Supplementary materials for "Detecting linear trend changes and point anomalies in data sequences"

Hyeyoung Maeng and Piotr Fryzlewicz

London School of Economics and Political Science

This document includes the following sections:

A. Lemmas

B. Additional simulation results

C. Additional data application results

D. Geometric interpretation of the TGUW transformation

A. Lemmas

Lemma 1 *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* $\psi_i^{(j,k)}$ *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* $\left\{ g_i^{(j,k)}{}_{i=1,\dots,I^{(j,k)}, j=1,\dots,J, k=1,\dots,K(j)} \right\}$ *and* $\{g_l\}$, *we then have* $P(A_T) \geq 1 - C_2 T^{-1}$ *where*

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

$\lambda = C_1 \sqrt{2 \log T}$ *and* C_1 *is a positive constant large enough 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, $\left(g_i^{(j,k)}\right)^T g_i^{(j,k)} = 1$, $\left(g_i^{(j,k)}\right)^T g_{i'}^{(j,k)} = 0$ and $\sum_i (\phi_i^{(j,k)})^2 = 1$, where $\psi^{(j,k)} = \sum_{i=1}^{I^{(j,k)}} \phi_i^{(j,k)} g_i^{(j,k)}$. Depending on the type of merge, $\psi^{(j,k)}$ fall into one of the followings,

$$\begin{aligned}
&\text{Type 1: } \psi_{p,q,r}^{(j,k)} = \alpha_1 e_p + \alpha_2 e_{p+1} + \alpha_3 e_{p+2}, \\
&\text{Type 2: } \psi_{p,q,r}^{(j,k)} = \beta_1 e_p + \beta_2 (\underbrace{0,\dots,0}_{p\times 1}, \boldsymbol{\ell}_{1,p+1,r}^T, \underbrace{0,\dots,0}_{(T-r)\times 1}) + \beta_3 (\underbrace{0,\dots,0}_{p\times 1}, \boldsymbol{\ell}_{2,p+1,r}^T, \underbrace{0,\dots,0}_{(T-r)\times 1}), \\
&\qquad\qquad \psi_{p,q,r}^{(j,k)} = \beta_4 (\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}_{1,p,r-1}^T, \underbrace{0,\dots,0}_{(T-r+1)\times 1}) + \beta_5 (\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}_{2,p,r-1}^T, \underbrace{0,\dots,0}_{(T-r+1)\times 1}) + \beta_6 e_r, \\
&\text{Type 3: } \psi_{p,q,r}^{(j,k)} = \gamma_1 (\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}_{1,p,q}^T, \underbrace{0,\dots,0}_{(T-q)\times 1}) + \gamma_2 (\underbrace{0,\dots,0}_{(p-1)\times 1}, \boldsymbol{\ell}_{2,p,q}^T, \underbrace{0,\dots,0}_{(T-q)\times 1}) \\
&\qquad + \gamma_3 (\underbrace{0,\dots,0}_{q\times 1}, \boldsymbol{\ell}_{1,q+1,r}^T, \underbrace{0,\dots,0}_{(T-r)\times 1}) + \gamma_4 (\underbrace{0,\dots,0}_{q\times 1}, \boldsymbol{\ell}_{2,q+1,r}^T, \underbrace{0,\dots,0}_{(T-r)\times 1}),
\end{aligned} \tag{A.2}$$

where e_i is a vector of length T with one 1 at i^{th} position and zero for the others. As will be shown in Section D., $\boldsymbol{\ell}_{1,i,j}$ and $\boldsymbol{\ell}_{2,i,j}$ are an arbitrary orthonormal basis of the subspace $\{(x_1, x_2, \ldots, x_{j-i+1}) \mid x_1 - x_2 = x_2 - x_3 = \cdots = x_{j-i} - x_{j-i+1}\}$ 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 (A.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. From the orthonormality of the basis $(\boldsymbol{\ell}_{1,\cdot,\cdot}, \boldsymbol{\ell}_{2,\cdot,\cdot})$, we see that the conditions, $\left(g_i^{(j,k)}\right)^T g_i^{(j,k)} = 1$ and $\left(g_i^{(j,k)}\right)^T g_{i'}^{(j,k)} = 0$, are satisfied for any (i, i', j, k) where $i \neq i'$. In addition, as $\psi^{(j,k)}$ keeps 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 (A.2).

If we predefine the pairs $(\boldsymbol{\ell}_{1,i,j}, \boldsymbol{\ell}_{2,i,j})$ for all i and j by choosing an orthonormal basis of the subspace $\{(x_1, x_2, \ldots, x_{j-i+1}) \mid x_1 - x_2 = x_2 - x_3 = x_{j-i} - x_{j-i+1}\}$ of $\mathbb{R}^{j-i+1}$, then there exist at most $2T^2$ vectors g_l in the set G. Now we are in position to show that $P(A_T) \geq 1 - C_2 T^{-1}$. Using a simple Bonferroni inequality, we have

$$1 - P(A_T) \leq \sum_G P(|Z| > \lambda) \leq 2T^2 \frac{\phi_Z(\lambda)}{\lambda} \leq \frac{C_2}{T} \tag{A.3}$$

where ϕ_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 $p < \eta_i - 0.5 < r$ for some $i = 1, \ldots, N\}$, and $\mathcal{S}_j^0 = \{1, \ldots, K(j)\} \setminus \mathcal{S}_j^1$. On the set A_T which satisfies $P(A_T) \to 1$ as $T \to \infty$, we have*

$$\max_{\substack{j=1,\ldots,J, \\ k \in \mathcal{S}_j^0}} \left|d^{(j,k)}\right| \leq \lambda, \tag{A.4}$$

where λ is as in Theorem 1.

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

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

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)^T$. The condition, $\sum_i \left(\phi_i^{(j,k)}\right)^2 = 1$ for any fixed (i, j, k), given in the proof of Lemma 1 implies $\max_i \left|\phi_i^{(j,k)}\right| \leq 1$ for any (j, k), thus we have (A.4) when the constant C_1 for λ in (A.4) is larger than or equal to 4 times C_1 used in (A.1).

B. Additional simulation results

In addition to the simulations in Section 4, here we present the results for three different distributions of the noise ε_t, (a) ε_t follows a stationary Gaussian AR(1) process of $\phi = 0.3$, with

zero-mean and unit-variance, (b) $\varepsilon_t \sim$ i.i.d. scaled t_5 distribution with unit-variance and (c) $\varepsilon_t \sim$ i.i.d. Laplace(0, $1/\sqrt{2}$). We use $C = 1.8$ as a default thresholding constant for TS. 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 1: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the noise term ϵ_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 $d_H(\times 10^2)$.

		$\hat{N} - N$									
Model	Method	≤-3	-2	-1	0	1	2	≥3	MSE	$d_H(\times 10^2)$	time
(M1)	TS	0	0	4	**93**	3	0	0	0.081	3.58	1.27
	NOT	0	0	0	**91**	9	0	0	0.072	2.95	0.26
	ID	0	0	0	82	14	4	0	0.067	**2.65**	0.38
	TF	0	0	0	0	0	0	100	0.532	5.00	36.61
	CPOP	0	0	0	7	15	6	72	0.080	3.69	4.95
	BUP	0	0	8	86	6	0	0	0.077	3.66	2.75
(M2)	TS	1	6	23	69	1	0	0	0.195	2.44	1.12
	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	1	5	**88**	6	0	0	0.052	4.16	1.56
	NOT	0	0	0	**89**	7	4	0	0.042	**3.40**	0.31
	ID	0	0	3	77	16	3	1	0.064	**3.12**	0.50
	TF	0	0	0	0	0	0	100	0.259	6.24	44.94
	CPOP	0	0	0	1	4	10	85	0.068	4.67	9.57
	BUP	0	0	0	0	3	18	79	0.056	5.57	3.56
(M4)	TS	0	6	23	**53**	18	0	0	0.058	**2.41**	1.53
	NOT	0	93	6	1	0	0	0	0.086	2.91	0.31
	ID	2	6	30	49	10	2	1	0.165	2.99	0.48
	TF	0	0	0	0	0	0	100	0.218	6.22	45.99
	CPOP	0	0	0	1	3	9	87	0.066	4.02	5.40
	BUP	0	0	0	11	35	37	17	0.109	5.64	3.77

Table 2: Distribution of $\hat{N}-N$ for models (M5)-(M8) and all methods with the noise term ϵ_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 $d_H(\times 10^2)$.

		$\hat{N}-N$									
Model	Method	≤-3	-2	-1	0	1	2	≥3	MSE	$d_H(\times 10^2)$	time
(M5)	TS	0	0	19	**54**	21	6	0	0.055	**1.87**	1.54
	NOT	0	0	91	6	3	0	0	0.060	**1.94**	0.28
	ID	0	0	9	23	23	18	27	0.402	9.47	0.46
	TF	0	0	0	0	0	0	100	0.182	6.21	42.65
	CPOP	0	0	0	0	2	4	94	0.068	3.70	4.08
	BUP	0	0	0	15	37	32	16	0.112	5.25	3.53
(M6)	TS	0	0	0	**98**	2	0	0	0.018	**0.06**	1.70
	NOT	68	9	10	4	1	3	5	0.257	21.63	0.25
	ID	20	10	0	0	11	0	59	0.164	12.83	0.63
	TF	0	0	0	0	0	0	100	0.332	11.04	47.43
	CPOP	0	0	0	5	11	17	67	0.056	4.86	5.31
	BUP	0	0	0	0	0	0	100	0.170	10.18	3.95
(M7)	TS	11	38	31	15	3	2	0	0.217	11.52	0.68
	NOT	5	12	19	**24**	22	7	11	0.158	7.69	0.24
	ID	32	1	18	**26**	14	5	4	0.511	17.54	0.03
	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	0.003	**0.00**	1.09
	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 3: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the noise term ϵ_t being i.i.d. 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 $d_H(\times 10^2)$.

		$\hat{N} - N$									
Model	Method	≤-3	-2	-1	0	1	2	≥3	MSE	$d_H(\times 10^2)$	time
(M1)	TS	0	0	3	55	32	5	5	0.083	3.41	1.09
	NOT	0	0	0	**98**	2	0	0	0.037	2.21	0.26
	ID	0	0	0	85	10	4	1	0.036	**1.87**	0.29
	TF	0	0	0	0	0	0	100	0.017	4.36	36.49
	CPOP	0	0	0	21	20	20	39	0.064	2.28	5.69
	BUP	0	4	13	78	5	0	0	0.071	3.84	2.62
(M2)	TS	0	3	11	70	10	5	1	0.164	2.18	1.03
	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	1	11	41	25	10	12	0.073	4.90	1.44
	NOT	0	0	0	**96**	3	1	0	0.021	**2.54**	0.31
	ID	0	0	1	73	19	3	4	0.053	**2.72**	0.44
	TF	0	0	0	0	0	0	100	0.024	5.92	46.35
	CPOP	0	0	0	9	10	11	70	0.065	3.57	11.71
	BUP	0	0	0	1	21	40	38	0.043	5.44	3.52
(M4)	TS	0	3	14	34	23	16	10	0.075	3.10	1.46
	NOT	0	97	3	0	0	0	0	0.066	**2.45**	0.28
	ID	1	12	22	**48**	10	3	4	0.159	**2.42**	0.42
	TF	0	0	0	0	0	0	100	0.081	6.06	45.74
	CPOP	0	0	0	4	4	15	77	0.062	3.37	5.15
	BUP	0	2	7	28	47	12	4	0.095	5.30	3.56

Table 4: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods with the noise term ϵ_t being i.i.d. 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 $d_H(\times 10^2)$.

		$\hat{N} - N$									
Model	Method	≤-3	-2	-1	0	1	2	≥3	MSE	$d_H(\times 10^2)$	time
(M5)	TS	0	0	9	**40**	19	21	11	0.069	2.72	1.49
	NOT	0	0	95	4	1	0	0	0.042	**1.29**	0.26
	ID	0	0	1	16	24	23	36	0.372	9.86	0.43
	TF	0	0	0	0	0	0	100	0.118	6.15	43.23
	CPOP	0	0	0	3	8	12	77	0.060	2.97	3.55
	BUP	0	0	10	**40**	43	6	1	0.083	4.76	3.51
(M6)	TS	0	0	0	**46**	2	39	13	0.035	**3.16**	1.66
	NOT	54	21	4	8	5	1	7	0.244	17.30	0.23
	ID	8	8	0	0	6	0	78	0.125	6.99	0.62
	TF	0	0	0	0	0	0	100	0.138	10.99	48.53
	CPOP	0	0	0	9	11	17	63	0.059	4.68	3.48
	BUP	0	0	0	0	0	0	100	0.145	10.27	3.92
(M7)	TS	14	28	32	14	8	4	0	0.204	11.21	0.65
	NOT	0	6	16	**30**	36	11	1	0.079	5.12	0.22
	ID	14	8	12	17	24	13	12	0.421	16.22	0.04
	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	49	0	43	8	0.030	14.86	1.01
	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**	0.45
	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

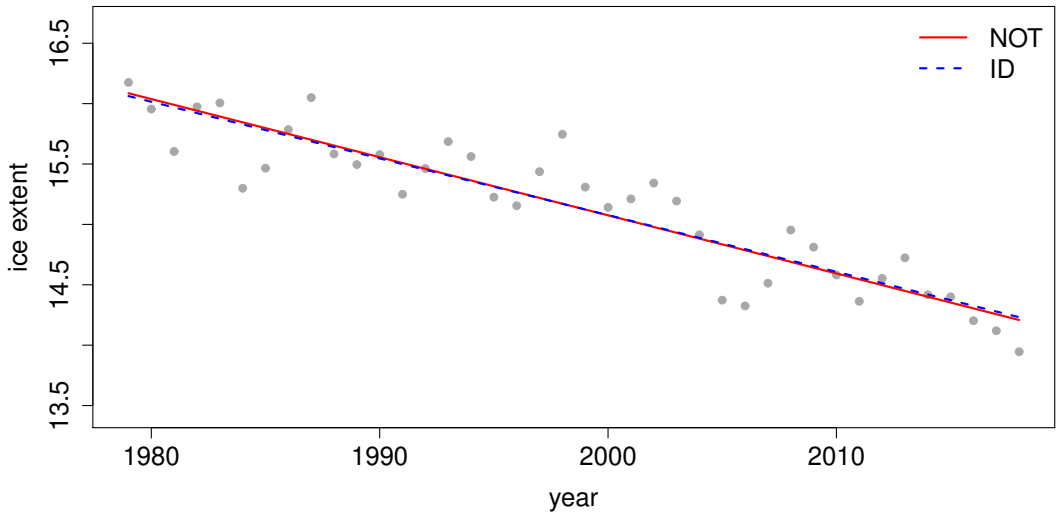
Table 5: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods with the noise term ϵ_t being i.i.d. Laplace over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance d_H and the average computational time in seconds 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 $d_H(\times 10^2)$.

		$\hat{N} - N$									
Model	Method	≤-3	-2	-1	0	1	2	≥3	MSE	$d_H(\times 10^2)$	time
(M1)	TS	5	17	36	33	8	1	0	0.167	5.68	1.10
	NOT	0	1	10	**86**	3	0	0	0.086	3.56	0.28
	ID	0	0	0	**89**	9	2	0	0.065	**2.33**	0.32
	TF	0	0	0	0	0	0	100	0.034	4.44	36.66
	CPOP	0	0	0	28	30	16	26	0.091	**2.21**	7.63
	BUP	0	0	0	0	0	4	96	0.080	4.64	2.64
(M2)	TS	69	16	9	4	2	0	0	0.466	5.81	1.03
	NOT	5	8	36	45	4	2	0	0.240	2.80	0.33
	ID	0	1	0	**84**	10	3	2	0.158	**1.87**	0.39
	TF	0	0	0	0	0	0	100	0.102	2.30	31.92
	CPOP	0	0	0	26	16	20	38	0.182	**1.71**	2.20
	BUP	64	21	14	1	0	0	0	0.375	3.71	2.23
(M3)	TS	2	16	30	33	17	2	0	0.127	7.75	1.45
	NOT	0	0	11	**83**	5	0	1	0.049	3.74	0.33
	ID	0	0	4	**83**	11	2	0	0.061	**3.29**	0.48
	TF	0	0	0	0	0	0	100	0.038	5.75	46.08
	CPOP	0	0	0	17	13	16	54	0.100	3.90	16.09
	BUP	0	0	0	0	0	0	100	0.072	5.98	3.48
(M4)	TS	11	42	32	9	5	1	0	0.154	4.94	1.46
	NOT	1	89	8	2	0	0	0	0.092	**3.12**	0.31
	ID	11	8	44	**27**	7	1	2	0.188	3.77	0.49
	TF	0	0	0	0	0	0	100	0.099	5.91	45.56
	CPOP	0	2	5	13	13	20	47	0.118	**3.31**	8.35
	BUP	0	0	0	0	0	0	100	0.132	6.00	3.51

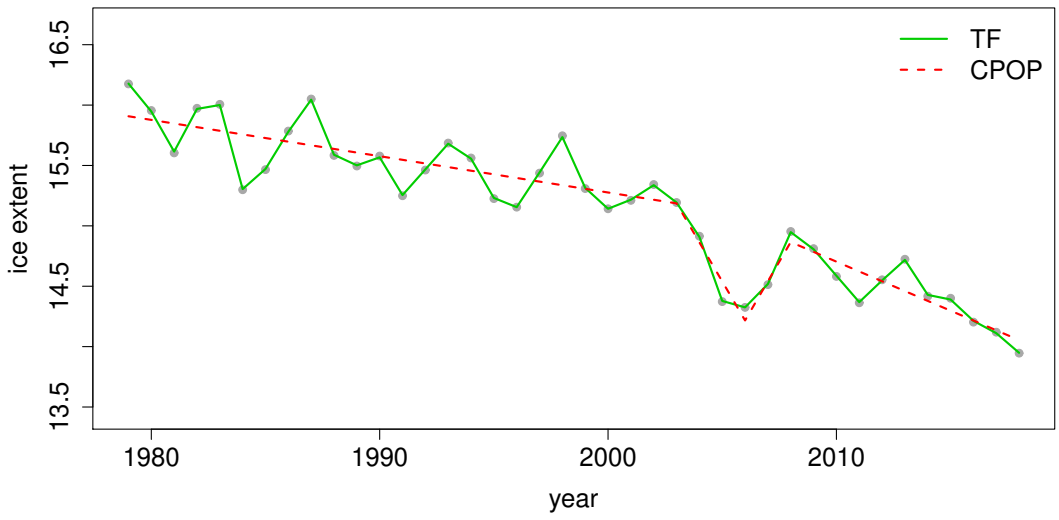
Table 6: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods with the noise term ϵ_t being i.i.d. Laplace over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal $\hat{f}_t$, the average Hausdorff distance d_H and the average computational time in seconds 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 $d_H(\times 10^2)$.

		$\hat{N} - N$									
Model	Method	≤-3	-2	-1	0	1	2	≥3	MSE	$d_H(\times 10^2)$	time
(M5)	TS	0	6	43	22	23	5	1	0.141	3.49	1.49
	NOT	0	0	86	13	1	0	0	0.074	**2.12**	0.30
	ID	0	4	12	**38**	20	17	9	0.448	9.50	0.46
	TF	0	0	0	0	0	0	100	0.168	6.14	44.16
	CPOP	0	0	1	4	12	16	67	0.113	3.10	5.21
	BUP	0	0	0	0	0	0	100	0.148	5.91	3.55
(M6)	TS	4	13	0	**57**	3	22	1	0.077	**6.14**	1.64
	NOT	76	11	11	1	1	0	0	0.260	27.86	0.27
	ID	25	13	1	0	23	0	38	0.181	16.77	0.64
	TF	2	0	0	0	0	0	98	0.230	11.40	48.91
	CPOP	0	0	0	6	13	16	65	0.092	**3.98**	3.63
	BUP	0	0	0	0	0	0	100	0.211	10.59	3.97
(M7)	TS	52	35	11	1	1	0	0	0.353	12.30	0.65
	NOT	42	41	8	5	2	2	0	0.232	11.71	0.25
	ID	45	5	21	12	11	4	2	0.607	24.48	0.04
	TF	0	0	0	0	0	0	100	0.147	**6.04**	22.68
	CPOP	0	0	3	9	12	15	61	0.260	9.59	1.39
	BUP	0	0	4	**35**	42	18	1	0.232	**5.64**	1.54
(M8)	TS	0	0	0	67	0	31	2	0.026	8.72	1.01
	NOT	0	0	0	**99**	1	0	0	0.003	**0.01**	0.19
	ID	0	0	0	**99**	1	0	0	0.003	**0.01**	0.49
	TF	0	0	0	61	10	6	23	0.006	16.88	36.09
	CPOP	0	0	0	47	2	25	26	0.047	15.32	4.04
	BUP	0	0	0	0	0	0	100	0.056	48.08	2.66

C. Additional data application results

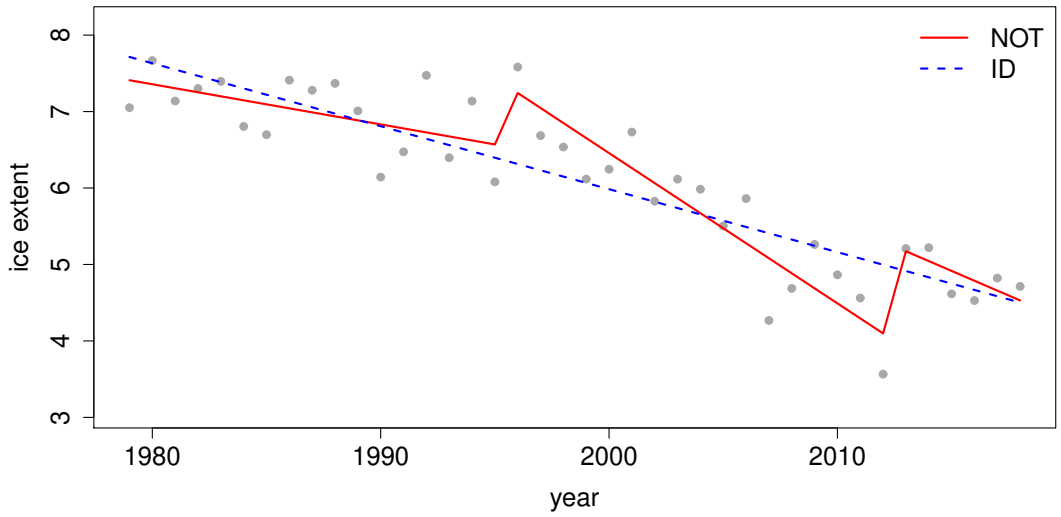


(a) NOT and ID

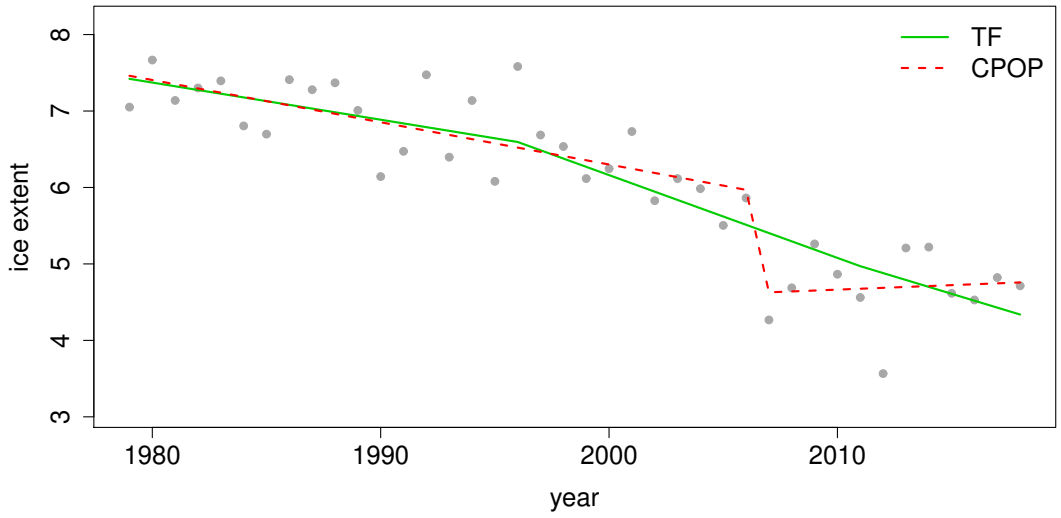


(b) TF and CPOP

Figure 1: Change-point analysis for the monthly average sea ice extent of the Arctic in February from 1979 to 2018 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 (- - -).

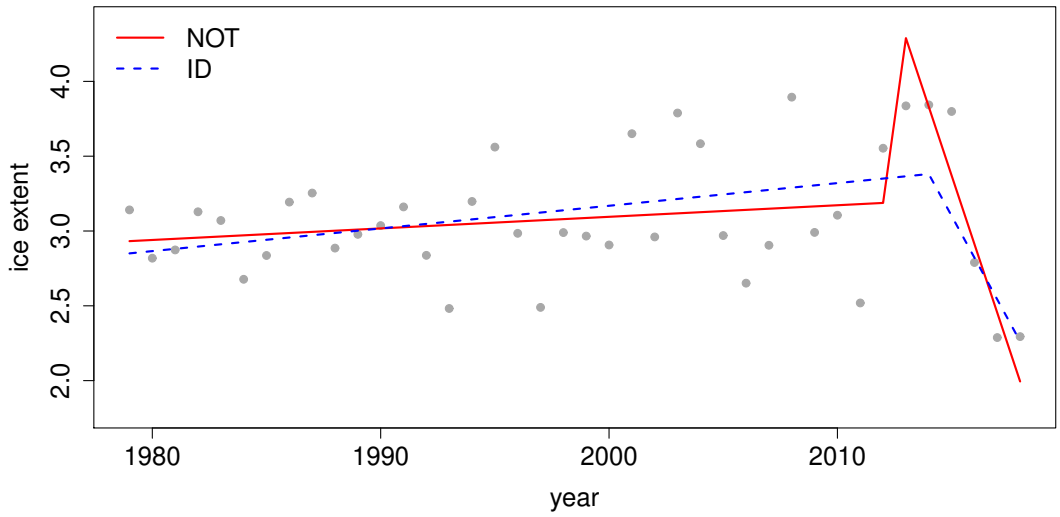


(a) NOT and ID

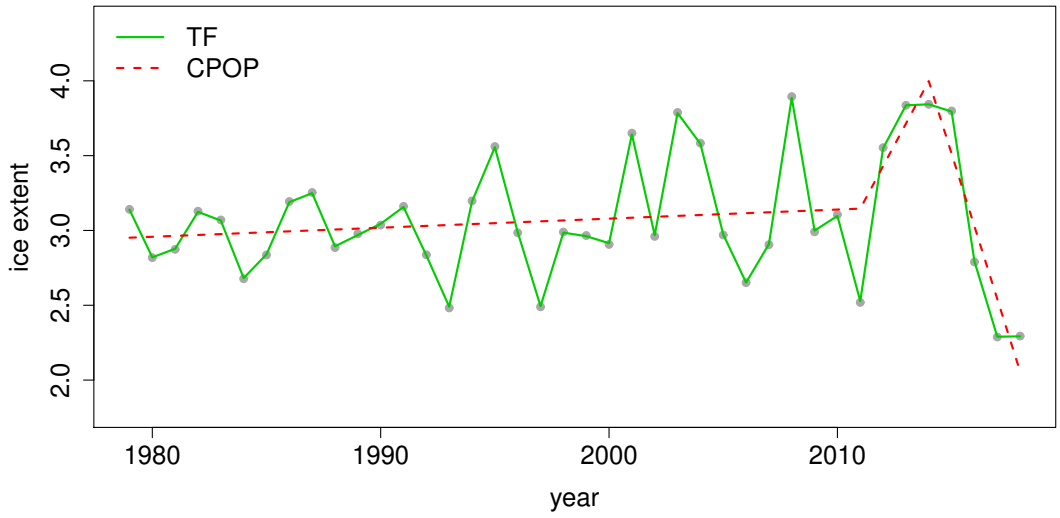


(b) TF and CPOP

Figure 2: Change-point analysis for the monthly average sea ice extent of the Arctic in September from 1979 to 2018 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 (- - -).

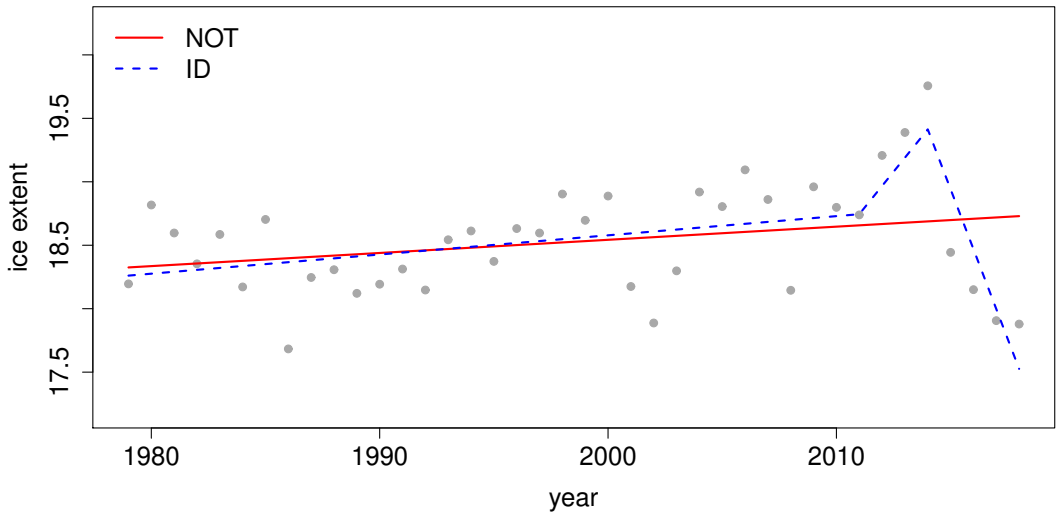


(a) NOT and ID

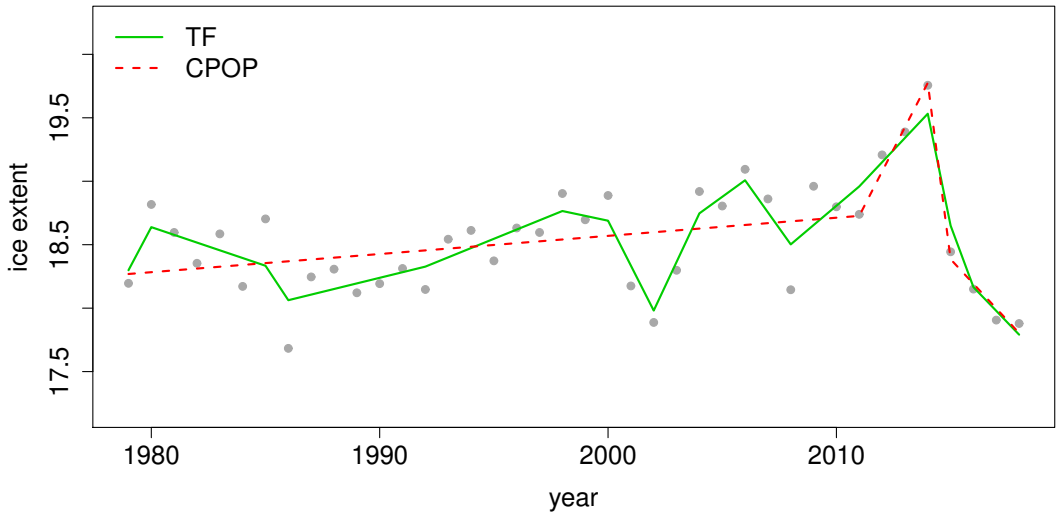


(b) TF and CPOP

Figure 3: Change-point analysis for the monthly average sea ice extent of the Antarctic in February from 1979 to 2018 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 4: Change-point analysis for the monthly average sea ice extent of the Antarctic in September from 1979 to 2018 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. Geometric interpretation of the TGUW transformation

In this section, we explore the shape of the unbalanced wavelet basis and offer the geometric interpretation of the TGUW transformation through an illustrative example. We recall that the TGUW transformation can be summarised as an orthonormal transformation of initial data X using orthonormal basis for $\mathbb{R}^T$ as follows,

$$\begin{pmatrix} s^1_{1,T} \\ s^2_{1,T} \\ \begin{pmatrix} d^{(j,k)}{}_{j=1,\ldots,J,k=1,\ldots,K(j)} \end{pmatrix} \end{pmatrix}_{T\times 1} = \begin{pmatrix} \psi^{(0,1)} \\ \psi^{(0,2)} \\ \begin{pmatrix} \psi^{(j,k)}{}_{j=1,\ldots,J,k=1,\ldots,K(j)} \end{pmatrix} \end{pmatrix}_{T\times T} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix}_{T\times 1} = \Psi_{T\times T} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix}, \tag{A.5}$$

where Ψ is an orthogonal matrix, $\{\psi^{(j,k)}\}$ is the set of vectors such that $d^{(j,k)} = \langle X, \psi^{(j,k)} \rangle$ and $(\psi^{(0,1)}, \psi^{(0,2)})$ are vectors satisfying $s^1_{1,T} = \langle X, \psi^{(0,1)} \rangle$ and $s^2_{1,T} = \langle X, \psi^{(0,2)} \rangle$, respectively. We note that there is a unique correspondence between $\psi^{(j,k)}_{p,q,r}$ and $\psi^{(j,k)}$ where $j = 1, \ldots, J$, $k = 1, \ldots, K(j)$ as is also the case for $d^{(j,k)}_{p,q,r}$ and $d^{(j,k)}$.

D.1 Shape of the unbalanced wavelet basis

The orthogonal transformation matrix Ψ in (A.5) contains T orthonormal basis in its rows which can be categorised into two: 1) $\left\{\psi^{(j,k)}{}_{j=1,\ldots,J,k=1,\ldots,K(j)}\right\}$ corresponding to detail coefficients $\left\{d^{(j,k)}{}_{j=1,\ldots,J,k=1,\ldots,K(j)}\right\}$ and 2) $\psi^{(0,1)}$ and $\psi^{(0,2)}$ corresponding to two smooth coefficients, $s^1_{1,T} = \langle X, \psi^{(0,1)} \rangle$ and $s^2_{1,T} = \langle X, \psi^{(0,2)} \rangle$, respectively.

We firstly investigate the shape of wavelet basis for detail coefficients as it plays an important role in the proof of theorems in Section B. The basis vector $\psi^{(j,k)}_{p,q,r}$ has length T but has non-zero values only from its p^{th} element to r^{th} element where $r - p \geq 2$ and the shape of non-zero region of $\psi^{(j,k)}_{p,q,r}$ depends on the type of merging which produces the corresponding detail coefficient. The general form in (A.2) implies that $\psi^{(j,k)}_{p,q,r}$ always has a shape of linear trend in previously merged region and this linearity is preserved as long as the later merges are performed under the "two together" rule. This is due to that the orthonormal transforms continue in a way of extending the dimension of space in which an orthonormal basis lives We now give a geometric interpretation of the TGUW transformation through an example to show how the linear trend in wavelet basis is preserved through the consecutive orthonormal transforms.

D.2 Example

We use a simple example to explain how the basis $\psi^{(j,k)}_{p,q,r}$ in (A.2) keep its linearity in subregions which are previously merged. Suppose we have the initial data sequence $\boldsymbol{s}^0 = (X_1, \ldots, X_5)$ and the

initial weight vectors for constancy and linearity are $\boldsymbol{w}^{c,0} = (1,1,1,1,1)^{\mathrm{T}}$ and $\boldsymbol{w}^{l,0} = (1,2,3,4,5)^{\mathrm{T}}$, respectively. As we have 5 initial observations, there would be 3 orthonormal transformations and the most important task for each transform is finding the appropriate orthonormal matrix as follows,

$$\Lambda = \begin{pmatrix} \ell_{1,1} & \ell_{1,2} & \ell_{1,3} \\ \ell_{2,1} & \ell_{2,2} & \ell_{2,3} \\ a & b & c \end{pmatrix} = \begin{pmatrix} \boldsymbol{\ell}_1^{\mathrm{T}} \\ \boldsymbol{\ell}_2^{\mathrm{T}} \\ \boldsymbol{h}^{\mathrm{T}} \end{pmatrix}. \tag{A.6}$$

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 matrix Λ, we firstly seek the detail filer, $\boldsymbol{h}$, which satisfies the conditions (1) $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{w}^{c,0}_{3,4,5} = 0$, (2) $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{w}^{l,0}_{3,4,5} = 0$ and (3) $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h} = 1$, where $\boldsymbol{w}^{\cdot}_{p,q,r}$ is $(p,q,r)^{th}$ element of $\boldsymbol{w}^{\cdot}$. Thus, $\boldsymbol{h}$ is obtained as a normal vector of the plane $\{(x,y,z) \mid x - 2y + z = 0\}$ as $\boldsymbol{w}^{c,0}_{3,4,5}$ and $\boldsymbol{w}^{l,0}_{3,4,5}$ are vectors of that plane. From the detail filter $\boldsymbol{h}$, two low filter vectors ($\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$) are obtained under the conditions, (1) $\boldsymbol{\ell}_1^{\mathrm{T}}\boldsymbol{h} = 0$, (2) $\boldsymbol{\ell}_2^{\mathrm{T}}\boldsymbol{h} = 0$, (3) $\boldsymbol{\ell}_1^{\mathrm{T}}\boldsymbol{\ell}_2 = 0$ and (4) $\boldsymbol{\ell}_1^{\mathrm{T}}\boldsymbol{\ell}_1 = \boldsymbol{\ell}_2^{\mathrm{T}}\boldsymbol{\ell}_2 = 1$. Therefore, $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ are obtained as an orthonormal basis of the plane $\{(x,y,z) \mid x - 2y + z = 0\}$ and this guarantees the linear trend of $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$. In summary, $\boldsymbol{w}^{c,0}_{3,4,5}, \boldsymbol{w}^{l,0}_{3,4,5}, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2$ are vectors of the plane $\{(x,y,z) \mid x - 2y + z = 0\}$ and $\boldsymbol{h}$ is a normal vector which is orthogonal to that plane. Now, the orthonormal transform updates the data sequence and weight vectors as follows,

$$\begin{aligned} \boldsymbol{s}^0 = (X_1, \ldots, X_5) \quad &\rightarrow \quad \boldsymbol{s} = (X_1, X_2, s^1_{3,4,5}, s^2_{3,4,5}, d_{3,4,5}), \\ \boldsymbol{w}^{c,0} = (1,1,1,1,1)^{\mathrm{T}} \quad &\rightarrow \quad \boldsymbol{w}^{c} = (1,1,e_{c_1}, e_{c_2}, 0)^{\mathrm{T}}, \\ \boldsymbol{w}^{l,0} = (1,2,3,4,5)^{\mathrm{T}} \quad &\rightarrow \quad \boldsymbol{w}^{l} = (1,2,e_{l_1}, e_{l_2}, 0)^{\mathrm{T}}, \end{aligned} \tag{A.7}$$

where constants (e_{c_1}, e_{c_2}) and (e_{l_1}, e_{l_2}) are obtained from $\Lambda\boldsymbol{w}^{c,0} =_{3,4,5} = (e_{c_1}, e_{c_2}, 0)^{\mathrm{T}}$ and $\Lambda\boldsymbol{w}^{l,0}_{3,4,5} = (e_{l_1}, e_{l_2}, 0)^{\mathrm{T}}$, respectively. As $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ are 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 $\boldsymbol{w}^{c,0}_{3,4,5}$ and $\boldsymbol{w}^{l,0}_{3,4,5}$ as a linear span of basis $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ as follows,

$$\boldsymbol{w}^{c,0}_{3,4,5} = e_{c_1}\boldsymbol{\ell}_1 + e_{c_2}\boldsymbol{\ell}_2, \quad \boldsymbol{w}^{l,0}_{3,4,5} = e_{l_1}\boldsymbol{\ell}_1 + e_{l_2}\boldsymbol{\ell}_2. \tag{A.8}$$

To show how the linear trend in the subregion of ψ is preserved through recursive orthonormal transforms, we now describe the process of obtaining the matrix $\Psi_{T\times T}$ in (A.5) (an orthonormal basis of $\mathbb{R}^5$ in this example) from the initial input $\Psi^0 = \mathbf{I}$ where $\mathbf{I}$ is 5×5 identity matrix. When $(p,q,r)^{th}$ values in $\boldsymbol{s}$ are selected to be merged, we extract the corresponding three columns of Ψ^{T} and update them by matrix multiplication with the orthogonal matrix Λ defined for that merge. For example, after the first orthonormal transform for (X_3, X_4, X_5) is performed as in (A.7), we

have the following updated matrix,

$$\Psi^{\mathrm{T}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ell_{1,1} & \ell_{2,1} & a \\ 0 & 0 & \ell_{1,2} & \ell_{2,2} & b \\ 0 & 0 & \ell_{1,3} & \ell_{2,3} & c \end{pmatrix}. \tag{A.9}$$

The 5$^{\text{th}}$ column of Ψ^{T} is now fixed (not going to be updated again) as it corresponds to the detail coefficient but other columns would be updated as the mergings continue.

Second merge From now on, the "two together" rule is applied. suppose that $(X_2, s^1_{3,4,5}, s^2_{3,4,5})$ are selected to be merged at next scale. Then we firstly select the corresponding weight vectors, $\boldsymbol{w}^c_{2,3,4} = (1, e_{c_1}, e_{c_2})^{\mathrm{T}}$ and $\boldsymbol{w}^l_{2,3,4} = (2, e_{l_1}, e_{l_2})^{\mathrm{T}}$ and construct the detail filter $\boldsymbol{h}^{*\mathrm{T}} = (a^*, b^*, c^*)$ which satisfies the conditions (1) $\boldsymbol{h}^{*\mathrm{T}}\boldsymbol{w}^c_{2,3,4} = 0$, (2) $\boldsymbol{h}^{*\mathrm{T}}\boldsymbol{w}^l_{2,3,4} = 0$ and (3) $\boldsymbol{h}^{*\mathrm{T}}\boldsymbol{h}^* = 1$. The detail filter is designed as a weight vector for investigating how strong linear trend exist in (X_2, X_3, X_4, X_5) as (e_{c_1}, e_{c_2}) and (e_{l_1}, e_{l_2}) contain the information of three values (X_3, X_4, X_5). Two low filters, $\boldsymbol{\ell}^*_1$ and $\boldsymbol{\ell}^*_2$, are obtained by satisfying the conditions, $\boldsymbol{\ell}^{*\mathrm{T}}_1\boldsymbol{h}^* = 0$, $\boldsymbol{\ell}^{*\mathrm{T}}_2\boldsymbol{h}^* = 0$, $\boldsymbol{\ell}^{*\mathrm{T}}_1\boldsymbol{\ell}^*_2 = 0$ and $\Lambda^{*\mathrm{T}}\Lambda^* = \mathbf{I}$, and the orthonormal transform is defined as follows,

$$\Lambda^* = \begin{pmatrix} \ell^*_{1,1} & \ell^*_{1,2} & \ell^*_{1,3} \\ \ell^*_{2,1} & \ell^*_{2,2} & \ell^*_{2,3} \\ a^* & b^* & c^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\ell}^{*\mathrm{T}}_1 \\ \boldsymbol{\ell}^{*\mathrm{T}}_2 \\ \boldsymbol{h}^{*\mathrm{T}} \end{pmatrix}, \tag{A.10}$$

Now the data sequence and weight vectors are updated as follows,

$$\begin{aligned} \boldsymbol{s} = (X_1, X_2, s^1_{3,4,5}, s^2_{3,4,5}, d_{3,4,5}) \quad &\rightarrow \quad \boldsymbol{s} = (X_1, s^1_{2,2,5}, s^2_{2,2,5}, d_{2,2,5}, d_{3,4,5}), \\ \boldsymbol{w}_c = (1, 1, e_{c_1}, e_{c_2}, 0)^{\mathrm{T}} \quad &\rightarrow \quad \boldsymbol{w}_c = (1, e^*_{c_1}, e^*_{c_2}, 0, 0)^{\mathrm{T}}, \\ \boldsymbol{w}_l = (1, 2, e_{l_1}, e_{l_2}, 0)^{\mathrm{T}} \quad &\rightarrow \quad \boldsymbol{w}_l = (1, e^*_{l_1}, e^*_{l_2}, 0, 0)^{\mathrm{T}}, \end{aligned} \tag{A.11}$$

and the wavelet matrix is also updated as follows,

$$\Psi^{\mathrm{T}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \ell^*_{1,1} & \ell^*_{2,1} & a^* & 0 \\ 0 & \begin{pmatrix} \\ \ell^*_{1,2}\boldsymbol{\ell}_1 + \ell^*_{1,3}\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} & \begin{pmatrix} \\ \ell^*_{2,2}\boldsymbol{\ell}_1 + \ell^*_{2,3}\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} & \begin{pmatrix} \\ b^*\boldsymbol{\ell}_1 + c^*\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} & a \\ 0 & & & & b \\ 0 & & & & c \end{pmatrix}. \tag{A.12}$$

At this scale, the 4$^{\text{th}}$ column of Ψ^{T} is fixed (not going to be updated again) and it corresponds to the second row of (A.2) whose non-zero subregion is composed of a single point (a^*) and a linear trend in three values $(b^*\boldsymbol{\ell}_1 + c^*\boldsymbol{\ell}_2)$.

Importantly, the orthonormal transform is performed in a way of returning an orthonormal basis of the expanded subspace e.g. 2nd and 3rd columns of (A.12) (which are referred to as $\boldsymbol{\ell}_1^{**}$ and $\boldsymbol{\ell}_2^{**}$ in (A.13)) 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 $\mathbf{S}$ in (A.13) which extends the dimension of the subspace from 3 to 4 but preserves the orthonormality of $\boldsymbol{\ell}_1^*$ and $\boldsymbol{\ell}_2^*$ which guarantees the properties, $\boldsymbol{\ell}_1^{**\mathrm{T}}\boldsymbol{\ell}_2^{**} = 0$ and $\boldsymbol{\ell}_1^{**\mathrm{T}}\boldsymbol{\ell}_1^{**} = \boldsymbol{\ell}_2^{**\mathrm{T}}\boldsymbol{\ell}_2^{**} = 1$, where

$$\boldsymbol{\ell}_1^{**} = \begin{pmatrix} \ell_{1,1}^* \\ \begin{pmatrix} \\ \ell_{1,2}^*\boldsymbol{\ell}_1 + \ell_{1,3}^*\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \end{pmatrix} = \mathbf{S}\begin{pmatrix} \ell_{1,1}^* \\ \ell_{1,2}^* \\ \ell_{1,3}^* \end{pmatrix}, \ \boldsymbol{\ell}_2^{**} = \begin{pmatrix} \ell_{2,1}^* \\ \begin{pmatrix} \\ \ell_{2,2}^*\boldsymbol{\ell}_1 + \ell_{2,3}^*\boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \end{pmatrix} = \mathbf{S}\begin{pmatrix} \ell_{2,1}^* \\ \ell_{2,2}^* \\ \ell_{2,3}^* \end{pmatrix}, \ \mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \begin{pmatrix} \\ \boldsymbol{\ell}_1 \\ \\ \end{pmatrix} & \begin{pmatrix} \\ \boldsymbol{\ell}_2 \\ \\ \end{pmatrix} \\ 0 & & \\ 0 & & \end{pmatrix}, \tag{A.13}$$

and $\mathbf{S}$ is obtained from the 2nd to 4th columns of (A.9) corresponding to the indices of smooth coefficients merged by the orthonormal matrix in (A.10).

As the linearity of $\boldsymbol{w}_{3,4,5}^{c,0}$ and $\boldsymbol{w}_{3,4,5}^{l,0}$ is expressed by a linear span of basis with unique constants in (A.8), now the extended subregion of the original weight vectors, $\boldsymbol{w}_{2:5}^{c,0}$ and $\boldsymbol{w}_{2:5}^{l,0}$, can also be written as a linear span of $\boldsymbol{\ell}_1^{**}$ and $\boldsymbol{\ell}_2^{**}$ (an orthonormal basis of the subspace $\{(w, x, y, z) \mid w - x = x - y = y - z\}$ of $\mathbb{R}^4$) as follows,

$$\boldsymbol{w}_{2:5}^{c,0} = e_{c_1}^*\boldsymbol{\ell}_1^{**} + e_{c_2}^*\boldsymbol{\ell}_2^{**}, \quad \boldsymbol{w}_{2:5}^{l,0} = e_{l_1}^*\boldsymbol{\ell}_1^{**} + e_{l_2}^*\boldsymbol{\ell}_2^{**}. \tag{A.14}$$

This can be simply shown by 1) expressing the weight vectors as a linear span of two low filters,

$$\begin{aligned} \boldsymbol{w}_{2,3,4}^{c} &= (1, e_{c_1}, e_{c_2})^{\mathrm{T}} = e_{c_1}^*\boldsymbol{\ell}_1^* + e_{c_2}^*\boldsymbol{\ell}_2^*, \\ \boldsymbol{w}_{2,3,4}^{l} &= (2, e_{l_1}, e_{l_2})^{\mathrm{T}} = e_{l_1}^*\boldsymbol{\ell}_1^* + e_{l_2}^*\boldsymbol{\ell}_2^*, \end{aligned} \tag{A.15}$$

and 2) performing the matrix multiplication with $\mathbf{S}$ in (A.13) to both sides of (A.15),

$$\begin{aligned} &\text{LHS} : \mathbf{S}\boldsymbol{w}_{2,3,4}^{c} = (1, e_{c_1}\boldsymbol{\ell}_1 + e_{c_2}\boldsymbol{\ell}_2)^{\mathrm{T}} = (1, 1, 1, 1)^{\mathrm{T}} = \boldsymbol{w}_{2:5}^{c,0}, \quad \text{RHS} : e_{c_1}^*\boldsymbol{\ell}_1^{**} + e_{c_2}^*\boldsymbol{\ell}_2^{**}, \\ &\text{LHS} : \mathbf{S}\boldsymbol{w}_{2,3,4}^{l} = (2, e_{l_1}\boldsymbol{\ell}_1 + e_{l_2}\boldsymbol{\ell}_2)^{\mathrm{T}} = (2, 3, 4, 5)^{\mathrm{T}} = \boldsymbol{w}_{2:5}^{l,0}, \quad \text{RHS} : e_{l_1}^*\boldsymbol{\ell}_1^{**} + e_{l_2}^*\boldsymbol{\ell}_2^{**}. \end{aligned} \tag{A.16}$$

Last merge In the same manner, if the last orthonormal transform is applied to $(X_1, s_{2,3,5}^1, s_{2,3,5}^2)$, we end up with the finalised Ψ^{T} which contains an orthonormal basis of the subspace $\{(v, w, x, y, z) \mid v - w = w - x = x - y = y - z\}$ in $\mathbb{R}^5$ in its first and second columns which correspond to two vectors, $\psi^{(0,1)}$ and $\psi^{(0,2)}$, in (A.5). In other words, regardless of the length of data T, first two wavelet vectors in Ψ which build two final smooth coefficients $(s_{1,T}^1, s_{1,T}^2)$ always have a linear trend with length T and other wavelet vectors corresponding to detail coefficients follow one of the forms in (A.2).

Although two low filters ($\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ in Λ) are not uniquely determined at each orthonormal transform, this has no effect on preserving the linearity of non-zero regions in a wavelet basis

vector as stated above. In simulation study, we empirically find that the choice of low filters has no effect on the results, thus use a fixed type of filter function rather than choosing an arbitrary filter every run which also saves the computational costs.