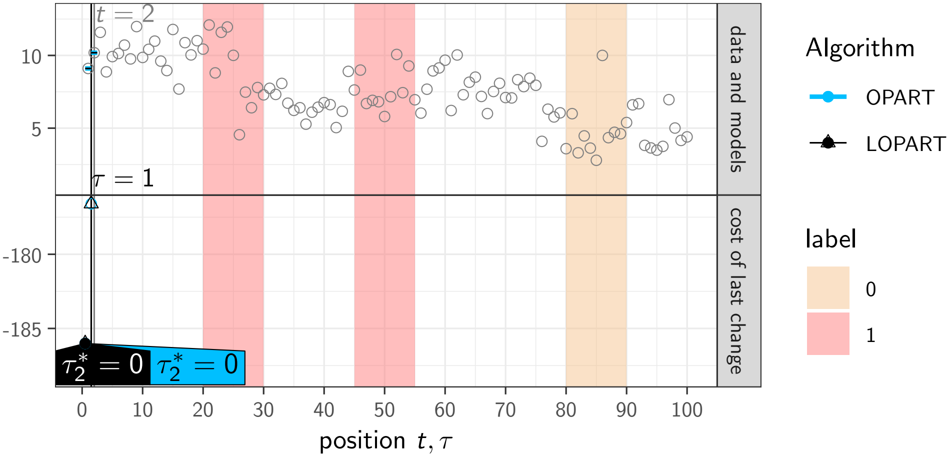
Results and Discussion

Demonstration of LOPART with penalty $\lambda = 10$



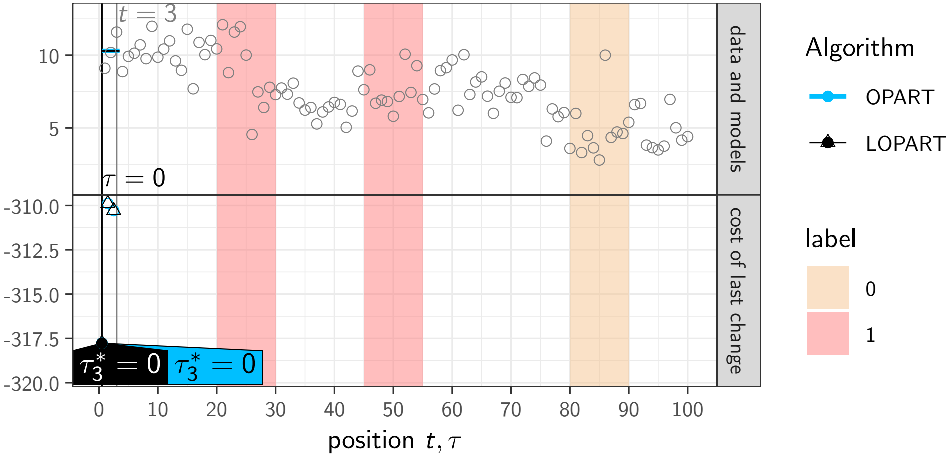
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



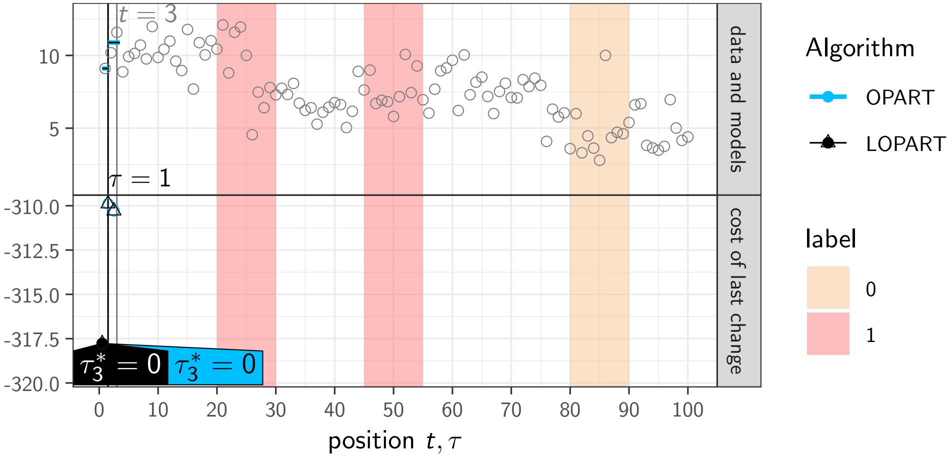
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



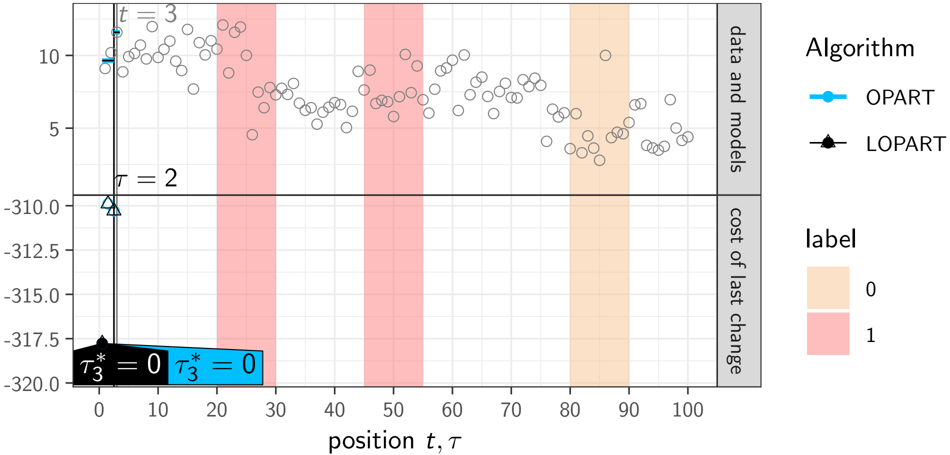
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



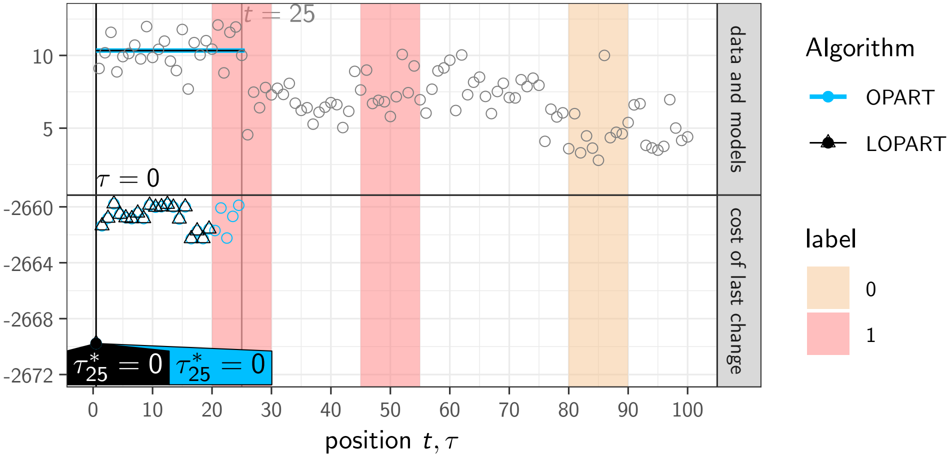
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



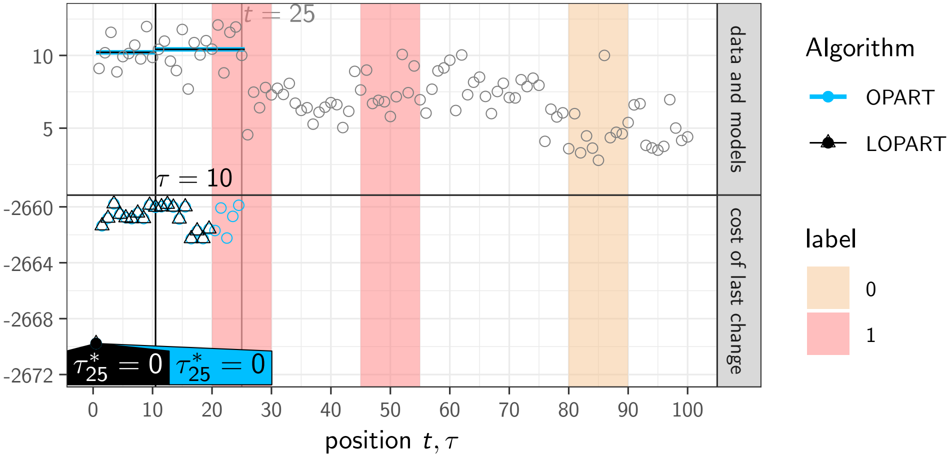
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



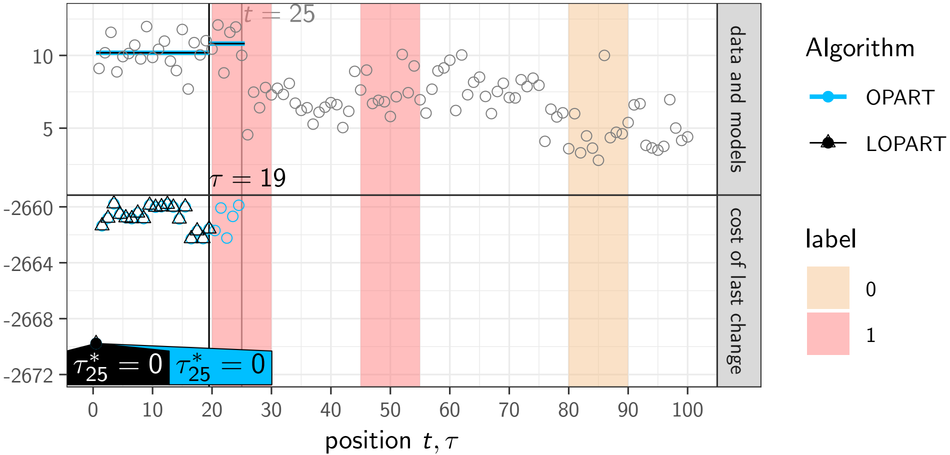
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



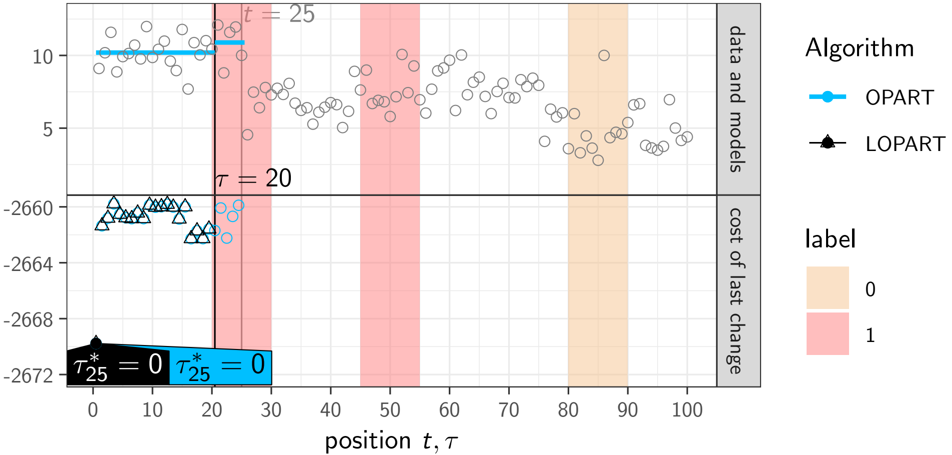
$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



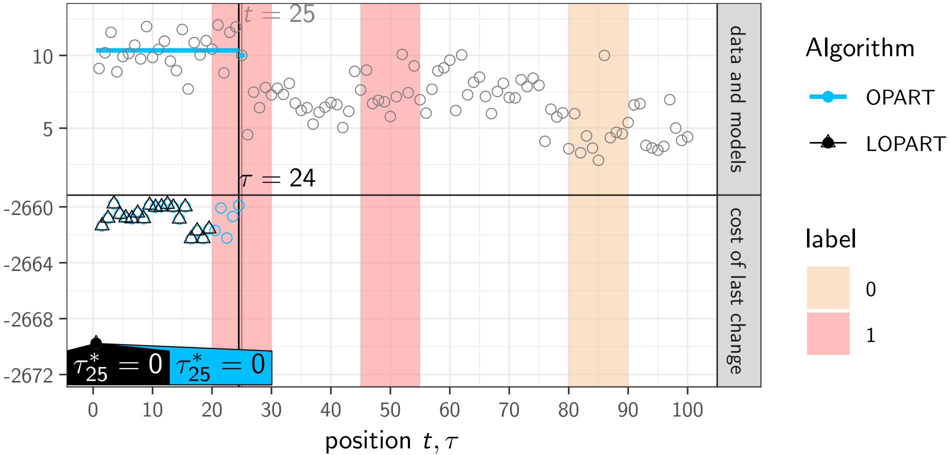
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



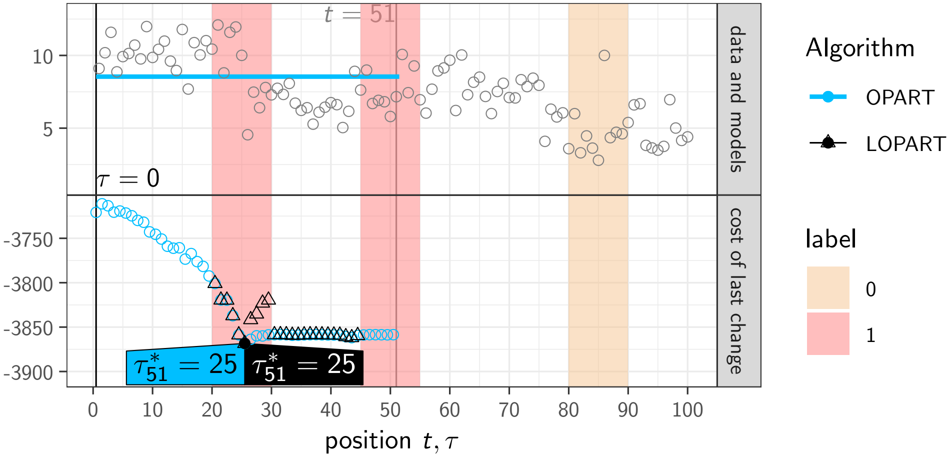
$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau + 1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



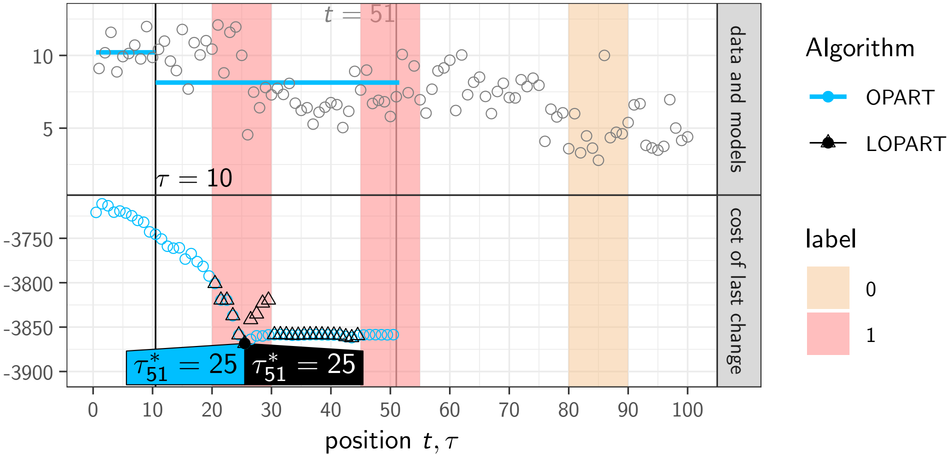
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



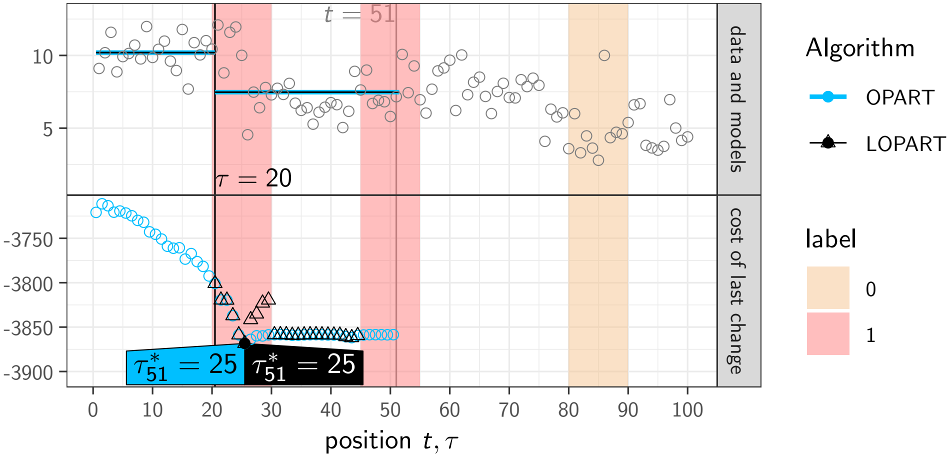
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



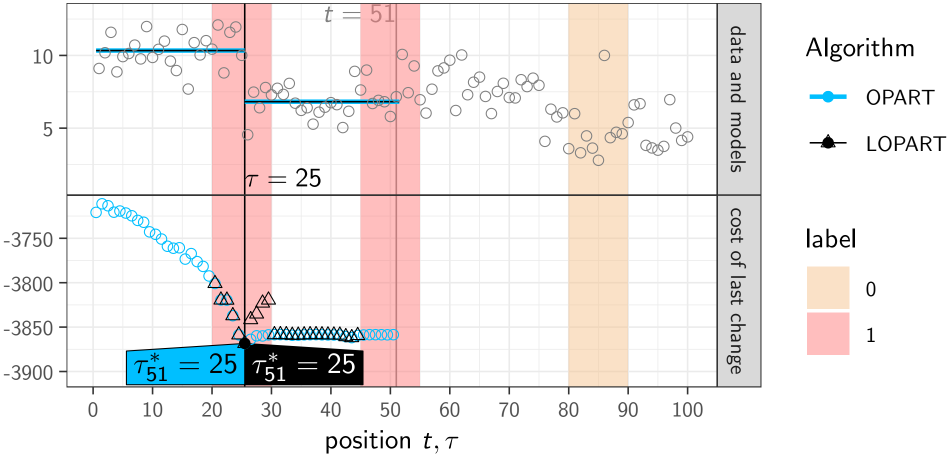
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



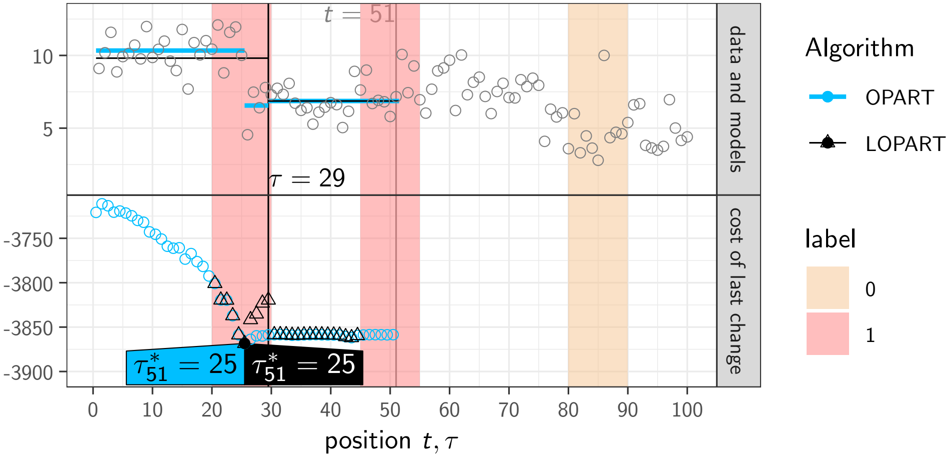
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



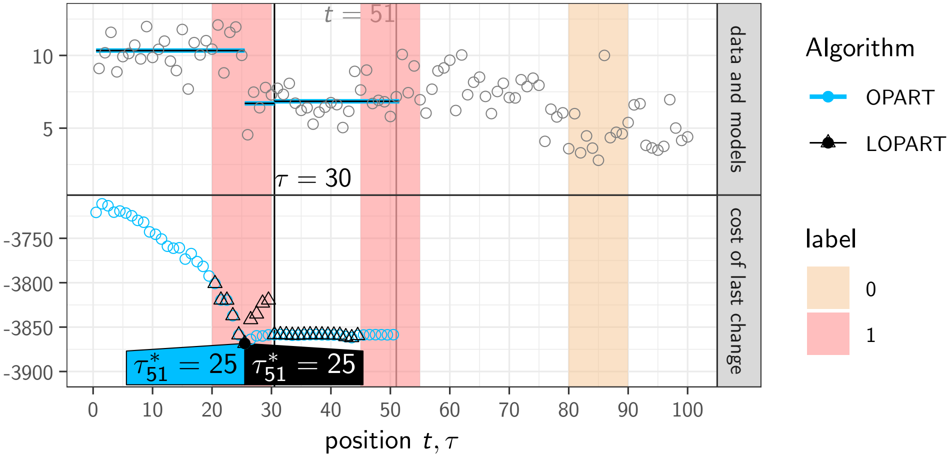
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



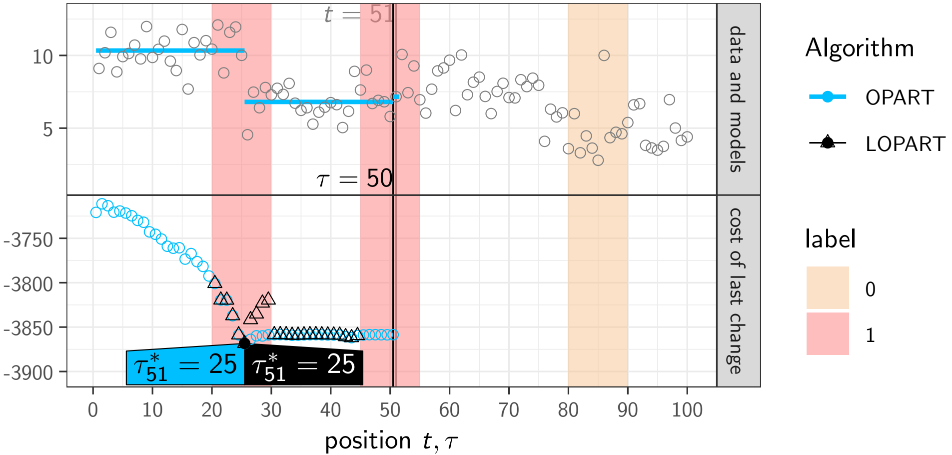
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



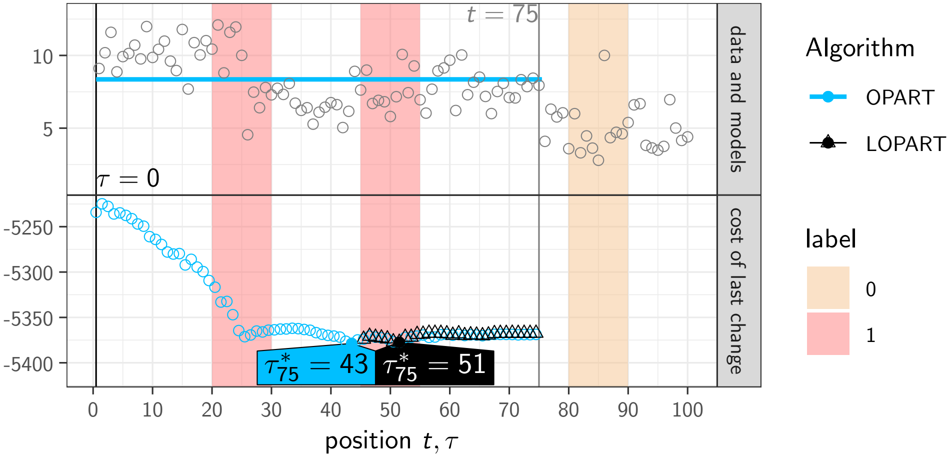
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



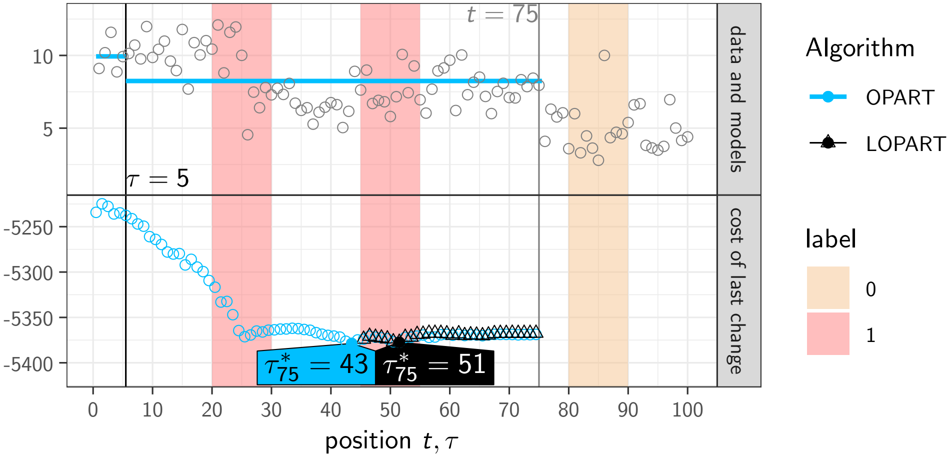
$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



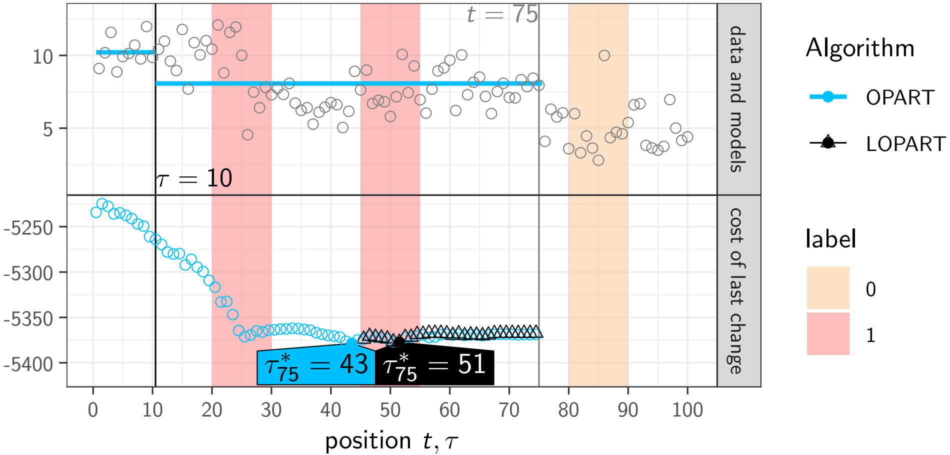
$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



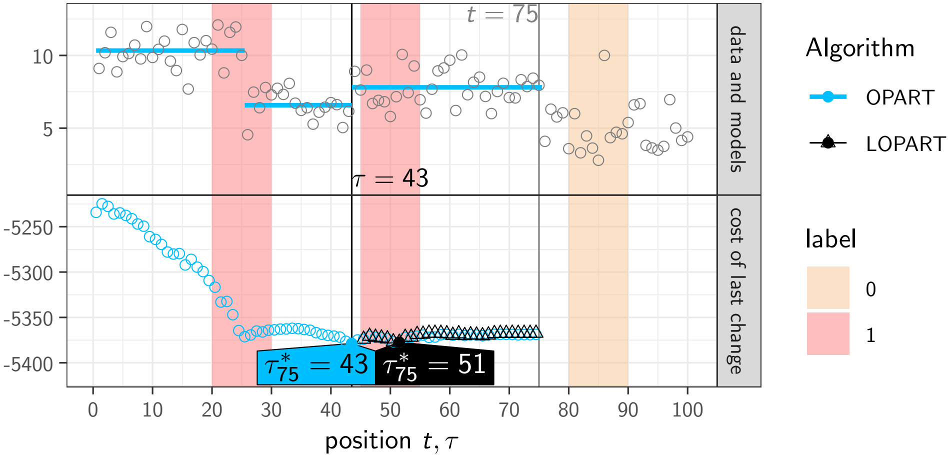
$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



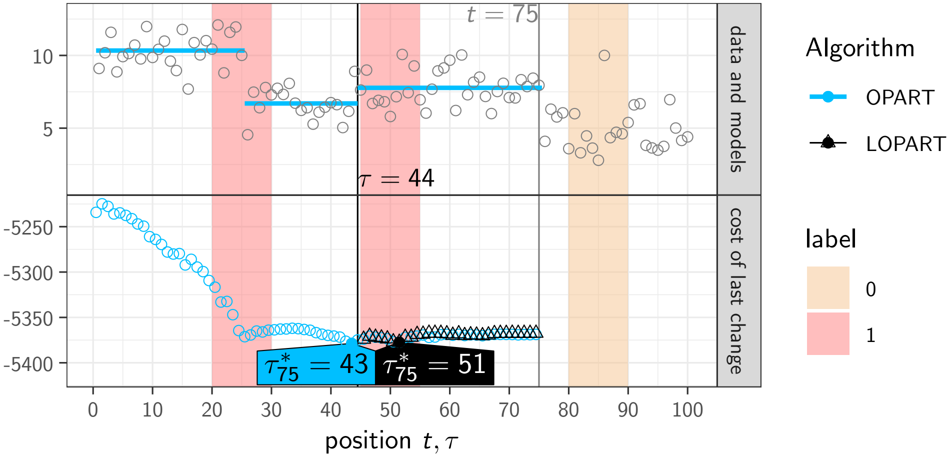
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



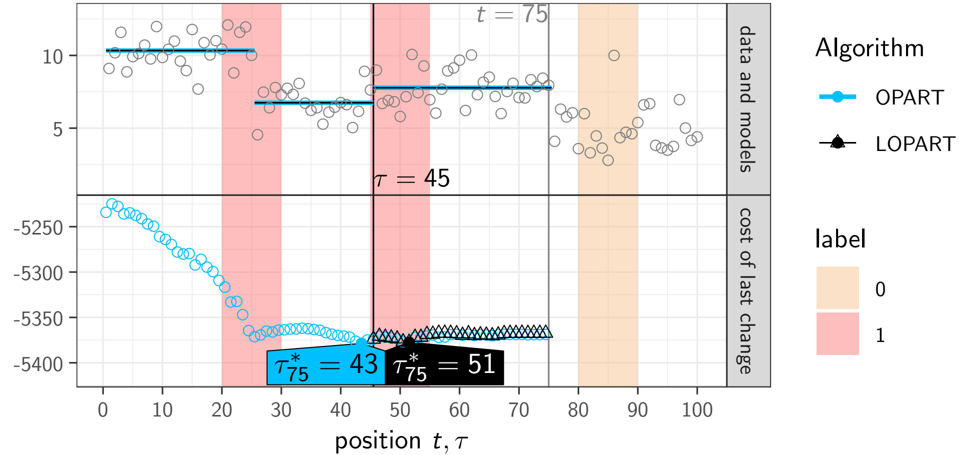
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



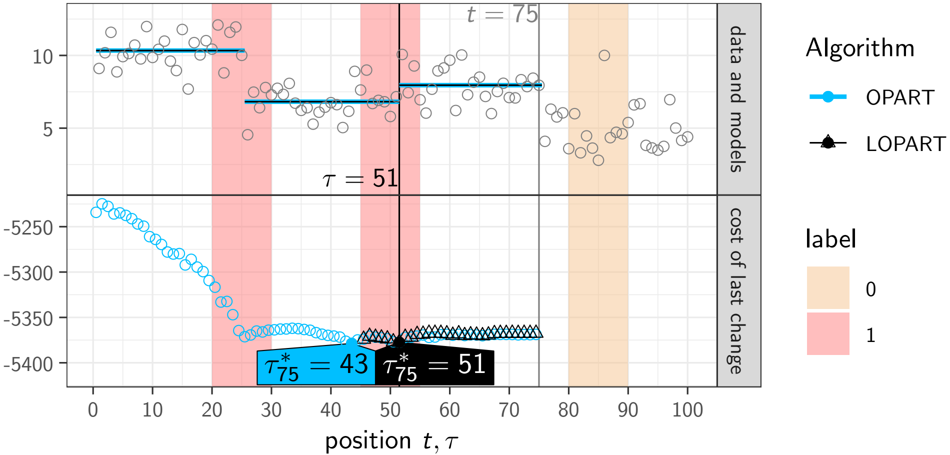
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



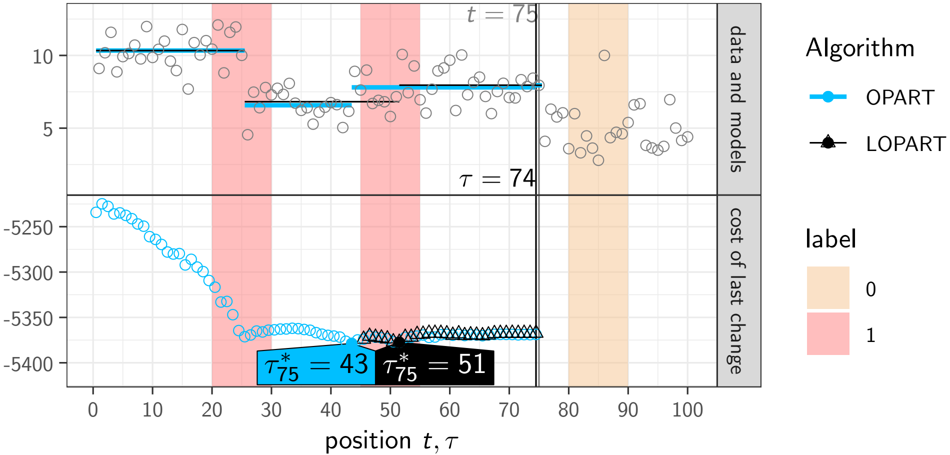
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



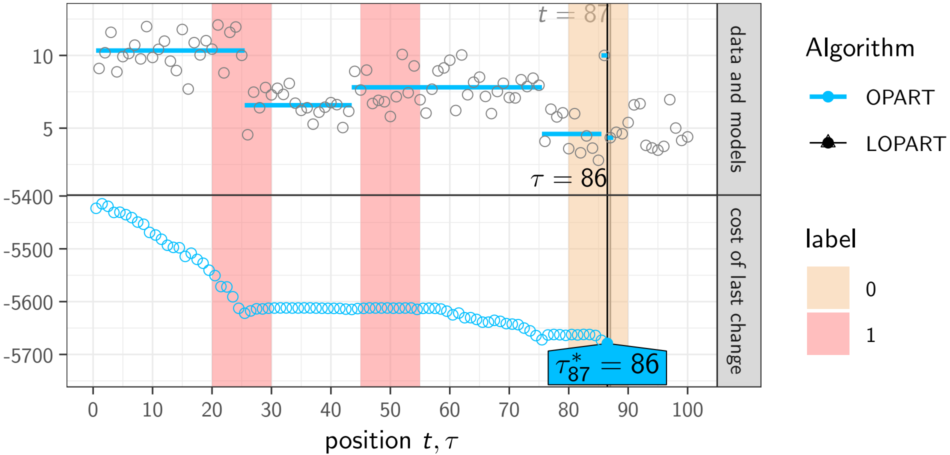
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



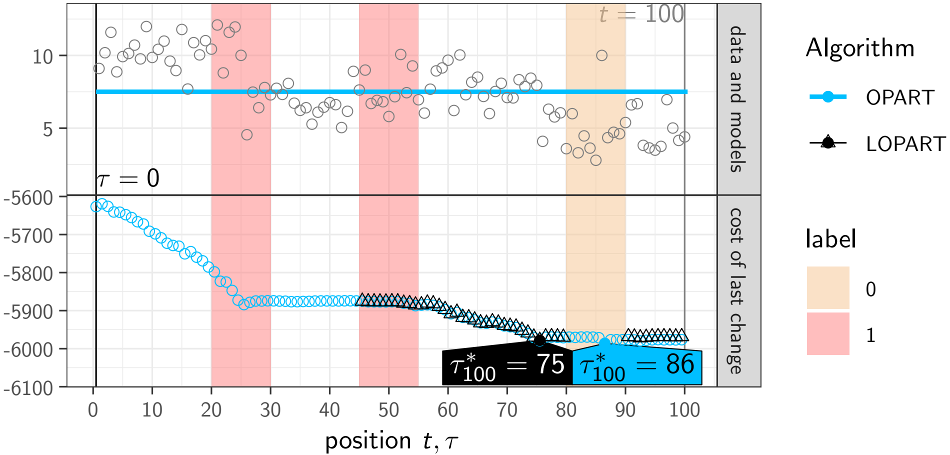
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



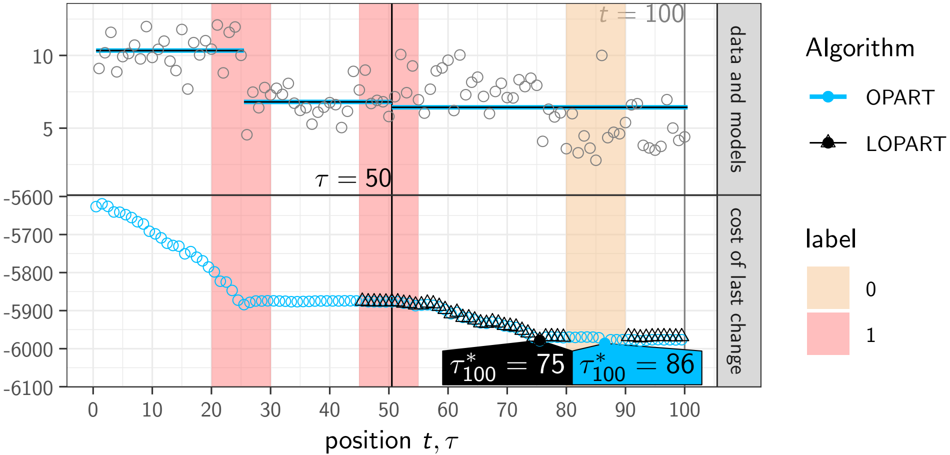
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



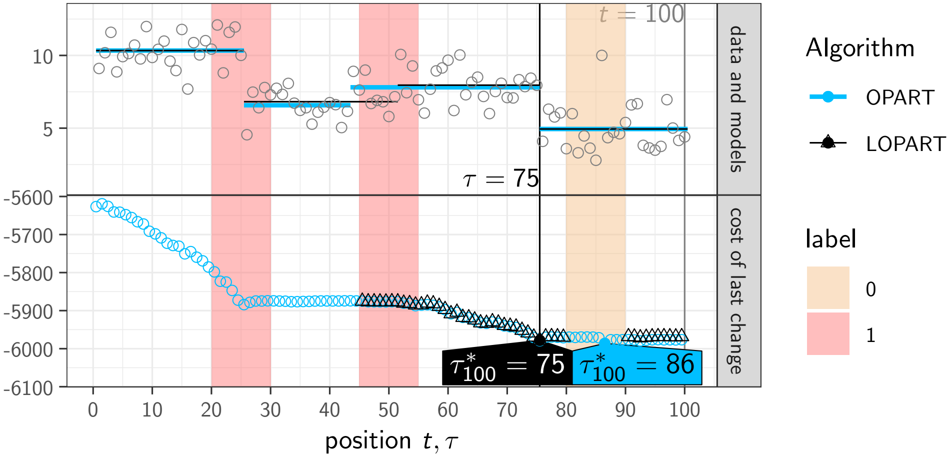
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



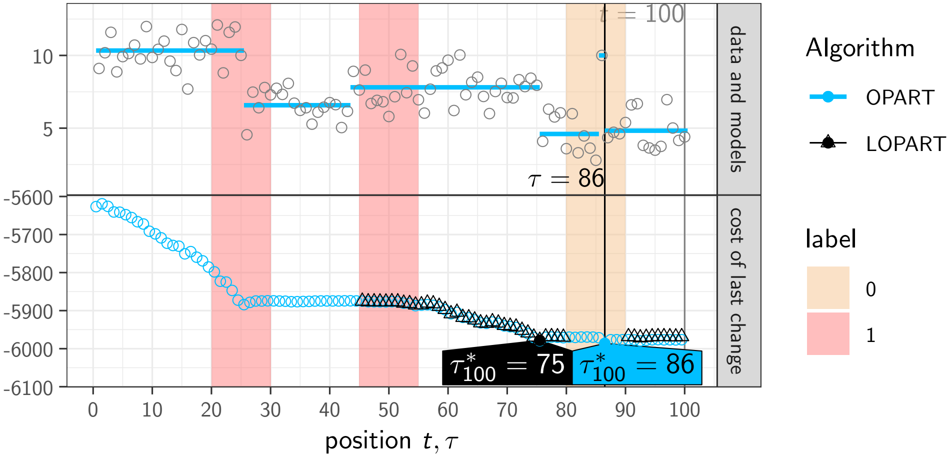
$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



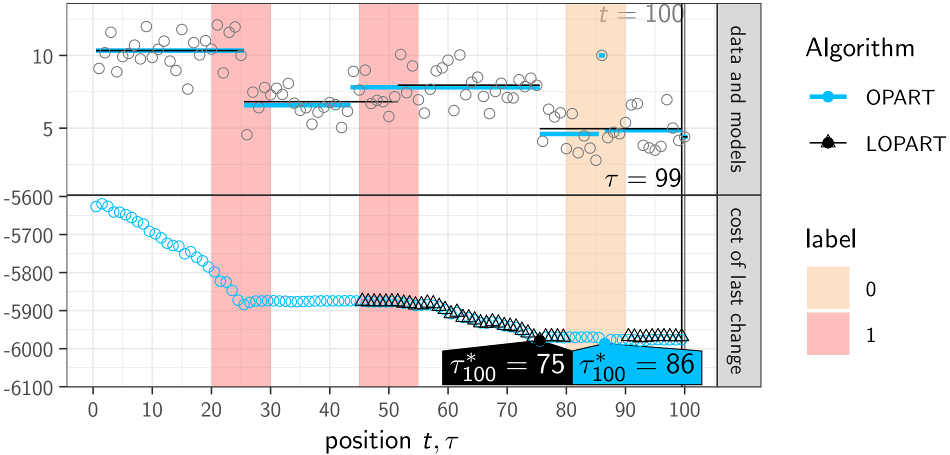
$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau + 1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Demonstration of LOPART with penalty $\lambda = 10$



$$\tau^*_t,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

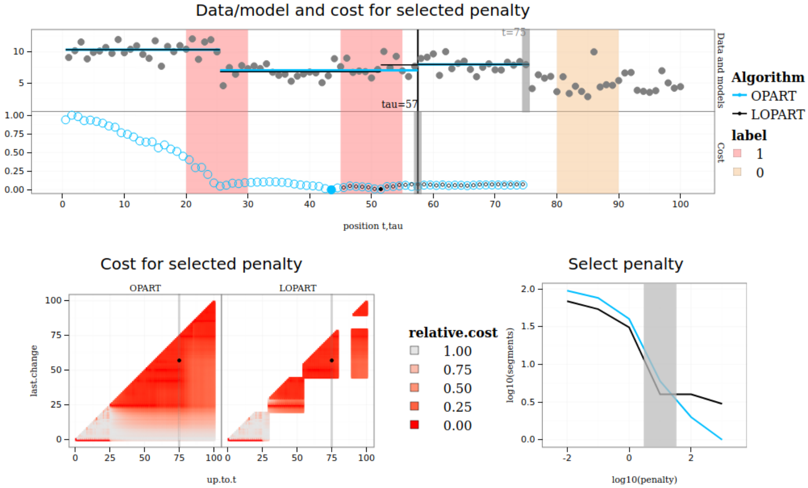
Demonstration of LOPART with penalty $\lambda = 10$



$$\tau_t^*,\ W_t = \arg\min,\ \min_{\tau \in T_t} \underbrace{W_\tau}_{\text{optimal cost up to } \tau} + \underbrace{\lambda}_{\text{penalty}} + \underbrace{L(\tau+1, t, \mathbf{x})}_{\text{optimal cost of last segment}}$$

Interactive version, new algorithm requires less computation

Try this at home:



http://members.cbio.mines-paristech.fr/~thocking/figure-candidates-interactive/

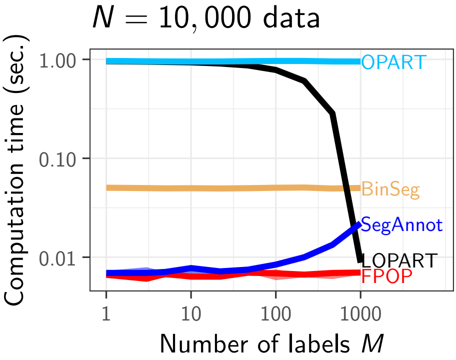
Introduction: supervised changepoint detection for cancer diagnosis with DNA copy number data

New Labeled Optimal Partitioning (LOPART) Algorithm

Demonstration of LOPART on example labeled data

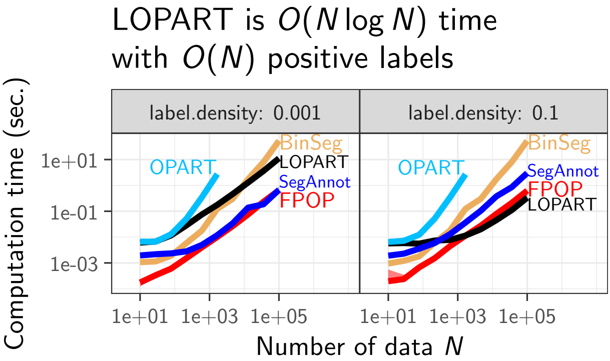
Results and Discussion

Empirical time complexity (labels)

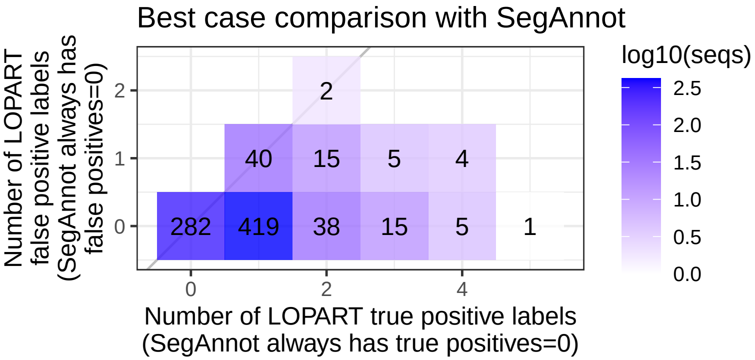


Random normal data simulations.

Empirical time complexity (data)



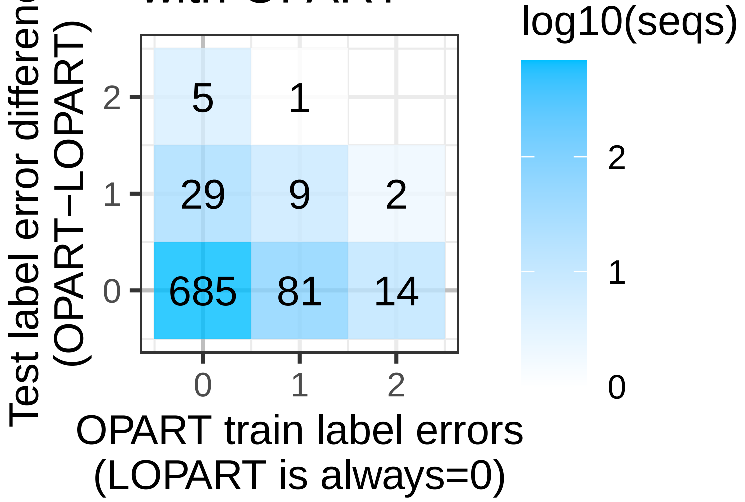
Random normal data simulations.



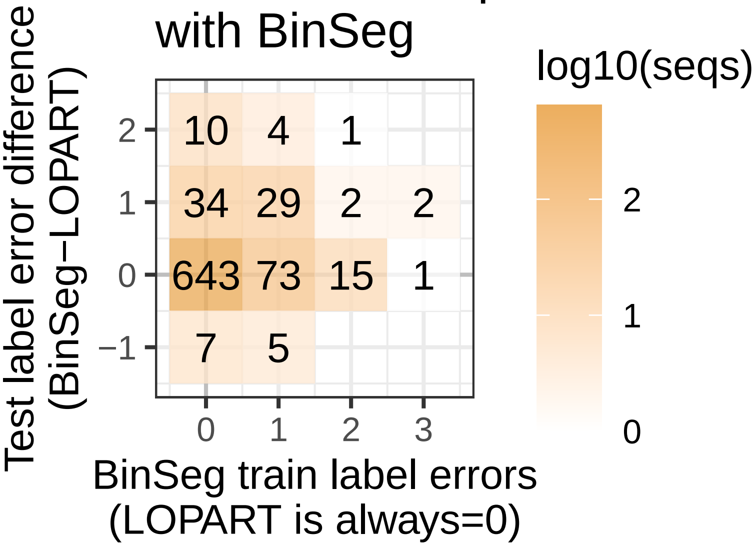
Real genomic data labeled by biologists (Hocking *et al.*, 2014).

- 413 sequences, $K = 2$ fold cross-validation over labels.
- Data per sequence $N = 39$ to 43628.
- Labels per sequence $M = 2$ to 12.

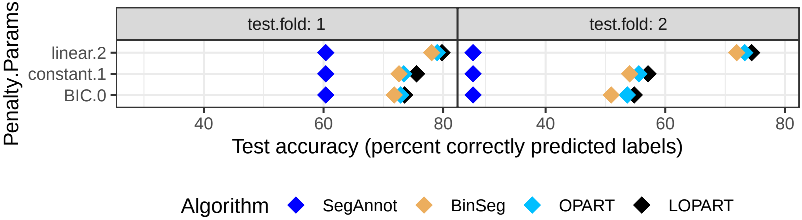
Best case comparison with OPART



Best case comparison with BinSeg



Test accuracy in cross-validation experiments



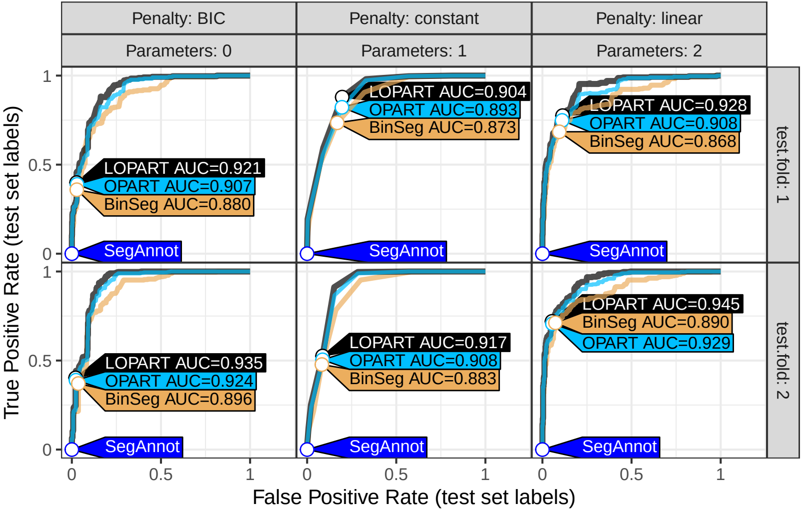
Penalty λ predicted on test set using either unsupervised or supervised (constant/linear) methods:

linear.2 linear $\log \lambda_i = b + w \log \log N_i$ (Hocking 2013), convex optimization of w, b (supervised; 2 learned parameters).

constant.1 grid search, constant λ (supervised, 1 learned parameter).

BIC.0 classical Bayesian Information Criterion (Schwarz 1978), $\lambda_i = \log N_i$ (unsupervised, 0 learned parameters).

Test ROC curves in cross-validation experiments



Dot = FPR/TPR at predicted threshold.

Summary and Discussion

- Proposed algo fixes issues with two previous algorithms (better train AND test accuracy).
- Results demonstrate improved speed and accuracy.
- R package on CRAN and `https://github.com/tdhock/LOPART`
- Figure/slide code on `https://github.com/tdhock/LOPART-paper`
- Future work: functional pruning algorithm, which can solve more complex constrained changepoint problems (e.g. change must be non-decreasing), and should be faster (log-linear instead of quadratic).