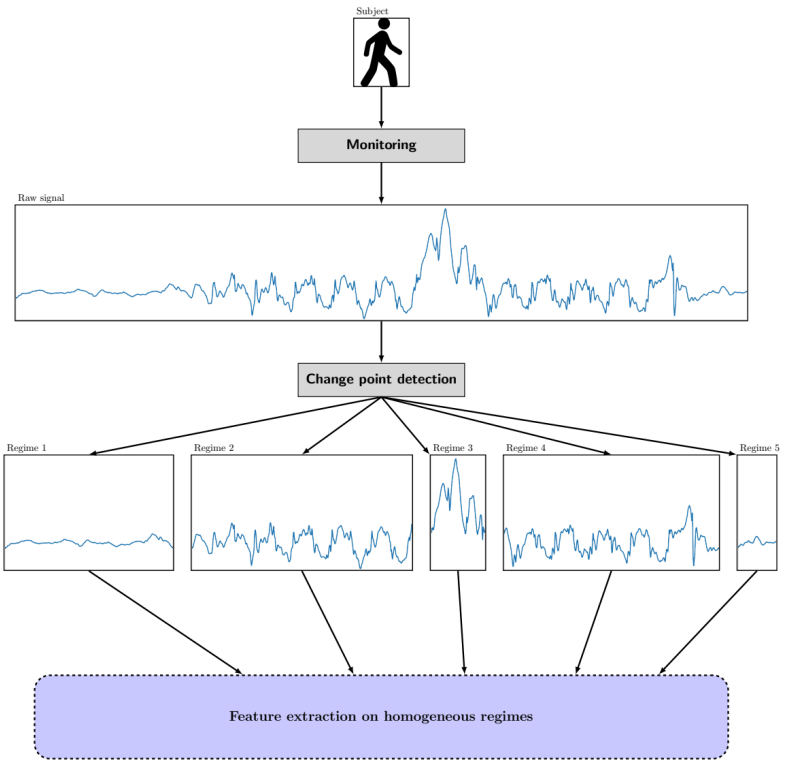
[Signal Processing'20] Selective review of offline change point detection methods

1. Link:
    1. Paper: https://arxiv.org/pdf/1801.00718
    2. Repo: https://centre-borelli.github.io/ruptures-docs/
2. Arthurs and institution: Charles Truonga, Laurent Oudreb, Nicolas Vayatis from CMLA, CNRS, ENS Paris Saclay and L2TI, University Paris 13

TL;DR A selective survey of algorithms for the offline detection of multiple change points in multivariate time series, while it not cover the Bayesian methods.



Thoughts and critisims

Problem formulation

1. data

Let us consider a multivariate non-stationary random process $y = \{y_1, \dots, y_T\}$ that takes value in $\mathbb{R}^d$ $(d \geq 1)$ and has T samples. The signal y is assumed to be piecewise stationary, meaning that some characteristics of the process change abruptly at some unknown instants $t_1^* < t_2^* < \cdots < t_{K^*}^*$. Change point detection consists in estimating the indexes t_k^*. Depending on the context, the number K^* of changes may or may not be known, in which case it has to be estimated too.

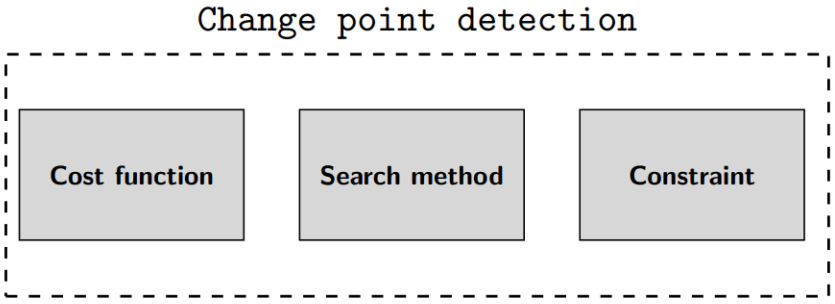
2. Objective

segmentation:

$$V(\mathcal{T}, y) := \sum_{k=0}^{K} c(y_{t_k..t_{k+1}}) \tag{1}$$

where $c(\cdot)$ is a cost function which measures goodness-of-fit of the sub-signal $y_{t_k..t_{k+1}} = \{y_t\}_{t_k+1}^{t_{k+1}}$ to a specific model. The "best segmentation" $\widehat{\mathcal{T}}$ is the minimizer of the criterion $V(\mathcal{T})$. In practice, depending on whether the number

3. Modules



1. Cost function. The cost function c(·) is a measure of "homogeneity"
2. Search method. The search method is the resolution procedure for the discrete optimization problems associated with Problem 1 (P1) and Problem 2 (P2).
3. Constraint (on the number of change points). When the number of changes is unknown (P2), a constraint is added, in the form of a complexity penalty pen(·) (P2), to balance out the goodness-of-fit term V (T , y). T

Key concepts

Cost function

1. property of estimation
    1. Asymptotic consistency

satisfies the following conditions, when $T \longrightarrow +\infty$*:*

(i) $P(|\widehat{\mathcal{T}}| = K^*) \longrightarrow 1$,

(ii) $\frac{1}{T}\left\|\widehat{\mathcal{T}} - \mathcal{T}^*\right\|_\infty \xrightarrow{p} 0$,

where the distance between two change point sets is defined by

$$\left\|\widehat{\mathcal{T}} - \mathcal{T}^*\right\|_\infty := \max\left\{ \max_{\hat{t}\in\widehat{\mathcal{T}}} \min_{t^*\in\mathcal{T}^*} |\hat{t} - t^*|,\ \max_{t^*\in\mathcal{T}^*} \min_{\hat{t}\in\widehat{\mathcal{T}}} |\hat{t} - t^*| \right\}. \tag{4}$$

1.

2. evaluation

1. AnnotationError

The AnnotationError is simply the difference between the predicted number of change points $|\widehat{\mathcal{T}}|$ and the true number of change points $|\mathcal{T}^\star|$:

$$\textsc{AnnotationError} := |\widehat{K} - K^*|. \tag{5}$$

2. Hausdorf

$$\textsc{Hausdorff}(\mathcal{T}^*, \widehat{\mathcal{T}}) := \max\left\{ \max_{\hat{t}\in\widehat{\mathcal{T}}} \min_{t^*\in\mathcal{T}^*} |\hat{t} - t^*|,\ \max_{t^*\in\mathcal{T}^*} \min_{\hat{t}\in\widehat{\mathcal{T}}} |\hat{t} - t^*| \right\}.$$

3. RandIndex

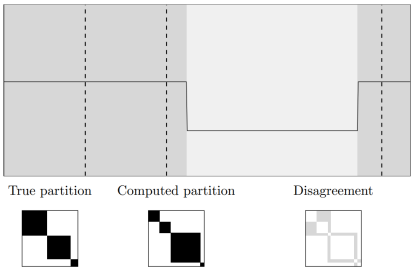


Figure 4: RandIndex. Top: alternating gray areas mark the segmentation $\mathcal{T}^*$; dashed lines mark the segmentation $\widehat{\mathcal{T}}$. Below: representations of associated adjacency matrices and disagreement matrix. The adjacency matrix of a segmentation is the $T \times T$ binary matrix with coefficient (s, t) equal to 1 if s and t belong to the same segment, 0 otherwise. The disagreement matrix is the $T \times T$ binary matrix with coefficient (s, t) equal to 1 where the two adjacency matrices disagree, and 0 otherwise. RandIndex is equal to the white area (where coefficients are 0) of the disagreement matrix.

The RandIndex is then defined as follows:

$$\textsc{RandIndex}(\mathcal{T}^*, \widehat{\mathcal{T}}) := \frac{|\mathrm{gr}(\widehat{\mathcal{T}}) \cap \mathrm{gr}(\mathcal{T}^*)| + |\mathrm{ngr}(\widehat{\mathcal{T}}) \cap \mathrm{ngr}(\mathcal{T}^*)|)}{T(T-1)}. \tag{6}$$

4. F1-score

$$\textsc{Tp}(\mathcal{T}^*, \widehat{\mathcal{T}}) := \{t^* \in \mathcal{T}^* \mid \exists \hat{t} \in \widehat{\mathcal{T}} \ \text{ s.t. } |\hat{t} - t^*| < M\}.$$

3. Cost functions

1. parametric

1. i.i.d.

Cost function 1 ($c_{\text{i.i.d.}}$)**.** *For a given parametric family of distribution densities* $\{f(\cdot|\theta)|\theta \in \Theta\}$ *where* Θ *is a compact subset of* $\mathbb{R}^p$ *(for a certain* p*), the cost function* $c_{i.i.d.}$ *is defined by*

$$c_{i.i.d.}(y_{a..b}) := -\sup_\theta \sum_{t=a+1}^{b} \log f(y_t|\theta). \tag{C1}$$

2. cost 2:

Cost function 2 (c_{L_2}). *The cost function c_{L_2} is given by*

$$c_{L_2}(y_{a..b}) := \sum_{t=a+1}^{b} \|y_t - \bar{y}_{a..b}\|_2^2$$

3. cost 3:

Cost function 3 (c_Σ). *The cost function c_Σ is given by*

$$c_\Sigma(y_{a..b}) := (b-a)\log\det\widehat{\Sigma}_{a..b} + \sum_{t=a+1}^{b} (y_t - \bar{y}_{a..b})'\widehat{\Sigma}_{a..b}^{-1}(y_t - \bar{y}_{a..b}) \quad \text{(C3)}$$

where $\bar{y}_{a..b}$ and $\widehat{\Sigma}_{a..b}$ are respectively the empirical mean and the empirical covariance matrix of the sub-signal $y_{a..b}$.

4. cost 4:

Cost function 4 ($c_{\mathbf{Poisson}}$). *The cost function $c_{Poisson}$ is given by*

$$c_{Poisson}(y_{a..b}) := -(b-a)\bar{y}_{a..b}\log\bar{y}_{a..b} \quad \text{(C4)}$$

where $\bar{y}_{a..b}$ is the empirical mean of the sub-signal $y_{a..b}$.

5. cost 5:

Cost function 5 ($c_{\mathbf{linear}}$). *For a signal y (response variable) and covariates x and z, the cost function c_{linear} is defined by*

$$c_{linear}(y_{a..b}) := \min_{u\in\mathbb{R}^p, v\in\mathbb{R}^q} \sum_{t=a+1}^{b} (y_t - x_t'u - z_t'v)^2. \quad \text{(C5)}$$

6. cost AR:

Cost function 7 (c_{AR}). *For a signal y and an order $p \geq 1$, the cost function c_{AR} is defined by*

$$c_{AR}(y_{a..b}) := \min_{u\in\mathbb{R}^p} \sum_{t=a+1}^{b} \|y_t - x_t'u\|^2 \quad \text{(C7)}$$

where $x_t := [y_{t-1}, y_{t-2}, \ldots, y_{t-p}]$ is the vector of lagged samples.

2. non-parametric

1. assume a emperical CDF

Signal model.. Assume that the observed signal $y = \{y_1, \ldots, y_T\}$ is composed of independent random variables, such that

$$y_t \sim \sum_{k=0}^{K^*} F_k \, \mathbb{1}(t_k^* < t \leq t_{k+1}^*) \quad \text{(M3)}$$

1.

2. MLE:

$$\forall u \in \mathbb{R}, \quad \widehat{F}_{a..b}(u) := \frac{1}{b-a}\left[\sum_{t=a+1}^{b} \mathbb{1}(y_t < u) + 0.5 \times \mathbb{1}(y_t = u)\right]. \quad (13)$$

1. cost function:

Cost function 9 ($c_{\widehat{F}}$). *The cost function $c_{\widehat{F}}$ is given by*

$$c_{\widehat{F}}(y_{a..b}) := -(b-a)\sum_{u=1}^{T}\frac{\widehat{F}_{a..b}(u)\log\widehat{F}_{a..b}(u)+(1-\widehat{F}_{a..b}(u))\log(1-\widehat{F}_{a..b}(u))}{(u-0.5)(T-u+0.5)} \tag{C9}$$

where the empirical cdf $\widehat{F}_{a..b}$ is defined by (13).

3. kernel method:

Cost function 11 ($c_{\mathbf{kernel}}$). *For a given kernel function $k(\cdot,\cdot):\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$, the cost function c_{kernel} is given by*

$$c_{kernel}(y_{a..b}) := \sum_{t=a+1}^{b}\|\phi(y_t)-\bar{\mu}_{a..b}\|_{\mathcal{H}}^2 \tag{C11}$$

where $\bar{\mu}_{a..b}\in\mathcal{H}$ is the empirical mean of the embedded signal $\{\phi(y_t)\}_{t=a+1}^{b}$ and $\|\cdot\|_{\mathcal{H}}$ is defined in (16).

1. rbf:

Cost function 12 ($c_{\mathbf{rbf}}$). *The cost function c_{rbf} is given by*

$$c_{rbf}(y_{a..b}) := (b-a) \;-\; \frac{1}{b-a}\sum_{s,t=a+1}^{b}\exp(-\gamma\|y_s-y_t\|^2) \tag{C12}$$

where $\gamma>0$ is the so-called bandwidth parameter.

The parametric cost function c_M (based on a Mahalanobis-type norm) can be extended to the non-parametric setting through the use of a kernel. Formally, the Mahalanobis-type norm $\|\cdot\|_{\mathcal{H},M}$ in the feature space $\mathcal{H}$ is defined by

$$\|\phi(y_s)-\phi(y_t)\|_{\mathcal{H},M}^2 = (\phi(y_s)-\phi(y_t))'\,M\,(\phi(y_s)-\phi(y_t)) \tag{20}$$

where M is a (possibly infinite dimensional) symmetric positive semi-definite matrix defined on $\mathcal{H}$. The associated cost function, denoted $c_{\mathcal{H},M}$, is defined below. Intuitively, using $c_{\mathcal{H},M}$ instead of c_M introduces a non-linear treatment of the data samples.

Search method

optimal detection

1. if K is known
    1. dynamic programming

$$\begin{aligned}\min_{|\mathcal{T}|=K} V(\mathcal{T}, y=y_{0..T}) &= \min_{0=t_0<t_1<\cdots<t_K<t_{K+1}=T}\sum_{k=0}^{K} c(y_{t_k..t_{k+1}}) \\ &= \min_{t\le T-K}\left[c(y_{0..t}) \;+\; \min_{t=t_0<t_1<\cdots<t_{K-1}<t_K=T}\sum_{k=0}^{K-1} c(y_{t_k..t_{k+1}})\right] \\ &= \min_{t\le T-K}\left[c(y_{0..t}) \;+\; \min_{|\mathcal{T}|=K-1} V(\mathcal{T}, y_{t..T})\right]\end{aligned} \tag{21}$$

2. if K is unknown
    1. dynamic programming wiht penalty PELT (pruned exact linear time)

is given by:

$$\text{if}\quad \left[\min_{\mathcal{T}} V(\mathcal{T}, y_{0..t})+\beta|\mathcal{T}|\right]+c(y_{t..s}) \ge \left[\min_{\mathcal{T}} V(\mathcal{T}, y_{0..s})+\beta|\mathcal{T}|\right]\quad\text{holds,}$$

$$\text{then } t \text{ cannot be the last change point prior to } T. \tag{23}$$

Approximate detection

1. window slicing

5.2.1. Window sliding

The window-sliding algorithm, denoted `Win`, is a fast approximate alternative to optimal methods. It consists in computing the discrepancy between two adjacent windows that slide along the signal y. For a given cost function $c(\cdot)$, this discrepancy between two sub-signals is given by

$$d(y_{a..t}, y_{t..b}) = c(y_{a..b}) - c(y_{a..t}) - c(y_{t..b}) \quad (1 \leq a < t < b \leq T). \tag{24}$$

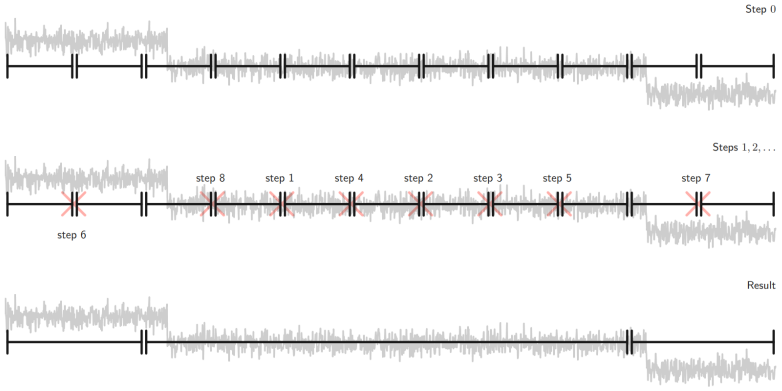
2. binary segmentation

Figure 9: Schematic example of `BinSeg`

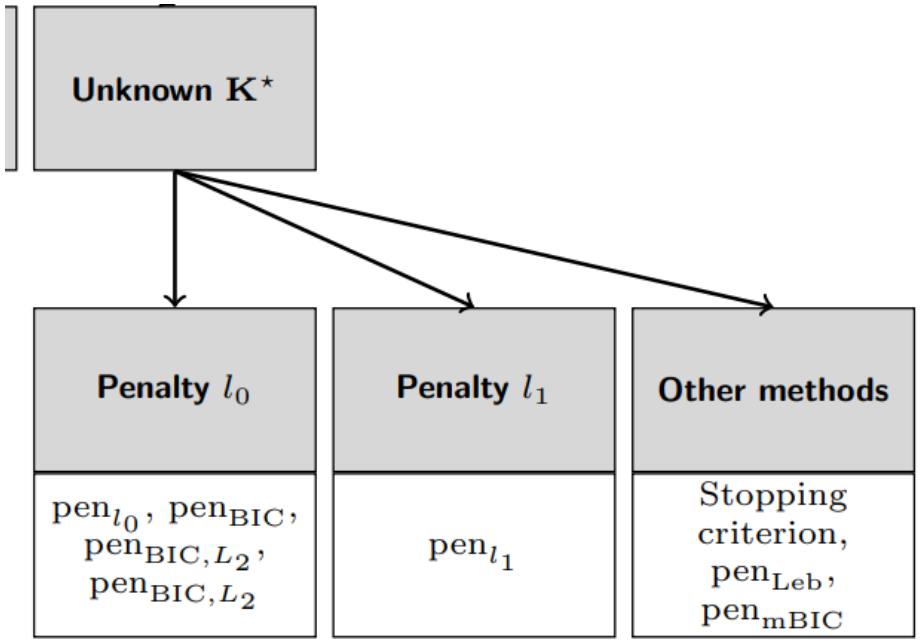
125]. `BinSeg` is a greedy sequential algorithm, outlined as follows. The first change point estimate $\hat{t}^{(1)}$ is given by

$$\hat{t}^{(1)} := \operatorname{argmin}_{1 \leq t < T-1} \underbrace{c(y_{0..t}) + c(y_{t..T})}_{V(\mathcal{T}=\{t\})}. \tag{27}$$

3. Bottom-up segmentation



Penalties


Unknown $\mathbf{K}^\star$
Penalty l_0
Penalty l_1
Other methods
pen_{l_0}, $\mathrm{pen}_{\mathrm{BIC}}$, $\mathrm{pen}_{\mathrm{BIC},L_2}$, $\mathrm{pen}_{\mathrm{BIC},L_2}$
pen_{l_1}
Stopping criterion, $\mathrm{pen}_{\mathrm{Leb}}$, $\mathrm{pen}_{\mathrm{mBIC}}$