Supplementary Appendix for "Multiple change point detection under serial dependence: Wild contrast maximisation and gappy Schwarz algorithm"

Haeran Cho[1] Piotr Fryzlewicz[2]

A Algorithms

A.1 Wild Binary Segmentation 2 algorithm

Algorithm 1 provides a pseudo code for the Wild Binary Segmentation 2 (WBS2) algorithm proposed in Fryzlewicz (2020).

We remark that WBS2 as defined in Fryzlewicz (2020) uses random sampling in line 7 of Algorithm 1, but our preference is for deterministic sampling as it generates reproducible results without having to fix a random seed. To obtain at least $\widetilde{R}$ intervals over an equispaced (or almost equispaced, if exactly equal spacing is not possible) grid on a generic interval $[s, e]$, we firstly select the smallest integer $\widetilde{K}$ for which the number of all intervals with start- and end-points in the set $\{1, \ldots, \widetilde{K}\}$ equals or exceeds $\widetilde{R}$. Next, we map (linearly with rounding) the integer grid $[1, \widetilde{K}]$ onto an integer grid within $[s, e]$, as $j \to [\frac{e-s}{\widetilde{K}-1} j + s - \frac{e-s}{\widetilde{K}-1}]$ for each $j \in \{1, \ldots, \widetilde{K}\}$, where $[\cdot]$ represents rounding to the nearest integer. We then use all start- and end-points on the resulting grid to obtain the required collection (s_m, e_m) in line 7 of Algorithm 1.

A.2 Gappy Schwarz algorithm

For each $l \geq 1$, we denote $\widehat{\Theta}_l = \{\widehat{\theta}_{l,j},\, 1 \leq j \leq \widehat{q}_l : \widehat{\theta}_{l,1} < \ldots < \widehat{\theta}_{l,\widehat{q}_l}\}$, and adopt the notational convention that $\widehat{\theta}_{l,0} = 0$ and $\widehat{\theta}_{l,\widehat{q}_l+1} = n$. Initialised with $l = M$, gSa performs the following steps.

Step 1: We identify $u \in \{0, \ldots, \widehat{q}_{l-1}\}$ with $\{\widehat{\theta}_{l-1,u} + 1, \ldots, \widehat{\theta}_{l-1,u+1} - 1\} \cap \widehat{\Theta}_l \neq \emptyset$; that is, the segment $\{\widehat{\theta}_{l-1,u} + 1, \ldots, \widehat{\theta}_{l-1,u+1} - 1\}$ defined by the consecutive elements of $\widehat{\Theta}_{l-1}$, has additional change points detected in $\widehat{\Theta}_l$ such that $\{\widehat{\theta}_{l-1,u} + 1, \ldots, \widehat{\theta}_{l-1,u+1} - 1\} \cap$

[1]School of Mathematics, University of Bristol. Email: haeran.cho@bristol.ac.uk.
[2]Department of Statistics, London School of Economics. Email: p.fryzlewicz@lse.ac.uk.

Algorithm 1: Wild Binary Segmentation 2

Input: Data $\{X_t\}_{t=1}^n$, the number of intervals R_n

Function $\mathsf{wbs2}(\{X_t\}_{t=1}^n, R_n, s, e)$**:**

- **if** $e - s \leq 1$ **then return** $\emptyset$
- Let $\mathcal{A}_{s,e} \leftarrow \{(\ell, r) \in \mathbb{Z}^2 : s \leq \ell < r \leq e \text{ and } r - \ell > 1\}$
- **if** $|\mathcal{A}_{s,e}| \leq R_n$ **then**
  - $\widetilde{R} \leftarrow |\mathcal{A}_{s,e}|$ and set $\mathcal{R}_{s,e} \leftarrow \mathcal{A}_{s,e}$
- **else**
  - $\widetilde{R} \leftarrow R_n$ and draw $\widetilde{R}$ elements from $\mathcal{A}_{s,e}$ deterministically over an equispaced grid, to form $\mathcal{R}_{s,e} = \{1 \leq m \leq \widetilde{R} : (s_m, e_m)\}$
- **end**
- Identify $(s_\circ, k_\circ, e_\circ) = \arg\max_{(s_m,k,e_m):\, 1\leq m\leq \widetilde{R},\, s_m<k<e_m} |\mathcal{X}_{s_m,k,e_m}|$
- **return** $(s_\circ, k_\circ, e_\circ, |\mathcal{X}_{s_\circ,k_\circ,e_\circ}|) \cup \mathsf{wbs2}(\{X_t\}_{t=1}^n, R_n, s, k_\circ) \cup \mathsf{wbs2}(\{X_t\}_{t=1}^n, R_n, k_\circ, e)$

$\mathcal{P}_0 \leftarrow \mathsf{wbs2}(\{X_t\}_{t=1}^n, R_n, 0, n)$

Output: $\mathcal{P}_0$

$(\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}) \neq \emptyset$. By construction, the set of such indices, $\mathcal{I}_l := \{u_1, \ldots, u_{q_l'}\}$, satisfies $|\mathcal{I}_l| \geq 1$. For each u_v, $v = 1, \ldots, q_l'$, we repeat the following steps with a logical vector of length q_l', $\mathbf{F} \in \{\mathtt{TRUE}, \mathtt{FALSE}\}^{q_l'}$, initialised as $\mathbf{F} = (\mathtt{TRUE}. \ldots, \mathtt{TRUE})$.

Step 1.1: Setting $\mathcal{A} = \{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\} \cap \widehat{\Theta}_l$, obtain $\widehat{p}$ that returns the smallest $\mathrm{SC}(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \mathcal{A}, r)$ over $r \in \{0, \ldots, p_{\max}\}$ as outlined in (9), and the corresponding AR parameter estimator $\widehat{\boldsymbol{\alpha}}(\widehat{p})$ via least squares estimation.

Step 1.2: If $\mathrm{SC}(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \mathcal{A}, \widehat{p}) < \mathrm{SC}_0(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \widehat{\boldsymbol{\alpha}}(\widehat{p}))$, update $F_v \leftarrow \mathtt{FALSE}$.

Step 2: If some elements of $\mathbf{F}$ satisfy $F_v = \mathtt{TRUE}$ and $l > 1$, update $l \leftarrow l - 1$ and go to Step 1. If $F_v = \mathtt{FALSE}$ for all $v = 1, \ldots, q_l'$, return $\widehat{\Theta}_l$ as the set of change point estimators. Otherwise, return $\widehat{\Theta}_0 = \emptyset$.

Theorem 3.1 shows that we have either $F_v = \mathtt{FALSE}$ for all $v = 1, \ldots, q_l'$ when the corresponding $\widehat{\Theta}_l = \widehat{\Theta}_{l^*}$ (see Assumption 3.2 for the definition of $\widehat{\Theta}_{l^*}$), or $F_v = \mathtt{TRUE}$ for all v when $l > l^*$ and thus all $\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}$ are spurious estimators. In implementing the methodology, we take a conservative approach in the above Step 2, to guard against the unlikely event where the output $\mathbf{F}$ contains mixed results.

B Refinement of change point estimators

Throughout this section, we condition on the event that $\widehat{\Theta}[q]$ is chosen at the model selection step, and discuss how the location estimators can further be refined; consistent model selection

based on the estimators of change point locations returned directly by WBS2 (without any additional refinement), is discussed in Section 3.

By Theorem 2.1 and Assumption 2.2, each $\widehat{\theta}_j$, $1 \le j \le q$, is sufficiently close to the corresponding change point θ_j in the sense that $|\widehat{\theta}_j - \theta_j| \le (f'_j)^{-2}\rho_n \le c\delta_j$ for some $c \in (0, 1/6)$ with probability tending to one, for n large enough. Defining $\ell_1 = 0$, $r_q = n$,

$$\ell_j = \left\lfloor \frac{2}{3}\widehat{\theta}_{j-1} + \frac{1}{3}\widehat{\theta}_j \right\rfloor, \quad j = 2, \ldots, q, \quad \text{and} \quad r_j = \left\lfloor \frac{1}{3}\widehat{\theta}_j + \frac{2}{3}\widehat{\theta}_{j+1} \right\rfloor, \quad j = 1, \ldots, q-1,$$

we have each interval (ℓ_j, r_j) sufficiently large and contain a single change point θ_j well within its interior, i.e.

$$\min(\theta_j - \ell_j, r_j - \theta_j) \ge (2/3 - c)\delta_j > \delta_j/2, \quad \text{and} \tag{B.1}$$

$$\min(\ell_j - \theta_{j-1}, \theta_{j+1} - r_j) \ge (1/3 - c)\delta_j > 0. \tag{B.2}$$

Then, we propose to further refine the location estimator $\widehat{\theta}_j$ by $\check{\theta}_j = \arg\max_{\ell_j < k < r_j} \left|\mathcal{X}_{\ell_j,k,r_j}\right|$, which generally improves the localisation rate. To see this, we impose the following assumption on the error distribution which, by its formulation, trivially holds under Assumption 2.1 with $\widetilde{\zeta}_n = \zeta_n$. However, we often have the assumption met with a much tighter bound as discussed in Remark B.1, which leads to the improvement in the localisation rate of the refined estimators $\check{\theta}_j$ as shown in Proposition B.1.

Assumption B.1. For any sequence $1 \le a_n \le \min_{1 \le j \le q}(f'_j)^2\delta_j$ and some $\widetilde{\zeta}_n$ satisfying $\widetilde{\zeta}_n = O(\zeta_n)$ (with ζ_n as in Assumption 2.1), let $\mathsf{P}(\widetilde{\mathcal{Z}}_n) \to 1$ where

$$\widetilde{\mathcal{Z}}_n = \left\{ \max_{1 \le j \le q} \max_{(f'_j)^{-2}a_n \le \ell \le \theta_j - \theta_{j-1}} \frac{\sqrt{(f'_j)^{-2}a_n}}{\ell} \left| \sum_{t=\theta_j - \ell + 1}^{\theta_j} Z_t \right| \le \widetilde{\zeta}_n \right\}$$
$$\bigcap \left\{ \max_{1 \le j \le q} \max_{(f'_j)^{-2}a_n \le \ell \le \theta_{j+1} - \theta_j} \frac{\sqrt{(f'_j)^{-2}a_n}}{\ell} \left| \sum_{t=\theta_j + 1}^{\theta_j + \ell} Z_t \right| \le \widetilde{\zeta}_n \right\}.$$

Proposition B.1. Let the assumptions of Theorem 2.1 and Assumption B.1 hold. Then, there exists $c_3 \in (0, \infty)$ such that

$$\mathsf{P}\left(\max_{1 \le j \le q} (f'_j)^2 |\check{\theta}_j - \theta_j| \le c_3 (\widetilde{\zeta}_n)^2 \right) \ge \mathsf{P}\left(\mathcal{Z}_n \cap \widetilde{\mathcal{Z}}_n \right) \to 1.$$

Remark B.1. When the number of change points q is bounded, Assumption B.1 holds with $\widetilde{\zeta}_n$ diverging at an arbitrarily slow rate, provided that

$$\mathsf{E}\left| \sum_{t=l+1}^{r} Z_t \right|^{\nu} \le C(r-l)^{\nu/2} \quad \text{for any} \quad -\infty < l < r < \infty \tag{B.3}$$

for some constant $C > 0$ and $\nu > 2$, see Proposition 2.1 (c.ii) of Cho and Kirch (2021). The condition (B.3) is satisfied by many time series models, see Appendix B.2 in Kirch (2006) and the references therein. On the other hand, Theorem 1 of Shao and Zhang (2010) indicates that the lower bound $\sqrt{\log(n)} = O(\zeta_n)$ cannot be improved. Therefore, Proposition B.1 shows that the extra step indeed improves upon the localisation rate attained by the WBS2 reported in Theorem 2.1 (i). In fact, for time series models satisfying (B.3), the refinement leads to $(f'_j)^2|\check{\theta}_j - \theta_j| = O_p(1)$, thus matching the minimax optimal rate of multiple change point localisation (see Proposition 6 of Verzelen et al. (2020)).

C Implementation and the choice of tuning parameters

In line with the condition (5) and Assumption 3.2, we set $Q_n = \lfloor \log^{1.9}(n) \rfloor$, which imposes an upper bound on the number of change points, and we allow for at most $M = 5$ nested change point models (in addition to the null model) to be considered by the model selection methodology. By default, the number of intervals drawn by the deterministic sampling in Algorithm 1 is set at $R_n = 100$, and the maximum AR order is set at $p_{\max} = 10$ unless stated otherwise when input time series is short. To ensure that there are enough observations over each interval defined by two adjacent candidate change point estimators for numerical stability, we set the minimum spacing to be $\max(20, p_{\max} + \lceil \log(n) \rceil)$ and feed this into Algorithm 1 in the solution path generation. Finally, the penalty of SC is given by $\xi_n = \log^{1.01}(n)$ which is in accordance with Assumption 3.4 when the innovations $\{\varepsilon_t\}$ are distributed as (sub-)Gaussian random variables such that $\omega_n \asymp \sqrt{\log(n)}$ fulfils Assumption 3.1 (iv).

D Complete simulation studies

In this section, we present the complete simulation results summarised in Section 4.1 of the main text.

D.1 Set-up

We consider a variety of data generating processes for $\{X_t\}$; in the following, we assume $\varepsilon_t \sim_{\text{iid}} \mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 1$ unless stated otherwise. In addition to (M1)–(M3), we simulate datasets under the following scenarios. We also consider the case where $f_t = 0$ in each setting, to evaluate the size control performance of the methods considered in the comparative simulation study (their descriptions are given below the list of data generating processes).

(M4) f_t undergoes $q = 5$ change points at $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (100, 300, 500, 550, 750)$ with $n = 1000$ and $(f_0, f'_1, f'_2, f'_3, f'_4, f'_5) = (0, 1, -1, 2, -2, -1)$, and $Z_t = \varepsilon_t$.

(M5) f_t undergoes $q = 2$ change points at $(\theta_1, \theta_2) = (75, 125)$ with $n = 200$ and $(f_0, f_1', f_2') = (0, 2.5, -2.5)$, and $\{Z_t\}$ follows an ARMA$(1, 1)$ model: $Z_t = a_1 Z_{t-1} + \varepsilon_t + b_1 \varepsilon_t$ with $a_1 = 0.5$, $b_1 = 0.3$ and $\sigma_\varepsilon = 1/2.14285$.

(M6) f_t undergoes $q = 2$ change points at $(\theta_1, \theta_2) = (50, 100)$ with $n = 150$ and $(f_0, f_1', f_2') = (0, 2.5, -2.5)$, and $\{Z_t\}$ follows an AR(1) model: $Z_t = a_1 Z_{t-1} + \varepsilon_t$ with $a_1 = 0.5$ and $\sigma_\varepsilon = \sqrt{1 - a_1^2}$.

(M7) f_t undergoes $q = 2$ change points at $(\theta_1, \theta_2) = (100, 200)$ with $n = 300$ and $(f_0, f_1', f_2') = (0, 1, -1)$, and $\{Z_t\}$ follows an ARMA(1, 1) model: $Z_t = a_1 Z_{t-1} + \varepsilon_t + b_1 \varepsilon_{t-1}$ with the ARMA parameters are generated as $a_1, b_1 \sim_{\text{iid}} \mathcal{U}(-0.9, 0.9)$ for each realisation, and $\sigma_\varepsilon = \sqrt{(1 - a_1^2)/(1 + a_1 b_1 + b_1^2)}$.

(M8) f_t undergoes $q = 5$ change points at $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (100, 300, 500, 550, 750)$ with $n = 1000$ and $(f_0, f_1', f_2', f_3', f_4', f_5') = (0, 1, -1, 2, -2, -1)$, and $\{Z_t\}$ follows an MA(1) model $Z_t = \varepsilon_t + b_1 \varepsilon_{t-1}$ with $b_1 = 0.3$.

(M9) f_t undergoes $q = 5$ change points as in (M4) with $n = 1000$ and $(f_0, f_1', f_2', f_3', f_4', f_5') = (0, 3, -3, 4, -4, -3)$, and $\{Z_t\}$ follows an MA(4) model: $Z_t = \varepsilon_t + 0.9\varepsilon_{t-1} + 0.8\varepsilon_{t-2} + 0.7\varepsilon_{t-3} + 0.6\varepsilon_{t-4}$.

(M10) f_t undergoes $q = 15$ change points at $\theta_j = \lceil nj/16 \rceil$, $j = 1, \ldots, 15$ with $n = 2000$, where the level parameters $f_{\theta_j + 1}$ are generated uniformly as $(-1)^j \cdot f_{\theta_j + 1} \sim_{\text{iid}} \mathcal{U}(1, 2)$, $j = 0, \ldots, 15$, for each realisation. $\{Z_t\}$ follows an AR(1) model as in (M6) with $a_1 = 0.5$.

(M11) f_t undergoes $q = 10$ change points at $\theta_j = 150j$, $j = 1, \ldots, 10$ with $n = 1650$ and $(f_0, f_1', f_2', f_3', f_4', f_5', f_6', f_7', f_8', f_9', f_{10}') = (0, 7, -7, 6, -6, 5, -5, 4, -4, 3, -3)$, and $\{Z_t\}$ follows an ARMA(2, 6) model as in (M2).

(M12) f_t is as in (M4) and $\{Z_t\}$ follows a time-varying AR(1) model: $Z_t = a_1(t) Z_{t-1} + \sigma(t)\varepsilon_t$ with $a_1(t) = 0.5 - 0.2\cos(2\pi t/n)$ and $\sigma(t) = \sqrt{1 - a_1(t)^2}$.

(M13) f_t is as in (M4) and $\{Z_t\}$ follows a time-varying AR(1) model: $Z_t = a_1(t) Z_{t-1} + \sigma(t)\varepsilon_t$ where $a_1(t)$ is piecewise constant with change points at θ_j, $j = 1, \ldots, q$ such that $a_1(t) = 0.3\mathbb{I}_{t \le \theta_1} + 0.4\mathbb{I}_{\theta_1 < t \le \theta_2} + 0.6\mathbb{I}_{\theta_2 < t \le \theta_3} + 0.7\mathbb{I}_{\theta_3 < t \le \theta_4} + 0.5\mathbb{I}_{\theta_4 < t \le \theta_5} + 0.3\mathbb{I}_{t > \theta_5}$ and $\sigma(t) = \sqrt{1 - a_1(t)^2}$.

Apart from Model (M4), all others model have serial correlations in $\{Z_t\}_{t=1}^n$. Models (M5) (motivated by an example in Wu and Zhou (2020)), (M6) and (M7) consider relatively short time series with $n \in [150, 300]$. Models (M2), (M8) and (M9) are taken from Dette et al. (2020). In (M1), the LRV is close to zero and thus its accurate estimation is difficult. Models (M3) and (M10) have a teeth-like signal containing frequent change points and the underlying $\{Z_t\}$ has strong autocorrelations in (M3), and (M11) considers frequent, heterogeneous changes in the mean. In Models (M12) and (M13), the noise $\{Z_t\}_{t=1}^n$ has time-varying serial dependence structure.

We generate 1000 realisations under each model. For each scenario, we additionally consider the case in which $f_t \equiv 0$ (thus $q = 0$) in order to evaluate the proposed methodology on its size control. On each realisation, we apply the proposed WCM.gSa with the tuning parameters are selected as described in Section C. For comparison, we consider a procedure that omits the gappy model sequence generation step from WCM.gSa: referred to as 'no gap', it applies the SC-based model selection procedure directly to the model sequence consisting of consecutive entries from the WBS2-generated solution path.

We include DepSMUCE (Dette et al., 2020), DeCAFS (Romano et al., 2021), MACE (Wu and Zhou, 2020) and SNCP (Zhao et al., 2021) in the simulation studies. DepSMUCE extends the SMUCE procedure (Frick et al., 2014) proposed for independent data, by estimating the LRV using a difference-type estimator. MACE is a multiscale moving sum-based procedure with self-normalisation-based scaling that accounts for serial correlations. SNCP is a time series segmentation methodology that combines self-normalisation and a nested local window-based algorithm, and is applicable to detect multiple change points in a broad class of parameters. Although not its primary objective, DeCAFS can be adopted for the problem of detecting multiple change points in the mean of an otherwise stationary AR(1) process, and we adapt the main routine of its R implementation (Romano et al., 2020) to change point analysis under (1) as suggested by the authors. For DepSMUCE and MACE, we consider $\alpha \in \{0.05, 0.2\}$ and for SNCP, $\alpha \in \{0.01, 0.05, 0.1\}$ as per the codes provided by the authors. MACE requires the selection of the minimum and the maximum bandwidths in the rescaled time $[0, 1]$ and moreover, the latter, say $s_{\max}$, controls the maximum detectable number of change points to be $(2s_{\max})^{-1}$; we set $s_{\max} = \min(1/(3q), n^{-1/6})$ for fair comparison, which varies from one model to another. Other tuning parameters not mentioned here are chosen as recommended by the authors.

D.2 Results

Table D.1 summarises the performance of different change point detection methodologies included in the comparative simulation study under the null model $H_0 : q = 0$ and the alternative $H_1 : q > 1$. More specifically, we report the proportion of falsely detecting one or more change points under H_0 (size), as well as the following statistics under H_1: the distribution of the estimated number of change points, the relative mean squared error (MSE):

$$\sum_{t=1}^{n}(\widehat{f}_t - f_t)^2 / \sum_{t=1}^{n}(\widehat{f}_t^* - f_t)^2$$

where $\widehat{f}_t$ is the piecewise constant signal constructed with the set of estimated change point locations $\widehat{\Theta}$, and $\widehat{f}_t^*$ is an oracle estimator constructed with the true θ_j, and the Hausdorff

distance (d_H) between $\widehat{\Theta}$ and Θ:

$$d_H(\widehat{\Theta}, \Theta) = \max\left(\max_{\theta\in\Theta}\min_{\widehat{\theta}\in\widehat{\Theta}}|\theta-\widehat{\theta}|, \max_{\widehat{\theta}\in\widehat{\Theta}}\min_{\theta\in\Theta}|\widehat{\theta}-\theta|\right),$$

averaged over 1000 realisations.

Overall, across the various scenarios, WCM.gSa performs well under both the null and the alternative scenarios. In particular, it keeps the size at bay under H_0 regardless of the underlying serial correlation structure; when the time series is sufficiently long ($n \geq 300$), the proportion of the events where WCM.gSa spuriously detects any change point under H_0 is strictly below 0.05 (often below 0.01). Even when the input time series is short as in (M6) with $n = 150$, the proportion of such events is smaller than 0.1. Controlling for the size under H_0, especially in the presence of serial correlations, is a difficult task and as shown below, other methods considered in the comparative study fail to do so by a large margin in some scenarios.

Under H_1, WCM.gSa performs well in most scenarios according to a variety of criteria, such as model selection accuracy measured by $|\widehat{q}-q|$ or the localisation accuracy measured by d_H. The results under (M12)–(M13) show that WCM.gSa is able to handle mild nonstationarities in $\{Z_t\}_{t=1}^n$. Without the gappy model sequence generation step, the procedure suffers from having to perform a large number of model comparison steps, and the 'no gap' procedure tends to over-estimate the number of change points when q is large, or in the presence of mild nonstationarities in the noise. From this, we conclude that the gappy model sequence generation step plays an important role in final model selection by removing those candidate models that are not likely to be the one correctly detecting all change points from consideration.

DepSMUCE performs well for short series (see (M6)) or in the presence of weak serial correlations as in (M8), but generally suffers from a calibration issue. That is, in order not to detect spurious change points under H_0, it requires the tuning parameter to be set conservatively at $\alpha = 0.05$; however, for improved detection power, $\alpha = 0.2$ is a better choice. In addition, the estimator of the LRV proposed therein tends to under-estimate the LRV when it is close to zero as in (M1), or when there are strong autocorrelations as in (M3), thus incurring a large number of falsely detected change points under H_0.

Similar sensitivity to the choice of the level α is observable in the case of SNCP, and it tends to return spurious change point estimators when the time series is short as in (M5)–(M6), or when autocorrelations are strong as in (M3), and tends to under-estimate the number of change points generally with the exception of (M1).

DeCAFS operates under the assumption that $\{Z_t\}_{t=1}^n$ is an AR(1) process. Therefore, it is applied under model mis-specification in some scenarios, but still performs reasonably well in not returning false positives under H_0. The exception is (M3) where, in the presence of strong autocorrelations, it returns spurious estimators over 50% of realisations even though the model

is correctly specified in this scenario. Its detection power suffers under model mis-specification in some scenarios such as (M2) and (M9) when compared to WCM.gSa, but DeCAFS tends to attain good MSE. MACE suffers from both size inflation and lack of power, possibly due to its sensitivity to choice of some tuning parameters such as the bandwidths.

Table D.1: We report the proportion of rejecting H_0 (by returning $\widehat{q} \geq 1$) under $H_0 : q = 0$ (size) and the summary of estimated change points under $H_1 : q > 1$ according to the distribution of $\widehat{q} - q$, relative MSE and the Hausdorff distance (d_H) over 1000 realisations. Methods that control the size under H_0 (according to the specified α for DepSMUCE, MACE and SNCP, and at 0.05 for WCM.gSa and DeCAFS), and that achieve the best performance under H_1 according to different criteria, are highlighted in **bold** for each scenario.

			$\widehat{q} - q$								
Model	Method	Size	≥ -3	-2	-1	0	1	2	$3 \leq$	RMSE	d_H
(M1)	WCM.gSa	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	68.720	1.988
	no gap	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	68.720	1.988
	DepSMUCE(0.05)	1.000	0.000	0.000	0.000	0.485	0.167	0.163	0.185	219.196	48.359
	DepSMUCE(0.2)	1.000	0.000	0.000	0.000	0.170	0.093	0.177	0.560	437.883	90.818
	DeCAFS	0.064	0.000	0.006	0.029	0.742	0.148	0.053	0.022	304.694	26.274
	MACE(0.05)	0.222	0.000	0.000	0.922	0.078	0.000	0.000	0.000	1729.645	56.939
	MACE(0.2)	0.515	0.000	0.000	0.805	0.187	0.008	0.000	0.000	1724.294	65.194
	SNCP(0.01)	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**35.512**	**1.06**
	SNCP(0.05)	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**35.512**	**1.06**
	SNCP(0.1)	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**35.512**	**1.06**
(M2)	WCM.gSa	**0.001**	0.000	0.000	0.019	**0.873**	0.092	0.014	0.002	4.907	**34.627**
	no gap	**0.020**	0.002	0.002	0.012	0.178	0.024	0.037	0.745	11.030	148.765
	DepSMUCE(0.05)	**0.031**	0.052	0.385	0.429	0.134	0.000	0.000	0.000	18.567	145.406
	DepSMUCE(0.2)	**0.142**	0.006	0.093	0.410	0.490	0.001	0.000	0.000	11.066	83.157
	DeCAFS	0.099	0.006	0.035	0.137	0.773	0.049	0.000	0.000	**3.891**	61.517
	MACE(0.05)	0.682	0.767	0.157	0.064	0.012	0.000	0.000	0.000	40.977	316.419
	MACE(0.2)	0.874	0.477	0.273	0.156	0.083	0.009	0.002	0.000	33.876	286.084
	SNCP(0.01)	0.022	0.423	0.323	0.193	0.060	0.000	0.001	0.000	24.928	249.412
	SNCP(0.05)	0.084	0.117	0.293	0.372	0.215	0.002	0.001	0.000	15.428	166.724
	SNCP(0.1)	0.152	0.044	0.192	0.404	0.349	0.010	0.001	0.000	11.839	126.588
(M3)	WCM.gSa	**0.000**	0.087	0.177	0.233	0.319	0.076	0.041	0.067	3.184	86.139
	no gap	0.058	0.000	0.000	0.000	0.000	0.000	0.000	1.000	4.498	92.759
	DepSMUCE(0.05)	0.936	0.767	0.153	0.070	0.010	0.000	0.000	0.000	8.655	139.298
	DepSMUCE(0.2)	0.989	0.276	0.320	0.303	0.101	0.000	0.000	0.000	5.537	108.339
	DeCAFS	0.565	0.000	0.004	0.019	**0.755**	0.203	0.017	0.002	**1.065**	**19.751**
	MACE(0.05)	1.000	0.053	0.059	0.084	0.129	0.169	0.170	0.336	7.092	126.325

	MACE(0.2)	1.000	0.008	0.007	0.024	0.041	0.092	0.111	0.717	5.804	107.392
	SNCP(0.01)	0.105	0.995	0.004	0.000	0.001	0.000	0.000	0.000	14.135	430.912
	SNCP(0.05)	0.258	0.956	0.034	0.007	0.003	0.000	0.000	0.000	11.698	290.266
	SNCP(0.1)	0.397	0.890	0.074	0.027	0.009	0.000	0.000	0.000	10.342	245.351
(M4)	WCM.gSa	**0.000**	0.000	0.000	0.002	**0.994**	0.003	0.001	0.000	4.881	7.892
	no gap	**0.009**	0.000	0.000	0.000	0.873	0.026	0.044	0.057	5.587	21.121
	DepSMUCE(0.05)	**0.006**	0.000	0.000	0.104	0.896	0.000	0.000	0.000	6.671	22.699
	DepSMUCE(0.2)	**0.062**	0.000	0.000	0.016	0.984	0.000	0.000	0.000	4.901	9.21
	DeCAFS	**0.008**	0.000	0.000	0.000	0.983	0.015	0.002	0.000	**4.837**	**7.823**
	MACE(0.05)	0.558	0.681	0.242	0.062	0.013	0.002	0.000	0.000	97.279	311.77
	MACE(0.2)	0.816	0.370	0.328	0.212	0.073	0.015	0.002	0.000	82.773	253.051
	SNCP(0.01)	**0.003**	0.000	0.023	0.251	0.726	0.000	0.000	0.000	11.718	57.614
	SNCP(0.05)	**0.028**	0.000	0.002	0.093	0.898	0.007	0.000	0.000	7.916	24.667
	SNCP(0.1)	**0.065**	0.000	0.000	0.053	0.937	0.010	0.000	0.000	6.859	17.656
(M5)	WCM.gSa	0.080	0.000	0.000	0.000	0.884	0.086	0.015	0.015	2.753	4.583
	no gap	0.105	0.000	0.000	0.000	0.839	0.102	0.041	0.018	2.936	6.554
	DepSMUCE(0.05)	**0.028**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	2.051	**0.166**
	DepSMUCE(0.2)	**0.098**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	2.051	**0.166**
	DeCAFS	0.107	0.000	0.000	0.000	0.873	0.088	0.028	0.011	**1.970**	6.203
	MACE(0.05)	0.482	0.000	0.006	0.115	0.761	0.114	0.004	0.000	24.515	11.421
	MACE(0.2)	0.747	0.000	0.000	0.040	0.743	0.201	0.016	0.000	12.031	11.458
	SNCP(0.01)	0.086	0.000	0.000	0.002	0.945	0.052	0.001	0.000	9.839	2.764
	SNCP(0.05)	0.220	0.000	0.000	0.000	0.851	0.138	0.011	0.000	9.367	5.774
	SNCP(0.1)	0.328	0.000	0.000	0.000	0.778	0.193	0.027	0.002	9.652	8.315
(M6)	WCM.gSa	0.067	0.000	0.000	0.000	0.865	0.119	0.016	0.000	5.993	4.782
	no gap	0.074	0.000	0.000	0.000	0.865	0.119	0.016	0.000	5.993	4.782
	DepSMUCE(0.05)	**0.025**	0.000	0.006	0.202	0.792	0.000	0.000	0.000	14.038	9.14
	DepSMUCE(0.2)	**0.104**	0.000	0.000	0.041	**0.959**	0.000	0.000	0.000	**5.876**	**3.057**
	DeCAFS	0.193	0.000	0.005	0.005	0.751	0.099	0.074	0.066	7.867	9.537
	MACE(0.05)	0.621	0.000	0.143	0.433	0.391	0.033	0.000	0.000	41.943	25.549
	MACE(0.2)	0.812	0.000	0.052	0.288	0.584	0.075	0.001	0.000	29.655	20.355
	SNCP(0.01)	0.161	0.000	0.018	0.167	0.744	0.069	0.002	0.000	18.362	12.366
	SNCP(0.05)	0.367	0.000	0.005	0.054	0.740	0.177	0.022	0.002	12.173	9.618
	SNCP(0.1)	0.503	0.000	0.001	0.017	0.669	0.253	0.053	0.007	10.201	10.529
(M7)	WCM.gSa	**0.027**	0.000	0.102	0.001	**0.852**	0.025	0.009	0.011	**13.490**	**7.821**
	no gap	**0.044**	0.000	0.089	0.011	0.783	0.038	0.039	0.040	14.067	12.69
	DepSMUCE(0.05)	0.266	0.000	0.091	0.196	0.565	0.030	0.031	0.087	202.355	29.781
	DepSMUCE(0.2)	0.361	0.000	0.043	0.150	0.591	0.047	0.036	0.133	294.382	30.141

	DeCAFS	0.188	0.000	0.114	0.048	0.613	0.057	0.031	0.137	403.467	26.973
	MACE(0.05)	0.303	0.000	0.266	0.283	0.423	0.026	0.002	0.000	60.194	34.062
	MACE(0.2)	0.491	0.000	0.132	0.272	0.532	0.058	0.006	0.000	41.137	36.826
	SNCP(0.01)	0.061	0.000	0.147	0.191	0.654	0.007	0.001	0.000	18.293	22.939
	SNCP(0.05)	0.115	0.000	0.066	0.150	0.755	0.021	0.007	0.001	15.908	21.198
	SNCP(0.1)	0.159	0.000	0.032	0.143	0.778	0.030	0.015	0.002	14.410	22.208
(M8)	WCM.gSa	**0.000**	0.000	0.000	0.012	**0.972**	0.016	0.000	0.000	5.053	16.36
	no gap	**0.007**	0.000	0.000	0.004	0.850	0.036	0.046	0.064	5.707	29.525
	DepSMUCE(0.05)	**0.007**	0.006	0.117	0.472	0.405	0.000	0.000	0.000	15.523	114.702
	DepSMUCE(0.2)	**0.063**	0.000	0.009	0.201	0.790	0.000	0.000	0.000	7.204	44.676
	DeCAFS	**0.016**	0.000	0.003	0.004	0.969	0.022	0.001	0.001	**4.957**	**15.207**
	MACE(0.05)	0.565	0.816	0.141	0.036	0.006	0.001	0.000	0.000	64.459	338.846
	MACE(0.2)	0.808	0.523	0.269	0.162	0.035	0.011	0.000	0.000	54.656	286.868
	SNCP(0.01)	**0.008**	0.064	0.216	0.447	0.272	0.001	0.000	0.000	18.386	162.591
	SNCP(0.05)	**0.034**	0.005	0.080	0.355	0.554	0.006	0.000	0.000	11.438	94.291
	SNCP(0.1)	**0.074**	0.002	0.024	0.269	0.693	0.011	0.001	0.000	8.825	64.143
(M9)	WCM.gSa	**0.003**	0.000	0.001	0.003	**0.926**	0.059	0.008	0.003	4.776	**21.35**
	no gap	**0.012**	0.001	0.015	0.020	0.632	0.023	0.042	0.267	7.121	68.784
	DepSMUCE(0.05)	**0.020**	0.051	0.233	0.546	0.170	0.000	0.000	0.000	16.374	87.334
	DepSMUCE(0.2)	**0.127**	0.003	0.052	0.406	0.537	0.002	0.000	0.000	9.544	37.717
	DeCAFS	0.097	0.001	0.061	0.019	0.863	0.055	0.001	0.000	**3.779**	31.135
	MACE(0.05)	0.670	0.779	0.167	0.041	0.012	0.001	0.000	0.000	49.668	334.816
	MACE(0.2)	0.870	0.462	0.275	0.192	0.059	0.011	0.001	0.000	39.156	285.542
	SNCP(0.01)	0.021	0.292	0.361	0.252	0.094	0.001	0.000	0.000	21.119	201.372
	SNCP(0.05)	0.077	0.093	0.258	0.343	0.296	0.010	0.000	0.000	14.061	126.391
	SNCP(0.1)	0.152	0.033	0.180	0.352	0.417	0.016	0.002	0.000	11.489	93.392
(M10)	WCM.gSa	**0.000**	0.000	0.000	0.008	**0.982**	0.006	0.003	0.001	2.425	**5.485**
	no gap	**0.006**	0.000	0.000	0.000	0.511	0.055	0.070	0.364	3.480	34.066
	DepSMUCE(0.05)	**0.020**	0.118	0.332	0.380	0.170	0.000	0.000	0.000	20.085	85.553
	DepSMUCE(0.2)	**0.133**	0.003	0.048	0.338	0.611	0.000	0.000	0.000	7.534	39.648
	DeCAFS	**0.023**	0.000	0.000	0.000	0.974	0.023	0.003	0.000	**2.112**	5.564
	MACE(0.05)	0.902	0.917	0.049	0.026	0.007	0.000	0.001	0.000	61.743	232.45
	MACE(0.2)	0.984	0.628	0.173	0.110	0.050	0.028	0.009	0.002	47.687	177.494
	SNCP(0.01)	0.011	0.035	0.106	0.292	0.567	0.000	0.000	0.000	13.030	60.337
	SNCP(0.05)	**0.043**	0.002	0.022	0.165	0.811	0.000	0.000	0.000	9.461	29.324
	SNCP(0.1)	0.104	0.000	0.006	0.096	0.898	0.000	0.000	0.000	8.556	18.968
(M11)	WCM.gSa	**0.001**	0.080	0.360	0.252	**0.287**	0.013	0.006	0.002	5.435	**180.548**
	no gap	0.012	0.003	0.014	0.003	0.069	0.022	0.021	0.868	8.287	105.137

	DepSMUCE(0.05)	**0.022**	0.912	0.081	0.007	0.000	0.000	0.000	0.000	15.463	351.082
	DepSMUCE(0.2)	**0.126**	0.562	0.345	0.088	0.005	0.000	0.000	0.000	10.991	258.122
	DeCAFS	0.077	0.221	0.474	0.063	0.234	0.008	0.000	0.000	**4.831**	286.997
	MACE(0.2)	0.839	0.994	0.005	0.000	0.001	0.000	0.000	0.000	32.807	565.07
	MACE(0.05)	0.960	0.925	0.049	0.020	0.004	0.002	0.000	0.000	29.778	424.598
	SNCP(0.01)	0.011	0.990	0.009	0.001	0.000	0.000	0.000	0.000	23.936	510.673
	SNCP(0.05)	0.070	0.862	0.113	0.023	0.002	0.000	0.000	0.000	17.976	349.351
	SNCP(0.1)	0.126	0.706	0.206	0.081	0.007	0.000	0.000	0.000	15.070	290.98
(M12)	WCM.gSa	**0.002**	0.000	0.002	0.061	**0.718**	0.151	0.048	0.020	5.828	**50.476**
	no gap	**0.031**	0.002	0.010	0.016	0.501	0.058	0.082	0.331	7.648	73.266
	DepSMUCE(0.05)	0.074	0.155	0.450	0.350	0.045	0.000	0.000	0.000	16.612	232.209
	DepSMUCE(0.2)	0.273	0.026	0.177	0.471	0.325	0.001	0.000	0.000	10.426	139.304
	DeCAFS	0.081	0.009	0.079	0.074	0.717	0.094	0.023	0.004	**5.727**	82.021
	MACE(0.05)	0.675	0.790	0.161	0.043	0.005	0.001	0.000	0.000	33.749	327.001
	MACE(0.2)	0.873	0.537	0.249	0.151	0.050	0.012	0.001	0.000	28.311	304.191
	SNCP(0.01)	0.020	0.645	0.224	0.103	0.028	0.000	0.000	0.000	24.165	303.019
	SNCP(0.05)	0.081	0.265	0.324	0.286	0.122	0.003	0.000	0.000	16.420	218.013
	SNCP(0.1)	0.152	0.131	0.283	0.363	0.217	0.006	0.000	0.000	13.713	166.677
(M13)	WCM.gSa	**0.001**	0.000	0.002	0.043	0.831	0.089	0.030	0.005	5.442	**38.565**
	no gap	**0.023**	0.000	0.008	0.007	0.613	0.056	0.086	0.230	6.880	57.405
	DepSMUCE(0.05)	0.053	0.093	0.381	0.423	0.103	0.000	0.000	0.000	16.547	202.408
	DepSMUCE(0.2)	0.205	0.012	0.113	0.445	0.430	0.000	0.000	0.000	9.754	112.529
	DeCAFS	**0.041**	0.003	0.043	0.049	**0.834**	0.059	0.012	0.000	**5.069**	50.936
	MACE(0.05)	0.646	0.819	0.133	0.044	0.003	0.001	0.000	0.000	38.863	329.921
	MACE(0.2)	0.855	0.543	0.255	0.141	0.051	0.008	0.002	0.000	32.993	301.344
	SNCP(0.01)	0.015	0.470	0.304	0.175	0.051	0.000	0.000	0.000	22.871	280.454
	SNCP(0.05)	0.064	0.161	0.282	0.375	0.179	0.003	0.000	0.000	15.759	184.029
	SNCP(0.1)	0.134	0.077	0.209	0.397	0.311	0.005	0.001	0.000	12.778	137.346

D.3 Numerical experiments motivating the use of SC_0

If any change point is ignored in fitting an AR model, the information criterion SC tends to over-compensate for the under-specification of mean shifts, which makes direct minimisation of SC unreliable as a model selection method. To illustrate this and motivate the use of SC_0 in gSa, we present a simulation study with datasets generated under the models (M9) and (M11) in Section D.1. Here, our aim is to compare a change point model $\widehat{\Theta}_1$ (correctly detecting all q change points) and the null model $\widehat{\Theta}_0 = \emptyset$ using two different approaches – one adopted in gSa comparing $\mathrm{SC}_0\left(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(\widehat{p})\right)$ and $\mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}_1, \widehat{p})$ with $\widehat{p} = \widehat{p}(\widehat{\Theta}_1)$ ('Method 1'), and the other selecting the model minimising SC by comparing $\mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}_0, \widehat{p}(\widehat{\Theta}_0))$ and $\mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}_1, \widehat{p})$ ('Method 2'). In both scenarios, the errors do not follow an AR model of

a finite order so we select $\widehat{p}(\widehat{\Theta}_0)$ and $\widehat{p}(\widehat{\Theta}_1)$ as described in (9).
For the choice of $\widehat{\Theta}_1$, we consider the *no bias* case $\widehat{\Theta}_1 = \{\theta_j,\, 1 \le j \le q\}$ and the *biased* case $\widehat{\Theta}_1 = \{\theta_j + s_j \cdot \lambda_j,\, 1 \le j \le q\}$, where $s_j \sim_{\text{iid}} \text{Uniform}\{-1, 1\}$ and $\lambda_j \sim_{\text{iid}} \text{Poisson}(5)$; the latter case reflects that the best localisation rate in change point problems is $O_p(1)$. The result is summarised in Table D.2 where we report the size (proportion of selecting $\widehat{\Theta}_1$ over $\widehat{\Theta}_0$ when there is no change point), as well as the power (proportion of correctly selecting $\widehat{\Theta}_1$) out of 1000 realisations. From the results, we conclude that Method 1, which adopts SC_0 as a proxy of the goodness-of-fit adjusted by model complexity under the no change point model, works well both in controlling the size and attaining good power. In comparison, Method 2 suffers from loss of power due to the bias in AR parameter estimators in the presence of mean shifts, and its performance worsens when the change point estimators do not exactly coincide with the true locations, which is often the case in change point problems when the magnitude of the jumps is small.

Table D.2: Size and power of Methods 1 and 2 under the models (M9) and (M11) when the change point model is specified without any bias in change point estimators ('no bias') and with bias.

	(M9)				(M11)			
	No bias		Bias		No bias		Bias	
	Size	Power	Size	Power	Size	power	Size	Power
Method 1	0	1	0	1	0	1	0	0 0.989
Method 2	0	0.876	0	0.202	0	0.793	0	0.015

E Additional real data analysis

E.1 Pre-processing of nitrogen oxides concentrations data

The concentration measurements are positive integers and possibly highly skewed, see top panels of Figure E.1. Also, the data exhibit seasonality as well as weekly patterns, the latter particularly visible from the autocorrelations (see middle panels of Figure E.1), and the level of concentrations drops sharply on bank holidays, in line with the behaviour of road traffic. We adopt the square root transform in order to bring the data to light-tailedness without masking any shift in the level greatly. Also, after visual inspection and preliminary research into the relevant literature, we select the period between January 2004 and December 2010 to estimate the seasonal, weekly and bank holiday patterns, which is achieved by regressing the square root transformed time series onto the indicator variables representing their effects. In summary, 19 parameters including the intercept were estimated from the 2508 observations, and all three factors (seasonal, daily and bank holiday effects) were deemed significant, with the models fitted to the NO_2 and NO_x concentrations attaining the adjusted R^2 coefficients

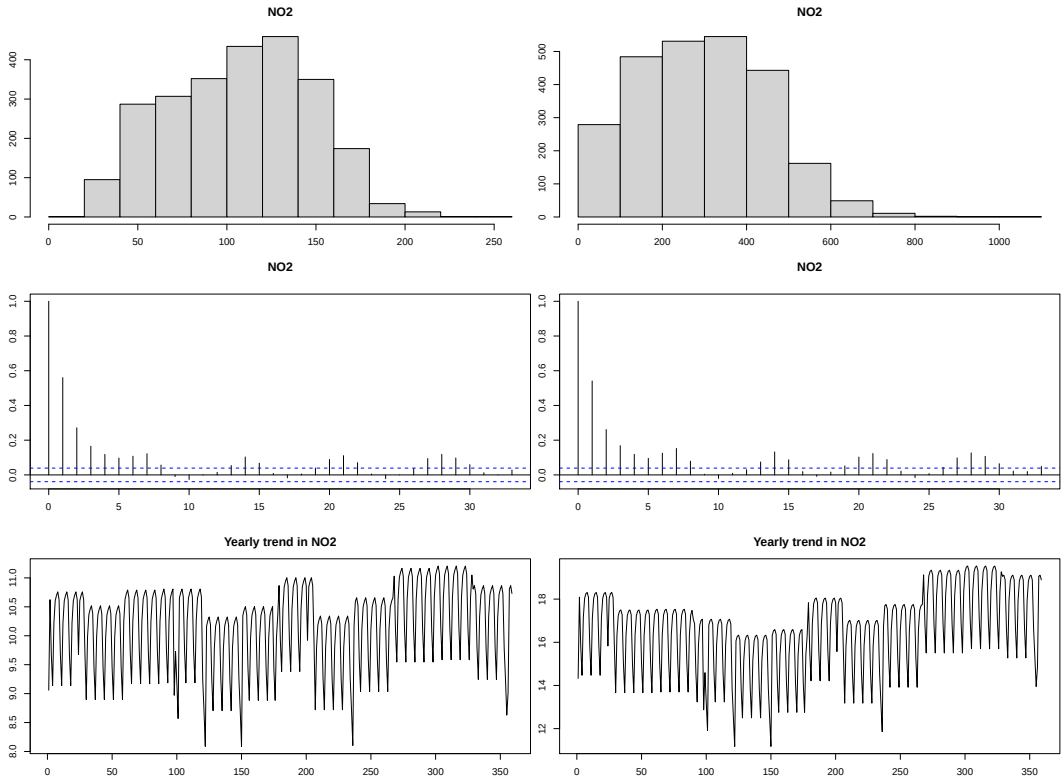

Figure E.1: Various statistical properties of the daily concentrations of NO_2(left) and NO_x (right) measured at Marylebone Road in London between January 2004 and December 2010. Top: histogram of raw concentrations. Middle: autocorrelations after square root transform. Bottom: yearly fitted patterns.

of 0.1077 and 0.1149, respectively. Bottom panels of Figure E.1 plot the fitted yearly trend, while Figure 1 in the main text plots the residuals, which we analyse for change points in the level.

E.2 Validating the number of change points detected from the NO_2 time series

Table 2 in the main paper shows a considerable variation in the number of detected change points in the NO_2 time series between the competing methods. To run an independent check for the number of change points, we firstly remove the bulk of the serial dependence of the data by fitting the AR(1) model to it and work with the empirical residuals from this fit. For this, we set the AR coefficient to 0.5, as suggested by the sample autocorrelation function in Figures E.1 and E.2. In particular, the latter figure confirms that the assumption of weak stationarity on the noise is well-satisfied by the NO_2 time series, with the leading autocorrelations remaining approximately the same across the segments defined by the change points estimated by WCM.gSa.

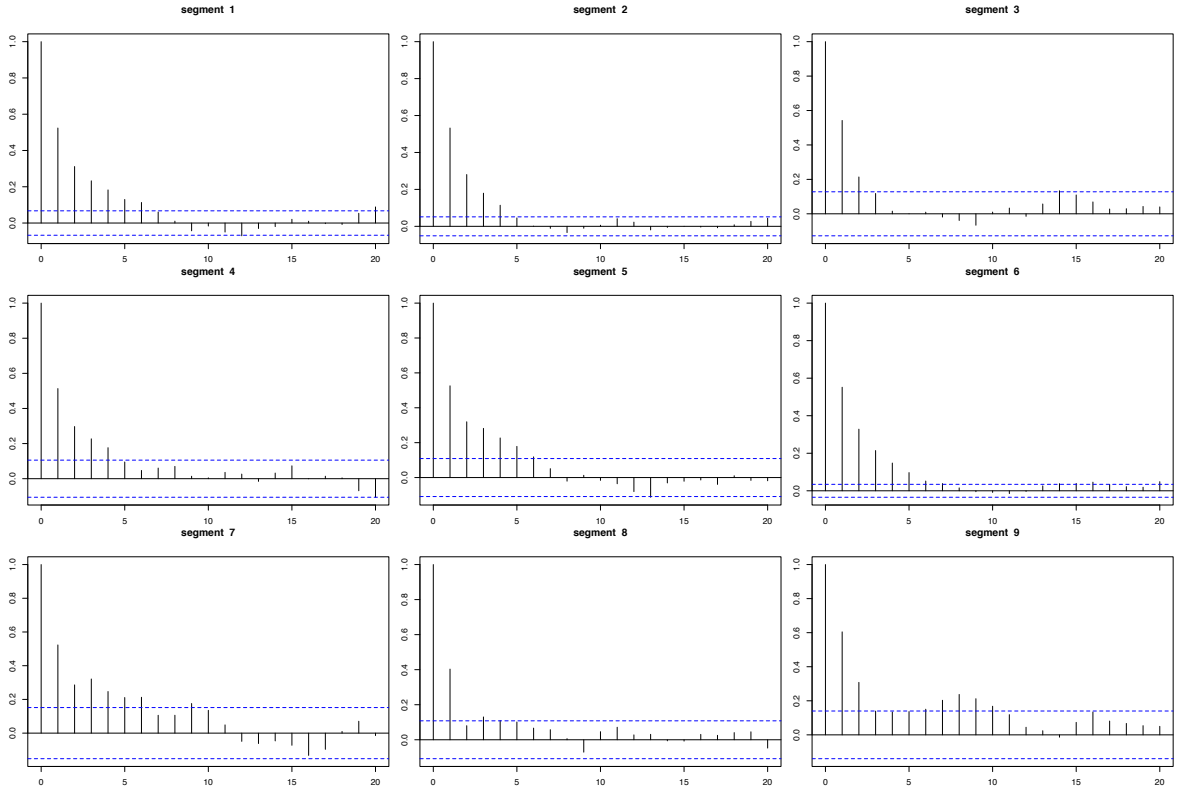

Figure E.2: Autocorrelations at 20 lags computed from the nine segments defined by the change point estimators returned by WCM.gSa when applied to the de-trended and transformed NO_2 measurements.

On these, we perform change point detection using a method suitable for multiple level-shift detection under serially uncorrelated noise. The method we use is the IDetect technique with the information-criterion-based model selection (Anastasiou and Fryzlewicz, 2020), as implemented in the R package `breakfast` (Anastasiou et al., 2020). The reason for the selection of this method is that it is possibly the best-performing method of the package overall (as reported in the package vignette available at `https://cran.r-project.org/web/packages/breakfast/vignettes/breakfast-vignette.html`), and it is independently commended in Fearnhead and Rigaill (2020) as having very strong performance overall.

The R execution `model.ic(sol.idetect(no2.res))$cpts`, where `no2.res` are the residuals obtained as above, returns 7 change point estimators, a number close to the 8 obtained by our WCM.gSa method. Out of the 7 locations estimated by IDetect, there is very good agreement with WCM.gSa for 6 out of these locations. The exception is the WCM.gSa-estimated change point at 2010-07-25, which IDetect estimates some 800 days later. However, IDetect also does not estimate the following WCM.gSa-estimated change point at 2018-10-13, which is a possible reason for IDetect to replace these two WCM.gSa-estimated change points by one in between them.

This, in our view, represents very good agreement on the whole, especially given that the two methods are entirely different in nature and worked with different time series on input. This result further enhances our confidence in the output of WCM.gSa for this dataset.

E.3 Hadley Centre central England temperature data analysis

The Hadley Centre central England temperature (HadCET) dataset (Parker et al., 1992) contains the mean, maximum and minimum daily and monthly temperatures representative of a roughly triangular area enclosed by Lancashire, London and Bristol, UK.

We analyse the yearly average of the monthly mean, maximum and minimum temperatures up to 2019 for change points using the proposed WCM.gSa methodology. The mean monthly data dates back to 1659, while the maximum and the minimum monthly data begins in 1878; we focus on the period of 1878–2019 ($n = 142$) for all three time series. To take into account that the time series are relatively short, we set $p_{\max} = 5$ (maximum allowable AR order) for WCM.gSa and the minimum spacing to be 10 (i.e. no change points occur within 10 years from one another), while the rest of the parameters are chosen as recommended in Section C; the results are invariant to the choice of the penalty $\xi_n \in \{\log^{1.01}(n), \log^{1.1}(n)\}$. Table E.1 reports the change points estimated by WCM.gSa as well as those detected by DepSMUCE and DeCAFS for comparison.

On all three datasets, WCM.gSa and DeCAFS return identical estimators, and the same change points are detected by DepSMUCE (with $\alpha = 0.2$). Figure E.3 shows that there appears to be a noticeable change in the persistence of the autocorrelations in the datasets before and after these shifts in the mean are accounted for, which further confirms that the

yearly temperatures undergo level shifts over the years. In particular, the second change point detected at 1987/88 coincides with the global regime shift in Earth's biophysical systems identified around 1987 (Reid et al., 2016), which is attributed to anthropogenic warming and a volcanic eruption.

Table E.1: Change points (in year) detected from the yearly average of the mean, maximum and minimum monthly temperatures from 1878 to 2019.

Method	Mean	Maximum	Minimum
WCM.gSa	1892, 1988	1892, 1988	1892, 1987
DepSMUCE(0.05)	1987	1988	1956
DepSMUCE(0.2)	1892, 1988	1988	1892, 1987
DeCAFS	1892, 1988	1892, 1988	1892, 1987

F Proofs

For any square matrix $\mathbf{B} \in \mathbb{R}^{p\times p}$, let $\lambda_{\max}(\mathbf{B})$ and $\lambda_{\min}(\mathbf{B})$ denote the maximum and the minimum eigenvalues of $\mathbf{B}$, respectively, and we define the operator norm $\|\mathbf{B}\| = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$. Let $\mathbf{1}$ denote a vector of ones, $\mathbf{0}$ a vector of zeros and $\mathbf{I}$ an identity matrix whose dimensions are determined by the context. The projection matrix onto the column space of a given matrix $\mathbf{A}$ is denoted by $\mathbf{\Pi_A} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$, provided that $\mathbf{A}^\top\mathbf{A}$ is invertible. We write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

F.1 Proof of the results in Section 2

Throughout the proofs, we work under the following non-asymptotic bound

$$\max\left(\frac{n^\varphi \zeta_n^2}{\min_{1\le j\le q}(f_j')^2\delta_j}, \frac{1}{\log(\zeta_n)}\right) \le \frac{1}{K} \tag{F.1}$$

for some $K > 0$, which holds for all $n \ge n(K)$ for some large enough $n(K)$, which replaces the asymptotic condition in Assumptions 2.2 and (5). The o-notation always refers to K in (F.1) being large enough, which in turn follows for large enough n. By $\mathcal{F}_{s,k,e}$ and $\mathcal{Z}_{s,k,e}$, we denote the CUSUM statistics defined with $\{f_t\}$ and $\{Z_t\}$ replacing $\{X_t\}$ in (2), respectively.

F.1.1 Preliminaries

Lemma F.1 (Lemma B.1 of Cho and Kirch (2021))**.** For $\max(s, \theta_{j-1}) < k < \theta_j < \min(e, \theta_{j+1})$, it holds that

$$\mathcal{F}_{s,k,e} = -\sqrt{\frac{(k-s)(e-k)}{e-s}}\left\{\frac{(e-\theta_j)\,f_j'}{e-k} + \frac{(e-\theta_{j+1})_+\,f_{j+1}'}{e-k} + \frac{(\theta_{j-1}-s)_+\,f_{j-1}'}{k-s}\right\},$$

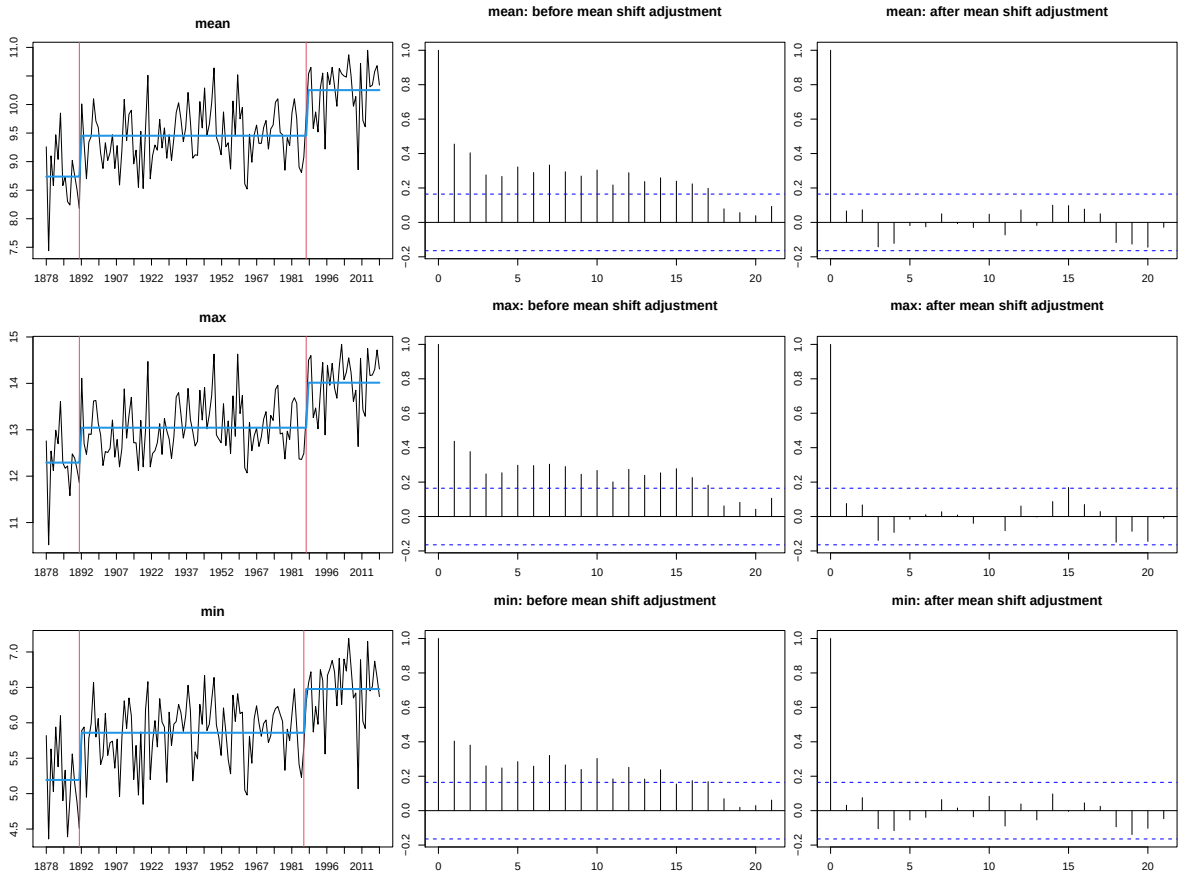

Figure E.3: Left: yearly average of the mean, maximum and minimum monthly temperatures (top to bottom), plotted together with the change points estimated by WCM.gSa (vertical lines) and piecewise constant mean (bold lines). Middle and right: autocorrelation function of the data without and with the time-varying mean adjusted.

where $a_+ = a \cdot \mathbb{I}_{a \geq 0}$. Similarly, for $\max(s, \theta_{j-1}) < \theta_j \leq k < \min(e, \theta_{j+1})$, it holds that

$$\mathcal{F}_{s,k,e} = -\sqrt{\frac{(k-s)(e-k)}{e-s}} \left\{ \frac{(\theta_j - s)\, f'_j}{k-s} + \frac{(e-\theta_{j+1})_+\, f'_{j+1}}{e-k} + \frac{(\theta_{j-1}-s)_+\, f'_{j-1}}{k-s} \right\}.$$

Lemma F.2 (Lemma 2.2 of Venkatraman (1992); Lemma 8 of Wang and Samworth (2018))**.** For some $0 \leq s < e \leq n$ with $e - s > 1$, let $\Theta \cap [s, e] = \{\theta^\circ_1, \ldots, \theta^\circ_m\}$ with $m \leq q$, and we adopt the notations $\theta^\circ_0 = s$ and $\theta^\circ_{m+1} = e$. If the series $\mathcal{F}_{s,k,e}$ is not constantly zero for $\theta^\circ_j + 1 \leq k \leq \theta^\circ_{j+1}$ for some $j = 0, \ldots, m$, one of the following is true:

(i) $j = 0$ and $\mathcal{F}_{s,k,e}$, $\theta^\circ_j + 1 \leq k \leq \theta^\circ_{j+1}$ does not change sign and has strictly increasing absolute values,

(ii) $j = m$ and $\mathcal{F}_{s,k,e}$, $\theta^\circ_j + 1 \leq k \leq \theta^\circ_{j+1}$ does not change sign and has strictly decreasing absolute values,

(iii) $1 \le j \le m-1$ and $\mathcal{F}_{s,k,e}$, $\theta_j^\circ + 1 \le k \le \theta_{j+1}^\circ$ is strictly monotonic,

(iv) $1 \le j \le m-1$ and $\mathcal{F}_{s,k,e}$, $\theta_j^\circ + 1 \le k \le \theta_{j+1}^\circ$ does not change sign and its absolute values are strictly decreasing then strictly increasing.

F.1.2 Proof of Theorem 2.1

Throughout the proofs, $C_0, C_1, \ldots$ denote some positive constants.
We define the following intervals for each $j = 0, \ldots, q_n$,

$$I_{L,j} = (\theta_{j-1}, \theta_j - \lceil \delta_j/3 \rceil) \quad \text{and} \quad I_{R,j} = (\theta_j + \lceil \delta_j/3 \rceil, \theta_{j+1}].$$

Let (s, e) denote an interval considered at some iteration of the WBS2 algorithm. By construction, the minimum length of the interval obtained by deterministic sampling is given by $\lfloor (e-s)/\widetilde{K} \rfloor$, where $\widetilde{K}$ satisfies $R_n \le \widetilde{K}(\widetilde{K}+1)/2$. Then, $\mathcal{R}_{s,e}$ drawn by the deterministic sampling contains at least one interval $(\ell_{m(j)}, r_{m(j)})$ satisfying $\ell_{m(j)} \in I_{L,j}$ and $r_{m(j)} \in I_{R,j}$ for any $\theta_j \in \Theta \cap (s, e)$ (if $\Theta \cap (s, e)$ is not empty), provided that $3\lfloor (e-s)/\widetilde{K} \rfloor \le 2 \min_{1 \le j \le q} \delta_j$. This condition in turn is met under (5). Then, it follows from the proof of Proposition B.1 of Cho and Kirch (2021) that there exists a permutation $\{\pi(1), \ldots, \pi(q)\}$ of $\{1, \ldots, q\}$ such that on $\mathcal{Z}_n$,

$$\max_{1 \le j \le q} (f'_{\pi(j)})^2 |k_{(j)} - \theta_{\pi(j)}| \le \rho_n = c_2 \zeta_n^2, \quad \text{and} \tag{F.2}$$

$$\exp(\mathcal{Y}_{(j)}) = \left|\mathcal{X}_{(j)}\right| \ge C_0 |f'_{\pi(j)}| \sqrt{\delta_{\pi(j)}} \ge C_1 n^{\varphi/2} \zeta_n \tag{F.3}$$

for $j = 1, \ldots, q$, by (F.1). From (F.2), the assertion in (i) follows readily. Also consequently, the intervals $(s_{(m)}, e_{(m)})$, $m = q+1, \ldots, n-1$ meet one of the followings:

(a) $(s_{(m)}, e_{(m)}) \cap \Theta = \emptyset$, or

(b) $(s_{(m)}, e_{(m)}) \cap \Theta = \{\theta_j\}$ and $(f'_j)^2 \min(\theta_j - s_{(m)}, e_{(m)} - \theta_j) \le \rho_n$, or

(c) $(s_{(m)}, e_{(m)}) \cap \Theta = \{\theta_j, \theta_{j+1}\}$ and $\max\{(f'_j)^2(\theta_j - s_{(m)}), (f'_{j+1})^2(e_{(m)} - \theta_{j+1})\} \le \rho_n$,

for some $j = 1, \ldots, q$. Under (a), from Assumption 2.1,

$$\exp(\mathcal{Y}_{(m)}) = |\mathcal{Z}_{s_{(m)}, k_{(m)}, e_{(m)}}| \le 2\zeta_n. \tag{F.4}$$

Under (b), supposing that $\theta_j \le k_{(m)}$, we obtain

$$\begin{aligned} \exp(\mathcal{Y}_{(m)}) &\le |\mathcal{F}_{s_{(m)}, k_{(m)}, e_{(m)}}| + |\mathcal{Z}_{s_{(m)}, k_{(m)}, e_{(m)}}| \\ &\le \sqrt{\frac{(k_{(m)} - s_{(m)})(e_{(m)} - k_{(m)})}{e_{(m)} - s_{(m)}}} \frac{(\theta_j - s_{(m)})|d_j|}{k_{(m)} - s_{(m)}} + 2\zeta_n \end{aligned}$$

$$\leq \sqrt{d_j^2 \min(\theta_j - s_{(m)}, e_{(m)} - \theta_j)} + 2\zeta_n \leq \sqrt{\rho_n} + 2\zeta_n \leq C_2\zeta_n \tag{F.5}$$

by Lemma F.1; the case when $\theta_j > k_{(m)}$ is handled analogously. Under (c), we obtain

$$\begin{aligned}\exp(\mathcal{Y}_{(m)}) &\leq \max\left\{|\mathcal{F}_{s_{(m)},\theta_j,e_{(m)}}|, |\mathcal{F}_{s_{(m)},\theta_{j+1},e_{(m)}}|\right\} + 2\zeta_n \\ &\leq \sqrt{d_j^2(\theta_j - s_{(m)})} + \sqrt{d_{j+1}^2(e_{(m)} - \theta_{j+1})} + 2\zeta_n \leq C_3\zeta_n \end{aligned} \tag{F.6}$$

where the first inequality follows from Lemma F.2 and the second inequality from Lemma F.1. From (F.3) and (F.4)–(F.6), and also that $\mathcal{X}_{(1)} \leq C_4\sqrt{n}$ due to $f_j' = O(1)$, we conclude that

$$\begin{aligned}&\mathcal{Y}_{(m)} = \gamma_m \log(n)(1 + o(1)) = \gamma_m \log(n)(1 + o(1)) + \log(\zeta_n) \quad \text{for} \quad m = 1, \ldots, q, \\ &\mathcal{Y}_{(m)} \leq \kappa_m \log(\zeta_n)(1 + o(1)) \quad \text{for} \quad m = q + 1, \ldots, P,\end{aligned}$$

where $\{\gamma_m\}$ and $\{\kappa_m\}$ meet the conditions in (ii).

F.2 Proof of the results in Section 3

We adopt the following notations throughout the proof: For a fixed integer $r \geq 1$ and an arbitrary set $\mathcal{A} = \{k_1, \cdots, k_m\} \subset \{1, \ldots, n\}$ satisfying $\min_{0\leq j\leq m}(k_{j+1} - k_j) \geq r + 1$ (with $k_0 = 0$ and $k_{m+1} = n$), we define $\mathbf{X} = \mathbf{X}(\mathcal{A}, r) = [\mathbf{L} : \mathbf{R}]$ and $\mathbf{Y}$ as in (8). Also we set $\mathbf{X}_{(j)} = [\mathbf{L}_{(j)} : \mathbf{1}]$ for each $j = 0, \ldots, m$, where $\mathbf{L}_{(j)}$ has $\mathbf{x}_t = (X_t, \ldots, X_{t-r+1})^\top$, $k_j \leq t \leq k_{j+1} - 1$ as its rows. Sub-vectors of $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$ corresponding to $k_j \leq t \leq k_{j+1} - 1$ are denoted by $\mathbf{Y}_{(j)}$ and $\boldsymbol{\varepsilon}_{(j)}$, respectively. When $r = 0$, we have $\mathbf{X} = \mathbf{R}$ and $\mathbf{X}_{(j)} = \mathbf{R}_{(j)}$,
Besides, we denote the (approximate) linear regression representation of (6) with the true change point locations θ_j and AR order p by

$$\mathbf{Y} = \mathbf{L}^\circ\boldsymbol{\alpha}^\circ + \boldsymbol{\nu}^\circ + \boldsymbol{\varepsilon} = \begin{bmatrix} \underbrace{\mathbf{L}^\circ}_{n\times p} & \underbrace{\mathbf{R}^\circ}_{n\times(q+1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^\circ \\ \boldsymbol{\mu}^\circ \end{bmatrix} + (\boldsymbol{\nu}^\circ - \mathbf{R}^\circ\boldsymbol{\mu}^\circ) + \boldsymbol{\varepsilon}, \tag{F.7}$$

where $\boldsymbol{\nu}^\circ = ((1 - a(B))f_t,\ 1 \leq t \leq n)^\top$. Correspondingly, $\mathbf{X}^\circ$ denotes an $n \times (p+q+1)$-matrix with its rows given by

$$\mathbf{x}_t = (X_{t-1}, \ldots, X_{t-p}, \mathbb{I}_{1\leq t\leq\theta_1}, \ldots, \mathbb{I}_{\theta_{q+1}\leq t\leq n})^\top$$

for $1 \leq t \leq n$, whereby $\mathbf{X}^\circ \equiv \mathbf{X}(\Theta, p)$. When $p = 0$, the matrix $\mathbf{L}^\circ$ is empty.

F.2.1 Preliminaries

The following results are frequently used throughout the proof.

Proposition F.3. Suppose that $p \geq 0$ and $r \in \{\max(p,1), \ldots, p_{\max}\}$ with $p_{\max} \geq \max(p,1)$ fixed. Also, let $\mathcal{A} = \{k_1, \ldots, k_m\}$ as an arbitrary subset of $\widehat{\Theta}_M$. With such $\mathcal{A}$, define $\mathbf{X} = \mathbf{X}(\mathcal{A}, r) = [\mathbf{L} : \mathbf{R}]$ as in (8), and also $\mathbf{X}_{(j)}$, $\mathbf{L}_{(j)}$, $\mathbf{R}_{(j)}$ and $\boldsymbol{\varepsilon}_{(j)}$, correspondingly, and let $N_j = k_{j+1} - k_j$. Then, under Assumption 3.1 (i)–(iii) and Assumption 3.2, we have the followings hold almost surely for all $j = 0, \ldots, m$ and $\mathcal{A} \subset \widehat{\Theta}_M$:

$$\mathrm{tr}(\mathbf{L}^\top \mathbf{L}) = O(n), \quad \mathrm{tr}(\mathbf{L}_{(j)}^\top \mathbf{L}_{(j)}) = O(N_j), \tag{F.8}$$

$$\liminf_{n\to\infty} n^{-1}\lambda_{\min}(\mathbf{L}^\top \mathbf{L}) > 0, \quad \liminf_{n\to\infty} N_j^{-1}\lambda_{\min}(\mathbf{L}_{(j)}^\top \mathbf{L}_{(j)}) > 0, \tag{F.9}$$

$$\mathrm{tr}(\mathbf{X}^\top \mathbf{X}) = O(n), \quad \liminf_{n\to\infty} n^{-1}\lambda_{\min}(\mathbf{X}^\top \mathbf{X}) > 0,$$

$$\mathrm{tr}(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)}) = O(N_j), \quad \liminf_{n\to\infty} N_j^{-1}\lambda_{\min}(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)}) > 0, \tag{F.10}$$

$$(\mathbf{L}^\top \mathbf{L})^{-1}\mathbf{L}^\top \boldsymbol{\varepsilon} = O\left(\sqrt{\frac{\log(n)}{n}}\right), \quad (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon} = O\left(\sqrt{\frac{\log(n)}{n}}\right),$$

$$(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)})^{-1}\mathbf{X}_{(j)}^\top \boldsymbol{\varepsilon}_{(j)} = O\left(\sqrt{\frac{\log(n)}{N_j}}\right). \tag{F.11}$$

Proof. The results in (F.8)–(F.9) follow from Theorem 3 (ii) of Lai and Wei (1983) and the finiteness of $\widehat{\Theta}_M$. By Corollary 2 of Lai and Wei (1982a), (F.10) follow from that $\mathsf{tr}(\mathbf{R}^\top \mathbf{R}) = n$ and $\mathbf{R}_{(j)}^\top \mathbf{R}_{(j)} = N_j$. By Lemma 1 of Lai and Wei (1982b), we have

$$\left\|(\mathbf{L}^\top \mathbf{L})^{-1/2}\mathbf{L}^\top \boldsymbol{\varepsilon}\right\| = O\left(\sqrt{\log(\lambda_{\max}(\mathbf{L}^\top \mathbf{L}))}\right) = O(\sqrt{\log(n)}) \quad \text{a.s.},$$

$$\left\|(\mathbf{X}^\top \mathbf{X})^{-1/2}\mathbf{X}^\top \boldsymbol{\varepsilon}\right\| = O\left(\sqrt{\log(\lambda_{\max}(\mathbf{X}^\top \mathbf{X}))}\right) = O(\sqrt{\log(n)}) \quad \text{a.s.},$$

$$\left\|(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)})^{-1/2}\mathbf{X}_{(j)}^\top \boldsymbol{\varepsilon}_{(j)}\right\| = O\left(\sqrt{\log(\lambda_{\max}(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)}))}\right) = O(\sqrt{\log(n)}) \quad \text{a.s.}$$

which, together with (F.8) and (F.10), leads to (F.11). □

Lemma F.4 (Lemma 3.1.2 of Csörgő and Horváth (1997))**.** For any $\mathbf{X} = [\mathbf{L} : \mathbf{R}]$, the OLS estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} = (\widehat{\boldsymbol{\alpha}}^\top, \widehat{\boldsymbol{\mu}}^\top)^\top$ satisfies $\widehat{\boldsymbol{\alpha}} = (\mathbf{L}^\top \mathbf{L})^{-1}\mathbf{L}^\top(\mathbf{Y} - \mathbf{R}\widehat{\boldsymbol{\mu}})$ and $\widehat{\boldsymbol{\mu}} = \{\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R}\}^{-1}\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{Y}$.

Lemma F.5. For some $\mathbf{R} = \mathbf{R}(\mathcal{A})$ constructed with a set $\mathcal{A} = \{k_1, \ldots, k_m\} \subset \{1, \ldots, n\}$ with $k_1 < \ldots < k_m$, we denote by $\mathbf{R}_{-j}$, for any $1 \leq j \leq m$, an $n \times m$-matrix formed by merging the j-th and the $(j+1)$-th columns of $\mathbf{R}$ via summing them up, while the rest of the columns of $\mathbf{R}$ are unchanged. Then,

$$\|(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{R}_{-j}})\mathbf{U}\|^2 - \|(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{R}})\mathbf{U}\|^2 = |\mathcal{C}_{k_{j-1},k_j,k_{j+1}}(\mathbf{U})|^2 \tag{F.12}$$

for any $\mathbf{U} = (U_1, \ldots, U_{n-(m+1)r})^\top$, where

$$\mathcal{C}_{k_{j-1},k_j,k_{j+1}}(\mathbf{U}) := \sqrt{\frac{(k_{j+1}-k_j)(k_j-k_{j-1})}{k_{j+1}-k_{j-1}}} \times \left(\frac{1}{k_j - k_{j-1}} \sum_{t=k_{j-1}+1}^{k_j} U_t - \frac{1}{k_{j+1}-k_j} \sum_{t=k_j+1}^{k_{j+1}} U_t \right).$$

Proof. Denote the $(j+1)$-th column of $\mathbf{R}$ by $\mathbf{R}_j$. Then, by simple calculations, we have

$$\|(\mathbf{I} - \mathbf{\Pi}_{\mathbf{R}})\mathbf{U}\|^2 = \mathbf{U}^\top(\mathbf{I} - \mathbf{\Pi}_{\mathbf{R}_{-j}})\mathbf{U} - \frac{(\mathbf{U}^\top(\mathbf{I} - \mathbf{\Pi}_{\mathbf{R}_{-j}})\mathbf{R}_j)^2}{\mathbf{R}_j^\top(\mathbf{I} - \mathbf{\Pi}_{\mathbf{R}_{-j}})\mathbf{R}_j}.$$

Also by construction,

$$\mathbf{R}_{-j}^\top \mathbf{R}_j = (\underbrace{0, \ldots, 0}_{j-1}, k_{j+1} - k_j, 0, \ldots, 0)^\top,$$
$$(\mathbf{R}_{-j}^\top \mathbf{R}_{-j})^{-1} = \operatorname{diag}\left(\frac{1}{k_1}, \ldots, \frac{1}{k_{j-1} - k_{j-2}}, \frac{1}{k_{j+1} - k_{j-1}}, \frac{1}{k_{j+2} - k_{j+1}}, \ldots, \frac{1}{n - k_m} \right).$$

Hence,

$$[\mathbf{R}_{-j}(\mathbf{R}_{-j}^\top \mathbf{R}_{-j})^{-1}\mathbf{R}_{-j}^\top \mathbf{R}_j]_i = \begin{cases} \frac{k_{j+1}-k_j}{k_{j+1}-k_{j-1}} & \text{for } k_{j-1}+1 \le i \le k_{j+1}, \\ 0 & \text{otherwise,} \end{cases}$$
$$[\mathbf{R}_j - \mathbf{R}_{-j}(\mathbf{R}_{-j}^\top \mathbf{R}_{-j})^{-1}\mathbf{R}_{-j}^\top \mathbf{R}_j]_i = \begin{cases} -\frac{k_{j+1}-k_j}{k_{j+1}-k_{j-1}} & \text{for } k_{j-1}+1 \le i \le k_j, \\ \frac{k_j-k_{j-1}}{k_{j+1}-k_{j-1}} & \text{for } k_j+1 \le i \le k_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbf{R}_j^\top(\mathbf{I} - \mathbf{\Pi}_{\mathbf{R}_{-j}})\mathbf{R}_j = \frac{(k_j - k_{j-1})(k_{j+1} - k_j)}{k_{j+1} - k_{j-1}},$$
$$\mathbf{U}^\top(\mathbf{I} - \mathbf{\Pi}_{\mathbf{R}_{-j}})\mathbf{R}_j = \frac{(k_j - k_{j-1})(k_{j+1} - k_j)}{k_{j+1} - k_{j-1}} \left(\frac{1}{k_{j+1} - k_j} \sum_{t=k_j+1}^{k_{j+1}} U_t - \frac{1}{k_j - k_{j-1}} \sum_{t=k_{j-1}+1}^{k_j} U_t \right),$$

which concludes the proof. □

F.2.2 Proof of Theorem 3.1

Throughout the proofs, $C_0, C_1, \ldots$ denote some positive constants. In what follows, we operate in $\mathcal{E}_n \cap \mathcal{M}_n$, and all big-$O$ notations imply that they hold a.s. due to Proposition F.3.
We briefly sketch the proof, which proceeds in four steps (i)–(iv) below. We first suppose

that Assumption 3.2 holds with $M = 1$, and also that p is known. Then, a single iteration of the gSa algorithm in Section A.2 boils down to choosing between $\widehat{\Theta}_0 = \emptyset$ and $\widehat{\Theta}_1$: If $\mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}_1, p) < \mathrm{SC}_0(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(p))$, we favour a change point model; if not, we conclude that there is no change point in the data. In (i), when $q = 0$, we show that $\mathbf{R}\widehat{\boldsymbol{\mu}} \approx \mathbf{1}\mu_0^\circ \approx \boldsymbol{\Pi}_\mathbf{1}(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})$ with $\mu_0^\circ = (1 - \sum_{i=1}^p a_i) f_0$ representing the time-invariant overall level, and therefore $\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \approx \|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2$ which leads to $\mathrm{SC}_0(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(p)) < \mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}_1, p)$ under Assumption 3.4. In (ii), when $q \geq 1$, we show that

$$\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \geq Cq \min_{1 \leq j \leq q} d_j^2 \delta_j \gg q\xi_n$$

for some fixed constant $C > 0$ and thus $\mathrm{SC}_0(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(p)) > \mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}_1, p)$, provided that $\widehat{\Theta}_1$ meets (11). In (iii), we show the consistency of the proposed order selection scheme. For the general case where $M > 1$, in (iv), we can repeatedly apply the above arguments for each call of Step 1 of the gSa algorithm: Under Assumption 3.2, when $l > l^*$, any $\widehat{\theta}_{l,j} \notin \widehat{\Theta}_{l^*}$ are spurious estimators and thus we have the gSa algorithm proceed to examine $\widehat{\Theta}_{l-1}$; when $l = l^*$, any $\widehat{\theta}_{l^*,j} \notin \widehat{\Theta}_{l^*-1}$ are detecting those change points undetected in $\widehat{\Theta}_{l^*-1}$ and thus the gSa algorithm returns $\widehat{\Theta}_{l^*}$.

As outlined above, in the following (i)–(iii), we only consider the case of $M = 1$ and consequently drop the subscript '1' from $\widehat{\Theta}_1$ and $\widehat{\theta}_{1,j}$ where there is no confusion.

For given $\widehat{\Theta}$, recall that $\mathbf{X} = \mathbf{X}(\widehat{\Theta}, p) = [\mathbf{L} : \mathbf{R}]$ and $N_j = \widehat{\theta}_{j+1} - \widehat{\theta}_j$. For $t = \theta_j + 1, \ldots, \theta_j + p$, we have

$$|[\boldsymbol{\nu}^\circ - \mathbf{R}^\circ \boldsymbol{\mu}^\circ]_t| \leq |d_j| \max_{1 \leq i \leq p} \left| \sum_{i'=i}^p a_{i'} \right| \leq |d_j|, \tag{F.13}$$

for all $1 \leq j \leq q$, while $[\boldsymbol{\nu}^\circ - \mathbf{R}^\circ \boldsymbol{\mu}^\circ]_t = 0$ elsewhere.

(i) When $q = 0$. We first note that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{L}\boldsymbol{\alpha}^\circ + \mu_0^\circ \mathbf{1} + \boldsymbol{\varepsilon})$$

such that by Proposition F.3, we have

$$\left\| \widehat{\boldsymbol{\beta}} - \underbrace{\begin{bmatrix} \boldsymbol{\alpha}^\circ \\ \mu_0^\circ \mathbf{1}_{\widehat{q}+1} \end{bmatrix}}_{\boldsymbol{\beta}^\circ(\widehat{q})} \right\| = \left\| (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \right\| = O\left(\sqrt{\frac{\log(n)}{n}} \right). \tag{F.14}$$

We decompose the residual sum of squares as

$$\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\varepsilon}\|^2 + \|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ(\widehat{q}))\|^2 - 2\boldsymbol{\varepsilon}^\top \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ(\widehat{q})) =: \|\boldsymbol{\varepsilon}\|^2 + \mathcal{R}_{11} + \mathcal{R}_{12}.$$

Invoking Proposition F.3 and (F.14),

$$\mathcal{R}_{11} \leq \|\mathbf{X}\|^2 \, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ(\widehat{q})\|^2 = O\left(\frac{n\log(n)}{n}\right) = O(\log(n)) \quad \text{a.s.},$$

and

$$|\mathcal{R}_{12}| \leq \|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\| \, \|\mathbf{X}^\top\mathbf{X}\| \, \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\circ(\widehat{q})\| = O\left(\sqrt{n\log(n)} \cdot \sqrt{\frac{\log(n)}{n}}\right) = O\left(\log(n)\right).$$

Putting together the bounds on $\mathcal{R}_{11}$–$\mathcal{R}_{12}$, we conclude that

$$\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O(\log(n)). \tag{F.15}$$

Next, note that

$$\begin{aligned}\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 &= \|\boldsymbol{\varepsilon}\|^2 - \boldsymbol{\varepsilon}^\top\boldsymbol{\Pi}_\mathbf{1}\boldsymbol{\varepsilon} + \|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\|^2 - 2\boldsymbol{\varepsilon}^\top(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) \\ &=: \|\boldsymbol{\varepsilon}\|^2 + \mathcal{R}_{21} + \mathcal{R}_{22} + \mathcal{R}_{23}.\end{aligned}$$

By the arguments similar to those adopted in Proposition F.3 and Lemma 1 of Lai and Wei (1982a), we have $|\mathcal{R}_{21}| = O(\log(n))$. Also, by Proposition F.3 and (F.14), $\mathcal{R}_{22} \leq \|\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\|^2 = O(\log(n))$. Next,

$$|\mathcal{R}_{23}| \leq 2\left|\boldsymbol{\varepsilon}^\top\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\right| + 2\left|\boldsymbol{\varepsilon}^\top\boldsymbol{\Pi}_\mathbf{1}\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\right|$$

where the first term is bounded by

$$2\|(\mathbf{L}^\top\mathbf{L})^{-1}\mathbf{L}^\top\boldsymbol{\varepsilon}\| \, \|\mathbf{L}^\top\mathbf{L}\| \, \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O(\log(n))$$

due to Proposition F.3 and Lemma 1 of Lai and Wei (1982a), and the second term is bounded by the bound on the first term and $\mathcal{R}_{21}$ as $O(\log(n))$. Therefore,

$$\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O(\log(n)). \tag{F.16}$$

Combining (F.15) and (F.16) with Assumption 3.1 (ii)–(iii), and noting that $\log(1+x) \leq x$ for all $x \geq 0$,

$$\begin{aligned}&\mathrm{SC}_0(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(p)) - \mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}, p) \\ &= \frac{n}{2}\log\left(1 + \frac{\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}\right) - \widehat{q}\xi_n = O(\log(n)) - \widehat{q}\xi_n < 0\end{aligned}$$

for n large enough, due to Assumption 3.4.

(ii) When $q \geq 1$. Recall that in $\mathcal{M}_n$, we have $\widehat{q} = q$. Below we use that by Proposition F.3,

$$\mathsf{tr}(\mathbf{L}^\top \mathbf{R}) = O(n) \quad \text{and} \quad [\mathbf{L}_{(j)}^\top \mathbf{1}]_i = O(N_j) \text{ for } i = 1, \ldots, p,\ j = 0, \ldots, q, \tag{F.17}$$

where $\bar{f} = \max_{0 \leq j \leq q} |f_{\theta_j + 1}|$. We first establish the consistency of $\widehat{\boldsymbol{\mu}}$ in estimating $\boldsymbol{\mu}^\circ$. Applying Lemma F.4, we write

$$\begin{aligned}\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\circ =& (\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R})^{-1}\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})(\boldsymbol{\nu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ) + \\ & (\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R})^{-1}\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\boldsymbol{\varepsilon} =: \mathcal{R}_{31} + \mathcal{R}_{32}.\end{aligned}$$

Since $(\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R})^{-1}$ is a sub-matrix of $(\mathbf{X}^\top\mathbf{X})^{-1}$, we have $\lambda_{\max}((\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R})^{-1}) \leq (\lambda_{\min}(\mathbf{X}^\top\mathbf{X}))^{-1}$ (Horn and Johnson, 1985, Theorem 4.2.2)) and thus $\liminf_{n\to\infty} n^{-1}\lambda_{\min}(\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R}) > 0$ by Proposition F.3. Also, since $\mathsf{tr}(\mathbf{R}^\top(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{L}})\mathbf{R}) \leq n$ trivially, we obtain $|\mathcal{R}_{32}| = O\left(\sqrt{\log(n)/n}\right)$ adopting the same arguments used in the proof of (F.11). Next, by (F.13) and since

$$[\mathbf{R}^\circ\boldsymbol{\mu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ]_t = \begin{cases} d_j & \text{for } \theta_j + 1 \leq t \leq \widehat{\theta}_j, \\ -d_j & \text{for } \widehat{\theta}_j + 1 \leq t \leq \theta_j, \end{cases} \quad \text{for } j = 1, \ldots, q$$

while $[\mathbf{R}^\circ\boldsymbol{\mu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ]_t = 0$ otherwise, we obtain

$$\|\boldsymbol{\nu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ\|^2 \leq 2\|\boldsymbol{\nu}^\circ - \mathbf{R}^\circ\boldsymbol{\mu}^\circ\|^2 + 2\|\mathbf{R}^\circ\boldsymbol{\mu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ\|^2 \leq 2\sum_{j=1}^{q} d_j^2 \cdot (p + d_j^{-2}\rho_n) = O(q\rho_n) \tag{F.18}$$

and therefore $|\mathcal{R}_{31}|^2 = O(q\rho_n/n)$. Putting together the bounds on $\mathcal{R}_{31}$–$\mathcal{R}_{32}$, we obtain

$$|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\circ| = O\left(\sqrt{\frac{\log(n) \vee q\rho_n}{n}}\right). \tag{F.19}$$

Also, note that by Lemma F.4,

$$\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ = (\mathbf{L}^\top\mathbf{L})^{-1}\mathbf{L}^\top\left\{\boldsymbol{\varepsilon} + (\boldsymbol{\nu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ) + \mathbf{R}(\boldsymbol{\mu}^\circ - \widehat{\boldsymbol{\mu}})\right\}.$$

Adopting Proposition F.3, (F.17), (F.18) and (F.19), we have

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O\left(\sqrt{\frac{\log(n) \vee q\rho_n}{n}}\right). \tag{F.20}$$

Next, we consider

$$\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) + (\mathbf{R}\widehat{\boldsymbol{\mu}} - \boldsymbol{\nu}^\circ) - \boldsymbol{\varepsilon}\|^2$$

$$
\begin{aligned}
=&\|\boldsymbol{\varepsilon}\|^2 + \|\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\|^2 + \|\mathbf{R}\widehat{\boldsymbol{\mu}} - \boldsymbol{\nu}^\circ\|^2 + 2(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)^\top \mathbf{L}^\top(\mathbf{R}\widehat{\boldsymbol{\mu}} - \boldsymbol{\nu}^\circ) \\
&- 2\boldsymbol{\varepsilon}^\top \mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) - 2\boldsymbol{\varepsilon}^\top(\mathbf{R}\widehat{\boldsymbol{\mu}} - \boldsymbol{\nu}^\circ) =: \|\boldsymbol{\varepsilon}\|^2 + \mathcal{R}_{41} + \mathcal{R}_{42} + \mathcal{R}_{43} + \mathcal{R}_{44} + \mathcal{R}_{45}.
\end{aligned}
$$

By Proposition F.3 and (F.20),

$$
\mathcal{R}_{41} = O\left(n \cdot \frac{\log(n) \vee q\rho_n}{n}\right) = O\left(\log(n) \vee q\rho_n\right).
$$

Also, due to (F.18) and (F.19),

$$
\mathcal{R}_{42} \leq 2\|\mathbf{R}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\circ)\|^2 + 2\|\mathbf{R}\boldsymbol{\mu}^\circ - \boldsymbol{\nu}^\circ\|^2 = O\left(\log(n) \vee q\rho_n\right) \tag{F.21}
$$

and we also obtain $\mathcal{R}_{43} = O\left(\log(n) \vee q\rho_n\right)$. By Proposition F.3 and (F.20),

$$
\mathcal{R}_{44} \leq \|(\mathbf{L}^\top\mathbf{L})^{-1}\mathbf{L}^\top\boldsymbol{\varepsilon}\| \, \|\mathbf{L}^\top\mathbf{L}\| \, \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O\left(\sqrt{\log(n)(\log(n) \vee q\rho_n)}\right) = O\left(\log(n) \vee \sqrt{q}\rho_n\right),
$$

while with (F.13), (F.19), Assumption 3.1 and Chebyshev's inequality,

$$
\begin{aligned}
|\mathcal{R}_{45}| &\leq 2|\boldsymbol{\varepsilon}^\top\mathbf{R}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\circ)| + 2|\boldsymbol{\varepsilon}^\top(\mathbf{R}\boldsymbol{\mu}^\circ - \mathbf{R}^\circ\boldsymbol{\mu}^\circ)| + 2|\boldsymbol{\varepsilon}^\top(\mathbf{R}^\circ\boldsymbol{\mu}^\circ - \boldsymbol{\nu}^\circ)| \\
&= O\left(\sqrt{n\log(n)} \cdot \sqrt{\frac{\log(n) \vee q\rho_n}{n}} + \sum_{j=1}^{q} |d_j| \cdot \sqrt{d_j^{-2}\rho_n}\omega_n + p\sqrt{\sum_{j=1}^{q} |d_j|^2}\right) \\
&= O\left(\log(n) \vee q(\rho_n \vee \omega_n^2)\right)
\end{aligned}
$$

on $\mathcal{E}_n$. Combining the bounds on $\mathcal{R}_{41}$–$\mathcal{R}_{45}$, we obtain

$$
\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O\left(\log(n) \vee q\left(\rho_n \vee \omega_n^2\right)\right). \tag{F.22}
$$

Next, note that

$$
\begin{aligned}
&\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \left(\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{R})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2\right) \\
&+ \left(\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{R})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2\right) =: \mathcal{R}_{51} + \mathcal{R}_{52}.
\end{aligned}
$$

Repeatedly invoking Lemma F.5, we have

$$
\begin{aligned}
\mathcal{R}_{51} =&\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{R}_{-\mathcal{I}_1}})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 + \sum_{j \in \mathcal{I}_1} \left|\mathcal{C}_{\widehat{\theta}_{j-1}, \widehat{\theta}_j, \widehat{\theta}_{j+1}}(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\right|^2 \\
&\geq \left\lceil \frac{q}{2} \right\rceil \min_{1 \leq j \leq q} \left|\mathcal{C}_{\widehat{\theta}_{j-1}, \widehat{\theta}_j, \widehat{\theta}_{j+1}}(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\right|^2
\end{aligned}
$$

where $\mathbf{R}_{-\mathcal{I}_1}$ denotes a matrix constructed by merging the j-th and the $(j+1)$-th columns of $\mathbf{R}$ via summing them up for all $j \in \mathcal{I}_1$, while the rest of the columns of $\mathbf{R}$ are unchanged, with

$\mathcal{I}_1$ denoting a subset of $\{1,\ldots,q\}$ consisting of all the odd indices. For notational simplicity, let $\mathcal{C}_j(\cdot) = \mathcal{C}_{\widehat{\theta}_{j-1},\widehat{\theta}_j,\widehat{\theta}_{j+1}}(\cdot)$ where there is no confusion. Note that

$$\mathcal{C}_j(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}}) = \mathcal{C}_j(\mathbf{R}^\circ\boldsymbol{\mu}^\circ) + \mathcal{C}_j(\boldsymbol{\nu}^\circ - \mathbf{R}^\circ\boldsymbol{\mu}^\circ) + \mathcal{C}_j(\boldsymbol{\varepsilon}) + \mathcal{C}_j(\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)).$$

Without loss of generality, suppose that $\widehat{\theta}_j \le \theta_j$. Analogous arguments apply when $\widehat{\theta}_j > \theta_j$. By Lemma F.1,

$$\begin{aligned}\mathcal{C}_j(\mathbf{R}^\circ\boldsymbol{\mu}^\circ) = &-\sqrt{\frac{N_{j-1}N_j}{N_{j-1}+N_j}}\left\{\frac{(N_j + \widehat{\theta}_j - \theta_j)d_j}{N_j} + \frac{(\widehat{\theta}_{j+1} - \theta_{j+1})_+ d_{j+1}}{N_j}\right.\\ &\left. + \frac{(\theta_{j-1} - \widehat{\theta}_{j-1})_+ d_{j-1}}{N_{j-1}}\right\} =: \mathcal{R}_{61} + \mathcal{R}_{62} + \mathcal{R}_{63}.\end{aligned}$$

Under Assumptions 3.2, 3.3 and 3.4, $\min(N_{j-1}, N_j)^{-1} d_j^2 |\widehat{\theta}_j - \theta_j| = O(\delta_j^{-1}\rho_n) = o(1)$ (due to $D_n^{-1}\rho_n \to 0$ as $n \to \infty$) and thus

$$|\mathcal{R}_{61}| = |d_j|\sqrt{\frac{N_{j-1}N_j}{N_{j-1}+N_j}}(1+o(1)) \ge |d_j|\sqrt{\frac{\min(N_{j-1},N_j)}{2}}(1+o(1)) \ge \sqrt{\frac{d_j^2\delta_j}{2}}(1+o(1)),$$

while

$$|\mathcal{R}_{62}| \le \frac{d_{j+1}^2(\widehat{\theta}_{j+1} - \theta_{j+1})}{\sqrt{d_{j+1}^2(\widehat{\theta}_{j+1} - \widehat{\theta}_j - p)}} \le \frac{\rho_n}{\sqrt{D_n}}(1+o(1)) = o(\sqrt{\rho_n})$$

and $\mathcal{R}_{63}$ is similarly bounded. Therefore, we conclude

$$\min_{1\le j\le q} |\mathcal{C}_j(\mathbf{R}^\circ\boldsymbol{\mu}^\circ)| \ge \sqrt{\frac{D_n}{2}}(1+o(1)). \tag{F.23}$$

Similarly, by (F.13) and Assumption 3.2, we derive

$$|\mathcal{C}_j(\boldsymbol{\nu}^\circ - \mathbf{R}^\circ\boldsymbol{\mu}^\circ)| \le p\sqrt{\frac{N_{j-1}N_j}{N_{j-1}+N_j}}\left\{\frac{|d_j| + |d_{j+1}|}{N_j} + \frac{|d_{j-1}|}{N_{j-1}}\right\} = o(1). \tag{F.24}$$

Invoking Assumption 3.1 (iv), it is easily seen that on $\mathcal{E}_n$,

$$|\mathcal{C}_j(\boldsymbol{\varepsilon})| \le 2\omega_n. \tag{F.25}$$

Finally, by (F.17) and (F.20),

$$|\mathcal{C}_j(\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ))| = \sqrt{\frac{N_{j-1}N_j}{N_{j-1}+N_j}}\left|\frac{1}{N_{j-1}}\mathbf{1}^\top\mathbf{L}_{(j-1)}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) - \frac{1}{N_j}\mathbf{1}^\top\mathbf{L}_{(j)}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\right|$$

$$= O\left(\sqrt{\min(N_{j-1}, N_j)} \cdot \sqrt{\frac{\log(n) \vee q\rho_n}{n}}\right) = O\left(\sqrt{\log(n) \vee q\rho_n}\right). \quad \text{(F.26)}$$

By (F.23)–(F.26), under Assumption 3.3, there exists some constant $C_0 > 0$ satisfying

$$\mathcal{R}_{51} \geq C_0 q D_n \qquad \text{for } n \text{ large enough.} \quad \text{(F.27)}$$

Next, we note that

$$\begin{aligned}
&\|(\mathbf{I} - \mathbf{\Pi_R})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 = \|\boldsymbol{\varepsilon}\|^2 - \boldsymbol{\varepsilon}^\top \mathbf{\Pi_R} \boldsymbol{\varepsilon} + \|(\mathbf{I} - \mathbf{\Pi_R})\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\|^2 + \|(\mathbf{I} - \mathbf{\Pi_R})\boldsymbol{\nu}^\circ\|^2 \\
&+ 2(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)^\top \mathbf{L}^\top (\mathbf{I} - \mathbf{\Pi_R})\boldsymbol{\nu}^\circ - 2\boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{\Pi_R})\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) - 2\boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{\Pi_R})\boldsymbol{\nu}^\circ \\
&=: \|\boldsymbol{\varepsilon}\|^2 - \mathcal{R}_{71} + \mathcal{R}_{72} + \mathcal{R}_{73} + \mathcal{R}_{74} + \mathcal{R}_{75} + \mathcal{R}_{76}.
\end{aligned}$$

First, by Assumption 3.1 (iv), $\mathcal{R}_{71} = O(\sum_{j=0}^{q} N_j \omega_n^2 \cdot N_j^{-1}) = O(q\omega_n^2)$ on $\mathcal{E}_n$. Also, from Proposition F.3 and (F.20), $\mathcal{R}_{72} \leq \|\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\|^2 = O(\log(n) \vee q\rho_n)$. In addition,

$$\mathcal{R}_{73} \leq 2\|\boldsymbol{\nu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ\|^2 + 2\|\mathbf{R}(\boldsymbol{\mu}^\circ - (\mathbf{R}^\top \mathbf{R})^{-1}\mathbf{R}^\top \nu^\circ)\|^2$$

where the first term is $O(q\rho_n)$ as in (F.18). From (F.13) and the definition of $\mathbf{R}$ and $\mathbf{R}^\circ$,

$$\boldsymbol{\mu}^\circ - (\mathbf{R}^\top \mathbf{R})^{-1}\mathbf{R}^\top \mathbf{R}^\circ \boldsymbol{\mu}^\circ = \begin{bmatrix} \frac{-(\widehat{\theta}_1 - \theta_1)_+ d_1}{\widehat{\theta}_1} \\ \frac{(\theta_1 - \widehat{\theta}_1)_+ d_1 - (\widehat{\theta}_2 - \theta_2)_+ d_2}{\widehat{\theta}_2 - \widehat{\theta}_1} \\ \vdots \\ \frac{(\theta_q - \widehat{\theta}_q)_+ d_q}{n - \widehat{\theta}_q} \end{bmatrix}, \quad \text{(F.28)}$$

$$\left|[(\mathbf{R}^\top \mathbf{R})^{-1}\mathbf{R}^\top (\mathbf{R}^\circ \boldsymbol{\mu}^\circ - \boldsymbol{\nu}^\circ)]_j\right| \leq \frac{p(|d_{j-1}| + |d_j|)}{\widehat{\theta}_j - \widehat{\theta}_{j-1}} \quad \text{(F.29)}$$

(recall that $\widehat{\theta}_0 = \theta_0 = 0$ and $\widehat{\theta}_{q+1} = \theta_{q+1} = n$) such that by Assumptions 3.2 and 3.3, we obtain

$$\|\mathbf{R}(\boldsymbol{\mu}^\circ - (\mathbf{R}^\top \mathbf{R})^{-1}\mathbf{R}^\top \boldsymbol{\nu}^\circ)\|^2 \leq C_1 \sum_{j=1}^{q} d_j^2 \cdot \frac{(d_j^{-2}\rho_n)^2 + p^2}{\widehat{\theta}_{j+1} - \widehat{\theta}_j} = o(q\rho_n)$$

for some constant $C_1 > 0$, hence $\mathcal{R}_{73} = O(q\rho_n)$. The bounds on $\mathcal{R}_{72}$ and $\mathcal{R}_{73}$ imply the $O(\log(n) \vee q\rho_n)$ bound on $\mathcal{R}_{74}$. Next, since $\lambda_{\max}((\mathbf{L}^\top(\mathbf{I} - \mathbf{\Pi_R})\mathbf{L})^{-1}) \leq \lambda_{\min}^{-1}(\mathbf{X}^\top \mathbf{X})$, we have

$$|\mathcal{R}_{75}| \leq \|(\mathbf{L}^\top(\mathbf{I} - \mathbf{\Pi_R})\mathbf{L})^{-1}\mathbf{L}^\top(\mathbf{I} - \mathbf{\Pi_R})\boldsymbol{\varepsilon}\| \, \|\mathbf{L}^\top(\mathbf{I} - \mathbf{\Pi_R})\mathbf{L}\| \, \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O\left(\log(n) \vee q\rho_n\right)$$

from Lemma 1 of Lai and Wei (1982a), Proposition F.3 and (F.20). Finally,

$$|\mathcal{R}_{76}| \leq 2|\boldsymbol{\varepsilon}^\top(\boldsymbol{\nu}^\circ - \mathbf{R}\boldsymbol{\mu}^\circ)| + 2|\boldsymbol{\varepsilon}^\top\mathbf{R}(\boldsymbol{\mu}^\circ - (\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\boldsymbol{\nu})|$$

where using the arguments involved in bounding $\mathcal{R}_{45}$, we have the first term bounded by $O(q(\rho_n \vee \omega_n^2))$, while the second term is bounded as

$$O\left(\sum_{j=1}^{q}\sqrt{N_j}\omega_n \cdot \frac{d_j^{-2}\rho_n \cdot |d_j|}{N_j}\right) = O\left(\sum_{j=1}^{q}\frac{\omega_n\rho_n}{\sqrt{D_n}}\right) = O(q\rho_n),$$

on $\mathcal{E}_n$, recalling (F.28)–(F.29) and by Assumptions 3.1 (iv), 3.2 and 3.3. Therefore, $\mathcal{R}_{76} = O(q(\rho_n \vee \omega_n^2))$. Collecting the bounds on $\mathcal{R}_{71}$–$\mathcal{R}_{76}$, we obtain

$$\|(\mathbf{I} - \mathbf{\Pi}_\mathbf{R})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O\left(\log(n) \vee q(\rho_n \vee \omega_n^2)\right). \tag{F.30}$$

From (F.22), (F.27) and (F.30),

$$\|(\mathbf{I} - \mathbf{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}^\circ\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \geq C_0 q D_n + O\left(\log(n) \vee q(\rho_n \vee \omega_n^2)\right). \tag{F.31}$$

Note that

$$\begin{aligned}
&\mathrm{SC}_0(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(p)) - \mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}, p) \\
&= \frac{n}{2}\log\left(1 + \frac{\|(\mathbf{I} - \mathbf{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}^\circ\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}\right) - q\xi_n =: \frac{n}{2}\log(1 + \mathcal{R}_8) - q\xi_n.
\end{aligned} \tag{F.32}$$

When $\mathcal{R}_8 \geq 1$, we have the RHS of (F.32) trivially bounded away from zero by Assumption 3.4. When $\mathcal{R}_8 < 1$, note that for $g(x) = \log(x)/(x-1)$, since $\lim_{x\downarrow 1} g(x) \to 1$ and from its continuity, there exists a constant $C_2 > 0$ such that $\inf_{1\leq x<2} g(x) \geq C_2$. Therefore,

$$\frac{n}{2}\log(1 + \mathcal{R}_8) - q\xi_n \geq C_3 q D_n + O\left(\log(n) \vee q(\rho_n \vee \omega_n^2)\right) - q\xi_n > 0,$$

invoking Assumption 3.1 (ii)–(iii), (F.22) and (F.31) for some $C_3 > 0$.

(iii) Order selection consistency. Thus far, we have assumed that the AR order p is known. We show next that for n large enough, the order p is consistently estimated by $\widehat{p}$ obtained as in (9). Recall the notation $\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, r) = (\widehat{\boldsymbol{\alpha}}^\top(r), \widehat{\boldsymbol{\mu}}^\top(\widehat{\Theta}))^\top$. Firstly, suppose that $r > p$ while $r \leq p_{\max}$. Then, by (F.14) when $q = 0$ or by (F.19) and (F.20) when $q \geq 1$ (here, q coincides

with the cardinality of $\widehat{\Theta}$), we have

$$\|\widehat{\boldsymbol{\alpha}}(r) - \boldsymbol{\alpha}^\circ(r)\| = O\left(\sqrt{\frac{\log(n) \vee q\rho_n}{n}}\right) \quad \text{with} \quad \boldsymbol{\alpha}^\circ(r) = (\boldsymbol{\alpha}^{\circ\top}, \underbrace{0, \ldots, 0}_{r-p})^\top$$

whether there are changes or not, see the steps leading to (F.20). Then, the arguments similar to those adopted in showing (F.15) or (F.22) establish that

$$\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, r)\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, r)\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O\left(\log(n) \vee q(\rho_n \vee \omega_n^2)\right)$$

and therefore, we have

$$\begin{aligned}
&\mathrm{SC}\left(\{X_t\}_{t=1}^n, \widehat{\Theta}, r\right) - \mathrm{SC}\left(\{X_t\}_{t=1}^n, \widehat{\Theta}, p\right) \\
=& -\frac{n}{2}\log\left(1 + \frac{\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, p)\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, p)\|^2 - \|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, r)\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, r)\|^2}{\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, r)\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, r)\|^2}\right) + (r-p)\xi_n \\
=& O\left(\log(n) \vee q(\rho_n \vee \omega_n^2)\right) + (r-p)\xi_n > 0
\end{aligned}$$

for n large enough, by Assumption 3.4.

Next, consider $r < p$. For notational convenience, let $\boldsymbol{\Pi}(r) = \boldsymbol{\Pi}_{\mathbf{X}(\widehat{\Theta}, r)}$, and the sub-matrix of $\mathbf{X}(\widehat{\Theta}, p)$ containing its columns corresponding to the i-th lags for $i = r+1, \ldots, p$ by $\mathbf{X}(p|r)$. Then, $[\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{X}(p|r)]^{-1}$ is a sub-matrix of $(\mathbf{X}(\widehat{\Theta}, p)^\top \mathbf{X}(\widehat{\Theta}, p))^{-1}$ and thus by Theorem 4.2.2 of Horn and Johnson (1985) and Proposition F.3, we have

$$\begin{aligned}
&\lambda_{\max}\left(\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{X}(p|r)\right) \leq \lambda_{\max}\left(\mathbf{X}(\widehat{\Theta}, p)^\top \mathbf{X}(\widehat{\Theta}, p)\right) \\
&\qquad \leq \mathsf{tr}\left(\mathbf{X}(\widehat{\Theta}, p)^\top \mathbf{X}(\widehat{\Theta}, p)\right) = O(n) \quad \text{and similarly,} \qquad \text{(F.33)} \\
&\lambda_{\min}\left(\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{X}(p|r)\right) \geq \lambda_{\min}\left(\mathbf{X}(\widehat{\Theta}, p)^\top \mathbf{X}(\widehat{\Theta}, p)\right) \quad \text{and thus} \\
&\liminf_{n\to\infty} n^{-1}\lambda_{\min}\left(\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{X}(p|r)\right) > 0. \qquad \text{(F.34)}
\end{aligned}$$

It then follows that

$$\begin{aligned}
&\left\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, r)\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, r)\right\|^2 - \left\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, p)\widehat{\boldsymbol{\beta}}(\widehat{\Theta}, p)\right\|^2 \\
=& \left\|\left[\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{X}(p|r)\right]^{-1/2}\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{Y}\right\|^2 \\
\geq& \lambda_{\min}\left(\mathbf{X}(p|r)^\top(\mathbf{I} - \boldsymbol{\Pi}(r))\mathbf{X}(p|r)\right) \left\|\begin{bmatrix}\alpha_{r+1}^\circ \\ \vdots \\ \alpha_p^\circ\end{bmatrix}\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&- \left\| \left[\mathbf{X}(p|r)^\top (\mathbf{I} - \mathbf{\Pi}(r)) \mathbf{X}(p|r) \right]^{-1/2} \mathbf{X}(p|r)^\top (\mathbf{I} - \mathbf{\Pi}(r)) \boldsymbol{\varepsilon} \right\|^2 \\
&- \left\| \left[\mathbf{X}(p|r)^\top (\mathbf{I} - \mathbf{\Pi}(r)) \mathbf{X}(p|r) \right]^{-1/2} \mathbf{X}(p|r)^\top (\mathbf{I} - \mathbf{\Pi}(r)) \left(\boldsymbol{\nu}^\circ - \mathbf{R}(\widehat{\Theta}) \boldsymbol{\mu}^\circ \right) \right\|^2 \\
\geq\, & C_4 n \sum_{i=r+1}^{p} (\alpha_i^\circ)^2 + O(\log(n)) + O(q\rho_n) \qquad \text{(F.35)}
\end{aligned}
$$

with some constant $C_4 > 0$ for n large enough, where the $O(\log(n))$ bound on the RHS of (F.35) is due to (F.33), (F.34) and Lemma 1 of Lai and Wei (1982a), while the $O(q\rho_n)$ bound from (F.18), regardless of whether there are change points or not. Therefore, we have

$$
\begin{aligned}
&\mathrm{SC}\left(\{X_t\}_{t=1}^n, \widehat{\Theta}, r \right) - \mathrm{SC}\left(\{X_t\}_{t=1}^n, \widehat{\Theta}, p \right) \\
=& \frac{n}{2} \log \left(1 + \frac{\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, r) \widehat{\boldsymbol{\beta}}(\widehat{\Theta}, r)\|^2 - \|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, p) \widehat{\boldsymbol{\beta}}(\widehat{\Theta}, p)\|^2}{\|\mathbf{Y} - \mathbf{X}(\widehat{\Theta}, p) \widehat{\boldsymbol{\beta}}(\widehat{\Theta}, p)\|^2} \right) - (p - r) \xi_n \\
\geq& C_5 n - (p - r) \xi_n > 0
\end{aligned}
$$

with some constant $C_5 > 0$ for n large enough, by Assumption 3.4, (F.15) and (F.22).

(iv) When $M > 1$. The above (i)–(iii) completes the proof in the special case when Assumption 3.2 is met with $M = 1$. In the general case where $M > 1$, the above proof is readily adapted to prove the claim of the theorem.

(a) First, note that for any $l \geq l^*$, the intervals examined in Step 1 of the gSa algorithm, $\{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$, $v = 1, \ldots, q_l'$, correspond to one of the following cases under Assumption 3.2: **Null case** with no 'detectable' change points, i.e. either $\Theta \cap \{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\} = \emptyset$, or all $\theta_j \in \Theta \cap \{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$ satisfy $d_j^2 \min(\theta_j - \widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1} - \theta_j) \leq \rho_n$, or **change point case** with $\Theta \cap \{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\} \neq \emptyset$ and $d_j^2 \min(\theta_j - \widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1} - \theta_j) \geq D_n - \rho_n$ for at least one $\theta_j \in \Theta \cap \{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$.

In fact, when $l = l^*$, all $\{\widehat{\theta}_{l^*-1,u_v} + 1, \ldots, \widehat{\theta}_{l^*-1,u_v+1} - 1\}$ for $v = 1, \ldots, q_{l^*}'$, correspond to the change point case, while when $l \geq l^* + 1$, they all correspond to the null case.

(b) In the null case, the set $\mathcal{A} = \widehat{\Theta}_l \cap \{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$ serves the role of the set of spurious estimators, $\widehat{\Theta}$, as in (i) with $|\mathcal{A}|$ serving as $\widehat{q}$. Besides, we account for the possible estimation bias in the boundary points $\widehat{\theta}_{l-1,u_v}$ and $\widehat{\theta}_{l-1,u_v+1}$ in the case of $q \geq 1$ (while there are no detectable change points within $\{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$), by replacing the bound (F.14) derived in (i), with (F.19) and (F.20) in (ii). Consequently, (F.15) and (F.16) are written with $O\left(\log(n) \vee \widehat{q}(\rho_n \vee \omega_n^2)\right)$ (see (F.22) and (F.30)),

which leads to

$$\mathrm{SC}_0\left(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \widehat{\boldsymbol{\alpha}}(p)\right) - \mathrm{SC}\left(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \mathcal{A}, p\right)$$
$$= O\left(\log(n) \vee |\mathcal{A}|(\rho_n \vee \omega_n^2)\right) - |\mathcal{A}|\xi_n < 0$$

for n large enough.

(c) In the change point case, the arguments under (ii) are applied analogously by regarding $\mathcal{A}$ as $\widehat{\Theta}$ therein, with $|\mathcal{A}|$ equal to the number of detectable change points in $\{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$ as defined in (a). Then, we obtain

$$\mathrm{SC}_0\left(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \widehat{\boldsymbol{\alpha}}(p)\right) - \mathrm{SC}\left(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \mathcal{A}, p\right)$$
$$\geq C_3|\mathcal{A}|D_n + O\left(\log(n) \vee |\mathcal{A}|(\rho_n \vee \omega_n^2)\right) - |\mathcal{A}|\xi_n > 0$$

for n large enough.

(d) The proof on order selection consistency in (iii) holds from regardless of whether there are detectable change points in $\{\widehat{\theta}_{l-1,u_v} + 1, \ldots, \widehat{\theta}_{l-1,u_v+1} - 1\}$ or not. Thus with (a)–(c) above, the proof is complete.

F.3 Proof of Proposition B.1

For a fixed $j = 1, \ldots, q$, we drop the subscript j and write $\check{\theta} = \check{\theta}_j$, $\ell = \ell_j$, $r = r_j$, $\theta = \theta_j$, $f' = f'_j$ and $\delta = \delta_j$. In what follows, we assume that $\mathcal{X}_{\ell,\check{\theta},r} > 0$; otherwise, consider $-X_t$ (resp. $-f_t$ and $-Z_t$) in place of X_t (f_t and Z_t). Then, on $\mathcal{Z}_n$, we have

$$\max_{\ell<k<r} |\mathcal{Z}_{\ell,k,r}| \leq \max_{\ell<k<r} \left(\sqrt{\frac{r-k}{r-\ell}} + \sqrt{\frac{k-\ell}{r-\ell}}\right)\zeta_n = \sqrt{2}\zeta_n, \tag{F.36}$$

while by (B.1)–(B.2),

$$|\mathcal{F}_{\ell,\theta,r}| \geq \sqrt{\frac{(f')^2\delta}{4}}. \tag{F.37}$$

By Lemma F.2 and (B.2), we have $\mathcal{F}_{\ell,k,r}$ strictly increases, peaks at $k = \theta$ and then decreases in modulus without changing signs. Also by Lemma 7 of Wang and Samworth (2018), we obtain

$$|\mathcal{F}_{\ell,\theta,r} - \mathcal{F}_{\ell,k,r}| \geq \frac{2}{3\sqrt{6}} \frac{|f'|\,|k-\theta|}{\sqrt{\min(\theta-\ell, r-\theta)}} \tag{F.38}$$

for $|k-\theta| \le \min(\theta-\ell, r-\theta)/2$. Then, from (F.1) and (F.36)–(F.37),

$$|\mathcal{F}_{\ell,\check{\theta},r}| \ge |\mathcal{F}_{\ell,\theta,r}| - 2\max_{\ell<k<r}|\mathcal{Z}_{\ell,k,r}| \ge \sqrt{\frac{(f')^2\delta}{4}} - 2\sqrt{2}\zeta_n > \frac{\sqrt{(f')^2\delta}}{4}, \tag{F.39}$$

which implies that $|\mathcal{Z}_{\ell,\check{\theta},r}|/|\mathcal{F}_{\ell,\check{\theta},r}| = o(1)$ and consequently that $\mathcal{F}_{\ell,\theta,r} > \mathcal{F}_{\ell,\check{\theta},r} > 0$ for n large enough. Below, we consider the case where $\check{\theta} \le \theta$; the case where $\check{\theta} > \theta$ can be handled analogously. We first establish that

$$\theta - \check{\theta} \le \min(\theta-\ell, r-\theta)/2. \tag{F.40}$$

If $\theta - \check{\theta} > \min(\theta-\ell, r-\theta)/2 \ge \delta/4$ (due to (B.1)), by Lemma F.2 and (F.38), we have

$$\mathcal{F}_{\ell,\theta,r} - \mathcal{F}_{\ell,\check{\theta},r} \ge \frac{1}{3\sqrt{3}}\sqrt{(f')^2\delta}$$

while $|\mathcal{Z}_{\ell,\theta,r} - \mathcal{Z}_{\ell,\check{\theta},r}| \le 2\sqrt{2}\zeta_n$, thus contradicting that $\mathcal{X}_{\ell,\check{\theta},r} \ge \mathcal{X}_{\ell,\theta,r}$ under (F.1). Next, for some $\widetilde{\rho}_n$ satisfying $(f')^{-2}\widetilde{\rho}_n \le \delta/4$, we have

$$\begin{aligned}
&\mathsf{P}\left(\arg\max_{\ell<k<r}|\mathcal{X}_{\ell,k,r}| \le \theta - (f')^{-2}\widetilde{\rho}_n\right) \le \mathsf{P}\left(\max_{\theta-\delta/4\le k\le\theta-(f')^{-2}\widetilde{\rho}_n} \mathcal{X}_{\ell,k,r} \ge \mathcal{X}_{\ell,\theta,r}\right)\\
&\le \mathsf{P}\left(\max_{\theta-\delta/4\le k\le\theta-(f')^{-2}\widetilde{\rho}_n} (\mathcal{F}_{\ell,k,r}+\mathcal{Z}_{\ell,k,r})^2 - (\mathcal{F}_{\ell,\theta,r}+\mathcal{Z}_{\ell,\theta,r})^2 \ge 0\right)\\
&= \mathsf{P}\left(\max_{\theta-\delta/4\le k\le\theta-(f')^{-2}\widetilde{\rho}_n} -D_1(k)D_2(k)\left(1+\frac{A_1(k)}{D_1(k)}\right)\left(1+\frac{A_2(k)}{D_2(k)}\right) \ge 0\right)\\
&\le \mathsf{P}\left(\max_{\theta-\delta/4\le k\le\theta-(f')^{-2}\widetilde{\rho}_n} \left|\frac{A_1(k)A_2(k)}{D_1(k)D_2(k)} + \frac{A_1(k)}{D_1(k)} + \frac{A_2(k)}{D_2(k)}\right| \ge 1\right)\\
&\le 2\mathsf{P}\left(\max_{\theta-\delta/4\le k\le\theta-(f')^{-2}\widetilde{\rho}_n} \frac{|A_1(k)|}{D_1(k)} \ge \frac{1}{3}\right) + 2\mathsf{P}\left(\max_{\theta-\delta/4\le k\le\theta-(f')^{-2}\widetilde{\rho}_n} \frac{|A_2(k)|}{D_2(k)} \ge \frac{1}{3}\right), \quad \text{where}
\end{aligned}$$

$D_1(k) = \mathcal{F}_{\ell,\theta,r} - \mathcal{F}_{\ell,k,r}$, $D_2(k) = \mathcal{F}_{\ell,\theta,r} + \mathcal{F}_{\ell,k,r}$, $A_1(k) = \mathcal{Z}_{\ell,\theta,r} - \mathcal{Z}_{\ell,k,r}$, $A_2(k) = \mathcal{Z}_{\ell,\theta,r} + \mathcal{Z}_{\ell,k,r}$.

Note that

$$\begin{aligned}
|A_1(k)| \le &\left|\left(\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}} - \sqrt{\frac{r-\ell}{(k-\ell)(r-k)}}\right)\sum_{t=\ell+1}^{k}(Z_t - \bar{Z}_{\ell:r})\right|\\
&+ \sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\left|\sum_{t=k+1}^{\theta}(Z_t - \bar{Z}_{\ell:r})\right| =: A_{11}(k) + A_{12}(k).
\end{aligned}$$

For $k < \theta$, we obtain

$$\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}} - \sqrt{\frac{r-\ell}{(k-\ell)(r-k)}} = \sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\left(1 - \sqrt{\frac{(\theta-\ell)(r-\theta)}{(k-\ell)(r-k)}}\right)$$

$$\leq \sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\left(1-\sqrt{1-\frac{\theta-k}{r-k}}\right) \leq \frac{1}{2}\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\frac{\theta-k}{r-k}$$

and similarly,

$$\sqrt{\frac{r-\ell}{(k-\ell)(r-k)}}-\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}} \leq \frac{1}{2}\sqrt{\frac{r-\ell}{(k-\ell)(r-k)}}\frac{\theta-k}{\theta-\ell},$$

such that on $\mathcal{Z}_n$, due to (B.1) and (F.40),

$$A_{11}(k) \leq \sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\frac{2(\theta-k)}{\min(\theta-\ell, r-\theta)}\left(\sqrt{k-\ell}\,\zeta_n + \frac{k-\ell}{\sqrt{r-\ell}}\zeta_n\right) \leq \frac{4(\theta-k)\zeta_n}{\delta}.$$

Also, by (B.1),

$$A_{12}(k) \leq \sqrt{\frac{2}{\delta}}\left(\left|\sum_{t=k+1}^{\theta} Z_t\right| + \frac{\theta-k}{\sqrt{r-\ell}}\zeta_n\right).$$

Then, by (F.38) and (F.1), there exists some $c_3 > 0$ such that setting $\widetilde{\rho}_n = c_3(\widetilde{\zeta}_n)^2$, we have

$$\begin{aligned}
&\mathsf{P}\left(\max_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}\frac{|A_1(k)|}{D_1(k)} \geq \frac{1}{3},\, \widetilde{\mathcal{Z}}_n\right) \\
\leq &\mathsf{P}\left(\max_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}\frac{\sqrt{(f')^{-2}\widetilde{\rho}_n}}{\theta-k}\sum_{t=k+1}^{\theta} Z_t \geq \sqrt{\widetilde{\rho}_n}\left(\frac{1}{3}-\frac{(2\sqrt{2}+1)\zeta_n}{\sqrt{(f')^2\delta}}\right),\, \widetilde{\mathcal{Z}}_n\right) = 0,
\end{aligned}$$

which holds uniformly over $j = 1, \ldots, q$. Next, note that from (F.36),

$$\max_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}|A_2(k)| \leq 2\sqrt{2}\zeta_n,$$

while from (F.37),

$$\min_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}|D_2(k)| \geq \frac{\sqrt{(f')^2\delta}}{2}$$

and thus

$$\mathsf{P}\left(\max_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}\frac{|A_2(k)|}{D_2(k)} \geq \frac{1}{3}, \mathcal{Z}_n\right) = 0$$

under (F.1), which completes the proof.

G Assumptions 2.1 and 3.1

In this section, we provide an example that fulfils Assumptions 2.1 and 3.1 (iv) motivated by the Nagaev-type tail probability inequalities derived in Zhang and Wu (2017) for dependent time series with sub-exponential innovations.

Suppose that $Z_t = \sum_{\ell=0}^{\infty} b_\ell \varepsilon_{t-\ell}$ where the innovations $\{\varepsilon_t\}$ are i.i.d. sub-exponential random variables with $\mathsf{E}(\varepsilon_t) = 0$. Further, we assume that the linear coefficients decay polynomially such that there exists some $\gamma > 0$ and $\beta > 1$ satisfying $|b_\ell| \leq \gamma \ell^{-\beta}$ for all $\ell \geq 1$. With $\nu = 1$, the dependence adjusted sub-exponential norm

$$\|Z_\cdot\|_{\psi_\nu,0} = \sup_{m \geq 2} m^{-\nu} \sum_{t=0}^{\infty} \left\{ \mathsf{E}\left(\left| Z_t - Z_{t,\{0\}} \right|^m \right) \right\}^{1/m},$$

is bounded from the above by some fixed constant $C_1 > 0$, where $Z_{t,\{0\}} = \sum_{\ell=0,\, \ell \neq t}^{\infty} b_\ell \varepsilon_{t-\ell} + b_t \varepsilon_0'$ with ε_0' an independent copy of ε_0. Then, by Lemma C.4 of Zhang and Wu (2017), there exists a fixed constant $C_2 > 0$ such that

$$\mathsf{P}\left(\max_{0 \leq s < e \leq n} \frac{1}{\sqrt{e-s}} \left| \sum_{t=s+1}^{e} Z_t \right| \geq \zeta_n \right) \leq C_2 n(n+1) \exp\left(-\frac{3\zeta_n^{2/3}}{4e\|Z_\cdot\|_{\psi_1,0}} \right),$$

i.e. we can set $\zeta_n = C_3 \log^{3/2}(n)$ with a large enough $C_3 > 0$ (depending only on $\|Z_\cdot\|_{\psi_1,0}$) and have $\mathsf{P}(\mathcal{Z}_n) \to 1$. Using similar arguments and Bernstein's inequality (see e.g. Theorem 2.8.1 of Vershynin (2018)), we have $\mathsf{P}(\mathcal{E}_n) \to 1$ with $\omega_n \asymp \log(n)$.

References

Anastasiou, A., Chen, Y., Cho, H., and Fryzlewicz, P. (2020). *breakfast: Methods for Fast Multiple Change-Point Detection and Estimation.* R package version 2.1.

Anastasiou, A. and Fryzlewicz, P. (2020). Detecting multiple generalized change-points by isolating single ones. *Preprint.*

Cho, H. and Kirch, C. (2021). Two-stage data segmentation permitting multiscale change points, heavy tails and dependence. *Annals of the Institute of Statistical Mathematics (to appear).*

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-point Analysis*, volume 18. John Wiley & Sons Inc.

Dette, H., Schüler, T., and Vetter, M. (2020). Multiscale change point detection for dependent data. *Scandinavian Journal of Statistics*, 47:1243–1274.

Fearnhead, P. and Rigaill, G. (2020). Relating and comparing methods for detecting changes in mean. *Stat*, 9:e291.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:495–580.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, pages 1–44.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Kirch, C. (2006). *Resampling methods for the change analysis of dependent data*. PhD thesis, Universität zu Köln.

Lai, T. and Wei, C. (1982a). Asymptotic properties of projections with applications to stochastic regression problems. *Journal of Multivariate Analysis*, 12:346–370.

Lai, T. and Wei, C. (1982b). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10:154–166.

Lai, T. and Wei, C. (1983). Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of Multivariate Analysis*, 13:1–23.

Parker, D. E., Legg, T. P., and Folland, C. K. (1992). A new daily central England temperature series, 1772–1991. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 12:317–342.

Reid, P. C., Hari, R. E., Beaugrand, G., Livingstone, D. M., Marty, C., Straile, D., Barichivich, J., Goberville, E., Adrian, R., Aono, Y., et al. (2016). Global impacts of the 1980s regime shift. *Global change Biology*, 22:682–703.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2020). *DeCAFS: Detecting Changes in Autocorrelated and Fluctuating Signals*. R package version 3.2.3.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2021). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *Journal of the American Statistical Association (to appear)*.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association*, 105:1228–1240.

Venkatraman, E. (1992). Consistency results in multiple change-point problems. *Technical Report No. 24, Department of Statistics, Stanford University*.

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.

Verzelen, N., Fromont, M., Lerasle, M., and Reynaud-Bouret, P. (2020). Optimal change-point detection and localization. *arXiv preprint arXiv:2010.11470*.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:57–83.

Wu, W. and Zhou, Z. (2020). Multiscale jump testing and estimation under complex temporal

dynamics. *arXiv preprint arXiv:1909.06307*.

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919.

Zhao, Z., Jiang, F., and Shao, X. (2021). Segmenting time series via self-normalization. *arXiv preprint arXiv:2112.05331*.