NOT: Narrowest-Over-Threshold Detection of Multiple Change-points and Change-point-like Features

Rafal Baranowski, Yining Chen and Piotr Fryzlewicz
(London School of Economics and Political Science)

ICMS, July 2018

Outline

- Introduction
- Binary Segmentation (BS)
- Narrowest-Over-Threshold (NOT) Detection
- Computational aspects
- Extensions

Introduction

In the univariate setting, consider the model

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

where the unobserved function f_t contains an unknown number of **features** at unknown locations, and ε_t is centered noise.

Examples:

- (canonical) change-point detection (f_t being piecewise constant)
- knot selection in spline smoothing
- trend changes in time series analysis

More broadly, a **feature** can be anything we know how to estimate the location of, if we know that there is only one present inside an interval.

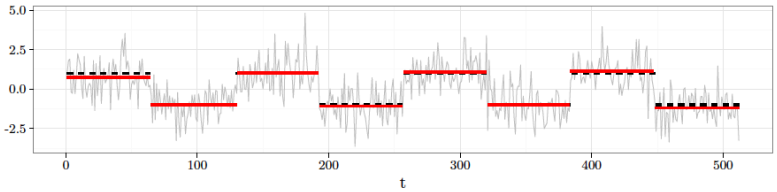
Goals

Objective: estimating the number and locations of these features

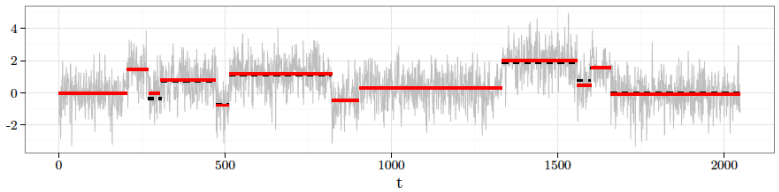
Goals:

- to consistently estimate the number of the features
- to consistently estimate the locations of the features, and ideally at minimax optimal rates (up to an $O(\log T)$ factor worse)
- to be computationally feasible
  (i.e. complexity is at most a logarithmic factor worse than $O(T)$)

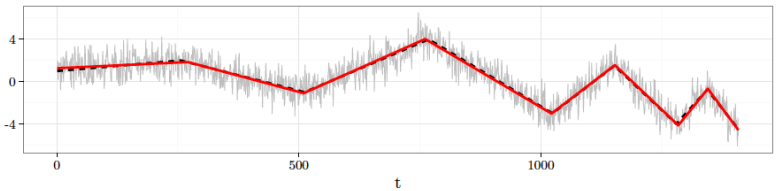
Our aim: a general framework



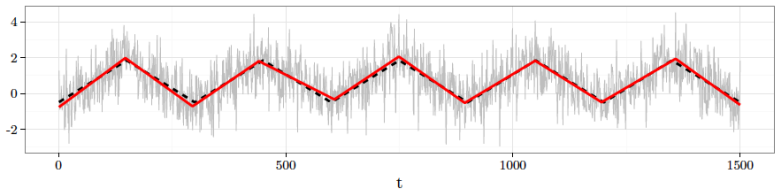
(a) (M1) teeth



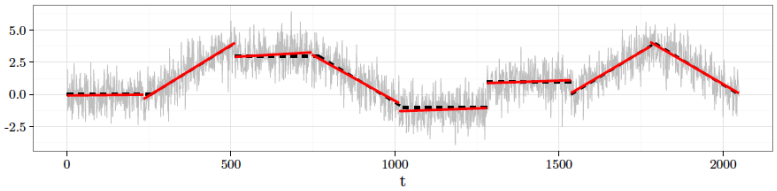
(b) (M2) blocks



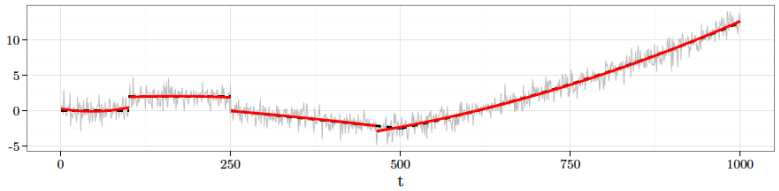
(c) (M3) wave1



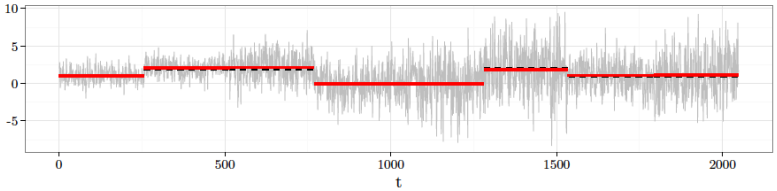
(d) (M4) wave2



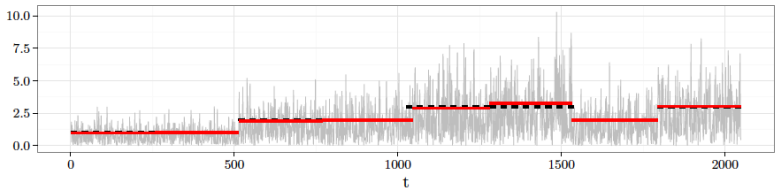
(e) (M5) mix



(f) (M7) quad



(g) (M6) vol: f_t



(h) (M6) vol: σ_t

Single feature detection

Suppose that we know there exists a single feature inside the interval $[s, e]$, then detection could be typically accomplished via (quasi-)log-likelihood-ratio-type statistics, i.e.

1. Find $\bar{f}$, a function with **only one** feature (at different locations from $s+1$ to $e-1$), minimising

$$\sum_{t=s}^{e} \{X_t - \bar{f}_t\}^2.$$

2. Denote the location of the feature of $\bar{f}$ by b.

Examples:

- piecewise constant
- knot of degree 1 (a.k.a. kink)
- piecewise linear
- .....

Single feature detection - piecewise constant

Denote by $\bar{\mathbf{f}}^b$ a step vector with a change-point at index b. We have that

$$\operatorname{argmin}_{s<b<e} \min_{\bar{\mathbf{f}}^b} \sum_{t=s}^{e} \left\{ X_t - \bar{\mathbf{f}}_t^b \right\}^2 \equiv \operatorname{argmax}_b |\langle \mathbf{X}, \psi_{s,e}^b \rangle|$$

where $\mathbf{X} = (X_1, \ldots, X_n)'$ and $\psi_{s,e}^b$ is an "Unbalanced Haar" vector, i.e. a vector which

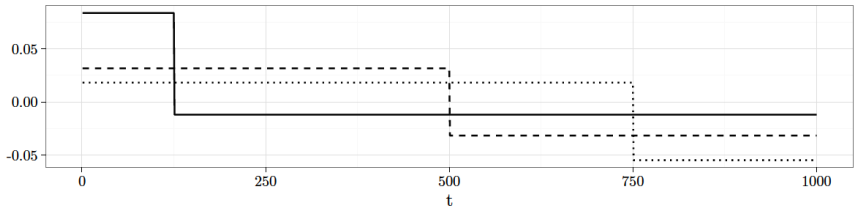
- is constant and positive for $i = s, \ldots, b$,
- is constant and negative for $i = b+1, \ldots, e$,
- sums to zero and sums to one when squared.

Thus, to locate the change-point, it is enough to only inspect the absolute maxima of $\langle \mathbf{X}, \psi_{s,e}^b \rangle$ over b, a.k.a. the CUSUM statistic.

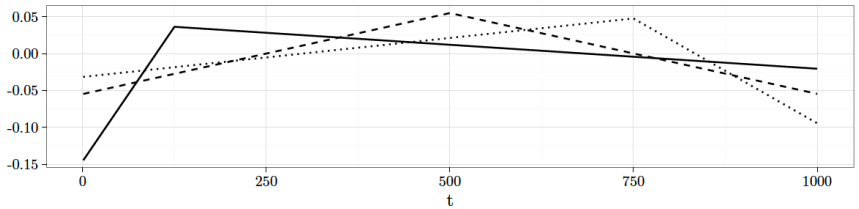
Single feature detection - knot of degree one

Similarly, to locate the kink, it is enough to only inspect the absolute maxima of the new CUSUM-type statistic (which we call **CONTRAST**), $|\langle \mathbf{X}, \phi^b_{s,e} \rangle|$ over b, where $\phi^b_{s,e}$ is a vector which

- is linear for $i = s, \ldots, b$,
- is linear for $i = b, \ldots, e$,
- sums to zero and sums to one when squared.
- $|\langle \gamma, \phi^b_{s,e} \rangle| = 0$ for any linear vector γ.



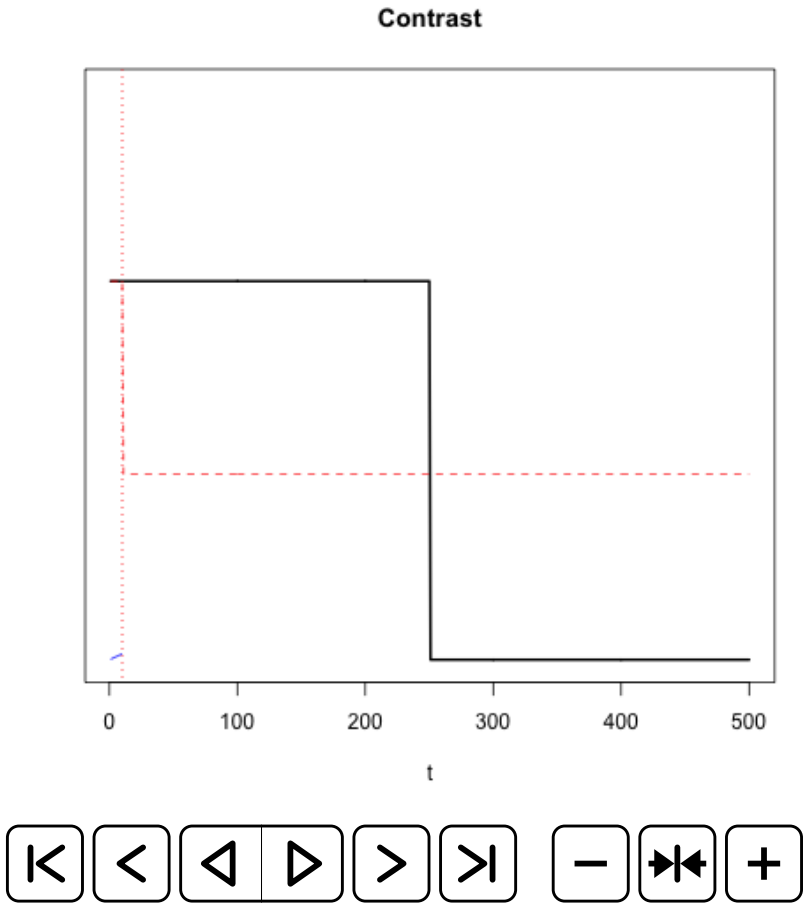
(a) $\psi^b_{s,e}$



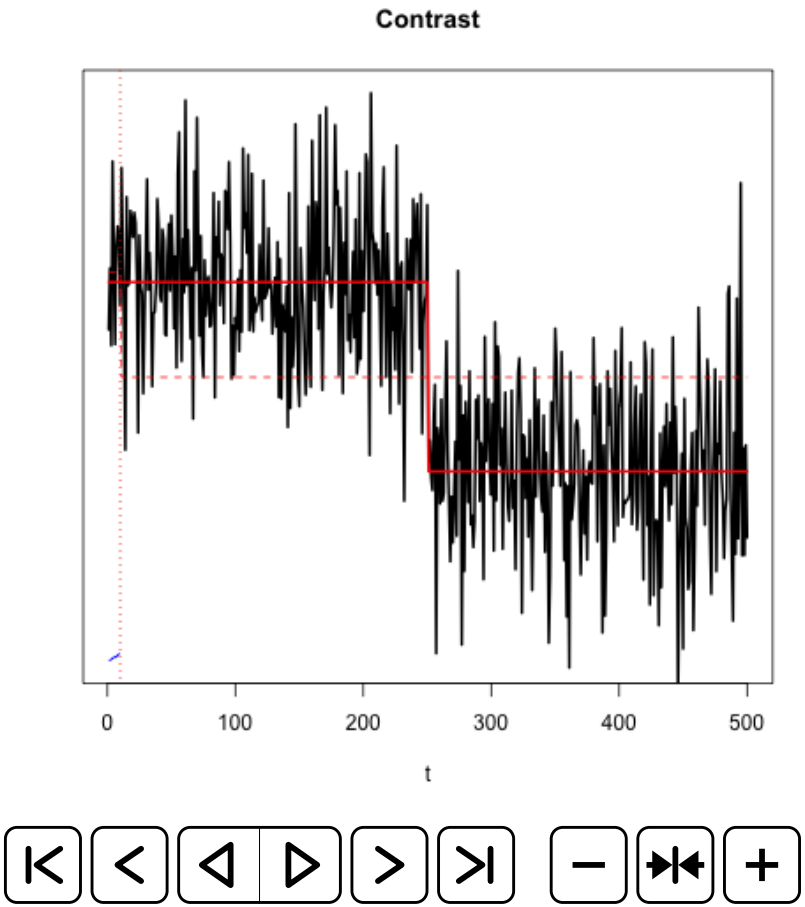
(b) $\phi^b_{s,e}$

Fig. Plots of $\psi^b_{s,e}$ and $\phi^b_{s,e}$ for $s = 1$, $e = 1000$ and several values of b. Solid line: $b = 125$; dashed line: $b = 500$; dotted line: $b = 750$.

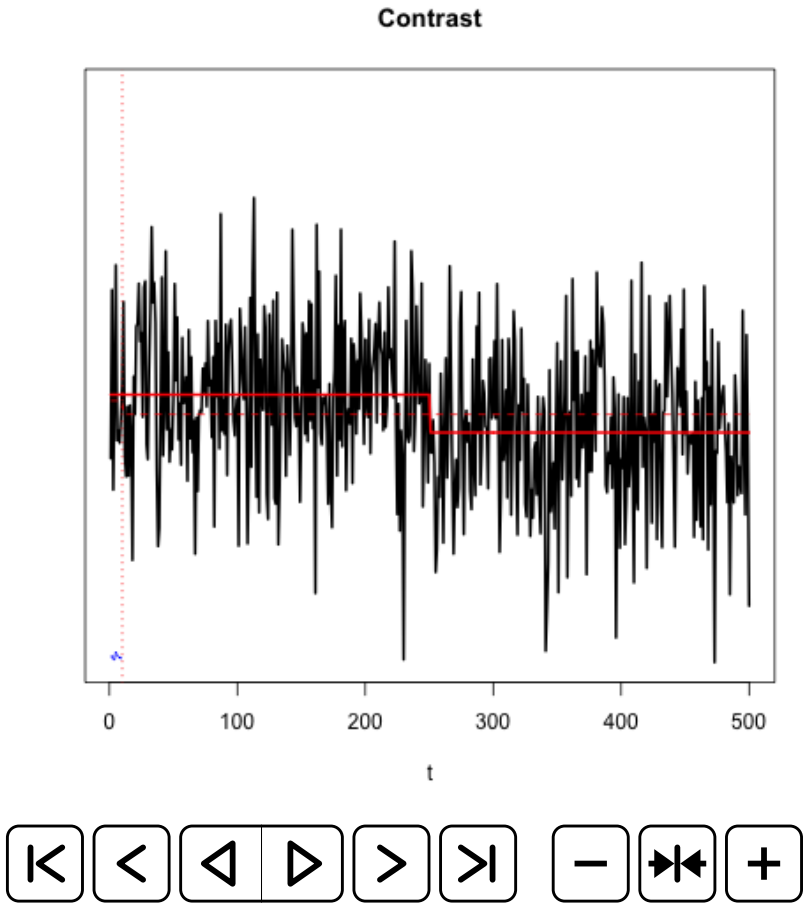
Single change-point detection: a noiseless example



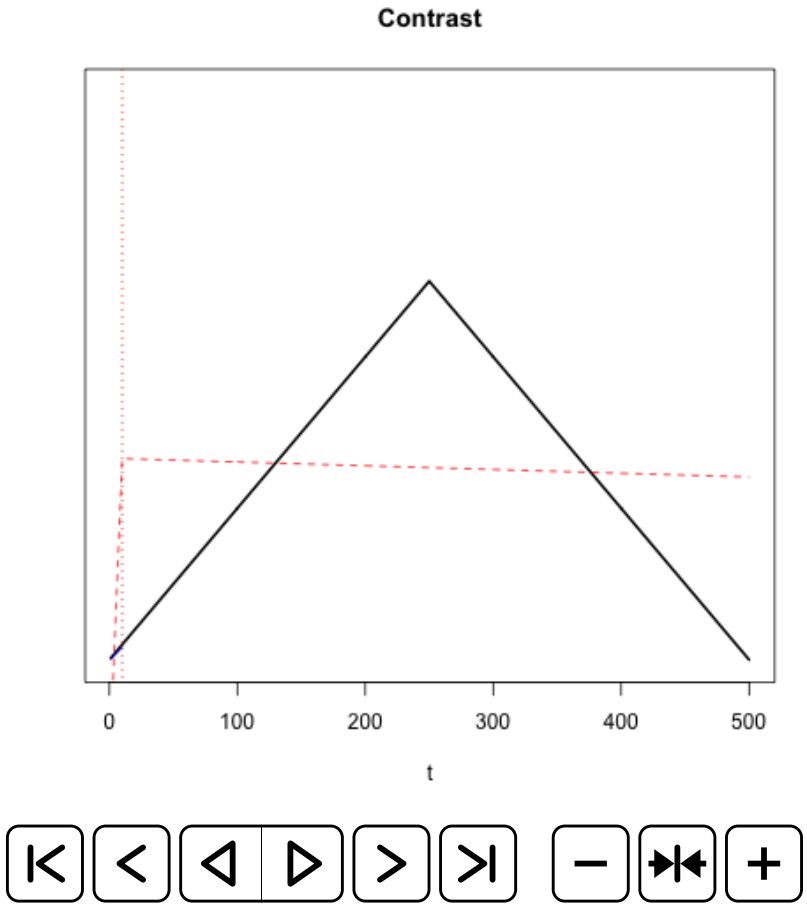
Single change-point detection: the same example with noise



Single change-point detection: with a lot more noise



Single kink detection: a noiseless example



From single feature to multiple features?

Question: how to deal with (unknown number of) multiple features?

Idea: make use of the "binary tree" structure of the problem and solve it via **divide-and-conquer**.

Suppose we **are** able to detect a feature at $b \in \{1, \cdots, T\}$. The problem can then be divided into two sub-problems:

- find multiple features in $\{1, \cdots, b-1\}$.
- find multiple features in $\{b+1, \cdots, T\}$.
- return the locations from the previous two steps together with b.

This approach is particularly popular in the canonical change-point detection literature; we will show that it could be useful for other more complicated problems too.

Change-point detection - overview - I

A substantial number of techniques. A brief (*but by no mean comprehensive*) literature review:

- Least-squares (or generally likelihood-type fit) + AIC or BIC-type penalty: Yao (1988), Yao and Au (1989), Lee (1995), Lavielle (1999, 2005), Lavielle & Moulines (2000), Lebarbier (2005), Pan & Chen (2006), Boysen et al. (2009).
- Minimum Description Length: Davis et al. (2006).
- L1-type penalties: Rinaldo (2009), Lin et al. (2017).
- Binary Segmentation: Vostrikova (1981), Venkatraman (1992), Bai (1997), Chen et al. (2011), Cho & Fryzlewicz (2012, 2013).

Change-point detection - overview - II

Some more comments:

- Least-squares (or generally likelihood-type fit) + AIC or BIC-type penalty: potentially slow computational speed, typically of order $O(T^2)$. However some serious efforts to reduce this, e.g. Rigail (2010) and Killick et al. (2012) (a.k.a. PELT, or pruned exact linear time)
- MDL: minimisation could be quite involved, via a genetic algorithm in Davis et al. (2006).
- L1-type penalties: not necessarily optimal for change-point detection, see Brodsky & Darkhovsky (1993). Often lead to spurious detections.

Binary Segmentation (BS)

Generic algorithm of BS, using canonical change-point detection as an example:

function $\mathrm{BS}(s,\, e,\, \zeta_T)$
  if $e - s \leq 1$ **then**
    STOP
  else
    Pick $b_0 \in \arg\max_{b \in \{s,\ldots,e-1\}} |\langle \mathbf{X}, \psi^b_{s,e} \rangle|$
    if $|\langle \mathbf{X}, \psi^{b_0}_{s,e} \rangle| > \zeta_T$ **then**
      Add b_0 to the index set of estimated features $\mathcal{S}$
      $\mathrm{BS}(s,\, b_0,\, \zeta_T)$
      $\mathrm{BS}(b_0 + 1,\, e,\, \zeta_T)$
    else
      STOP
    end if
  end if
end function
$\mathcal{S} = \emptyset$; $\mathrm{BS}(1,\, T,\, \zeta_T)$

BS – handle multiple change-points?

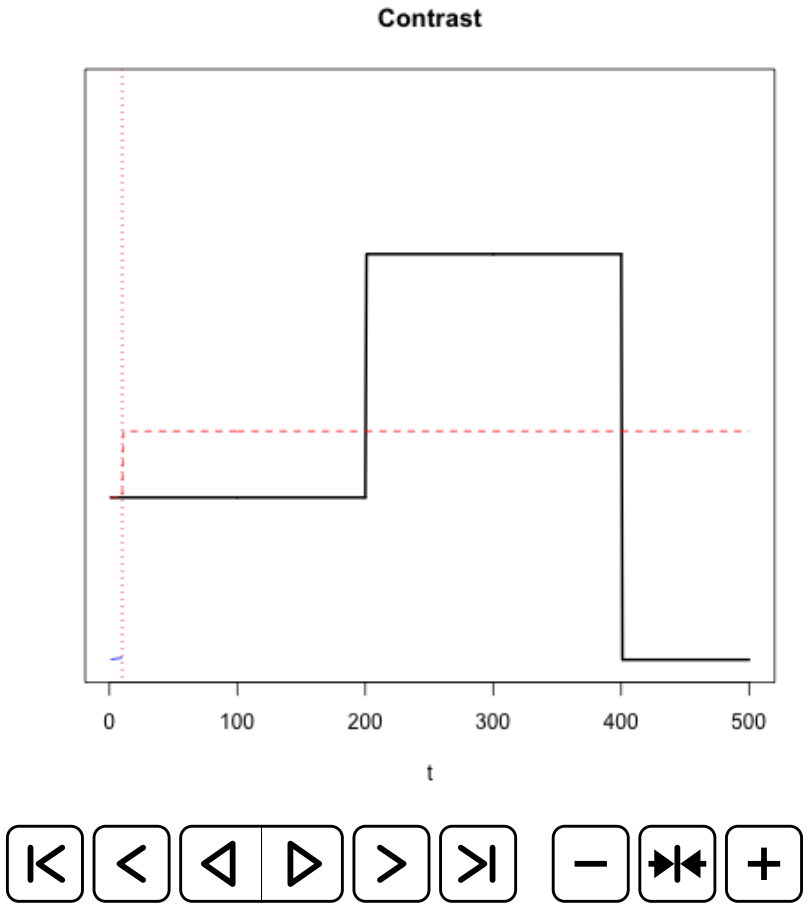
In principle, BS is fast (typically $O(T \log T)$), conceptually simple, easy to code, and tractable theoretically.

Since BS fits a one-step function to the current interval $\{s, \ldots, e\}$, we can expect the performance to be good if $\{s, \ldots, e\}$ contains no more than one change-point.

If the current interval $\{s, \ldots, e\}$ contains more than one change-point, things are still okay in the canonical setting (Venkatraman, 1992). Consider the noiseless case where $\mathbf{f} = (f_1, \ldots, f_T)'$:

- even if there are multiple change-points in index from s to e, $\operatorname{argmax}_b |\langle \mathbf{f}, \psi^b_{s,e} \rangle|$ must belong to the set that contains all change-points of $\mathbf{f}$ from index s to e.

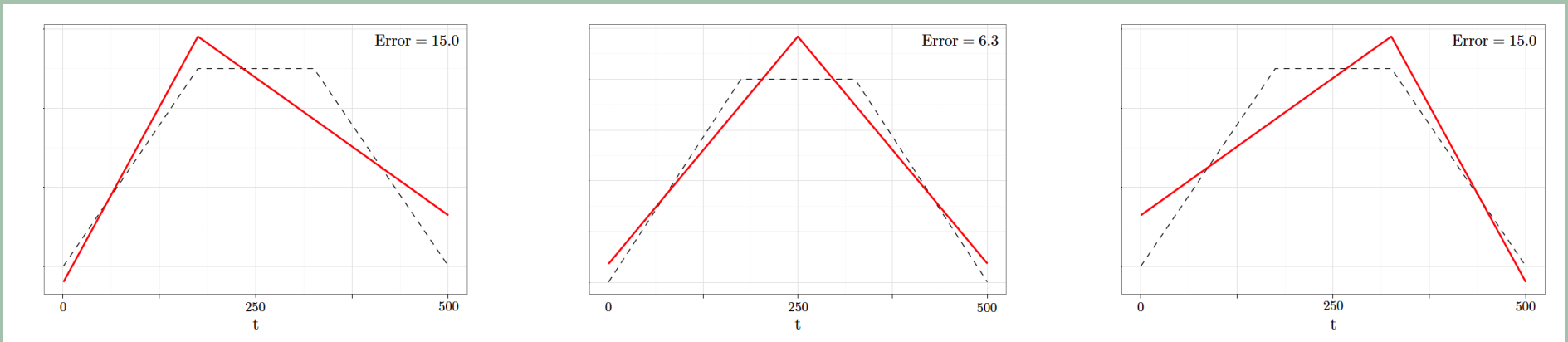
BS – handle multiple change-points: a noiseless example



Note: we are *lucky* here, because this property does not hold in general.

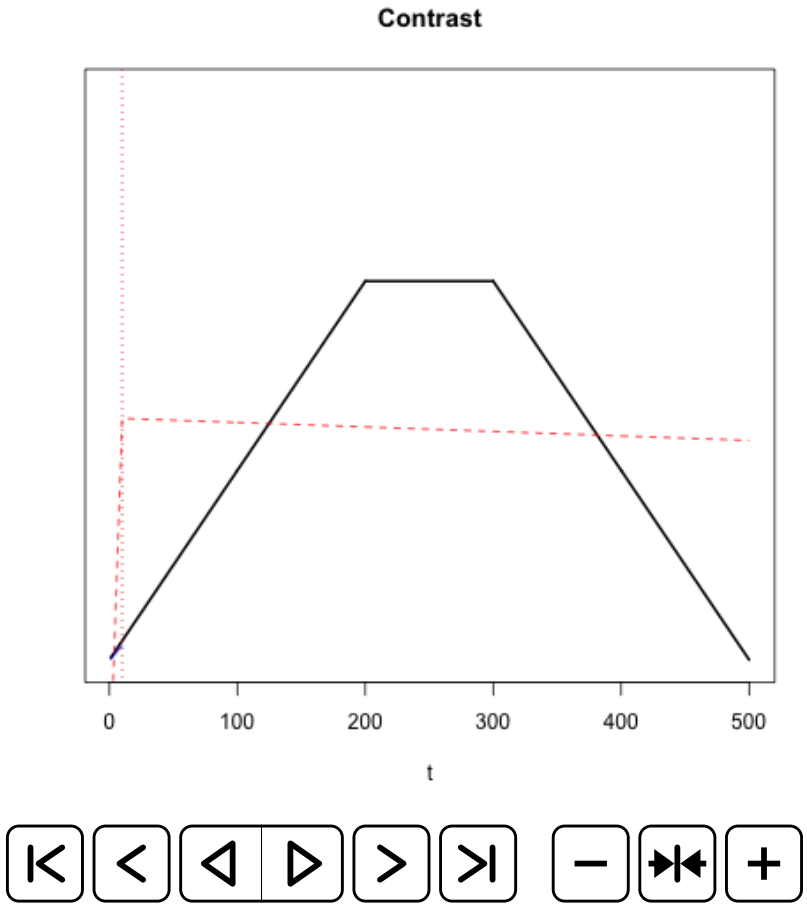
BS fails to detect certain features - a noiseless example

Observation: if the current interval contains two or more features (of ever-so-slightly more complicated nature), it may happen that the best approximation by one feature will not indicate any of them:



Best ℓ_2 approximation of the true signal (dashed) via a triangular signal with a single feature.

BS fails to detect certain features - a noiseless example



Narrowest-over-threshold (NOT)

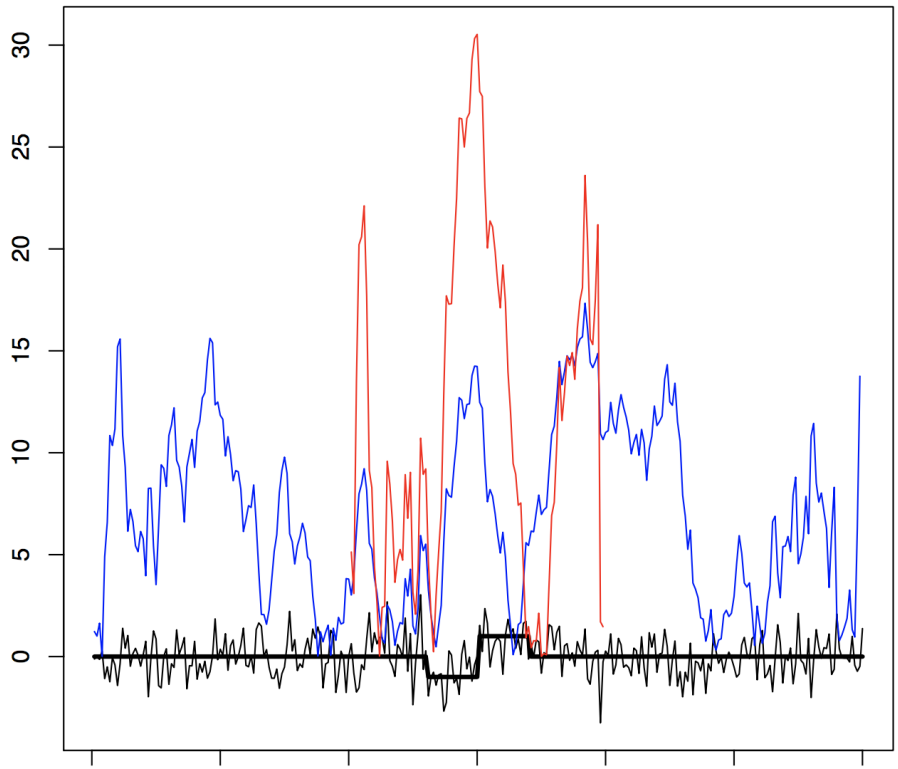
Aims:

- we want to deal with intervals with **only one** feature;
- the location of the true feature in any choosen interval is sufficiently far away from the two ending points.

One possible solution:

1. randomly pick the starting and ending points of the intervals, s and e, uniformly with replacement over $\{1, \ldots, T\}$, a suitable number of times (often $O(\log T)$ is sufficient); See also Fryzlewicz (2014);
2. only keep the intervals with the value of the summary statistic **over the threshold**, e.g. $\max_{s<b<e} \mathrm{CONTRAST}^b_{s,e} > \zeta_T$;
3. then concentrate on the one with the **narrowest** width.

Narrowest-over-threshold (NOT) - intuitions



Example of global (blue) and local (red) $|\langle \mathbf{X}, \psi^{s,b,e} \rangle|$ as a function of b, on data $\mathbf{X}$ in black.

Narrowest-over-threshold (NOT) - intuitions

1. randomly pick the starting and ending points of the intervals, s and e, uniformly over $\{1, \ldots, T\}$ a suitable number of times;
2. keep those intervals with the value of the statistic **over the threshold**;
3. then concentrate on the one with the **narrowest** length, $e - s$.

Intuitions:

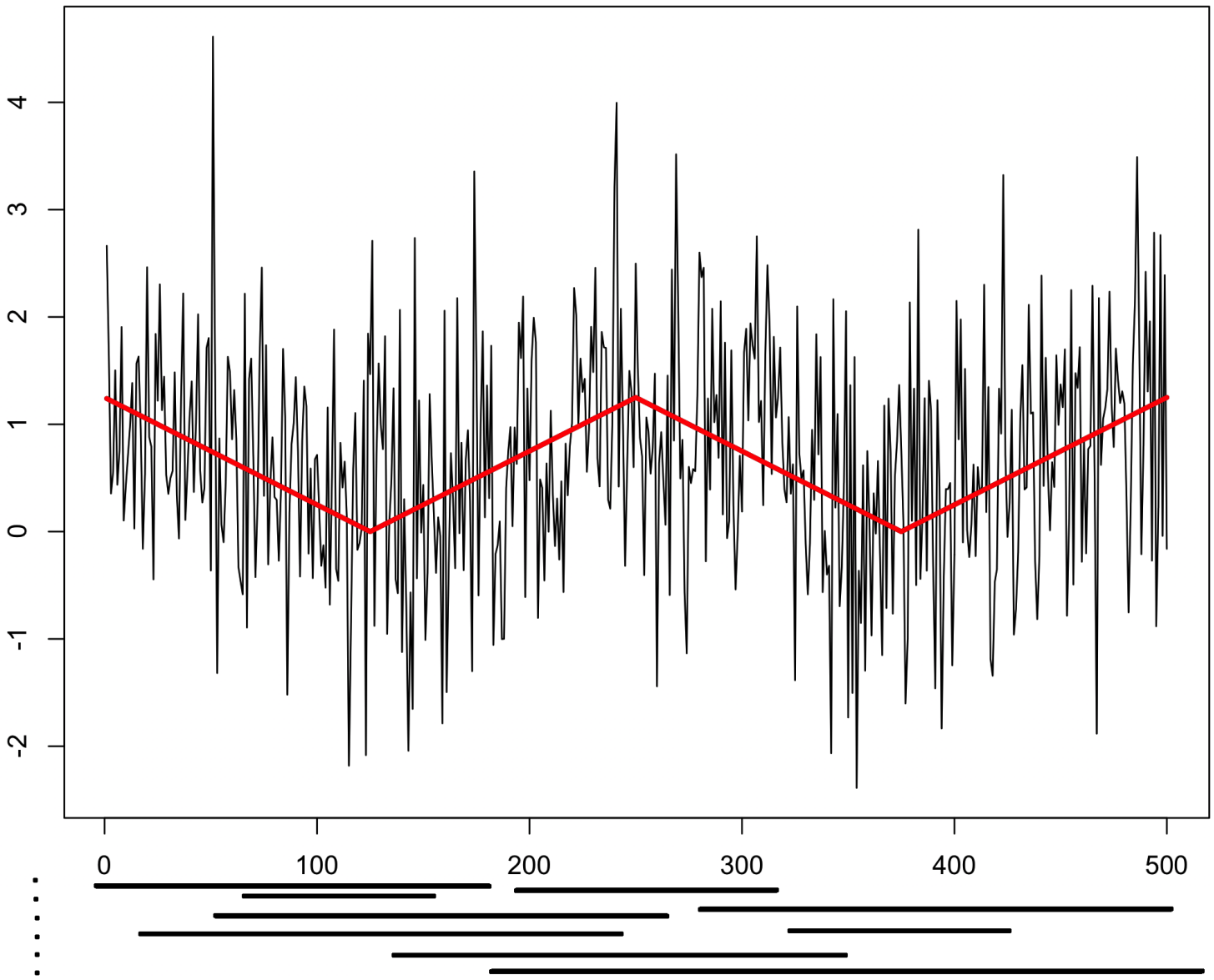
1. better mixture of subintervals that represents both local and global properties;
2. to make sure that the intervals has *at least* one feature;
3. to make sure that the intervals has *at most* one feature.

NOT - generic algorithm

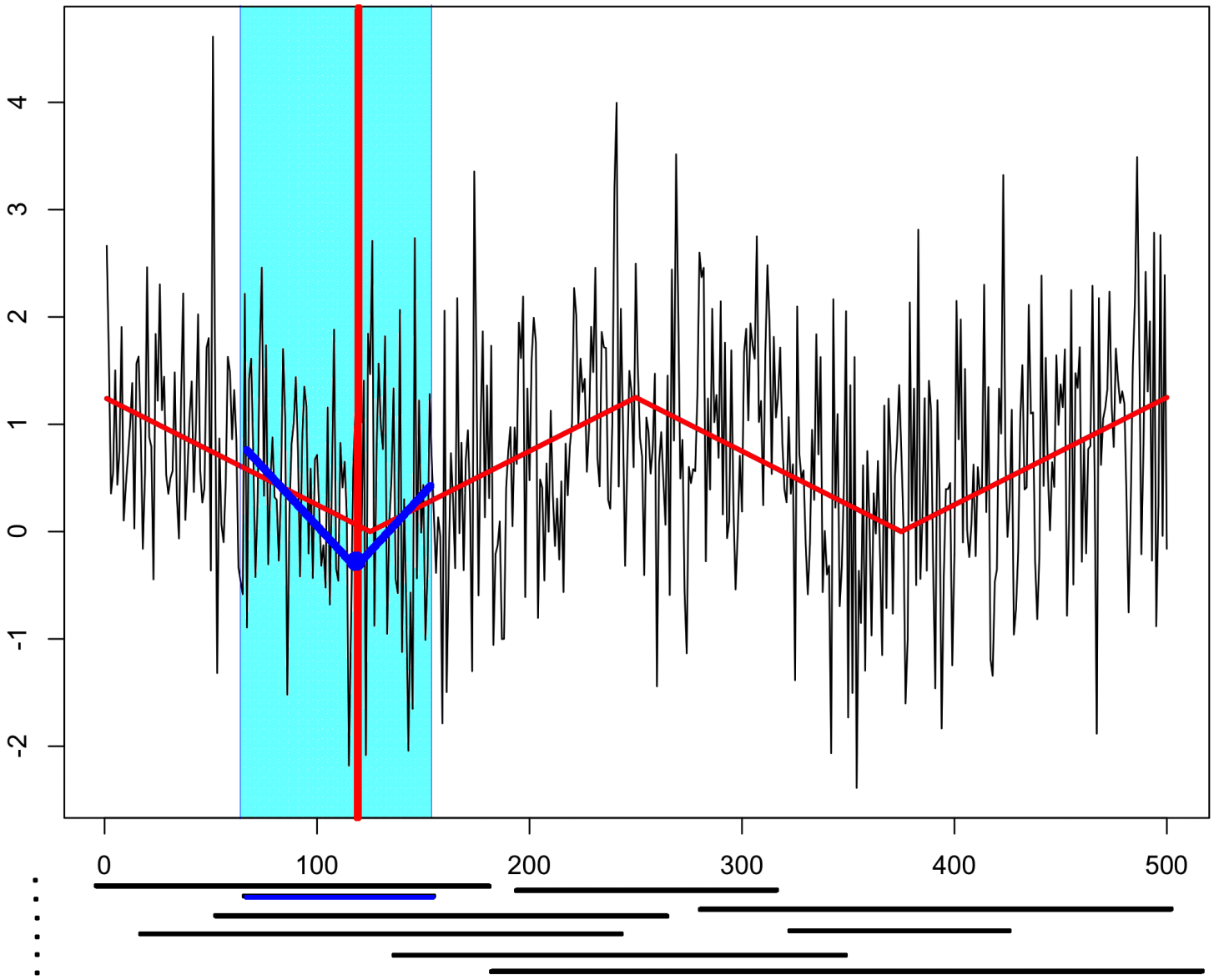
Given a data vector $\mathbf{X} = (X_1, \ldots, X_T)'$, F_T^M is a set of M intervals, with start- and end-points drawn independently and uniformly from $\{1, \ldots, T\}$, $\mathcal{S} = \emptyset$.

procedure $\mathrm{NOT}(s, e, \zeta_T)$
  if $e - s < 1$ **then** STOP
  else
    $\mathcal{M}_{s,e} := \{m : [s_m, e_m] \in F_T^M, [s_m, e_m] \subset [s, e]\}$
    if $\mathcal{M}_{s,e} = \emptyset$ **then** STOP
    else
      $\mathcal{O}_{s,e} := \{m \in \mathcal{M}_{s,e} : \max_{s_m \le b \le e_m} \mathrm{CONTRAST}_{s_m, e_m}^b(\mathbf{X}) > \zeta_T\}$
      if $\mathcal{O}_{s,e} = \emptyset$ **then** STOP
      else
        $m^* :\in \arg\min_{m \in \mathcal{O}_{s,e}} |e_m - s_m|$
        $b^* := \arg\max_{s_{m^*} \le b \le e_{m^*}} \mathrm{CONTRAST}_{s_{m^*}, e_{m^*}}^b(\mathbf{X})$
        $\mathcal{S} := \mathcal{S} \cup \{b^*\}$
        $\mathrm{NOT}(s, b^*, \zeta_T)$
        $\mathrm{NOT}(b^* + 1, e, \zeta_T)$
      end if
    end if
  end if
end procedure

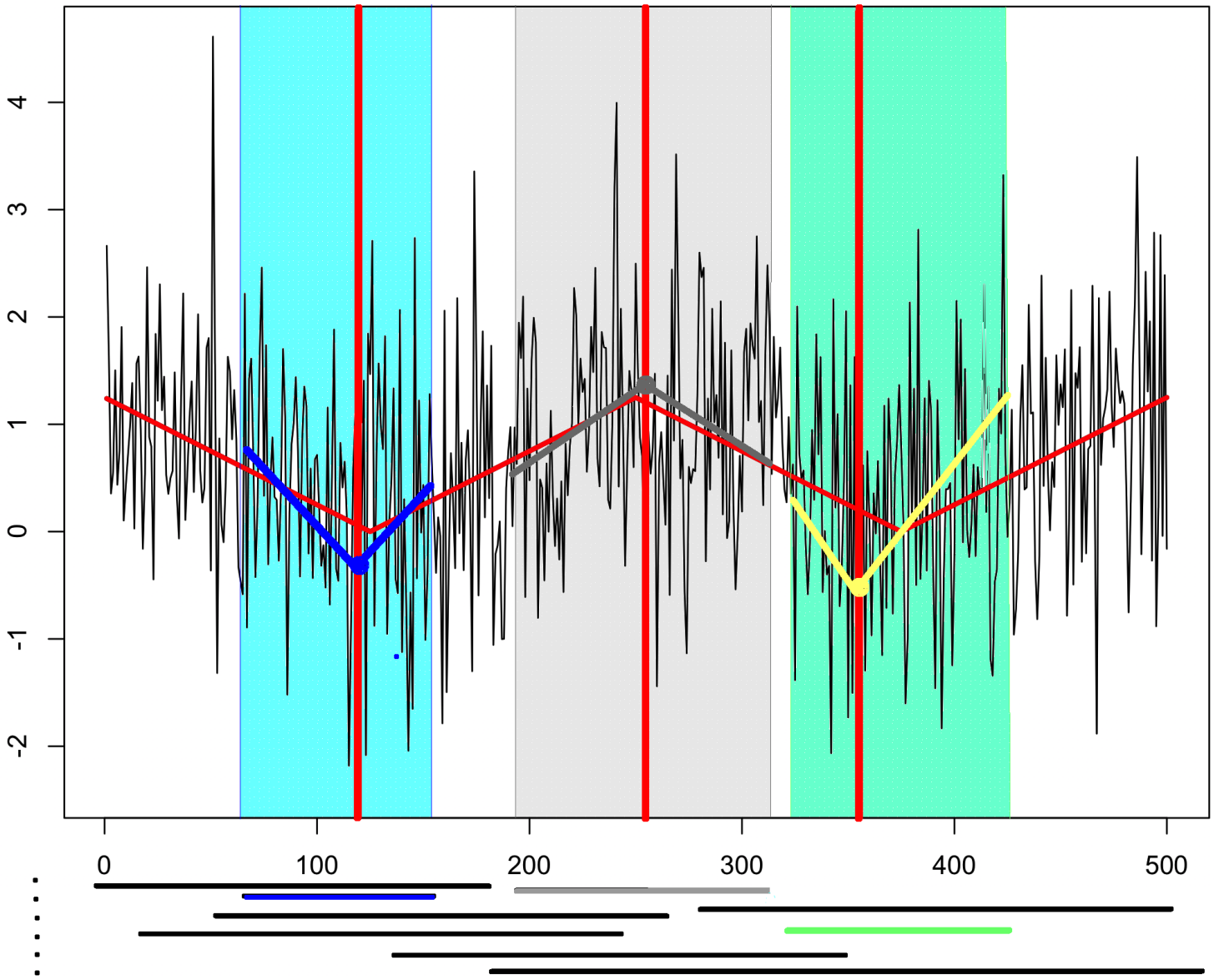
NOT – toy example demonstration



NOT – toy example demonstration



NOT – toy example demonstration



Choosing ζ_T via strengthened SIC

Note that each threshold $\zeta_T^{(k)}$ is associated with a fitted model $\mathcal{M}_k$.

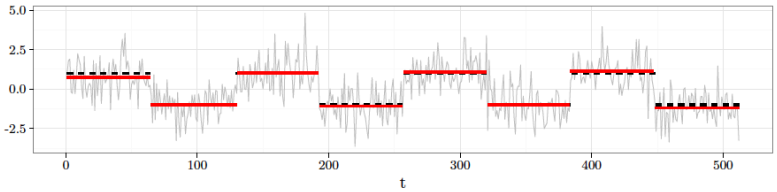
1. We first perform NOT on all possible thresholds on $(0, \infty)$, getting a series of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ along the **solution path**.
2. We then select the k_*-th model such that

$$k_* = \text{argmin}_k \Big\{ -2\text{loglik}(\mathcal{M}_k) + D_k \log^\alpha n \Big\}$$

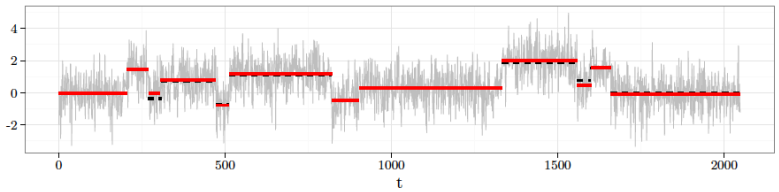
where D_k is the degree of freedom of the corresponding model $\mathcal{M}_k$, and $\alpha > 1$. We call it *NOT with sSIC*.

NOT solution path can also be viewed as an efficient way of reducing the number of candidate models.

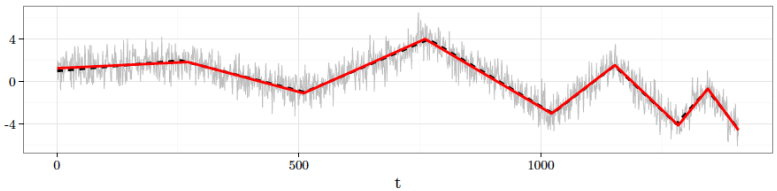
General framework that works in many scenarios



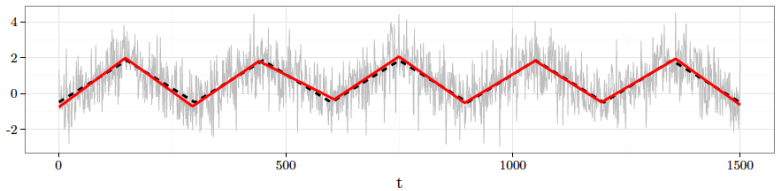
(a) (M1) teeth



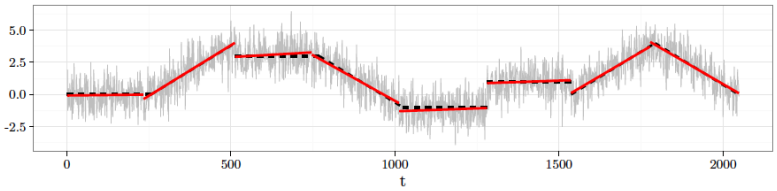
(b) (M2) blocks



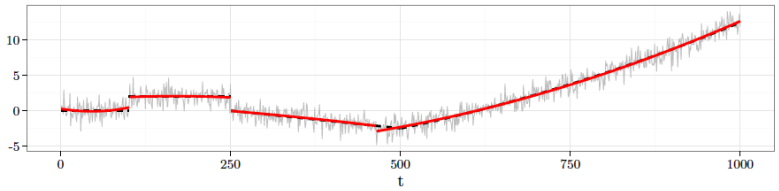
(c) (M3) wave1



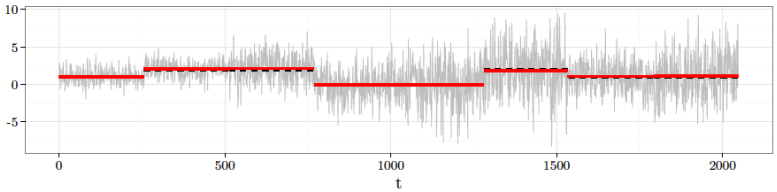
(d) (M4) wave2



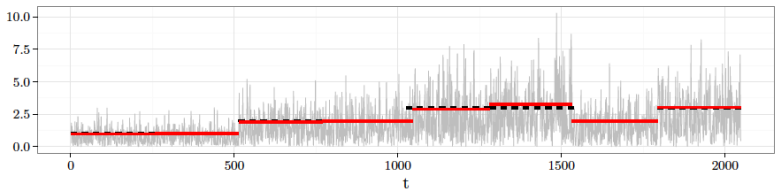
(e) (M5) mix



(f) (M7) quad



(g) (M6) vol: f_t



(h) (M6) vol: σ_t

Some theory - jumps

Theorem (Consistency and Convergence rates)

Suppose the true jumps are at $\tau_1, \ldots, \tau_q$ *(with the convention of* $\tau = 0$ *and* $\tau_{q+1} = T$*) where* q *is fixed, and* $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$*. Let* $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$*,* $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$ *and* $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^{\mathbf{f}}$*. Furthermore,* $\delta_T / T \geq \underline{C}_1$*,* $\underline{f}_T \geq \underline{C}_2$ *and* $\max_{i=1,\ldots,T} |f_i| \leq \bar{C}$ *for some* $\underline{C}_1, \underline{C}_2, \bar{C} > 0$*.*

Let $\hat{q}$ *and* $\hat{\tau}_1, \ldots, \hat{\tau}_q$ *denote, respectively, the number and locations of change-points, sorted in increasing order, estimated by NOT with sSIC with* $\alpha > 1$*. Then there exist constants* C *such that given* $M \geq 36 \underline{C}_1^{-2} \log(\underline{C}_1^{-1} T)$*, as* $T \to \infty$*,*

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

Some theory - jumps - more details

About the key assumptions:

- q is fixed
- Gaussian noise
- Spacing between consecutive jumps: $\frac{\min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})}{T} \geq \underline{C}_1$
- Size of the jumps: $\min_{j=1,\ldots,q+1} |f_{\tau_j+1} - f_{\tau_j}| \geq \underline{C}_2$

About the convergence rate:

- Optimal rate: $O_p(1)$
- NOT with sSIC: $O_p(\log T)$

Some theory - kinks

Theorem (Consistency and Convergence rates)

Suppose the true kinks are at $\tau_1, \ldots, \tau_q$ *where* q *is fixed, and* $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$. *Let* $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j - 1} - f_{\tau_j + 1}|$, $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^{\mathbf{f}}$. *Furthermore, assume that* $\delta_T / T \geq \underline{C}_1$, $\underline{f}_T T \geq \underline{C}_2$ *and* $\max_{i=1,\ldots,T} |f_i| \leq \bar{C}$ *for some* $\underline{C}_1, \underline{C}_2, \bar{C} > 0$.

Let $\hat{q}$ *and* $\hat{\tau}_1, \ldots, \hat{\tau}_q$ *denote, respectively, the number and locations of features, sorted in increasing order, estimated by sSIC using* $\alpha > 1$. *Then there exist constants* C *such that given* $M \geq 36 \underline{C}_1^{-2} \log(\underline{C}_1^{-1} T)$, *as* $T \to \infty$,

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

Some theory - kinks - more details

About the key assumptions:

- q is fixed
- Gaussian noise
- Spacing between consecutive kinks: $\frac{\min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})}{T} \geq \underline{C}_1$
- Size of the change in slope: $\min_{j=1,\ldots,q+1} |f_{\tau_j+1} + f_{\tau_j-1} - 2f_{\tau_j}| T \geq \underline{C}_2$

About the convergence rate:

- Optimal rate: $O_p(\sqrt{T})$
- NOT with sSIC: $O_p(\sqrt{T \log T})$

Computational complexity - I

Making use of the recurrence relationship of CONTRAST statistics over b, we could compute

$$\langle \mathbf{X}, \psi_{s,e}^{s+1} \rangle, \ldots, \langle \mathbf{X}, \psi_{s,e}^{e-1} \rangle$$

or

$$\langle \mathbf{X}, \phi_{s,e}^{s+1} \rangle, \ldots, \langle \mathbf{X}, \phi_{s,e}^{e-1} \rangle$$

at the cost of $O(s - e)$.

- If we take $M = O(\log T)$ intervals, we could deal with all of them in $O(T \log T)$.
- Moreover, the cost of constructing the entire solution path (with respect to all possible thresholds) is at most $O(M^3)$ (much faster in practice).
  - At most M different models on the solution path.
  - The binary tree corresponding to each model has at most depth M.
  - Construction of the tree at each depth level costs at most $O(M)$.
- Therefore, the complexity NOT (with the entire solution path) is $O(T \log T)$.

Software: R package **not**

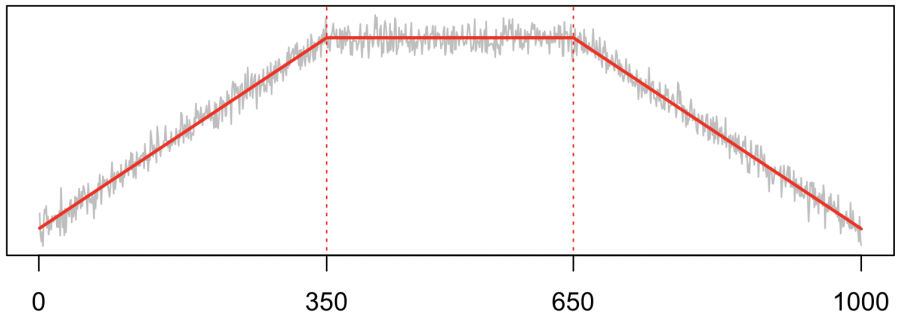
Computational complexity - II - the solution path algorithm

To see why the entire solution path can be computed in much faster manner in practice:

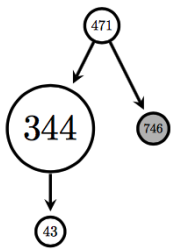
Observations:
Suppose that we have already constructed the binary tree for the model at threshold $\zeta_T^{(k)}$.

- The next thresold $\zeta_T^{(k+1)}$ on the path (with $> \zeta_T^{(k)}$) could be computed by going through all leaves of the previous model tree.
- The new model at $\zeta_T^{(k+1)}$ should be *typically* quite similar to the old one.

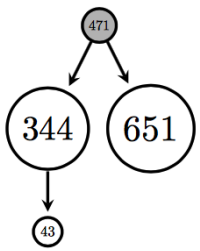
Computational complexity - III - toy example



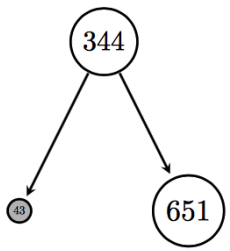
s	e	$e-s+1$	$\text{argmax}_{s\le b\le e} \mathcal{C}^b_{s,e}(\mathbf{Y})$	$\max_{s\le b\le e} \mathcal{C}^b_{s,e}(\mathbf{Y})$
1	1000	1000	490	10.19
10	245	236	43	0.08
225	450	226	344	0.76
500	750	251	651	0.83
740	950	211	746	0.03
450	550	101	471	0.07



(a) $\zeta_T^{(1)} = 0$



(b) $\zeta_T^{(2)} = 0.03$



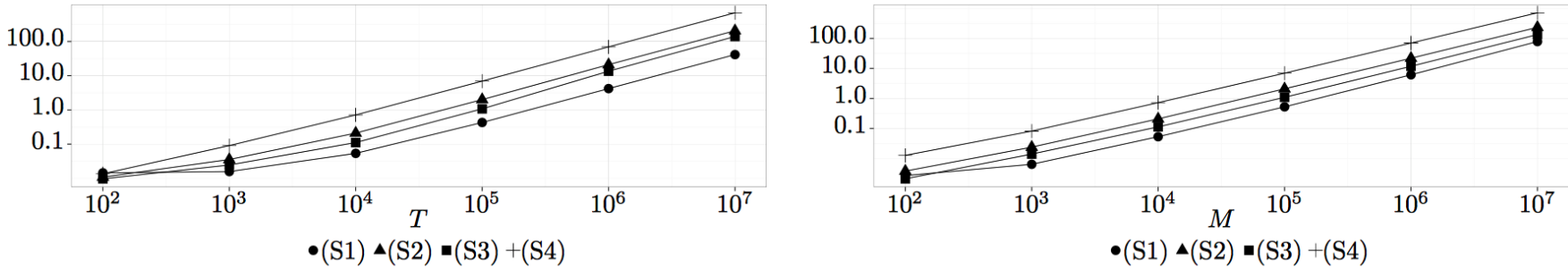
(c) $\zeta_T^{(3)} = 0.07$

(d) $\zeta_T^{(4)} = 0.08$

Computational complexity - IV - empirical evidence

Computational complexity empirically: like $O(MT)$

- (S1): jumps in f_t
- (S2): kinks in f_t
- (S3): piecewise linear in f_t
- (S4): jumps in both f_t and σ_t



(a) fixed $M = 10000$ (b) fixed $T = 10000$

Fig. Execution times (in seconds)

Other extension: unknown (global) degree of polynomials

For possible degrees of $0, 1, 2, \ldots, K$,

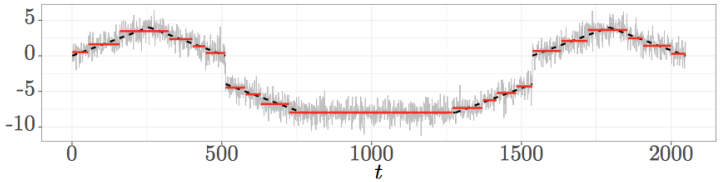
1. Denote the sSIC scores corresponding to the estimates from $\mathrm{NOT}_0, \ldots, \mathrm{NOT}_K$ by sSIC(NOT_0) ,..., sSIC(NOT_K) respectively.
2. Pick the estimator produced by NOT_{i^*} with

$$i^* = \mathrm{argmin}_{i \in \{0,\ldots,K\}} \mathrm{sSIC}(\mathrm{NOT}_i).$$

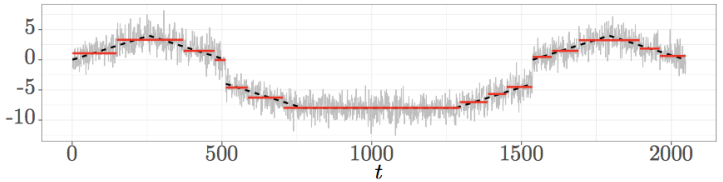
Model selection consistency can also be proved.

How about allowing the degree of polynomials to change locally...

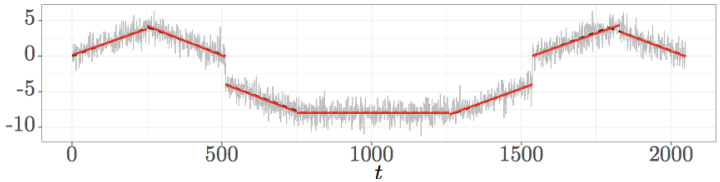
Other extension: an example



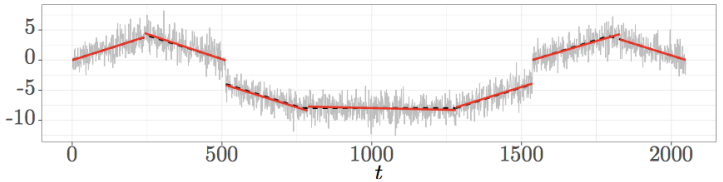
(a) NOT_0, $\mathcal{N}(0,1)$ noise



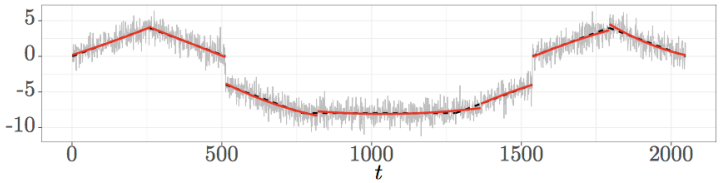
(b) NOT_0, $\mathcal{N}(0,2)$ noise



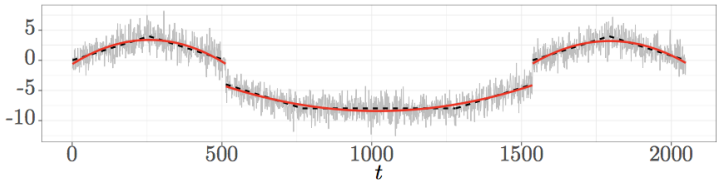
(c) NOT_1, $\mathcal{N}(0,1)$ noise



(d) NOT_1, $\mathcal{N}(0,2)$ noise



(e) NOT_2, $\mathcal{N}(0,1)$ noise



(f) NOT_2, $\mathcal{N}(0,2)$ noise

Noise	Method	$\hat{q}-q$ ≤ -3	-2	-1	0	1	2	≥ 3	MSE	Number of times selected by sSIC
$\mathcal{N}(0,1)$	NOT_0	0	0	0	0	0	0	100	0.120	0
	NOT_1	0	0	0	**99**	1	0	0	**0.015**	**100**
	NOT_2	0	4	18	78	0	0	0	0.024	0
$\mathcal{N}(0,2)$	NOT_0	0	0	0	0	0	0	100	0.188	0
	NOT_1	0	0	0	**100**	0	0	0	**0.032**	**94**
	NOT_2	57	23	14	6	0	0	0	0.078	6

Summary

- NOT solution path can be viewed as a fast device of reducing the number of potential candidate models to a manageable level, before a proper model selection step (e.g. via sSIC) is performed.
- NOT could be applied to a varity of multiple change-point and change-point-like feature detection problems.
- NOT typically offers near-optimal detection rates with feasible computational costs.
- Please try our R package **not**.