Online supplementary materials for “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

This document contains the following parts:

A. Simulation models

B. More details on the contrast functions and their construction

C. More details on the computational aspects of NOT and its solution path

D. Further extension of NOT: noise with slow-varying variance

E. Additional simulation results

F. Additional numerical experiments on the choice of M

G. More on model misspecification and model selection

H. Additional real data example: oil price

I. Proofs

Address for correspondence: Yining Chen, Department of Statistics, Columbia House, Houghton Street, London, WC2A 2AE, U.K.
E-mail: y.chen101@lse.ac.uk

A. Simulation models

(M1) `teeth`: piecewise-constant f_t (in Scenario (S1)), $T = 512$, $q = 7$ change-points at $\tau = 64, 128, \ldots, 448$, with the corresponding jump sizes $-2, 2, -2, \ldots, -2$, starting intercept $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \ldots, T$.

(M2) `blocks`: piecewise-constant f_t (in Scenario (S1)), $T = 2024$, $q = 11$ change-points at $\tau = 205, 267, 308, 472, 512, 820, 902, 1332, 1557, 1598, 1659$, with the corresponding jump sizes $1.464, -1.830, 1.098, -1.464, 1.830, -1.537, 0.768, 1.574, -1.135, 0.769, -1.537$, starting intercept $f_1 = 0$, $\sigma_t = 1$ for $t = 1, \ldots, T$. This signal is widely analysed in the literature, see e.g. Fryzlewicz (2014).

(M3) `wave1`: piecewise-linear f_t without jumps in the intercept (in Scenario (S2)), $T = 1408$, $q = 7$ change-points at $\tau = 256, 512, 768, 1024, 1152, 1280, 1344$, with the corresponding changes in slopes $1 \cdot 2^{-6}, -2 \cdot 2^{-6}, 3 \cdot 2^{-6} \ldots, -7 \cdot 2^{-6}$, starting intercept $f_1 = 1$ and slope $f_2 - f_1 = 2^{-8}$, $\sigma_t = 1$ for $t = 1, \ldots, T$.

(M4) `wave2`: piecewise-linear f_t without jumps in the intercept (in Scenario (S2)), $T = 1500$, $q = 9$ change-points at $\tau = 150, 300, \ldots, 1350$, with the corresponding changes in slopes $2^{-5}, -2^{-5}, 2^{-5}, \ldots, -2^{-5}$, starting intercept $f_1 = 2^{-1}$ and slope $f_2 - f_1 = 2^{-6}$, $\sigma_t = 1$ for $t = 1, \ldots, T$.

(M5) `mix`: piecewise-linear f_t with possible jumps at change-points (in Scenario (S3)), length $T = 2048$, $q = 7$ change-points at $\tau = 256, 512, \ldots, 1792$, with the corresponding sizes of jump $0, -1, 0, 0, 2, -1, 0$ and changes in the slope $2^{-6}, -2^{-6}, -2^{-6}, 2^{-6}, 0, 2^{-6}, -2^{-5}$, starting value for the intercept $f_1 = 0$ and slope $f_2 - f_1 = 0$, $\sigma_t = 1$ for $t = 1, \ldots, T$.

(M6) `vol`: piecewise-constant f_t and σ_t (in Scenario (S4)), $T = 2048$, $q = 7$ change-points at $\tau = 256, 512, \ldots, 1792$ with the corresponding jumps in f_t and σ_t being $1, 0, -2, 0, 2, -1, 0$ and $0, 1, 0, 1, 0, -1, 1$, respectively, initial values $f_1 = \sigma_1 = 1$.

(M7) `quad`: piecewise-quadratic f_t (in Scenario (S5)), $T = 1000$, $q = 3$ change-points at $\tau = 100, 250, 500$, with the corresponding changes in the intercept $2, -2, 0$, in the slope $0, -10^{-1}, 10^{-1}$ and in the quadratic coefficient $0, 0, 2 \times 10^{-5}$, the initial values $f_1 = f_2 - f_1 = f_3 - 2f_2 + f_1 = 0$, $\sigma_t = 1$ for all $t = 1, \ldots, T$.

(M8) `smile`: piecewise-linear f_t with possible jumps at change-points (designed to test NOT under misspecification), $T = 2048$, $q = 6$ change-points at $\tau = 256, 512, 768, 1280, 1536, 1792$, with the corresponding sizes of jump $0, -4, 0, 0, 4, 0$ and changes in the slope $-2^{-5}, 0, 2^{-6}, 2^{-6}, 0, -2^{-5}$, starting value for the intercept $f_1 = 0$ and slope $f_2 - f_1 = 2^{-6}$, $\sigma_t = 1$ for $t = 1, \ldots, T$.

B. More details on the contrast functions and their construction

B.1. Scenario (S1)

Here f_t is piecewise-constant. For any integer triple (s, e, b) with $0 \le s < e \le T$ and $s < b < e$, recalling that we have defined the contrast vector $\boldsymbol{\psi}_{s,e}^b = \big(\psi_{s,e}^b(1), \ldots, \psi_{s,e}^b(T)\big)'$ as

$$\psi_{(s,e]}^b(t) = \begin{cases} \sqrt{\frac{e-b}{(e-s)(b-s)}}, & t = s+1, \ldots, b \\ -\sqrt{\frac{b-s}{(e-s)(e-b)}}, & t = b+1, \ldots, e \\ 0, & \text{otherwise.} \end{cases}$$

Also, if $b \notin \{s+1, \ldots, e-1\}$, then we set $\psi_{(s,e]}^b(t) = 0$ for all t.

For any vector $\mathbf{v} = (v_1, \ldots, v_T)'$, we define the contrast function as $\mathcal{C}_{(s,e]}^b(\mathbf{v}) = \left|\left\langle \mathbf{v}, \boldsymbol{\psi}_{(s,e]}^b \right\rangle\right|$. Therefore, if $s < b < e$, then

$$\mathcal{C}_{(s,e]}^b(\mathbf{v}) = \left| \sqrt{\frac{e-b}{(e-s)(b-s)}} \sum_{t=s+1}^{b} v_t - \sqrt{\frac{b-s}{(e-s)(e-b)}} \sum_{t=b+1}^{e} v_t \right|.$$

Otherwise, $\mathcal{C}_{(s,e]}^b(\mathbf{v}) = 0$. This recovers the well-known CUSUM statistic in the change-point detection literature. It can be shown that $\left[\mathcal{C}_{(s,e]}^b(\mathbf{Y})\right]^2 = \sigma_0^2 \mathcal{R}_{(s,e]}^b(\mathbf{Y})$ for every (s, e, b) with $0 \le s < b < e \le T$, thus $\mathcal{C}_{(s,e]}^b(\cdot)$ fulfills the requirements for the contrast function listed in Section 2.3.

In addition, for any $0 \le s < e \le T$, we define the constant vector for the interval $(s, e]$ as

$$\mathbf{1}_{(s,e]}(t) = \begin{cases} (e-s)^{-1/2}, & t = s+1, \ldots, e \\ 0, & \text{otherwise} \end{cases},$$

and write $\mathbf{1}_{(s,e]} = \big(\mathbf{1}_{(s,e]}(1), \ldots, \mathbf{1}_{(s,e]}(T)\big)'$. Then it is easy to check that $\mathbf{1}_{(s,e]}$ and $\boldsymbol{\psi}_{(s,e]}^b$ are orthonormal. This explains why the CUSUM is invariant to shifts in the mean.

B.2. Scenario (S2)

Here f_t is piecewise-linear and continuous. For any triple (s, e, b) with $0 \le s < e \le T$ and $s + 1 < b < e$, consider the contrast vector $\boldsymbol{\phi}_{(s,e]}^b = \big(\phi_{(s,e]}^b(1), \ldots, \phi_{(s,e]}^b(T)\big)'$ with

$$\phi_{(s,e]}^b(t) = \begin{cases} \alpha_{(s,e]}^b \beta_{(s,e]}^b \Big[\big\{3(b-s) + (e-b) - 1\big\}t - \big\{b(e-s-1) + 2(s+1)(b-s)\big\}\Big], & t = s+1, \ldots, b \\ -\frac{\alpha_{(s,e]}^b}{\beta_{(s,e]}^b} \Big[\big\{3(e-b) + (b-s) + 1\big\}t - \big\{b(e-s-1) + 2e(e-b+1)\big\}\Big], & t = b+1, \ldots, e, \\ 0, & \text{otherwise.} \end{cases}$$

where $\alpha_{s,e}^b = \left(\frac{6}{l(l^2-1)\big(1+(e-b+1)(b-s)+(e-b)(b-s-1)\big)}\right)^{1/2}$, $\beta_{s,e}^b = \left(\frac{(e-b+1)(e-b)}{(b-s-1)(b-s)}\right)^{1/2}$ and $l = e - s$. If $b \notin \{s+2, \ldots, e-1\}$, then we set $\phi_{(s,e]}^b(t) = 0$ for all t. The contrast function is then defined as

$$\mathcal{C}_{(s,e]}^b(\mathbf{v}) = \left|\left\langle \mathbf{v}, \boldsymbol{\phi}_{(s,e]}^b \right\rangle\right|.$$

To explain the rationale behind $\boldsymbol{\phi}^b_{(s,e]}$, we first define the "linear" vector for the interval $(s,e]$ with $e-s>1$, $\boldsymbol{\gamma}_{(s,e]} = \big(\gamma_{(s,e]}(1), \ldots, \gamma_{(s,e]}(T)\big)'$, as

$$\gamma_{(s,e]}(t) = \begin{cases} \Big\{\frac{1}{12}(e-s-1)(e-s)(e-s+1)\Big\}^{-1/2}\big(t - \frac{e+s+1}{2}\big), & t = s+1, \ldots, e; \\ 0, & \text{otherwise.} \end{cases}$$

Then we have that $\boldsymbol{\phi}^b_{(s,e]}$ is orthonormal to both $\mathbf{1}_{(s,e]}$ and $\boldsymbol{\gamma}_{(s,e]}$ (note that $\boldsymbol{\gamma}_{(s,e]}$ itself is orthonormal to $\mathbf{1}_{(s,e]}$). The orthonormality of the vectors $\mathbf{1}_{(s,e]}$, $\boldsymbol{\gamma}_{(s,e]}$ and $\boldsymbol{\phi}^b_{(s,e]}$ is important in deriving the identity $\sigma_0^2\mathcal{R}^b_{(s,e]}(\mathbf{Y}) = \mathcal{C}^b_{(s,e]}(\mathbf{Y})^2$ below, and helps improve the numerical efficiency and stability in our implementation of NOT. In particular, it means that the contrast function is invariant to both mean shifts and slope shifts on a given interval. In fact, $\boldsymbol{\phi}^b_{(s,e]}$ can be derived by (i) applying the Gram–Schmidt process on the following vector (linear with a kink at b on $(s,e]$)

$$\tilde{\phi}^b_{(s,e]}(t) = \begin{cases} t-b, & t = b+1, \ldots, e \\ 0, & \text{otherwise} \end{cases}$$

with respect to $\mathbf{1}_{(s,e]}$ and $\boldsymbol{\gamma}_{(s,e]}$, and (ii) normalisation such that $\|\cdot\|_2 = 1$. Now write the restriction of $\mathbf{v}$ on the interval $(s,e]$ as $\mathbf{v}|_{(s,e]} = (0, \ldots, 0, v_{s+1}, \ldots, v_e, 0, \ldots, 0)'$. Fix any (s,e,b), given the restriction imposed on $\boldsymbol{\Theta}$ in (S2), the best approximation of $\mathbf{Y}|_{(s,e]}$ (in the ℓ_2 distance) with a single kink at b is a linear combination of $\mathbf{1}_{(s,e]}$, $\boldsymbol{\gamma}_{(s,e]}$ and $\boldsymbol{\phi}^b_{(s,e]}$ (all mutually orthonormal). Therefore,

$$\begin{aligned}
&\sigma_0^2\mathcal{R}^b_{(s,e]}(\mathbf{Y}) \\
&= \min_{a_0,a_1\in\mathbb{R}} \|\mathbf{Y}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]}\|_2^2 - \min_{a_0,a_1,a_2\in\mathbb{R}} \|\mathbf{Y}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]} - a_2\boldsymbol{\phi}^b_{(s,e]}\|_2^2 \\
&= \|\mathbf{Y}|_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\gamma}_{(s,e]}\rangle\boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{Y}, \mathbf{1}_{(s,e]}\rangle\mathbf{1}_{(s,e]}\|^2 \\
&\quad - \|\mathbf{Y}|_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]}\rangle\boldsymbol{\phi}^b_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\gamma}_{(s,e]}\rangle\boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{Y}, \mathbf{1}_{(s,e]}\rangle\mathbf{1}_{(s,e]}\|^2 \\
&= \langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]}\rangle^2 = \mathcal{C}^b_{(s,e]}(\mathbf{Y})^2.
\end{aligned}$$

Thus the aforementioned requirements for the contrast function are satisfied.

B.3. Scenario (S3)

Here f_t is a piecewise-linear but not necessarily continuous function. We use the following contrast function for any $s+1<b<e-1$:

$$\mathcal{C}^b_{(s,e]}(\mathbf{v}) = \left(\Big\langle \mathbf{v}, \boldsymbol{\psi}^b_{(s,e]}\Big\rangle^2 + \Big\langle \mathbf{v}, \boldsymbol{\gamma}_{(s,b]}\Big\rangle^2 + \Big\langle \mathbf{v}, \boldsymbol{\gamma}_{(b,e]}\Big\rangle^2 - \Big\langle \mathbf{v}, \boldsymbol{\gamma}_{(s,e]}\Big\rangle^2\right)^{1/2}. \tag{13}$$

Otherwise, for $b \notin \{s+2, \ldots, e-2\}$, we set $\mathcal{C}^b_{(s,e]}(\mathbf{v}) = 0$.

This construction is justified by noting that

$$\begin{aligned}
\sigma_0^2 \mathcal{R}^b_{(s,e]}(\mathbf{Y}) &= \min_{a_0,a_1\in\mathbb{R}} \|\mathbf{Y}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]}\|_2^2 \\
&\quad - \left(\min_{a_0,a_1\in\mathbb{R}} \|\mathbf{Y}|_{(s,b]} - a_0\mathbf{1}_{(s,b]} - a_1\boldsymbol{\gamma}_{(s,b]}\|_2^2 + \min_{a_0,a_1\in\mathbb{R}} \|\mathbf{Y}|_{(b,e]} - a_0\mathbf{1}_{(b,e]} - a_1\boldsymbol{\gamma}_{(b,e]}\|_2^2 \right) \\
&= \min_{a_0,a_1\in\mathbb{R}} \|\mathbf{Y}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]}\|_2^2 \\
&\quad - \min_{a_0,a_1,a_2,a_3\in\mathbb{R}} \|\mathbf{Y}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,b]} - a_2\boldsymbol{\gamma}_{(b,e]} - a_3\boldsymbol{\psi}^b_{(s,e]}\|_2^2 \\
&= \mathcal{C}^b_{(s,e]}(\mathbf{Y})^2,
\end{aligned}$$

where we also used the orthonormality among $\mathbf{1}_{(s,e]}$, $\boldsymbol{\psi}^b_{(s,e]}$, $\boldsymbol{\gamma}_{(s,b]}$ and $\boldsymbol{\gamma}_{(b,e]}$ in the above derivation.

B.4. Scenario (S4)

Here both f_t and σ_t are piecewise-constant. For any $s+1 < b < e-1$, we propose

$$\mathcal{C}^b_{(s,e]}(\mathbf{v}) = (e-s)\log\left(\hat{\sigma}_{(s,e]}(\mathbf{v})\right) - (b-s)\log\left(\hat{\sigma}_{(s,b]}(\mathbf{v})\right) - (e-b)\log\left(\hat{\sigma}_{(b,e]}(\mathbf{v})\right), \tag{14}$$

where

$$\hat{\sigma}^2_{(s,e]}(\mathbf{v}) = \frac{1}{e-s}\sum_{t=s+1}^{e}\left(v_t - \frac{1}{e-s}\sum_{t=s+1}^{e} v_t\right)^2 = \left\langle \mathbf{v}^2, \mathbf{1}^2_{(s,e]} \right\rangle - \left\langle \mathbf{v}, \mathbf{1}^2_{(s,e]} \right\rangle^2.$$

Otherwise, for $b \notin \{s+2,\ldots,e-2\}$, we set $\mathcal{C}^b_{(s,e]}(\mathbf{v}) = 0$. In this Scenario, it is straightforward to verify that $\mathcal{C}^b_{(s,e]}(\mathbf{Y}) = \mathcal{R}^b_{(s,e]}(\mathbf{Y})$. (N.B. $\mathbf{1}^2_{(s,e]} \neq \mathbf{1}_{(s,e]}$ due to the normalising constant.) In practice, for numerical stability, we use $\log_\epsilon(\cdot) := \log\{\max(\cdot,\epsilon)\}$ instead of $\log(\cdot)$ in (14) with a pre-given small $\epsilon > 0$.

C. More details on the compututational aspects of NOT and its solution path

C.1. Computing contrast functions in a linear time

The practical performance (in terms of computational cost) of Algorithm 1 relies on the fast computation of the contrast functions discussed in Section 2.3 on any given interval $(s,e]$. Here we show that in all scenarios listed in Section 2.3, the cost of computing $\{\mathcal{C}^b_{(s,e]}(\mathbf{Y})\}_{b=s+1}^{e-1}$ is $O(e-s)$.

Note that the key ingredients in $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ under the different scenarios are functions of the inner products, i.e. $\left\langle \mathbf{Y}, \boldsymbol{\psi}^b_{(s,e]} \right\rangle$, $\left\langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]} \right\rangle$, $\left\langle \mathbf{Y}, \boldsymbol{\gamma}_{(s,b]} \right\rangle$, $\left\langle \mathbf{Y}, \boldsymbol{\gamma}_{(b,e]} \right\rangle$, $\left\langle \mathbf{Y}, \mathbf{1}^2_{(s,b]} \right\rangle$, $\left\langle \mathbf{Y}, \mathbf{1}^2_{(b,e]} \right\rangle$, $\left\langle \mathbf{Y}^2, \mathbf{1}^2_{(s,b]} \right\rangle$ and $\left\langle \mathbf{Y}^2, \mathbf{1}^2_{(b,e]} \right\rangle$ for $b = s+1,\ldots,e-1$. For a fixed interval $(s,e]$, by simple algebra, we observe that $\left\langle \mathbf{Y}, \boldsymbol{\psi}^b_{(s,e]} \right\rangle$ and $\left\langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]} \right\rangle$ can be decomposed as

$$\begin{aligned}
\left\langle \mathbf{Y}, \boldsymbol{\psi}^b_{(s,e]} \right\rangle &= \overleftarrow{a}_{\boldsymbol{\psi},b} \sum_{t=s+1}^{b} Y_t - \overrightarrow{a}_{\boldsymbol{\psi},b} \sum_{t=b+1}^{e} Y_t \\
&:= \overleftarrow{a}_{\boldsymbol{\psi},b} \overleftarrow{\pi}^{(0)}_b(\mathbf{Y}) - \overrightarrow{a}_{\boldsymbol{\psi},b} \overrightarrow{\pi}^{(0)}_b(\mathbf{Y}), \\
\left\langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]} \right\rangle &= \overleftarrow{a}^{(1)}_{\boldsymbol{\phi},b} \sum_{t=s+1}^{b} tY_t - \overrightarrow{a}^{(1)}_{\boldsymbol{\phi},b} \sum_{t=b+1}^{e} tY_t + \overleftarrow{a}^{(0)}_{\boldsymbol{\phi},b} \sum_{t=s+1}^{b} Y_t - \overrightarrow{a}^{(0)}_{\boldsymbol{\phi},b} \sum_{t=b+1}^{e} Y_t \\
&:= \overleftarrow{a}^{(1)}_{\boldsymbol{\phi},b} \overleftarrow{\pi}^{(1)}_b(\mathbf{Y}) - \overrightarrow{a}^{(1)}_{\boldsymbol{\phi},b} \overrightarrow{\pi}^{(1)}_b(\mathbf{Y}) + \overleftarrow{a}^{(0)}_{\boldsymbol{\phi},b} \overleftarrow{\pi}^{(0)}_b(\mathbf{Y}) - \overrightarrow{a}^{(0)}_{\boldsymbol{\phi},b} \overrightarrow{\pi}^{(0)}_b(\mathbf{Y}),
\end{aligned}$$

where $\overleftarrow{a}_{\boldsymbol{\psi},b}$, $\overrightarrow{a}_{\boldsymbol{\psi},b}$, $\overleftarrow{a}^{(1)}_{\boldsymbol{\phi},b}$, $\overrightarrow{a}^{(1)}_{\boldsymbol{\phi},b}$, $\overleftarrow{a}^{(0)}_{\boldsymbol{\phi},b}$ and $\overrightarrow{a}^{(0)}_{\boldsymbol{\phi},b}$ are scalars that do not depend on $\mathbf{Y}$, and can all be computed at the cost of $O(1)$ using equations given in Section 2.3. Here for notational convenience, we use overhead arrows to indicate whether a scalar or a function is associated with observations with indices $\leq b$ (i.e. over $(s, b]$, using $\overleftarrow{\cdot}$) or with indices $> b$ (i.e. over $(b, e]$, using $\overrightarrow{\cdot}$). We also suppress their dependence on s and e in the notation. In addition, the following recursive formulae hold

$$\begin{aligned}
\overleftarrow{\pi}^{(k)}_{b+1}(\mathbf{Y}) &= \overleftarrow{\pi}^{(k)}_b(\mathbf{Y}) + (b+1)^k Y_{b+1}, \\
\overrightarrow{\pi}^{(k)}_b(\mathbf{Y}) &= \overrightarrow{\pi}^{(k)}_{b+1}(\mathbf{Y}) + (b+1)^k Y_{b+1},
\end{aligned}$$

with $\overleftarrow{\pi}^{(k)}_s(\mathbf{Y}) = \overrightarrow{\pi}^{(k)}_e(\mathbf{Y}) = 0$ for $k = 0, 1$. Consequently, $\overleftarrow{\pi}^{(k)}_b(\mathbf{Y})$ and $\overrightarrow{\pi}^{(k)}_b(\mathbf{Y})$ for all $b \in \{s+1, \ldots, e-1\}$ and $k = 0, 1$ (thereby $\left\langle \mathbf{Y}, \boldsymbol{\psi}^b_{(s,e]} \right\rangle$ and $\left\langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]} \right\rangle$) can be computed in a single pass through $Y_{s+1}, \ldots, Y_e$. Similar approach can be applied to the remaining inner products involved in the definitions of the contrast functions given in Section 2.3, which demonstrates that in all these cases the computation of $\{\mathcal{C}^b_{(s,e]}(\mathbf{Y})\}_{b=s+1}^{e-1}$ scales linearly with the length of the sub-interval.

C.2. Details of the NOT solution path algorithm

As mentioned in Section 3.2 of the main paper, we have developed Algorithm 2 that computes the entire threshold-indexed solution path $\{\mathcal{T}(\zeta_T)\}_{\zeta_T \geq 0}$ quickly, and have implemented it in our R package **not**. Detailed pseudo-code is provided on the next page.

The construction of Algorithm 2 stems from two observations. First, for any fixed threshold ζ_T, Algorithm 1 implies a binary tree data structure that is constructed according to the order of the detection of each change-point. More specifically, in our implementation, each tree node `N` contains the following information.

(a) The current interval of interest is (`N.s`, `N.e`].

(b) From all elements in F_T^M that are also subsets of (`N.s`, `N.e`], we find the narrowest-over-threshold sub-interval. Within that sub-interval, let `N.c` be the maximum achieved value of the contrast function over all possible locations of the feature, and `N.b` be the corresponding location (i.e. the detected change-point location over (`N.s`, `N.e`]).

(c) `N.Left` and `N.Right` point to the nodes of the next detected change-points in (`N.s`, `N.b`] and (`N.b`, `N.e`], respectively.

Algorithm 2 NOT solution path

Input: Data vector $\mathbf{Y}$, all sub-intervals $(s_m, e_m] \in F_T^M$ together with

$$b_m := \mathrm{argmax}_{s_m < b \le e_m} \mathcal{C}^b_{(s_m,e_m]}(\mathbf{Y}), \quad c_m := \mathcal{C}^{b_m}_{(s_m,e_m]}(\mathbf{Y}) \quad \text{and} \quad l_m := e_m - s_m.$$

Output: Thresholds $0 = \zeta_T^{(1)} < \ldots < \zeta_T^{(N)}$ and sets of estimated change-points $\mathcal{T}(\zeta_T^{(1)}), \ldots, \mathcal{T}(\zeta_T^{(N)})$.

To start the algorithm: Call SolutionPath()

```
procedure BuildBinaryTree((s, e], ζ_T, N)
    M_(s,e] := set of those m ∈ {1, ..., M} such that (s_m, e_m] ⊂ (s, e]
    O_(s,e] := set of m ∈ M_(s,e] such that c_m > ζ_T
    if O_(s,e] = ∅ then N = NULL
    else
        k := any element of argmin_{m ∈ O_(s,e]} l_m
        N.b := b_k, N.c := c_k, N.Left := NULL, N.Right := NULL
        BuildBinaryTree((s, N.b], ζ_T, N.Left)
        BuildBinaryTree((N.b, e], ζ_T, N.Right)
    end if
end procedure

procedure UpdateBinaryTree((s, e], ζ_T, N)
    if N.c ≤ ζ_T then
        BuildBinaryTree((s, e], ζ_T, N)
    else
        if N.Left ≠ NULL then
            UpdateBinaryTree((s, N.b], ζ_T, N.Left)
        end if
        if N.Right ≠ NULL then
            UpdateBinaryTree((N.b, e], ζ_T, N.Right)
        end if
    end if
end procedure

procedure SolutionPath()
    Set N_r := NULL, i := 1, ζ_T^(1) := 0
    BuildBinaryTree((0, T], ζ_T^(1), N_r)
    while N_r ≠ NULL do
        D := {N_r and all its children nodes}
        T(ζ_T^(i)) := {N.b | N ∈ D}
        ζ_T^(i+1) := min_{N ∈ D} {N.c}
        UpdateBinaryTree((0, T], ζ_T^(i+1), N_r)
        i := i + 1
    end while
end procedure
```

Table 2. Intervals considered in Figure 7a and corresponding maxima of the contrast function $\mathcal{C}^b_{(s,e]}(\cdot)$ given by (8), all calculated for a sample path of Y_t, $t = 1, \ldots, 1000$ generated from model (1) with the signal f_t given by (2) and the noise $\varepsilon_t \sim \mathcal{N}(0, 0.05^2)$.

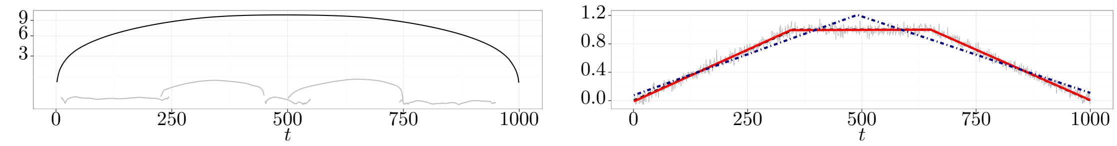
s	e	$e-s$	$\text{argmax}_{s<b\le e}\, \mathcal{C}^b_{(s,e]}(\mathbf{Y})$	$\max_{s<b\le e} \mathcal{C}^b_{(s,e]}(\mathbf{Y})$
0	1000	1000	490	10.19
9	245	236	43	0.08
224	450	226	344	0.76
499	750	251	651	0.83
749	950	211	746	0.03
449	550	101	471	0.07

We then treat the first detected change-point over $(0, T]$ as the root of the tree and construct its branches in a recursive fashion afterwards. Second, suppose that we have already constructed the tree for ζ_T with root $\mathtt{N_r}$. For $\zeta'_T > \zeta_T$, the new tree's root is unchanged if $\mathtt{N_r.c} > \zeta'_T$. This observation remains valid for $\mathtt{N_r.Left}$ and $\mathtt{N_r.Right}$ and all subsequent nodes. Therefore, a branch of the tree has to be reconstructed only if $\mathtt{N.c} \le \zeta'_T$ for some node N. In this way, the tree constructed for ζ_T can be used as a starting point to finding the tree corresponding to ζ'_T, thus significantly reducing the computational time in comparison to constructing the tree from scratch.

Next, we elaborate on the complexity of Algorithm 2. As explained previously, finding solutions of Algorithm 1 for a single threshold ζ_T is equivalent to the construction of a binary tree, which can be performed with the BuildBinaryTree routine given in Algorithm 2. Computational cost of this operation is no larger than $O(MK_{\zeta_T})$, where K_{ζ_T} denotes the height of the constructed binary tree with the threshold ζ_T. The computational complexity of finding the entire solution path using Algorithm 2 is therefore (in the worst case) $O(MKN)$, where N and K are, respectively, the number of solutions and the maximum tree depth over the entire solution path. However, this is a rough estimate which assumes that for each threshold on the path the binary tree has a different root node, which, from our empirical experience, is highly unlikely to occur in practice. Typically, the consecutive trees on the path differ just slightly (see e.g. our next Section C.3), which significantly reduces the amount of computation that Algorithm 2 requires. As such, we find that the computational complexity of Algorithm 2 is more like $O(MT)$ in practice.

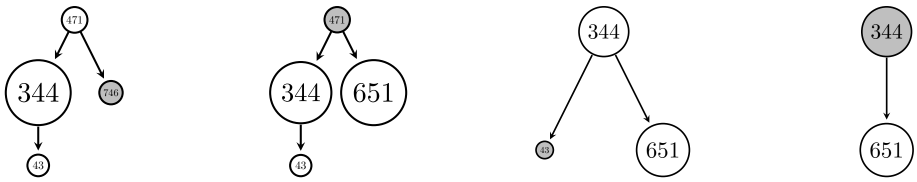
C.3. An illustrative example

In this part, we revisit the example shown in the Introduction of our paper, and provide a simple illustration of how Algorithm 1 and Algorithm 2 work on a simulated dataset. Figure 7 shows the generated data $\{Y_t\}_{t=1}^{1000}$ following Scenario (S2), where the signal f_t is as in (2) and $\sigma_t = 0.05$. The contrast function (8) is evaluated for 5 intervals. We observe that the contrast function corresponding to $(0, 1000]$, being the longest interval here, attains its maximum at $b = 490$, which is far from the true change-points located at $\tau = 350$ and $\tau = 650$. Furthermore, $\max_b \mathcal{C}^b_{(0,1000]}(\mathbf{Y})$ is much larger than the corresponding value for the other intervals considered in Table 2. However, thanks to the fact that we focus on the narrowest-over-threshold intervals, Algorithm 1 (for any $\zeta_T \in (0.08, 0.83)$) picks at its first iteration an interval with exactly one change-point (depending on ζ_T, it is either $(224, 450]$ or $(499, 750]$) and the maximum of the contrast function computed is close to one of the true change-points.



(a) (b)

Fig. 7. An application of the NOT methodology to Y_t generated from model (1) with the signal f_t given by (2) and i.i.d. $\varepsilon_t \sim \mathcal{N}(0, 0.05^2)$. Figure 7a: contrast function $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ given by (8) evaluated for all $b \in (s, e]$ with intervals $(s, e]$ specified in Table 2. For intervals containing one change-point, $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ attains its maximum at b close to the actual change-point. When there are two change-points (black solid line), the maximum is far from both change-points, despite $\max_b \mathcal{C}^b_{(s,e]}(\mathbf{Y})$ being large. Figure 7b: observed Y_t (thin grey), true signal (thick dashed black), signal estimated picking the change-point candidate based on the interval corresponding to the largest contrast function (dotted-dashed navy) and the *narrowest-over-threshold* intervals (dashed red).



(a) $\zeta_T^{(1)} = 0$ (b) $\zeta_T^{(2)} = 0.03$ (c) $\zeta_T^{(3)} = 0.07$ (d) $\zeta_T^{(4)} = 0.08$

Fig. 8. First four segmentation trees obtained by Algorithm 2 applied to a realization of $(Y_1, \ldots, Y_{1000})'$ presented in Figure 7. The larger the node, the larger the corresponding value of $\max_b \mathcal{C}^b_{(s,e]}(\mathbf{Y})$. Here $\mathcal{C}^b_{(s,e]}(\cdot)$ is given by (8). The grey nodes correspond to the smallest contrast function for each tree that are updated as Algorithm 2 proceeds.

Figure 8 shows how Algorithm 2 proceeds in the example presented in Figure 7. At the initial stage that can be seen in Figure 8a, the threshold is set to $\zeta_T^{(1)} = 0$ and $b = 471$, the maximum of the contrast function computed for the shortest interval $(449, 550]$ is taken as the root of the binary tree. Then we construct its left and right branches by considering only those intervals specified in Table 2 with $(s, e] \subset (0, 471]$ and $(s, e] \subset (471, 1000]$, respectively, and the procedure continues for the resulting nodes. Next, the node with the smallest value of the contrast function is determined $(b = 746)$ and the threshold is set to the corresponding minimum $\zeta_T^{(2)} = 0.03$. This guarantees that as Algorithm 2 proceeds, there will be at least one update in the binary tree. In our example, the $b = 746$ node is removed and, as the maximum for $(499, 750] \subset (471, 1000]$ exceeds the threshold, the $b = 651$ node is inserted its place. Subsequently, we identify the node with the smallest contrast again $(b = 471)$, update the threshold to $\zeta_T^{(3)} = 0.07$ and reconstruct the entire tree, as $b = 471$ in Figure 8b constitutes its root. Algorithm 2 keeps running until the resulting tree shrinks to `NULL`. In this example, the fourth solution on the path (Figure 8d) contains exactly two nodes being close to the true change-points.

D. Further extension of NOT: noise with slow-varying variance

In all scenarios considered previously, we assumed a constant or piecewise constant σ_t. Now we discuss how NOT can be extended to handle σ_t of a more general form. We model $\mathbf{Y}$ through (3) with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. To fix ideas, here we focus on the case of piecewise constant signal f_t (i.e. similar to (S1)), but with a slowly-varying σ_t, i.e. $\sigma_t = \sigma(t/T)$ with $\sigma(\cdot)$ being an unknown smooth function from $[0,1] \to (0,\infty)$.

D.1. Methodology

We propose to estimate the change-points in three steps:

(a) We estimate $\sigma(\cdot)$ using a standard nonparametric method, such as spline smoothing, on $\big\{(t/T, \sqrt{\pi}|Y_{t+1} - Y_t|/2)\big\}_{t=2}^{T}$, which we denote as $\hat{\sigma}(\cdot)$. Also, we write $\hat{\sigma}_t = \hat{\sigma}(t/T)$ for $t = 1, \ldots, T$.

(b) We perform our NOT solution path algorithm using the contrast function $\hat{\mathcal{C}}^b_{(s,e]}(\mathbf{Y}) = \left|\left\langle \mathbf{Y}, \hat{\boldsymbol{\psi}}^b_{(s,e]} \right\rangle\right|$ for any tuple (s,b,e) with $0 \le s < b < e \le T$ in our NOT procedure, with $\hat{\boldsymbol{\psi}}^b_{(s,e]} = \big(\hat{\psi}^b_{(s,e]}(1), \ldots, \hat{\psi}^b_{(s,e]}(T)\big)'$ and

$$\hat{\psi}^b_{(s,e]}(t) = \begin{cases} \hat{\sigma}_t^{-2}\sqrt{(\Omega_e^2 - \Omega_b^2)(\Omega_e^2 - \Omega_s^2)^{-1}(\Omega_b^2 - \Omega_s^2)^{-1}}, & t = s+1, \ldots, b \\ -\hat{\sigma}_t^{-2}\sqrt{(\Omega_b^2 - \Omega_s^2)(\Omega_e^2 - \Omega_s^2)^{-1}(\Omega_e^2 - \Omega_b^2)^{-1}}, & t = b+1, \ldots, e \\ 0, & \text{otherwise.} \end{cases}$$

where $\Omega_t^2 = \sum_{i=1}^{t} 1/\hat{\sigma}_i^2$ (and by default $\Omega_0^2 = 0$). As before, if $b \notin \{s+1, \ldots, e-1\}$, then we set $\hat{\psi}^b_{(s,e]}(t) = 0$ for all t. We remark that this contrast function originates from the generalised log-likelihood ratio, and can be viewed as a weighted and scaled version of (6) based on CUSUM statistic. Its derivation can be found in Section D.2. In addition, when $\hat{\sigma}_t$ is constant (say $\hat{\sigma}_t = 1$ for all t), we would recover (6).

(c) We pick the best model along the solution path via the sSIC criterion, with the log-likelihood for each segment given by

$$\log \ell(Y_{\hat{\tau}_{j-1}+1}, \ldots, Y_{\hat{\tau}_j}; \hat{\boldsymbol{\Theta}}_j) = \sum_{t=\hat{\tau}_{j-1}+1}^{\hat{\tau}_j} \left\{ -\frac{(Y_t - \hat{Y}_{(\hat{\tau}_{j-1},\hat{\tau}_j]})^2}{2\hat{\sigma}_t^2} - \frac{\log(2\pi\hat{\sigma}_t^2)}{2} \right\},$$

where $\hat{Y}_{(\hat{\tau}_{j-1},\hat{\tau}_j]} = \big(\sum_{t=\hat{\tau}_{j-1}+1}^{\hat{\tau}_j} \hat{\sigma}_t^{-2} Y_t\big) / \big(\sum_{t=\hat{\tau}_{j-1}+1}^{\hat{\tau}_j} \hat{\sigma}_t^{-2}\big)$.

To make this solution complete, a suitable choice of the smoothing parameter would have to be considered in the first step. This is a standard problem in nonparametrics, and several solutions exist, e.g. those based on (leave-one-out) cross-validation. We leave a detailed study of this issue for future research.

D.2. Detailed derivation of the corresponding contrast function

Here we derive the contrast function from the generalised log-likelihood ratio.

Given an interval $(s, e]$. Suppose that there is a change-point at b, then under the normality assumption, the log-likelihood is

$$-\frac{1}{2}\sum_{t=s+1}^{b}\frac{(Y_t-\mu_L)^2}{\hat{\sigma}_t^2}-\frac{1}{2}\sum_{t=b+1}^{e}\frac{(Y_t-\mu_R)^2}{\hat{\sigma}_t^2}-\frac{1}{2}\sum_{t=s+1}^{e}\log(2\pi\hat{\sigma}_t^2),$$

which is maximised at

$$\mu_L=\left(\sum_{t=s+1}^{b}\frac{1}{\hat{\sigma}_t^2}\right)^{-1}\left(\sum_{t=s+1}^{b}\frac{Y_t}{\hat{\sigma}_t^2}\right)\quad\text{and}\quad\mu_R=\left(\sum_{t=b+1}^{e}\frac{1}{\hat{\sigma}_t^2}\right)^{-1}\left(\sum_{t=b+1}^{e}\frac{Y_t}{\hat{\sigma}_t^2}\right).$$

Now suppose there is no change-point over $(s, e]$, then the log-likelihood is

$$-\frac{1}{2}\sum_{t=s+1}^{e}\frac{(Y_t-\mu)^2}{\hat{\sigma}_t^2}-\frac{1}{2}\sum_{t=s+1}^{e}\log(2\pi\hat{\sigma}_t^2),$$

which is maximised at

$$\mu=\left(\sum_{t=s+1}^{e}\frac{1}{\hat{\sigma}_t^2}\right)^{-1}\left(\sum_{t=s+1}^{e}\frac{Y_t}{\hat{\sigma}_t^2}\right)=\frac{\Omega_b^2-\Omega_s^2}{\Omega_e^2-\Omega_s^2}\mu_L+\frac{\Omega_e^2-\Omega_b^2}{\Omega_e^2-\Omega_s^2}\mu_R,$$

where $\Omega_t^2=\sum_{i=1}^{t}1/\hat{\sigma}_i^2$ (and by default $\Omega_0^2=0$). After some algebraic manupulation, we have that the generalised log-likelihood is

$$\begin{aligned}
\mathcal{R}_{(s,e]}^{b}(\mathbf{Y}) &= \frac{1}{2}\Big\{(\Omega_b^2-\Omega_s^2)\mu_L^2+(\Omega_e^2-\Omega_b^2)\mu_R^2-(\Omega_e^2-\Omega_s^2)\mu^2\Big\}\\
&= \frac{1}{2}\frac{(\Omega_b^2-\Omega_s^2)(\Omega_e^2-\Omega_s^2)}{\Omega_e^2-\Omega_b^2}(\mu_L-\mu_R)^2\\
&= \frac{1}{2}\left(\sqrt{\frac{\Omega_e^2-\Omega_b^2}{(\Omega_e^2-\Omega_s^2)(\Omega_b^2-\Omega_s^2)}}\sum_{t=s+1}^{b}\frac{Y_t}{\hat{\sigma}_t^2}-\sqrt{\frac{\Omega_b^2-\Omega_s^2}{(\Omega_e^2-\Omega_s^2)(\Omega_e^2-\Omega_b^2)}}\sum_{t=b+1}^{e}\frac{Y_t}{\hat{\sigma}_t^2}\right)^2\\
&= \frac{1}{2}\left|\left\langle\mathbf{Y},\hat{\boldsymbol{\psi}}_{(s,e]}^{b}\right\rangle\right|^2\\
&= \frac{1}{2}\left\{\hat{\mathcal{C}}_{(s,e]}^{b}(\mathbf{Y})\right\}^2.
\end{aligned}$$

E. Additional simulation results

In addition to the results presented in Section 5, here we present Tables 3–6 that summarise the results for three different distributions of the noise ε_t, where (b) $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2)$, (c) $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \text{Laplace}(0, 2^{-1/2})$, (d) i.i.d. scaled Student-t_5 in Table 5, and (e) ε_t follows zero-mean unit-variance Gaussian AR(1) with $\varphi = 0.3$.

Table 3. Distribution of $\hat{q} - q$ for data generated according to (3) with the noise term ε_t being i.i.d. $\mathcal{N}(0, 2)$ for various choices of f_t and σ_t given in Section A and competing methods listed in Section 5. Also, the average Mean-Square Error of the resulting estimate of the signal f_t, average inverse V-measure d_H, average V distance d_V and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average of d_H or d_V, and those within 10% of the highest or lowest accordingly.

		$\hat{q} - q$										
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	$d_H \times 10^2$	d_V	time
B&P		0	0	0	**100**	0	0	0	0.107	**0.93**	**0.037**	1.337
e-cp3o		0	0	0	**100**	0	0	0	0.132	**0.98**	0.052	0.12
FDRSeg		0	0	0	83	14	1	2	0.134	1.51	0.054	0.093
NMCD		0	0	0	**92**	8	0	0	0.151	1.44	0.059	1.067
NOT		0	0	0	**97**	3	0	0	0.111	1.05	**0.038**	0.044
NOT HT	(M1)	0	0	0	**97**	3	0	0	0.126	1.25	0.044	0.058
NP-PELT		0	0	0	73	25	2	0	0.141	1.55	0.048	0.019
PELT		0	2	0	**98**	0	0	0	0.115	1.22	**0.039**	0.002
S3IB		0	0	0	**90**	8	1	1	0.114	1.13	**0.038**	0.076
SMUCE		0	2	14	84	0	0	0	0.185	2.14	0.064	0.056
WBS		0	0	0	**93**	7	0	0	0.113	1.15	**0.038**	0.074
B&P		100	0	0	0	0	0	0	0.145	8.78	0.155	30.413
e-cp3o		100	0	0	0	0	0	0	0.223	7.74	0.15	2.425
FDRSeg		20	33	36	7	2	2	0	0.073	**3.31**	0.066	2.576
NMCD		46	27	21	6	0	0	0	0.076	4.29	0.074	4.324
NOT		28	30	27	13	2	0	0	0.066	**3.4**	**0.059**	0.077
NOT HT	(M2)	49	27	19	2	3	0	0	0.083	4.26	0.077	0.138
NP-PELT		4	9	30	**21**	21	10	5	0.068	3.74	**0.062**	0.239
PELT		91	7	2	0	0	0	0	0.114	8.21	0.122	0.004
S3IB		37	34	17	10	1	1	0	0.071	4.15	0.068	0.342
SMUCE		100	0	0	0	0	0	0	0.144	5.95	0.13	0.022
WBS		26	32	29	13	0	0	0	0.067	**3.55**	**0.062**	0.145
B&P		0	0	100	0	0	0	0	0.258	4.25	0.155	54.381
NOT	(M3)	0	0	0	**97**	2	1	0	0.033	**1.59**	**0.073**	0.35
TF		0	0	0	0	0	0	100	0.032	8.42	0.216	46.038
B&P		13	53	28	6	0	0	0	0.322	6.11	0.204	62.421
NOT	(M4)	0	0	0	**100**	0	0	0	0.037	**2.01**	**0.097**	0.335
TF		0	0	0	0	0	0	100	0.03	4.47	0.151	47.536
B&P		0	0	9	**91**	0	0	0	0.046	**3.52**	**0.115**	119.454
NOT	(M5)	0	0	7	**92**	1	0	0	0.047	**3.65**	**0.117**	0.334
TF		0	0	0	0	0	0	100	0.041	5.9	0.24	57.36
e-cp3o		11	12	12	33	20	5	7	0.145	6.91	0.164	1.738
HSMUCE		97	3	0	0	0	0	0	0.091	12.77	0.209	0.051
NMCD		0	0	18	70	12	0	0	0.06	4.04	0.068	4.135
NOT	(M6)	0	0	13	**85**	2	0	0	0.047	2.6	**0.048**	0.455
NP-PELT		0	0	1	19	26	24	30	0.126	3.17	0.068	0.279
PELT		9	18	31	37	5	0	0	0.069	8.17	0.087	0.011
SegNeigh		0	0	3	49	36	10	2	0.053	**1.98**	**0.048**	17.211
B&P		0	0	42	**58**	0	0	0	0.073	**6.21**	**0.132**	29.222
NOT	(M7)	0	0	43	**57**	0	0	0	0.071	**6.13**	**0.122**	0.225
TF		0	0	0	0	0	0	100	0.08	22.86	0.399	43.198

Table 4. Distribution of $\hat{q} - q$ for data generated according to (3) with the noise term ε_t being i.i.d. Laplace $\left(0, (\sqrt{2})^{-1}\right)$ (N.B. $\mathrm{Var}(\varepsilon_t) = 1$) for various choices of f_t and σ_t given in Section A and competing methods listed in Section 5. Also, the average Mean-Square Error of the resulting estimate of the signal f_t, average Hausdorff distance d_H, average inverse V-measure d_V and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average of d_H or d_V, and those within 10% of the highest or lowest accordingly.

		$\hat{q} - q$										
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	$d_H \times 10^2$	d_V	time
B&P	(M1)	0	0	0	**100**	0	0	0	0.105	1.02	**0.035**	1.37
e-cp3o		0	0	0	**100**	0	0	0	0.123	**0.91**	0.049	0.116
FDRSeg		0	0	0	2	1	5	92	0.207	5.16	0.116	0.078
NMCD		0	0	0	**92**	7	1	0	0.15	1.55	0.059	1.053
NOT		0	0	0	**93**	4	3	0	0.114	1.3	0.038	0.043
NOT HT		0	0	0	**100**	0	0	0	0.099	**0.96**	**0.033**	0.058
NP-PELT		0	0	0	55	31	12	2	0.154	2.07	0.05	0.018
PELT		0	0	0	58	17	16	9	0.17	1.89	0.042	0.001
S3IB		0	0	0	70	12	12	6	0.154	1.64	0.04	0.068
SMUCE		0	0	0	48	20	22	10	0.154	2.68	0.066	0.057
WBS		0	0	0	60	4	22	14	0.173	2.18	0.044	0.073
B&P	(M2)	100	0	0	0	0	0	0	0.144	8.7	0.151	32.221
e-cp3o		100	0	0	0	0	0	0	0.208	7.62	0.145	2.313
FDRSeg		0	0	0	0	0	0	100	0.1	7.58	0.158	1.909
NMCD		24	36	35	4	1	0	0	0.06	3.57	0.056	4.354
NOT		63	14	12	6	4	1	0	0.085	5.09	0.082	0.078
NOT HT		31	27	34	8	0	0	0	0.056	**3.2**	**0.049**	0.139
NP-PELT		1	0	12	**22**	24	16	25	0.084	4.17	0.06	0.223
PELT		25	23	15	15	10	8	4	0.112	5.35	0.081	0.004
S3IB		90	7	1	2	0	0	0	0.125	9.55	0.142	0.315
SMUCE		21	15	19	11	13	13	8	0.111	6.03	0.12	0.019
WBS		25	11	15	13	14	5	17	0.11	5.34	0.083	0.143
B&P	(M3)	0	0	100	0	0	0	0	0.255	4.1	0.153	54.071
NOT		0	0	0	**93**	5	2	0	0.038	**1.93**	**0.078**	0.345
TF		0	0	0	0	0	0	100	0.035	8.42	0.224	46.298
B&P	(M4)	10	49	35	6	0	0	0	0.311	6.27	0.204	61.911
NOT		0	0	1	**93**	6	0	0	0.042	**2.26**	**0.1**	0.309
TF		0	0	0	0	0	0	100	0.03	4.41	0.153	47.57
B&P	(M5)	0	0	10	**90**	0	0	0	0.044	**3.47**	**0.112**	118.603
NOT		0	0	5	**92**	3	0	0	0.045	**3.62**	**0.112**	0.329
TF		0	0	0	0	0	0	100	0.041	5.87	0.232	57.763
e-cp3o	(M6)	34	19	9	11	6	7	14	0.304	9.94	0.225	1.693
HSMUCE		100	0	0	0	0	0	0	0.199	15.6	0.275	0.064
NMCD		4	18	44	31	3	0	0	0.114	9.25	0.116	4.085
NOT		2	6	33	**38**	18	2	1	0.185	7.82	0.107	0.451
NP-PELT		0	1	0	14	24	23	38	0.364	**5.44**	0.109	0.32
PELT		26	13	35	22	4	0	0	0.226	13.41	0.148	0.013
SegNeigh		0	0	7	30	38	13	12	0.176	**5.23**	**0.094**	17.316
B&P	(M7)	0	0	39	**60**	1	0	0	0.067	**6.23**	**0.123**	28.903
NOT		0	2	50	48	0	0	0	0.077	7.42	0.136	0.211
TF		0	0	0	0	0	0	100	0.074	22.81	0.406	43.53

Table 5. Distribution of $\hat{q} - q$ for data generated according to (3) with the noise term ε_t being i.i.d. $(3/5)^{1/2}t_5$ (N.B. $\mathrm{Var}(\varepsilon_t) = 1$) for various choices of f_t and σ_t given in Section A of the online supplementary materials and competing methods listed in Section 5. Also, the average Mean-Square Error of the resulting estimate of the signal f_t, average Hausdorff distance d_H, average inverse V-measure d_V and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q} - q = 0$ or smallest average of d_H or d_V, and those within 10% of the highest or lowest accordingly.

		$\hat{q} - q$										
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	$d_H \times 10^2$	d_V	time
B&P		0	0	0	**100**	0	0	0	0.046	**0.45**	**0.016**	1.36
e-cp3o		0	0	0	**100**	0	0	0	0.087	0.58	0.04	0.119
FDRSeg		0	0	0	8	2	5	85	0.113	4.67	0.089	0.07
NMCD		0	0	0	**97**	3	0	0	0.089	0.67	0.041	1.07
NOT		0	0	0	**96**	4	0	0	0.049	0.53	**0.017**	0.047
NOT HT	(M1)	0	0	0	**99**	1	0	0	0.045	**0.48**	**0.016**	0.057
NP-PELT		0	0	0	75	12	12	1	0.081	1.35	0.031	0.015
PELT		0	0	0	53	7	25	15	0.106	1.89	0.026	0.002
S3IB		0	0	0	50	10	28	12	0.105	1.97	0.026	0.066
SMUCE		0	0	0	43	13	21	23	0.093	2.65	0.054	0.056
WBS		0	0	0	43	3	29	25	0.12	2.45	0.031	0.071
B&P		100	0	0	0	0	0	0	0.126	5.71	0.128	33.68
e-cp3o		100	0	0	0	0	0	0	0.186	6.77	0.129	1.996
FDRSeg		0	0	0	0	0	2	98	0.042	7.02	0.11	1.56
NMCD		0	6	55	39	0	0	0	0.03	1.8	0.032	4.355
NOT		3	10	51	20	13	3	0	0.029	3.49	0.038	0.077
NOT HT	(M2)	0	3	52	**44**	1	0	0	0.023	**1.48**	**0.022**	0.136
NP-PELT		0	0	13	22	19	23	23	0.043	3.98	0.039	0.2
PELT		1	5	16	28	18	12	20	0.056	3.63	0.04	0.003
S3IB		26	18	23	21	9	3	0	0.058	4.21	0.054	0.299
SMUCE		1	9	10	22	24	6	28	0.05	5.49	0.074	0.016
WBS		2	3	24	7	22	11	31	0.058	4.49	0.046	0.143
B&P		0	0	100	0	0	0	0	0.221	3.67	0.132	53.919
NOT	(M3)	0	0	0	**97**	3	0	0	0.016	**1.05**	**0.054**	0.395
TF		0	0	0	0	0	0	100	0.019	8.36	0.221	46.891
B&P		0	0	9	**91**	0	0	0	0.082	2.85	0.143	61.857
NOT	(M4)	0	0	0	**98**	1	1	0	0.017	**1.29**	**0.07**	0.371
TF		0	0	0	0	0	0	100	0.018	4.41	0.151	48.119
B&P		0	0	0	**100**	0	0	0	0.018	**2.17**	**0.082**	118.05
NOT	(M5)	0	0	2	**90**	7	1	0	0.021	2.53	**0.086**	0.368
TF		0	0	0	0	0	0	100	0.026	5.98	0.26	59.006
e-cp3o		19	4	12	34	19	7	5	0.141	6.83	0.17	1.695
HSMUCE		100	0	0	0	0	0	0	0.098	12.68	0.212	0.052
NMCD		0	13	40	42	5	0	0	0.056	7.67	0.088	4.123
NOT	(M6)	0	3	11	**51**	23	9	3	0.08	5.09	**0.084**	0.463
NP-PELT		0	0	3	15	19	19	44	0.194	5.08	0.089	0.281
PELT		5	16	27	40	9	3	0	0.09	7.71	0.099	0.012
SegNeigh		0	0	7	26	28	20	19	0.094	**4.33**	**0.077**	17.3
B&P		0	0	0	**99**	1	0	0	0.022	**2.26**	**0.071**	28.876
NOT	(M7)	0	0	6	86	8	0	0	0.027	3.03	**0.078**	0.226
TF		0	0	0	0	0	0	100	0.049	23.29	0.442	42.538

Table 6. Distribution of $\hat{q}-q$ for data generated according to (3) with the noise term ε_t being a zero-mean unit-variance Gaussian AR(1) process with $\varphi = 0.3$ for various choices of f_t and σ_t given in Section A and competing methods listed in Section 5. Also, the average Mean-Square Error of the resulting estimate of the signal f_t, average Hausdorff distance d_H, average inverse V-measure d_V and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q}-q=0$ or smallest average of d_H or d_V, and those within 10% of the highest or lowest accordingly.

		$\hat{q}-q$										
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	$d_H \times 10^2$	d_V	time
B&P		0	0	0	**100**	0	0	0	0.088	**0.84**	**0.028**	1.361
e-cp3o		0	0	0	**100**	0	0	0	0.126	0.99	0.05	0.116
FDRSeg		0	0	0	1	1	4	94	0.199	5.59	0.128	0.07
NMCD		0	0	0	63	29	6	2	0.145	2.36	0.06	1.048
NOT		0	0	0	64	18	7	11	0.113	2.13	0.04	0.046
NOT HT	(M1)	0	0	0	78	19	2	1	0.104	1.67	0.036	0.058
NP-PELT		0	0	0	39	31	20	10	0.134	2.63	0.05	0.017
PELT		0	0	0	73	21	3	3	0.106	1.88	0.036	0.001
S3IB		0	0	0	73	22	3	2	0.102	1.79	0.034	0.069
SMUCE		0	0	0	56	30	10	4	0.136	2.52	0.059	0.053
WBS		0	0	0	63	20	7	10	0.11	2.18	0.038	0.072
B&P		100	0	0	0	0	0	0	0.136	6.67	0.14	30.394
e-cp3o		100	0	0	0	0	0	0	0.202	6.81	0.137	2.046
FDRSeg		0	0	0	0	0	0	100	0.121	8.87	0.209	1.401
NMCD		1	9	37	**35**	13	4	1	0.056	**2.92**	0.055	4.316
NOT		1	8	34	25	9	12	11	0.053	3.63	0.053	0.082
NOT HT	(M2)	5	14	39	24	8	7	3	0.056	3.34	0.056	0.136
NP-PELT		0	1	1	10	17	14	57	0.067	4.98	0.074	0.192
PELT		1	11	30	**38**	10	9	1	0.048	**2.73**	**0.045**	0.003
S3IB		11	27	39	20	3	0	0	0.05	3.1	**0.048**	0.34
SMUCE		0	12	36	26	21	3	2	0.057	4.45	0.066	0.015
WBS		2	10	29	27	11	12	9	0.052	3.41	0.051	0.141
B&P		0	0	91	9	0	0	0	0.245	4.37	0.147	53.676
NOT	(M3)	0	0	0	**96**	4	0	0	0.03	**1.51**	**0.07**	0.394
TF		0	0	0	0	0	0	100	0.465	9.08	0.519	46.654
B&P		0	1	25	74	0	0	0	0.136	3.74	0.159	61.576
NOT	(M4)	0	0	0	**97**	2	1	0	0.035	**2.03**	**0.095**	0.378
TF		0	0	0	0	0	0	100	0.479	5	0.462	47.875
B&P		0	0	0	**98**	2	0	0	0.04	**3.28**	**0.113**	117.832
NOT	(M5)	0	0	0	**89**	8	2	1	0.043	**3.55**	**0.115**	0.346
TF		0	0	0	0	0	0	100	0.218	6.24	0.461	56.926
e-cp3o		19	9	16	23	13	10	10	0.224	8.25	0.19	1.659
HSMUCE		65	30	5	0	0	0	0	0.117	12.78	0.196	0.05
NMCD		1	0	5	28	29	18	19	0.178	5.36	0.093	4.097
NOT	(M6)	0	2	23	**56**	13	5	1	0.123	5.29	0.074	0.455
NP-PELT		0	0	0	0	1	1	98	0.482	5.49	0.127	0.219
PELT		9	17	28	40	6	0	0	0.126	8.78	0.1	0.011
SegNeigh		0	0	2	39	24	23	12	0.12	**3.1**	**0.066**	17.242
B&P		0	0	2	**86**	11	1	0	0.045	4.13	0.101	28.884
NOT	(M7)	0	0	4	**89**	5	2	0	0.043	**3.39**	**0.089**	0.232
TF		0	0	0	0	0	0	100	0.11	24.39	0.537	42.602

F. Additional numerical experiments on the choice of M

F.1. Setup

We now elaborate on the effect of the choice of M, the number of randomly drawn sub-intervals. We focus on Scenario (S1) and consider the models based on variations of (M1). All models listed below have piecewise-constant f_t with equal-spaced change-points.

(M1-1) `teeth-1`: $T = 512$, $q = 1$ change-points at $\tau = 256$, with the corresponding jump sizes -2, $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$.

(M1-2) `teeth-2`: $T = 512$, $q = 3$ change-points at $\tau = 128, 256, 384$, with the corresponding jump sizes $-2, 2, -2$, $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$.

(M1-3) `teeth-3`: $T = 512$, $q = 7$ change-points at $\tau = 64, 128, \dots, 448$, with the corresponding jump sizes $-2, 2, -2, 2, -2$, $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$. Note that this model is the same as (M1) `teeth` listed in Section A.

(M1-4) `teeth-4`: $T = 512$, $q = 15$ change-points at $\tau = 32, 64, \dots, 480$, with the corresponding jump sizes $-2, 2, -2, \dots, -2$, $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$.

(M1-5) `teeth-5`: $T = 512$, $q = 31$ change-points at $\tau = 16, 32, \dots, 496$, with the corresponding jump sizes $-2, 2, -2, \dots, -2$, $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$.

(M1-6) `teeth-6`: $T = 512$, $q = 63$ change-points at $\tau = 8, 16, \dots, 504$, with the corresponding jump sizes $-2, 2, -2, \dots, -2$, $f_1 = 1$, $\sigma_t = 1$ for $t = 1, \dots, T$.

We take $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, run our NOT procedure using contrast function given by (6), and the threshold picked by the SIC, but with different $M = 10, 10^2, 10^3, 10^4, 10^5$.

F.2. Results

In Table 7, we report the performance of NOT with the SIC in terms of estimates of $\mathbb{P}(\hat{q} = q)$, $\mathbb{E}(\hat{q}/q)$, d_H and d_V after 500 realisations. Here d_H and d_V are, respectively, the scaled Hausdorff distance measure and the inverse V-measure, both given in Section 5.3.

We see that when there is only a small number of change-points in the signal, a moderate M would be sufficient for the purpose of identifying all the change-points. In this case, having a larger M will be more computationally intensive, but will not necessarily improve the performance of the NOT procedure. As we increase the number of change-points (while fixing T, as well as the size of the jumps, etc), we see that a larger M would be needed for satisfactory performance. For example, in Model (M1-4) (with $q = 15$), our procedure with $M = 10^4$ or 10^5 estimates the number of the change-points correctly more than 95% of the time, while this proportion reduces to 87% for $M = 10^3$, and virtually 0% for $M = 10^2$ or smaller. In cases like that, increasing M would be helpful.

Table 7. Performance of NOT with ζ_T picked via the SIC and different $M = 10, 10^2, 10^3, 10^4, 10^5$. Here data is generated according to (3) with the noise term ε_t being i.i.d. $\mathcal{N}(0,1)$ for various models given in Section F.1. The estimated values of $\mathbb{P}(\hat{q}=q)$, $\mathbb{E}(\hat{q}/q)$, d_H and d_V are reported, all calculated over 500 simulated data sets.

Model ($T=512$)	Measure	M				
		10	10^2	10^3	10^4	10^5
(M1-1): $q=1$	$\mathbb{P}(\hat{q}=q)$	0.978	0.978	0.968	0.972	0.976
(M1-2): $q=3$		0.418	0.988	0.982	0.976	0.978
(M1-3): $q=7$		0.004	0.566	0.976	0.972	0.974
(M1-4): $q=15$		0	0.006	0.872	0.958	0.956
(M1-5): $q=31$		0	0	0.002	0.45	0.424
(M1-6): $q=63$		0	0	0	0	0
(M1-1): $q=1$	$\mathbb{E}(\hat{q}/q)$	1.042	1.026	1.04	1.036	1.036
(M1-2): $q=3$		0.925	1.004	1.006	1.009	1.008
(M1-3): $q=7$		0.419	0.98	1.003	1.004	1.004
(M1-4): $q=15$		0.06	0.565	1.003	1.003	1.003
(M1-5): $q=31$		0.002	0.007	0.055	0.837	0.845
(M1-6): $q=63$		0	0	0	0	0
(M1-1): $q=1$	$d_H \times 10^2$	0.53	0.36	0.45	0.46	0.42
(M1-2): $q=3$		14.23	0.33	0.36	0.39	0.38
(M1-3): $q=7$		22.71	5.03	0.53	0.53	0.53
(M1-4): $q=15$		43.45	14.31	1.25	0.79	0.79
(M1-5): $q=31$		49.82	49.04	45.49	8.4	7.92
(M1-6): $q=63$		50	49.98	49.97	49.98	49.98
(M1-1): $q=1$	d_V	0.019	0.014	0.015	0.015	0.015
(M1-2): $q=3$		0.142	0.013	0.013	0.013	0.013
(M1-3): $q=7$		0.349	0.043	0.018	0.018	0.018
(M1-4): $q=15$		0.848	0.242	0.029	0.026	0.027
(M1-5): $q=31$		0.995	0.977	0.905	0.168	0.155
(M1-6): $q=63$		1	0.999	0.999	0.999	0.999

Table 8. Performance of NOT with ζ_T picked via the AIC and different $M = 10, 10^2, 10^3, 10^4, 10^5$. Here data is generated according to (3) with the noise term ε_t being i.i.d. $\mathcal{N}(0,1)$ for Model (M1-6) given in Section F.1. The estimated values of $\mathbb{P}(\hat{q} = q)$, $\mathbb{E}(\hat{q}/q)$, d_H and d_V are reported, all calculated over 500 simulated data sets.

	M	$\mathbb{P}(\hat{q} = q)$	$\mathbb{E}(\hat{q}/q)$	$d_H \times 10^2$	d_V
(M1-6) : $q = 63$	10	0	0.007	47.06	0.948
	10^2	0	0.041	37.29	0.776
	10^3	0	0.276	13.29	0.331
	10^4	0.08	0.937	1.76	0.063
	10^5	0.106	0.988	1.58	0.057

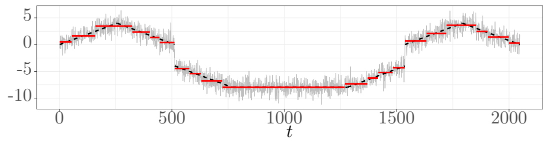
On the other hand, caution must be exercised for signals with an extremely large number of change-points, or the spacing of change-points to be highly non-homogeneous. For example, in Model (M1-6) (with $q = 63$) where jumps in the signal occur at every 8 observations (which is itself a difficult problem), having $M = 10^5$ or larger will not lead to any improvement of NOT with the SIC. However, we believe that this is partially due to the fact that the SIC penalty is no longer approperiate for this extreme scenario. One could alleviate the issue by using NOT with other less harsh penalty, or use methods designed to tackle frequent change-points such as Fryzlewicz (2018). For instance, by changing the SIC to the AIC in our procedure, we can see from Table 8 that with $M = 10^5$, $\mathbb{P}(\hat{q} = q)$ increases from 0% to around 10%. More importantly, there is huge improvement in terms of $\mathbb{E}(\hat{q}/q)$, d_H and d_V. In fact, a close inspection indicates that $\hat{q}/q \in [0.9, 1.1]$ more than 90% of the time.

G. More on model misspecification and model selection

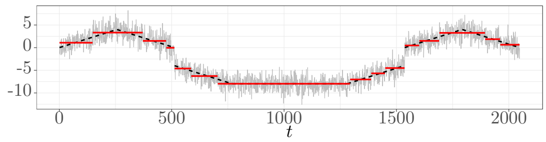
We have demonstrated that NOT is relatively robust against the misspecification in the distribution of ε_t, when the truth is either correlated or heavy-tailed. Now we investigate the case where the signal f_t is misspecified. In particular, we focus on the misspecification of the degree of the polynomials between consecutive change-points.

We simulate data according to (3) using the signal (M8) `smile` and noise of (a) i.i.d. $\mathcal{N}(0,1)$ and (b) i.i.d. $\mathcal{N}(0,2)$. Here the true signal is piecewise-linear but not necessarily continuous (i.e. from Scenario (S3)). We test NOT with the sSIC using contrast functions constructed from Scenarios (S1), (S3) and (S5), where the estimators are denoted by NOT_0, NOT_1 and NOT_2, respectively. Again we take $\alpha = 1$. Figure 9 shows a typical realisation of the estimates produced by NOT with different contrast functions, while Table 9 summarises the results.

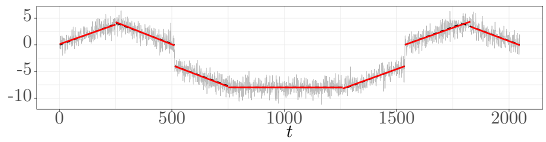
For NOT_0 (suitable for piecewise-constant signal), we see that unsurprisingly NOT_0 significantly overestimates the number of change-points q. This is due to the bias-variance tradeoff in the sSIC, where the bias term only approaches zero as the estimated number of change-points $\hat{q} \to \infty$. Nevertheless, we observe that the set of change-point estimates from NOT_0 typically includes the true change-points with jump, even though the construction of the contrast function (wrongly) assumes that the signal is piecewise-constant in the neighbourhood of these change-points. Furthermore, under the higher signal-to-noise ratio setting, NOT_2, which is designed for piecewise-quadratic signal, is able to estimate the number of change-points q correctly most of the time. However, since NOT_2 is overparameterised in this setting of Scenario (S3), it tends to perform slightly worse than NOT_1 in terms of both the MSE for the estimated signal, and the accuracy of the estimated locations of the change-points. Finally, under the lower signal-to-noise ratio setting, NOT_2 tends to underestimate the number of change-points, thanks to the bias-variance tradeoff in the sSIC. Nevertheless, as is illustrated in Figures 9f, the estimated f_t is quite close to the truth in terms of the ℓ_2 distance. These findings suggest that NOT could still provide valuable insights in certain misspecified circumstances.



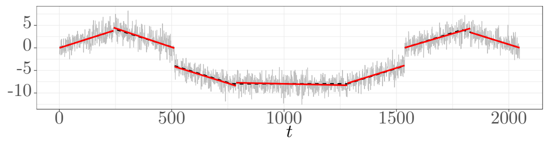
(a) NOT_0, $\mathcal{N}(0,1)$ noise



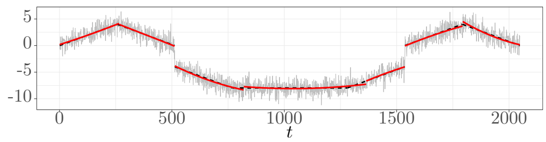
(b) NOT_0, $\mathcal{N}(0,2)$ noise



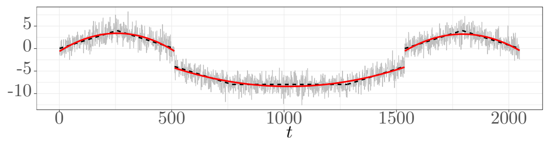
(c) NOT_1, $\mathcal{N}(0,1)$ noise



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



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



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

Fig. 9. Typical realisation of the estimates produced by different NOTs, with data generated from (M8) `smile`. Figure 9a– 9f: data series Y_t (thin grey), true signal f_t (dashed black), $\hat{f}_t$ being the LS estimate of f_t with the change-points estimated by NOT (thick red). Higher signal-to-noise ratio setting (with $\mathcal{N}(0,1)$ errors) in Figures 9a, 9c and 9e; lower signal-to-noise ratio setting (with $\mathcal{N}(0,2)$ errors) in Figures 9b, 9d and 9f. Here NOT_0, NOT_1 and NOT_2 denote estimates from NOT with sSIC using contrast functions constructed from Scenarios (S1), (S3) and (S5), respectively.

In the same example, we also demonstrate that one could empirically select the degree of the polynomial for the NOT's contrast function via sSIC. Denote the sSIC scores corresponding to the estimates from NOT_0, NOT_1 and NOT_2 by $\mathrm{sSIC}(\mathrm{NOT}_0)$, $\mathrm{sSIC}(\mathrm{NOT}_1)$ and $\mathrm{sSIC}(\mathrm{NOT}_2)$ respectively. We propose to pick the estimator produced by NOT_{i^*} with

$$i^* = \mathrm{argmin}_{i \in \{0,1,2\}} \mathrm{sSIC}(\mathrm{NOT}_i).$$

As shown in Table 9, empirical results suggest that we are able to select the correct order of the polynomial for our NOT approach using the sSIC (with $\alpha = 1$), especially when the signal-to-noise ratio is high.

Table 9. Distribution of $\hat{q}-q$ obtained by $\mathrm{NOT}_0, \mathrm{NOT}_1, \mathrm{NOT}_2$ for data generated according to (3) with the signal (M8) and the noise $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\mathcal{N}(0,2)$, the average Mean-Square Error of the resulting estimate of the signal over 100 simulations. The number of times each method selected by sSIC is also reported.

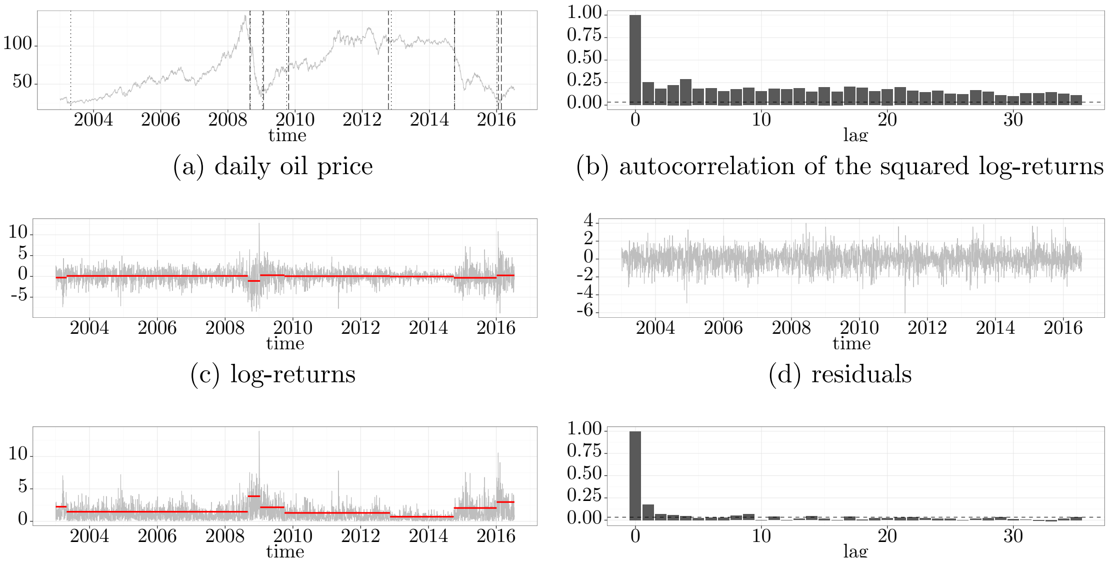
		$\hat{q}-q$								Number of times
Noise	Method	≤ -3	-2	-1	0	1	2	≥ 3	MSE	selected by sSIC
$\mathcal{N}(0,1)$	NOT_0	0	0	0	0	0	0	100	0.120	0
	NOT_1	0	0	0	**99**	1	0	0	**0.015**	**100**
	NOT_2	0	4	18	78	0	0	0	0.024	0
$\mathcal{N}(0,2)$	NOT_0	0	0	0	0	0	0	100	0.188	0
	NOT_1	0	0	0	**100**	0	0	0	**0.032**	**94**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

H. Additional real data example: OPEC Reference Basket oil price

We perform change-point analysis on the daily Organisation of the Petroleum Exporting Countries (OPEC) Reference Basket oil price from 1 January, 2003 to 15 July, 2016. The data were obtained from the OPEC database through the R package **Quandl** (McTaggart *et al.*, 2016). Instead of working with the raw price series, we analyse the log-returns series $Y_t = 100 \log(P_t/P_{t-1})$, where P_t denotes the daily oil price. One of the stylised facts of the financial time series data is that the autocorrelation of assets returns are weak, while squared returns tend to exhibit strong autocorrelation, which is the case for the oil price time series (see Figure 10b). This phenomenon can be possibly explained by the existence of the structural breaks in the mean and variance structure of the data series (Mikosch and Stărică, 2004; Fryzlewicz *et al.*, 2006). In this study, we apply NOT with the contrast function given by (14), which is designed to detect changes in both the mean and the volatility, as in Scenario (S4). For comparison, we also report change-points detected with the NMCD method of Zou *et al.* (2014).

We apply Algorithm 2 to compute the NOT solution path and choose the model achieving the lowest SIC given by (11), setting the number of intervals drawn $M = 10000$ and the maximum number of change-points $q_{max} = 25$. Computations for the solution path and model selection are performed using the R package **not** (Baranowski *et al.*, 2016). For the NMCD procedure, we use the `nmcd` routine from the R package **nmcdr** (Zou and Lancezhange, 2014), setting the maximum number of change-points to $q_{max} = 25$ as well.

Figure 10 illustrates the results of our analysis. The oil price time series and the locations of the change-points identified by NOT and NMCD can be seen in Figure 10a. Both methods discover 7 change-points, largely agreeing on their locations, in the sense that for 6 out of 7 features NOT detects, NMCD detects a change-point nearby. However, NMCD does not indicate any change-point around the first change-point identified by NOT on 29 April 2003. This date could potentially be related to the end of the 2003 invasion of Iraq, which initiated the upward trend in the oil price lasting almost ceaselessly until the beginning of the 2008–09 financial crisis. On the other hand, NMCD indicates two change-points in the first quarter of 2016, while NOT only finds one in that period. Table 10 lists the exact locations of the change-points detected by the two methods and the events that coincide with some of them. Figure 10f shows the autocorrelation function for the squared residuals obtained by subtracting the sample mean and dividing by the standard deviations from the data in each segment. It appears that there is little autocorrelation in the squares of the residuals, suggesting that Scenario (S4) fits the data reasonably well.



(a) daily oil price (b) autocorrelation of the squared log-returns

(c) log-returns (d) residuals

(e) centred log-returns (f) autocorrelation of the squared residuals

Fig. 10. Change-point analysis on the daily OPEC Reference Basket oil price in USD from 1 January, 2003 to 15 July, 2016. Figure 10a: price series P_t (thin grey), locations of the change-points detected with NOT (vertical dotted lines) and NMCD (vertical dashed lines). Figure 10b: autocorrelation function of Y_t^2. Figure 10c: log-returns $Y_t = 100\log(P_t/P_{t-1})$ (thin grey), the fitted piecewise-constant mean via NOT, $\hat{f}_t$ (thick red). Figure 10d: estimated residuals via NOT, $\hat{\varepsilon}_t = (Y_t - \hat{f}_t)/\hat{\sigma}_t$. Figure 10e: the centred log-returns $|Y_t - \hat{f}_t|$ (thin grey), fitted piecewise-constant volatility $\hat{\sigma}_t$ (thick red). Figure 10f: autocorrelation of $\hat{\varepsilon}_t^2$. The exact locations of the change-points detected via NOT are given in Table 10.

Table 10. Change-points detected using NOT and NMCD methods in the daily OPEC Reference Basket oil price data from 1 January 2003 to 15 July 2016, with some of them dated.

NOT	NMCD	Event that coincides
29 April 2003	N/A	Invasion of Iraq
1 September 2008	28 August 2008	critical stage of the subprime mortgage crisis
27 January 2009	22 January 2009	tensions in the Gaza Strip
1 October 2009	23 October 2009	
12 November 2012	12 October 2012	beginning of a period of low volatility
30 September 2014	1 October 2014	
5 January 2016	21 January 2016	beginning of a sell-off leading the price to 12-year low
N/A	22 February 2016	

I. Proofs

I.1. Some useful lemmas

I.1.1. ***The piecewise-constant case***

Lemma 1. *Let* $g(x,y) = \frac{xy}{x+y}$ *and suppose that* $\min(x,y) > 0$. *Then*

$$g(x,y) \geq \frac{1}{2}\min(x,y).$$

Proof. Without loss of generality, assume that $x \geq y$. Then $g(x,y) \geq \frac{xy}{2x} \geq y/2 = \min(x,y)/2$. □

Lemma 2. *Suppose* $\mathbf{f} = (f_1,\ldots,f_T)'$ *is piecewise-constant vector as in Scenario (S1), and* $\tau_1,\ldots,\tau_q$ *are the locations of the change-points. Suppose* $0 \leq s < e \leq T$, *such that* $\tau_{j-1} \leq s < \tau_j < e \leq \tau_{j+1}$ *for some* $j = 1\ldots,q$. *Let* $\eta = \min\{\tau_j - s, e - \tau_j\}$ *and* $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$. *Then*

$$\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}) = \max_{s<b<e} \mathcal{C}^{b}_{(s,e]}(\mathbf{f}) \begin{cases} \geq \frac{1}{\sqrt{2}}\eta^{1/2}\Delta_j^{\mathbf{f}}, \\ \leq \eta^{1/2}\Delta_j^{\mathbf{f}}. \end{cases}$$

Proof. For any $s < b < e$, by simple algebra, we have

$$\mathcal{C}^{b}_{(s,e]}(\mathbf{f}) = \begin{cases} \sqrt{\frac{b-s}{(e-s)(e-b)}}(e-\tau_j)|f_{\tau_j+1} - f_{\tau_j}|, & b \leq \tau_j; \\ \sqrt{\frac{(\tau_j-s)(e-\tau_j)}{e-s}}|f_{\tau_j+1} - f_{\tau_j}|, & b = \tau_j; \\ \sqrt{\frac{e-b}{(e-s)(b-s)}}(\tau_j - s)|f_{\tau_j+1} - f_{\tau_j}|, & b \geq \tau_j. \end{cases} \tag{15}$$

Now $\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}) = \max_{s<b<e} \mathcal{C}^{b}_{(s,e]}(\mathbf{f})$ follows from the fact that $\mathcal{C}^{b}_{(s,e]}(\mathbf{f})$ is increasing (as a function of b) for $s < b \leq \tau_j$ and decreasing for $\tau_j \leq b < e$. To prove the lower bound, we set $\eta_L = \tau_j - s$ and $\eta_R = e - \tau_j$ and observe that $\eta_L \geq \eta$ and $\eta_R \geq \eta$. Therefore by Lemma 1, $\frac{\eta_L\eta_R}{\eta_L+\eta_R} \geq \frac{\eta}{2}$. Noting that $e - s = \eta_L + \eta_R$ we bound

$$\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}) = \sqrt{\frac{(\tau_j - s)(e-\tau_j)}{e-s}}|f_{\tau_j+1} - f_{\tau_j}| \begin{cases} \geq (\eta/2)^{1/2}\Delta_j^{\mathbf{f}}; \\ \leq \eta^{1/2}\Delta_j^{\mathbf{f}}. \end{cases}$$

which completes the proof. □

Lemma 3. *Suppose* $\mathbf{f} = (f_1,\ldots,f_T)'$ *is piecewise-constant vector as in Scenario (S1), and* $\tau_1,\ldots,\tau_q$ *are the locations of the change-points. Suppose* $0 \leq s < e \leq T$ *such that* $\tau_{j-1} \leq s \leq \tau_j$ *and* $\tau_{j+1} \leq e \leq \tau_{j+2}$ *for some* $j = 1\ldots,q-1$. *Then*

$$\max_{s<b<e} \mathcal{C}^{b}_{(s,e]}(\mathbf{f}) \leq (\tau_j - s)^{1/2}\Delta_j^{\mathbf{f}} + (e - \tau_{j+1})^{1/2}\Delta_{j+1}^{\mathbf{f}}$$

where $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$.

Proof. Suppose that $b^* = \mathrm{argmax}_{s<b<e}\, \mathcal{C}^{b}_{(s,e]}(\mathbf{f})$. Then

$$\begin{aligned} 0 &\leq \|\mathbf{f} - \langle \mathbf{f}, \boldsymbol{\psi}^{b^*}_{(s,e]}\rangle\boldsymbol{\psi}^{b^*}_{(s,e]} - \langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle\mathbf{1}_{(s,e]}\|^2 = \|\mathbf{f} - \langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle\mathbf{1}_{(s,e]}\|^2 - \langle \mathbf{f}, \boldsymbol{\psi}^{b^*}_{(s,e]}\rangle^2 \\ &\leq \|\mathbf{f} - f_{\tau_j+1}\sqrt{e-s}\mathbf{1}_{(s,e]}\|^2 - \langle \mathbf{f}, \boldsymbol{\psi}^{b^*}_{(s,e]}\rangle^2 \\ &= (\tau_j - s)(\Delta_j^{\mathbf{f}})^2 + (e - \tau_{j+1})(\Delta_{j+1}^{\mathbf{f}})^2 - \Big(\max_{s<b<e} \mathcal{C}^{b}_{(s,e]}(\mathbf{f})\Big)^2. \end{aligned}$$

It then follows that

$$\max_{s<b<e} \mathcal{C}^b_{(s,e]}(\mathbf{f}) \leq \sqrt{(\tau_j - s)(\Delta^{\mathbf{f}}_j)^2 + (e - \tau_{j+1})(\Delta^{\mathbf{f}}_{j+1})^2} \leq (\tau_j - s)^{1/2}\Delta^{\mathbf{f}}_j + (e - \tau_{j+1})^{1/2}\Delta^{\mathbf{f}}_{j+1}.$$

□

Lemma 4. *Suppose $\mathbf{f} = (f_1, \ldots, f_T)'$ is piecewise-constant vector as in Scenario (S1). Pick any interval $(s, e] \subset (0, T]$ such that $[s + 1, e - 1]$ contains exactly one change-point τ_j. Let $\rho = |\tau_j - b|$, $\Delta^{\mathbf{f}}_j = |f_{\tau_j+1} - f_{\tau_j}|$, $\eta_L = \tau_j - s$ and $\eta_R = e - \tau_j$. Then,*

$$\|\boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle\|_2^2 = (\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2.$$

Moreover,

(a) for any $\tau_j \leq b < e$, $(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2 = \frac{\rho\, \eta_L}{\rho + \eta_L}(\Delta^{\mathbf{f}}_j)^2$;

(b) for any $s < b \leq \tau_j$, $(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2 = \frac{\rho\, \eta_R}{\rho + \eta_R}(\Delta^{\mathbf{f}}_j)^2$.

Proof. First, we note that since there is only one change-point in $[s+1, e-1]$, the restriction of $\mathbf{f}$ on $(s, e]$, i.e. $\mathbf{f}|_{(s,e]} = (0, \ldots, 0, f_{s+1}, \ldots, f_e, 0, \ldots, 0)'$ can be decomposed into

$$\mathbf{f}|_{(s,e]} = \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle + \mathbf{1}_{(s,e]}\langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle,$$

where we also used the fact that $\boldsymbol{\psi}^{\tau_j}_{(s,e]}$ and $\mathbf{1}_{(s,e]}$ are orthonormal. Note that $\boldsymbol{\psi}^b_{(s,e]}$ and $\mathbf{1}_{(s,e]}$ are also orthonormal, it follows that

$$\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle = \langle \mathbf{f}|_{(s,e]}, \boldsymbol{\psi}^b_{(s,e]}\rangle = \left\langle \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle + \mathbf{1}_{(s,e]}\langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle, \boldsymbol{\psi}^b_{(s,e]}\right\rangle = \langle \boldsymbol{\psi}^{\tau_j}_{(s,e]}, \boldsymbol{\psi}^b_{(s,e]}\rangle\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle.$$

Therefore,

$$\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle^2 = \langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle\langle \boldsymbol{\psi}^{\tau_j}_{(s,e]}, \boldsymbol{\psi}^b_{(s,e]}\rangle\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle,$$

and thus

$$\begin{aligned}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle^2 - \langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle^2 &= \langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle^2 + \langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle^2 - 2\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle\langle \boldsymbol{\psi}^{\tau_j}_{(s,e]}, \boldsymbol{\psi}^b_{(s,e]}\rangle\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle \\ &= \|\boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle\|_2^2.\end{aligned}$$

Here in the above final step, we used the fact that $\|\boldsymbol{\psi}^{\tau_j}_{(s,e]}\|_2^2 = \|\boldsymbol{\psi}^b_{(s,e]}\|_2^2 = 1$.

Second, for the sake of brevity, we only prove the case of $b \geq \tau_j$. Let $l = e - s$, $x = b - s$, and thus $\rho = x - \eta_L$. Using (15), we get

$$\begin{aligned}(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2 &= \left(\frac{\eta_L(l - \eta_L)}{l} - \frac{\eta_L^2(l - x)}{lx}\right)|f_{\tau_j+1} - f_{\tau_j}|^2 \\ &= \frac{\eta_L(x - \eta_L)}{x}(\Delta^{\mathbf{f}}_j)^2 = \left(\frac{\rho\eta_L}{\eta_L + \rho}\right)(\Delta^{\mathbf{f}}_j)^2.\end{aligned}$$

□

I.1.2. ***The piecewise-linear continuous case***

Lemma 5. *Suppose* $\mathbf{f} = (f_1, \ldots, f_T)'$ *is piecewise-linear vector as in Scenario (S2), and* $\tau_1, \ldots, \tau_q$ *are the locations of the change-points. Suppose* $0 \le s < e \le T$*, such that* $\tau_{j-1} \le s+1 < \tau_j < e \le \tau_{j+1}$ *for some* $j = 1 \ldots, q$*. Let* $\eta = \min\{\tau_j - s - 1, e - \tau_j\}$ *and* $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j - 1} - f_{\tau_j+1}|$*. Then*

$$\mathcal{C}_{(s,e]}^{\tau_j}(\mathbf{f}) = \max_{s+1<b<e} \mathcal{C}_{(s,e]}^{b}(\mathbf{f}) \begin{cases} \ge \frac{1}{\sqrt{24}}\eta^{3/2}\Delta_j^{\mathbf{f}}, \\ \le \frac{1}{\sqrt{3}}(\eta+1)^{3/2}\Delta_j^{\mathbf{f}}. \end{cases}$$

Proof. First, we show that $\mathcal{C}_{(s,e]}^{b}(\mathbf{f})$ is maximised at $b = \tau_j$. Using the notation from the proof of Lemma 4, we have that

$$\mathbf{f}|_{(s,e]} = \boldsymbol{\phi}_{(s,e]}^{\tau_j}\langle \mathbf{f}, \boldsymbol{\phi}_{(s,e]}^{\tau_j}\rangle + \boldsymbol{\gamma}_{(s,e]}\langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle + \mathbf{1}_{(s,e]}\langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle.$$

Therefore, it follows that

$$\|\mathbf{f}|_{(s,e]}\|_2^2 = \langle \mathbf{f}, \boldsymbol{\phi}_{(s,e]}^{\tau_j}\rangle^2 + \langle \mathbf{f}, \boldsymbol{\gamma}_{(s,e]}\rangle^2 + \langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle^2. \tag{16}$$

For any $b \in \{s+2, \ldots, \tau_j - 1, \tau_j + 1, \ldots, e-1\}$, it is clear that $\mathbf{f}|_{(s,e]}$ does not lie in the span of $\boldsymbol{\phi}_{(s,e]}^{b}$, $\boldsymbol{\gamma}_{(s,e]}$ and $\mathbf{1}_{(s,e]}$. Consequently, by projecting $\mathbf{f}|_{(s,e]}$ onto these three bases, we have that

$$\|\mathbf{f}|_{(s,e]}\|^2 > \langle \mathbf{f}, \boldsymbol{\phi}_{(s,e]}^{b}\rangle^2 + \langle \mathbf{f}, \boldsymbol{\gamma}_{(s,e]}\rangle^2 + \langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle^2. \tag{17}$$

Comparing (17) with (16) entails that $|\langle \mathbf{f}, \boldsymbol{\phi}_{(s,e]}^{\tau_j}\rangle| > \left|\langle \mathbf{f}, \boldsymbol{\phi}_{(s,e]}^{b}\rangle\right|$ for any $b \ne \tau_j$.

Secondly, set $\eta_L = \tau_j - s - 1$ and $\eta_R = e - \tau_j$. After some calculation, we get that

$$\mathcal{C}_{(s,e]}^{\tau_j}(\mathbf{f}) = \left\{\frac{\eta_L(\eta_L+1)\eta_R(\eta_R+1)(2\eta_L\eta_R + \eta_L + \eta_R + 2)}{6l(l^2-1)}\right\}\Delta_j^{\mathbf{f}},$$

where $l = e - s$. Also, we have $\eta_L \ge \eta$, $\eta_R \ge \eta$ and $l = \eta_L + \eta_R + 1$. To prove the lower bound, we observe that

$$\begin{aligned} &\left\{\frac{\eta_L(\eta_L+1)\eta_R(\eta_R+1)(2\eta_L\eta_R + \eta_L + \eta_R + 2)}{6l(l^2-1)}\right\} \\ &\ge \left\{\frac{1}{6}\frac{(\eta_L+1)\eta_R}{l}\frac{\eta_L(\eta_R+1)}{l}\frac{2\min(\eta_L,\eta_R)\{\max(\eta_L,\eta_R)+1\}}{l}\right\} \ge \left\{\frac{\eta^3}{24}\right\}, \end{aligned}$$

where the last inequality is obtained applying Lemma 1 three times. For the upper bound, we notice that $2\eta_L\eta_R + \eta_L + \eta_R + 2 \le 2(\eta_L+1)(\eta_R+1)$ which implies

$$\left\{\frac{\eta_L(\eta_L+1)\eta_R(\eta_R+1)(2\eta_L\eta_R + \eta_L + \eta_R + 2)}{6l(l^2-1)}\right\} \le \left\{\frac{1}{3}\frac{\eta_L\eta_R(\eta_L+1)^2(\eta_R+1)^2}{(l-1)l^2}\right\} \le \left\{\frac{(\eta+1)^3}{3}\right\}.$$

□

Lemma 6. *Suppose* $\mathbf{f} = (f_1, \ldots, f_T)'$ *is piecewise-linear vector as in Scenario (S2), and* $\tau_1, \ldots, \tau_q$ *are the locations of the change-points. Suppose* $0 \le s < e \le T$ *such that* $\tau_{j-1} \le s+1 \le \tau_j$ *and* $\tau_{j+1} \le e \le \tau_{j+2}$ *for some* $j = 1 \ldots, q-1$*. Then,*

$$\max_{s+1<b<e} \mathcal{C}_{(s,e]}^{b}(\mathbf{f}) \le \frac{1}{\sqrt{3}}(\tau_j - s)^{3/2}\Delta_j^{\mathbf{f}} + \frac{1}{\sqrt{3}}(e - \tau_{j+1} + 1)^{3/2}\Delta_{j+1}^{\mathbf{f}}$$

and

$$\max_{s+1<b<e} \mathcal{C}^b_{(s,e]}(\mathbf{f}) \le (\tau_j - s - 1)^{3/2}\Delta^{\mathbf{f}}_j + (e-\tau_{j+1})^{3/2}\Delta^{\mathbf{f}}_{j+1},$$

where $\Delta^{\mathbf{f}}_j = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$.

Proof. Suppose that $b^* = \mathrm{argmax}_{s\le b\le e}\, \mathcal{C}^b_{(s,e]}(\mathbf{f})$. Then

$$\begin{aligned}
0 &\le \|\mathbf{f}|_{(s,e]} - \langle \mathbf{f}, \boldsymbol{\phi}^{b^*}_{(s,e]}\rangle \boldsymbol{\phi}^{b^*}_{(s,e]} - \langle \mathbf{f}, \boldsymbol{\gamma}_{(s,e]}\rangle \boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle \mathbf{1}_{(s,e]}\|^2 \\
&= \|\mathbf{f}|_{(s,e]} - \langle \mathbf{f}, \boldsymbol{\gamma}_{(s,e]}\rangle \boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle \mathbf{1}_{(s,e]}\|^2 - \langle \mathbf{f}, \boldsymbol{\phi}^{b^*}_{(s,e]}\rangle^2 \\
&= \frac{1}{6}(\tau_j - s - 1)(\tau_j - s)(2\tau_j - 2s - 1)(\Delta^{\mathbf{f}}_j)^2 + \frac{1}{6}(e - \tau_{j+1})(e - \tau_{j+1} + 1)(2e - 2\tau_{j+1} + 1)(\Delta^{\mathbf{f}}_{j+1})^2 \\
&\quad - \Big(\max_{s+1<b<e} \mathcal{C}^b_{(s,e]}(\mathbf{f})\Big)^2.
\end{aligned}$$

It then follows that

$$\begin{aligned}
\max_{s+1<b<e} \mathcal{C}^b_{(s,e]}(\mathbf{f}) &\le \Big\{(\tau_j - s)^3(\Delta^{\mathbf{f}}_j)^2/3 + (e - \tau_{j+1}+1)^3(\Delta^{\mathbf{f}}_{j+1})^2/3\Big\}^{1/2} \\
&\le \frac{1}{\sqrt{3}}(\tau_j - s)^{3/2}\Delta^{\mathbf{f}}_j + \frac{1}{\sqrt{3}}(e - \tau_{j+1}+1)^{3/2}\Delta^{\mathbf{f}}_{j+1}.
\end{aligned}$$

For the second claim, we note that $(\tau_j - s)(2\tau_j - 2s - 1) \le 6(\tau_j - s - 1)^2$ and $(e - \tau_{j+1} + 1)(2e - 2\tau_{j+1} + 1) \le 6(e - \tau_{j+1})^2$, so

$$\begin{aligned}
\max_{s+1<b<e} \mathcal{C}^b_{(s,e]}(\mathbf{f}) &\le \Big\{(\tau_j - s - 1)^3(\Delta^{\mathbf{f}}_j)^2 + (e - \tau_{j+1})^3(\Delta^{\mathbf{f}}_{j+1})^2\Big\}^{1/2} \\
&\le (\tau_j - s - 1)^{3/2}\Delta^{\mathbf{f}}_j + (e - \tau_{j+1})^{3/2}\Delta^{\mathbf{f}}_{j+1}.
\end{aligned}$$

□

Lemma 7. *Suppose* $\mathbf{f} = (f_1, \ldots, f_T)'$ *is piecewise-linear vector as in Scenario (S2), and* $\tau_1, \ldots, \tau_q$ *are the locations of the change-points. Suppose* $0 \le s < e \le T$, *such that* $\tau_{j-1} \le s+1 < \tau_j < e \le \tau_{j+1}$ *for some* $j = 1 \ldots, q$. *Let* $\rho = |\tau_j - b|$, $\eta_L = \tau_j - s - 1$, $\eta_R = e - \tau_j$ *and* $\Delta^{\mathbf{f}}_j = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$. *Then,*

$$\|\boldsymbol{\phi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^b_{(s,e]}\rangle - \boldsymbol{\phi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^{\tau_j}_{(s,e]}\rangle\|_2^2 = (\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2. \tag{18}$$

Moreover,

(a) for any $\tau_j \le b < e$, $(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2 \ge \frac{1}{63}\min(\rho, \eta_L)^3(\Delta^{\mathbf{f}}_j)^2$;

(b) for any $s+1 < b \le \tau_j$, $(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{f}))^2 \ge \frac{1}{63}\min(\rho, \eta_R)^3(\Delta^{\mathbf{f}}_j)^2$.

Proof. The proof of (18) is very similar to that shown in Lemma 4, so is omitted for brevity. In the following, we only deal with the case of $\tau_j \le b < e$. Note that

$$\begin{aligned}
&\|\boldsymbol{\phi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^b_{(s,e]}\rangle - \boldsymbol{\phi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^{\tau_j}_{(s,e]}\rangle\|_2^2 \\
&= \big\|\boldsymbol{\phi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^b_{(s,e]}\rangle + \boldsymbol{\gamma}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\gamma}_{(s,e]}\rangle + \mathbf{1}_{(s,e]}\langle \mathbf{f}, \mathbf{1}_{(s,e]}\rangle - \mathbf{f}|_{(s,e]}\big\|_2^2 \\
&\ge \min_{a_0,a_1\in\mathbb{R}} \big\|\mathbf{f}|_{(s,b]} - a_0\mathbf{1}_{(s,b]} - a_1\boldsymbol{\gamma}_{(s,b]}\big\|_2^2 + \min_{a_0,a_1\in\mathbb{R}} \big\|\mathbf{f}|_{(b,e]} - a_0\mathbf{1}_{(b,e]} - a_1\boldsymbol{\gamma}_{(b,e]}\big\|_2^2 \\
&\ge \min_{a_0,a_1\in\mathbb{R}} \big\|\mathbf{f}|_{(s,b]} - a_0\mathbf{1}_{(s,b]} - a_1\boldsymbol{\gamma}_{(s,b]}\big\|_2^2.
\end{aligned}$$

Recalling the definitions of $\alpha^{\tau_j}_{(s,b]}$ and $\beta^{\tau_j}_{(s,b]}$ in (7), and writing $d = b - s$. After some calculations (similar to what has already been carried out in deriving $\boldsymbol{\phi}^b_{(s,e]}$, as demonstrated in Section B), we obtain that

$$
\begin{aligned}
&\min_{a_0,a_1\in\mathbb{R}} \left\|\mathbf{f}|_{(s,b]} - a_0\mathbf{1}_{(s,b]} - a_1\boldsymbol{\gamma}_{(s,b]}\right\|_2^2 \\
&= \left[(3\eta_L + \rho + 2)\alpha^{\tau_j}_{(s,b]}\beta^{\tau_j}_{(s,b]} + (3\rho + \eta_L + 2)\alpha^{\tau_j}_{(s,b]}(\beta^{\tau_j}_{(s,b]})^{-1}\right]^{-2}(\Delta^{\mathbf{f}}_j)^2 \\
&= \frac{1}{6}(\Delta^{\mathbf{f}}_j)^2 d(d^2-1)\big[1 + \rho\eta_L + (\rho+1)(\eta_L+1)\big]\times \\
&\qquad \left[(d+2\eta_L+1)^2\frac{\rho(\rho+1)}{\eta_L(\eta_L+1)} + (d+2\rho+1)^2\frac{\eta_L(\eta_L+1)}{\rho(\rho+1)} + 2(d+2\eta_L+1)(d+2\rho+1)\right]^{-1}.
\end{aligned}
$$

Notice that the above equation is symmetric with respect to η_L and ρ. Without loss of generality, here we proceed by assuming that $\eta_L \geq \rho$. Since $(d+2\eta_L+1)+(d+2\rho+1) = 4d$, it follows that $(d + 2\eta_L + 1)(d + 2\rho + 1) \leq 4d^2$. Therefore,

$$
\begin{aligned}
&\min_{a_0,a_1\in\mathbb{R}} \left\|\mathbf{f}|_{(s,b]} - a_0\mathbf{1}_{(s,b]} - a_1\boldsymbol{\gamma}_{(s,b]}\right\|_2^2 \\
&\geq \frac{1}{6}(\Delta^{\mathbf{f}}_j)^2 d(d^2-1)[2(\eta_L+1)\rho]\left[(3d)^2 + (2d)^2\frac{(\eta_L+1)^2}{\rho^2} + 8d^2\right]^{-1} \\
&\geq \frac{1}{6}(\Delta^{\mathbf{f}}_j)^2 d^2(d-1)[2(\eta_L+1)\rho]\left[21d^2\frac{(\eta_L+1)^2}{\rho^2}\right]^{-1} \geq \frac{1}{63}\rho^3(\Delta^{\mathbf{f}}_j)^2,
\end{aligned}
$$

where in the last step, we used the fact that $\frac{d-1}{\eta_L+1} \geq 1$ for $\rho \geq 1$ (and note that the last above-displayed equation also holds if $\rho = 0$).

Finally, we remark that the case of $s + 1 < b \leq \tau_j$ can be handled in a similar manner by symmetry. □

Lemma 8. *Suppose $\mathbf{f} = (f_1, \ldots, f_T)'$ is piecewise-linear vector as in Scenario (S2), and $\tau_1, \ldots, \tau_q$ are the locations of the change-points. Suppose $0 \leq s < e \leq T$, such that $\tau_{j-1} \leq s + 1 < \tau_j < e \leq \tau_{j+1}$ for some $j = 1 \ldots, q$. Let $\rho = |\tau_j - b|$, $\eta_L = \tau_j - s - 1$, $\eta_R = e - \tau_j$ and $\Delta^{\mathbf{f}}_j = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$. Then, for any b satisfying $\tau_j - \min(\eta_L, \eta_R)/2 < b < \tau_j + \min(\eta_L, \eta_R)/2$, we have that*

$$
(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^{b}_{(s,e]}(\mathbf{f}))^2 \geq \frac{(\Delta^{\mathbf{f}}_j)^2}{48}\big\{\min(\eta_L, \eta_R) - 1\big\}\rho^2.
$$

Proof. Here we focus on the scenario where $b > \tau_j$. By Lemma 7,

$$
\begin{aligned}
(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^{b}_{(s,e]}(\mathbf{f}))^2 &= \|\boldsymbol{\phi}^b_{s,e}\langle \mathbf{f}, \boldsymbol{\phi}^b_{(s,e]}\rangle - \boldsymbol{\phi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^{\tau_j}_{(s,e]}\rangle\|_2^2 \\
&= \min_{a_0,a_1,a_2\in\mathbb{R}} \left\|\mathbf{f}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]} - a_2\boldsymbol{\phi}^b_{(s,e]}\right\|_2^2 \\
&= (\Delta^{\mathbf{f}}_j)^2 \min_{a_0,a_1,a_2\in\mathbb{R}} \left\|\tilde{\mathbf{f}}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]} - a_2\boldsymbol{\phi}^b_{(s,e]}\right\|_2^2,
\end{aligned}
$$

where $\tilde{\mathbf{f}}|_{(s,e]} := (0, \ldots, 0, 1, \ldots, e - \tau_j, 0, \ldots, 0)'$, in which "1" appears at the $(\tau_j + 1)$-th position. In the following, our aim is to bound the residual sum of squares resulted from fitting $\tilde{\mathbf{f}}|_{(s,e]}$ via a piecewise-linear and continuous function with only one kink at b over

$(s, e]$. Assuming that the fitted value of this vector at the b-th position is m, then, we have that

$$\min_{a_0,a_1,a_2\in\mathbb{R}} \big\|\tilde{\mathbf{f}}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]} - a_2\boldsymbol{\phi}^b_{(s,e]}\big\|_2^2$$
$$\geq \Big(\frac{2m}{\eta_L + 2\rho}\Big)^2 \times \frac{1}{6}\Big(\frac{\eta_L - 1}{2}\Big)\Big(\frac{\eta_L + 1}{2}\Big)\eta_L + \Big\{\frac{2(\rho - m)}{e - b}\Big\}^2 \times \frac{1}{6}\Big(\frac{e - b - 1}{2}\Big)\Big(\frac{e - b + 1}{2}\Big)(e - b).$$

Since $b < \tau_j + \eta_R/2$, it follows that $e - b > \eta_R/2$, and thus $e - b - 1 \geq (\eta_R - 1)/2$. Moreover, the fact of $\rho < \min(\eta_L, \eta_R)/2$ yields $\eta_L + 2\rho \leq 2\eta_L$. Plugging these two inequalities into the previous equation, we have that

$$\min_{a_0,a_1,a_2\in\mathbb{R}} \big\|\tilde{\mathbf{f}}|_{(s,e]} - a_0\mathbf{1}_{(s,e]} - a_1\boldsymbol{\gamma}_{(s,e]} - a_2\boldsymbol{\phi}^b_{(s,e]}\big\|_2^2$$
$$\geq m^2\frac{\eta_L - 1}{24} + (\rho - m)^2\frac{\eta_R - 1}{12} \geq \frac{1}{2}\min\Big(\frac{\eta_L - 1}{24}, \frac{\eta_R - 1}{12}\Big)\rho^2$$

Consequently,

$$(\mathcal{C}^{\tau_j}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^{b}_{(s,e]}(\mathbf{f}))^2 \geq \frac{(\Delta^{\mathbf{f}}_j)^2}{48}\big\{\min(\eta_L, \eta_R) - 1\big\}\rho^2.$$

By symmetry, the scenario of $b < \tau_j$ can be dealt with in a similar fashion. Finally, we remark that the constants here are not sharp, as we will only use this lemma to establish rate-type results later. □

I.2. Proof of Theorem 1

Here we informally discuss our proof strategy, which could be generalised to other scenarios.

- Intuitively speaking, lemmas from Appendix I.1 deal with noiseless versions of the change-point estimation problems. In order to apply these results to show the consistency of estimated number of change-points, we need to control $\|\mathcal{C}^b_{(s,e]}(\mathbf{Y}) - \mathcal{C}^b_{(s,e]}(\mathbf{f})\|$ for every tuple (s, e, b), which can be achieved using Bonferroni in Step One.
- Note that for any fixed left-open and right-closed interval with start-point s and end-point e, to decide whether b_1 or b_2 is a more suitable change-point candidate inside this interval, we only need to look at the value of $\mathcal{C}^{b_1}_{(s,e]}(\mathbf{Y}) - \mathcal{C}^{b_2}_{(s,e]}(\mathbf{Y})$. Therefore, when establishing the convergence rate of the estimated change-point location , we control the distance between $\mathcal{C}^{b_1}_{(s,e]}(\mathbf{Y}) - \mathcal{C}^{b_2}_{(s,e]}(\mathbf{Y})$ and its noiseless analogue $\mathcal{C}^{b_1}_{(s,e]}(\mathbf{f}) - \mathcal{C}^{b_2}_{(s,e]}(\mathbf{f})$ (after proper normalisation) for all tuples (s, e, b_1, b_2) in Step Two.
- In Step Three, we show that given a properly chosen threshold and a large enough M, both bounds in Step One and Step Two hold, and for each change-point τ_j, there exists an interval from F^M_T that contains only this change-point and both its start- and end-points are sufficiently far away from other change-points. Since we are dealing with the narrowest-over-threshold intervals, the actual intervals that our NOT algorithm pick must have length no longer than the ones we considered in Step Three, thus could only contain precisely one change-point.

- So in Step Four, it suffices to investigate a single change-point detection problem, where we can use lemmas from Appendix I.1 and the bound in Step Two to establish the convergence rate for its location estimation.

- Finally, in Step Five, we show that after detecting all the change-points, the NOT algorithm stops with no further detection. This is because the remaining elements $(s,e] \in F_T^M$ to be considered either have no change-point inside, or have one/two change-points that are very close to its start- or/and end- points, thus their corresponding $\max_b \mathcal{C}^b_{(s,e]}(\mathbf{Y})$ cannot exceed the given threshold in views of the property of its noiseless analogue and the bound from Step One.

Now we proceed to the technical details.

Proof. We shall prove the following more specific result, which in turn implies (9).

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\ldots,q}\left(|\hat{\tau}_j - \tau_j|(\Delta_j^{\mathbf{f}})^2\right) \le C_3 \log T\right) \ge 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1-\delta_T^2 T^{-2}/36)^M, \tag{19}$$

Step One.
Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)'$ and $\lambda_T = \sqrt{8\log T}$. Define the set

$$A_T = \left\{\max_{s,b,e:0\le s<b<e\le T} |\mathcal{C}^b_{(s,e]}(\boldsymbol{\varepsilon})| \le \lambda_T\right\}.$$

Note that for any $0 \le s < b < e \le T$, $\mathcal{C}^b_{(s,e]}(\boldsymbol{\varepsilon})$ follows a standard normal distribution. Therefore, using the Bonferroni bound, we get

$$\mathbb{P}(A_T^c) \le \frac{T^3}{6}\frac{2e^{-(\sqrt{8\log T})^2/2}}{\sqrt{8\log T}\sqrt{2\pi}} \le \frac{T^{-1}}{12\sqrt{\pi}}.$$

Moreover, because $\mathcal{C}^b_{(s,e]}(\mathbf{Y}) - \mathcal{C}^b_{(s,e]}(\mathbf{f}) = \mathcal{C}^b_{(s,e]}(\boldsymbol{\varepsilon})$, so A_T also implies that

$$\left\{\max_{s,b,e:0\le s<b<e\le T} |\mathcal{C}^b_{(s,e]}(\mathbf{Y}) - \mathcal{C}^b_{(s,e]}(\mathbf{f})| \le \lambda_T\right\}.$$

We remark that though the constant in λ_T (i.e. $\sqrt{8}$) does not appear sharp (as it is rooted in the simple Bonferroni bound), it is sufficient for our purpose of establishing consistency and rate-type results later. We refer the readers to Dümbgen and Spokoiny (2001) and Rufibach and Walther (2010) for possible improvement over this constant.

Step Two.
Define the set

$$B_T = \left\{\max_{j=1,\ldots,q}\ \max_{\substack{\tau_{j-1}\le s<\tau_j\\ \tau_j<e\le\tau_{j+1}\\ s<b<e}} \frac{\left|\left\langle \boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle, \boldsymbol{\varepsilon}\right\rangle\right|}{\|\boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle\|_2} \le \lambda_T\right\}.$$

Again, for any $0 \le s < b < e \le T$, $\frac{|\langle \boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle, \boldsymbol{\varepsilon}\rangle|}{\|\boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle\|_2}$ follows a standard normal distribution, so using a similar argument, we get

$$\mathbb{P}\left(B_T^c\right) \le \frac{T^{-1}}{12\sqrt{\pi}}.$$

Step Three.
To fix the ideas, for $j = 1, \ldots, q$, we define intervals

$$\mathcal{I}_j^L = (\tau_j - \delta_T/3 - 1, \tau_j - \delta_T/6 - 1] \tag{20}$$
$$\mathcal{I}_j^R = (\tau_j + \delta_T/6, \tau_j + \delta_T/3] \tag{21}$$

Note that these intervals all contain at least one integer as long as $\delta_T > 6$. This is always true for sufficiently large T, as it follows from Conditions 1 and 2 that $\delta_T > \underline{C} \log T / \underline{f}$. Recall that F_T^M is the set of M randomly drawn intervals with pairs of endpoints in $\{0, \ldots, T-1\} \times \{1, \ldots, T\}$. Denote by $(s_1, e_1], \ldots, (s_M, e_M]$ the elements of F_T^M and let

$$D_T^M = \left\{\forall j = 1, \ldots, q, \exists k \in \{1, \ldots, M\}, \text{ s.t. } s_k \in \mathcal{I}_j^L \text{ and } e_k \in \mathcal{I}_j^R\right\}. \tag{22}$$

We have that

$$\begin{aligned}\mathbb{P}\left((D_T^M)^c\right) &\le \sum_{j=1}^q \Pi_{m=1}^M \Big(1 - \mathbb{P}\left(s_m \times e_m \in \mathcal{I}_j^L \times \mathcal{I}_j^R\right)\Big)\\ &\le q\left(1 - \frac{\delta_T^2}{6^2 T^2}\right)^M \le \frac{T}{\delta_T}\left(1 - \frac{\delta_T^2}{36T^2}\right)^M.\end{aligned}$$

Therefore, $\mathbb{P}\left(A_T \cap B_T \cap D_T^M\right) \ge 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M$.

In the rest of the proof, we assume that A_T, B_T and D_T^M all hold. We give the constants as follows:

$$\underline{C} = \sqrt{6}\big(2\sqrt{C_3} + 4\sqrt{2}\big) + 1, \quad C_1 = 2\sqrt{C_3} + 2\sqrt{2}, \quad C_2 = \frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{\underline{C}}, \quad C_3 = 32\sqrt{2} + 48.$$

These constants could be further refined by applying the Bonferroni bound more carefully. See also our remark at the end of Step One. But since our main aim is to establish the rate, we chose not to pursue this direction further. In addition, here we set $\underline{C}$ in such a way that $\underline{C}C_2 > C_1$ (as well as $C_2 > 0$). This means that given $\delta_T^{1/2}\underline{f}_T \ge \underline{C}\sqrt{\log T}$, one have that $C_2\delta_T^{1/2}\underline{f}_T > C_1\sqrt{\log T}$, i.e. we can select $\zeta_T \in [C_1\sqrt{\log T}, C_2\delta_T^{1/2}\underline{f}_T)$.

Step Four.
We focus on a generic interval $(s, e]$ such that

$$\exists j \in \{1, \ldots, q\},\ \exists k \in \{1, \ldots, M\}, \text{ s.t. } (s_k, e_k] \subset (s, e] \text{ and } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R \tag{23}$$

Fix such an interval $(s, e]$ and let $j \in \{1, \ldots, q\}$ and $k \in \{1, \ldots, M\}$ be such that (23) is satisfied. Let $b_k^* = \operatorname{argmax}_{s_k < b < e_k} \mathcal{C}^b_{(s_k, e_k]}(\mathbf{Y})$. By construction, $(s_k, e_k]$ satisfies $\tau_j - s_k >$

$\delta_T/6$ and $e_k - \tau_j > \delta_T/6$. Denote by

$$\mathcal{M}_{(s,e]} = \left\{m : (s_m, e_m] \in F_T^M, (s_m, e_m] \subset (s, e]\right\};$$
$$\mathcal{O}_{(s,e]} = \{m \in \mathcal{M}_{(s,e]} : \max_{s_m < b < e_m} \mathcal{C}^b_{(s_m,e_m]}(\mathbf{Y}) > \zeta_T\}$$

Our first aim is to show that $\mathcal{O}_{(s,e]}$ is non-empty. This follows from Lemma 2 and the calculation below.

$$\begin{aligned}\mathcal{C}^{b_k^*}_{(s_k,e_k]}(\mathbf{Y}) &\geq \mathcal{C}^{\tau_j}_{(s_k,e_k]}(\mathbf{Y})\\ &\geq \mathcal{C}^{b_k^*}_{(s_k,e_k]}(\mathbf{f}) - \lambda_T \geq \left(\frac{\delta_T}{6}\right)^{1/2} |f_{\tau_j+1} - f_{\tau_j}| - \lambda_T \geq \left(\frac{\delta_T}{6}\right)^{1/2} \underline{f}_T - \lambda_T\\ &= \left(\frac{1}{\sqrt{6}} - \frac{\lambda_T}{\delta_T^{1/2}\underline{f}_T}\right)\delta_T^{1/2}\underline{f}_T \geq \left(\frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{\underline{C}}\right)\delta_T^{1/2}\underline{f}_T = C_2\delta_T^{1/2}\underline{f}_T > \zeta_T.\end{aligned}$$

Let $m^* = \text{argmin}_{m \in \mathcal{O}_{s,e}}(e_m - s_m)$ and $b^* = \text{argmax}_{s_{m^*} < b < e_{m^*}} \mathcal{C}^b_{(s_{m^*},e_{m^*}]}(\mathbf{Y})$. Observe that (s_{m^*}, e_{m^*}) must contain at least one change-point. Indeed, if that was not the case, we would have $\mathcal{C}^b_{(s_{m^*},e_{m^*}]}(\mathbf{f}) = 0$ and

$$\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) = \left|\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) - \mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{f})\right| \leq \lambda_T \leq \zeta_T$$

which contradicts $\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) > \zeta_T$. On the other hand, $[s_{m^*}, e_{m^*})$ cannot contain more than one change-points, because $e_{m^*} - s_{m^*} \leq e_k - s_k \leq \delta_T$, as we picked the *narrowest-over-threshold* interval.

Without loss of generality, assume $\tau_j \in (s_{m^*}, e_{m^*})$. Denote by $\eta_L = \tau_j - s_{m^*}$, $\eta_R = e_{m^*} - \tau_j$ and $\eta_T = (C_1 - \sqrt{8})^2(\Delta_j^{\mathbf{f}})^{-2}\log T$, where $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$. We claim that $\min(\eta_L, \eta_R) > \eta_T$, because otherwise $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 2 would result in

$$\begin{aligned}\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) &\leq \mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{f}) + \lambda_T \leq \mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{f}) + \lambda_T \leq \eta_T^{1/2}\Delta_j^{\mathbf{f}} + \lambda_T\\ &= (C_1 - \sqrt{8} + \sqrt{8})\sqrt{\log T} = C_1\sqrt{\log T} \leq \zeta_T,\end{aligned}$$

which would contradict $\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) > \zeta_T$.

We are now in the position to prove $|b^* - \tau_j| \leq C_3 \log T/(\Delta_j^{\mathbf{f}})^2$. The arguments we use here are simpler and slightly more general than Lemma A.3 of Fryzlewicz (2014). Our aim is to find ϵ_T such that for any $b \in \{s_{m^*}+1, \ldots, e_{m^*}-1\}$ with $|b - \tau_j| > \epsilon_T$, we always have

$$(\mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}))^2 - (\mathcal{C}^{b}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}))^2 > 0. \tag{24}$$

This would then imply that $|b^* - \tau_j| \leq \epsilon_T$. By expansion and rearranging the terms (using the fact that $f_t = Y_t + \varepsilon_t$), we see that (24) is equivalent to

$$\begin{aligned}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*},e_{m^*}]}\rangle^2 - \langle \mathbf{f}, \boldsymbol{\psi}^{b}_{(s_{m^*},e_{m^*}]}\rangle^2 >& \langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}^{b}_{(s_{m^*},e_{m^*}]}\rangle^2 - \langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*},e_{m^*}]}\rangle^2\\ &+ 2\left\langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}^{b}_{(s_{m^*},e_{m^*}]}\langle \mathbf{f}, \boldsymbol{\psi}^{b}_{(s_{m^*},e_{m^*}]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s_{m^*},e_{m^*}]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*},e_{m^*}]}\rangle\right\rangle.\end{aligned} \tag{25}$$

In the following, we assume that $b \geq \tau_j$. The case that $b < \tau_j$ can be handled in a similar fashion. By Lemma 4, we have

$$\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*}, e_{m^*}]} \rangle^2 - \langle \mathbf{f}, \boldsymbol{\psi}^{b}_{(s_{m^*}, e_{m^*}]} \rangle^2 = (\mathcal{C}^{\tau_j}_{(s_{m^*}, e_{m^*}]}(\mathbf{f}))^2 - (\mathcal{C}^{b}_{(s_{m^*}, e_{m^*}]}(\mathbf{f}))^2 = \frac{|b-\tau_j|\eta_L}{|b-\tau_j|+\eta_L}(\Delta^{\mathbf{f}}_j)^2 := \kappa.$$

In addition, since A_T and B_T hold, we have that

$$\begin{aligned}
&\langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}^{b}_{(s_{m^*}, e_{m^*}]} \rangle^2 - \langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*}, e_{m^*}]} \rangle^2 \leq \lambda_T^2, \\
&2\left\langle \boldsymbol{\varepsilon}, \boldsymbol{\psi}^{b}_{(s_{m^*}, e_{m^*}]} \langle \mathbf{f}, \boldsymbol{\psi}^{b}_{(s_{m^*}, e_{m^*}]} \rangle - \boldsymbol{\psi}^{\tau_j}_{(s_{m^*}, e_{m^*}]} \langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*}, e_{m^*}]} \rangle \right\rangle \\
&\qquad\qquad \leq 2\| \boldsymbol{\psi}^{b}_{(s_{m^*}, e_{m^*}]} \langle \mathbf{f}, \boldsymbol{\psi}^{b}_{(s_{m^*}, e_{m^*}]} \rangle - \boldsymbol{\psi}^{\tau_j}_{(s_{m^*}, e_{m^*}]} \langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s_{m^*}, e_{m^*}]} \rangle \|_2 \lambda_T = 2\kappa^{1/2}\lambda_T,
\end{aligned}$$

where the last equality also comes from Lemma 4. Consequently, (25) can be deducted from the stronger inequality $\kappa - 2\lambda_T\kappa^{1/2} - \lambda_T^2 > 0$. This quadratic inequality is implied by $\kappa > (\sqrt{2}+1)^2\lambda_T^2$, and could be restricted further to

$$\frac{2|b-\tau_j|\eta_L}{|b-\tau_j|+\eta_L} \geq \min(|b-\tau_j|, \eta_L) > (32\sqrt{2}+48)(\Delta^{\mathbf{f}}_j)^{-2}\log T = C_3(\Delta^{\mathbf{f}}_j)^{-2}\log T. \tag{26}$$

But since

$$\eta_L \geq \eta_T = (C_1 - \sqrt{8})^2(\Delta^{\mathbf{f}}_j)^{-2}\log T = (2\sqrt{C_3})^2(\Delta^{\mathbf{f}}_j)^{-2}\log T > C_3(\Delta^{\mathbf{f}}_j)^{-2}\log T,$$

we see that the inequality in (26) is equivalent to $|b-\tau_j| > C_3(\Delta^{\mathbf{f}}_j)^{-2}\log T$. To sum up, $|b^*-\tau_j|(\Delta^{\mathbf{f}}_j)^2 > C_3\log T$ would result in (24), a contradiction. So we have proved that $|b^*-\tau_j|(\Delta^{\mathbf{f}}_j)^2 \leq C_3\log T$.

Step Five.
Using the arguments given above which are valid on the event $A_T \cap B_T \cap D^M_T$, we can now proceed with the proof of the theorem as follows. At the start of Algorithm 1 we have $s = 0$ and $e = T$ and, provided that $q \geq 1$, condition (23) is satisfied. Therefore the algorithm detects a change-point b^* in that interval such that $|b^* - \tau_j| \leq C_3 \log T(\Delta^{\mathbf{f}}_j)^{-2}$ for some j. By construction, we also have that $|b^* - \tau_j| < 2/3\delta_T$. This in turn implies that for all $l = 1, \ldots, q$ such that $\tau_l \in (s, e)$ and $l \neq j$ we have either $\mathcal{I}^L_l, \mathcal{I}^R_l \subset (s, b^*]$ or $\mathcal{I}^L_l, \mathcal{I}^R_l \subset (b^*, e]$. Therefore (23) is satisfied within each segment containing at least one change-point. Note that before all q change-points are detected, each change-point will not be detected twice. To see this, we suppose that τ_j has already been detected by b^*, then for all intervals $(s_k, e_k] \subset (\tau_j - C_3\log T(\Delta^{\mathbf{f}}_j)^{-2}, \tau_j + 2/3\delta_T] \cup (\tau_j - 2/3\delta_T, \tau_j + C_3\log T(\Delta^{\mathbf{f}}_j)^{-2}]$, Lemma 2, together with the event A_T, guarantees that

$$\max_{s_k<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{Y}) \leq \max_{s<b<e} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{f}) + \lambda_T \leq \sqrt{C_3\log T(\Delta^{\mathbf{f}}_j)^{-2}}\Delta^{\mathbf{f}}_j + \lambda_T \leq C_1\sqrt{\log T} \leq \zeta_T.$$

Once all the change-points are detected, we then only need to consider $(s_k, e_k]$ such that

$$(s_k, e_k] \subset (\tau_j - C_3\log T(\Delta^{\mathbf{f}}_j)^{-2}, \tau_{j+1} + C_3\log T(\Delta^{f}_{j+1})^{-2}]$$

for $j = 0, \ldots, q$, where we set $\Delta_0^f = \Delta_{q+1}^f = \infty$ for notational convenience. It follows from Lemma 3 (within the event A_T) that

$$\begin{aligned}\max_{s_k<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{Y}) &\leq \max_{s_k<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{f}) + \lambda_T \\ &\leq \sqrt{C_3 \log T (\Delta_j^{\mathbf{f}})^{-2}}\Delta_j^{\mathbf{f}} + \sqrt{C_3 \log T (\Delta_{j+1}^{f})^{-2}}\Delta_{j+1}^{\mathbf{f}} + \lambda_T \\ &< (2\sqrt{C_3} + \sqrt{8})\sqrt{\log T} = C_1\sqrt{\log T} \leq \zeta_T.\end{aligned}$$

Hence the algorithm terminates with no further change-point detection. □

I.3. Proof of Theorem 2

Proof. The proof proceeds in analogy to the proof of Theorem 1. In five steps we shall establish the following result,

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\ldots,q}\left(|\hat{\tau}_j - \tau_j|(\Delta_j^{\mathbf{f}})^{2/3}\right) \leq C_3(\log T)^{1/3}\right) \geq 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M, \tag{27}$$

which in turn implies (10).

Step One and Step Two

We define the following two events

$$A_T = \left\{\max_{s,b,e:1\leq s+1<b<e\leq T} |\mathcal{C}^b_{(s,e]}(\boldsymbol{\varepsilon})| \leq \lambda_T\right\},$$

$$B_T = \left\{\max_{j=1,\ldots,q}\ \max_{\substack{\tau_{j-1}\leq s+1<\tau_j \\ \tau_j<e\leq\tau_{j+1} \\ s+1<b<e}} \frac{\left|\left\langle \boldsymbol{\phi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^b_{(s,e]}\rangle - \boldsymbol{\phi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^{\tau_j}_{(s,e]}\rangle, \boldsymbol{\varepsilon}\right\rangle\right|}{\|\boldsymbol{\phi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^b_{(s,e]}\rangle - \boldsymbol{\phi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\phi}^{\tau_j}_{(s,e]}\rangle\|_2} \leq \lambda_T\right\},$$

where $\lambda_T = \sqrt{8\log T}$. Arguments as those used in Step One and Step Two of the proof of Theorem 1 show that $\mathbb{P}(A_T^c) \leq \frac{T^{-1}}{12\sqrt{\pi}}$ and $\mathbb{P}(B_T^c) \leq \frac{T^{-1}}{12\sqrt{\pi}}$.

Step Three

Here $\mathcal{I}_j^L$, $\mathcal{I}_j^L$ and D_T^M are as defined in the proof of Theorem 1. In the rest of the proof, we assume that A_T, B_T and D_T^M all hold, where the last event is given by (22). Exactly as in the proof of Theorem 9, we show that $\mathbb{P}\left(A_T \cap B_T \cap D_T^M\right) \geq 1 - T^{-1}/(6\sqrt{\pi}) - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/36)^M$.

We give the constants as follows:

$$\underline{C} = 72\left(4\sqrt{2}+2C_3^{3/2}\right)+1, \quad C_1 = 2C_3^{3/2}+2\sqrt{2}, \quad C_2 = \frac{1}{72} - \frac{2\sqrt{2}}{\underline{C}}, \quad C_3 = 2\sqrt[3]{7}\left(3\left(1+\sqrt{2}\right)\right)^{2/3}.$$

Here we set $\underline{C}$ in such a way that $\underline{C}C_2 > C_1$ (which also implies that $C_2 > 0$). Consequently, given $\delta_T^{3/2}\underline{f}_T \geq \underline{C}\sqrt{\log T}$ it is possible to select $\zeta_T \in \left[C_1\sqrt{\log T}, C_2\delta_T^{3/2}\underline{f}_T\right)$.

Again, these constants could be further refined. But since our main aim is to establish the rate, we chose not to pursue this direction here.

Step Four
Consider a generic interval $(s, e]$ satisfying

$$\exists j \in \{1, \ldots, q\},\ \exists k \in \{1, \ldots, M\},\ \text{s.t. } (s_k, e_k] \subset (s, e] \text{ and } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R \tag{28}$$

and define events

$$\begin{aligned} \mathcal{M}_{(s,e]} &= \left\{m : (s_m, e_m] \in F_T^M, (s_m, e_m] \subset (s, e]\right\}, \\ \mathcal{O}_{(s,e]} &= \{m \in \mathcal{M}_{(s,e]} : \max_{s_m+1<b<e_m} \mathcal{C}^b_{(s_m,e_m]}(\mathbf{Y}) > \zeta_T\}. \end{aligned}$$

Let $b_k^* = \operatorname{argmax}_{s_k+1<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{Y})$. We have

$$\begin{aligned} \mathcal{C}^{b_k^*}_{(s_k,e_k]}(\mathbf{Y}) &\geq \mathcal{C}^{\tau_j}_{(s_k,e_k]}(\mathbf{Y}) \\ &\geq \mathcal{C}^{b_k^*}_{(s_k,e_k]}(\mathbf{f}) - \lambda_T \geq \frac{1}{\sqrt{24}}(\delta_T/6)^{3/2}\Delta_j^{\mathbf{f}} - \lambda_T \geq \frac{1}{72}\delta_T^{3/2}\underline{f}_T - \lambda_T \\ &= \left(\frac{1}{72} - \frac{\lambda_T}{\delta_T^{3/2}\underline{f}_T}\right)\delta_T^{3/2}\underline{f}_T \geq \left(\frac{1}{72} - \frac{2\sqrt{2}}{\underline{C}}\right)\delta_T^{3/2}\underline{f}_T = C_2\delta_T^{3/2}\underline{f}_T > \zeta_T, \end{aligned}$$

where the third inequality above follows from Lemma 5, therefore $\mathcal{O}_{s,e}$ is non-empty.

Let $m^* = \operatorname{argmin}_{m\in\mathcal{O}_{(s,e]}}(e_m - s_m)$ and $b^* = \operatorname{argmax}_{s_{m^*}+1<b<e_{m^*}} \mathcal{C}^b_{(s_{m^*},e_{m^*}]}(\mathbf{Y})$. Arguing exactly as in Step Four in the proof of Theorem 1, we show that $(s_{m^*} + 1, e_{m^*})$ must contain exactly one change-point. Further, without loss of generality, assume that $\tau_j \in (s_{m^*} + 1, e_{m^*})$. Let $\eta_L = \tau_j - s_{m^*} - 1$, $\eta_R = e_{m^*} - \tau_j$ and

$$\eta_T = \left(\sqrt{3}(C_1 - \sqrt{8})\sqrt{\log T}(\Delta_j^{\mathbf{f}})^{-1})\right)^{2/3} - 1.$$

We observe that $\min(\eta_L, \eta_R) > \eta_T$, as otherwise $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 5 would imply

$$\begin{aligned} \mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) &\leq \mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{f}) + \lambda_T \leq \mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{f}) + \lambda_T \leq \frac{1}{\sqrt{3}}(\eta_T + 1)^{3/2}\Delta_j^{\mathbf{f}} + \lambda_T \\ &= (C_1 - \sqrt{8} + \sqrt{8})\sqrt{\log T} = C_1\sqrt{\log T} \leq \zeta_T, \end{aligned}$$

contradicting $\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) > \zeta_T$.

We are now in the position to prove that $|b^* - \tau_j| \leq C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} := \epsilon_T$. Let $b \in \{s_{m^*} + 2, \ldots, e_{m^*} - 1\}$. Our aim is to claim that when $|b - \tau_j| > \epsilon_T$,

$$(\mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}))^2 - (\mathcal{C}^{b}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}))^2 > 0. \tag{29}$$

Since inequality (29) does not hold for $b = b^*$, proving this claim consequently demonstrates that $|b^* - \tau_j| \leq \epsilon_T$.

Without loss of generality, we consider the case of $b > \tau_j$. Using arguments as those in Step Four of the proof of Theorem 1 we can show that (29) is implied by $\kappa > (\sqrt{2}+1)^2\lambda_T^2$, where $\kappa = (\mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{f}))^2 - (\mathcal{C}^{b}_{(s_{m^*},e_{m^*}]}(\mathbf{f}))^2$. By Lemma 7, $\kappa > (\sqrt{2}+1)^2\lambda_T^2$ is implied by

$$\min(|b - \tau_j|, \eta_L) > \left(63(\Delta_j^{\mathbf{f}})^{-2} \cdot 8(\sqrt{2}+1)^2 \log T\right)^{1/3} = C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}$$

However, for sufficiently large T,

$$\eta_L > \eta_T = (\sqrt{3}(C_1 - \sqrt{8}))^{2/3}(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} - 1 > (C_1 - \sqrt{8})^{2/3}(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}$$
$$> (C_3^{3/2} + \sqrt{8} - \sqrt{8})^{2/3}(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} = C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} = \epsilon_T,$$

hence $|b - \tau_j| > \epsilon_T$ implies (29), so it must hold that $|b^* - \tau_j| \le \epsilon_T$.

Step Five

Using the arguments given above which are valid on the event $A_T \cap B_T \cap D_T^M$, we can now proceed with the proof of the theorem as follows. At the start of Algorithm 1 we have $s = 0$ and $e = T$ and, provided that $q \ge 1$, condition (23) is satisfied. Therefore the algorithm detects a change-point b^* in that interval such that $|b^* - \tau_j| \le C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}$ for some j. By construction, we also have that $|b^* - \tau_j| < 2/3\delta_T$. This in turn implies that for all $l = 1, \ldots, q$ such that $\tau_l \in (s+1, e)$ and $l \ne j$ we have either $\mathcal{I}_l^L, \mathcal{I}_l^R \subset (s, b^*]$ or $\mathcal{I}_l^L, \mathcal{I}_l^R \subset (b^*, e]$. Therefore (23) is satisfied within each segment containing at least one change-point. Note that before all q change-points are detected, each change-point will not be detected twice. To see this, we suppose that τ_j has already been detected by b^*, then for all intervals $(s_k, e_k] \subset (\tau_j - \epsilon_T, \tau_j + 2/3\delta_T] \cup (\tau_j - 2/3\delta_T, \tau_j + \epsilon_T]$, Lemma 5, together with the event A_T, guarantees that for sufficiently large T

$$\begin{aligned}\max_{s_k+1<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{Y}) &\le \max_{s+1<b<e} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{f}) + \sqrt{8\log T} \\ &\le \frac{1}{\sqrt{3}}(C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3} + 1)^{3/2}\Delta_j^{\mathbf{f}} + \sqrt{8\log T} \\ &\le (2C_3^{3/2} + \sqrt{8})\sqrt{\log T} = C_1\sqrt{\log T} \le \zeta_T\end{aligned}$$

Once all the change-points are detected, we then only need to consider $(s_k, e_k]$ such that

$$(s_k, e_k] \subset (\tau_j - C_3\big(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}, \tau_{j+1} + C_3(\Delta_{j+1}^{\mathbf{f}})^{-2/3}(\log T)^{1/3}\big]$$

for $j = 0, \ldots, q$, where we set $\Delta_0^f = \Delta_{q+1}^f = \infty$ for notational convenience. It follows from Lemma 6 (within the event A_T) that

$$\begin{aligned}\max_{s_k+1<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{Y}) &\le \max_{s_k+1<b<e_k} \mathcal{C}^b_{(s_k,e_k]}(\mathbf{f}) + \sqrt{8\log T} \\ &\le (C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3})^{3/2}\Delta_j^{\mathbf{f}} + (C_3(\Delta_{j+1}^{\mathbf{f}})^{-2/3}(\log T)^{1/3})^{3/2}\Delta_{j+1}^{\mathbf{f}} + \sqrt{8\log T} \\ &= (2C_3^{3/2} + \sqrt{8})\sqrt{\log T} \le C_1\sqrt{\log T} \le \zeta_T.\end{aligned}$$

Hence the algorithm terminates and no further change-points will be detected. □

I.4. Proof of Theorem 3

Proof. Recall that $\{\varepsilon_t\}_{t=1}^T$ are i.i.d. $N(0, \sigma_0^2)$ with $\sigma_0 = 1$. For any candidate $\mathcal{T}(\zeta^{(k)})$ on the NOT solution path, the sSIC criterion function in (S1) can be written as

$$T\hat{\sigma}_k^2 + (2\hat{q}_k + 1)\log^\alpha(T) + \text{constant}$$

where $\hat{\sigma}_k^2$ is the estimated variance of the noise (i.e. the residual sum of squares divided by T) based on $\mathcal{T}(\zeta^{(k)})$, and $\hat{q}_k$ is the estimated number of change-points.

We now divide our proof into three parts.

Part I. About a particular model candidate on the NOT solution path
By Theorem 1, we know that with arbitrarily high probability for sufficiently large T, there exists k^* such that $\mathcal{T}(\zeta^{(k^*)})$ on the NOT solution path is a "good" candidate with $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_{k^*}} \in \mathcal{T}(\zeta^{(k^*)})$ satisfying $\hat{q}_{k^*} = q$ and $\max_{i=1}^q |\hat{\tau}_i - \tau_i| \le C' \log T$ for some $C' > 0$. In the rest of the proof, for presentational convenience, we condition on the event that such k^* does exist throughout our analysis.

In addition, we recall that $\mathbf{1}_{(s,e]} = \big(\mathbf{1}_{(s,e]}(1), \ldots, \mathbf{1}_{(s,e]}(T)\big)'$ with

$$\mathbf{1}_{(s,e]}(t) = \begin{cases} (e-s)^{-1/2}, & t = s+1, \ldots, e \\ 0, & \text{otherwise} \end{cases}, \tag{30}$$

and define the set

$$E_T = \Big\{ \max_{s,e: 0 \le s < e \le T} |\langle \mathbf{1}_{(s,e]}, \boldsymbol{\varepsilon} \rangle| \le \sqrt{6 \log T} \Big\}.$$

Using an argument similar to Step One of the proof of Theorem 1, we see that $\mathbb{P}\left(E_T^c\right) = O(T^{-1})$. Since we are only interested in proving a certain type of probabilistic statement for $T \to \infty$, here we could also assume that E_T holds.

Let $\{\hat{f}_t\}_{t=1}^T$ be the fitted values using the candidate on the solution path with $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_{k^*}} \in \mathcal{T}(\zeta^{(k^*)})$, and define $\tilde{f}_t = f_{\tau_j}$ for $t = \hat{\tau}_j, \ldots, \hat{\tau}_{j+1} - 1$ for every $j = 0, 1, \ldots q$. Here for notational convenience, we suppressed the dependence of $\{\hat{f}_t\}_{t=1}^T$ and $\{\tilde{f}_t\}_{t=1}^T$ on k^*. It is easy to see that $f_t - \tilde{f}_t$ is piecewise-constant, only non-zero for t between the true location of the change-point τ_j and its estimation $\hat{\tau}_j$, and exactly zero elsewhere. Write $\tilde{\mathbf{f}} = (\tilde{f}_1, \ldots, \tilde{f}_T)'$. Then

$$\begin{aligned} T\hat{\sigma}_{k^*}^2 &= \sum_{t=1}^T (\varepsilon_t + f_t - \hat{f}_t)^2 \\ &\le \sum_{t=1}^T (\varepsilon_t + f_t - \tilde{f}_t)^2 = \sum_{t=1}^T \varepsilon_t^2 + 2\langle \boldsymbol{\varepsilon}, \mathbf{f} - \tilde{\mathbf{f}} \rangle + \|\mathbf{f} - \tilde{\mathbf{f}}\|^2 \\ &= \sum_{t=1}^T \varepsilon_t^2 + 4q\bar{C}\sqrt{6\log T}\sqrt{C'\log T} + q(2\bar{C})^2 C' \log T \\ &= \sum_{t=1}^T \varepsilon_t^2 + (4q\bar{C}\sqrt{6C'} + 4qC'\bar{C}^2)\log T \end{aligned}$$

where the second last step follows from E_T, linearity of the inner product, and the fact that $\max_{i=1}^q |\hat{\tau}_i - \tau_i| \le C' \log T$.

Part II. Estimation of the number of change-points
In this part, we prove that for NOT with the sSIC, $\mathbb{P}\left(\hat{q} = q\right) \to 1$ as $T \to \infty$. We accomplish this by showing separately that (i) $\mathbb{P}\left(\hat{q} > q\right) \to 0$ and (ii) $\mathbb{P}\left(\hat{q} < q\right) \to 0$.

First, for all k with $\hat{q}_k > q$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k} \in \mathcal{T}(\zeta^{(k)})$, we consider a "saturated oracle" candidate model with $\hat{q}_k + q$ change-points at $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k}, \tau_1, \ldots, \tau_q$ respectively. We reorder these $\hat{q}_k + q$ locations as $0 = \mathring{\tau}_0 < \mathring{\tau}_1 \le \ldots \le \mathring{\tau}_{\hat{q}_k+q} < \mathring{\tau}_{\hat{q}_k+q+1} = T$, and denote the estimated variance of the errors corresponding this saturated oracle candidate by $\mathring{\sigma}_k^2$. Since for each

$j = 0, \ldots, \hat{q}_k + q$, f_t is constant over $\{1 + \mathring{\tau}_j, \ldots, \mathring{\tau}_{j+1}\}$, it then follows that

$$\begin{aligned}
T\hat{\sigma}_k^2 \geq T\mathring{\sigma}_k^2 &= \sum_{j=0}^{\hat{q}_k+q} \sum_{t=1+\mathring{\tau}_j}^{\mathring{\tau}_{j+1}} \Big\{\varepsilon_t - \frac{1}{\mathring{\tau}_{j+1} - \mathring{\tau}_j} \sum_{b=1+\mathring{\tau}_j}^{\mathring{\tau}_{j+1}} \varepsilon_b\Big\}^2 \\
&= \sum_{t=1}^{T} \varepsilon_t^2 - \sum_{j=0}^{\hat{q}_k+q} \langle \boldsymbol{\varepsilon}, \mathbf{1}_{1+\mathring{\tau}_j, \mathring{\tau}_{j+1}}\rangle^2 \geq \sum_{t=1}^{T} \varepsilon_t^2 - 6(q + \hat{q}_k + 1)\log T,
\end{aligned}$$

where the last line again follows from E_T. Note that in the above, for notational convenience, we have implicitly assumed that $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k}, \tau_1, \ldots, \tau_q$ are distinct. It is clear to see that the same argument holds even if $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k}\} \cap \{\tau_1, \ldots, \tau_q\} \neq \emptyset$. This means that for all k with $\hat{q}_k > q$,

$$\begin{aligned}
\mathrm{sSIC}(k) - \mathrm{sSIC}(k^*) &\geq T(\mathring{\sigma}_k^2 - \hat{\sigma}_{k^*}^2) + 2(\hat{q}_k - q)\log^\alpha(T) \\
&\geq \Big\{\sum_{t=1}^{T} \varepsilon_t^2 - 6(q + \hat{q}_k + 1)\log T\Big\} - \Big\{\sum_{t=1}^{T} \varepsilon_t^2 + (4q\bar{C}\sqrt{6C'} + 4qC'\bar{C}^2)\log T\Big\} \\
&\quad + 2(\hat{q}_k - q)\log^\alpha(T) \\
&= (\hat{q}_k - q)\{2\log^\alpha(T) - 6\log T\} - (12q + 4q\bar{C}\sqrt{6C'} + 4qC'\bar{C}^2 + 6)\log T \\
&\geq \{2\log^\alpha(T) - 6\log T\} - (12q + 4q\bar{C}\sqrt{6C'} + 4qC'\bar{C}^2 + 6)\log T > 0
\end{aligned}$$

for large enough T, which implies $\mathbb{P}(\hat{q} > q) \to 0$.

Second, for all k with $\hat{q}_k < q$, it must be the case that one can find some $j^* \in \{1, \ldots, q\}$ such that the corresponding $\hat{f}_t$ is constant over $\big(\tau_{j^*} - \lfloor \delta_T/2 \rfloor, \tau_{j^*} + \lfloor \delta_T/2 \rfloor\big]$. Now consider the "intermediate" candidate model with $\hat{q}_k + 3$ change-points at $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k}, \tau_{j^*} - \lfloor \delta_T/2 \rfloor, \tau_{j^*}, \tau_{j^*} + \lfloor \delta_T/2 \rfloor$, and denote the corresponding estimated variance of errors by $\tilde{\sigma}_k^2$. Without loss of generality, assume that $f_{\tau_{j^*}+1} > f_{\tau_{j^*}}$. Then,

$$\begin{aligned}
T\hat{\sigma}_k^2 - T\tilde{\sigma}_k^2 \geq &\sum_{t=\tau_{j^*}-\lfloor \delta_T/2 \rfloor+1}^{\tau_{j^*}} \Big\{\varepsilon_t - \frac{\Delta_{j^*}^{\mathbf{f}}}{2} - \frac{1}{2\lfloor \delta_T/2 \rfloor} \sum_{b=\tau_{j^*}-\lfloor \delta_T/2 \rfloor+1}^{\tau_{j^*}+\lfloor \delta_T/2 \rfloor} \varepsilon_b\Big\}^2 \\
&+ \sum_{t=\tau_{j^*}+1}^{\tau_{j^*}+\lfloor \delta_T/2 \rfloor} \Big\{\varepsilon_t + \frac{\Delta_{j^*}^{\mathbf{f}}}{2} - \frac{1}{2\lfloor \delta_T/2 \rfloor} \sum_{b=\tau_{j^*}-\lfloor \delta_T/2 \rfloor+1}^{\tau_{j^*}+\lfloor \delta_T/2 \rfloor} \varepsilon_b\Big\}^2 \\
&- \sum_{t=\tau_{j^*}-\lfloor \delta_T/2 \rfloor+1}^{\tau_{j^*}} \varepsilon_t^2 - \sum_{t=\tau_{j^*}+1}^{\tau_{j^*}+\lfloor \delta_T/2 \rfloor} \varepsilon_t^2 \\
= &\, 2(\Delta_{j^*}^{\mathbf{f}}/2)^2 \lfloor \delta_T/2 \rfloor - \Big\langle \boldsymbol{\varepsilon}, \mathbf{1}_{\left(\tau_{j^*}-\lfloor \delta_T/2 \rfloor, \tau_{j^*}+\lfloor \delta_T/2 \rfloor\right]} \Big\rangle^2 \\
&- \Delta_{j^*}^{\mathbf{f}} \sqrt{\lfloor \delta_T/2 \rfloor} \Big\{\Big\langle \boldsymbol{\varepsilon}, \mathbf{1}_{\left(\tau_{j^*}-\lfloor \delta_T/2 \rfloor, \tau_{j^*}\right]} \Big\rangle - \Big\langle \boldsymbol{\varepsilon}, \mathbf{1}_{\left(\tau_{j^*}, \tau_{j^*}+\lfloor \delta_T/2 \rfloor\right]} \Big\rangle\Big\} \\
\geq &\, \frac{1}{2}\big(\Delta_{j^*}^{\mathbf{f}} \sqrt{\lfloor \delta_T/2 \rfloor} - 2\sqrt{6\log T}\big)^2 - 12\log T - 6\log T
\end{aligned}$$

In the mean time, by adding $q - 1$ more change-points, $\tau_1, \ldots, \tau_{j^*-1}, \tau_{j^*+1}, \ldots, \tau_q$, to the intermediate candidate model, we can show that using the same argument as in the first half of Part II that

$$T\tilde{\sigma}_k^2 \geq \sum_{t=1}^{T} \varepsilon_t^2 - 6(q + \hat{q}_k + 3)\log T.$$

Since $\delta_T \geq \underline{C}_1(\log T)^{\alpha'}$ with $\alpha' > 1$, $\Delta_{j^*}^{\mathbf{f}}\sqrt{\lfloor \delta_T/2 \rfloor} \geq \underline{C}_2\sqrt{\lfloor \delta_T/2 \rfloor} \geq 2\sqrt{6\log T}$ for large enough T. Consequently, combining the previous two displayed equations lead to

$$T\hat{\sigma}_k^2 \geq \sum_{t=1}^{T} \varepsilon_t^2 - 6(q + \hat{q}_k + 6)\log T + \frac{1}{2}\big(\underline{C}_2\sqrt{\lfloor \delta_T/2 \rfloor} - 2\sqrt{6\log T}\big)^2$$

for large enought T. This means that for all k with $\hat{q}_k < q$,

$$\begin{aligned}
\mathrm{sSIC}(k) - \mathrm{sSIC}(k^*) &= T(\hat{\sigma}_k^2 - \hat{\sigma}_{k^*}^2) + 2(\hat{q}_k - q)\log^{\alpha}(T) \\
&\geq \frac{1}{2}\big(\underline{C}_2\sqrt{\lfloor \delta_T/2 \rfloor} - 2\sqrt{6\log T}\big)^2 \\
&\qquad - (4q\bar{C}\sqrt{6C'} + 4qC'\bar{C}^2 + 6q + 6\hat{q}_k + 36)\log T - 2q\log^{\alpha}(T) \\
&> 0
\end{aligned}$$

for sufficiently large T, where we again used that fact that $\delta_T \geq \underline{C}_1(\log T)^{\alpha'}$ with $\alpha' > \alpha > 1$, so $\frac{1}{2}\big(\underline{C}_2\sqrt{\lfloor \delta_T/2 \rfloor} - 2\sqrt{6\log T}\big)^2$ is at least of order $(\log T)^{\alpha'}$. This implies $\mathbb{P}(\hat{q} < q) \to 0$.

In conclusion, we have established $\mathbb{P}(\hat{q} = q) \to 1$.

Part III. Estimation of the change-point locations

In view of the conclusion of Part II, in the rest of the proof we could assume that E_T holds and $\hat{q} = q$. Suppose that the model picked via NOT with the sSIC is $\hat{\tau}_1, \ldots, \hat{\tau}_q \in \mathcal{T}(\zeta^{(\hat{k})})$. Furthermore, let

$$j^* = \mathrm{argmax}_{j=1,\ldots,q} \min_{i=1,\ldots,q} |\hat{\tau}_i - \tau_j| \quad \text{and} \quad C := \frac{\min\big(\lfloor \delta_T/2 \rfloor, \min_{i=1,\ldots,q}|\hat{\tau}_i - \tau_{j^*}|\big)}{\log T}.$$

Our aim is to show that C is finite (more precisely, has an upper bound independent of T). Now consider a "near-saturated oracle" candidate model with $2q+1$ change-points at

$$\{\hat{\tau}_1, \ldots, \hat{\tau}_q, \tau_1, \ldots, \tau_{j^*-1}, \tau_{j^*+1}, \ldots, \hat{\tau}_q, \tau_{j^*} - C\log T, \tau_{j^*} + C\log T\}$$

with the corresponding estimated variance of the errors denoted as $\dot{\sigma}_{\hat{k}}^2$. So here instead of adding all the true change-points to the set of estimated change-points as before (which generates the so-called "saturated oracle"), we add all true change-points apart from τ_{j^*}, and replace it by $\tau_{j^*} \pm C\log T$.

Note that by construction (i.e. via δ_T in the definition of C), f_t is constant over $\big(\tau_{j^*} - C\log T, \tau_{j^*}\big]$ and $\big(\tau_{j^*}, \tau_{j^*} + C\log T\big]$. In addition, $\Delta_{j^*}^{\mathbf{f}} = |f_{\tau_{j^*}+1} - f_{\tau_{j^*}}| \geq \underline{f}_T$. Write

$$\bar{\varepsilon}_* = \frac{1}{2C\log T}\sum_{t=\tau_{j^*}-C\log T+1}^{\tau_{j^*}+C\log T} \varepsilon_t.$$

Without loss of generality, assume that $f_{\tau_{j^*}+1} > f_{\tau_{j^*}}$. Now using the arguments similar to those in Part II, we see that

$$
\begin{aligned}
T\hat{\sigma}^2_{\hat{k}} \geq T\dot{\sigma}^2_{\hat{k}} \geq & \sum_{t=1}^{\tau_{j^*}-C\log T} \varepsilon_t^2 + \sum_{t=\tau_{j^*}+C\log T+1}^{T} \varepsilon_t^2 - (2q)6\log T \\
& + \sum_{t=\tau_{j^*}-C\log T+1}^{\tau_{j^*}} (\varepsilon_t - \Delta^{\mathbf{f}}_{j^*}/2 - \bar{\varepsilon}_*)^2 + \sum_{t=\tau_{j^*}+1}^{\tau_{j^*}+C\log T} (\varepsilon_t + \Delta^{\mathbf{f}}_{j^*}/2 - \bar{\varepsilon}_*)^2 \\
= & \sum_{t=1}^{T} \varepsilon_t^2 - 12q\log T + \Delta^{\mathbf{f}}_{j^*}\Big(\sum_{t=\tau_{j^*}+1}^{\tau_{j^*}+C\log T} \varepsilon_t - \sum_{t=\tau_{j^*}-C\log T+1}^{\tau_{j^*}} \varepsilon_t \Big) \\
& + (\Delta^{\mathbf{f}}_{j^*}/2)^2(2C\log T) - (2C\log T)\bar{\varepsilon}_*^2 \\
= & \sum_{t=1}^{T} \varepsilon_t^2 - 12q\log T + \Delta^{\mathbf{f}}_{j^*}\sqrt{C\log T}\Big\{ \langle \boldsymbol{\varepsilon}, \mathbf{1}_{\tau_{j^*}+1,\tau_{j^*}+C\log T}\rangle - \langle \boldsymbol{\varepsilon}, \mathbf{1}_{\tau_{j^*}-C\log T+1,\tau_{j^*}}\rangle \Big\} \\
& + (\Delta^{\mathbf{f}}_{j^*}/2)^2(2C\log T) - \langle \boldsymbol{\varepsilon}, \mathbf{1}_{\tau_{j^*}-C\log T+1,\tau_{j^*}+C\log T}\rangle^2 \\
\geq & \sum_{t=1}^{T} \varepsilon_t^2 - \{6(2q+1) + 2\sqrt{6C}\Delta^{\mathbf{f}}_{j^*}\}\log T + (\Delta^{\mathbf{f}}_{j^*}/2)^2(2C\log T)
\end{aligned}
$$

However,

$$
T\hat{\sigma}^2_{\hat{k}} \leq T\hat{\sigma}^2_{k^*} \leq \sum_{t=1}^{T} \varepsilon_t^2 + (4q\bar{C}\sqrt{6C'} + 4qC'\bar{C}^2)\log T
$$

Combining the above two inequalities, and after some algebraic manipulations, we get

$$
2q\bar{C}\sqrt{6C'} + 2qC'\bar{C}^2 \geq C(\Delta^{\mathbf{f}}_{j^*}/2)^2 - 3(2q+1) - \sqrt{6C}\Delta^{\mathbf{f}}_{j^*},
$$

and thus

$$
2q\bar{C}\sqrt{6C'} + 2qC'\bar{C}^2 + 3(2q+1) + 6 \geq (\sqrt{C}\Delta^{\mathbf{f}}_{j^*}/2 - \sqrt{6})^2,
$$

which entails

$$
C \leq 4\Big[\Big\{2q\bar{C}\sqrt{6C'} + 2qC'\bar{C}^2 + 3(2q+1) + 6\Big\}^{1/2} + \sqrt{6}\Big]^2/\underline{C}_2^2.
$$

Finally, we remark that since $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1}) \geq \underline{C}_1(\log T)^{\alpha'}$, for sufficiently large T,

$$
C\log T \geq \min\Big(\lfloor \delta_T/2 \rfloor, \max_{j=1,\ldots,q}\min_{i=1,\ldots,q}|\hat{\tau}_i - \tau_j|\Big) = \max_{j=1,\ldots,q}|\hat{\tau}_j - \tau_j|.
$$

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

I.5. Proof of Theorem 4

First, we strengthen Theorem 2 in the scenario where the true signal has finitely many kinks with spacing $\sim T$.

Lemma 9. *Under the assumptions of Theorem 4, there exist constants C' and $\tilde{C}$ such that by setting $\zeta_T = \tilde{C}\sqrt{T}$ and $M \geq 36\underline{C}_1^{-2}\log(\underline{C}_1^{-1}T)$, we have that*

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\ldots,q} |\hat{\tau}_j - \tau_j| \leq C'\sqrt{T\log T}\right) \to 1, \tag{31}$$

as $T \to \infty$.

Proof. Let $\underline{C}, C_1, C_2, C_3 > 0$ be the constants upon applying Theorem 2. For simplicity, here we shall take

$$\widetilde{C} = C_2\underline{C}_1^{3/2}\underline{C}_2/2 \quad \text{and} \quad C' = \frac{32\sqrt{3}(\sqrt{2}+1)}{\underline{C}_2\{\sqrt{3}\underline{C}_1\widetilde{C}/\bar{C}\}^{1/3}}$$

First, we verify that the conditions in Theorem 2 are satisfied. Specifically, we note that under the additional assumptions of Theorem 4, for sufficiently large T,

(a) $\delta_T^{3/2}\underline{f}_T \geq \underline{C}_1^{3/2}\underline{C}_2\sqrt{T} > \underline{C}\sqrt{\log T}$,

(b) $\zeta_T = \widetilde{C}\sqrt{T} \in [C_1\sqrt{\log T}, C_2\delta_T^{3/2}\underline{f}_T)$,

(c) $M \geq 36\underline{C}_1^{-2}\log(\underline{C}_1^{-1}T) \geq 36(T/\delta_T)^2\log\{(T/\delta_T)T\}$.

This means that

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\ldots,q} |\hat{\tau}_j - \tau_j| \leq C_3\underline{C}_2^{-2/3}(T^2\log T)^{1/3}\right) \to 1.$$

Second, to strengthen the convergence rate of $\max_{j=1,\ldots,q}|\hat{\tau}_j - \tau_j|$, we make some minor modifications to Step Four in the proof of Theorem 2.

We still let $m^* = \operatorname{argmin}_{m\in\mathcal{O}_{(s,e]}}(e_m - s_m)$ and $b^* = \operatorname{argmax}_{s_{m^*}+1<b<e_{m^*}} \mathcal{C}^b_{(s_{m^*},e_{m^*}]}(\mathbf{Y})$, where $(s_{m^*}+1, e_{m^*})$ must contain exactly one change-point. Again, we consider $\tau_j \in (s_{m^*}+1, e_{m^*})$, and let $\eta_L = \tau_j - s_{m^*} - 1$ and $\eta_R = e_{m^*} - \tau_j$. Note that

$$\max_{j=1,\ldots,q} \Delta_j^{\mathbf{f}} \leq \frac{4\max_{i=1,\ldots,T}|f_i|}{\delta_T} \leq \frac{4\bar{C}}{\underline{C}_1}\frac{1}{T}$$

By setting $\eta_T = \{\sqrt{3}\underline{C}_1\widetilde{C}/(8\bar{C})\}^{2/3}T - 1$ (*different* from the proof of Theorem 2), we observe that $\min(\eta_L, \eta_R) > \eta_T$ for sufficiently large T (satisfying $8\log T < \widetilde{C}^2T/4$). It is because otherwise $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 5 would imply that

$$\begin{aligned}\mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}) &\leq \mathcal{C}^{b^*}_{(s_{m^*},e_{m^*}]}(\mathbf{f}) + \lambda_T \leq \mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{f}) + \lambda_T \leq \frac{1}{\sqrt{3}}(\eta_T+1)^{3/2}\frac{4\bar{C}}{\underline{C}_1}\frac{1}{T} + \lambda_T \\ &= \frac{\widetilde{C}}{2}\sqrt{T} + \sqrt{8\log T} < \widetilde{C}\sqrt{T} = \zeta_T,\end{aligned}$$

which leads to a contradiction.

We are now in the position to prove that $|b^* - \tau_j| \leq C'\sqrt{T\log T} := \epsilon_T$. Note that in view of Theorem 2, it suffices to only consider

$$b \in \left\{s_{m^*}+2, \ldots, e_{m^*}-1\right\} \cap \left\{\tau_j - \lceil C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}\rceil, \ldots, \tau_j + \lceil C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}\rceil\right\}$$

Our aim is to show that given $|b-\tau_j| > \epsilon_T$ (as well as $|b-\tau_j| \le C_3(\Delta_j^{\mathbf{f}})^{-2/3}(\log T)^{1/3}$ in view of Theorem 2),

$$(\mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}))^2 - (\mathcal{C}^{b}_{(s_{m^*},e_{m^*}]}(\mathbf{Y}))^2 > 0. \tag{32}$$

Inequality (32) does not hold for $b=b^*$ by the definition of b^*, so proving this claim would demonstrate that $|b^*-\tau_j| \le \epsilon_T$.

Using arguments as those in Step Four of the proof of Theorem 1 (or Theorem 2), we can show that (32) is implied by $\kappa > (\sqrt{2}+1)^2\lambda_T^2$, where $\kappa = (\mathcal{C}^{\tau_j}_{(s_{m^*},e_{m^*}]}(\mathbf{f}))^2 - (\mathcal{C}^{b}_{(s_{m^*},e_{m^*}]}(\mathbf{f}))^2$. By Lemma 8, $\kappa > (\sqrt{2}+1)^2\lambda_T^2$ is implied by

$$\frac{(\Delta_j^{\mathbf{f}})^2}{48}\big\{\min(\eta_L,\eta_R)-1\big\}|b-\tau_j|^2 > (\sqrt{2}+1)^2\lambda_T^2, \tag{33}$$

In view of the fact that

$$\min(\eta_L,\eta_R)-1 > \eta_T - 2 = \big\{\sqrt{3}\underline{C}_1\widetilde{C}/(8\bar{C})\big\}^{2/3}T - 2 > \big\{\sqrt{3}\underline{C}_1\widetilde{C}/(8\bar{C})\big\}^{2/3}T/2$$

for sufficiently large T, (33) is further implied by

$$|b-\tau_j| > \frac{8\sqrt{6}(\sqrt{2}+1)\sqrt{\log T}}{\underline{C}_2/T\big\{\sqrt{3}\underline{C}_1\widetilde{C}/(8\bar{C})\big\}^{1/3}\sqrt{T/2}} = \frac{32\sqrt{3}(\sqrt{2}+1)}{\underline{C}_2\big\{\sqrt{3}\underline{C}_1\widetilde{C}/\bar{C}\big\}^{1/3}}\sqrt{T\log T} = C'\sqrt{T\log T}.$$

In conclusion, $|b-\tau_j| > \epsilon_T$ implies (32), leading to a contradiction. So it must hold that $|b^*-\tau_j| \le \epsilon_T$ for large T.

Finally, since $\mathbb{P}(\hat{q}=q) \to 1$, we have that

$$\mathbb{P}\left(\hat{q}=q,\ \max_{j=1,\ldots,q}|\hat{\tau}_j-\tau_j| \le C'\sqrt{T\log T}\right) \to 1,$$

as required.

□

Now we are in the position to prove Theorem 4.

Proof. The proof proceeds in analogy to the proof of Theorem 3. In the following, we present details of the main steps.

Again, thanks to the standard Gaussianity of the noise, for any candidate $\mathcal{T}(\zeta^{(k)})$ on the NOT solution path, the sSIC criterion function in (S2) can be written as

$$T\hat{\sigma}_k^2 + (2\hat{q}_k+2)\log^{\alpha}(T) + \text{constant}$$

where $\hat{\sigma}_k^2$ is the estimated variance of the noise (i.e. the residual sum of squares divided by T) based on $\mathcal{T}(\zeta^{(k)})$, and $\hat{q}_k$ is the estimated number of kinks.

Part I. About a particular model candidate on the NOT solution path

By Lemma 9, we know that with arbitrarily high probability for sufficiently large T, there exists k^* such that $\mathcal{T}(\zeta^{(k^*)})$ on the NOT solution path is a "good" candidate with $\hat{\tau}_1,\ldots,\hat{\tau}_{\hat{q}_{k^*}} \in \mathcal{T}(\zeta^{(k^*)})$ satisfying $\hat{q}_{k^*}=q$ and $\max_{i=1}^{q}|\hat{\tau}_i-\tau_i| \le C'\sqrt{T\log T}$ for some $C'>0$. In the rest of the proof, for presentational convenience, we assume the existence of such k^*.

Define the set

$$E_T = \left\{ \max_{s,e:0\leq s<e\leq T} \max\left(|\langle \boldsymbol{\gamma}_{(s,e]}, \boldsymbol{\varepsilon}\rangle|, |\langle \mathbf{1}_{(s,e]}, \boldsymbol{\varepsilon}\rangle| \right) \leq \sqrt{6\log T} \right\}.$$

Using the Bonferroni bound, we see that $\mathbb{P}\left(E_T^c\right) = O(T^{-1})$. Again, in the following, we could assume that E_T holds.

Let $\{\hat{f}_t\}_{t=1}^T$ be the fitted values using the candidate on the solution path with $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_{k^*}} \in \mathcal{T}(\zeta^{(k^*)})$, and define $\tilde{f}_1 = \hat{f}_1$, $\tilde{f}_{t+1} = \tilde{f}_t + (f_{\tau_j+1} - f_{\tau_j})$ for $t = \hat{\tau}_j, \ldots, \hat{\tau}_{j+1} - 1$ for every $j = 0, 1, \ldots q$. Again, here for notational convenience, we suppressed the dependence of $\{\hat{f}_t\}_{t=1}^T$ and $\{\tilde{f}_t\}_{t=1}^T$ on k^*. It is easy to see that $f_t - \tilde{f}_t$ is piecewise-linear and continuous, with at most $2q$ kinks and

$$\max_{t=1,\ldots,T} |f_t - \tilde{f}_t| \leq q \max_j(\Delta_j^{\mathbf{f}}) C' \sqrt{T\log T} \leq \frac{4\bar{C}}{\underline{C}_1 T} C' q \sqrt{T\log T} = \frac{4q\bar{C}C'}{\underline{C}_1}\sqrt{\log T/T}.$$

Write $\tilde{\mathbf{f}} = (\tilde{f}_1, \ldots, \tilde{f}_T)'$, then $\|\mathbf{f} - \tilde{\mathbf{f}}\|^2 \leq (4q\bar{C}C'/\underline{C}_1)^2 \log T$. Furthermore, it is easy to verify (under E_T) that

$$\begin{aligned} T\hat{\sigma}_{k^*}^2 &= \sum_{t=1}^T (\varepsilon_t + f_t - \hat{f}_t)^2 \leq \sum_{t=1}^T (\varepsilon_t + f_t - \tilde{f}_t)^2 = \sum_{t=1}^T \varepsilon_t^2 + 2\langle \boldsymbol{\varepsilon}, \mathbf{f} - \tilde{\mathbf{f}}\rangle + \|\mathbf{f} - \tilde{\mathbf{f}}\|^2 \\ &\leq \sum_{t=1}^T \varepsilon_t^2 + M'\log T \end{aligned}$$

for some positive constant M' that does not depend on T. Consequently, as $T \to \infty$, it follows that $\mathbb{P}\left(\hat{\sigma}_{k^*}^2 < 1 + \delta/2\right) = 1$ for any $\delta > 0$.

Part II. Estimation of the number of change-points

Our aim in this part is to show that $\mathbb{P}\left(\hat{q} = q\right) \to 1$ as $T \to \infty$. We accomplish this by showing separately that (i) $\mathbb{P}\left(\hat{q} < q\right) \to 0$ and (ii) $\mathbb{P}\left(\hat{q} > q\right) \to 0$.

First, we note that it follows from Lemma 5.3 and 5.4 of Liu *et al.* (1997) that there exists $\delta > 0$ such that as $T \to \infty$,

$$\min_{k:\hat{q}_k<q} \mathbb{P}\left(\hat{\sigma}_k^2 > 1 + \delta\right) \to 1.$$

This means that for all k with $\hat{q}_k < q$,

$$\mathrm{sSIC}(k) - \mathrm{sSIC}(k^*) = T(\hat{\sigma}_k^2 - \hat{\sigma}_{k^*}^2) + 2(\hat{q}_k - q)\log^\alpha(T) \geq \delta T/2 - 2q\log^\alpha(T) > 0$$

for large enough T, which implies $\mathbb{P}\left(\hat{q} < q\right) \to 0$.

Second, for all k with $\hat{q}_k > q$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k} \in \mathcal{T}(\zeta^{(k)})$, we consider a "saturated oracle" candidate model with $\hat{q}_k + q$ kinks at $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k}, \tau_1, \ldots, \tau_q$ respectively. We reorder these $\hat{q}_k + q$ locations as $0 = \mathring{\tau}_0 < \mathring{\tau}_1 \leq \ldots \leq \mathring{\tau}_{\hat{q}_k+q} < \mathring{\tau}_{\hat{q}_k+q+1} = T$, and denote by $\mathring{\sigma}_k^2$ the estimated variance of the errors corresponding to a piecewise-linear model with features at these locations but **without** the continuity constraint (so effectively the way of estimating this quantity under Scenario (S3)). Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)'$,

$$\boldsymbol{\Gamma}_{(s,e]} := \left[\mathbf{1}_{(s,e]}, \boldsymbol{\gamma}_{(s,e]}\right] \quad \text{and} \quad \mathbf{H}_{(s,e]} = \boldsymbol{\Gamma}_{(s,e]}\left(\boldsymbol{\Gamma}_{(s,e]}'\boldsymbol{\Gamma}_{(s,e]}\right)^{-1}\boldsymbol{\Gamma}_{(s,e]}'$$

for $0 \le s < e \le T$, where $\mathbf{\Gamma}_{(s,e]}$ is a $T \times 2$ matrix and $\mathbf{H}_{(s,e]}$ is a $T \times T$ matrix. Furthermore, denote by $\mathbf{D}_{(s,e]}$ a $T \times T$ diagonal matrix with 1 in the $(s+1, s+1)$-th to the (e,e)-th entries and zero elsewhere. Here both $\mathbf{H}_{(s,e]}$ and $\mathbf{H}_{(s,e]} - \mathbf{D}_{(s,e]}$ are idempotent matrices.

Then the residual sum of squares for fitting a linear line over $(\mathring{\tau}_j, \mathring{\tau}_{j+1}]$ (on which f_t is linear as well) is

$$(\mathbf{f} + \boldsymbol{\varepsilon})'\{\mathbf{D}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]} - \mathbf{H}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]}\}(\mathbf{f} + \boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}'\{\mathbf{D}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]} - \mathbf{H}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]}\}\boldsymbol{\varepsilon}.$$

It then follows that

$$\begin{aligned} T\hat{\sigma}_k^2 \ge T\mathring{\sigma}_k^2 &= \sum_{j=0}^{\hat{q}_k + q} \boldsymbol{\varepsilon}'\{\mathbf{D}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]} - \mathbf{H}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]}\}\boldsymbol{\varepsilon}. \\ &= \sum_{t=1}^{T} \varepsilon_t^2 - \sum_{j=0}^{\hat{q}_k + q} \boldsymbol{\varepsilon}'\mathbf{H}_{(\mathring{\tau}_j, \mathring{\tau}_{j+1}]}\boldsymbol{\varepsilon}. \end{aligned}$$

Note that $\boldsymbol{\varepsilon}'\mathbf{H}_{(s,e]}\boldsymbol{\varepsilon}$ follows a χ_2^2 distribution. For any $Z \sim \chi_2^2$, $\mathbb{P}(Z > z) \le e^{-z/2}$. Therefore, by defining the set

$$G_T = \Big\{ \max_{s,e:0 \le s < e \le T} \boldsymbol{\varepsilon}'\mathbf{H}_{(s,e]}\boldsymbol{\varepsilon} \le 6 \log T \Big\},$$

we have that $\mathbb{P}(G_T^c) = O(T^{-1})$ using the Bonferroni bound. Now assume that G_T holds, it follows that

$$T\hat{\sigma}_k^2 \ge \sum_{t=1}^{T} \varepsilon_t^2 - 6(\hat{q}_k + q + 1)\log T$$

This means that for all k with $\hat{q}_k > q$,

$$\begin{aligned} \mathrm{sSIC}(k) - \mathrm{sSIC}(k^*) &\ge T(\mathring{\sigma}_k^2 - \hat{\sigma}_{k^*}^2) + 2(\hat{q}_k - q)\log^\alpha(T) \\ &\ge 2(\hat{q}_k - q)\log^\alpha(T) - \{6(\hat{q}_k + q + 1) + M'\}\log T \\ &= 2(\hat{q}_k - q)\{\log^\alpha(T) - 3\log T\} - (12q + 6 + M')\log T \\ &\ge 2\log^\alpha(T) - (12q + 12 + M')\log T > 0 \end{aligned}$$

for large enough T, which in turn implies $\mathbb{P}(\hat{q} > q) \to 0$.

In conclusion, we have established that $\mathbb{P}(\hat{q} = q) \to 1$.

Part III. Estimation of the change-point locations

In view of the conclusion of Part II, in the rest of the proof we could assume that $A_T \cap B_T \cap D_T \cap E_T \cap G_T$ holds and $\hat{q} = q$.

Suppose that the model picked via NOT with the sSIC is $\hat{\tau}_1, \ldots, \hat{\tau}_q \in \mathcal{T}(\zeta^{(\hat{k})})$. Comparing the residual sum of squares of this candidate with $\mathcal{T}(\zeta^{(k^*)})$ yields that $\hat{\tau}_j \in \{\tau_j - \lfloor \delta_T/6 \rfloor + 1, \ldots, \tau_j + \lfloor \delta_T/6 \rfloor - 1\}$. It is because otherwise one could find an interval of length roughly $\delta_T/3$ (i.e. $\sim T$) with a true kink in the middle of but with no kinks in its estimates, leading to $\hat{\sigma}^2 > 1 + \delta$ for some $\delta > 0$ (see Lemma 5.3 and 5.4 of Liu *et al.* (1997)), and thus a contradiction (as the sSIC would clearly prefer $\mathcal{T}(\zeta^{(k^*)})$). Likewise, since $\hat{q} = q$, it is easy to see that $\hat{\tau}_j$ is the only estimated kink over $\big(\tau_j - \lceil \delta_T/3 \rceil - 2, \tau_j + \lceil \delta_T/3 \rceil\big]$ for every $j = 1, \ldots, q$.

Let

$$j^* = \text{argmax}_{j=1,\ldots,q} |\hat{\tau}_j - \tau_j|.$$

Now consider a "near-saturated oracle" candidate model with $2q+1$ kinks at

$$\{\hat{\tau}_1, \ldots, \hat{\tau}_q, \tau_1, \ldots, \tau_{j^*-1}, \tau_{j^*+1}, \ldots, \hat{\tau}_q, \tau_{j^*} - \lceil \delta_T/3 \rceil - 2, \tau_{j^*} + \lceil \delta_T/3 \rceil + 1\}$$

with the corresponding estimated variance of the errors denoted as $\dot{\sigma}^2_{\hat{k}}$. So again, instead of adding all the true kinks to the set of estimated kinks as before (which generates the so-called "saturated oracle"), we add all true kinks apart from τ_{j^*}, and replace it by $\tau_{j^*} - (\lceil \delta_T/3 \rceil + 2)$ and $\tau_{j^*} + (\lceil \delta_T/3 \rceil + 1)$.

Note that $\dot{\sigma}^2_{\hat{k}}$ is no smaller than the estimated variance of the errors from a model with the features at the same $2q+1$ locations, but with the continuity constraint only enforced at $\hat{\tau}_{j^*}$. More precisely, in the rest of the proof we could effectively follow a model with the signal following Scenario (S2) over $\{\tau_{j^*} - \lceil \delta_T/3 \rceil - 1, \ldots, \tau_{j^*} + \lceil \delta_T/3 \rceil + 1\}$ and Scenario (S3) elsewhere.

In addition, for any $1 \le s+1 < b < e \le T$,

$$\begin{aligned}
&\left\| \mathbf{Y}|_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]} \rangle \boldsymbol{\phi}^b_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\gamma}_{(s,e]} \rangle \boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{Y}, \mathbf{1}_{(s,e]} \rangle \mathbf{1}_{(s,e]} \right\|^2 \\
&= \left\| \mathbf{Y}|_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\gamma}_{(s,e]} \rangle \boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{Y}, \mathbf{1}_{(s,e]} \rangle \mathbf{1}_{(s,e]} \right\|^2 - \langle \mathbf{Y}, \boldsymbol{\phi}^b_{(s,e]} \rangle^2 \\
&= \left\| \mathbf{Y}|_{(s,e]} - \langle \mathbf{Y}, \boldsymbol{\gamma}_{(s,e]} \rangle \boldsymbol{\gamma}_{(s,e]} - \langle \mathbf{Y}, \mathbf{1}_{(s,e]} \rangle \mathbf{1}_{(s,e]} \right\|^2 - (\mathcal{C}^b_{(s,e]}(\mathbf{Y}))^2
\end{aligned}$$

Applying this result on $s = \tau_{j^*} - \lceil \delta_T/3 \rceil - 2$, $e = \tau_{j^*} + \lceil \delta_T/3 \rceil + 1$, $b = \tau_{j^*}$ or $\hat{\tau}_{j^*}$, and using the argument similar to that in Part II, we obtain that

$$\begin{aligned}
T\hat{\sigma}^2_{\hat{k}} \ge T\dot{\sigma}^2_{\hat{k}} \ge &\sum_{t=1}^{\tau_{j^*} - \lceil \delta_T/3 \rceil - 2} \varepsilon_t^2 + \sum_{t=\tau_{j^*} + \lceil \delta_T/3 \rceil + 2}^{T} \varepsilon_t^2 - (2q) 6 \log T \\
&+ (\mathcal{C}^{\tau_{j^*}}_{(s,e]}(\mathbf{Y}))^2 - (\mathcal{C}^{\hat{\tau}_{j^*}}_{(s,e]}(\mathbf{Y}))^2 + \Big(\sum_{\tau_{j^*} - \lceil \delta_T/3 \rceil - 1}^{\tau_{j^*} + \lceil \delta_T/3 \rceil + 1} \varepsilon_t^2 - 12 \log T \Big),
\end{aligned}$$

where $\sum_{\tau_{j^*} - \lceil \delta_T/3 \rceil - 1}^{\tau_{j^*} + \lceil \delta_T/3 \rceil + 1} \varepsilon_t^2 - 12 \log T$ is the lower-bound of the residual sum of squares for fitting a piecewise-linear function over $\{\tau_{j^*} - \lceil \delta_T/3 \rceil - 1, \ldots, \tau_{j^*} + \lceil \delta_T/3 \rceil + 1\}$ with only one feature at τ_{j^*}. Consequently, it follows from an argument similar to that in Step Four of the proof of Theorem 1 that

$$\begin{aligned}
T\hat{\sigma}^2_{\hat{k}} &\ge \sum_{t=1}^{T} \varepsilon_t^2 - 6(2q+2) \log T + (\mathcal{C}^{\tau_{j^*}}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^{\hat{\tau}_{j^*}}_{(s,e]}(\mathbf{f}))^2 \\
&\quad - 2\sqrt{8 \log T} \sqrt{(\mathcal{C}^{\tau_{j^*}}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^{\hat{\tau}_{j^*}}_{(s,e]}(\mathbf{f}))^2} - 8 \log T \\
&= \sum_{t=1}^{T} \varepsilon_t^2 - 6(2q+2) \log T + \left(\sqrt{(\mathcal{C}^{\tau_{j^*}}_{(s,e]}(\mathbf{f}))^2 - (\mathcal{C}^{\hat{\tau}_{j^*}}_{(s,e]}(\mathbf{f}))^2} - \sqrt{8 \log T} \right)^2 - 16 \log T
\end{aligned}$$

Using the fact that $|\hat{\tau}_{j^*} - \tau_{j^*}| < \delta_T/6 \le (\lceil \delta_T/3 \rceil + 1)/2$ and Lemma 8, we have that in the case where $\frac{\underline{C}_1\underline{C}_2^2}{144T}|\hat{\tau}_{j^*} - \tau_{j^*}|^2 \geq 8\log T$,

$$T\hat{\sigma}^2_{\hat{k}} \geq \sum_{t=1}^{T} \varepsilon_t^2 - (12q+28)\log T + \Big(\frac{\underline{C}_2}{\sqrt{48T}}(\underline{C}_1 T/3 + 1 - 1)^{1/2}|\hat{\tau}_{j^*} - \tau_{j^*}| - \sqrt{8\log T}\Big)^2$$
$$= \sum_{t=1}^{T} \varepsilon_t^2 - (12q+28)\log T + \Big(\sqrt{\frac{\underline{C}_1\underline{C}_2^2}{144T}}|\hat{\tau}_{j^*} - \tau_{j^*}| - \sqrt{8\log T}\Big)^2.$$

However,

$$T\hat{\sigma}^2_{\hat{k}} \leq T\hat{\sigma}^2_{k^*} \leq \sum_{t=1}^{T} \varepsilon_t^2 + M' \log T.$$

Combining the above two inequalities, and after some algebraic manipulations, we get

$$|\hat{\tau}_{j^*} - \tau_{j^*}| \leq \frac{12}{\sqrt{\underline{C}_1}\underline{C}_2}(\sqrt{M' + 12q + 28} + \sqrt{8})\sqrt{T\log T} =: C\sqrt{T\log T}.$$

On the other hand, in the case where $\frac{\underline{C}_1\underline{C}_2^2}{144T}|\hat{\tau}_{j^*} - \tau_{j^*}|^2 < 8\log T$, we directly have that

$$|\hat{\tau}_{j^*} - \tau_{j^*}| < \frac{24\sqrt{2}}{\sqrt{\underline{C}_1}\underline{C}_2}\sqrt{T\log T} < C\sqrt{T\log T}.$$

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

I.6. Proof of Corollary 1

Proof. We set $P := \sum_{k=-\infty}^{\infty} |\rho_k|$, where ρ_k is the auto-correlation function of $\{\varepsilon_t\}$. Now we modify our proof of Theorem 1 as follows:

Step One and Two

Let $\lambda_T = \sqrt{8P\log T}$ and define the set A_T as before. Denote the autocorrelation matrix of $\{\varepsilon_t\}$ by $\mathbf{P}_T = [\rho_{i-j}]_{i,j=1,\dots,T}$ (which is also the autocovariance matrix, since ε_t has unit-variance). Then since $\mathbf{P}_T$ is symmetric, we have that

$$\|\mathbf{P}_T\|_\infty = \|\mathbf{P}_T\|_1 = \max_j \sum_i |P_{ij}| \leq P,$$

where $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are the operator norms of a matrix. Consequently, by Hölder's inequality, $\|\mathbf{P}_T\|_2 \leq \sqrt{\|\mathbf{P}_T\|_1 \, \|\mathbf{P}_T\|_\infty} \leq P$, i.e., the largest eigenvalue of $\mathbf{P}_T$ is bounded above by P, which is irrelevant of T.

For any s, b, e such that $0 \leq s < b < e \leq T$, since $\left\langle \boldsymbol{\psi}^b_{(s,e]}, \boldsymbol{\varepsilon} \right\rangle$ has a normal distribution, with zero-mean and

$$\mathrm{Var}(\langle \boldsymbol{\psi}^b_{(s,e]}, \boldsymbol{\varepsilon} \rangle) = (\boldsymbol{\psi}^b_{(s,e]})^T \mathbf{P}_T \boldsymbol{\psi}^b_{(s,e]} \leq P\|\boldsymbol{\psi}^b_{(s,e]}\|_2^2 \leq P,$$

we have that

$$\mathbb{P}\left(|\mathcal{C}^b_{(s,e]}(\boldsymbol{\varepsilon})| \geq \lambda_T\right) = \mathbb{P}\left(|\mathcal{C}^b_{(s,e]}(\boldsymbol{\varepsilon})|/\sqrt{P} \geq \sqrt{8\log T}\right) \leq \frac{2e^{-8\log T/2}}{\sqrt{8\log T}\sqrt{2\pi}}.$$

It follows from the Bonferroni bound that $\mathbb{P}\left(A_T^c\right) \leq 12\sqrt{\pi}T^{-1}$.

Using the same argument as above, we can show that for any $0 \leq s < b < e \leq T$, $\frac{\left\langle \boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle, \boldsymbol{\varepsilon}\right\rangle}{\|\boldsymbol{\psi}^b_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^b_{(s,e]}\rangle - \boldsymbol{\psi}^{\tau_j}_{(s,e]}\langle \mathbf{f}, \boldsymbol{\psi}^{\tau_j}_{(s,e]}\rangle\|_2}$ is normal distributed, with zero-mean and variance bounded above by P. Thus, $\mathbb{P}\left(B_T^c\right) \leq 12\sqrt{\pi}T^{-1}$.

Step Three, Four and Five
The rest of the proof goes through by simply changing the constants as

$$\underline{C} = \sqrt{6}\left(2\sqrt{C_3}+\sqrt{32P}\right)+1, \quad C_1 = 2\sqrt{C_3}+\sqrt{8P}, \quad C_2 = \frac{1}{\sqrt{6}}-\frac{\sqrt{8P}}{\underline{C}}, \quad C_3 = (32\sqrt{2}+48)P$$

and setting

$$\eta_T = (C_1 - \sqrt{8P})^2.$$

□

Finally, we remark that the proof of Corollary 2 is similar to that of Corollary 1, so is omitted for brevity.

References

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016). not: Narrowest-over-threshold change-point detection. URL `https://cran.r-project.org/web/packages/not`. R package version 1.0.

Dümbgen, L. and Spokoiny, V. G. (2001) Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29**, 124–152.

Fryzlewicz, P., Sapatinas, T. and Rao, S. S. (2006). A Haar–Fisz technique for locally stationary volatility estimation. *Biometrika*, **93**, 687–704.

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

Fryzlewicz, P. (2018). Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Preprint*, URL `http://stats.lse.ac.uk/fryzlewicz/wbs2/wbs2.pdf`.

Liu. J., Wu, S. and Zidek, J. V. (1997). On segmented nultivariate regression. *Statistica Sinica*, **7**, 497–525.

McTaggart, R., Daroczi, G. and Leung, C. (2016). Quandl: Api wrapper for quandl.com. URL `https://CRAN.R-project.org/package=Quandl`. R package version 2.8.0.

Mikosch, T. and Stărică, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics*, **86**, 378–390.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics*, **19**, 175–190.

Zou, C. and Lancezhange. (2014). nmcdr: Non-parametric multiple change-points detection. URL `https://CRAN.R-project.org/package=nmcdr`. R package version 0.3.0.

Zou, C., Yin, G., Feng, L. and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics*, **42**, 970–1002.