Locating Changes in Highly Dependent Data with an Unknown Number of Change-Points

Azadeh Khaleghi and Daniil Ryabko
INRIA-Lille, Université de Lille, France

Problem

Setup

We are given a sequence

$$\mathbf{x} := X_1, \dots, X_n \in \mathbb{R}^n, \; n \in \mathbb{N}$$

which is formed as the concatenation of an **unknown number** $k+1$ of sequences

$$\underbrace{X_1, \dots, X_{\pi_1}}_{\sim\rho}, \underbrace{X_{\pi_1+1}, \dots, X_{\pi_2}}_{\sim\rho'}, \underbrace{X_{\pi_2+1}, \dots, X_{\pi_3}}_{\sim\rho}, \dots, \underbrace{X_{\pi_k+1}, \dots, X_n}_{\sim\rho''}$$

- Each sequence is generated by an **unknown** discrete-time stochastic process.
- The consecutive segments separated by $\pi_i, \; i = 1..k$ are generated by **different processes**.
- The indices $\pi_i, \; i = 1..k$ are called **change-points** and are **unknown**.
- The segments have lengths linear in n i.e.,

$$\pi_i := n\theta_i, \; i = 1..k, \; \theta_i \in (0,1)$$
$$\lambda_{\min} := \min_{\substack{i=1..k+1 \\ \theta_0:=0,\; \theta_{k+1}:=1}} \theta_i - \theta_{i-1} > 0$$

where $\theta_i, \; i = 1..k$ and $\lambda_{\min}$ are **unknown**.

Our goal is to estimate every change-point consistently.

Objective

We seek an **asymptotically consistent estimate** $\hat{\theta}_i(n)$ for every $\theta_i, \; i = 1..k$ so that with probability one we have

$$\lim_{n\to\infty} |\hat{\theta}_i(n) - \theta_i| = 0.$$

Main Result

Theorem (The proposed algorithm is asymptotically consistent).
Let $\mathbf{x} := X_1, \dots, X_n, \; n \in \mathbb{N}$ be a sequence with an **unknown number** k of change-points, $\pi_i := n\theta_i, \; i = 1..k$ and assume that the process distributions that generate $\mathbf{x}$ are stationary ergodic. The proposed algorithm takes the sequence $\mathbf{x}$ along with a parameter $\lambda \in (0,1)$ to produce a list $\hat{\theta}_1(n), \dots, \hat{\theta}_{1/\lambda}(n)$ of estimates. For all $\lambda \in (0, \lambda_{\min}]$, the first k elements of the produced list converge to some permutation of $\theta_1, \dots, \theta_k$ so that with probability one we have

$$\lim_{n\to\infty} \sup_{i=1..k} |\hat{\theta}_{[i]}(n) - \theta_i| = 0.$$

Assumptions

We consider an extremely general nonparametric framework.

- We allow the samples to be **dependent** and the dependence can be **arbitrary**.
- Our only assumption on the **unknown distributions** that generate the data is that they are **stationary ergodic**.
  $\Rightarrow$ We make no such assumptions as iid, Markov etc.
- We do **not require** the finite-dimensional marginals of any fixed size to be different.

We consider the **most general case**: **the process distributions change.**

This framework is similar to that of [1] where the single change-point problem was considered. It turns out that extensions to the multiple change-point problem is non-trivial.

Remark

The assumption that the process-distributions are stationary ergodic is one of the weakest assumptions in statistics. Typically in the change-point literature the **samples are assumed iid within segments**, the distributions have **known forms** and the **change is in the mean**. In non-parametric settings the **form of the change** and/or the **nature of dependence** are usually restricted. *For example the processes are assumed to be strongly mixing.* Moreover, it is *almost exclusively* assumed that the **finite-dimensional marginals** are *different*.

Number of Change-Points

An Impossibility Theorem [2]: For a pair of sequences generated by stationary ergodic processes, it is *impossible to distinguish* between the case where they are generated by *the same* process or by *different* ones.

It is therefore **impossible to estimate k** in this setting.

With the number k of change-points unknown, we have two choices
$\rightarrow$ ~~Make stronger assumptions~~
$\rightarrow$ Produce **a sorted list of change points** whose **first k elements** converge to some permutation of the **true change points**.

Distance Measure

We measure the distance between two sequences $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^{n'}$ as

$$\hat{d}(\mathbf{y}, \mathbf{z}) := \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in \mathcal{B}^{m,l}} |\nu(\mathbf{y}, B) - \nu(\mathbf{z}, B)|$$

where $\mathcal{B}^{m,l}$ $m, l \in \mathbb{N}$ is the set of all hypercubes of dimension m and edge-length 2^{-l} and $\nu(\mathbf{x}, B)$ is the frequency with which $\mathbf{x}$ crosses B; and $w_i := 2^{-i}$. As shown in [4] that if $\mathbf{y}$ and $\mathbf{z}$ are generated by **stationary ergodic** processes ρ and ρ', then $\hat{d}(\mathbf{y}, \mathbf{z})$ converges to the so-called **distributional-distance** [3] given by

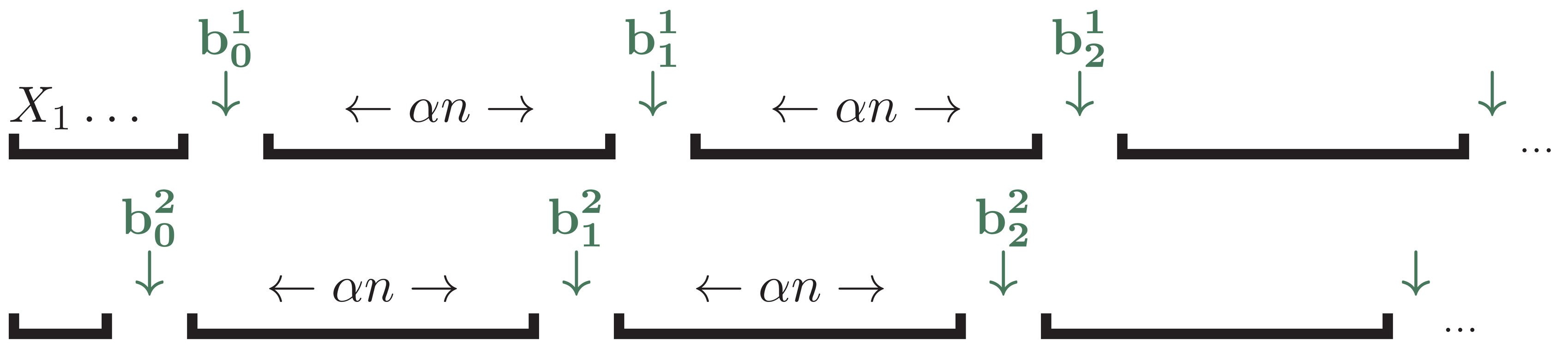
$$d(\rho, \rho') := \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in \mathcal{B}^{m,l}} |\rho(B) - \rho'(B)|.$$

Algorithm

input: $\mathbf{x} := X_1, \ldots, X_n$, $\lambda \in (0, \lambda_{\min}]$

1. Set interval size $\alpha \leftarrow \lambda/3$ and generate two sets of separators

$$b_i^t \leftarrow n\alpha(i + \frac{1}{t+1}),\ i = 0..\frac{1}{\alpha},\ t = 1, 2$$



where $b_0^1 := n\alpha/2,\ b_0^2 := n\alpha/3$

2. Estimate a change-point $\hat{\theta}_i^t$ in every segment as

$$\hat{\theta}_i^t := \frac{1}{n} \underset{t' \in b_i^t..b_{i+1}^t}{\operatorname{argmax}} \hat{d}(X_{b_{i-1}^t..t'}, X_{t'..b_{i+2}^t})$$

3. Calculate a performance score for every estimate $\hat{\theta}_i^t$ as

$$\Delta_{\mathbf{x}}(b_i^t, b_{i+1}^t) := \hat{d}(X_{b_i^t..c_i^t}, X_{c_i^t..b_{i+1}^t})$$

$$\text{where } c_i^t := \frac{b_i^t + b_{i+1}^t}{2}$$

4. Start from the set of all estimates

Do (While estimates are still available)

i. **Add** to the output list *an available estimate* $\hat{\theta}$ of **highest score**

ii. **Remove** all **estimates within** $\lambda/2$ from $\hat{\theta}$

output: A (sorted) list of change-point estimates.

Proof Sketch

• Since $\alpha \in (0, \lambda_{\min}/3]$ if a change-point $\pi_j := n\theta_j$ for some $j \in 1..k$ is contained within a segment $X_{b_i^t..b_{i+1}^t}$ for some $i \in 1..\alpha^{-1}$ (i.e. $\pi_j \in [b_i^t, b_{i+1}^t]$) then we have $[\pi_{j-1}, \pi_{j+1}] \subseteq [b_{i-1}^t, b_{i+2}^t]$.

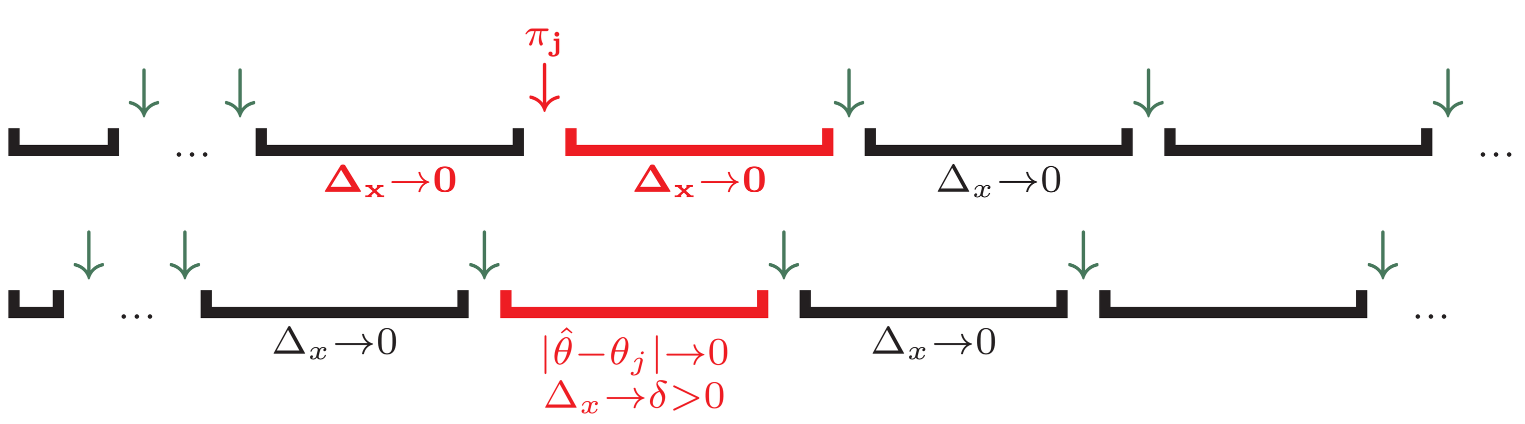
$$\ldots \underbrace{X_{b_{i-1}^t..b_i^t} \downarrow X_{b_i^t..\pi_j}}_{\sim\rho} \underbrace{X_{\pi_j..b_{i+1}^t} \downarrow X_{b_{i+1}^t..b_{i+2}^t}}_{\sim\rho'} \ldots$$

In this case (by the consistency of $\hat{d}(\cdot,\cdot)$) we can show that θ_i^t is a consistent estimate of θ_j i.e. $\hat{\theta}_{\mathbf{i}}^{\mathbf{t}} \rightarrow \theta_{\mathbf{j}}$, and is further assigned a score that converges to a non-zero constant i.e. $\mathbf{\Delta_x(b_i^t, b_{i+1}^t)} \rightarrow \delta > \mathbf{0}$.

• If $X_{b_i^t..b_{i+1}^t}$ does not contain any change-points then its performance score converges to 0, i.e. $\mathbf{\Delta_x(b_i^t, b_{i+1}^t) \rightarrow 0}$.

$$\ldots\ldots \underbrace{X_{b_i} \downarrow X_{b_i+1} \ldots X_{b_{i+1}} \downarrow X_{b_{i+1}+1}}_{\sim\rho} \ldots.$$

• Every change-point is (consistently) estimated at least once.



• Since $\lambda \in (0, \lambda_{\min}]$ the estimate of every true change-point appears at most once in the output.

Therefore,

◇ The algorithm provides a list of change-point estimates.

◇ The estimates are sorted according to their performance scores.

◇ The first k estimates converge to some permutation of the true change-points.

Experimental Results

Time-Series Generation

1. Fix some parameter $\alpha \in (0, 1)$, and select some length $n \in \mathbb{N}$.
2. Select $r_0 \in [0, 1]$ at random.
3. For each $i = 1..n$ obtain $r_i := r_{i-1} + \alpha - \lfloor r_{i-1} + \alpha \rfloor$.
4. Let $X_i := \mathbb{I}\{r_i > 0.5\}$ to generate $\mathbf{x} = X_1, \cdots, X_n$.

If α is irrational then $\mathbf{x}$ forms a **stationary-ergodic time-series** which does **not** belong to any "simpler" class. In particular, it cannot be modeled by a hidden Markov process with a finite state-space [5]. We simulate α by a `longdouble` with a long mantissa.

In our experiments we fixed $\lambda_{\min} = 0.23$ and generated a sequence with $k = 3$ change-points using $\alpha_1 := 0.30...$, $\alpha_2 := 0.35...$, $\alpha_3 := 0.40...$, $\alpha_4 := 0.45...$ (*with long mantissae*).

Consistency

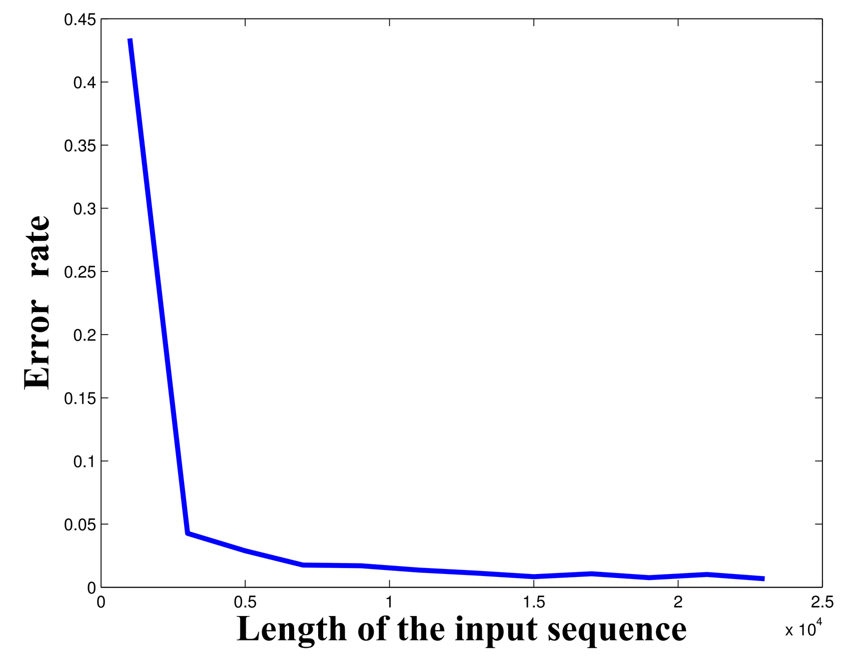


Figure 1: Error as a function of the sequence-length (avg. over 20 runs).

Dependance on λ

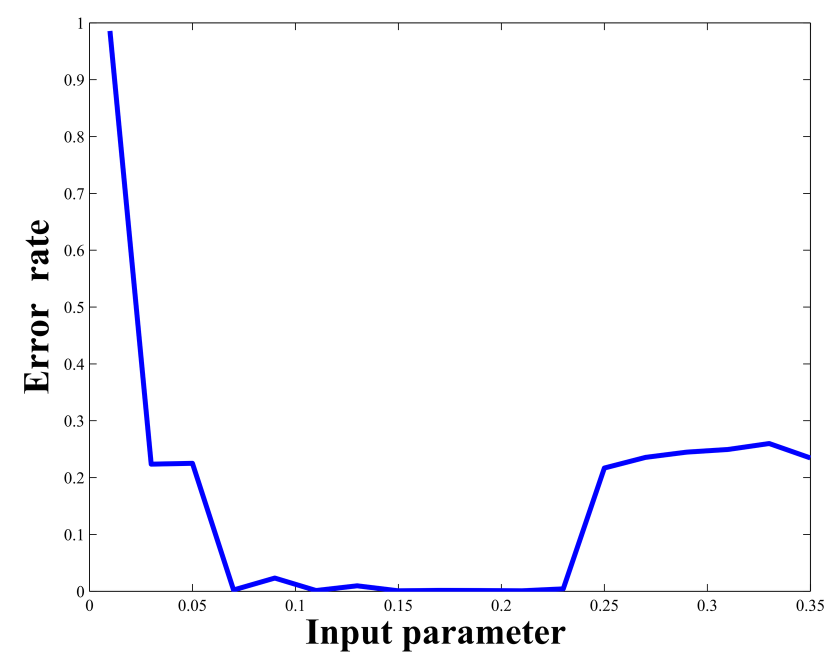


Figure 2: Error as a function of λ (avg. over 25 runs). The sequence-length is fixed to 20000, $\lambda_{\min} := 0.23$ and λ is varied.

Computational Complexity

$\boxed{\text{The computational complexity of the algorithm is } \mathcal{O}(n^2 \operatorname{polylog} n)}$

Even though the distance $\hat{d}(\cdot,\cdot)$ involves infinite summations it can be calculated efficiently.

• All summands corresponding to $m > n$ equal 0.

• All summands corresponding to $l > s_{\min}$ are equal where

$$s_{\min} := \min_{i,j \in 1..n,\ X_i \neq X_j} |X_i - X_j|$$

corresponds to the partition in which each cell contains at most one point. On the other hand, the frequencies of cells in $\mathcal{B}^{m,l}$ corresponding to higher values of m are not consistent estimates of their probabilities. Thus we may take m upto $\log n$ and still obtain consistent results; see also [4] and [6]. Therefore, the computational complexity of calculating the distance becomes $n \operatorname{polylog} n$ and that of the algorithm $n^2 \operatorname{polylog} n$.

References

[1] D. Ryabko, B. Ryabko, Nonparametric Statistical Inference for Ergodic Processes. IEEE Transactions on Information Theory. 2010.

[2] D. Ryabko. Discrimination between B-processes is impossible. Journal of Theoretical Probability. 2010.

[3] R. Gray. Prob. Random Processes, & Ergodic Properties Springer Verlag. 1988.

[4] D. Ryabko. Clustering processes. ICML 2010.

[5] P. Shields. Prob. The Ergodic Theory of Discrete Sample Paths AMS Bookstore. 1996.

[6] A. Khaleghi, et. al. Online Clustering of Processes. AISTATS 2012.