# Supplementary materials for "Detecting linear trend changes in data sequences"

Hyeyoung Maeng[*] and Piotr Fryzlewicz[†]

This document includes the following sections:

**A.** Proofs

**B.** Extension to dependent non-Gaussian noise

**C.** Additional simulation results

**D.** Additional data application results

**E.** Shape of the unbalanced wavelet basis

**F.** A practical way to implement the TGUW transformation

**G.** Extension to piecewise-quadratic signal

## A Proofs

### A.1 Some useful lemmas for Theorems 1-3 of the main article

**Lemma 1** Let the distribution of $\varepsilon_t$ in model (1) of the main article be iid standard Gaussian. Let $\psi^{(j,k)} = \sum_{i=1}^{I^{(j,k)}} \phi_i^{(j,k)} g_i^{(j,k)}$ where $\phi_i^{(j,k)}$ are constants and $g_i^{(j,k)}$ are vectors of equal length with $\psi^{(j,k)}$ where $I^{(j,k)} \in \{3,4\}, j = 1,\dots,J,\ k = 1,\dots,K(j)$. If we define the set $G = \{g_l\}$ where there is a unique correspondence between $\{g_i^{(j,k)}{}_{i=1,\dots,I^{(j,k)},j=1,\dots,J,k=1,\dots,K(j)}\}$ and $\{g_l\}$, we then have $P(A_T) \geq 1 - C_2 T^{-1}$ where

$$A_T = \left\{ \max_{g_l \in G} |g_l^\top \boldsymbol{\varepsilon}| \leq \lambda \right\}, \tag{1}$$

$\lambda$ is as in Theorem 1 and $C_2$ is a positive constant.

**Proof.** We firstly show that for any fixed $(j,k)$, $g_i^{(j,k)}$ and $\phi_i^{(j,k)}$ satisfy the conditions, $(g_i^{(j,k)})^\top g_i^{(j,k)} = 1$, $(g_i^{(j,k)})^\top g_{i'}^{(j,k)} = 0$ and $\sum_i (\phi_i^{(j,k)})^2 = 1$, where $\psi^{(j,k)} = \sum_{i=1}^{I^{(j,k)}} \phi_i^{(j,k)} g_i^{(j,k)}$. Depend-

[*] Department of Mathematical Sciences, Durham University. Email: hyeyoung.maeng@durham.ac.uk

[†] Department of Statistics, London School of Economics. Email: p.fryzlewicz@lse.ac.uk

ing on the type of merge, $\psi^{(j,k)}$ fall into one of the followings,

$$\text{Type 1: } \psi^{(j,k)}_{p,q,r} = \alpha_1 e_p + \alpha_2 e_{p+1} + \alpha_3 e_{p+2}, \tag{2}$$

$$\text{Type 2: } \psi^{(j,k)}_{p,q,r} = \beta_1 e_p + \beta_2(\underbrace{0,\dots,0}_{p\times 1}, \boldsymbol{\ell}^\top_{1,p+1,r}, \underbrace{0,\dots,0}_{(T-r)\times 1}) + \beta_3(\underbrace{0,\dots,0}_{p\times 1}, \boldsymbol{\ell}^\top_{2,p+1,r}, \underbrace{0,\dots,0}_{(T-r)\times 1}), \tag{3}$$

$$\psi^{(j,k)}_{p,q,r} = \beta_4(\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}^\top_{1,p,r-1}, \underbrace{0,\dots,0}_{(T-r+1)\times 1}) + \beta_5(\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}^\top_{2,p,r-1}, \underbrace{0,\dots,0}_{(T-r+1)\times 1}) + \beta_6 e_r,$$

$$\text{Type 3: } \psi^{(j,k)}_{p,q,r} = \gamma_1(\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}^\top_{1,p,q}, \underbrace{0,\dots,0}_{(T-q)\times 1}) + \gamma_2(\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}^\top_{2,p,q}, \underbrace{0,\dots,0}_{(T-q)\times 1}) \tag{4}$$

$$+ \gamma_3(\underbrace{0,\dots,0}_{q\times 1}, \boldsymbol{\ell}^\top_{1,q+1,r}, \underbrace{0,\dots,0}_{(T-r)\times 1}) + \gamma_4(\underbrace{0,\dots,0}_{q\times 1}, \boldsymbol{\ell}^\top_{2,q+1,r}, \underbrace{0,\dots,0}_{(T-r)\times 1}),$$

where $e_i$ is a vector of length $T$ having 1 only at $i^{th}$ element and zero for the others. As will be shown in Section E, $\boldsymbol{\ell}_{1,i,j}$ and $\boldsymbol{\ell}_{2,i,j}$ are an arbitrary orthonormal basis of the subspace $\{(x_1, x_2, \dots, x_{j-i+1}) \mid x_1 - x_2 = x_2 - x_3 = \cdots = x_{j-i} - x_{j-i+1}\}$ of $\mathbb{R}^{j-i+1}$.

In any case, we can obtain the representation $\psi^{(j,k)} = \sum_{i=1}^{I^{(j,k)}} \phi_i^{(j,k)} g_i^{(j,k)}$ from (2) if the constants $\phi_i^{(j,k)}$ correspond to $\{\alpha_i\}_{i=1}^3$ in Type 1, $\{\beta_i\}_{i=1}^3$ or $\{\beta_i\}_{i=4}^6$ in Type 2 and $\{\gamma_i\}_{i=1}^4$ in Type 3 and $g_i^{(j,k)}$ is the corresponding vector. From the orthonormality of the basis $(\boldsymbol{\ell}_{1,m,n}, \boldsymbol{\ell}_{2,m,n})$ for any $(m, n)$, we see that the conditions, $(g_i^{(j,k)})^\top g_i^{(j,k)} = 1$ and $(g_i^{(j,k)})^\top g_{i'}^{(j,k)} = 0$, are satisfied for any $(i, i', j, k)$ where $i \neq i'$. In addition, as $\psi^{(j,k)}$ keep orthonormality, we can argue that $\phi_i^{(j,k)}$ is bounded by the condition $\sum_i (\phi_i^{(j,k)})^2 = 1$ for any $(i, j, k)$ which implies $\sum_{i=1}^3 \alpha_i^2 = \sum_{i=1}^3 \beta_i^2 = \sum_{i=4}^6 \beta_i^2 = \sum_{i=1}^4 \gamma_i^2 = 1$ in (2).

If we predefine the pairs $(\boldsymbol{\ell}_{1,m,n}, \boldsymbol{\ell}_{2,m,n})$ for any $(m, n)$ by choosing an orthonormal basis of the subspace $\{(x_1, x_2, \dots, x_{n-m+1}) \mid x_1 - x_2 = x_2 - x_3 = \cdots = x_{n-m} - x_{n-m+1}\}$ of $\mathbb{R}^{n-m+1}$, then there exist at most $T^2$ vectors $g_l$ in the set $G$. This is because $m$ and $n$ can be randomly chosen from $\{1, 2, \dots, T\}$ with replacement and if $m \neq n$, the two drawn pairs, $(m, n)$ and $(n, m)$, correspond to the same basis vectors, $(\boldsymbol{\ell}_{1,m,n}, \boldsymbol{\ell}_{2,m,n})$, while $(m, m)$ correspond to one vector $e_m$. Now we are in position to show that $P(A_T) \geq 1 - C_2 T^{-1}$. Using a simple Bonferroni inequality, we have

$$1 - P(A_T) \leq \sum_G P(|Z| > \lambda) \leq 2T^2 \frac{\phi_Z(\lambda)}{\lambda} = \frac{1}{C_1 \sqrt{\pi} T^{C_1^2 - 2} \sqrt{\log T}} \leq \frac{C_2}{T} \tag{5}$$

where $\phi_Z$ is the p.d.f. of a standard normal $Z$. This completes the proof.

**Lemma 2** Let $\mathcal{S}_j^1 = \{1 \leq k \leq K(j) : d^{(j,k)}$ is $d_{p,q,r}$ such that $p < \eta_i + 1/2 < r$ for some $i = 1, \dots, N$ $\}$, and $\mathcal{S}_j^0 = \{1, \dots, K(j)\} \setminus \mathcal{S}_j^1$. On the set $A_T$ in (1) which satisfies $P(A_T) \to 1$ as $T \to \infty$, we have

$$\max_{\substack{j=1,\dots,J,\\ k\in\mathcal{S}_j^0}} \left|d^{(j,k)}\right| \le \lambda, \tag{6}$$

where $\lambda$ is as in Theorem 1.

**Proof.** On the set $A_T$, the following holds for $j = 1, \dots, J, k \in \mathcal{S}_j^0$,

$$\begin{aligned}\left|d^{(j,k)}\right| &= \left|(\psi^{(j,k)})^\top \boldsymbol{\varepsilon}\right| \\ &= \left|\phi_1^{(j,k)}(g_1^{(j,k)})^\top \boldsymbol{\varepsilon} + \phi_2^{(j,k)}(g_2^{(j,k)})^\top \boldsymbol{\varepsilon} + \phi_3^{(j,k)}(g_3^{(j,k)})^\top \boldsymbol{\varepsilon} + \phi_4^{(j,k)}(g_4^{(j,k)})^\top \boldsymbol{\varepsilon}\right| \\ &\le \max_{j,\,k}\left(\left|\phi_1^{(j,k)}\right| + \left|\phi_2^{(j,k)}\right| + \left|\phi_3^{(j,k)}\right| + \left|\phi_4^{(j,k)}\right|\right)\cdot\left(\max_{l:\, g_l\in G}\ \left|g_l^\top \boldsymbol{\varepsilon}\right|\right),\end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_T)^\top$ and $\psi_{p,q,r}^{(j,k)}$ are as in (2). The condition, $\sum_i (\phi_i^{(j,k)})^2 = 1$ for any fixed $(j,k)$, given in the proof of Lemma 1 implies that $\max_i \left|\phi_i^{(j,k)}\right| \le 1$ for any $(j,k)$, thus we have (6) when the constant $C_1$ for $\lambda$ in (6) is larger than or equal to 4 times $C_1$ used in (1).

# B Extension to dependent non-Gaussian noise

In this section, we extend the TGUW methodology to more realistic settings when the noise $\varepsilon_t$ is possibly dependent and/or non-Gaussian. We borrow the idea proposed in the supplementary material of Fryzlewicz (2018) in the sense that the extension is performed in a way of altering the estimators $\tilde{f}, \tilde{\tilde{f}}$ and $\hat{f}$ and keeping the rate of threshold, $O((\log T)^{1/2})$, used in Theorems 1-3 of the main article established under the iid Gaussian noise. However, our technique is distinguished from Fryzlewicz (2018) in that we put an additional step which ensures that only the detail coefficients $d_{p,q,r}^{(j,k)}$ corresponding to a long enough interval $[p, r]$ are survived, while Fryzlewicz (2018) gives a condition that both the left ($[p, q]$) and the right ($[q + 1, r]$) segments should be long enough. This enables us to use the same size of threshold, $O((\log T)^{1/2})$, used in the iid Gaussian model without any further procedure such as basis rearrangement proposed in Fryzlewicz (2018).

We now define the sets of short-segment and long-segment coefficients at each scale $j$ as follows:

$$\begin{aligned}\mathcal{W}_j^S(a) &= \{1 \le k \le K(j) : d_{p,q,r}^{(j,k)} \text{ is such that } r - p \le a\}, \\ \mathcal{W}_j^L(a) &= \{1 \le k \le K(j)\} \setminus \mathcal{W}_j^S(a),\end{aligned} \tag{7}$$

where $a$ will be specified later this section. Those detail coefficients obtained from short segments are set to zero in the construction of the new estimators $\tilde{f}^L, \tilde{\tilde{f}}^L$ and $\hat{f}^L$, where $L$ in $f^L$ stands for "Long-segment". The initial estimator $\tilde{f}^L$ is obtained from the estimator of $\mu^{(j,k)}$ for $j \geq 1$ by applying the "connected" rule that is modified from the original one in Section 2.3 of the main article to satisfy the condition that the minimum segment length is longer than $a$:

$$\hat{\mu}^{(j,k)} = d^{(j,k)}_{p,q,r} \cdot \mathbb{I}\left\{\exists (j', k') \in C_{j,k} \quad \left|d^{(j',k')}_{p',q',r'}\right| > \lambda \quad \text{and} \quad k' \in \mathcal{W}^L_{j'}(a)\right\}, \tag{8}$$

where $\mathbb{I}$ is an indicator function and

$$C_{j,k} = \{(j', k'), j' = 1, \ldots, j, k' = 1, \ldots, K(j') : d^{(j',k')}_{p',q',r'} \text{ is such that } [p', r'] \subseteq [p, r]\}.$$

We then apply the "two together" rule to (8) in which both of the paired detail coefficients (formed by Type 3 mergings) should be survived if at least one is survived as done in thresholding of the main article. Compared to the estimator $\hat{\mu}^{(j,k)}$ obtained under the iid Gaussian setting, the only added step is setting all short-segment coefficient $d^{(j,k)}_{p,q,r}$ to zero.

## B.1 Preparatory lemmas

**Lemma 3** Let the distribution of $\varepsilon_t$ in model (1) of the main article as in Theorem B.1. Then for a constant $C_3 > 0$ and $\lambda$ as in Theorem B.1, we have $P(A_T) \geq 1 - C_3 T^{-1}$ for a constant $C_3 > 0$, where

$$A^L_T = \left\{\forall 1 \leq t_1 \leq t_2 \leq T, \ \forall k \in \{1, 2\} \quad s.t. \quad t_2 - t_1 \geq C_1 \log T \quad \left|\sum_{t=t_1}^{t_2} \boldsymbol{\ell}^t_{k,t_1,t_2} \varepsilon_t\right| \leq \lambda\right\}, \tag{9}$$

and $\boldsymbol{\ell}^t_{k,t_1,t_2}$ is the $t$-th element of the vector $\boldsymbol{\ell}_{k,t_1,t_2}$ of length $t_2 - t_1 + 1$ and the pairs $(\boldsymbol{\ell}_{1,t_1,t_2}, \boldsymbol{\ell}_{2,t_1,t_2})$ are predetermined for any $(t_1, t_2)$ by choosing an orthonormal basis of the subspace $\{(x_1, x_2, \ldots, x_{t_2-t_1+1}) \mid x_1 - x_2 = x_2 - x_3 = \cdots = x_{t_2-t_1} - x_{t_2-t_1+1}\}$ of $\mathbb{R}^{t_2-t_1+1}$.

**Proof.** In the following, we consider the single sum $\sum_{t=1}^{a} w_t \varepsilon_t$ from the interval $[1, a]$ where $w_t = \boldsymbol{\ell}^t_{k,1,a}$ for a fixed $k \in \{1, 2\}$. The results can in principle be applied to an interval with different ends given that the length of the interval is at least $a$. Since $\varepsilon_t$ is $m$-dependent, we have $\alpha(l) = 0$ for $l > m$ where $\alpha(\cdot)$ is the $\alpha$-mixing coefficients of $\varepsilon_t$.

From Theorem 1.4 in Bosq (1998), if $m_2^2 < \infty$, for each $\epsilon > 0$ and for a constant $c > 0$, we obtain

$$P\left(\sqrt{a}\left|\sum_{t=1}^{a} w_t \varepsilon_t\right| > a\epsilon\right) \leq a_1 \exp\left(-\frac{q\epsilon^2}{25 m_2^2 + 5c\epsilon}\right) + a_2(k)\alpha\left(\left[\frac{a}{q+1}\right]\right)^{\frac{2k}{2k+1}}, \tag{10}$$

where

$$a_1 = 2\frac{a}{q} + 2\left(1 + \frac{\varepsilon^2}{25m_2^2 + 5c\epsilon}\right), \quad \text{with} \quad m_2^2 = \max_{1\le t\le a} E\left[\left(\sqrt{a}w_t\varepsilon_t\right)^2\right],$$
$$a_2(k) = 11n\left(1 + \left(\frac{5m_k}{\epsilon}\right)^{\frac{2k}{2k+1}}\right), \quad \text{with} \quad m_k = \max_{1\le t\le a} \left\| \sqrt{a}w_t\varepsilon_t\right\|_k.$$

The assumption $m_2^2 < \infty$ is reasonably achievable as we can show $m_2^2 = a\max_t(w_t^2)$ is bounded by a constant from two conditions given on $w_t$, 1)$\{w_1 - w_2 = \cdots = w_{a-1} - w_a\}$ and 2) $\sum_{t=1}^a w_t^2 = 1$.

By setting $\epsilon = \lambda/\sqrt{a}$, $a = C\log(T)$ and $\lambda = C_1\log^{1/2}T$ for large enough $C > 0$ and $C_1 > 0$, and setting $q = [c_1a]$ with a small $c_1$ (which gives $\left[\frac{a}{q+1}\right] \ge m + 1$), we have that $a_1$ is bounded by a constant and $\alpha([a/(q+1)]) = 0$, thus (10) can be bounded as

$$P\left(\left|\sum_{t=1}^{a} w_t\varepsilon_t\right| > \lambda\right) \le \exp\left(-\frac{q\frac{\lambda^2}{a}}{25m_2^2 + 5c\frac{\lambda}{\sqrt{a}}}\right) \le \exp\{-C_2\log T\} = T^{-C_2},$$

where $C_2 > 0$ is suitably large. Since there exist at most $T^2$ sub-intervals $[t_1, t_2]$, applying a simple Bonferroni inequality, we have

$$P\left(\forall 1 \le t_1 \le t_2 \le T,\ \forall k \in \{1,2\} \quad s.t. \quad t_2 - t_1 \ge C\log T \quad \left|\sum_{t=t_1}^{t_2} \ell^t_{k,t_1,t_2}\varepsilon_t\right| \le \lambda\right) \ge 1 - \frac{C_3}{T}$$

as $T \to \infty$ for a large enough $C > 0$ and a certain constant $C_3 > 0$.

**Lemma 4** Let $\mathcal{S}_j^1$ and $\mathcal{S}_j^0$ as in Lemma 2. On the set $A_T^L$ in (9) that satisfies $P(A_T^L) \to 1$ as $T \to \infty$, we have

$$\max_{\substack{j=1,\dots,J,\\ k\in\mathcal{S}_j^0}} \left|d^{(j,k)}\right| \le \lambda,$$

where $\lambda$ is as in Theorem B.1.

**Proof.** The argument follows the proof of Lemma 2.

## B.2 Theoretical results of the length-lowerbounded-basis estimators

We now describe the behaviour of the initial estimator $\tilde{f}^L$ that is built from the basis vectors whose non-zero elements have length larger than $a$.

**Theorem B.1** Let the distribution of $\varepsilon_t$ in model (1) of the main article as follows:

(a) $\varepsilon_t$ has mean zero and satisfies Cramer's conditions that

$$E|\varepsilon_t|^k \leq c^{k-2}k!E(\varepsilon_t^2) < \infty, \quad t = 1, \ldots, T, \quad k = 3, 4, \ldots,$$

where $c > 0$.

(b) $\{\varepsilon_t\}_t$ is the stationary sequence and $m$-dependent i.e. $\sigma(\varepsilon_s, s \leq t)$ and $\sigma(\varepsilon_s, s \geq t + k)$ are independent for $k > m$.

Let $\bar{f} = \max_t f_t - \min_t f_t$ be bounded and let the estimator $\tilde{f}^L$ is obtained from the estimator $\hat{\mu}^{(j,k)}$ in (8), with $a = C \log(T)$ and the threshold $\lambda = C_1 \log^{1/2}(T)$, for large enough $C$ and $C_1$. Then on the set $A_T^L$ in (9), we have

$$\|\tilde{f}^L - f\|_T^2 \;\leq\; \tilde{C}\,\frac{1}{T}\,N\;\log^2(T)\,\lceil\,\log(T)/\log(1-\rho)^{-1}\,\rceil,$$

for a constant $\tilde{C} > 0$.

**Proof.** Let $\mathcal{S}_j^1$ and $\mathcal{S}_j^0$ as in Lemma 2. From the conditional orthonormality of the unbalanced wavelet transform, on the set $A_T^L$ in (9), we have

$$\begin{aligned}
\|\tilde{f}^L - f\|_T^2 \;&= \frac{1}{T}\sum_{j=1}^{J}\sum_{k=1}^{K(j)}\Big(d^{(j,k)}\cdot\mathbb{I}\{\,\exists(j',k')\in C_{j,k}\quad |d^{(j',k')}| > \lambda\;\text{ and }\;k'\in\mathcal{W}_{j'}^L(a)\,\} - \mu^{(j,k)}\Big)^2\\
&\quad+\;T^{-1}(s_{1,T}^{[1]} - \mu^{(0,1)})^2 +\; T^{-1}(s_{1,T}^{[2]} - \mu^{(0,2)})^2\\
&\leq \frac{1}{T}\sum_{j=1}^{J}\Bigg(\sum_{k\in\mathcal{S}_j^0} + \sum_{k\in\mathcal{S}_j^1\cap\mathcal{W}_j^S(a)} + \sum_{k\in\mathcal{S}_j^1\cap\mathcal{W}_j^L(a)}\Bigg)\\
&\quad\Big(d^{(j,k)}\cdot\mathbb{I}\{\,\exists(j',k')\in C_{j,k}\quad |d^{(j',k')}| > \lambda\;\text{ and }\;k'\in\mathcal{W}_{j'}^L(a)\,\} - \mu^{(j,k)}\Big)^2\\
&\quad+\;4C_1^2T^{-1}\log T\\
&=: I + II + III + 4C_1^2T^{-1}\log T,
\end{aligned} \tag{11}$$

where $\mu^{(0,1)} = \langle f, \psi^{(0,1)}\rangle$, $\mu^{(0,2)} = \langle f, \psi^{(0,2)}\rangle$ and $\mathcal{W}_{j'}^S(a)$ and $\mathcal{W}_{j'}^L(a)$ are as in (7). We note that $(s_{1,T}^{[1]} - \mu^{(0,1)})^2 \leq 2C_1^2 \log T$ is simply obtained by combining Lemma 4 and the fact that $s_{1,T}^{[1]} - \mu^{(0,1)} = \langle \boldsymbol{\varepsilon}, \psi^{(0,1)}\rangle$, which can also be applied to obtain $(s_{1,T}^{[2]} - \mu^{(0,2)})^2 \leq 2C_1^2 \log T$. We now examine the terms $I, II$ and $III$ in (11).

**Term** $I$: By Lemma 4, on the set $A_T^L$, $\mathbb{I}\{\,\exists(j',k') \in C_{j,k}\quad |d^{(j',k')}| > \lambda\} = 0$ for $k \in \mathcal{S}_j^0$ if $k' \in \mathcal{W}_{j'}^L(a)$. Also by the fact that $\mu^{(j,k)} = 0$ for $j = 1, \ldots, J, k \in \mathcal{S}_j^0$, we obtain $I = 0$.

**Term** *II*: As there is no short-segment parent coefficient whose children is from long-segment due to the principle of bottom-up merging, the indicator function returns zero and the term *II* is simplified to $\frac{1}{T}\sum_{j=1}^{J}\sum_{k\in\mathcal{S}_j^1\cap\mathcal{W}_j^S(a)}\left(\mu^{(j,k)}\right)^2$.

We now examine the bound of individual $\mu_{p,q,r}^{(j,k)}$. Note that only Type 2 and Type 3 basis vectors are considered due to the minimum length constraint given on the set $A_T^L$. Borrowing the generalised form of $\psi_{p,q,r}^{(j,k)}$ in (2), for Type 3 basis vector, we obtain

$$\begin{aligned}\mu_{p,q,r}^{(j,k)} = \langle f, \psi_{p,q,r}^{(j,k)}\rangle &= \gamma_1\boldsymbol{\ell}_{1,p,q}^{\top}f_{p:q} + \gamma_2\boldsymbol{\ell}_{2,p,q}^{\top}f_{p:q} + \gamma_3\boldsymbol{\ell}_{1,q+1,r}^{\top}f_{q+1:r} + \gamma_4\boldsymbol{\ell}_{2,q+1,r}^{\top}f_{q+1:r}\\ &\leq \gamma_1\|f_{p:q}\| + \gamma_2\|f_{p:q}\| + \gamma_3\|f_{q+1:r}\| + \gamma_4\|f_{q+1:r}\|, \qquad (12)\end{aligned}$$

where $f_{p:q}$ is the subvector of $f$ containing $q-p+1$ elements. The inequality (12) is obtained from the orthonormality of $\boldsymbol{\ell}_{1,p,q}, \boldsymbol{\ell}_{2,p,q}, \boldsymbol{\ell}_{1,q+1,r}, \boldsymbol{\ell}_{2,q+1,r}$ and the definition of inner product $a\cdot b = \|a\|\cdot\|b\|\cdot\cos(\theta)$, where $\theta$ is the angle between $a$ and $b$. Note that if $f_{p:q}$ does not contain a change point, the corresponding $\cos(\theta) = 0$ as $f_{p:q}$ has a perfect linear trend and in the case when $f_{p:q}$ includes a change point, the size of angle is bounded as $|\cos(\theta)| \leq 1$. As $\bar{f} = \max_t f_t - \min_t f_t$ is assumed to be bounded and $\|f_{p:q}\|^2 \leq C[(q-p+1)^2 + \bar{f}^2]$ regardless of whether there exists a true change point in $[p,q]$, we have

$$\begin{aligned}(\mu_{p,q,r}^{(j,k)})^2 \leq& (\gamma_1+\gamma_2)^2\cdot\|f_{p:q}\|^2 + (\gamma_3+\gamma_4)^2\cdot\|f_{q+1:r}\|^2 + 2\cdot(\gamma_1+\gamma_2)\cdot(\gamma_3+\gamma_4)\cdot\|f_{p:q}\|\cdot\|f_{q+1:r}\|\\ \leq& c_1[(q-p+1)^2 + (f_q - f_p)^2] + c_2[(r-q)^2 + (f_r - f_{q+1})^2]\\ &+ c_3\sqrt{(q-p+1)^2 + (f_q-f_p)^2}\cdot\sqrt{(r-q)^2 + (f_r - f_{q+1})^2}\\ \leq& c_4(r-p+1)^2 + c_5\bar{f}^2\\ \leq& C(r-q+1)^2\end{aligned}$$

where $c_i > 0$ and $C > 0$. Without loss of generality, we assume $r-q+1 \leq a$, then using $a = O(\log(T))$ and applying the same upper bounds, $J \leq \lceil \log(T)/\log((1-\rho)^{-1}) + \log(2)/\log(1-\rho)\rceil$ and $|\mathcal{S}_j^1| \leq N$, used in the proof of Theorem 1 in the main article, we obtain

$$II \leq \frac{1}{T}N\lceil\log(T)/\log((1-\rho)^{-1}) + \log(2)/\log(1-\rho)\rceil(\log(T))^2$$

**Term** *III*: Denote $\mathcal{B} = \{\exists(j',k') \in C_{j,k} \quad |d^{(j',k')}| > \lambda \quad \text{and} \quad k' \in \mathcal{W}_{j'}^L(a)\}$ and on the set $A_T^L$ in

(9) we have

$$\begin{aligned}(d^{(j,k)} \cdot \mathbb{I}\{\mathcal{B}\} - \mu^{(j,k)})^2 &= (d^{(j,k)} \cdot \mathbb{I}\{\mathcal{B}\} - d^{(j,k)} + d^{(j,k)} - \mu^{(j,k)})^2 \\ &\leq (d^{(j,k)})^2 \mathbb{I}\Big(\big|d^{(j',k')}\big| \leq \lambda \ \text{ or } \ k' \in \mathcal{W}^S_{j'}(a)\Big) + \ (d^{(j,k)} - \mu^{(j,k)})^2 \\ &\quad + \ 2\big|d^{(j,k)}\big| \, \mathbb{I}\Big(\big|d^{(j',k')}\big| \leq \lambda \ \text{ or } \ k' \in \mathcal{W}^S_{j'}(a)\Big) \, \big|d^{(j,k)} - \mu^{(j,k)}\big| \\ &\leq \lambda^2 + 2C_1^2 \log T + 2\lambda C_1 \{2 \log T\}^{1/2}. \end{aligned} \tag{13}$$

Following the same argument used in the proof of Theorem 1 in the main article, we have

$$III \leq \frac{1}{T} N \log(T) \lceil \log(T)/\log((1-\rho)^{-1}) + \log(2)/\log(1-\rho) \rceil.$$

To complete the proof, considering all terms in (11), we finally obtain

$$\|\tilde{f}^L - f\|_T^2 \ \leq \ \tilde{C}\, T^{-1}\, N\, (\log(T))^2\, \lceil (\log(T)/\log((1-\rho)^{-1}) + \log(2)/\log(1-\rho) \rceil, \tag{14}$$

where $\tilde{C} > 0$. Comparing it with Theorem 1 of the main article that is presented under the iid Gaussian noise assumption, the $\ell_2$ rate in (14) is different by only a logarithmic factor.

**Theorem B.2** $X_t$ follows model (1) with $\sigma = 1$. Let the distribution of $\varepsilon_t$ and the threshold $\lambda$ be as in Theorem B.1. Further, let $\bar{f} = \max_t f_t - \min f_t$ be bounded. Then we have $\left\|\tilde{\tilde{f}}^L - f\right\|_T^2 = O(NT^{-1} \log^3(T))$ with probability approaching 1 as $T \to \infty$, where $\tilde{\tilde{f}}^L$ is the estimator constructed from $\tilde{f}^L$ through Stage 1 of the post-processing described in Section 2.5 of the main article. And there exist at most two estimated change-points between each pair of true change-points $(\eta_i, \eta_{i+1})$ for $i = 0, \ldots, N$, where $\eta_0 = 0$ and $\eta_{N+1} = T$. Therefore $\tilde{\tilde{N}} \leq 2(N+1)$, where $\tilde{\tilde{N}}$ is the number of estimated change points in $\tilde{\tilde{f}}^L$.

**Proof.** The proof proceeds the same as the proof of Theorem 2 of the main article.

**Theorem B.3** $X_t$ follows model (1) with $\sigma = 1$. Let the distribution of $\varepsilon_t$ and the threshold $\lambda$ be as in Theorem B.1. Further, let the number of true change-points, $N$, have the order of $logT$ and let $\bar{f} = \max_t f_t - \min f_t$ be bounded. Let the estimators $\hat{f}^L$, $\hat{N}$ and $(\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}})$ are constructed through Stage 2 of the post-processing described in Section 2.5 of the main article. Let $\Delta_T = \min_{i=1,\ldots,N} \left\{ \left(\underline{f}_T^i\right)^{2/3} \cdot \delta_T^i \right\}$ where $\underline{f}_T^i = \min\left(|f_{\eta_{i+1}} - 2f_{\eta_i} + f_{\eta_{i-1}}|, |f_{\eta_{i+2}} - 2f_{\eta_{i+1}} + f_{\eta_i}|\right)$ and $\delta_T^i = \min\left(|\eta_i - \eta_{i-1}|, |\eta_{i+1} - \eta_i|\right)$. Assume that $T^{1/3} R_T^{1/3} = o\left(\Delta_T\right)$ where $\left\|\tilde{\tilde{f}}^L - f\right\|_T^2 = O_p(R_T)$ is as in Theorem B.2. Then we have

$$\mathbb{P}\left(\hat{N} = N, \quad \max_{i=1,\ldots,N} \left\{ |\hat{\eta}_i - \eta_i| \cdot \left(\underline{f}_T^i\right)^{2/3} \right\} \leq C T^{1/3} R_T^{1/3} \right) \ \to \ 1, \tag{15}$$

as $T \to \infty$ where $C$ is a constant.

**Proof.** The proof proceeds the same as the proof of Theorem 3 of the main article.

# C Threshold selection and additional simulation results

## C.1 Simulation results for non-Gaussian and/or dependent noise

In addition to the simulations in Section 4 of the main article, here we present the results for the cases when $\varepsilon_t$ is possibly dependent and/or non-Gaussian. Including the standard Gaussian noise, we consider the following six scenarios for $\varepsilon_t$:

(i) standard Gaussian,

(ii) iid $t_5$ distribution with unit-variance,

(iii) a stationary Gaussian AR(1) process of $\phi = 0.3$, with zero-mean and unit-variance,

(iv) the same setting as in (iii) except $\phi = 0.6$,

(v) a stationary AR(1) process of $\phi = 0.3$ with the noise term following $t_5$,

(vi) the same setting as in (v) except $\phi = 0.6$.

In summary, (ii) is iid but heavy-tailed, (iii) and (iv) are Gaussian AR(1) error with relatively mild and strong dependence, respectively, and (v) and (vi) are both heavy-tailed but different strength of dependence, where the summary of the simulation results can be found in Tables C.1-C.10.

Following the theoretical results presented in Sections B.1-B.2, we need to set the minimum segment length to be an order of $\log(T)$. As already used in the main paper, we set $\lfloor 0.9 \log(T) \rfloor$ as a default minimum segment length. We follow the Algorithm 1 introduced in the main paper and use $\lambda^{\text{Robust}}$ as a default threshold, as it is designed to work well in all circumstances.

The simulation results under this robust threshold selection are presented in Tables C.1-C.10 and TrendSegment generally outperforms over all scenarios of noise and over almost all simulation models considered in this paper. Among other competitors, only ID provides the option for heavy-tailed noise in their R package IDetect and other methods are set to their default settings.

Table C.1: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the noise term $\varepsilon_t \overset{\text{iid}}{\sim} t_5$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using an Intel Core i5 2.9 GHz CPU with 8 GB of RAM, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M1) | TS | 0 | 0 | 1 | **88** | 9 | 2 | 0 | 0.24 | 3.10 | 0.09 |
| | NOT | 0 | 0 | 0 | **92** | 6 | 2 | 0 | 0.20 | 2.51 | 0.22 |
| | ID | 0 | 0 | 0 | **91** | 9 | 0 | 0 | 0.14 | **1.69** | 0.01 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.10 | 4.40 | 3.22 |
| | CPOP | 0 | 0 | 0 | 78 | 12 | 9 | 1 | 0.13 | **1.44** | 0.04 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 2.63 | 10.61 | 0.35 |
| (M2) | TS | 0 | 0 | 4 | **83** | 9 | 2 | 2 | 0.13 | 2.05 | 0.24 |
| | NOT | 0 | 0 | 3 | **85** | 11 | 0 | 1 | 0.098 | **1.69** | 0.29 |
| | ID | 0 | 0 | 0 | **77** | 21 | 2 | 0 | 0.102 | **1.36** | 0.38 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.067 | 2.29 | 31.41 |
| | CPOP | 0 | 0 | 0 | 14 | 23 | 25 | 38 | 0.119 | **1.54** | 1.66 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0.752 | 4.69 | 2.18 |
| (M3) | TS | 0 | 0 | 8 | 81 | 8 | 2 | 1 | 0.04 | 4.43 | 0.29 |
| | NOT | 0 | 0 | 1 | **97** | 2 | 0 | 0 | 0.021 | **2.71** | 0.31 |
| | ID | 0 | 0 | 0 | 85 | 10 | 2 | 3 | 0.023 | **2.40** | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.010 | 5.20 | 28.83 |
| | CPOP | 0 | 0 | 0 | 32 | 25 | 24 | 19 | 0.039 | **2.51** | 13.06 |
| | BUP | 0 | 0 | 0 | 2 | 26 | 46 | 26 | 0.032 | 5.39 | 2.18 |
| (M4) | TS | 0 | 0 | 0 | **91** | 7 | 0 | 2 | 0.11 | 3.39 | 0.09 |
| | NOT | 0 | 0 | 0 | **98** | 2 | 0 | 0 | 0.08 | **2.57** | 0.24 |
| | ID | 0 | 0 | 0 | 87 | 12 | 1 | 0 | 0.08 | **2.21** | 0.01 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.05 | 5.54 | 8.73 |
| | CPOP | 0 | 0 | 0 | 62 | 22 | 8 | 8 | 0.08 | **2.24** | 0.38 |
| | BUP | 2 | 73 | 24 | 1 | 0 | 0 | 0 | 0.52 | 10.80 | 0.57 |

Table C.2: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods with the noise term $\varepsilon_t \overset{\text{iid}}{\sim} t_5$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using an Intel Core i5 2.9 GHz CPU with 8 GB of RAM, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N}-N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | | | | $\hat{N} - N$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M5) | TS | 0 | 0 | 2 | **75** | 18 | 4 | 1 | 0.04 | 2.17 | 0.31 |
| | NOT | 0 | 11 | 10 | 63 | 10 | 3 | 3 | 0.049 | **1.29** | 0.25 |
| | ID | 0 | 0 | 0 | 0 | 0 | 7 | 93 | 0.332 | 9.58 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.145 | 6.14 | 28.33 |
| | CPOP | 0 | 0 | 0 | 4 | 6 | 20 | 70 | 0.064 | 2.61 | 3.07 |
| | BUP | 0 | 0 | 0 | 32 | 44 | 20 | 4 | 0.097 | 4.62 | 2.22 |
| (M6) | TS | 0 | 4 | 1 | **88** | 3 | 1 | 3 | 0.02 | **1.23** | 0.35 |
| | NOT | 6 | 10 | 26 | 44 | 6 | 2 | 6 | 0.071 | 3.53 | 0.24 |
| | ID | 0 | 3 | 0 | 0 | 19 | 0 | 78 | 0.129 | 4.73 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.136 | 9.88 | 30.26 |
| | CPOP | 0 | 0 | 0 | 8 | 19 | 20 | 53 | 0.053 | 3.15 | 2.45 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.132 | 9.23 | 2.47 |
| (M7) | TS | 5 | 15 | 28 | **36** | 10 | 4 | 2 | 0.16 | 8.51 | 0.13 |
| | NOT | 0 | 6 | 16 | **30** | 36 | 11 | 1 | 0.079 | 5.12 | 0.22 |
| | ID | 6 | 3 | 9 | 18 | 19 | 15 | 30 | 0.385 | 12.40 | 0.01 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.098 | 6.08 | 23.86 |
| | CPOP | 0 | 0 | 0 | 0 | 4 | 5 | 91 | 0.102 | **3.01** | 0.81 |
| | BUP | 69 | 28 | 3 | 0 | 0 | 0 | 0 | 0.266 | 12.12 | 1.47 |
| (M8) | TS | 0 | 0 | 0 | **99** | 0 | 0 | 1 | 0.00 | **0.50** | 0.19 |
| | NOT | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.001 | **0.00** | 0.17 |
| | ID | 0 | 0 | 0 | **99** | 1 | 0 | 0 | 0.001 | **0.00** | 0.03 |
| | TF | 0 | 0 | 0 | 65 | 12 | 9 | 14 | 0.003 | 14.63 | 36.03 |
| | CPOP | 0 | 0 | 0 | 35 | 0 | 34 | 31 | 0.042 | 20.53 | 3.91 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.014 | 46.80 | 2.62 |

Table C.3: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the noise term $\epsilon_t$ being $AR(1)$ process of $\phi = 0.3$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using an Intel Core i5 2.9 GHz CPU with 8 GB of RAM, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M1) | TS | 0 | 1 | 13 | **82** | 4 | 0 | 0 | 0.39 | 3.65 | 0.07 |
| | NOT | 0 | 0 | 0 | **87** | 8 | 2 | 3 | 0.35 | **3.10** | 0.23 |
| | ID | 0 | 0 | 0 | 62 | 27 | 9 | 2 | 0.27 | **2.70** | 0.02 |
| | TF | 3 | 0 | 0 | 0 | 0 | 0 | 97 | 0.61 | 6.18 | 3.31 |
| | CPOP | 0 | 0 | 0 | 53 | 35 | 10 | 2 | 0.23 | **2.52** | 0.05 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 2.64 | 10.96 | 0.36 |
| (M2) | TS | 4 | 9 | 30 | 57 | 0 | 0 | 0 | 0.20 | 2.56 | 0.24 |
| | NOT | 0 | 0 | 8 | **83** | 6 | 2 | 1 | 0.182 | 2.11 | 0.31 |
| | ID | 0 | 0 | 0 | 69 | 24 | 5 | 2 | 0.155 | **1.75** | 0.40 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.600 | 2.38 | 32.03 |
| | CPOP | 0 | 0 | 0 | 1 | 6 | 8 | 85 | 0.163 | **1.98** | 1.50 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0.717 | 4.63 | 2.39 |
| (M3) | TS | 0 | 0 | 17 | 79 | 4 | 0 | 0 | 0.05 | 4.79 | 0.30 |
| | NOT | 0 | 0 | 1 | **89** | 7 | 2 | 1 | 0.045 | 3.81 | 0.32 |
| | ID | 0 | 0 | 1 | **83** | 14 | 1 | 1 | 0.037 | 2.98 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.258 | 6.24 | 28.76 |
| | CPOP | 0 | 0 | 0 | 76 | 10 | 9 | 5 | 0.022 | **2.14** | 15.35 |
| | BUP | 0 | 0 | 0 | 0 | 6 | 23 | 71 | 0.040 | 5.59 | 2.31 |
| (M4) | TS | 0 | 0 | 2 | **95** | 3 | 0 | 0 | 0.15 | 3.61 | 0.09 |
| | NOT | 0 | 0 | 0 | **86** | 9 | 3 | 2 | 0.16 | **3.50** | 0.23 |
| | ID | 0 | 0 | 0 | 84 | 14 | 0 | 2 | 0.13 | **2.87** | 0.01 |
| | TF | 1 | 0 | 1 | 0 | 1 | 0 | 97 | 0.64 | 6.76 | 8.67 |
| | CPOP | 0 | 0 | 0 | 51 | 24 | 15 | 10 | 0.11 | **3.17** | 0.39 |
| | BUP | 1 | 61 | 38 | 0 | 0 | 0 | 0 | 0.50 | 10.43 | 0.58 |

Table C.4: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods with the noise term $\epsilon_t$ being $AR(1)$ process of $\phi = 0.3$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using an Intel Core i5 2.9 GHz CPU with 8 GB of RAM, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M5) | TS | 0 | 0 | 0 | **84** | 15 | 1 | 0 | 0.05 | **1.85** | 0.32 |
| | NOT | 0 | 6 | 13 | **74** | 4 | 3 | 0 | 0.062 | **1.59** | 0.26 |
| | ID | 0 | 0 | 0 | 1 | 4 | 17 | 78 | 0.347 | 8.87 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.192 | 6.16 | 28.45 |
| | CPOP | 0 | 0 | 0 | 2 | 14 | 20 | 64 | 0.059 | **2.02** | 3.64 |
| | BUP | 0 | 0 | 0 | 11 | 32 | 30 | 27 | 0.131 | 5.19 | 2.32 |
| (M6) | TS | 0 | 6 | 0 | **93** | 1 | 0 | 0 | 0.02 | **1.34** | 0.35 |
| | NOT | 6 | 18 | 28 | 31 | 4 | 2 | 11 | 0.094 | 5.30 | 0.25 |
| | ID | 1 | 10 | 0 | 0 | 21 | 0 | 68 | 0.149 | 6.75 | 0.04 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.324 | 9.93 | 29.97 |
| | CPOP | 0 | 0 | 0 | 7 | 32 | 28 | 23 | 0.043 | **1.04** | 3.58 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.159 | 9.09 | 2.82 |
| (M7) | TS | 20 | 47 | 24 | 7 | 2 | 0 | 0 | 0.23 | 11.74 | 0.12 |
| | NOT | 5 | 12 | 19 | **24** | 22 | 7 | 11 | 0.158 | 7.69 | 0.24 |
| | ID | 11 | 3 | 15 | **22** | 18 | 16 | 15 | 0.405 | 14.22 | 0.01 |
| | TF | 3 | 0 | 0 | 0 | 0 | 0 | 97 | 0.623 | 7.01 | 23.25 |
| | CPOP | 0 | 0 | 0 | 0 | 0 | 1 | 99 | 0.162 | **5.27** | 0.85 |
| | BUP | 54 | 43 | 3 | 0 | 0 | 0 | 0 | 0.283 | 11.92 | 1.55 |
| (M8) | TS | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.00 | **0.00** | 0.19 |
| | NOT | 0 | 0 | 0 | **93** | 3 | 3 | 1 | 0.005 | **2.02** | 0.19 |
| | ID | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.003 | **0.00** | 0.51 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.551 | 49.94 | 35.81 |
| | CPOP | 0 | 0 | 0 | 30 | 10 | 3 | 57 | 0.035 | 19.71 | 7.55 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.025 | 46.73 | 2.72 |

Table C.5: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the noise term $\epsilon_t$ being $AR(1)$ process of $\phi = 0.6$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using 10 cores of Apple M1 Pro with 16 GB of RAM on mac OS, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M1) | TS | 2 | 2 | 22 | **67** | 7 | 0 | 0 | 0.82 | **4.82** | 0.07 |
| | NOT | 0 | 1 | 4 | **50** | 15 | 9 | 21 | 0.86 | **4.19** | 0.09 |
| | ID | 0 | 0 | 0 | 5 | 16 | 20 | 59 | 0.70 | **4.29** | 0.01 |
| | TF | 28 | 0 | 0 | 1 | 0 | 1 | 70 | 1.44 | 14.23 | 1.84 |
| | CPOP | 0 | 0 | 0 | 4 | 11 | 18 | 67 | 0.76 | **4.31** | 0.05 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 2.55 | 11.22 | 0.15 |
| (M2) | TS | 30 | 34 | 26 | 8 | 2 | 0 | 0 | 0.50 | 3.69 | 0.23 |
| | NOT | 0 | 4 | 13 | 21 | 19 | 18 | 25 | 0.51 | 2.94 | 0.14 |
| | ID | 2 | 5 | 4 | **33** | 23 | 17 | 16 | 0.41 | 3.20 | 0.02 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 1.23 | **2.38** | 12.86 |
| | CPOP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.54 | **2.48** | 0.87 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0.70 | 4.43 | 0.70 |
| (M3) | TS | 0 | 4 | 23 | **45** | 16 | 8 | 4 | 0.14 | 6.77 | 0.31 |
| | NOT | 0 | 0 | 4 | 24 | 7 | 6 | 59 | 0.21 | **5.95** | 0.20 |
| | ID | 1 | 5 | 11 | 28 | 11 | 16 | 28 | 0.12 | 6.08 | 0.04 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.58 | 6.23 | 19.54 |
| | CPOP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.38 | 6.08 | 3.31 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.13 | **5.92** | 1.22 |
| (M4) | TS | 0 | 1 | 16 | **57** | 18 | 4 | 4 | 0.42 | **5.63** | 0.09 |
| | NOT | 0 | 0 | 3 | 26 | 16 | 11 | 44 | 0.56 | **5.61** | 0.12 |
| | ID | 0 | 0 | 8 | 41 | 24 | 17 | 10 | 0.35 | **4.78** | 0.01 |
| | TF | 25 | 0 | 2 | 0 | 1 | 0 | 72 | 1.46 | 13.64 | 5.63 |
| | CPOP | 0 | 0 | 0 | 0 | 2 | 0 | 98 | 0.59 | 5.72 | 0.26 |
| | BUP | 1 | 37 | 58 | 4 | 0 | 0 | 0 | 0.57 | 9.45 | 0.31 |

Table C.6: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods with the noise term $\epsilon_t$ being $AR(1)$ process of $\phi = 0.6$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using 10 cores of Apple M1 Pro with 16 GB of RAM on mac OS, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | | | | $\hat{N} - N$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M5) | TS | 0 | 0 | 10 | **40** | 32 | 10 | 8 | 0.14 | **3.62** | 0.33 |
| | NOT | 0 | 3 | 10 | 11 | 11 | 10 | 55 | 0.22 | 4.46 | 0.18 |
| | ID | 2 | 3 | 1 | 4 | 6 | 5 | 79 | 0.42 | 7.28 | 0.04 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.40 | 6.18 | 18.88 |
| | CPOP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.47 | 5.96 | 2.18 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.24 | 5.90 | 1.24 |
| (M6) | TS | 0 | 3 | 0 | **65** | 12 | 11 | 9 | 0.07 | **3.09** | 0.36 |
| | NOT | 13 | 8 | 16 | 22 | 6 | 3 | 32 | 0.19 | 8.10 | 0.15 |
| | ID | 39 | 26 | 2 | 0 | 19 | 1 | 13 | 0.28 | 24.64 | 0.04 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.69 | 9.91 | 197.40 |
| | CPOP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.52 | 9.50 | 2.91 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.30 | 9.41 | 1.36 |
| (M7) | TS | 25 | 25 | 26 | 14 | 5 | 4 | 1 | 0.40 | 11.82 | 0.13 |
| | NOT | 3 | 4 | 4 | 11 | 13 | 3 | 62 | 0.49 | 8.03 | 0.11 |
| | ID | 10 | 3 | 9 | **20** | 19 | 15 | 24 | 0.47 | 13.15 | 0.02 |
| | TF | 24 | 0 | 2 | 0 | 0 | 0 | 74 | 1.47 | 13.14 | 9.90 |
| | CPOP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.59 | **7.02** | 0.44 |
| | BUP | 3 | 30 | 42 | **21** | 4 | 0 | 0 | 0.35 | 8.97 | 0.46 |
| (M8) | TS | 0 | 0 | 0 | 63 | 7 | 26 | 4 | 0.03 | 10.42 | 0.19 |
| | NOT | 0 | 0 | 0 | 7 | 8 | 4 | 81 | 0.19 | 37.77 | 0.10 |
| | ID | 0 | 0 | 0 | **96** | 3 | 0 | 1 | 0.01 | **0.61** | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 1.10 | 49.94 | 15.54 |
| | CPOP | 0 | 0 | 0 | 0 | 1 | 0 | 99 | 0.38 | 45.54 | 2.08 |
| | BUP | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.11 | 47.44 | 0.85 |

Table C.7: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the $\varepsilon_t$ being $AR(1)$ process of $\phi = 0.3$ with noise term following $t_5$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using 10 cores of Apple M1 Pro with 16 GB of RAM on mac OS, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M1) | TS | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.07 | **1.72** | 0.10 |
| | NOT | 0 | 0 | 0 | 90 | 7 | 2 | 1 | 0.06 | **1.61** | 0.09 |
| | ID | 0 | 0 | 0 | 49 | 29 | 9 | 13 | 0.06 | 2.08 | 0.01 |
| | TF | 17 | 0 | 0 | 0 | 1 | 0 | 82 | 0.13 | 9.94 | 1.64 |
| | CPOP | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.03 | **0.87** | 0.05 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 3.14 | 12.19 | 0.14 |
| (M2) | TS | 0 | 0 | 2 | **94** | 2 | 0 | 2 | 0.04 | 1.32 | 0.25 |
| | NOT | 0 | 0 | 0 | **95** | 4 | 1 | 0 | 0.03 | **1.10** | 0.16 |
| | ID | 0 | 0 | 0 | 47 | 25 | 14 | 14 | 0.08 | **1.18** | 0.02 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.14 | 2.38 | 12.86 |
| | CPOP | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.04 | **0.82** | 1.38 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 1.22 | 5.08 | 0.69 |
| (M3) | TS | 0 | 0 | 0 | **99** | 0 | 0 | 1 | 0.01 | 2.52 | 0.31 |
| | NOT | 0 | 0 | 0 | 88 | 9 | 2 | 1 | 0.01 | 2.04 | 0.24 |
| | ID | 0 | 0 | 0 | 56 | 29 | 9 | 6 | 0.01 | 2.42 | 0.02 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.04 | 6.23 | 20.08 |
| | CPOP | 0 | 0 | 0 | **99** | 0 | 1 | 0 | 0.00 | **0.67** | 21.78 |
| | BUP | 5 | 82 | 13 | 0 | 0 | 0 | 0 | 0.14 | 11.18 | 1.10 |
| (M4) | TS | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.03 | **2.00** | 0.10 |
| | NOT | 0 | 0 | 0 | 88 | 8 | 4 | 0 | 0.03 | **1.81** | 0.18 |
| | ID | 0 | 0 | 0 | 54 | 27 | 15 | 4 | 0.05 | **2.02** | 0.01 |
| | TF | 5 | 0 | 0 | 0 | 0 | 0 | 95 | 0.13 | 7.75 | 5.95 |
| | CPOP | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.03 | **1.62** | 0.34 |
| | BUP | 85 | 15 | 0 | 0 | 0 | 0 | 0 | 0.85 | 12.55 | 0.30 |

Table C.8: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods with the $\varepsilon_t$ being $AR(1)$ process of $\phi = 0.3$ with noise term following $t_5$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using 10 cores of Apple M1 Pro with 16 GB of RAM on mac OS, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M5) | TS | 0 | 0 | 0 | **98** | 0 | 0 | 2 | 0.01 | **0.66** | 0.34 |
| | NOT | 0 | 12 | 7 | 66 | 6 | 4 | 5 | 0.03 | **0.74** | 0.17 |
| | ID | 0 | 0 | 0 | 0 | 0 | 2 | 98 | 0.20 | 5.02 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.27 | 5.92 | 18.55 |
| | CPOP | 0 | 0 | 0 | 50 | 38 | 9 | 3 | 0.02 | 1.43 | 3.27 |
| | BUP | 16 | 84 | 0 | 0 | 0 | 0 | 0 | 0.16 | **0.89** | 1.12 |
| (M6) | TS | 0 | 0 | 1 | **97** | 0 | 0 | 2 | 0.01 | **0.20** | 0.37 |
| | NOT | 0 | 3 | 44 | 48 | 2 | 0 | 3 | 0.05 | **0.80** | 0.14 |
| | ID | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.09 | **0.39** | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.07 | 9.90 | 21.12 |
| | CPOP | 0 | 0 | 0 | 72 | 24 | 4 | 0 | 0.01 | **0.09** | 2.01 |
| | BUP | 23 | 36 | 30 | 11 | 0 | 0 | 0 | 0.18 | **0.27** | 1.42 |
| (M7) | TS | 32 | 35 | 26 | 6 | 0 | 0 | 1 | 0.13 | 11.69 | 0.13 |
| | NOT | 0 | 0 | 0 | **74** | 20 | 4 | 2 | 0.01 | **0.81** | 0.11 |
| | ID | 2 | 2 | 3 | 10 | 16 | 33 | 34 | 0.35 | 10.75 | 0.01 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.14 | 6.24 | 11.24 |
| | CPOP | 0 | 0 | 0 | 10 | 30 | 28 | 32 | 0.02 | **0.65** | 0.76 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 12.50 | 0.53 |
| (M8) | TS | 0 | 0 | 0 | **97** | 0 | 0 | 3 | 0.00 | **1.50** | 0.21 |
| | NOT | 0 | 0 | 0 | 88 | 4 | 7 | 1 | 0.00 | **3.46** | 0.09 |
| | ID | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.00 | **0.00** | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.11 | 49.93 | 15.78 |
| | CPOP | 0 | 0 | 0 | **99** | 0 | 1 | 0 | 0.00 | **0.05** | 5.50 |
| | BUP | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0.00 | 39.45 | 0.84 |

Table C.9: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the $\varepsilon_t$ being $AR(1)$ process of $\phi = 0.6$ with noise term following $t_5$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using 10 cores of Apple M1 Pro with 16 GB of RAM on mac OS, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | $\hat{N} - N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M1) | TS | 0 | 0 | 1 | **99** | 0 | 0 | 0 | 0.15 | **2.18** | 0.10 |
| | NOT | 0 | 0 | 0 | 55 | 12 | 8 | 25 | 0.17 | **2.90** | 0.10 |
| | ID | 0 | 0 | 0 | 7 | 9 | 11 | 73 | 0.15 | 3.98 | 0.01 |
| | TF | 43 | 0 | 1 | 0 | 0 | 0 | 56 | 0.29 | 18.34 | 1.74 |
| | CPOP | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0.10 | **1.21** | 0.06 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 3.03 | 11.77 | 0.15 |
| (M2) | TS | 1 | 7 | 19 | 72 | 0 | 0 | 1 | 0.11 | 2.11 | 0.24 |
| | NOT | 0 | 0 | 0 | 59 | 8 | 7 | 26 | 0.09 | 1.68 | 0.17 |
| | ID | 0 | 0 | 0 | 15 | 11 | 14 | 60 | 0.11 | 1.80 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.25 | 2.38 | 12.96 |
| | CPOP | 0 | 0 | 0 | **83** | 15 | 1 | 1 | 0.07 | **1.13** | 1.34 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 1.15 | 5.42 | 0.71 |
| (M3) | TS | 0 | 5 | 17 | 77 | 0 | 0 | 1 | 0.04 | 4.38 | 0.31 |
| | NOT | 0 | 0 | 0 | 16 | 7 | 8 | 69 | 0.05 | 5.16 | 0.22 |
| | ID | 0 | 0 | 0 | 24 | 18 | 14 | 44 | 0.03 | 4.20 | 0.03 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.10 | 6.23 | 19.46 |
| | CPOP | 0 | 0 | 0 | **96** | 2 | 2 | 0 | 0.01 | **1.39** | 19.42 |
| | BUP | 0 | 43 | 49 | 8 | 0 | 0 | 0 | 0.10 | 9.61 | 1.10 |
| (M4) | TS | 0 | 0 | 3 | **97** | 0 | 0 | 0 | 0.08 | **2.82** | 0.09 |
| | NOT | 0 | 0 | 0 | 22 | 14 | 10 | 54 | 0.12 | 4.68 | 0.18 |
| | ID | 0 | 0 | 0 | 27 | 19 | 21 | 33 | 0.09 | 3.71 | 0.01 |
| | TF | 38 | 0 | 0 | 1 | 1 | 0 | 60 | 0.29 | 17.07 | 6.15 |
| | CPOP | 0 | 0 | 0 | **95** | 3 | 2 | 0 | 0.05 | **2.03** | 0.35 |
| | BUP | 72 | 28 | 0 | 0 | 0 | 0 | 0 | 0.79 | 12.41 | 0.31 |

Table C.10: Distribution of $\hat{N}-N$ for models (M5)-(M8) and all methods with the $\varepsilon_t$ being $AR(1)$ process of $\phi = 0.6$ with noise term following $t_5$ over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance $d_H$ and the average computational time in seconds using 10 cores of Apple M1 Pro with 16 GB of RAM on mac OS, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N}-N=0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$. Note that TrendSegment is shortened to TS.

| | | | | | $\hat{N}-N$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | ≤-3 | -2 | -1 | 0 | 1 | 2 | ≥3 | MSE | $d_H(\times 10^2)$ | time |
| (M5) | TS | 0 | 2 | 15 | **81** | 1 | 0 | 1 | 0.03 | **1.11** | 0.33 |
| | NOT | 0 | 5 | 10 | 37 | 7 | 4 | 37 | 0.05 | 2.80 | 0.18 |
| | ID | 0 | 0 | 0 | 0 | 0 | 1 | 99 | 0.19 | 4.57 | 0.04 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.26 | 6.02 | 18.55 |
| | CPOP | 0 | 0 | 0 | 29 | 38 | 28 | 5 | 0.03 | **1.55** | 3.44 |
| | BUP | 13 | 86 | 1 | 0 | 0 | 0 | 0 | 0.16 | **1.32** | 1.17 |
| (M6) | TS | 0 | 0 | 0 | **99** | 0 | 0 | 1 | 0.01 | **0.10** | 0.36 |
| | NOT | 1 | 2 | 36 | 46 | 4 | 3 | 8 | 0.06 | 1.98 | 0.14 |
| | ID | 0 | 2 | 0 | 0 | 8 | 1 | 89 | 0.12 | 3.02 | 0.04 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.15 | 9.90 | 20.25 |
| | CPOP | 0 | 0 | 0 | 81 | 17 | 1 | 1 | 0.01 | **0.15** | 2.92 |
| | BUP | 0 | 2 | 24 | 74 | 0 | 0 | 0 | 0.08 | **0.24** | 1.22 |
| (M7) | TS | 73 | 25 | 1 | 0 | 0 | 0 | 1 | 0.20 | 12.44 | 0.12 |
| | NOT | 0 | 0 | 0 | **7** | 5 | 16 | 72 | 0.08 | **4.50** | 0.11 |
| | ID | 0 | 0 | 2 | 3 | 7 | 8 | 80 | 0.24 | 10.13 | 0.01 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.29 | 6.26 | 11.19 |
| | CPOP | 0 | 0 | 0 | 2 | 17 | 24 | 57 | 0.07 | **4.48** | 0.84 |
| | BUP | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0.26 | 12.51 | 0.52 |
| (M8) | TS | 0 | 0 | 0 | **99** | 0 | 0 | 1 | 0.00 | **0.50** | 0.20 |
| | NOT | 0 | 0 | 0 | 3 | 2 | 13 | 82 | 0.04 | 40.33 | 0.10 |
| | ID | 0 | 0 | 0 | 86 | 4 | 1 | 9 | 0.00 | **2.89** | 0.02 |
| | TF | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0.22 | 49.93 | 15.67 |
| | CPOP | 0 | 0 | 0 | **97** | 0 | 3 | 0 | 0.00 | **0.72** | 8.40 |
| | BUP | 0 | 0 | 0 | 0 | 80 | 20 | 0 | 0.00 | 40.07 | 0.83 |

## C.2 Choice of the threshold

In this section, we describe the details of how the thresholds, $\lambda^{\text{Naïve}}$ and $\lambda^{\text{Robust}}$, are built under different scenarios of noise introduced in Section C.1.

**The naive threshold selection** We first explain how the best performing threshold constant $C$ in the naive threshold. To cover all the noise scenario settings including dependent and/or non-Gaussian noise, here we use a simpler version of the follwoing naïve threshold:

$$\lambda^{\text{Naïve}} = C\sqrt{2\log T}, \tag{16}$$

by considering that the $\sigma$ in $\lambda^{\text{Naïve}} = C\sigma\sqrt{2\log T}$ can be absorbed into the constant $C$. To find the best performing constant $C$ over different noise scenarios introduced in Section C.1, we repeat the simulations with a range of $C$, $[0.5, 3.5]$. The performance can be evaluated by the accuracy of detecting number and location of change-points. For the number of change-point, we define

$$C^{\eta}_{\text{min}} = \text{Median}(\tilde{C}^{\eta}_{\text{min},1}, \ldots, \tilde{C}^{\eta}_{\text{min},8}), \tag{17}$$

where $\tilde{C}^{\eta}_{\text{min},j}$ is the minimum of those constants $C$ that give the maximum number of the case $\{\hat{N} - N = 0\}$ for the $j^{\text{th}}$ model from 100 simulation runs. The minimum condition is actually used when there is more than one constant giving the same maximum number of $\{\hat{N} - N = 0\}$. Similarly, we can define $C^{\eta}_{\text{med}}$ and $C^{\eta}_{\text{max}}$ by replacing the minimum condition with median and maximum respectively. For evaluating the performance of change-point location, we define

$$C^{d_H}_{\text{max}} = \text{Median}(\tilde{C}^{d_H}_{\text{max},1}, \ldots, \tilde{C}^{d_H}_{\text{max},8}), \tag{18}$$

where $\tilde{C}^{d_H}_{\text{max},j}$ is the maximum of those constants $C$ that give the minimum value of the average Hausdorff distance for the $j^{\text{th}}$ model computed from 100 simulation runs. Note that in contrast to that the maximum number of case $\{\hat{N} - N = 0\}$ is considered in (17), the minimum value of Hausdorff distance is used in (18) as the smaller the Hausdorff distance, the better the estimation of the change-point locations. Similar to (17), the maximum condition actually works when there is more than one constant giving the same minimum average Hausdorff distance, and $C^{d_H}_{\text{med}}$ and $C^{d_H}_{\text{min}}$ can be defined by replacing the maximum condition with median and minimum respectively.

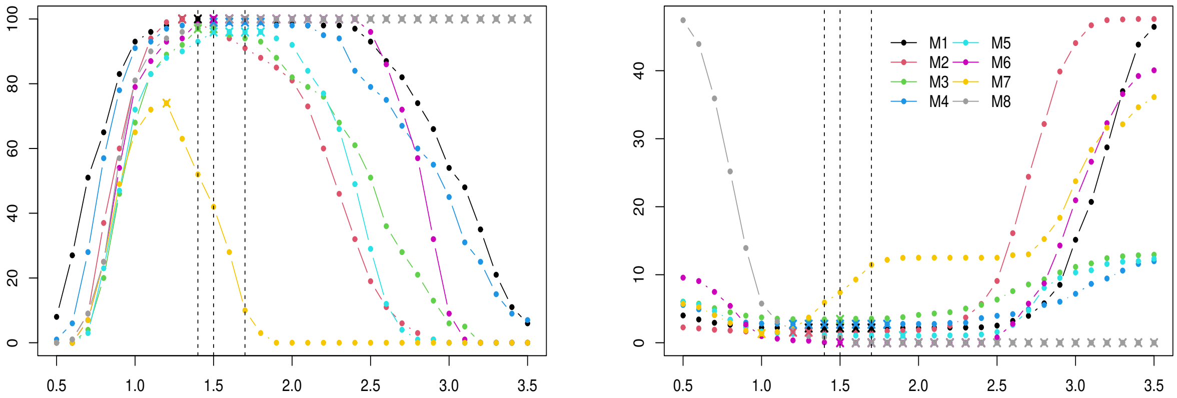
The best performing constants over all noise scenarios are reported in Table C.11. Compared to the iid Gaussian noise, it seems that a larger threshold constant tends to chosen when the noise is heavy-tailed and/or dependent. Also, compared to the other noise scenarios, when the noise is dependent but generated with Gaussian innovation ((iii) and (iv)), the best performing constant

Table C.11: The default thresholding constant $C$ with its range examined for five scenarios.

| $\varepsilon_t$ | $C^{\eta}_{\text{min}}$ | $C^{\eta}_{\text{med}}$ | $C^{\eta}_{\text{max}}$ | $C^{d_H}_{\text{min}}$ | $C^{d_H}_{\text{med}}$ | $C^{d_H}_{\text{max}}$ |
|---|---|---|---|---|---|---|
| (i) | 1.2 | 1.4 | 1.5 | 1.4 | 1.5 | 1.7 |
| (ii) | 1.4 | 1.5 | 1.6 | 1.3 | 1.5 | 1.5 |
| (iii) | 1.5 | 1.5 | 1.5 | 1.4 | 1.4 | 1.4 |
| (iv) | 1.8 | 1.8 | 1.8 | 1.3 | 1.3 | 1.3 |
| (v) | 1.0 | 1.6 | 1.9 | 1.0 | 1.4 | 1.9 |
| (vi) | 1.4 | 1.5 | 1.7 | 1.4 | 1.5 | 1.7 |

has a narrower range of $C^{\cdot}_{\text{max}}$ - $C^{\cdot}_{\text{min}}$. Similar interpretations and behaviours can be found from Figures C.1-C.6.

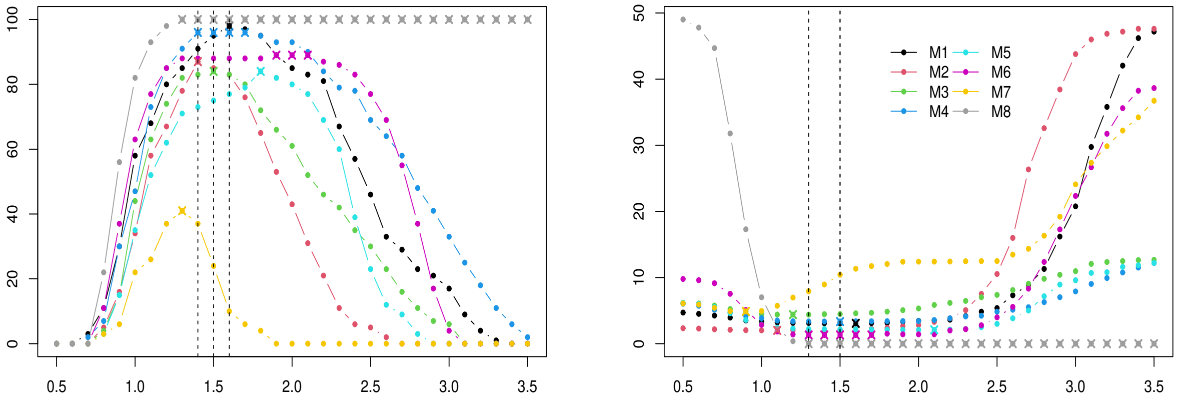
The naïve threshold, $\lambda^{\text{Naïve}}$, is an essential element in building the robust threshold, $\lambda^{\text{Robust}}$, as shown in Step 5 of Algorithm 1 in the main paper. For this, we use the default constants $C^{\eta}_{\text{med}}$ in Table C.11, where there is not much difference in simulation performance presented in Tables C.1-C.10 when $C^{\eta}_{\text{max}}$ is used instead.



(a) Number of the case, $\hat{N} - N = 0$.

(b) Mean of $d_H(\times 10^2)$.

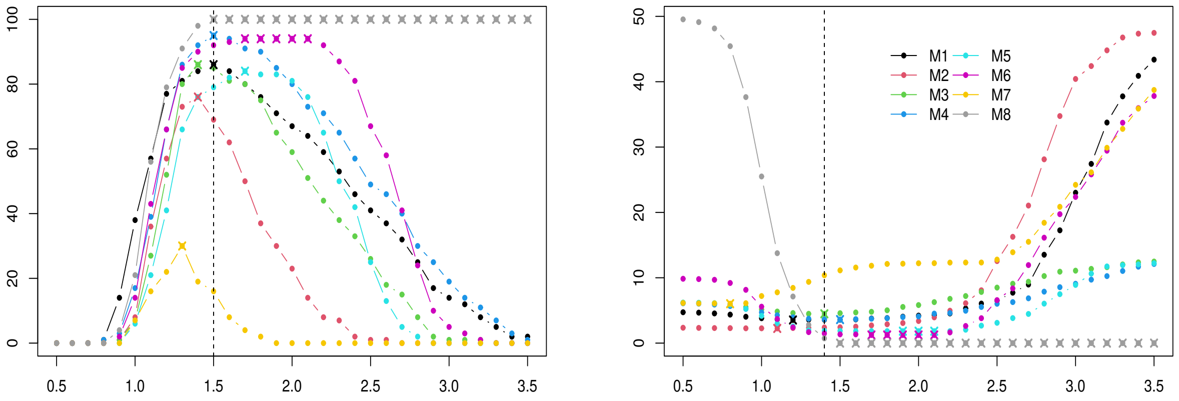
Figure C.1: Summary of the results over a sequence of the threshold constant $C$ from 0.8 to 1.8 for all models (M1)-(M8), where $\varepsilon_t \sim N(0, 1)$. (a) The number of the case $\hat{N} - N = 0$ from 100 simulation runs and (b) Mean of $d_H(\times 10^2)$ from 100 simulation runs. **X** symbols show where the maximum and minimum is obtained in (a) and (b), respectively, over a sequence of $C$s. The black vertical dashed lines present $C^{\eta}_{\text{min}} \leq C^{\eta}_{\text{med}} \leq C^{\eta}_{\text{max}}$ in (a) and $C^{d_H}_{\text{min}} \leq C^{d_H}_{\text{med}} \leq C^{d_H}_{\text{max}}$ in (b) defined in (17) and (18), respectively.



(a) Number of the case, $\hat{N} - N = 0$.

(b) Mean of $d_H(\times 10^2)$.

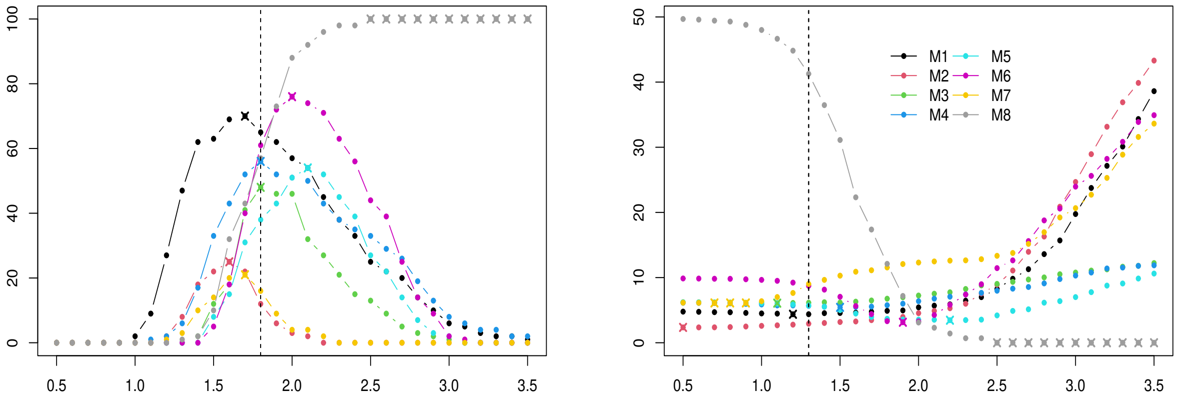
Figure C.2: Summary of the simulation results obtained over a sequence of the threshold constant $C$ from 1.2 to 2.2 for all models (M1)-(M8), where $\varepsilon_t \overset{\text{iid}}{\sim} t_5$. (a) The number of the case $\hat{N} - N = 0$ and (b) Mean of $d_H(\times 10^2)$. **X** symbols show where the maximum and minimum is obtained in (a) and (b), respectively, over a sequence of $C$s. The black vertical dashed lines present $C^{\eta}_{\text{min}} \leq C^{\eta}_{\text{med}} \leq C^{\eta}_{\text{max}}$ in (a) and $C^{d_H}_{\text{min}} \leq C^{d_H}_{\text{med}} \leq C^{d_H}_{\text{max}}$ in (b) defined in (17) and (18), respectively.



(a) Number of the case, $\hat{N} - N = 0$.

(b) Mean of $d_H(\times 10^2)$.

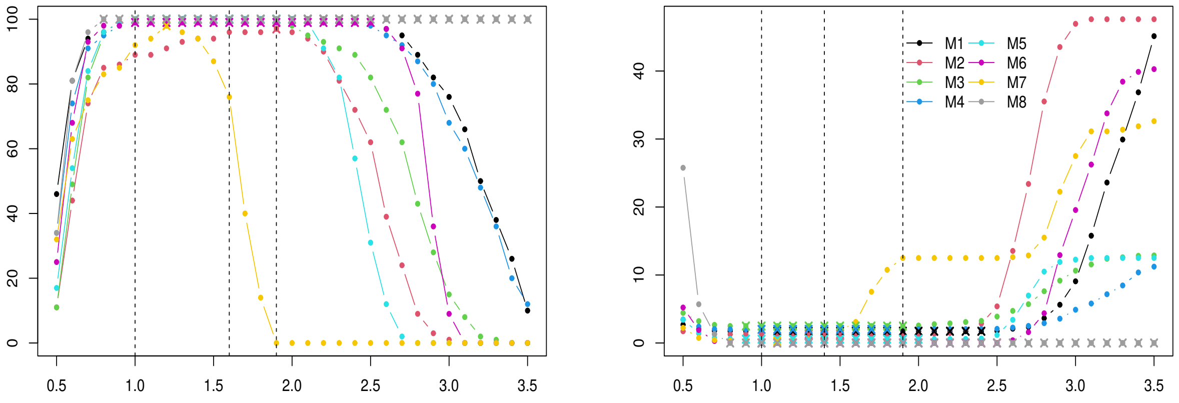
Figure C.3: Summary of the simulation results obtained over a sequence of the threshold constant $C$ from 1.2 to 2.2 for all models (M1)-(M8), where $\varepsilon_t \sim AR(1)$ with $\phi = 0.3$. (a) The number of the case $\hat{N} - N = 0$ and (b) Mean of $d_H(\times 10^2)$. **X** symbols show where the maximum and minimum is obtained in (a) and (b), respectively, over a sequence of $C$s. The black vertical dashed lines present $C^{\eta}_{\text{min}} \leq C^{\eta}_{\text{med}} \leq C^{\eta}_{\text{max}}$ in (a) and $C^{d_H}_{\text{min}} \leq C^{d_H}_{\text{med}} \leq C^{d_H}_{\text{max}}$ in (b) defined in (17) and (18), respectively.



(a) Number of the case, $\hat{N} - N = 0$.

(b) Mean of $d_H(\times 10^2)$.

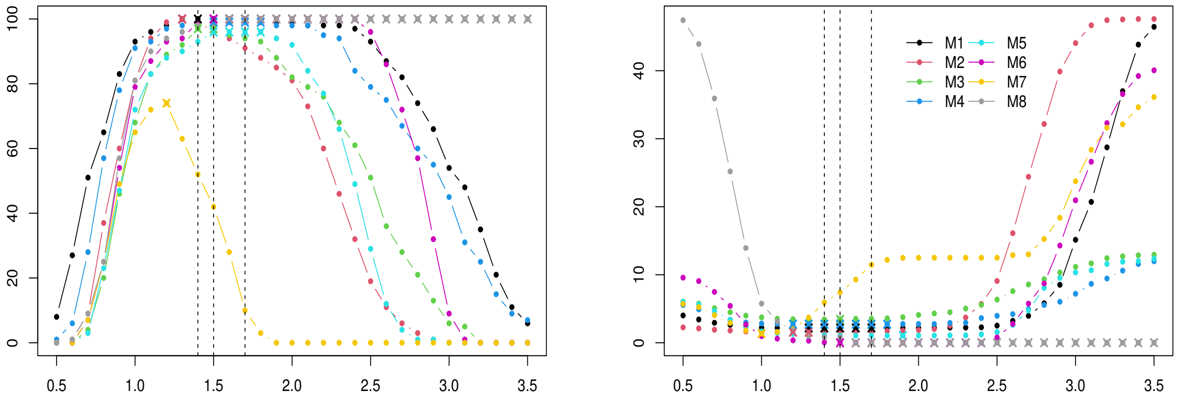
Figure C.4: Summary of the simulation results obtained over a sequence of the threshold constant $C$ from 1.5 to 4.5 for all models (M1)-(M8), where $\varepsilon_t \sim AR(1)$ with $\phi = 0.6$. (a) The number of the case $\hat{N} - N = 0$ and (b) Mean of $d_H(\times 10^2)$. **X** symbols show where the maximum and minimum is obtained in (a) and (b), respectively, over a sequence of $C$s. The black vertical dashed lines present $C^{\eta}_{\text{min}} \leq C^{\eta}_{\text{med}} \leq C^{\eta}_{\text{max}}$ in (a) and $C^{d_H}_{\text{min}} \leq C^{d_H}_{\text{med}} \leq C^{d_H}_{\text{max}}$ in (b) defined in (17) and (18), respectively.



(a) Number of the case, $\hat{N} - N = 0$.

(b) Mean of $d_H(\times 10^2)$.

Figure C.5: Summary of the simulation results obtained over a sequence of the threshold constant $C$ from 1.5 to 5 for all models (M1)-(M8), where $\varepsilon_t$ being $AR(1)$ process of $\phi = 0.3$ with noise term following $t_5$. (a) The number of the case $\hat{N} - N = 0$ and (b) Mean of $d_H(\times 10^2)$. **X** symbols show where the maximum and minimum is obtained in (a) and (b), respectively, over a sequence of $C$s. The black vertical dashed lines present $C^{\eta}_{\text{min}} \leq C^{\eta}_{\text{med}} \leq C^{\eta}_{\text{max}}$ in (a) and $C^{d_H}_{\text{min}} \leq C^{d_H}_{\text{med}} \leq C^{d_H}_{\text{max}}$ in (b) defined in (17) and (18), respectively.



(a) Number of the case, $\hat{N} - N = 0$.

(b) Mean of $d_H(\times 10^2)$.

Figure C.6: Summary of the simulation results obtained over a sequence of the threshold constant $C$ from 3 to 6 for all models (M1)-(M8), where $\varepsilon_t$ being $AR(1)$ process of $\phi = 0.6$ with noise term following $t_5$. (a) The number of the case $\hat{N} - N = 0$ and (b) Mean of $d_H(\times 10^2)$. **X** symbols show where the maximum and minimum is obtained in (a) and (b), respectively, over a sequence of $C$s. The black vertical dashed lines present $C^{\eta}_{\min} \leq C^{\eta}_{\text{med}} \leq C^{\eta}_{\max}$ in (a) and $C^{d_H}_{\min} \leq C^{d_H}_{\text{med}} \leq C^{d_H}_{\max}$ in (b) defined in (17) and (18), respectively.

**The robust threshold selection** We now describe the details of how $\lambda^{\text{Robust}}$ is built. We first justify using the ratio,

$$\frac{\lambda^{\text{Naïve}}}{1.3\hat{\mathcal{I}}\sqrt{2\log T}} \tag{19}$$

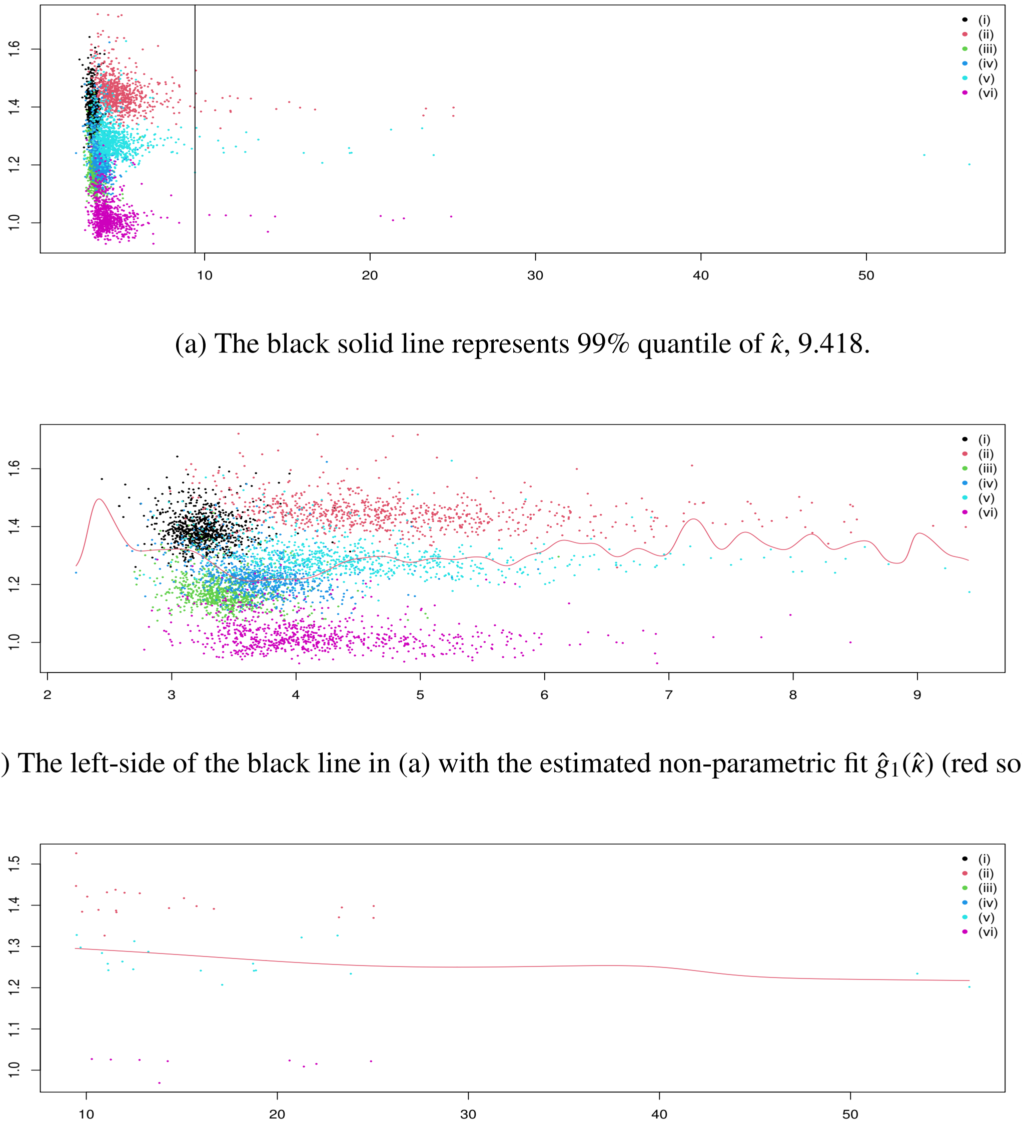
in the process of building the kurtosis function $g$ in Step 5 of Algorithm 1 in the main paper.

We first recall that the ratio in (19) corresponds to the kurtosis function $g(\mathcal{K})$ in the following robust threshold:

$$\lambda^{\text{Robust}} = C\mathcal{I}g(\mathcal{K})\sqrt{2\log T}. \tag{20}$$

Figure C.7 shows that $g(\hat{\mathcal{K}})$ behaves like constant over a range of the $\hat{\mathcal{K}}$ under all models and noise scenarios we considered. This is due to the condition on the minimum segment length imposed for stable and good performance. With this condition, we found that the constant-like behaviour is also observed in case noise has relatively extreme heavy-tail e.g. $t_{2.1}$, however we do not include such an extreme case in estimating the non-parametric function $g$.

We are now ready to estimate the function $g(\cdot)$. To avoid the situation that the estimation of non-parametric fit is affected a lot by extremely large size of $\hat{\mathcal{K}}$, we split $\hat{\mathcal{K}}$ into two with the 99% quantile of $\hat{\kappa}$ as shown in Figure C.7. Then we estimate the non-parametric regression fit for each interval, $g_1(\mathcal{K})$ and $g_2(\mathcal{K})$ respectively, and use these functions for computing the robust threshold.



(a) The black solid line represents 99% quantile of $\hat{\kappa}$, 9.418.

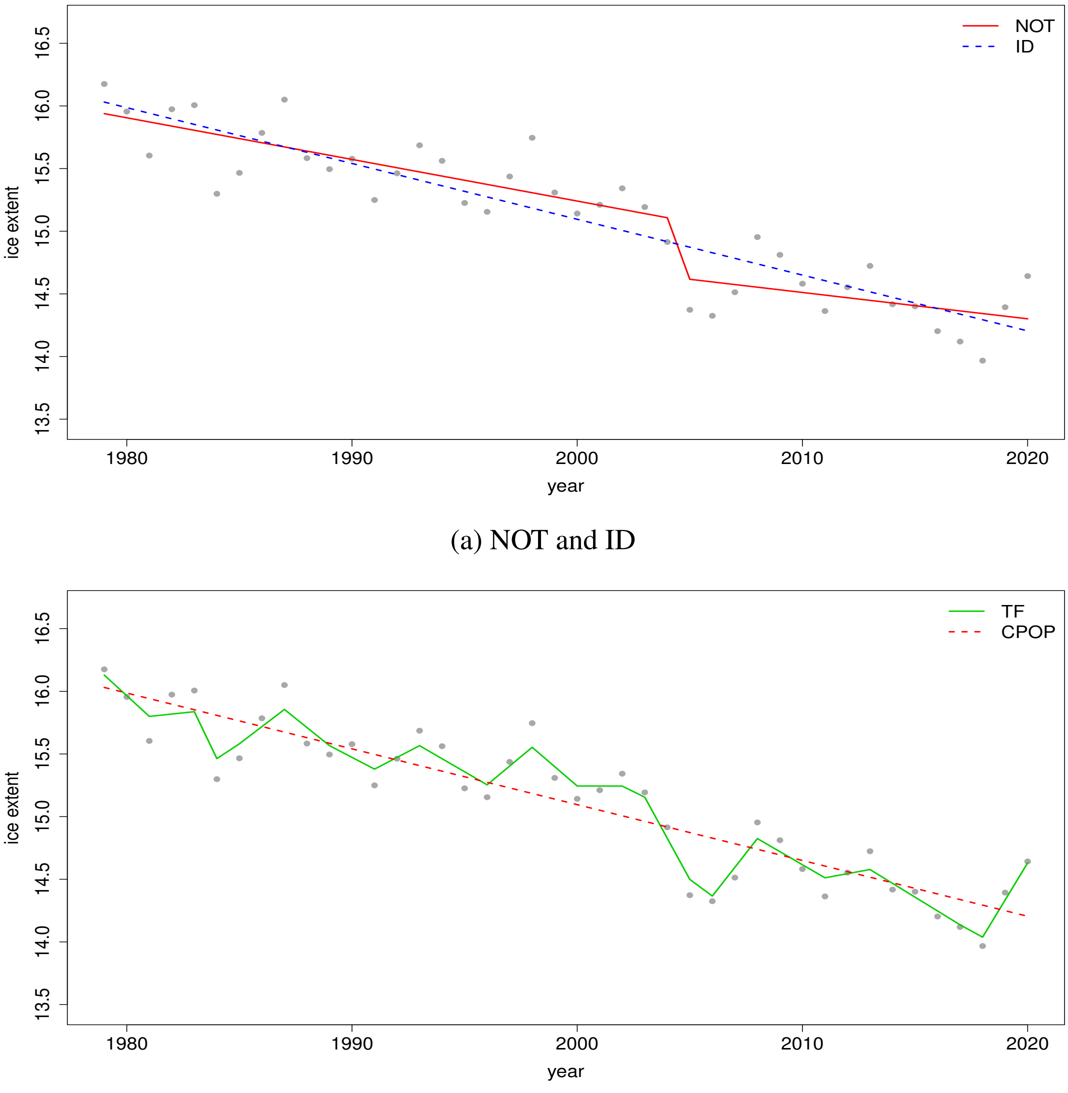
(b) The left-side of the black line in (a) with the estimated non-parametric fit $\hat{g}_1(\hat{\kappa})$ (red solid).

(c) The right-side of the black line in (a) with the estimated non-parametric fit $\hat{g}_2(\hat{\kappa})$ (red solid).

Figure C.7: The estimated kurtosis, $\hat{\kappa}$, of $\hat{\varepsilon}_t$ (x-axis) and the ratio in (19) (y-axis) over all models considered and over 6 different noise scenarios, (i)-(vi), presented in Section C.1, where each combination of model and noise scenario has $N = 100$ repetitions (dots) in (a). $\hat{\varepsilon}_t$ is obtained from the pre-fit in Algorithm 1 of the main paper.

# D Additional data application results

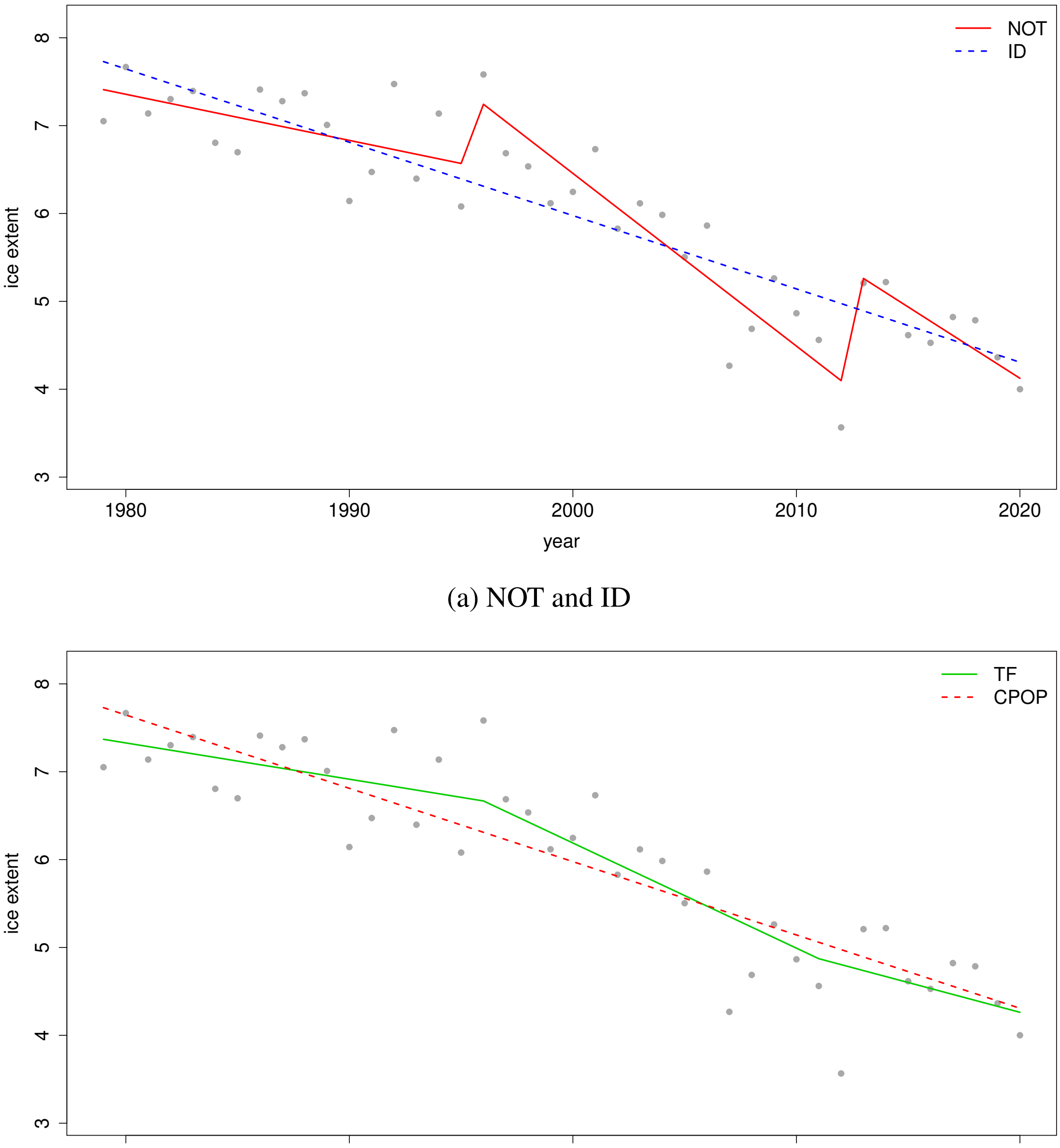
## D.1 Monthly average sea ice extent of Arctic and Antarctic data



(a) NOT and ID

(b) TF and CPOP

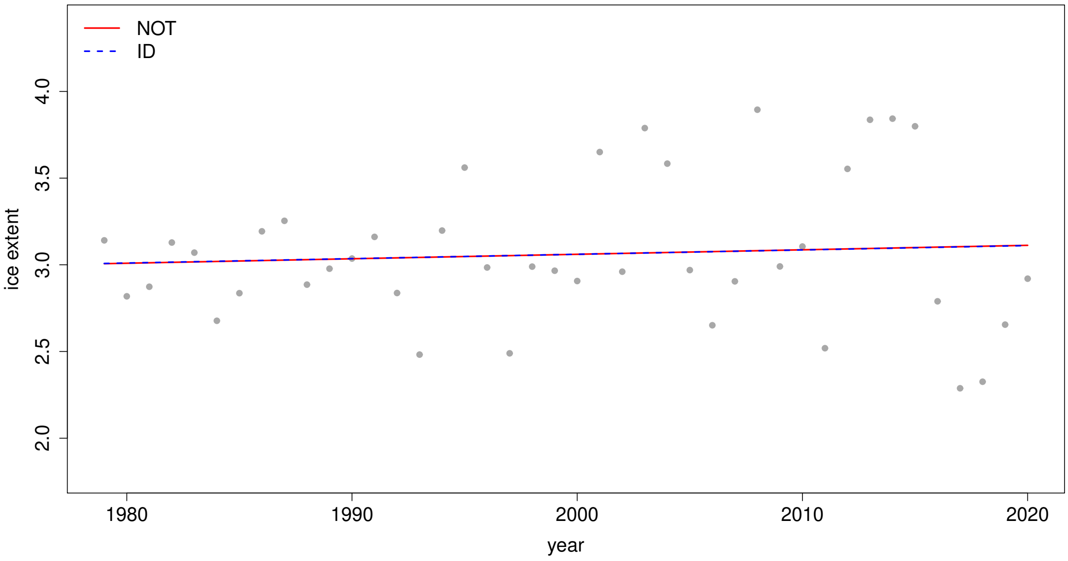
Figure D.1: Change-point analysis for the monthly average sea ice extent of the Arctic in February from 1979 to 2020 in Section 5.2. (a) the data series (grey dots) and estimated signal with change-points returned by NOT (—) and ID (- - -), (b) estimated signal with change-points returned by TF (—) and CPOP (-·-·).



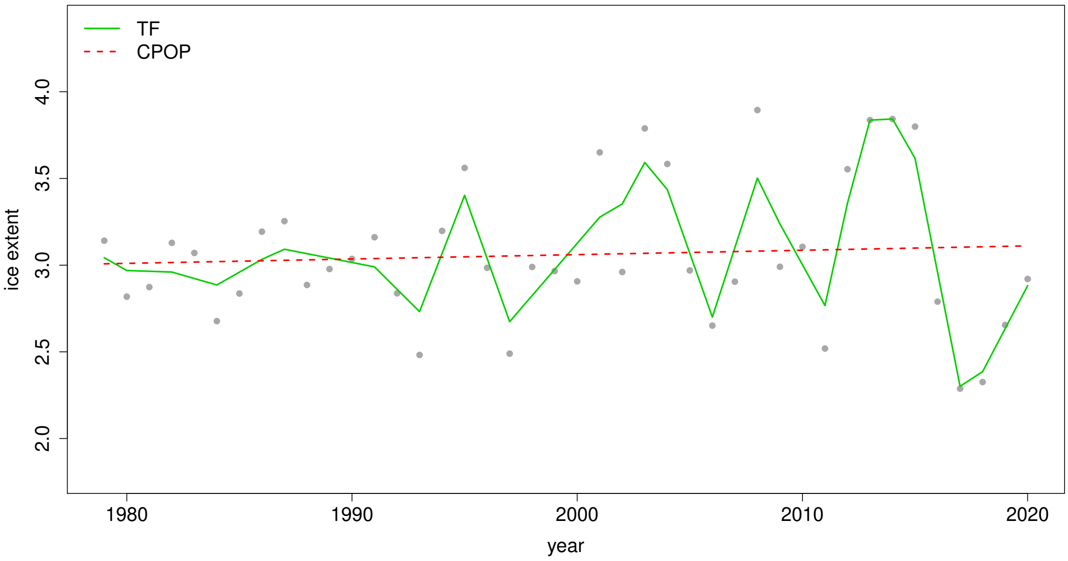
(a) NOT and ID

(b) TF and CPOP

Figure D.2: Change-point analysis for the monthly average sea ice extent of the Arctic in September from 1979 to 2020 in Section 5.2. (a) the data series (grey dots) and estimated signal with change-points returned by NOT (——) and ID (- - -), (b) estimated signal with change-points returned by TF (——) and CPOP (-·-·).

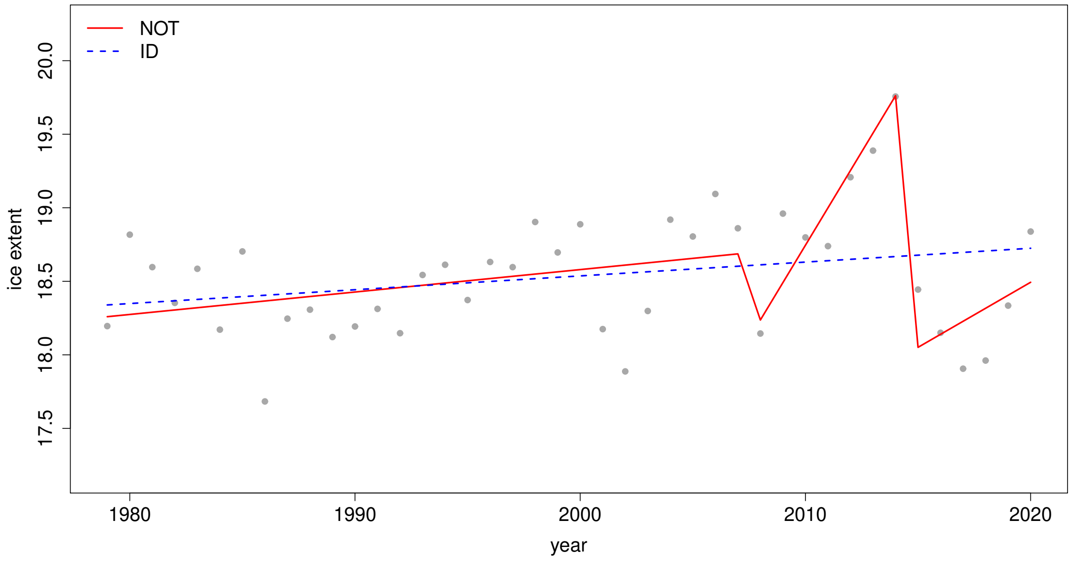


(a) NOT and ID

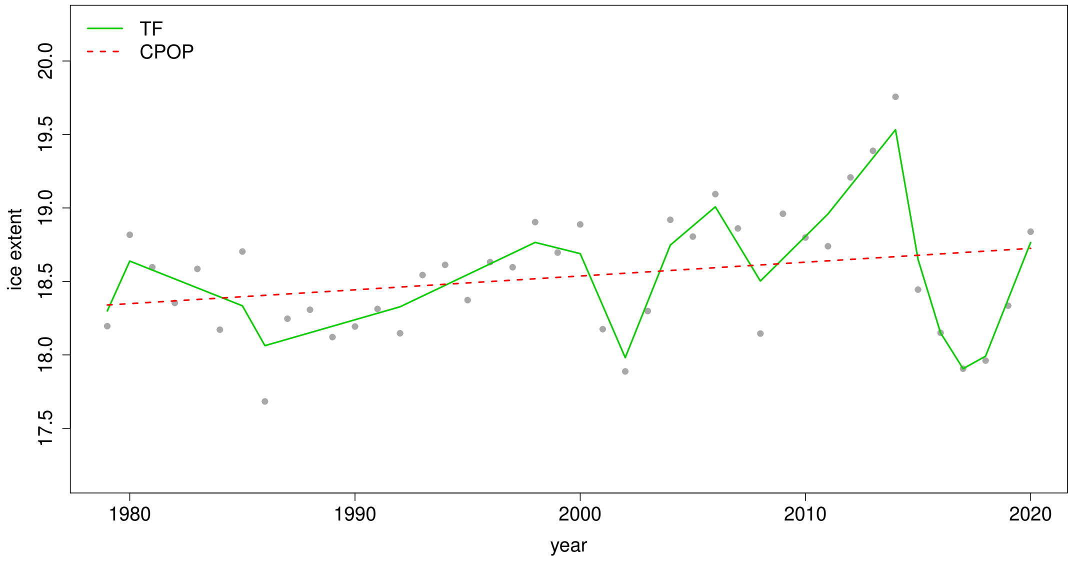


(b) TF and CPOP

Figure D.3: Change-point analysis for the monthly average sea ice extent of the Antarctic in February from 1979 to 2020 in Section 5.2. (a) the data series (grey dots) and estimated signal with change-points returned by NOT (—–) and ID (- - -), (b) estimated signal with change-points returned by TF (—–) and CPOP (-·-·).



(a) NOT and ID



(b) TF and CPOP

Figure D.4: Change-point analysis for the monthly average sea ice extent of the Antarctic in September from 1979 to 2020 in Section 5.2. (a) the data series (grey dots) and estimated signal with change-points returned by NOT (—) and ID (- - -), (b) estimated signal with change-points returned by TF (—) and CPOP (-·-·).

## D.2 Nitrogen oxides concentrations

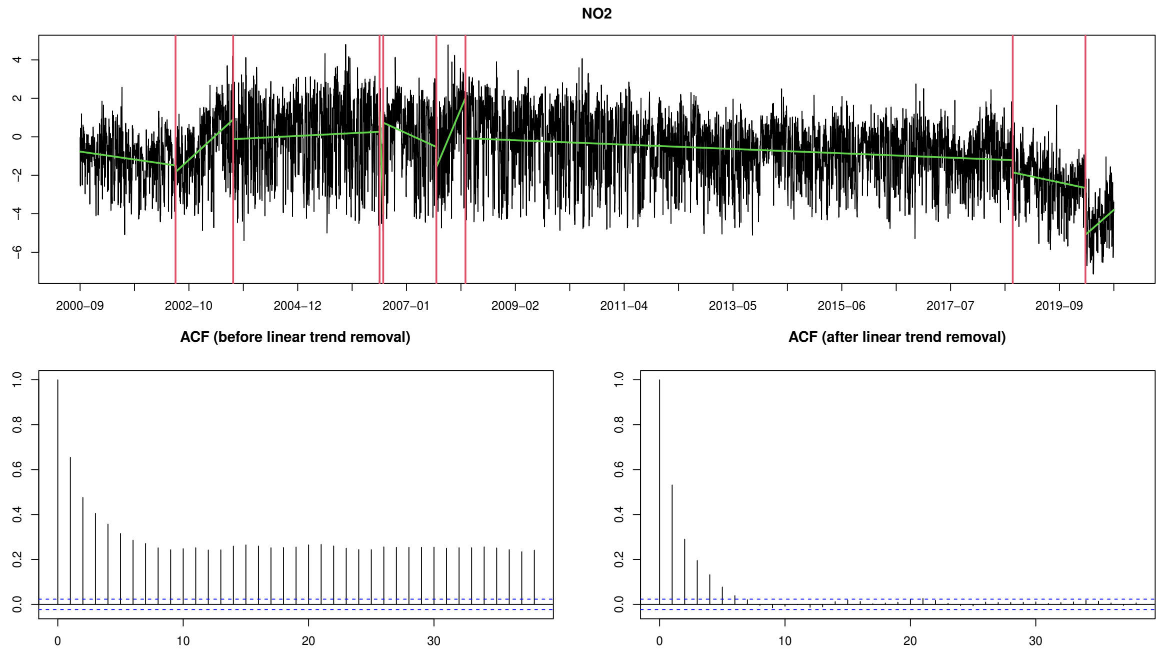


Figure D.5: (Top) Daily average concentrations of $NO_2$ (black) with the detected change-points by TrendSegment (red) and the estimated piecewise-linear trend (green). (Bottom) Autocorrelation function of $NO_2$ before (left) and after (right) linear trend removal.

In this section, we demonstrate that our TrendSegment algorithm shows a good performance on a real-world dataset that possibly has some nonnegligible autocorrelation. London air quality data is recently studied by Cho and Fryzlewicz (2020) in the context of proposing a methodology for detecting multiple changes in mean of a possibly autocorrelated time series. Using the same data but in a different context, we now detect changes in linear trend. We use the daily average concentrations of nitrogen dioxides ($NO_2$) measured from September 1, 2000 to September 30, 2020 at Marylebone Road in London, United Kingdom, which results in $T = 7139$ time points. The data is downloaded from `https://github.com/haeran-cho/wem.gsc`, where the original data can be obtained from Defra (`https://uk-air.defra.gov.uk/`). We follow the pre-processing steps used in Cho and Fryzlewicz (2020) by taking the square root transform and by removing weekly, seasonal and bank holiday effects.

Considering that the data possibly has serial dependent and/or heavy-tailedness, we use the robust threshold ($\lambda^{\text{Robust}}$) introduced in Section 4.1.5 of the main paper. The top plot in Figure D.5 shows the detected change-points using the robust threshold selection. From the two bottom plots, we see that the persistent autocorrelations are not observable anymore after removing the linear trends, although a certain amount of autocorrelations still exists.

# E Shape of the unbalanced wavelet basis

We now explore the shape of the adaptively constructed unbalanced wavelet basis. First, we denote that $\psi^{(j,k)}$ is sometimes referred to as $\psi^{(j,k)}_{p,q,r}$. One of the important properties of the unbalanced wavelet basis is that $\psi^{(j,k)}_{p,q,r}$ always has a shape of linear trend in regions that are previously merged and this linearity will also be preserved in future merges, as long as later transforms are performed under the "two together" rule. For example, two vectors $(\psi^{(0,1)}, \psi^{(0,2)})$ corresponding to the two smooth coefficients $s^1_{1,T}$ and $s^{[2]}_{1,T}$, have linear trends in the region $[1, T]$ as they form an orthonormal basis of the subspace $\{(x_1, x_2, \ldots, x_T) \mid x_1 - x_2 = x_2 - x_3 = \cdots = x_{T-1} - x_T\}$ of $\mathbb{R}^T$. This is due to the fact that the local orthonormal transforms continue in a way of extending the geometric dimension of subspace in which an orthonormal basis lives.

Through an illustrative example, we now show how a basis vector $\psi^{(j,k)}_{p,q,r}$ keeps its linearity in subregions that are already merged in previous scales, which includes a geometric interpretation of the TGUW transformation. Suppose that the initial data sequence is $\boldsymbol{s}^0 = (X_1, \ldots, X_5)$ and the initial weight vectors of constancy and linearity are $\boldsymbol{w}^c_0 = (1, 1, 1, 1, 1)^\top$ and $\boldsymbol{w}^l_0 = (1, 2, 3, 4, 5)^\top$, respectively. As we have the data sequence of length 5, the complete TGUW transform consists of 3 orthonormal transformations and the most important task for each transform is finding an appropriate orthonormal matrix.

**First merge**. Assume that $(X_3, X_4, X_5)$ is chosen as the first triplet to be merged. To find the values of the transform matrix $\Lambda$,

$$\Lambda = \begin{pmatrix} \ell_{1,1} & \ell_{1,2} & \ell_{1,3} \\ \ell_{2,1} & \ell_{2,2} & \ell_{2,3} \\ a & b & c \end{pmatrix} = \begin{pmatrix} \boldsymbol{\ell}_1^\top \\ \boldsymbol{\ell}_2^\top \\ \boldsymbol{h}^\top \end{pmatrix}, \tag{21}$$

we first seek the detail filter, $\boldsymbol{h}$, which satisfies the conditions (1) $\boldsymbol{h}^\top \boldsymbol{w}^c_{0,3:5} = 0$, (2) $\boldsymbol{h}^\top \boldsymbol{w}^l_{0,3:5} = 0$ and (3) $\boldsymbol{h}^\top \boldsymbol{h} = 1$, where $\boldsymbol{w}^\cdot_{0,p:r}$ is the subvector of length $r - p + 1$. Thus, $\boldsymbol{h}$ is obtained as a normal vector to the plane $\{(x, y, z) \mid x - 2y + z = 0\}$. Then, two low filter vectors ($\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$) are obtained under the conditions, (1) $\boldsymbol{\ell}_1^\top \boldsymbol{h} = 0$, (2) $\boldsymbol{\ell}_2^\top \boldsymbol{h} = 0$, (3) $\boldsymbol{\ell}_1^\top \boldsymbol{\ell}_2 = 0$ and (4) $\boldsymbol{\ell}_1^\top \boldsymbol{\ell}_1 = \boldsymbol{\ell}_2^\top \boldsymbol{\ell}_2 = 1$ which implies that $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ form an arbitrary orthonormal basis of the plane $\{(x, y, z) \mid x - 2y + z = 0\}$ and this guarantees the linear trend of $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$. Now, the orthonormal transform updates the data sequence and weight vectors as follows,

$$\begin{aligned}
\boldsymbol{s}^0 = (X_1, \ldots, X_5) &\quad \rightarrow \quad \boldsymbol{s} = (X_1, X_2, s^{[1]}_{3,5}, s^{[2]}_{3,5}, d_{3,4,5}), \\
\boldsymbol{w}^c_0 = (1,1,1,1,1)^\top &\quad \rightarrow \quad \boldsymbol{w}^c = (1, 1, e_{c_1}, e_{c_2}, 0)^\top, \\
\boldsymbol{w}^l_0 = (1,2,3,4,5)^\top &\quad \rightarrow \quad \boldsymbol{w}^l = (1, 2, e_{l_1}, e_{l_2}, 0)^\top,
\end{aligned} \tag{22}$$

where the constants $(e_{c_1}, e_{c_2})$ and $(e_{l_1}, e_{l_2})$ are obtained by $\Lambda \boldsymbol{w}^c_{0,3:5} = (e_{c_1}, e_{c_2}, 0)^\top$ and $\Lambda \boldsymbol{w}^l_{0,3:5} = (e_{l_1}, e_{l_2}, 0)^\top$, respectively. As $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ form an orthonormal basis of the plane $\{(x, y, z) \mid x-2y+z = 0\}$, $e_{c_1}, e_{c_2}$ and $e_{l_1}, e_{l_2}$ are unique constants which represent $\boldsymbol{w}^c_{0,3:5}$ and $\boldsymbol{w}^l_{0,3:5}$ as a linear span of basis vectors $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ as follows:

$$\boldsymbol{w}^c_{0,3:5} = e_{c_1}\boldsymbol{\ell}_1 + e_{c_2}\boldsymbol{\ell}_2, \quad \boldsymbol{w}^l_{0,3:5} = e_{l_1}\boldsymbol{\ell}_1 + e_{l_2}\boldsymbol{\ell}_2. \tag{23}$$

Importantly, the orthonormal transform matrix $\Psi_{T\times T}$ introduced in (5) (i.e. an orthonormal basis in $\mathbb{R}^5$ in this example) is constructed by recursively updating its initial input $\Psi_0 = \mathbf{I}_{5\times 5}$ through local orthonormal transforms. For example, if $(p, q, r)^{\text{th}}$ elements in $\boldsymbol{s}$ are selected to be merged, then we extract the corresponding $(p, q, r)^{\text{th}}$ columns of $\Psi^\top$ and update them through the matrix multiplication with $\Lambda$ used in that merge. Therefore, the first orthonormal transform performed in (22) updates the initial matrix $\Psi^\top_0$ by multiplying $\Lambda$ to the corresponding $(3, 4, 5)^{th}$ columns of $\Psi^\top_0$ which returns the following,

$$\Psi^\top = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ell_{1,1} & \ell_{2,1} & a \\ 0 & 0 & \ell_{1,2} & \ell_{2,2} & b \\ 0 & 0 & \ell_{1,3} & \ell_{2,3} & c \end{pmatrix}. \tag{24}$$

The 5th column of $\Psi^\top$ is now fixed (not going to be updated again) as it corresponds to the detail coefficient but other four columns corresponding to the smooth coefficients in $\boldsymbol{s}$ would be updated as the merging continues.

**Second merge.** Suppose that $(X_2, s^{[1]}_{3,4,5}, s^{[2]}_{3,4,5})$ are selected to be merged next under the "two together" rule. Then we need to find the following orthonormal transform matrix,

$$\Lambda^* = \begin{pmatrix} \ell^*_{1,1} & \ell^*_{1,2} & \ell^*_{1,3} \\ \ell^*_{2,1} & \ell^*_{2,2} & \ell^*_{2,3} \\ a^* & b^* & c^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\ell}_1^{*\top} \\ \boldsymbol{\ell}_2^{*\top} \\ \boldsymbol{h}^{*\top} \end{pmatrix}, \tag{25}$$

where its elements would be different from those in (21). The detail filter $\boldsymbol{h}^{*\top} = (a^*, b^*, c^*)$ is constructed from the corresponding weight vectors, $\boldsymbol{w}^c_{2:4} = (1, e_{c_1}, e_{c_2})^\top$ and $\boldsymbol{w}^l_{2:4} = (2, e_{l_1}, e_{l_2})^\top$, by satisfying the conditions (1) $\boldsymbol{h}^{*\top}\boldsymbol{w}^c_{2:4} = 0$, (2) $\boldsymbol{h}^{*\top}\boldsymbol{w}^l_{2:4} = 0$ and (3) $\boldsymbol{h}^{*\top}\boldsymbol{h}^* = 1$. The detail filter is a weight vector designed for indicating the strength of linearity in $(X_2, X_3, X_4, X_5)$ as $(e_{c_1}, e_{c_2})$ and $(e_{l_1}, e_{l_2})$ already contain the information of three raw observations $(X_3, X_4, X_5)$. Then, two low filters, $\boldsymbol{\ell}^*_1$ and $\boldsymbol{\ell}^*_2$, are obtained by satisfying the conditions, $\boldsymbol{\ell}_1^{*\top}\boldsymbol{h}^* = 0$, $\boldsymbol{\ell}_2^{*\top}\boldsymbol{h}^* = 0$, $\boldsymbol{\ell}_1^{*\top}\boldsymbol{\ell}^*_2 = 0$ and $\Lambda^{*\top}\Lambda^* = \mathbf{I}$. Now the data sequence and the weight vectors are updated as follows,

$$\begin{aligned} \boldsymbol{s} = (X_1, X_2, s^{[1]}_{3,5}, s^{[2]}_{3,5}, d_{3,4,5}) \quad &\rightarrow \quad \boldsymbol{s} = (X_1, s^{[1]}_{2,5}, s^{[2]}_{2,5}, d_{2,2,5}, d_{3,4,5}), \\ \boldsymbol{w}_c = (1, 1, e_{c_1}, e_{c_2}, 0)^\top \quad &\rightarrow \quad \boldsymbol{w}_c = (1, e^*_{c_1}, e^*_{c_2}, 0, 0)^\top, \\ \boldsymbol{w}_l = (1, 2, e_{l_1}, e_{l_2}, 0)^\top \quad &\rightarrow \quad \boldsymbol{w}_l = (1, e^*_{l_1}, e^*_{l_2}, 0, 0)^\top, \end{aligned} \tag{26}$$

and $\Psi^\top$ is also updated into

$$\Psi^\top = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \ell^*_{1,1} & \ell^*_{2,1} & a^* & 0 \\ 0 & & & & a \\ 0 & \left(\ell^*_{1,2}\boldsymbol{\ell}_1 + \ell^*_{1,3}\boldsymbol{\ell}_2\right) & \left(\ell^*_{2,2}\boldsymbol{\ell}_1 + \ell^*_{2,3}\boldsymbol{\ell}_2\right) & \left(b^*\boldsymbol{\ell}_1 + c^*\boldsymbol{\ell}_2\right) & b \\ 0 & & & & c \end{pmatrix}. \tag{27}$$

At this scale, the 4th column of $\Psi^\top$ is fixed. This corresponds to the Type 2 basis vector in (2) whose non-zero subregion is composed of a single point $(a^*)$ and a linear trend $(b^*\boldsymbol{\ell}_1 + c^*\boldsymbol{\ell}_2)$.

Importantly, the orthonormal transform at this scale is performed in a way of returning an orthonormal basis of the expanded subspace e.g. 2nd and 3rd columns of (27) (which are referred to as $\boldsymbol{\ell}^{**}_1$ and $\boldsymbol{\ell}^{**}_2$ in (28)) are obtained as an arbitrary orthonormal basis of the subspace $\{(w, x, y, z) \mid w - x = x - y = y - z\}$ of $\mathbb{R}^4$. This is due to the semi-orthogonality of the transformation matrix $\mathbf{\Pi}$ in (28) which extends the dimension from $\mathbb{R}^3$ to $\mathbb{R}^4$ but preserves the fact that $(\boldsymbol{\ell}^*_1, \boldsymbol{\ell}^*_2)$ and $(\boldsymbol{\ell}^{**}_1, \boldsymbol{\ell}^{**}_2)$ form an arbitrary orthonormal basis of the corresponding subspaces. This guarantees the properties, $\boldsymbol{\ell}_1^{**\top}\boldsymbol{\ell}_2^{**} = 0$ and $\boldsymbol{\ell}_1^{**\top}\boldsymbol{\ell}_1^{**} = \boldsymbol{\ell}_2^{**\top}\boldsymbol{\ell}_2^{**} = 1$, where

$$\begin{aligned}
\boldsymbol{\ell}_1^{**} &= \begin{pmatrix} \ell^*_{1,1} \\ \begin{pmatrix} \\ \ell^*_{1,2}\boldsymbol{\ell}_1 + \ell^*_{1,3}\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \begin{matrix} 0 \\ 0 \\ 0 \end{matrix} & \begin{pmatrix} \\ \boldsymbol{\ell}_1 \\ \\ \end{pmatrix} & \begin{pmatrix} \\ \boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \end{pmatrix} \begin{pmatrix} \ell^*_{1,1} \\ \ell^*_{1,2} \\ \ell^*_{1,3} \end{pmatrix} = \boldsymbol{\Pi} \begin{pmatrix} \ell^*_{1,1} \\ \ell^*_{1,2} \\ \ell^*_{1,3} \end{pmatrix}, \\
\boldsymbol{\ell}_2^{**} &= \begin{pmatrix} \ell^*_{2,1} \\ \begin{pmatrix} \\ \ell^*_{2,2}\boldsymbol{\ell}_1 + \ell^*_{2,3}\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \begin{matrix} 0 \\ 0 \\ 0 \end{matrix} & \begin{pmatrix} \\ \boldsymbol{\ell}_1 \\ \\ \end{pmatrix} & \begin{pmatrix} \\ \boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \end{pmatrix} \begin{pmatrix} \ell^*_{2,1} \\ \ell^*_{2,2} \\ \ell^*_{2,3} \end{pmatrix} = \boldsymbol{\Pi} \begin{pmatrix} \ell^*_{2,1} \\ \ell^*_{2,2} \\ \ell^*_{2,3} \end{pmatrix},
\end{aligned} \tag{28}$$

and $\boldsymbol{\Pi}$ is obtained from the 2nd to 4th columns of (24) and the selected rows correspond to the indices of smooth coefficients associated in the orthonormal transformation in (25).

As is in (23), now the extended subregions of the original weight vectors, $\boldsymbol{w}^c_{0,2:5}$ and $\boldsymbol{w}^l_{0,2:5}$, can also be presented as a linear combination of $\boldsymbol{\ell}_1^{**}$ and $\boldsymbol{\ell}_2^{**}$ as follows:

$$\boldsymbol{w}^c_{0,2:5} = e^*_{c_1}\boldsymbol{\ell}_1^{**} + e^*_{c_2}\boldsymbol{\ell}_2^{**}, \quad \boldsymbol{w}^l_{0,2:5} = e^*_{l_1}\boldsymbol{\ell}_1^{**} + e^*_{l_2}\boldsymbol{\ell}_2^{**}, \tag{29}$$

where $\boldsymbol{\ell}_1^{**}$ and $\boldsymbol{\ell}_2^{**}$ form an orthonormal basis of the subspace $\{(w, x, y, z) \mid w - x = x - y = y - z\}$ of $\mathbb{R}^4$. This can be simply shown by 1) expressing the weight vectors as a linear combination of two low filters,

$$\begin{aligned}
\boldsymbol{w}^c_{2:4} &= (1, e_{c_1}, e_{c_2})^\top = e^*_{c_1}\boldsymbol{\ell}_1^* + e^*_{c_2}\boldsymbol{\ell}_2^*, \\
\boldsymbol{w}^l_{2:4} &= (2, e_{l_1}, e_{l_2})^\top = e^*_{l_1}\boldsymbol{\ell}_1^* + e^*_{l_2}\boldsymbol{\ell}_2^*,
\end{aligned} \tag{30}$$

and 2) performing the matrix multiplication with $\boldsymbol{\Pi}$ in (28) to both sides of (30),

$$\begin{aligned}
&\text{LHS} : \boldsymbol{\Pi}\boldsymbol{w}^c_{2:4} = (1, e_{c_1}\boldsymbol{\ell}_1 + e_{c_2}\boldsymbol{\ell}_2)^\top = (1, 1, 1, 1)^\top = \boldsymbol{w}^c_{0,2:5}, \quad \text{RHS} : e^*_{c_1}\boldsymbol{\ell}_1^{**} + e^*_{c_2}\boldsymbol{\ell}_2^{**}, \\
&\text{LHS} : \boldsymbol{\Pi}\boldsymbol{w}^l_{2:4} = (2, e_{l_1}\boldsymbol{\ell}_1 + e_{l_2}\boldsymbol{\ell}_2)^\top = (2, 3, 4, 5)^\top = \boldsymbol{w}^l_{0,2:5}, \quad \text{RHS} : e^*_{l_1}\boldsymbol{\ell}_1^{**} + e^*_{l_2}\boldsymbol{\ell}_2^{**}.
\end{aligned} \tag{31}$$

**Last merge.** In the same manner, after the last orthonormal transform is applied to $(X_1, s^{[1]}_{2,5}, s^{[2]}_{2,5})$, we end up with the finalised $\Psi^\top$ in which an orthonormal basis of the subspace $\{(v, w, x, y, z) \mid v - w = w - x = x - y = y - z\}$ of $\mathbb{R}^5$ is shown in its first and second columns where these two columns correspond to two basis vectors, $\psi^{(0,1)}$ and $\psi^{(0,2)}$, in (5). Regardless of the length of data ($T$), the first two columns of the finalised $\Psi^\top$ build two smooth coefficients ($s^{[1]}_{1,T}, s^{[2]}_{1,T}$) and always keep a linear trend with length $T$, while the shape of other columns of $\Psi^\top$ corresponding to the detail coefficients depends on the type of merge and follows one of the forms in (2).

As shown above, the non-uniqueness of the low filters has no effect on preserving the linearity of the subregions that are already merged. In simulation studies, we empirically found that the choice of low filters has no qualitative effect on the results as long as they are chosen by satisfying the orthonormality condition of the transform, thus we used a fixed type of function for choosing a set of low filters rather than choosing an arbitrary set of low filters that satisfies the orthonormal condition every run which also saves the computational costs.

# F A practical way to implement the TGUW transformation

In this section, we explore a way of implementing the TGUW transform. As briefly mentioned in Section 2.2.3, it is implemented by consecutively updating so-called weight vectors of constancy and linearity. These two weight vectors are initially used in the first stage of the TGUW transform for obtaining the detail filter $\boldsymbol{h}$ and updated through the orthonormal transform. In detail, Steps 1 and 5 of the TGUW algorithm presented in Section 2.2.3 can be reformulated by weight vectors as follows.

**Step 1**. At each scale $j$, find the set of triplets that are candidates for merging under the "two together" rule and compute the corresponding detail coefficients. Regardless of the type of merge, a detail coefficient $d^{\cdot}_{p,q,r}$ is, in general, obtained as

$$d^{\cdot}_{p,q,r} = a\boldsymbol{s}^1_{p:r} + b\boldsymbol{s}^2_{p:r} + c\boldsymbol{s}^3_{p:r}, \tag{32}$$

where $p \leq q < r$, $\boldsymbol{s}^k_{p:r}$ is the $k^{\text{th}}$ smooth coefficient of the subvector $\boldsymbol{s}_{p:r}$ with a length of $r - p + 1$ and the constants $a, b, c$ are the elements of the detail filter $\boldsymbol{h} = (a, b, c)^\top$. Specifically, the detail filter $\boldsymbol{h}$ is established by solving the following equations,

$$\begin{aligned} &a\boldsymbol{w}^{c,1}_{p:r} + b\boldsymbol{w}^{c,2}_{p:r} + c\boldsymbol{w}^{c,3}_{p:r} = 0, \\ &a\boldsymbol{w}^{l,1}_{p:r} + b\boldsymbol{w}^{l,2}_{p:r} + c\boldsymbol{w}^{l,3}_{p:r} = 0, \\ &a^2 + b^2 + c^2 = 1, \end{aligned} \tag{33}$$

where $\boldsymbol{w}^{\cdot,k}_{p:r}$ is $k^{\text{th}}$ non-zero element of the subvector $\boldsymbol{w}^{\cdot}_{p:r}$ with a length of $r - p + 1$, and $\boldsymbol{w}^c$ and $\boldsymbol{w}^l$ are weight vectors of constancy and linearity, respectively, in which the initial inputs have a form of $\boldsymbol{w}_0^c = (1, 1, \ldots, 1)^\top, \boldsymbol{w}_0^l = (1, 2, \ldots, T)^\top$. The last condition in (33) is to preserve the orthonormality of the transform and the detail filter $\boldsymbol{h}$ becomes a unit normal vector of the plane $\{(x, y, z) \mid x - 2y + z = 0\}$. The solution to (33) is unique up to multiplication by $-1$ and this can be simply shown by solving the equations e.g. $a + b + c = 0$, $a + 2b + 3c = 0$ and $a^2 + b^2 + c^2 = 1$.

More specifically, the detail coefficient in (32) is formulated for each type of merging introduced in Section 2.2.1 as follows.
Type 1: merging three initial smooth coefficients $(s^0_{p,p}, s^0_{p+1,p+1}, s^0_{p+2,p+2})$,

$$d_{p,p+1,p+2} = a_{p,p+1,p+2} s^0_{p,p} + b_{p,p+1,p+2} s^0_{p+1,p+1} + c_{p,p+1,p+2} s^0_{p+2,p+2}. \tag{34}$$

Type 2: merging one initial and a paired smooth coefficient $(s^0_{p,p}, s^{[1]}_{p+1,r}, s^{[2]}_{p+1,r})$,

$$d_{p,p,r} = a_{p,p,r} s^0_{p,p} + b_{p,p,r} s^{[1]}_{p+1,r} + c_{p,p,r} s^{[2]}_{p+1,r}, \quad \text{where} \quad p + 2 < r, \tag{35}$$

similarly, when merging a paired smooth coefficient and one initial, $(s^{[1]}_{p,r-1}, s^{[2]}_{p,r-1}, s^0_{r,r})$,

$$d_{p,r-1,r} = a_{p,r-1,r} s^{[1]}_{p,r-1} + b_{p,r-1,r} s^{[2]}_{p,r-1} + c_{p,r-1,r} s^0_{r,r}, \quad \text{where} \quad p + 2 < r. \tag{36}$$

Type 3: merging two sets of (paired) smooth coefficients, $(s^{[1]}_{p,q}, s^{[2]}_{p,q})$ and $(s^{[1]}_{q+1,r}, s^{[2]}_{q+1,r})$,

$$\begin{aligned} d^{[1]}_{p,q,r} &= a^1_{p,q,r} s^{[1]}_{p,q} + b^1_{p,q,r} s^{[2]}_{p,q} + c^1_{p,q,r} s^{[1]}_{q+1,r} \\ d^{[2]}_{p,q,r} &= a^2_{p,q,r} s^{01}_{p,r} + b^2_{p,q,r} s^{02}_{p,r} + c^2_{p,q,r} s^{[2]}_{q+1,r} \end{aligned} \quad \Longrightarrow \quad d_{p,q,r} = \max(|d^{[1]}_{p,q,r}|, |d^{[2]}_{p,q,r}|), \tag{37}$$

where $q > p + 1$ and $r > q + 2$. Importantly, the two consecutive merges in (37) are achieved by visiting the same two adjacent data regions twice. In this case, after the first detail coefficient, $d^{[1]}_{p,q,r}$, has been obtained, we instantly update the corresponding triplets $\boldsymbol{s}$, $\boldsymbol{w}^c$ and $\boldsymbol{w}^l$ via an orthonormal transform as defined in (38) and (39). Therefore, the second detail filter, $(a^2_{p,q,r}, b^2_{p,q,r}, c^2_{p,q,r})$, is constructed with the updated $\boldsymbol{w}^c$ and $\boldsymbol{w}^l$ in a way that satisfies the conditions (33).

**Step 5**. For each $|d^{\cdot}_{p,q,r}|$ extracted in step 4, merge the corresponding smooth coefficients by updating the corresponding triplet in $\boldsymbol{s}$, $\boldsymbol{w}^c$ and $\boldsymbol{w}^l$ through the orthonormal transform as follows,

$$\begin{pmatrix} s_{p,r}^{[1]} \\ s_{p,r}^{[2]} \\ d_{p,q,r}^{\cdot} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\ell}_1^\top \\ \boldsymbol{\ell}_2^\top \\ \boldsymbol{h}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{s}_{p:r}^1 \\ \boldsymbol{s}_{p:r}^2 \\ \boldsymbol{s}_{p:r}^3 \end{pmatrix} = \Lambda \begin{pmatrix} \boldsymbol{s}_{p:r}^1 \\ \boldsymbol{s}_{p:r}^2 \\ \boldsymbol{s}_{p:r}^3 \end{pmatrix}, \tag{38}$$

$$\begin{pmatrix} w_{p,r}^{c,1} \\ w_{p,r}^{c,2} \\ 0 \end{pmatrix} = \Lambda \begin{pmatrix} \boldsymbol{w}_{p:r}^{c,1} \\ \boldsymbol{w}_{p:r}^{c,2} \\ \boldsymbol{w}_{p:r}^{c,3} \end{pmatrix}, \quad \begin{pmatrix} w_{p,r}^{l,1} \\ w_{p,r}^{l,2} \\ 0 \end{pmatrix} = \Lambda \begin{pmatrix} \boldsymbol{w}_{p:r}^{l,1} \\ \boldsymbol{w}_{p:r}^{l,2} \\ \boldsymbol{w}_{p:r}^{l,3} \end{pmatrix}. \tag{39}$$

The key step is finding the $3 \times 3$ orthonormal matrix, $\Lambda$, which is composed of one detail and two low-pass filter vectors in its rows. Firstly the detail filter $\boldsymbol{h}^\top$ is determined to satisfy the conditions in (33), and then the two low-pass filters $(\boldsymbol{\ell}_1^\top, \boldsymbol{\ell}_2^\top)$ are obtained by satisfying the orthonormality of $\Lambda$. There is no uniqueness in the choice of $(\boldsymbol{\ell}_1^\top, \boldsymbol{\ell}_2^\top)$, but as described in Section E, this has no effect on the orthonormal transformation itself.

# G Extension to piecewise-quadratic signal

In this section, we explore how the TGUW transform can be extended to handle piecewise-quadratic signals. Considering the fact that we perform an orthonormal transformation to the chosen pair (triplet) to deal with piecewise-constant (piecewise-linear) signals, it is natural to perform a transform to the chosen quadraplet of the smooth conefficients in the process of establishing a data-adaptive unbalanced wavelet basis. In each merge, four adjacent smooth coefficients are selected and the orthonormal transformation converts them into one detail and three (updated) smooth coefficients. Those three updated smooth coefficients are tripled in the sense that they contain information about one local quadratic regression fit. Therefore, any such triplet of smooth coefficients cannot be separated when choosing quadruplet in any subsequent merges which can be called as “three together” rule (instead of “two together” rule invented for piecewise-linear model). We now give a simple example to illustrate how the TGUW transform for piecewise-quadratic siganal works. Figure G.1 shows the merging history of the modified TGUW transform which follows the “three together” rule. Three different types of merges are similary defined as for piecewise-linear signal except the fact that the merges are performed on quadraplet instead of triplet. The tree structure show that the modified TGUW transform performs well in detecting a single change-point in piecewise-quadratic scenario as the last type 3 merge is corresponding to the true change-point.

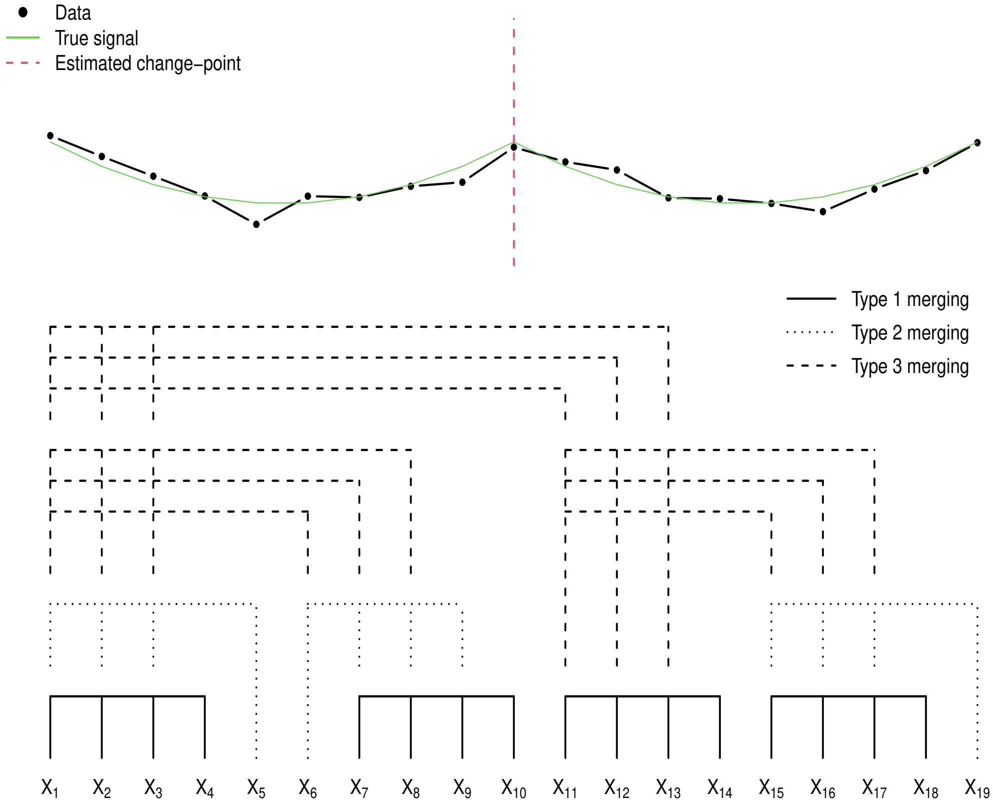


Figure G.1: Example of data with one true change-point at $t = 10$ in its underlying piecewise-quadratic signal (top) along with the tree structure constructed in TGUW transform by merging (bottom).

# References

Bosq, D. (1998). *Nonparametric statistics for stochastic processes.* Springer, New York.

Cho, H. and Fryzlewicz, P. (2020). Multiple change point detection under serial dependence: Wild energy maximisation and gappy schwarz criterion. *arXiv preprint arXiv:2011.13884*.

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast mulitple change-point detection. *The Annals of Statistics*, 46:3390–3421.