Non-parametric change-point detection via string matching

Dino Sejdinovic

School of Mathematics
University of Bristol

Joint work with:
Oliver Johnson, Ayalvadi Ganesh (Bristol Maths),
James Cruise (Heriot Watt),
Robert Piechocki (Bristol CCR)
Overview

1. Match lengths and entropy
2. Using match locations to detect change-points
3. Simulation results
4. Consistency
Data sources

- Consider observing a finite-alphabet source of data with a change-point, i.e., at an unknown time the statistical properties of the source change.
- We do not know statistical properties of source and do not want to assume particular parametric family of distributions.
- However, we need to make inference about it.
Change-point detection

**Parametric framework:**
- postulate a parametric family model: data comes from a model with some parameters $\theta$
- detect changes in these parameters, e.g., in mean and variance of normal samples
- can use maximum likelihood principle

**Non-parametric framework:**
- monitoring changes in the empirical mean
- comparing empirical distribution before and after a putative changepoint

[Horvath, 1993]
[Brodsky, Darkhovsky, 1993]
Detecting change in entropy?

- 0/1: We could estimate long-term density of heads by counting, but we might also want to know ‘how random’ it is.
- Randomness is expressed through the entropy of source.

**Example**

Consider two binary sequences:

1. $x$: 01010101010101010110
2. $y$: 00101101011000101011

- Both $x$ and $y$ have 10 0’s and 10 1’s.
- However, first has a long periodic substring, the second seems random.
Detecting change in entropy? (2)

- How can we detect a change-point when the source switches from a boring to an interesting state or vice-versa?
- Similar examples can be constructed on which the crude bigram and trigram strategies fail.
- Need a systematic way to take into account all features.
Definition

Given sequence \((x_0, \ldots, x_{n-1})\) of length \(n\), write
\[x_{i}^{i+L-1} = (x_i, \ldots, x_{i+L-1})\]
for substring of length \(L\) starting at \(i\). For each \(i\), the match length at \(i\) is given by:

\[L_i^n(x) = \min\{L : x_{i}^{i+L-1} \neq x_{j}^{j+L-1} \text{ for all } i \neq j\}.

- \(L_i^n\) is the length of a shortest unique prefix starting at \(i\).
Substring matches

Example

Consider two binary sequences:

1. $x$: 01010101010101010110
2. $y$: 001011010110001011

- Substring $x_0^{15} =$ 0101010101010110 (length 16) seen again at $x_2^{17} : L_0^{20}(x) = 17$.
- Substring $y_0^{4} =$ 00101 (length 5) seen again at $y_2^{16}$, but nothing longer: $L_0^{20}(y) = 6$.

- “More random” sources explore bigger set of substrings and have shorter repeats than simpler ones.
- How large do we expect $L_i^n$ to be as $n$ grows?
Theorem

[Shannon-MacMillan-Breiman] Given stationary source of entropy $H$, there exists a ‘typical set’ $\mathcal{T}$ of strings of length $m$ such that:

1. A random string lies in $\mathcal{T}$ with probability $\geq 1 - \epsilon$.
2. Any individual string in $\mathcal{T}$ has probability $\sim 2^{-mH}$.

Heuristically, we can predict the size of match lengths as follows:

- If string length $m$ at point $i$ is typical, it has probability $\sim 2^{-mH}$, so we expect to see it $\sim n2^{-mH}$ times.
- Hence by choosing $m = \frac{\log n}{H}$, expect to see it once:

$$L_i^n \sim \frac{\log n}{H}.$$
Estimating entropy with match lengths

**Theorem**

[Shields 1992, Shields 1997] If match lengths $L_i^n$ are calculated for an IID or mixing Markov source with entropy $H$,

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} L_i^n}{n \log n} = \frac{1}{H}, \quad (a.s.).$$

- [Kontoyiannis and Suhov 1993] extends the convergence for a broad class of stationary sources.

- Non-parametric, computationally efficient entropy estimators with fast convergence in $n$ (they out-perform plug-in estimators).
Source model with a changepoint

**Definition**

Sample two independent sequences $x(1), x(2)$, where $x(i) \sim \mu_i$ for a stationary process $\mu_i$ with $i = 1, 2$. Then, given length and change point parameters $n$ and $\gamma$, define the concatenated process $x$ by:

$$x_i = \begin{cases} 
    x(1)_i & \text{if } 0 \leq i \leq n\gamma - 1, \\
    x(2)_i & \text{if } n\gamma \leq i \leq n - 1.
\end{cases}$$

- Given $x$, we hope to detect the change point – that is, to estimate the true value of $\gamma$. 

Match locations

- Consider match locations – for each $i$, write $T_i^n$ for a position of longest substring that agrees with $i$.

**Example**

Consider two binary sequences:

1. $x$: 01010101010101010110
2. $y$: 001011010110001011

- Substring $x_0^{15}$: 01010101010101 (length 16) seen again at $x_2^{17}$: $T_0^{20}(x) = 2$.
- Substring $y_0^4$: 00101 (length 5) seen again at $y_{12}^{16}$: $T_0^{20}(y) = 12$.

$T_i^n$ need not be unique: in the event of a tie, choose random one.
Using match locations to detect change points

- Idea: substrings of $x(1)$ likely to be similar to other substrings of $x(1)$.
- The same is true for $x(2)$.
- Expect that if $i < n\gamma$ then $T_{i}$ will tend to be $< n\gamma$.
- Similarly, for $i \geq n\gamma$, expect $T_{i}$ will tend to be $\geq n\gamma$. 
Grassberger tree of shortest prefixes

- Grassberger Tree is a q-ary labelled tree $T_n(x)$ which encodes the shortest unique prefixes of each substring.
- The set of all matches of substring at $i \equiv$ the set of leaves in a subtree rooted at a parent of $i$ (excluding $i$).
Grassberger tree of shortest prefixes

We choose a match location $T^n_i$ to be an element from the set of all matches chosen uniformly at random.
Counting crossings

Figure: Directed graph formed by linking $i$ to $T_i^n$
Counting crossings (2)

**Definition**

Given a putative change point $0 \leq j \leq n - 1$, we write

- $C_{LR}(j) = \# \{ k : k < j \leq T^n_k \}$ for the number of left-right crossings of $j$,
- $C_{RL}(j) = \# \{ k : T^n_k < j \leq k \}$ for the number of right-left crossings of $j$. 
Counting crossings (3)

- $C_{LR}(2) = 2, \; C_{RL}(2) = 3$.
- Intuitively, we look for index $j$ such that both $C_{LR}(j)$ and $C_{RL}(j)$ are small.
- However, $C_{LR}(j)$ and $C_{RL}(j)$ will be highest around the middle of the sequence. Normalization?
Definition

For $0 \leq j \leq n - 1$, define the normalized crossing processes:

$$\psi_{LR}(j) = \frac{C_{LR}(j)}{n-j} - \frac{j}{n}$$

and

$$\psi_{RL}(j) = \frac{C_{RL}(j)}{j} - \frac{n-j}{n},$$

and

$$\psi(j) = \max(\psi_{LR}(j), \psi_{RL}(j)).$$

CRECHE estimator of $\gamma$ is given by

$$\hat{\gamma} = \frac{1}{n} \arg\min_{0 \leq j \leq n-1} \psi(j).$$

- The processes $\psi_{LR}(j)$ and $\psi_{RL}(j)$ are designed via subtracting off the mean of $C_{LR}(j)$ and $C_{RL}(j)$
- Related to the conductance of the directed graph
Results for IID sources – no change point

- 50,000 symbols with distribution (0.5, 0.25, 0.25)
Results for IID sources – with change-point

- 10,000 symbols with distribution (0.1, 0.3, 0.6) vs.
- 40,000 symbols with distribution (0.5, 0.25, 0.25)
IID vs. Markov

- Markov chain with a stationary distribution (0.3, 0.4, 0.3) vs. IID with distribution (0.3, 0.4, 0.3): (1) $\gamma = 1/3$, (2) $\gamma = 2/3$. Plot based on 1000 trials
Markov chain with a stationary distribution \((0.3, 0.4, 0.3)\) vs. IID with distribution \((0.3, 0.4, 0.3)\): (3) \(\gamma = 1/2\), (4) empirical average of \(\psi\).
Results for text files – German vs. English

Excerpts from German original and English translation of Goethe’s Faust
Excerpts from English text by two different authors
Audio: speaker turn detection

Original    Speaker 1    Speaker 2
Analysis of a related toy problem

- Would like to theoretically analyse performance of estimator $\hat{\gamma}$ for this source and matching model.
- To show $\psi$ is minimised close to change point $n\gamma$, we need uniform control of $\psi_{LR}$ and $\psi_{RL}$.
- However, dependencies make analysis tricky.
- Match locations tend to be roughly independent and uniform, so we analyse related toy source model instead.
For each $i \in \{0,1,\ldots,n-1\}$, define $T_i^n$ to be independently uniformly distributed on $\{0,1,\ldots,n-1\}$.

- For each $j = 1, \ldots, n-1$, as before define

$$C_{LR}(j) = \# \{ k: k \leq j < T_k^n \}$$

for the number of LR crossings of $j$. Denote $\psi_{LR}$ and $\psi_{RL}$ as before.
Simple toy problem: confidence region

**Theorem**

Let $T^n_i$ be independently uniformly distributed on $\{0, 1, \ldots, n - 1\}$. For any $0 \leq \delta \leq 1$ and $s > 0$,

$$\mathbb{P}\left(\sup_{1 \leq j \leq n(1-\delta)} |\psi_{LR}(j)| \geq \frac{s}{\sqrt{n}}\right) \leq \frac{(1 - \delta)^2}{\delta s^2}.$$

**Proof Sketch:**

- We characterize the distribution of the crossing process $C_{LR}$ using Rényi’s thinning operation.

- This allows us to show that $\psi_{LR}$ is a martingale.

- Doob’s submartingale inequality allows us to uniformly bound the fluctuations of $\psi_{LR}$, as required.
Toy problem vs. simulation results

- Form of bound on $\psi_{LR}$ explains high values seen at RH end of the 'no change point' curve.
- By symmetry, form of bound on $\psi_{RL}$ explains high values on LH end.
- Considering the maximum of $\psi_{LR}$ and $\psi_{RL}$ ensures that the curve is close to zero in the middle: maximal fluctuations are of the order $O\left(\frac{1}{\sqrt{n}}\right)$. 

![Graph showing $\psi(j)$ vs. $j$](image-url)
Toy problem with a changepoint

Toy model: For a change location $\gamma$, and parameters $\alpha_L, \alpha_R \in [0, 1]$, define independent random variables $T_i^n$ such that:

1. for each $0 \leq i \leq n\gamma - 1$, 
$$
\mathbb{P}(T_i^n = j) \propto \begin{cases} 
1, & 0 \leq j \leq n\gamma - 1, \\
\alpha_L, & n\gamma \leq j \leq n - 1.
\end{cases}
$$

2. for each $n\gamma \leq i \leq n - 1$, 
$$
\mathbb{P}(T_i^n = j) \propto \begin{cases} 
\alpha_R, & 0 \leq j \leq n\gamma - 1, \\
1, & n\gamma \leq j \leq n - 1.
\end{cases}
$$
Toy problem with a changepoint (2)

\( \psi_{LR} \) is close to its deterministic mean function:

\[
\psi_{LR}(j) \sim \begin{cases} 
-\frac{C_0 j^2}{n(n-j)} & \text{for } j \leq n\gamma, \\
C_1 \frac{j}{n} - C_2 & \text{for } j \geq n\gamma,
\end{cases}
\]

for certain explicit constants \( C_0, C_1, C_2 \), depending on \( \alpha_L, \alpha_R \) and \( \gamma \).
Fluctuations from the mean

\begin{align*}
\Psi_{LR}(j) &
\end{align*}

\begin{align*}
\Psi_{RL}(j) &
\end{align*}

\begin{align*}
\psi(j) &
\end{align*}

\begin{align*}
d_1(j) & \\
 d_2(j) & \\
 \psi_{LR}(j) &
\end{align*}
Toy problem vs. simulation results

- Form of mean functions explain form of curves seen in change-point graphs
Theorem

The estimator $\hat{\gamma}$ is $\sqrt{n}$-consistent: there exists a constant $K$, depending on $\alpha_L$, $\alpha_R$ and $\gamma$, such that for all $s$:

$$\mathbb{P}\left( |\hat{\gamma} - \gamma| \geq \frac{s}{\sqrt{n}} \right) \leq \frac{K}{s^2}.$$ 

Proof sketch:

- Use the insights from the no-changepoint case - scaled version of the crossings process minus the deterministic part is a martingale.

- The proof follows from Doob’s submartingale inequality and the union bound.
Conclusions

- A new fully non-parametric, model-free change-point estimator, based on ideas from information theory
- Promising performance for a variety of data sources
- $\sqrt{n}$- consistency in a related toy problem
- Multiple change-points? Streaming?
References