A short introduction to Wild Binary Segmentation, Fryzlewicz [2014]

Sungmin Park

School of Statistics
University of Minnesota

November 3, 2022

Change point detection

We have a sequence of data $\{X_1, \ldots, X_T\}$. There may exist a change point in the data s.t.

$$X_t \sim \begin{cases} f_0, & t \leq \tau \\ f_1, & t > \tau \end{cases}$$

The goal is to determine whether a change exists and accurately estimate τ when it exists.

Single change point

Consider the following example from Yu [2020]. Suppose $X_t \sim N(\mu_t, 1)$. We would like to test whether there is a change point in $\{X_t\}_{t=1}^{T}$. If there is no change point:

$$H_0 : \mu_1 = \ldots = \mu_n.$$

If there is a change point at time point t:

$$H_{1,\mathrm{t}} : \mu_1 = \ldots = \mu_t \neq \mu_{t+1} = \ldots = \mu_n.$$

The alternative hypothesis is

$$H_1 = \bigcup_{t=1}^{n-1} H_{1,t}.$$

Single change point

Define three different sample means:

$$\bar{X}_{0,\tau} = \frac{1}{\tau}\sum_{t=1}^{\tau} X_t, \quad \bar{X}_{1,\tau} = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} X_t, \quad \bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t,$$

the pre-change, post-change, and overall sample mean according to change point at time τ. The likelihood ratio statistic of H_0 vs. $H_{1,\tau}$ is

$$-\frac{1}{2}\left\{\sum_{t=1}^{\tau}(X_t - \bar{X}_{0,\tau})^2 + \sum_{t=\tau+1}^{T}(X_t - \bar{X}_{1,\tau})^2\right\} + \frac{1}{2}\sum_{t=1}^{T}(X_t - \bar{X})^2$$

$$= \frac{\tau(T-\tau)}{2T}\left(\bar{X}_{0,\tau} - \bar{X}_{1,\tau}\right)^2$$

To test H_0 vs H_1, one would reject H_0 when

$$\max_{1\le\tau\le n-1}\left\{\sqrt{\frac{\tau(T-\tau)}{T}}\left|\bar{X}_{0,\tau} - \bar{X}_{1,\tau}\right|\right\} \ge c.$$

Multiple change point

- Now, suppose there exist multiple change points $\tau_1, \ldots, \tau_K$ s.t.
$$X_t \sim N(\mu_k, 1), \quad \text{if } \tau_k < t \leq \tau_{k+1}$$
with $\tau_0 = 0$ and $\tau_{K+1} = n$. Moreover, the number of change points K may be unknown.
- While there are other methods that aim to identify all the change points simultaneously, this presentation will focus on methods that sequentially find one change point at a time.
- *Binary segmentation* applies the previous likelihood ratio test procedure recursively to identify all the change points.

Binary segmentation

Let $\tilde{X}_{s,e}^b = \sqrt{\frac{(e-b)}{(e-s+1)(b-s+1)}} \sum_{t=s}^{b} X_t - \sqrt{\frac{b-s+1}{(e-s+1)(e-b)}} \sum_{t=b+1}^{e} X_t$.

function BinSeg(s, e, ζ_T)
    if $e - s < 1$ **then**
        STOP
    else
        $b_0 := \arg\max_{b \in \{s,\ldots,e-1\}} |\tilde{X}_{s,e}^b|$
        if $|\tilde{X}_{s,e}^{b_0}| > \zeta_T$ **then**
            add b_0 to the set of estimated change points
            BinSeg(s, b_0, ζ_T)
            BinSeg($b_0 + 1, e, \zeta_T$)
        else
            STOP
        end if
    end if
end function

Binary segmentation

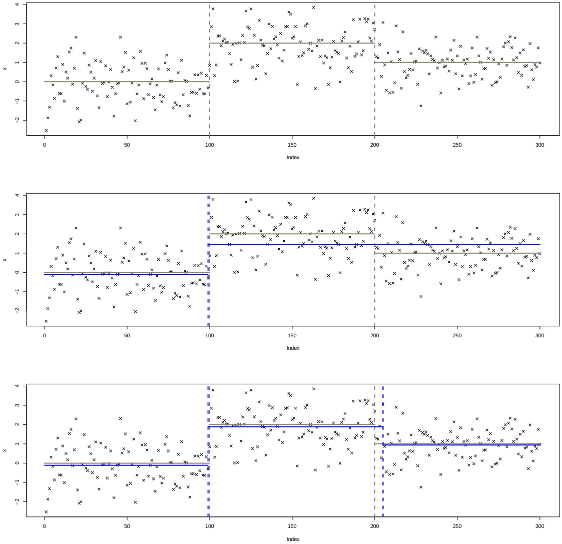

Figure: An example of binary segmentation

Limitations

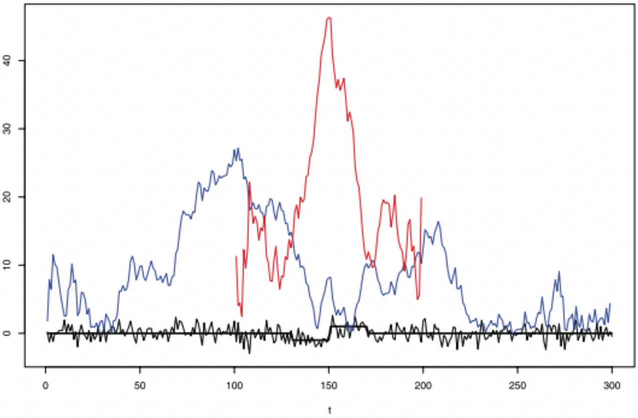

FIG. 1. *True function* f_t, $t = 1, \ldots, T = 300$ (*thick black*), *observed* X_t (*thin black*), $|\tilde{X}^b_{1,300}|$ *plotted for* $b = 1, \ldots, 299$ (*blue*), *and* $|\tilde{X}^b_{101,200}|$ *plotted for* $b = 101, \ldots, 199$ (*red*).

Figure: An example from Fryzlewicz [2014]

Limitations

Problem:

- At each step, binary segmentation assumes that the given segment contains *at most one change point*.
- It can be problematic when a segment contains multiple change points and signals that offset one another.

Solution:

- Narrow your search to smaller local segments that only contain one change point

Signal strength at local segments

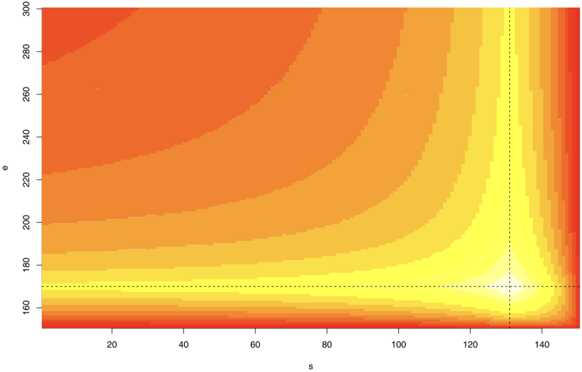

FIG. 2. *Heat map of the values of* $|\tilde{f}_{s,e}^{150}|$ *as a function of* s *and* e*: the lighter the colour, the higher the value. The two dashed lines indicate the location of the maximum,* $(s, e) = (131, 170)$.

Figure: Signal strength at local segments

Wild binary segmentation

However,

- We do not know in advance which local segments will have the strongest signal.
- Checking every local segment is not feasible as well.

Fryzlewicz [2014] introduced *Wild binary segmentation*:

1. Generate random intervals $[s_m, e_m]$, $m = 1, \ldots, M$
2. Compute $(m_0, b_0) = \arg\max_m \arg\max_b |\tilde{X}^b_{s_m,e_m}|$ where $\tilde{X}^b_{s_m,e_m} = \sqrt{\frac{(e_m-b)(b-s_m+1)}{(e_m-s_m+1)}} \left(\frac{1}{b-s_m+1} \sum_{t=s_m}^{b} X_t - \frac{1}{e_m-b} \sum_{t=b+1}^{e_m} X_t \right)$.
3. Add a change point if $|\tilde{X}^{b_0}_{s_{m_0},e_{m_0}}| \geq \zeta_T$

Wild binary segmentation

function WildBinSeg(s, e, M, ζ_T)
  if $e - s < 1$ **then**
    STOP
  else
    Randomly generate $[s_m, e_m] \subseteq [s, e]$ for $1 \leq m \leq M$
    $(m_0, b_0) := \arg\max_{1 \leq m \leq M} \arg\max_{b \in \{s_m, \ldots, e_m - 1\}} |\tilde{X}^b_{s_m, e_m}|$
    if $|\tilde{X}^{b_0}_{s_{m_0}, e_{m_0}}| > \zeta_T$ **then**
      add b_0 to the set of estimated change points
      WildBinSeg(s, b_0, M, ζ_T)
      WildBinSeg($b_0 + 1, e, M, \zeta_T$)
    else
      STOP
    end if
  end if
end function

Consistency

Consider a model

$$X_t = f_t + \epsilon_t, \quad t = 1, \ldots, T$$

where f_t is a bounded piece-wise constant signal and ϵ_t is i.i.d. $N(0, 1)$ random noise. Let $N, \eta_1, \ldots, \eta_N$ be the number and location of true change points in f_t.

Theorem (Consistency)

Under regularity conditions (on change magnitude, distance between change points, decision threshold, standard Gaussian error, etc.), wild binary segmentation is consistent. That is, for estimated change points $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$, we have

$$\hat{N} = N; \quad \max_{1 \leq i \leq N} \frac{|\eta_i - \hat{\eta}_i|}{T} = o(1)$$

with high probability.

Example

Observe $X_t = f_t + e_t$, $t = 1, \ldots, 300$ with $e_t \sim N(0,1)$ i.i.d. and $f_t = -\mathbb{I}(130 < t \leq 150) + \mathbb{I}(150 < t \leq 170)$

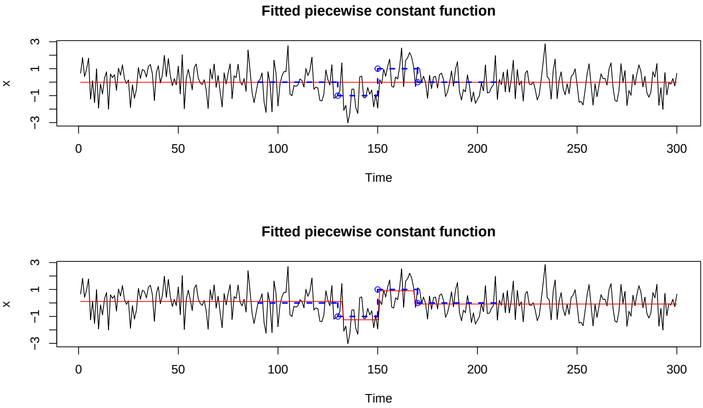

Figure: Above: binary segmentation, below: wild binary segmentation

Reference

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6): 2243–2281, 2014.

Yi Yu. A review on minimax rates in change point detection and localisation. *arXiv preprint arXiv:2011.01857*, 2020.