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

Azadeh Khaleghi and Daniil Ryabko

INRIA-Lille, Université de Lille, France

Spotlight Presentation at NIPS 2012
Poster ID: W68

The Change-Point Estimation Problem

We have a sequence

$$\mathbf{x} := X_1, \dots, X_n$$

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** time-series distribution.
- The **change-points** are **unknown** and we aim to estimate them.

Our results are vastly more general than the existing literature!

Framework

Change-point estimation turned into a real machine learning problem!

We consider an extremely general framework where

- The distributions are **completely unknown**.
- The samples can be **arbitrarily dependent**.
- Our only is that the **unknown distributions** are **stationary ergodic**.

  This is one of the weakest assumptions in statistics!

  ex. no such assumptions as i.i.d, Markov etc.

- We make **no assumptions** on the **form of the change** and/or **nature of dependence**.

  We simply search for the change in distribution.

- We do **not require** the **fixed-sized finite-dimensional marginals** before and after the change to be different.

Main Results

An approach analogous to Hierarchical Clustering

In this framework it is provably impossible to estimate k. We produce **a (sorted) list of estimates** whose first **first k elements** converge to the **true change-points**.

Theorem (Consistency)

The algorithm that we propose is asymptotically consistent.