Narrowest-over-threshold detection of multiple change points and change-point-like features

by Rafal Baranowski, Yining Chen, Piotr Fryzlewicz
Presenter: Sungmin Park

School of Statistics
University of Minnesota

November 29, 2022

Problem setup

Consider a univariate statistical model

$$Y_t = f_t + \sigma_t \epsilon_t, \quad t = 1, \ldots, T.$$

where the signal f_t is believed to display some regularity across the index t, ϵ_t is a random noise with mean zero and unit variance, and σ_t is the magnitude of the noise.

A popular choice of f_t is a piecewise constant function for which the problem becomes a change point detection in the mean.

Here, we allow f_t to be any piecewise parametric function.

Examples of f_t

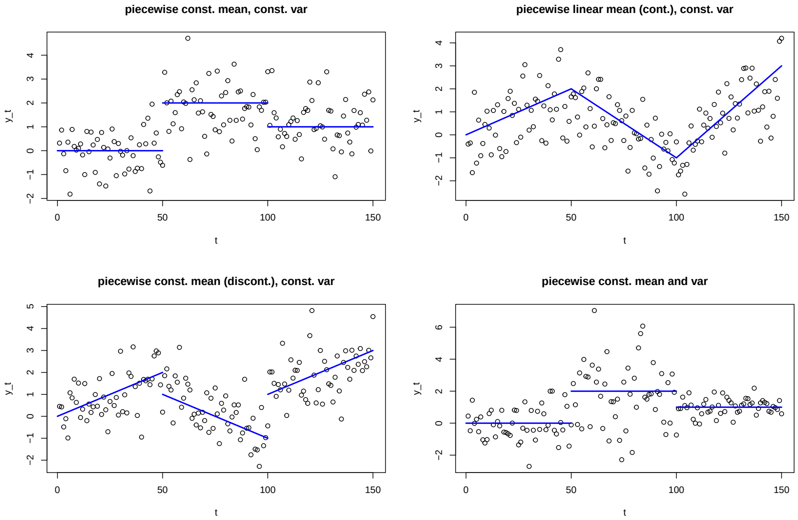


Figure: Examples of f_t

Overview of the change point literature

- Single change point / piecewise constant model
  - Generalized likelihood ratio test
- Multiple change points / piecewise constant model
  - Binary segmentation
  - Wild binary segmentation
- Multiple change points / piecewise parametric model
  - Narrowest-over-threshold (NOT) ($\star$)

Single change / piecewise constant model

Assume $f_t = \mu_0 I(t \leq \tau) + \mu_1 I(t > \tau)$ and $\epsilon_i \sim N(0, \sigma^2)$ i.i.d. where σ^2 is known, i.e.

$$Y_t \sim \begin{cases} N(\mu_0, \sigma^2), & t \leq \tau, \\ N(\mu_1, \sigma^2), & t > \tau. \end{cases}$$

The (generalized) likelihood ratio (LR) statistic for testing

$$H_0 : \mu_0 = \mu_1 \quad \text{vs.} \quad H_1 : \mu_0 \neq \mu_1 \text{ (location } \tau \text{ unknown)}$$

is $\max_{1 \leq b \leq T-1} C^b$ where

$$C^b = \sqrt{\frac{b(T-b)}{T}} \left| \left(\frac{1}{b} \sum_{t=1}^{b} X_t \right) - \left(\frac{1}{T-b} \sum_{t=b+1}^{T} X_t \right) \right|.$$

where C^b corresponds to the LR statistic corresponding to

$$H_1^b : \mu_0 \neq \mu_1, \ \tau = b.$$

Generalized likelihood ratio test

$$C^b = \sqrt{\frac{b(T-b)}{T}} \left| \left(\frac{1}{b} \sum_{t=1}^{b} X_t \right) - \left(\frac{1}{T-b} \sum_{t=b+1}^{T} X_t \right) \right|.$$

Given some threshold ζ_T, the generalized likelihood ratio (GLR) statistic rejects H_0 if

$$\max_b C^b > \zeta_T,$$

and place a change point at

$$b^* = \arg\max_b C^b.$$

The generalized likelihood ratio test will be the central component in various methods introduced in this presentation.

Multiple change point / piecewise constant model

Assume $f_t = \sum_{j=1}^{q+1} \mu_j I(t \in R_j)$ piecewise constant function with multiple change points. One can

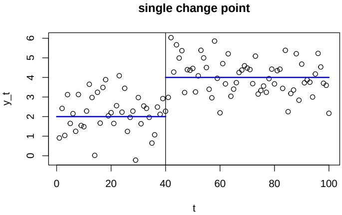
- identify q change points simultaneously,
- recursively find one change point at a time

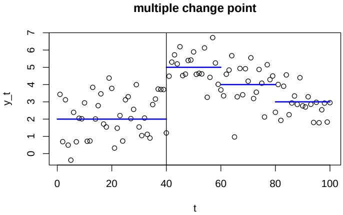
The latter (called *binary segmentation*) is popular for its computational efficiency.

function $\textsc{BinSeg}(s, e, \zeta_T)$
    $b^* := \arg\max_{b \in \{s,\dots,e-1\}} C_{s,e}^{b}$
    if $C_{s,e}^{b^*} > \zeta_T$ **then**
        add b^* to the set of estimated change points
        $\textsc{BinSeg}(s, b^*, \zeta_T)$
        $\textsc{BinSeg}(b^* + 1, e, \zeta_T)$
    end if
end function

Binary segmentation

Let us revisit the GLR test. The GLR test was designed for the problem when there is at most one change point. It may not work under multiple change points setting.





The choice of sub-segments to detect change point $t = 40$,

- $[1, 100]$ contains multiple change points,
- $[30, 50]$ contains not enough samples,
- $[1, 60]$ is the largest interval with a single change point.

Wild binary segmentation

Problem We do not know which segment is ideal (largest interval with a single change point).
Solution Randomly draw sub-intervals from [1,T]. Hopefully, some of them will be close to ideal, i.e. produce large test statistic. (*Wild binary segmentation*, Fryzlewicz [2014])

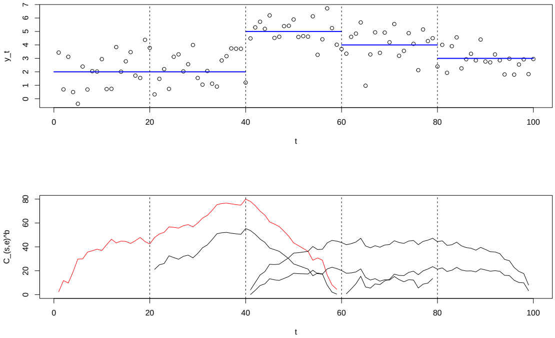


Figure: $C^{b}_{(s,e)}$ for different sub-segments

Wild binary segmentation

function WildBinSeg(s, e, M, ζ_T)

Randomly generate $[s_m, e_m] \subseteq [s, e]$ for $m = 1, \ldots, M$

$(m^*, b^*) := \arg\max_{1\le m\le M} \arg\max_{b\in\{s_m,\ldots,e_m-1\}} C^b_{s_m,e_m}$

if $C^{b^*}_{s_{m^*},e_{m^*}} > \zeta_T$ **then**

add b^* to the set of estimated change points

WildBinSeg(s, b^*, M, ζ_T)

WildBinSeg$(b^* + 1, e, M, \zeta_T)$

end if

end function

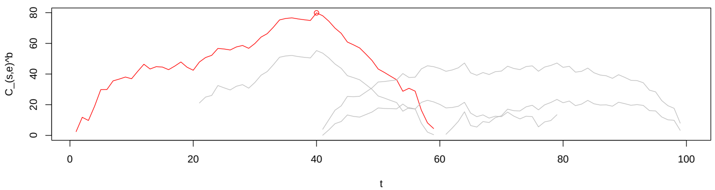


Figure: Illustration of wild binary segmentation

Multiple change / piecewise parametric model

Suppose the distribution of y_t can be modelled parametrically by a local real-valued d-dimensional parameter vector Θ_j:

$$y_t \sim g(\Theta_j) \quad \text{for} \quad \tau_{j-1} < t \leq \tau_j, \quad (1 \leq j \leq q+1),$$

with $\tau_0 = 0$ and $\tau_{q+1} = T$ (e.g. cubic spline).

For identifiability, each segment between change points will require at least d observations.

Multiple change / piecewise parametric model

If $q = 1$, the GLR statistic on the interval $(s, e]$ is

$$R^b_{(s,e]}(Y) = 2 \log \left[\frac{\sup_{\Theta^1,\Theta^2} \{ l(Y_{(s+1):b}; \Theta^1) l(Y_{(b+1):e}; \Theta^2) \}}{\sup_{\Theta} l(Y_{(s+1):e}; \Theta)} \right],$$
$$R_{(s,e]}(Y) = \max_{s+d \le b \le e-d} R^b_{(s,e]}(Y), \tag{1}$$

Place a change point at $\arg\max_b R^b_{(s,e]}(Y)$ if $R_{(s,e]}(Y) > \zeta_T$.

If constraints are added on Θ_1 and Θ_2 (e.g. continuity at the knots), supremum is taken over the constrained region in the numerator of (1).

Contrast functions

In many applications the GLR statistic (1) can be simplified with the help of *contrast functions* under the setting of Gaussian noise.

More precisely, for every (s, e, b) with $s < e$, find $C^b_{(s,e]}(Y)$ s.t.

1. $\arg\max_b C^b_{(s,e]}(Y) = \arg\max_b R^b_{(s,e]}(Y)$,
2. heuristically speaking, $C^b_{(s,b]}(Y)$ is relatively small if there is no change point in $(s, e]$,
3. the formulation of $C^b_{(s,e]}(Y)$ mainly consists of taking inner products between the data and some deterministic contrast vector.

Contrast functions

Scenario 1

$y_t = f_t + \epsilon_t$ with piecewise constant f_t and $\epsilon_t \sim N(0, \sigma^2)$ i.i.d. Assume σ^2 is known. Then

$$
\begin{aligned}
C^b_{(s,e]}(v) &= \left| \langle v, \psi^b_{(s,e]} \rangle \right| \\
&= \sqrt{\frac{(b-s)(e-b)}{(e-s)}} \left| \frac{1}{b-s} \sum_{t=s+1}^{b} v_t - \frac{1}{e-b} \sum_{t=b+1}^{e} v_t \right|,
\end{aligned}
$$

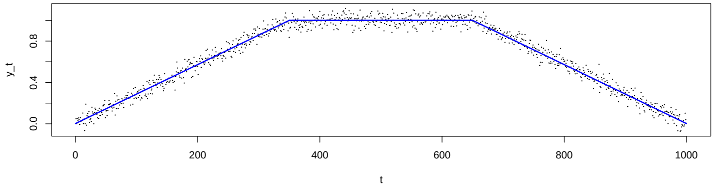
$$
\psi^b_{(s,e]}(t) = \begin{cases} \sqrt{\frac{e-b}{(e-s)(b-s)}}, & s < t \le b, \\ -\sqrt{\frac{e-b}{(b-s)(e-b)}}, & b < t \le e, \\ 0, & \text{o.w.} \end{cases}
$$

Contrast functions

Scenario 2

$y_t = f_t + \epsilon_t$ with piecewise linear f_t that is continuous at the change points, and $\epsilon_t \sim N(0, \sigma^2)$ i.i.d. Assume σ^2 is known. Then there exists a contrast vector $\phi^b_{(s,e]}$ that is completely characterized by (s, e, b) s.t.

$$C^b_{(s,e]}(v) = \left| \langle v, \phi^b_{(s,e]} \rangle \right|$$



Multiple change point detection in scenario 2

function WILDBINSEG(s, e, M, ζ_T)

    Randomly generate $[s_m, e_m] \subseteq [s, e]$ for $m = 1, \ldots, M$

    $(m^*, b^*) := \arg\max_{1 \le m \le M} \arg\max_{b \in \{s_m, \ldots, e_m - 1\}} C^b_{s_m, e_m}(Y)$

    ⋮

end function

Can we apply the ideas of wild binary segmentation to detect the change points in scenario 2? The answer is **NO.**

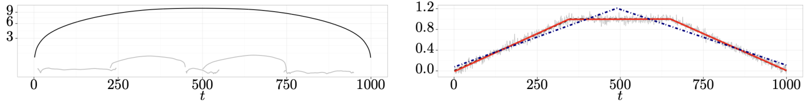


Figure: Values of $C^b_{(s,e]}$ for different intervals (left), observed data and the underlying model (right). The whole segment produces the strongest signal and is maximized at $t = 500$.

Goldilocks problem

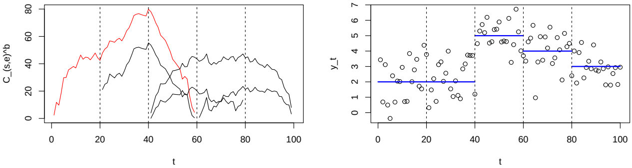


Figure: Values of $C^b_{(s,e]}$ for piecewise constant model

Recall the problem of detecting change point at $t = 40$:

- $[1, 100]$ contains multiple change points (too large),
- $[30, 50]$ contains not enough samples (too short),
- $[1, 60]$ contains sufficiently large sample while having a single change point (just right).

Goldilocks problem

The success of wild binary segmentation depneds on the ability to identify the interval with a *single change point* from a random pool of intervals.

function $\textsc{WildBinSeg}(s, e, M, \zeta_T)$

  Randomly generate $[s_m, e_m] \subseteq [s, e]$ for $m = 1, \ldots, M$

  $(m^*, b^*) := \arg\max_{m=1,\ldots,M} \arg\max_{b \in \{s_m,\ldots,e_m-1\}} C^b_{s_m,e_m}$

  ⋮

end function

Wild binary segmentation implicitly assumes that an interval with a single change point will produce the strongest signal.

It may be true for piecewise constant models, but does not hold in general for piecewise parametric models.

Narrowest-over-threshold method

To identify an interval with a single change point, we need to make sure that

- the interval produces a signal above some threshold,
- but is short enough to contain only one change point,

which leads to the narrowest-over-threshold (NOT) method, Baranowski et al. [2019].

function NOT(s, e, M, ζ_T)

  Randomly generate $[s_m, e_m] \subseteq [s, e]$ for $m = 1, \ldots, M$

  $O_{(s,e]} = \{m : \max_b C^b_{s_m,b_m}(Y) > \zeta_T\}$

  if $O_{(s,e]} \neq \varnothing$ **then**

    $m^* \in \arg\min_{m \in O_{(s,e]}} |e_m - s_m|$,

    $b^* = \arg\max_b C^b_{(s_{m^*}, e_{m^*}]}(Y)$

    add b^* to the set of estimated change points

    NOT(s, b^*, M, ζ_T)

    NOT$(b^* + 1, e, M, \zeta_T)$

  end if

end function

Narrowest-over-threshold method

Let us go back to the example in scenario 2.

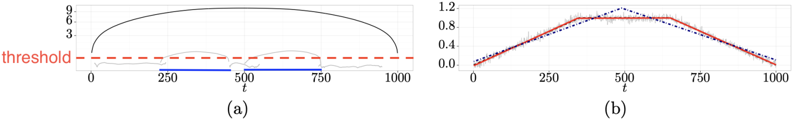


Figure: Threshold ζ_T that will identify the "right" sub-intervals containing single change points

Choosing appropriate threshold ζ_T will correctly identify the sub-intervals containing single change points.

Now, how do we choose the right threshold ζ_T?

Consistency

Theorem

Suppose f_t is a piecewise linear function that is continuous at the knots and $\epsilon_t \sim N(0,1)$ i.i.d. Let $\delta_T = \min_{j=1,\dots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^f = |(f_{\tau_j} - f_{\tau_j - 1}) - (f_{\tau_j+1} - f_{\tau_j})|$, $\underline{f}_T = \min_{j=1,\dots,q} \Delta_j^f$. Let $\hat{q}$ and $\hat{\tau}_1, \dots, \hat{\tau}_q$ denote the number and locations of change points from the NOT algorithm. Then there are constants, $\underline{C}, C_1, C_2, C_3 > 0$ (not depending on T) such that, given $\delta_T^{3/2}\underline{f}_T \geq \underline{C}\sqrt{\log(T)}$, $C_1\sqrt{\log(T)} \leq \zeta_T \leq C_2\delta_T^{3/2}\underline{f}_T$ and $M \geq 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1})$, as $T \to \infty$,

$$\mathbb{P}\left[\hat{q} = q, \max_{j=1,\dots,q}\left\{|\hat{\tau}_j - \tau_j|(\Delta_j^f)^{2/3}\right\} \leq C_3 \log(T)^{1/3}\right] \to 1.$$

- $\delta_T^{3/2}\underline{f}_T$ determines the difficulty of the problem.
- Wild binary segmentation is not consistent under the assumptions of this theorem.

Consistency

Corollary

On top of the assumptions from the previous theorem, assume additionally that we have finitely many change points, so that $\delta_T \asymp T$. *Then* $M \geq 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1}) \asymp \log(T)$. *In addition, when* $\underline{f}_t \asymp T^{-1}$, *as* $T \to \infty$,

$$\mathbb{P}\left[\hat{q} = q, \max_{j=1,\ldots,q}\left\{|\hat{\tau}_j - \tau_j|\right\} \leq C_3 T^{2/3}\log(T)^{1/3}\right] \to 1,$$

so

$$\max_{j=1,\ldots,q}\left\{|\hat{\tau}_j - \tau_j|\right\} = o_p(T).$$

Strengthened Schwarz information criterion

Although the theorem provides condition for the threshold ζ_T, they depend on unknown constants, namely $C_1\sqrt{\log(T)} \leq \zeta_T \leq C_2\delta_T^{3/2}\underline{f}_T$.

As a data-driven method to choose the threshold ζ_T, information-based criterion is considered.

strengthened Schwarz information criterion (sSIC)

Given a threshold ζ, let $\hat{q}(\zeta)$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}(\zeta)}$ be the number and the location of change points, and $n(\zeta)$ the number of estimated parameters in all $(\hat{\tau}_j, \hat{\Theta}_j)$'s. Define

$$\mathrm{sSIC}(\zeta) = -2\sum_{j=1}^{\hat{q}(\zeta)+1} \log\{l(Y_{\hat{\tau}_{j-1}+1:\hat{\tau}_j}; \hat{\Theta}_j)\} + n(\zeta)\log^{\alpha}(T),$$

for some pregiven $\alpha \geq 1$ with $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{q}(\zeta)+1} = T$. When $\alpha = 1$, this is the Schwarz/Bayesian information criterion.

Narrowest-over-threshold method with sSIC

Theorem

Suppose f_t is a piecewise linear function that is continuous at the knots and $\epsilon_t \sim N(0,1)$ i.i.d. Let $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^f = |(f_{\tau_j} - f_{\tau_j-1}) - (f_{\tau_j+1} - f_{\tau_j})|$, $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^f$. Furthermore, assume that q does not increase with T, $\delta_T/T \geq \underline{C}_1$, $\underline{f}_T^T \geq \underline{C}_2$ and $\max_{t=1,\ldots,T}|f_t| \leq \bar{C}$ for some $\underline{C}_1, \underline{C}_2, \bar{C} > 0$. Let $\hat{q}$ and $\hat{\tau}_1, \ldots, \hat{\tau}_q$ denote the number and locations of change points estimated with ζ_T picked via sSIC using $\alpha > 1$. Then there is a constant C (not depending on T) s.t. given $M \geq 36\underline{C}_1^{-2}\log(\underline{C}_1^{-1}T)$, as $T \to \infty$,

$$\mathbb{P}\left[\hat{q} = q, \max_{j=1,\ldots,q} |\hat{\tau}_j - \tau_j| \leq C\sqrt{T\log(T)}\right] \to 1.$$

Although the theorem requires $\alpha > 1$, the authors simply use $\alpha = 1$ (SIC / BIC) in their simulation studies.

Narrowest-over-threshold solution path algorithm

To choose the best threshold according to sSIC, we need to compute the solution path of

$$\mathcal{T}(\zeta_T) = \{\hat{\tau}_1(\zeta_T), \dots \hat{\tau}_{\hat{q}(\zeta_T)}(\zeta_T)\}$$

the location of chnage points estimated using ζ_T. However, calculating sSIC for all ζ_T using a grid search is impractical.

To compute the entire path efficiently, the authors note that

1. as a function, $\zeta_T \to \mathcal{T}(\zeta_T)$ changes its value only at discrete points.
2. there exists $\zeta_T^{\varnothing} > 0$ s.t. $\mathcal{T}(\zeta_T) = \varnothing$ for all $\zeta_T \geq \zeta_T^{\varnothing}$.

Narrowest-over-threshold solution path algorithm

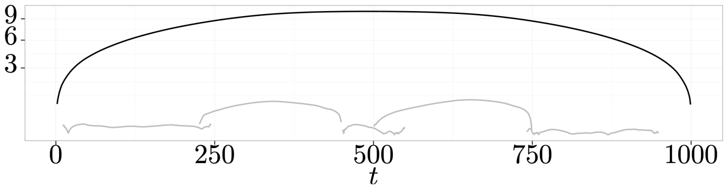


Figure: Realizations of $C^b_{(s,e]}(Y)$

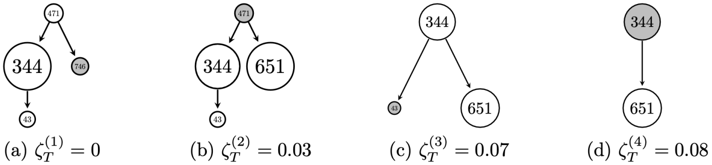


Figure: Sequence of $T(\zeta_T)$ with ζ_T increasing from 0 to 0.08

Narrowest-over-threshold with unknown σ

Although the theorem required $\sigma_t = 1$ for simplicity, it can be estimated accurately by using median absolute deviation (MAD) method for i.i.d. Gaussian errors. For $Z \sim N(0, 1)$,

$$\text{median}(|Z|) \approx \Phi^{-1}(3/4) \approx 0.674.$$

Observe that

1. for piecewise constant model, $(Y_{i-1} - Y_i) \sim N(0, 2)$,
2. for piecewise linear model, $(Y_{i-1} - 2Y_i + Y_{i+1}) \sim N(0, 6)$,

i.i.d. for all i except at, or near, change points.

1. $\hat{\sigma} = \text{median}\{|Y_{i-1} - Y_i|\}/\{\Phi^{-1}(3/4)\sqrt{2}\}$,
2. $\hat{\sigma} = \text{median}\{|Y_{i-1} - 2Y_i + Y_{i+1}|\}/\{\Phi^{-1}(3/4)\sqrt{6}\}$.

Narrowest-over-threshold under heavy-tailed noise

Assume a piecewise constant f_t but a heavy-tailed noise for ϵ_t. The authors propose a new contrast function

$$\tilde{C}^b_{(s,e]} = \langle S_{(s,e]}(Y), \psi^b_{(s,e]} \rangle$$

where

$$\left[S_{(s,e]}(Y)\right]_i = \text{sign}\left\{ v_i - \frac{1}{e-s} \sum_{t=s+1}^{e} v_t \right\}.$$

The idea is to transform the heavy-tailed observation into one that is not heavy-tailed, observation of 1's and -1's.

Simulation study

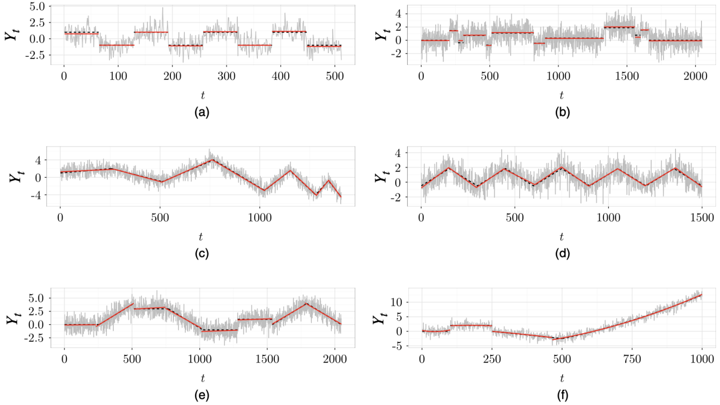


Figure: Simulated data. grey solid line: observed data, black dotted line: true signal, red solid line: fitted model using NOT

Computation time

- T: the number of observations,
- M: the number of random intervals drawn in NOT.

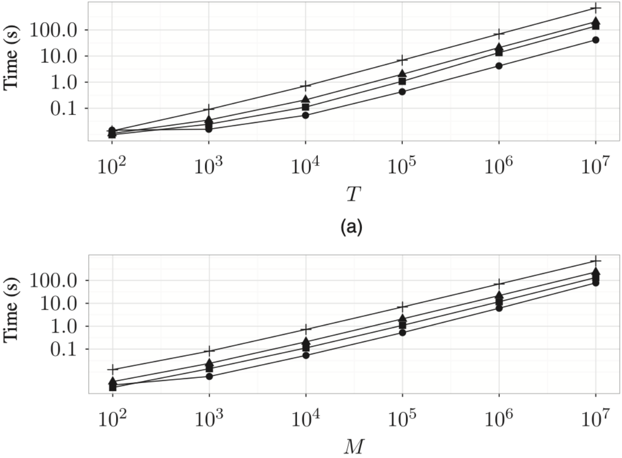


Figure: Execution time to compute the entire solution path $\mathcal{T}(\zeta_T)$

Real data

Goddard Institute for Space Studies surface temperature anomalies data consists of monthly global surface temperature anomalies recorded from January 1880 to June 2016.

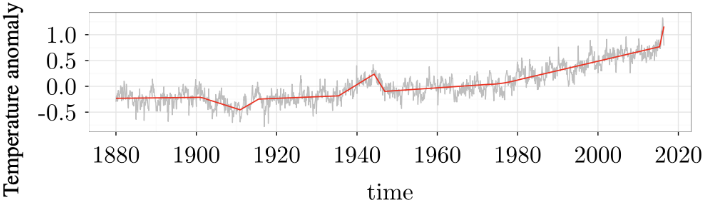


Figure: model: piecewise linear mean with constant variance
grey: observed value, red: fitted model using NOT

Real data 2

UK house price index (HPI) provides an overall estimate of the changes in house prices across the UK.

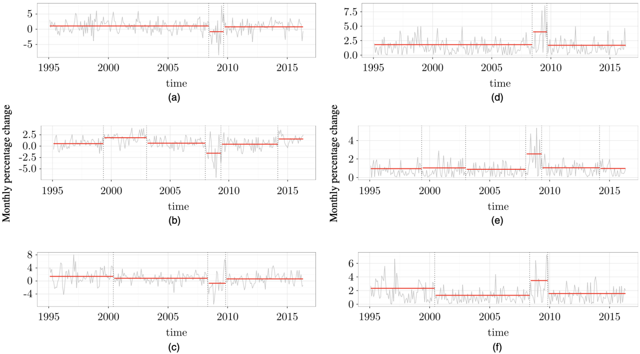


Figure: model: piecewise constant mean and variance
grey: observed value, red: fitted model using NOT

Extra simulation for heavy-tailed distribution

$y_t = \sum_{k=1}^{2} I(t \leq 500k) + \epsilon_t$, $\epsilon_t \sim t(3)$ i.i.d. $t = 1, \ldots, 5000$.

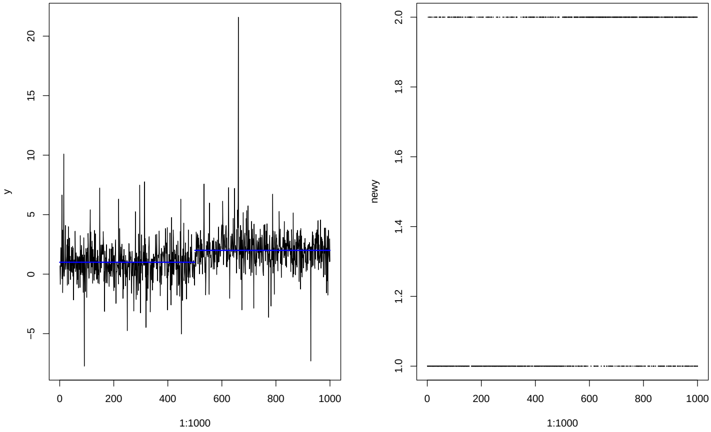


Figure: left: observed data and the true signal, right: heavy-tailed transformation

Extra simulation

$$y_t = \sum_{k=1}^{10} I(t \leq 500k) + \epsilon_t,\ \epsilon_t \sim t(df) \text{ i.i.d. } t = 1, \ldots, 5000.$$

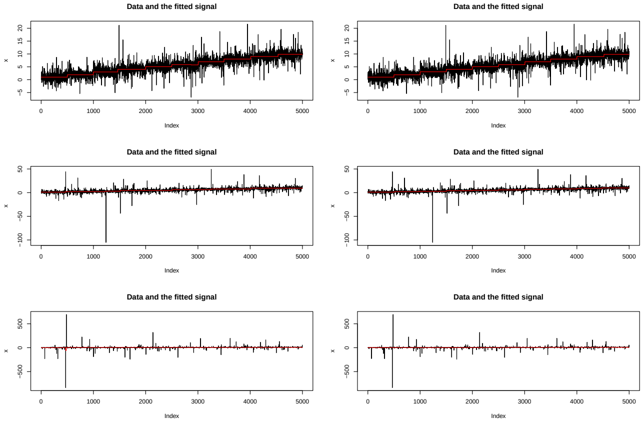


Figure: Top: $df = 3$, middle: $df = 2$, bottom: $df = 1$,
left: NOT, right: NOT for heavy-tailed distribution

Extra simulation

$y_t = \sum_{k=1}^{10} I(t \le 500k) + \epsilon_t$, $\epsilon_t \sim t(df)$ i.i.d. $t = 1, \ldots, 5000$.

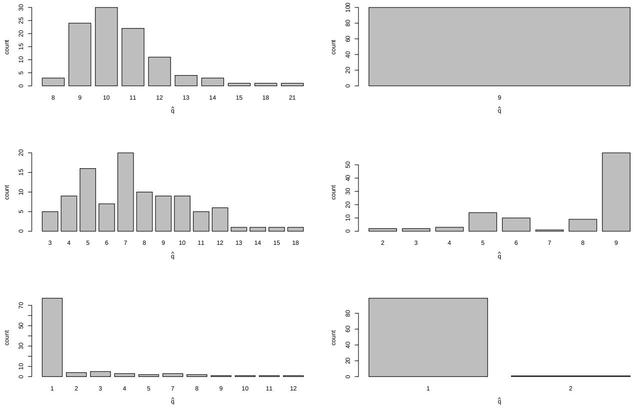


Figure: Estimated number of change points from 100 replications,
Top: $df = 3$, middle: $df = 2$, bottom: $df = 1$,
left: NOT, right: NOT for heavy-tailed distribution

Reference

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

Rafal Baranowski, Yining Chen, and Piotr Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3): 649–672, 2019.