Multiple change point detection under serial dependence: Wild energy maximisation and gappy Schwarz criterion

Haeran Cho[1] Piotr Fryzlewicz[2]

March 10, 2021

Abstract

We propose a methodology for detecting multiple change points in the mean of an otherwise stationary, autocorrelated, linear time series. It combines solution path generation based on the wild energy maximisation principle, and an information criterion-based model selection strategy termed gappy Schwarz criterion. The former is well-suited to separating shifts in the mean from fluctuations due to serial correlations, while the latter simultaneously estimates the dependence structure and the number of change points without performing the difficult task of estimating the level of the noise as quantified e.g. by the long-run variance. We provide modular investigation into their theoretical properties and show that the combined methodology, named WEM.gSC, achieves consistency in estimating both the total number and the locations of the change points. The good performance of WEM.gSC is demonstrated via extensive simulation studies, and we further illustrate its usefulness by applying the methodology to London air quality data.

Keywords: data segmentation, wild binary segmentation, information criterion, autoregressive time series

1 Introduction

The objective of this paper is to propose a new methodology for detecting possibly multiple change points in the piecewise constant mean of an otherwise stationary, linear time series. This is a known difficult problem in multiple change point analysis, whose challenge comes from the fact that change points can mask as natural fluctuations in a serially dependent process and vice versa, an observation made by several authors including Mikosch and Stărică (2004), Berkes et al. (2006) and Norwood and Killick (2018). Disentangling these two effects usually requires performing a statistically difficult task, such as accurately estimating the long-run variance (LRV) of the process in question in the presence of multiple shifts in the mean.

[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.

Our proposed methodology rests on two important ingredients, both of which are designed to mitigate the adverse effect of serial dependence.

The first ingredient relates to how we construct a solution path by iteratively locating the next most likely change points in the data. At each iteration, the next change point candidate is chosen as the one accounting for the most 'energy' in the data sections between the previously proposed candidates, in the sense of maximising absolute cumulative sum (CUSUM) statistics over a collection of intervals of varying lengths and locations. This approach is particularly useful under serial dependence since with a high probability, it generates a large gap between the max-CUSUMs attributed to change points and the remaining ones, which helps separate the effect of change points from that of serial dependence later at the model selection stage. We refer to this adaptive CUSUM selection as 'wild energy maximisation' (WEM) and justify the label later on. The thus-constructed solution path is a decreasing sequence of max-CUSUMs corresponding to each successively proposed change point candidate.

The second ingredient relates to how we select the preferred model along the solution path. To this end, we propose a new Schwarz-like (Schwarz, 1978) information criterion constructed under a parametric modelling assumption, and combine it with a novel 'backward elimination' strategy for estimating the dependence structure and the number of change points simultaneously. Information criteria have been widely adopted for model selection in change point problems (Yao, 1988; Kühn, 2001). However, through its application on the WEM-generated solution path, our proposal is different from the conventional use of an information criterion in the change point literature. More specifically, thanks to the WEM principle, only some candidate models, i.e. those corresponding to large drops in the decreasing sequence of max-CUSUMs, are seen as serious contenders for the final model. Therefore, our model selection strategy only considers a small subset of model candidates located (possibly non-consecutively) on the solution path, reducing the number of model candidates and facilitating the final model choice; hence the label of 'gappy Schwarz criterion' (gSC). The evaluation of the information criterion starts from the largest to the smallest (null model corresponding to mean stationarity) of the nested candidate models, which has an advantage over the direct minimisation of the information criterion on a solution path, by avoiding the substantial technical challenges linked to dealing with under-specified models in the presence of serial dependence.

The two ingredients: WEM solution path generation and the gSC criterion make up the WEM.gSC algorithm and throughout the paper, we highlight the important roles played by these two components and argue that WEM.gSC offers state-of-the-art performance in the problem of multiple change point detection under serially dependent noise.

We briefly review the existing literature on multiple change point detection in the presence of serial dependence and situate WEM.gSC in this context; see also Aue and Horváth (2013) for a review. One line of research extends the applicability of the test statistics developed

for independent data, such as the CUSUM (Csörgő and Horváth, 1997) and moving sum (MOSUM, Hušková and Slabý; 2001) statistics, to time series setting. Their performance depends on the estimated level of noise quantified e.g. by the LRV, and the estimators of the latter in the presence of multiple change points have been proposed (Tecuapetla-Gómez and Munk, 2017; Eichinger and Kirch, 2018; Dette et al., 2020). The estimation of the LRV, even when the mean changes are not present, has long been noted as a difficult problem (Robbins et al., 2011); the popularly adopted kernel estimator of LRV tends to incur downward bias (den Haan and Levin, 1997; Chan and Yau, 2017), and can even take negative values when the LRV is small (Hušková and Kirch, 2010). It becomes even more challenging in the presence of (possibly) multiple change points, and the estimators may be sensitive to the choice of tuning parameters which are often related to the frequency of change points. Indeed, Perron (2006) notes that 'there is no reliable method to appropriately choose this parameter in the context of structural changes'. Self-normalisation of test statistics avoids direct estimation of this nuisance parameter (Shao and Zhang, 2010; Pešta and Wendler, 2020) but theoretical investigation into its validity is often limited to change point testing, i.e. when there is at most a single change point, with the exception of Wu and Zhou (2020). Consistency of the methods utilising penalised least squares estimation (Lavielle and Moulines, 2000) or the Schwarz criterion (Cho and Kirch, 2020b) constructed without further parametric assumptions, has been established under general conditions permitting serial dependence and heavy-tails, but their consistency relies on the choice of the penalty, which in turn depends on the level of the noise.

The second line of research utilises particular linear or non-linear time series models such as the autoregressive (AR) and conditionally heteroscedastic models, and estimates the serial dependence and change point structures simultaneously. AR(1)-type dependence has often been adopted to describe the serial correlations in this context: Chakar et al. (2017) and Romano et al. (2020b) propose to minimise the penalised cost function for detection of multiple change points in the mean of AR(1) processes via dynamic programming, while Fang and Siegmund (2020) study a pseudo-sequential approach to change point detection in the level or slope of the data. Fryzlewicz (2020b) proposes to circumvent the need to estimate the AR parameters accurately through the use of a multi-resolution sup-norm (rather than the ordinary least squares) in fitting the postulated AR model, but this is only possible because the goal of the method is purely inferential and therefore different from ours.

More generally, Davis et al. (2006, 2008), Cho and Fryzlewicz (2012), Bardet et al. (2012), Chan et al. (2014), Yau and Zhao (2016) and Korkas and Fryzlewicz (2017), among others, study multiple change point detection under piecewise stationary, univariate time series models, and Cho and Fryzlewicz (2015), Safikhani and Shojaie (2020) and Wang et al. (2019) under high-dimensional time series models.

We now describe the novelty of WEM.gSC against this literature background and sum-

marise the main contributions of this paper.

1. The WEM principle has been adopted in the i.i.d. noise setting in Fryzlewicz (2014) and Fryzlewicz (2020a), but its benefits have not been noted or exploited in the presence of serial dependence. We make the key observation that WEM, i.e. fitting the CUSUM statistic over a representative range of sub-samples of the data, is particularly useful for change point detection in difficult dependent-data problems, achieving the (inherently difficult) disentanglement of the large max-CUSUMs attributed to change points, from those attributed to the fluctuations due to serial correlations.

2. The 'gappy' application of the information criterion, which facilitates model selection by reducing the space of models under consideration, is, to the best of our knowledge, new. The gSC model selection assumes a parametric AR(p) model in line with some of the existing literature (see the references earlier), but the sequential evaluation of the information criterion starting from the largest model candidate sets our proposed methodology apart from the commonly used methods involving global (Davis et al., 2006; Killick et al., 2012; Maidstone et al., 2017; Romano et al., 2020b) or approximate (Chan et al., 2014) minimisation of an objective function, such as penalised cost functions or information criteria.

WEM.gSC is modular in the sense that both WEM and gSC can be combined with other model selection and solution path procedures, respectively. For example, instead of the maximally selected CUSUMs as in WEM, we could build a solution path out of MOSUMs as outlined in Cho and Kirch (2020a), or residual sums of squares from the best signal fits with varying number of change points, the latter being frequently used in 'slope heuristics' and related adaptive methods for penalty selection, see Baudry et al. (2012) and the references therein. Similarly, we can perform the model selection e.g. by applying a suitable threshold on the max-CUSUMs, extending the approach commonly adopted in the i.i.d. setting (Fryzlewicz, 2014, 2020a). We provide separate theoretical analyses of WEM and gSC so that they can readily be fed into the analysis of such modifications.

The paper is organised as follows. In Sections 2 and 3, we introduce the two ingredients of WEM.gSC individually, and show its consistency in multiple change point detection in the presence of serial dependence. In Section 4, we provide the applications of the WEM.gSC to London air quality datasets. The Supplementary Appendix contains comprehensive simulation studies, further real data analysis and the proofs of the theoretical results. The R software implementing WEM.gSC is available from `https://github.com/haeran-cho/wem.gsc`.

2 Change point solution path via WEM principle

2.1 WEM principle

We consider the canonical change point model

$$X_t = f_t + Z_t = f_0 + \sum_{j=1}^{q} f'_j \cdot \mathbb{I}(t \geq \theta_j + 1) + Z_t, \quad t = 1, \ldots, n. \tag{1}$$

Under model (1), the set $\Theta = \Theta_n := \{\theta_1, \ldots, \theta_q\}$ with $\theta_j = \theta_{j,n}$, contains q change points (with $\theta_0 = 0$ and $\theta_{q+1} = n$) at which the mean of X_t undergoes changes of size $f'_j = f'_{j,n}$. We assume that the number of change points q does not vary with the sample size n, and we allow serial dependence in the sequence of errors $\{Z_t\}_{t=1}^n$ with $\mathsf{E}(Z_t) = 0$.

A large number of multiple change point detection methodologies have been proposed for a variant of model (1) in which the errors $\{Z_t\}_{t=1}^n$ are serially independent. In particular, a popular class of multiscale methods aim to isolate change points for their detection by drawing a large number of sub-samples of the data living on sub-intervals of $[1, n]$; when a sufficient number of sub-samples are drawn, there exists at least one interval which is well-suited for the detection and localisation of θ_j for each $\theta_j, j = 1, \ldots, q$. Methods in this category include the Wild Binary Segmentation (WBS, Fryzlewicz; 2014), the Narrowest-Over-Threshold method (Baranowski et al., 2019), the Seeded Binary Segmentation (Kovács et al., 2020) and the WBS2 (Fryzlewicz, 2020a). In all of the above, theoretical properties have been established assuming i.i.d. (sub-)Gaussianity of $\{Z_t\}_{t=1}^n$.

In the remainder of this paper, we focus on WBS2, which produces a complete solution path to the change point detection problem. It leaves open the possibility of estimating any number (from 0 to $n-1$) of change points, and this decision is left to the subsequent model selection procedure, which chooses a suitable model along the solution path. A key feature of the WBS2 is that for any given $0 \leq s < e \leq n$, we identify the sub-interval $(s_\circ, e_\circ] \subset (s, e]$ and its inner point $k_\circ \in (s_\circ, e_\circ)$, which obtains a local split of the data that contains the most energy. More specifically, let $\mathcal{R}_{s,e}$ denote a set of intervals drawn from $\mathcal{A}_{s,e} := \{(\ell, r) \in \mathbb{Z}^2 : s \leq \ell < r \leq e \text{ and } r - \ell > 1\}$, either randomly or deterministically, with $|\mathcal{R}_{s,e}| = \min(R_n, |\mathcal{A}_{s,e}|)$ for some given $R_n \leq n(n-1)/2$. Then, we identify $(s_\circ, e_\circ] \in \mathcal{R}_{s,e}$ that achieves the maximum absolute CUSUM statistic, as

$$\begin{aligned}(s_\circ, k_\circ, e_\circ) &= \arg\max_{\substack{(\ell,k,r):\, \ell<k<r \\ (\ell,r)\in\mathcal{R}_{s,e}}} |\mathcal{X}_{\ell,k,r}|, \quad \text{where} \\ \mathcal{X}_{\ell,k,r} &= \sqrt{\frac{(k-\ell)(r-k)}{r-\ell}} \left(\frac{1}{k-\ell} \sum_{t=\ell+1}^{k} X_t - \frac{1}{r-k} \sum_{t=k+1}^{r} X_t \right). \end{aligned} \tag{2}$$

Starting with $(s, e) = (0, n)$, recursively repeating the above operation over the segments

defined by the thus-identified $k_\circ$, i.e. $(s, k_\circ]$ and $(k_\circ, e]$, generates a complete solution path that attaches an order of importance to $\{1, \ldots, n-1\}$ as change point candidates; see Algorithm 1 in Appendix A for the pseudo code of the WBS2 algorithm, and for how to perform a deterministic sampling of $\mathcal{R}_{s,e}$ from $\mathcal{A}_{s,e}$.

We denote by $\mathcal{P}_0$ the output generated by the WBS2: each element of $\mathcal{P}_0$ contains the triplet of the beginning and the end of the interval and the break that returns the maximum energy (measured as in (2)) at a particular iteration, and the corresponding max-CUSUM statistic. We refer to this maximal selection of the CUSUM statistic as wild energy maximisation (WEM). The order of the sorted max-CUSUMs (in decreasing order) provides a natural ordering of the candidate change points, which gives rise to the following solution path $\mathcal{P} := \{(s_{(m)}, k_{(m)}, e_{(m)}, \mathcal{X}_{(m)}) : m = 1, \ldots, P\}$, where

$$\mathcal{X}_{(m)} := |\mathcal{X}_{s_{(m)}, k_{(m)}, e_{(m)}}| \quad \text{satisfying} \quad \mathcal{X}_{(1)} \geq \mathcal{X}_{(2)} \geq \ldots \geq \mathcal{X}_{(P)} > 0; \tag{3}$$

if $\mathcal{X}_{(m)} = 0$ for some $m \leq |\mathcal{P}_0|$, then $(s_{(m)}, k_{(m)}, e_{(m)})$ is not associated with any change point and thus such entries are excluded from the solution path $\mathcal{P}$.

The WEM principle provides a good basis for model selection, i.e. selecting the correct number of change points, in the presence of serially dependent noise. This is due to the iterative identification of the local split with the maximum energy, which helps separate the large max-CUSUMs attributed to changes in the mean, from those which are not. This is not to say that the WEM principle is best suited for estimating the locations of the change points (see e.g. Proposition D.1 of Cho and Kirch (2020a)). Nonetheless, its possible lack of location-estimation optimality is a price worth paying since the problem of estimating the number of change points is typically more difficult than that of estimating their locations, particularly in the dependent data setting in which change points can easily be mistaken for fluctuations in the serially dependent noise, and vice versa. In Appendix B, we propose a straightforward location refinement step, which achieves minimax rate optimality in multiple change point localisation under general conditions permitting serial correlations.

In light of the WEM property of the WBS2, we expect to observe a large gap between the CUSUM statistics $\mathcal{X}_{(m)}$ computed over those intervals $(s_{(m)}, e_{(m)})$ that contain change points well within their interior, and the remaining CUSUMs. Therefore, for the purpose of model selection, we can exploit this large gap in $\mathcal{X}_{(m)}$, $1 \leq m \leq P$ or equivalently, in $\mathcal{Y}_{(m)} := \log(\mathcal{X}_{(m)})$; we later show that under certain assumptions on the size of changes and the level of noise, the large log-CUSUMs $\mathcal{Y}_{(m)}$ attributed to change points scale as $\log(n)$ while the rest scale as $\log\log(n)$. For the identification of a large gap in the sorted log-CUSUM statistics and, consequently, the selection of the 'best' change point model for the data under (1), we consider two approaches that iteratively generate a sequence of nested

change point models

$$\emptyset = \widehat{\Theta}_0 \subset \widehat{\Theta}_1 \subset \ldots \subset \widehat{\Theta}_M \subset \{0, \ldots, n-1\} \quad \text{with} \quad \widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1} \neq \emptyset \quad \forall\, l = 1, \ldots, M, \tag{4}$$

for some M determined by a pre-specified upper bound $Q = Q_n \leq P$ on the number of change points. These two approaches are described next, under the labels of 'largest difference' and 'double CUSUM'.

Largest difference (LD)

The simplest way of identifying a large gap in $\mathcal{Y}_{(m)}$, $m = 1, \ldots, P$, is to look for the large difference $\mathcal{Y}_{(m)} - \mathcal{Y}_{(m+1)}$, $m = 1, \ldots, P-1$. However, the largest gap may not necessarily correspond to the difference between the max-CUSUMs attributed to mean shifts and spurious ones attributed to fluctuations in the errors, but simply be due to the heterogeneity in the change points (i.e. some changes being more pronounced and therefore easier to detect than others). Therefore, we iteratively identify large gaps and generate g_l^{LD}, $1 \leq l \leq M$ such that

$$\mathcal{Y}_{(g_1^{\mathrm{LD}})} - \mathcal{Y}_{(g_1^{\mathrm{LD}}+1)} \geq \mathcal{Y}_{(g_2^{\mathrm{LD}})} - \mathcal{Y}_{(g_2^{\mathrm{LD}}+1)} \geq \ldots \geq \mathcal{Y}_{(g_M^{\mathrm{LD}})} - \mathcal{Y}_{(g_M^{\mathrm{LD}}+1)} \tag{5}$$

with $g_0^{\mathrm{LD}} = 0$ and $g_l^{\mathrm{LD}} < Q$ for all $1 \leq l \leq M$. This returns a sequence of nested models $\emptyset = \widehat{\Theta}_0^{\mathrm{LD}} \subset \widehat{\Theta}_1^{\mathrm{LD}} \subset \ldots \subset \widehat{\Theta}_M^{\mathrm{LD}}$ with $\widehat{\Theta}_l^{\mathrm{LD}} = \widehat{\Theta}_{l-1}^{\mathrm{LD}} \cup \{k_{(g_{l-1}^{\mathrm{LD}}+1)}, \ldots, k_{(g_l^{\mathrm{LD}})}\}$, which can be considered in the model selection stage.

Double CUSUM (DC)

We adapt the DC methodology, originally proposed in Cho (2016) for high-dimensional panel data segmentation, to identify a large gap in the ordered log-CUSUMs $\mathcal{Y}_{(m)}$, by sequentially maximising the DC statistic

$$\mathbb{Y}_{i,m,Q} := \sqrt{\frac{(m-i)(Q-m)}{Q-i}} \left(\frac{1}{m-i} \sum_{r=i+1}^{m} \mathcal{Y}_{(r)} - \frac{1}{Q-m} \sum_{r=m+1}^{Q} \mathcal{Y}_{(r)} \right),$$

over $m = i+1, \ldots, Q-1$. By construction, $\mathbb{Y}_{0,m,Q}$ contrasts the m largest log-CUSUMs $\mathcal{Y}_{(r)}$, $1 \leq r \leq m$, against the remaining ones $\mathcal{Y}_{(r)}$, $m+1 \leq r \leq Q$, and thus is well-suited to separating log-CUSUMs attributed to change points from those that are not. Then, $g_1^{\mathrm{DC}} = \arg\max_{0<m<Q} \mathbb{Y}_{0,m,Q}$ indicates where the gap in the sorted log-CUSUMs is large, and thus $\widehat{\Theta}_1^{\mathrm{DC}} = \{k_{(1)}^{\mathrm{DC}}, \ldots, k_{(g_1^{\mathrm{DC}})}^{\mathrm{DC}}\}$ can serve as a candidate change point model. As with g_1^{LD}, however, this large gap may merely be due to change point heterogeneity, and therefore we adopt an iterative approach as follows: with $g_0^{\mathrm{DC}} = 0$, we sequentially generate

$$g_{l+1}^{\mathrm{DC}} = \arg\max_{g_l^{\mathrm{DC}}<m<Q} \mathbb{Y}_{g_l^{\mathrm{DC}},m,Q} \tag{6}$$

until, for some M, we have $g_M^{\mathrm{DC}} < Q$ while $g_{M+1}^{\mathrm{DC}} \geq Q$. This results in a sequence of nested models $\emptyset = \widehat{\Theta}_0^{\mathrm{DC}} \subset \widehat{\Theta}_1^{\mathrm{DC}} \subset \ldots \subset \widehat{\Theta}_M^{\mathrm{DC}}$, where $\widehat{\Theta}_l^{\mathrm{DC}} = \widehat{\Theta}_{l-1}^{\mathrm{DC}} \cup \{k_{(g_{l-1}^{\mathrm{DC}}+1)}, \ldots, k_{(g_l^{\mathrm{DC}})}\}$.

Applying either LD or DC iteratively yields a sequence of nested change point models (4), on which we perform model selection without ruling out the null model $\widehat{\Theta}_0 = \emptyset$. Typically, the thus-constructed sequence of model candidates is much sparser than the sequence of all possible models $\mathcal{K}_1 \subset \mathcal{K}_2 \subset \ldots$ with $\mathcal{K}_l = \{k_{(1)}, \ldots, k_{(l)}\}$ and therefore, intuitively, our model selection task should be easier than if we worked with the entire solution path of all nested models.

Remark 2.1. The DC approach in (6) requires the selection of a tuning parameter $Q = Q_n$, which imposes an upper bound on the maximum number of change points. This bound should not be too large, as it would otherwise be too challenging to disentangle the effect of very frequent change points from that of the serial dependence. Later, we present Theorem 2.1 which permits $Q_n \to \infty$ at a logarithmic rate, and discuss its choice in accordance with this theoretical requirement in Appendix C.

2.2 Theoretical properties

In this section, we establish the theoretical properties of the sequences of nested change point models. The following assumptions are, respectively, on the distribution of $\{Z_t\}_{t=1}^n$ and the size of changes under $H_1 : q \geq 1$.

Assumption 2.1. Let $\{Z_t\}_{t=1}^n$ be a sequence of random variables satisfying $\mathsf{E}(Z_t) = 0$ and $\mathsf{Var}(Z_t) = \sigma_Z^2$ with $\sigma_Z \in (0, \infty)$. Also, let $\mathsf{P}(\mathcal{Z}_n) \to 1$ with ζ_n satisfying $\sqrt{\log(n)} = O(\zeta_n)$ and $\zeta_n = O(\log^{\kappa}(n))$ for some $\kappa \in [1/2, \infty)$, where

$$\mathcal{Z}_n = \left\{ \max_{0 \leq s < e \leq n} (e-s)^{-1/2} \left| \sum_{t=s+1}^{e} Z_t \right| \leq \zeta_n \right\}.$$

Assumption 2.2. Let $\delta_j = \delta_{j,n} := \min(\theta_j - \theta_{j-1}, \theta_{j+1} - \theta_j)$ and recall that $f_j' = f_{\theta_j+1} - f_{\theta_j}$ for $j = 1, \ldots, q$. Then, $\max_{1 \leq j \leq q} |f_j'| = O(1)$. Also, there exists some $c_1 \in (0, 1)$ such that $\min_{1 \leq j \leq q} \delta_j \geq c_1 n$, and for some $\varphi > 0$, we have $\zeta_n^2 / (\min_{1 \leq j \leq q} (f_j')^2 \delta_j) = O(n^{-\varphi})$.

Remark 2.2. Assumption 2.1 permits $\{Z_t\}_{t=1}^n$ to have heavier tails than sub-Gaussian such as sub-exponential (Vershynin, 2018) or sub-Weibull (Vladimirova et al., 2019). Bernstein-type concentration inequalities have been developed in time series settings which, together with the arguments similar to those adopted in Lemma 1 of Boysen et al. (2007), are applicable to show that Assumption 2.1 holds with a logarithmic ζ_n for light-tailed $\{Z_t\}_{t=1}^n$: Doukhan and Neumann (2007) derive a Bernstein-type inequality for weakly dependent time series with $\mathsf{E}(|Z_t|^k) \leq (k!)^{\nu} C^k$ for all $k \geq 1$ and some $\nu \geq 0$ and $C > 0$; the results from Merlevède et al. (2011) apply to geometrically strong mixing sequences with sub-exponential tails. Alternatively, under the invariance principle, if there exists (possibly after enlarging the probability

space) a standard Wiener process $W(\cdot)$ such that $\sum_{t=1}^{\ell} Z_t - W(\ell) = O(\log^{\kappa'}(\ell))$ a.s. with $\kappa' \geq 1$, then Assumption 2.1 holds with $\zeta_n \asymp \log^{\kappa}(n)$ for any $\kappa > \kappa'$, where we denote by $a_n \asymp b_n$ to indicate that $a_n = O(b_n)$ and $b_n = O(a_n)$. Such invariance principles have been derived for dependent data under weak dependence such as mixing (Kuelbs and Philipp, 1980) and functional dependence measure (Berkes et al., 2014) conditions, to name but a few. As remarked in Proposition 2.1 (c.i) of Cho and Kirch (2020b), the thus-derived ζ_n usually does not provide the tightest upper bound, but it suits our purpose in controlling the level of noise.

Theorem 2.1. Let Assumptions 2.1 and 2.2 hold. Suppose that R_n, the number of intervals at each iteration, and $Q = Q_n$, the upper bound on the number of change points, satisfy

$$R_n \geq \frac{9}{8}\left(\frac{n}{\min_{1\leq j\leq q}\delta_j}\right)^2 + 1, \quad \text{and} \quad \frac{Q_n \log^2(\zeta_n)}{\log^2(n)} = o(1) \text{ with } Q_n > q. \tag{7}$$

Then, on $\mathcal{Z}_n$, the following statements hold for n large enough and some $c_2 \in (0, \infty)$.

(i) Let $\widehat{\Theta}[q] = \{\widehat{\theta}_j,\, 1 \leq j \leq q : \widehat{\theta}_1 < \ldots < \widehat{\theta}_q\}$ denote the set of q change point location estimators corresponding to the q largest max-CUSUMs $\mathcal{X}_{(m)}$, $1 \leq m \leq q$ obtained as in (3). Then, $\max_{1\leq j\leq q}(f'_j)^2|\widehat{\theta}_j - \theta_j| \leq c_2\zeta_n^2$.

(ii) The sorted log-CUSUMs $\mathcal{Y}_{(m)}$ satisfy $\mathcal{Y}_{(m)} = \gamma_m \log(n)(1 + o(1))$ for $m = 1, \ldots, q$, while $\mathcal{Y}_{(m)} \leq \kappa_m \log(\zeta_n)(1 + o(1))$ for $m \geq q + 1$, where $\{\gamma_m\}_{m=1}^{q}$ and $\{\kappa_m\}_{m\geq q+1}$ are non-increasing sequences with $0 < \gamma_m \leq 1/2$ and $0 \leq \kappa_m \leq 1$.

(iii) For any $i = 0, \ldots, q-1$, we have $q \geq \arg\max_{i<m<Q} \mathbb{Y}_{i,m,Q}$.

(iv) For some $i = 0, \ldots, q-1$, if $\gamma_{i+1} = \ldots = \gamma_q$, then $q = \arg\max_{i<m<Q} \mathbb{Y}_{i,m,Q}$.

Statements (i)–(ii) in Theorem 2.1 establish that for the solution path $\mathcal{P}$ obtained according to the WEM principle, the entries corresponding to the q largest max-CUSUMs contain the estimators of all q change points θ_j. Besides, the q largest log-CUSUMs are of order $\log(n)$ and are therefore distinguished from the rest of the log-CUSUMs bounded as $O(\log\log(n))$ under Assumption 2.1. This implies that the sequence of nested change point models (4), generated by sequentially identifying the largest difference in $\mathbb{Y}_{(m)}$ as in (5), contains the consistent model $\widehat{\Theta}[q]$ as a candidate model. Theorem 2.1 (iii)–(iv) show that sequential maximisation of the DC statistics as in (6) also obtains the model sequence containing $\widehat{\Theta}[q]$. In particular, when the change points are homogeneous in the sense that $\gamma_1 = \ldots = \gamma_q$, a single step leads to $\widehat{q}_1^{\mathrm{DC}} = \arg\max_{0<m<Q} \mathbb{Y}_{0,m,Q} = q$, such that $\widehat{\Theta}_1 = \widehat{\Theta}[q]$ consistently estimates Θ. For example, when $d_j^{-1} = O(1)$ for all j, the change points belong to the homogeneous case. Typically, it is unknown whether the change points are homogeneous, and therefore it is of importance to develop a consistent model selection methodology applicable to the sequence of models in (4); this is achieved in Section 3.

3 Model selection with gSC

We now discuss how to consistently estimate the number and the locations of change points by choosing an appropriate model from the sequence of nested change point models (4). As mentioned in Introduction, achieving this by estimating the LRV is difficult, particularly in the presence of multiple mean shifts. Instead, we work with a parametric model imposing an AR structure on $\{Z_t\}_{t=1}^n$, and propose a model selection strategy based on an information criterion. In doing so, the problem of consistently estimating the noise level characterised by the LRV is effectively reduced to that of estimating the innovation variance, which is considerably easier. We propose a novel, backward elimination-type application of the information criterion and show its usefulness when the model selection is performed simultaneously with the modelling of the serial dependence.

Section 3.1 introduces the Schwarz criterion constructed under a parametric model on the error sequence $\{Z_t\}_{t=1}^n$ in (1). Section 3.2 describes the proposed model selection methodology termed gSC (gappy Schwarz criterion), which applies to a sequence of nested change point models and consistently estimate the number of change points as well as their locations as shown in Section 3.3, provided that the model sequence (including the null model) contains such a consistent change point model. Since this is the case for the model sequence generated by the WEM principle introduced in Section 2 (see Theorem 2.1), the combined methodology WEM.gSC achieves consistency in multiple change point detection.

3.1 Schwarz criterion in the presence of autoregressive errors

We assume that $\{Z_t\}_{t=1}^n$ in (1) is a stationary AR process of order p, i.e.

$$Z_t = \sum_{i=1}^{p} a_i Z_{t-i} + \varepsilon_t \quad \text{such that} \quad X_t = (1 - a(B))f_t + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t, \tag{8}$$

where $a(B) = \sum_{i=1}^p a_i B^i$ is defined with the backshift operator B. The innovations $\{\varepsilon_t\}_{t=1}^n$ satisfy $\mathsf{E}(\varepsilon_t) = 0$ and $\mathsf{Var}(\varepsilon_t) = \sigma_\varepsilon^2 \in (0, \infty)$, and are assumed to have no serial correlations; further assumptions on $\{\varepsilon_t\}_{t=1}^n$ are made in Assumption 3.1. We denote by $\mu_j^\circ := (1 - \sum_{i=1}^p a_i) f_{\theta_j+1}$ the effective mean level over each interval $\theta_j + p + 1 \le t \le \theta_{j+1}$, for $j = 0, \ldots, q$, and by $d_j = d_{j,n} := \mu_j^\circ - \mu_{j-1}^\circ$ the effective size of change correspondingly. Also recall that $\delta_j = \min(\theta_j - \theta_{j-1}, \theta_{j+1} - \theta_j)$. In the model selection procedure, we do not assume that the AR order p is known; rather, its data-driven choice is incorporated into the model selection methodology as described later.

Suppose that the AR order is set to be some integer $r \ge 0$, and a change point model given by a set of change point candidates $\mathcal{A} = \{k_j,\, 1 \le j \le m : k_1 < \ldots < k_m\} \subset \{1, \ldots, n\}$.

Then, the Schwarz criterion (Schwarz, 1978, SC) is defined as

$$\mathrm{SC}\left(\{X_t\}_{t=1}^n, \mathcal{A}, r\right) = \frac{n}{2}\log\left(\widehat{\sigma}_n^2\left(\{X_t\}_{t=1}^n, \mathcal{A}, r\right)\right) + (|\mathcal{A}| + r)\xi_n, \tag{9}$$

where $\widehat{\sigma}_n^2(\mathcal{A}, r)$ is an estimator of the innovation variance σ_ε^2 serving as a goodness-of-fit measure, and the penalty is imposed on the model complexity given by the AR order and the number of change points; the requirement on the penalty parameter in relation to the distribution of $\{\varepsilon_t\}$ is discussed in Assumption 3.4. For notational convenience, we assume that $X_0, \ldots, X_{-r+1}$ are available and their means remain constant such that $\mathsf{E}(X_t) = \mathsf{E}(X_1)$ for $t \leq 0$.

Ignoring the boundaries $t = \theta_j + 1, \ldots, \theta_j + p$, $1 \leq j \leq q$ over which $(1 - a(B))f_t$ is not exactly piecewise constant, we propose to measure the goodness-of-fit as

$$\widehat{\sigma}_n^2(\{X_t\}_{t=1}^n, \mathcal{A}, r) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2, \quad \text{where} \quad \mathbf{Y} = (X_1, \ldots, X_n)^\top \quad \text{and}$$

$$\mathbf{X} = \mathbf{X}(\mathcal{A}, r) = \begin{bmatrix} \underbrace{\mathbf{L}(r)}_{n\times r} & \underbrace{\mathbf{R}(\mathcal{A})}_{n\times(m+1)} \end{bmatrix} = \begin{bmatrix} X_0 & \cdots & X_{1-r} & 1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & \\ X_{k_1-1} & \cdots & X_{k_1-r} & 1 & 0 & 0 & \cdots & 0 \\ X_{k_1} & \cdots & X_{k_1-r+1} & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & & & & \\ X_{n-1} & \cdots & X_{n-r} & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{10}$$

The design matrix $\mathbf{X}$ contains the AR part of (8) in $\mathbf{L}$ and the deterministic part in $\mathbf{R}$. The vector of regression parameters is partitioned accordingly into the AR parameters and time-varying levels. To obtain its estimator, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathcal{A}, r)$, we propose the following scheme: for each $j = 0, \ldots, m$, let $\mathbf{X}_{(j)} = \mathbf{X}_{(j)}(r)$ denote $(k_{j+1} - k_j) \times (r + 1)$-dimensional matrix with $(\mathbf{x}_t^\top, 1)$, $k_j \leq t \leq k_{j+1} - 1$ as its rows where $\mathbf{x}_t = \mathbf{x}_t(r) = (X_t, \ldots, X_{t-r+1})^\top$, and let $\mathbf{Y}_{(j)} = (X_{k_j+1}, \ldots, X_{k_{j+1}})^\top$. Then, $\widehat{\boldsymbol{\beta}}_{(j)} = \widehat{\boldsymbol{\beta}}_{(j)}(r) = (\mathbf{X}_{(j)}^\top\mathbf{X}_{(j)})^{-1}\mathbf{X}_{(j)}^\top\mathbf{Y}_{(j)}$ denotes the ordinary least squares (OLS) estimator from the j-th segment, and it is partitioned into $\widehat{\boldsymbol{\beta}}_{(j)} = (\widehat{\boldsymbol{\alpha}}_{(j)}(r)^\top, \widehat{\mu}_j)^\top$. With $\widehat{\boldsymbol{\beta}}_{(j)}$, $j = 0, \ldots, m$, we set

$$\widehat{\boldsymbol{\alpha}}(r) = \widehat{\boldsymbol{\alpha}}_{(j^*)}(r) \quad \text{and} \quad \widehat{\boldsymbol{\beta}}(\mathcal{A}, r) = (\widehat{\boldsymbol{\alpha}}(r)^\top, \widehat{\mu}_0, \ldots, \widehat{\mu}_m)^\top, \tag{11}$$

where $j^* = \arg\max_{0\leq j\leq m}(k_{j+1} - k_j)$, i.e. the index of the longest segment defined by $\mathcal{A}$. In other words, the effective mean level over each segment $[k_j + 1, k_{j+1}]$ is estimated locally from the observations therein, and the AR parameters are estimated from the longest segment. In doing so, estimation errors of the time-varying levels are related to the localisation rates of the corresponding change point estimators, while the estimation error of the AR parameters is the best attainable among those of the local estimators $\widehat{\boldsymbol{\alpha}}_{(j)}(r)$ provided that r is set adequately, an aspect we cover below. We prefer this approach over estimating the AR parameters globally for

the ease of theoretical analysis, but we use the global approach in the practical implementation as motivated in Appendix C. This performs well in practice, see Section 4 and Appendix D.

Since the AR order p is typically unknown, we integrate its selection procedure in estimation as follows: AR models of varying orders, $r \in \{0, \ldots, p_{\max}\}$ with a fixed upper bound $p_{\max} \geq p$, are fitted to the j^*-th segment (recall that $j^* = \arg\max_{0 \leq j \leq m}(k_{j+1} - k_j)$), such that p is estimated from the data by

$$\widehat{p} = \widehat{p}(\mathcal{A}) = \arg\min_{r \in \{0,\ldots,p_{\max}\}} \mathrm{SC}\left(\{X_t\}_{t=k_{j^*}+1}^{k_{j^*+1}}, \emptyset, r\right), \quad \text{where}$$

$$\mathrm{SC}\left(\{X_t\}_{t=k_{j^*}+1}^{k_{j^*+1}}, \emptyset, r\right) = \frac{(k_{j^*+1} - k_{j^*})}{2} \log\left(\frac{\|\mathbf{Y}_{(j^*)} - \mathbf{X}_{(j^*)}(r)\widehat{\boldsymbol{\beta}}_{(j^*)}(r)\|^2}{k_{j^*+1} - k_{j^*}}\right) + r\xi_n. \tag{12}$$

3.2 gSC: sequential model selection using the SC

We first narrow down the model selection problem to that of determining between a given change point model $\mathcal{A}$ and the null model without any change points, and describe how the proposed SC is adopted for the purpose.

The AR parameters are well-estimated by $\widehat{\boldsymbol{\alpha}}(\widehat{p})$ given in (11)–(12), whether the mean remains constant or not, provided that their number and locations are consistently estimated by some subset of $\mathcal{A}$ (in the sense made clear in Assumption 3.2 below). Therefore, the proposed criterion $\mathrm{SC}(\{X_t\}_{t=1}^n, \mathcal{A}, \widehat{p})$ gives a suitable indicator of the goodness-of-fit of the change point model $\mathcal{A}$ offset by the increased model complexity. On the other hand, if any shift in the mean is ignored in fitting an AR model, the resultant coefficient estimators are biased and, consequently, SC evaluated at the null model as proposed in Section 3.1 is unreliable in such a situation. Instead, we propose to compare

$$\mathrm{SC}_0\left(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(\widehat{p})\right) := \frac{n}{2} \log\left(\frac{\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})\left(\mathbf{Y} - \mathbf{L}(\widehat{p})\widehat{\boldsymbol{\alpha}}(\widehat{p})\right)\|^2}{n}\right) + \widehat{p}\,\xi_n$$

against $\mathrm{SC}(\{X_t\}_{t=1}^n, \mathcal{A}, \widehat{p})$, where $\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1}$ denotes the projection matrix removing the sample mean from the right-multiplied vector. By having the plug-in estimator $\widehat{\boldsymbol{\alpha}}(\widehat{p})$ in its definition, SC_0 avoids the above-mentioned difficulty arising when evaluating the SC at a model underestimating the number of change points. We conclude that the data is well-described by the change point model $\mathcal{A}$ if

$$\mathrm{SC}_0\left(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(\widehat{p})\right) > \mathrm{SC}(\{X_t\}_{t=1}^n, \mathcal{A}, \widehat{p}), \tag{13}$$

and if the converse holds, we prefer the null model over the change point model.

This SC-based model selection strategy is extended to be applicable with a sequence of nested change point models $\emptyset = \widehat{\Theta}_0 \subset \widehat{\Theta}_1 \subset \ldots \subset \widehat{\Theta}_M$, such as that obtained in (4) by the WEM principle, even when $M > 1$. Referred to as the gSC in the remainder of the paper,

the proposed methodology performs a backward search along the sequence from the largest model $\widehat{\Theta}_l$ with $l = M$, sequentially evaluating whether the reduction in the goodness of fit (i.e. increase in the residual sum of squares) by moving from $\widehat{\Theta}_l$ to $\widehat{\Theta}_{l-1}$, is sufficiently offset by the decrease in model complexity. More specifically, let $s, e \in \widehat{\Theta}_{l-1} \cup \{0, n\}$ denote two candidates satisfying $(s, e) \cap \widehat{\Theta}_{l-1} = \emptyset$, and suppose that $\mathcal{A} = (s, e) \cap (\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1})$ is not empty (by definition, $\{s, e\} \subset \widehat{\Theta}_l \cup \{0, n\}$). In other words, $\mathcal{A}$ contains candidate estimators detected within the local environment (s, e), which appear in $\widehat{\Theta}_l$ but do not appear in the subsequent smaller models $\widehat{\Theta}_{l'}$, $l' \leq l - 1$. Then, we compare $\mathrm{SC}(\{X_t\}_{t=s+1}^e, \mathcal{A}, \widehat{p}_{s:e})$ against $\mathrm{SC}_0(\{X_t\}_{t=s+1}^e, \widehat{\boldsymbol{\alpha}}_{s:e}(\widehat{p}_{s:e}))$ as in (13), with the AR parameter estimator $\widehat{\boldsymbol{\alpha}}_{s:e}(\widehat{p}_{s:e})$ and its dimension $\widehat{p}_{s:e}$ obtained locally as described in (11)–(12) using the longest interval determined by $\mathcal{A}$ within (s, e). If $\mathrm{SC}(\{X_t\}_{t=s+1}^e, \mathcal{A}, \widehat{p}_{s:e}) < \mathrm{SC}_0(\{X_t\}_{t=s+1}^e, \widehat{\boldsymbol{\alpha}}_{s:e}(\widehat{p}_{s:e}))$, the change point estimators in $\mathcal{A}$ are deemed important; if this is the case for all estimators in $\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}$, we return $\widehat{\Theta}_l$ as the final model. If not, we update $l \leftarrow l - 1$ and repeat the same procedure until some $\widehat{\Theta}_l$, $l \geq 1$, is selected as the final model, or the null model $\widehat{\Theta}_0 = \emptyset$ is reached. The full algorithmic description of the gSC is provided in Appendix A.2.

In summary, the gSC methodology does not directly minimise SC but starting from the largest model, searches for the first largest model $\widehat{\Theta}_l$ where all candidate estimators in $\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}$ are deemed important as described above. Fang and Siegmund (2020) note that the bias in AR parameter estimation under (8) due to multiple shifts in the mean, adversely affects the model selection consistency. In view of this, the proposed backward approach is particularly advantageous by avoiding the evaluation of SC at a model that under-estimates the number of change points.

3.3 Theoretical properties

For the theoretical analysis of gSC, we make a set of assumptions and remark on their relationship to those made in Section 2.2. Assumption 3.1 is imposed on the stochastic part of model (8).

Assumption 3.1. (i) The characteristic polynomial $a(z) = 1 - \sum_{i=1}^p a_i z^i$ has all of its roots outside the unit circle $|z| = 1$.

(ii) $\{\varepsilon_t\}$ is an ergodic and stationary martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{F}_t$, such that ε_t and X_t are $\mathcal{F}_t$-measurable and $\mathsf{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$.

(iii) There exists some $\Delta > 0$ such that $\sup_t \mathsf{E}(|\varepsilon_t|^{2+\Delta}|\mathcal{F}_{t-1}) < \infty$ a.s.

(iv) Let $\mathsf{P}(\mathcal{E}_n) \to 1$ with ω_n satisfying $\sqrt{\log(n)} = O(\omega_n)$ and $\omega_n^2 = O(\min_{1 \leq j \leq q} \delta_j)$, where

$$\mathcal{E}_n = \Big\{ \max_{0 \leq s < e \leq n} (e - s)^{-1/2} \Big| \sum_{t=s+1}^e \varepsilon_t \Big| \leq \omega_n \Big\}.$$

Assumption 3.1 (i)–(iii) are taken from Lai and Wei (1982a,b, 1983), where the strong consistency in stochastic regression problems is established. In particular, Assumption 3.1 (i) indicates that $\{Z_t\}_{t=1}^n$ is a short-memory linear process. The bound in Assumption 3.1 (iv) is related to the detectability of change points, and gives a lower bound on the penalty parameter ξ_n of SC, see Assumption 3.4. Theorem 1.2A of De la Peña (1999) derives a Bernstein-type inequality for a martingale difference sequence satisfying $\mathsf{E}|\varepsilon_t|^k \leq (k!/2)c_\varepsilon^k \mathsf{E}(\varepsilon_t^2)$ for all $k \geq 3$ and some $c_\varepsilon \in (0, \infty)$, from which we readily obtain $\omega_n \asymp \log(n)$. Under a more stringent condition that $\{\varepsilon_t\}$ is a sequence of i.i.d sub-Gaussian random variables, it suffices to set $\omega_n \asymp \sqrt{\log(n)}$ (e.g. see Proposition 2.1 (a) of Cho and Kirch (2020b)).

Remark 3.1 (Links between Assumptions 2.1 and 3.1). Assumption 2.1 does not impose any parametric condition on the dependence structure of $\{Z_t\}_{t=1}^n$. For linear, short memory processes (implied by Assumption 3.1 (i)), Peligrad and Utev (2006) show that the invariance principle for the linear process is inherited from that of the innovations at no extra cost. Then, as discussed in Remark 2.2, a logarithmic bound $\omega_n \asymp \log^\kappa(n)$ follows from $\sum_{t=1}^\ell \varepsilon_t - W(\ell) = O(\log^{\kappa'}(n))$ for some $\kappa' \in [1, \kappa)$, which in turn leads to $\zeta_n \asymp \omega_n$. In view of Assumptions 2.1 and 2.2, the condition that $\omega_n^2 = O(\min_{1\leq j\leq q} \delta_j)$ is a mild one.

We impose the following assumption on the nested model sequence $\widehat{\Theta}_0 \subset \ldots \subset \widehat{\Theta}_M$, where $\widehat{\Theta}_l = \{\widehat{\theta}_{l,j},\, 1 \leq j \leq \widehat{q}_l : \widehat{\theta}_{l,1} < \ldots < \widehat{\theta}_{l,\widehat{q}_l}\}$ for $l \geq 1$.

Assumption 3.2. Let $\mathcal{M}_n$ denote the following event: for a given penalty ξ_n, we have $\xi_n^{-1} \min_{0\leq j\leq \widehat{q}_M}(\widehat{\theta}_{M,j+1} - \widehat{\theta}_{M,j}) = o(1)$ and $\widehat{q}_M = |\widehat{\Theta}_M|$ is fixed for all n. Additionally, under $H_1 : q \geq 1$, there exists $l^* \in \{1, \ldots, M\}$ such that

$$\widehat{q}_{l^*} = q \quad \text{and} \quad \max_{1\leq j\leq q} d_j^2 \left|\widehat{\theta}_{l^*,j} - \theta_j\right| \leq \rho_n \tag{14}$$

for some ρ_n satisfying $(\min_{1\leq j\leq q} d_j^2 \delta_j)^{-1}\rho_n \to 0$. Then, $\mathsf{P}(\mathcal{M}_n) \to 1$.

By Theorem 2.1, we have Assumption 3.2 satisfied by the model sequence generated by the WEM principle with either the LD or the DC approach (see (5)–(6)) with $\rho_n \asymp \zeta_n^2$. We state this result as an assumption so that if gSC were to be applied with an alternative solution path algorithm, our results would be directly applicable if the latter satisfied Assumption 3.2. Since the serial dependence structure is learned from the data by fitting an AR model to each segment, the requirement on the minimum spacing of the largest model $\widehat{\Theta}_M$ is a natural one and it may be hard-wired into the solution path generation.

Assumption 3.3 is on the effective size of changes under (8), and Assumption 3.4 on the choice of the penalty parameter ξ_n. In particular, the choice of ξ_n connects the detectability of change points with the level of noise remaining in the data after accounting for the autoregressive dependence structure.

Assumption 3.3. $\max_{1\leq j\leq q} |d_j| = O(1)$ and $D_n := \min_{1\leq j\leq q} d_j^2\, \delta_j \to \infty$ as $n \to \infty$.

Assumption 3.4. ξ_n satisfies $D_n^{-1}\xi_n = o(1)$ and $\xi_n^{-1}\max(\omega_n^2, \rho_n) = o(1)$.

By Assumption 3.1 (i), the effective change size d_j is of the same order as f_j' since $d_j = (1 - \sum_{i=1}^p a_i)f_j'$. Therefore, Assumption 3.3 on the detection lower bound formulated with d_j, together with Assumption 3.4, is closely related to Assumption 2.2 formulated with f_j'. In fact, we can select ξ_n such that Assumption 3.4 follows immediately from Assumption 2.2 when gSC is applied in combination with the sequence of nested models generated by the WEM principle, recalling that the rate of localisation attained by the latter is $\rho_n \asymp \zeta_n^2$ and $\omega_n = O(\zeta_n)$.

Theorem 3.1. Let Assumptions 3.1–3.4 hold. Then, on $\mathcal{E}_n \cap \mathcal{M}_n$, the gSC methodology returns $\widehat{\Theta} = \{\widehat{\theta}_j,\, 1 \le j \le q : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}}\}$ satisfying

$$\widehat{q} = q \quad \text{and} \quad \max_{1 \le j \le q} d_j^2 \left|\widehat{\theta}_j - \theta_j\right| \le \rho_n$$

for n large enough.

Theorem 3.1 establishes that gSC achieves model selection consistency. Together, Theorems 2.1–3.1 lead to the consistency of WEM.gSC. Once the number of change points and their locations are consistently estimated, we can further improve the location estimators in $\widehat{\Theta}$; Appendix B discusses a simple refinement procedure which achieves the minimax optimal localisation rate.

4 Nitrogen oxides concentrations in London

Appendix C discusses in detail the choice of the tuning parameters for WEM.gSC. In Appendix D, we provide extensive simulation studies where WEM.gSC, combined with the default choice of tuning parameters, is shown to perform well in comparison with DepSMUCE (Dette et al., 2020), DeCAFS (Romano et al., 2020b) and MACE (Wu and Zhou, 2020). In particular, WEM.gSC is shown not to return false positives in the absence of mean shifts, and attains good power and localisation accuracy for a variety of change point configurations and serial dependence scenarios. In this section, we further demonstrate the good performance of WEM.gSC on London air quality data.

NO_x is a generic term for the nitrogen oxides that are the most relevant for air pollution, namely nitric oxide (NO) and nitrogen dioxide (NO_2). The main anthropogenic sources of NO_x are mobile and stationary combustion sources, and its acute and chronic health effects have been well-documented (Kampa and Castanas, 2008). We analyse the daily average concentrations of NO_2 and NO_x measured (in $\mu g/m^3$) at Marylebone Road in London, U.K., from September 1, 2000 to September 30, 2020; the datasets are retrieved from Defra (`https://uk-air.defra.gov.uk/`). The concentration measurements are positive integers and exhibit seasonality and weekly patterns as well as distinguished behaviour on bank holidays, since road

traffic is the principal outdoor source of NO_x in a busy London road. To correct for possible heavy-tailedness of the raw measurements, we take the square root transform and further remove seasonal and weekly trends and bank holiday effects from the transformed data using a model fitted on the observations from January 2004 to December 2010; for details of the pre-processing steps, see Appendix E.1. The resulting time series are plotted in Figure 1, where it is also seen that the thus-transformed data exhibit persistent autocorrelations.

We analyse the transformed time series from NO_2 and NO_x concentrations for change points in the level, with the tuning parameters for WEM.gSC chosen as recommended in Appendix C apart from M, the number of candidate models considered; given the large number of observations ($n = 7139$), we allow for $M = 10$ instead of the default choice $M = 5$. The change points detected by WEM.gSC with the LD for model sequence generation, are plotted in Figure 1. For comparison, we also report the change points estimated by DepSMUCE and DeCAFS, see Table 1.

From the NO_2 concentrations, WEM.gSC detects different sets of estimators depending on whether the LD or the DC approaches were adopted for model sequence generation. In Appendix E.2, we validate the set of change point estimators detected by WEM.gSC(LD) from the NO_2 time series (by attempting to remove the bulk of serial dependence from the data and then applying an existing procedure for change point detection for uncorrelated data), based on which we conclude that WEM.gSC(DC) possibly under-estimates the number of change points on this dataset. On the other hand, from the NO_x concentrations, WEM.gSC produces identical sets of change point estimators regardless of the choice of the model sequence generating methods. Figure 1 shows that a good deal of autocorrelations remain in the data after removing the estimated mean shifts, but the persistent autocorrelations are no longer observed, which supports the hypothesis that NO_2 and NO_x concentrations undergo changes in their levels over the period in consideration.

In February 2003, a programme of traffic management measures was introduced in central London including the installation of particulate traps on most London buses and other heavy duty diesel vehicles, which convert NO in the exhaust stream to NO_2 and thus bring in the increase of primary NO_2 emissions from such vehicles (Air Quality Expert Group, 2004). This accounts for the prominent increase in the concentration of NO_2 detected around January 2003 by WEM.gSC (also by DepSMUCE and DeCAFS) which, however, is not observed from NO_x, since the latter contains the combined concentrations of NO and NO_2. The two series share the common change point detected at the end of March 2019 (not detected by DepSMUCE or DeCAFS). The Ultra Low Emission Zone in central London was launched on 8 April 2019, which includes Marylebone Road where the measurements were taken. and its introduction coincides with the decline in the concentrations of both NO_2 and NO_x. Another common change point is detected on March 18, 2020 (also detected by DepSMUCE and DeCAFS) which confirms that the nation-wide COVID-19 lockdown on March 23, 2020 led to the substantial

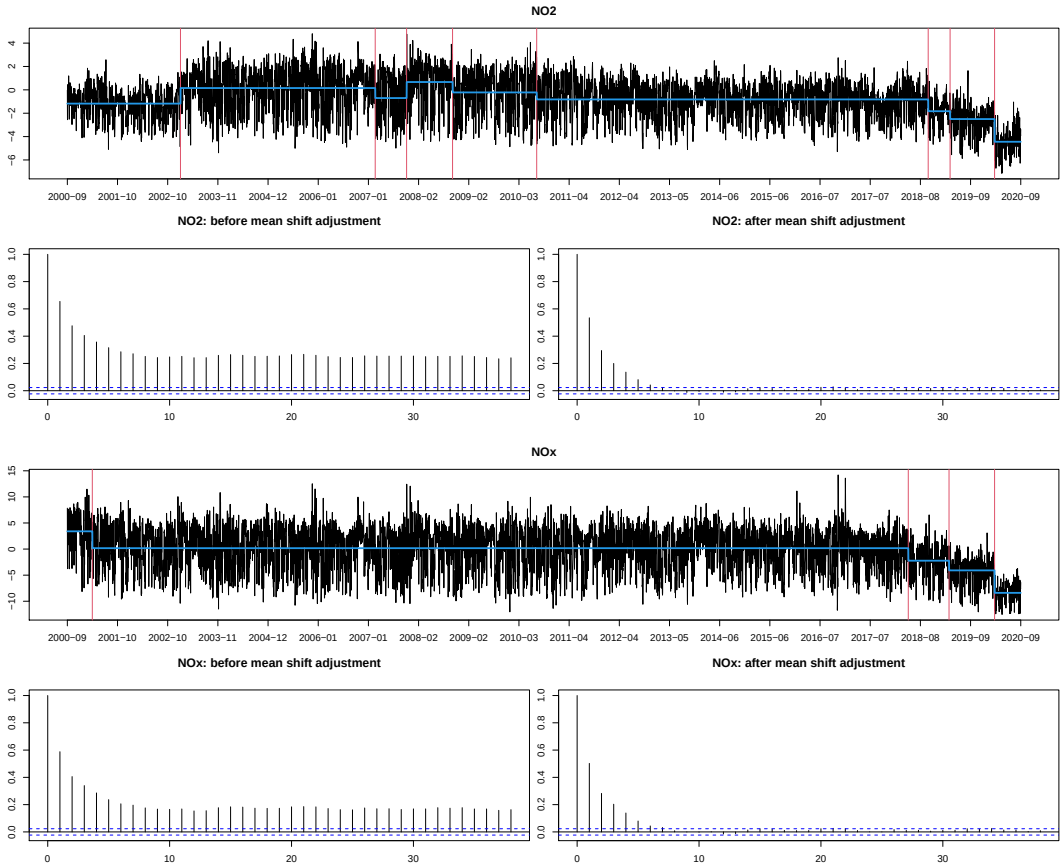

Figure 1: First (third) panel: daily average concentrations of NO_2 (NO_x) after transformation and de-trending, plotted together with the change points detected by WEM.gSC(LD) (vertical lines) and estimated piecewise constant mean (bold lines). Second (fourth) panel: autocorrelation function of transformed and de-trended NO_2 (NO_x) without (left) and with (right) the time-varying mean adjusted.

Table 1: Change points detected from the daily average concentrations of NO_2 and NO_x measured at Marylebone Road in London from September 1, 2000 to September 30, 2020. Any location estimators commonly detected from both NO_2 and NO_x concentrations (within 10 days from one another) are highlighted in bold. For DepSMUCE, parameterised by the significance level α, identical estimators are returned with either of $\alpha \in \{0.05, 0.2\}$.

Method	NO_2	NO_x
WEM.gSC(DC)	2003-01-31, 2010-07-25, **2019-03-30**, **2020-03-18**	2001-03-15, 2018-05-13, **2019-03-22**, **2020-03-18**
WEM.gSC(LD)	2003-01-31, 2007-03-17, 2007-11-15, 2008-10-26, 2010-07-25, 2018-10-13, **2019-03-30**, **2020-03-18**	2001-03-15, 2018-05-13, **2019-03-22**, **2020-03-18**
DepSMUCE	2003-01-31, 2010-07-25, 2018-10-14, **2020-03-18**	2001-03-15, 2018-05-13, **2020-03-18**
DeCAFS	2003-02-05, **2005-12-11**, **2005-12-17** 2007-04-25, 2007-05-05, 2007-12-10 2008-03-03, 2008-03-04, 2009-09-08 2009-09-20, 2012-10-20, 2012-10-27 2018-10-14, **2020-03-18**	2001-11-07, 2001-11-09, 2005-12-08 **2005-12-11**, **2005-12-17**, 2008-12-06 2008-12-08, 2018-05-13, **2020-03-18**

reduction of NO_x levels across the country (Higham et al., 2020).

Acknowledgements

We thank Paul Fearnhead and Gaetano Romano for their constructive comments. Haeran Cho was supported by the Leverhulme Trust Research Project Grant (RPG-2019-390).

References

Air Quality Expert Group (2004). Nitrogen dioxide in the United Kingdom. `https://uk-air.defra.gov.uk/library/assets/documents/reports/aqeg/nd-chapter2.pdf`. Accessed: 2020-11-04.

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*.

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, 34:1–16.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:649–672.

Bardet, J.-M., Kengne, W., and Wintenberger, O. (2012). Multiple breaks detection in general

causal time series using penalized quasi-likelihood. *Electronic Journal of Statistics*, 6:435–477.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22:455–470.

Berkes, I., Horváth, L., Kokoszka, P., Shao, Q.-M., et al. (2006). On discriminating between long-range dependence and changes in mean. *The Annals of Statistics*, 34:1140–1165.

Berkes, I., Liu, W., and Wu, W. B. (2014). Komlós-Major-Tusnády approximation under dependence. *The Annals of Probability*, 42:794–817.

Boysen, L., Liebscher, V., Munk, A., and Wittich, O. (2007). Scale space consistency of piecewise constant least squares estimators–another look at the regressogram. In *Asymptotics: Particles, Processes and Inverse Problems*, pages 65–84. Institute of Mathematical Statistics.

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2017). A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli*, 23:1408–1447.

Chan, K. W. and Yau, C. Y. (2017). High-order corrected estimator of asymptotic variance with optimal bandwidth. *Scandinavian Journal of Statistics*, 44:866–898.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). Group LASSO for structural break time series. *Journal of the American Statistical Association*, 109:590–599.

Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, 10:2000–2038.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22:207–229.

Cho, H. and Fryzlewicz, P. (2015). Multiple change-point detection for high-dimensional time series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:475–507.

Cho, H. and Kirch, C. (2020a). Discussion of 'Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection'. *Journal of the Korean Statistical Society*, 49:1076–1080.

Cho, H. and Kirch, C. (2020b). Two-stage data segmentation permitting multiscale change points, heavy tails and dependence. *arXiv preprint arXiv: 1910.12486*.

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

Davis, R., Lee, T., and Rodriguez-Yam, G. (2006). Structural break estimation for nonstationary time series. *Journal of the American Statistical Association*, 101:223–239.

Davis, R., Lee, T., and Rodriguez-Yam, G. (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29:834–867.

De la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27:537–564.

den Haan, W. J. and Levin, A. T. (1997). A practitioner's guide to robust covariance matrix estimation. *Handbook of Statistics*, 15:299 – 342.

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

Doukhan, P. and Neumann, M. H. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117:878–903.

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24:526–564.

Fang, X. and Siegmund, D. (2020). Detection and estimation of local signals. *arXiv preprint arXiv:2004.08159*.

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.

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

Fryzlewicz, P. (2014). Wild Binary Segmentation for multiple change-point detection. *The Annals of Statistics*, 42:2243–2281.

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

Fryzlewicz, P. (2020b). Narrowest Significance Pursuit: inference for multiple change-points in linear models. *Preprint*.

Higham, J., Ramírez, C. A., Green, M., and Morse, A. (2020). UK COVID-19 lockdown: 100 days of air pollution reduction. *Air Quality, Atmosphere & Health*, pages 1–8.

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

Hušková, M. and Kirch, C. (2010). A note on studentized confidence intervals for the change-point. *Computational Statistics*, 25:269–289.

Hušková, M. and Slabý, A. (2001). Permutation tests for multiple changes. *Kybernetika*, 37:605–622.

Kampa, M. and Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution*, 151:362–367.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:1590–1598.

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

Korkas, K. K. and Fryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, 27:287–311.

Kovács, S., Li, H., Bühlmann, P., and Munk, A. (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633*.

Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing B-valued random variables. *The Annals of Probability*, pages 1003–1036.

Kühn, C. (2001). An estimator of the number of change points based on a weak invariance principle. *Statistics & Probability Letters*, 51:189–196.

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.

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21:33–59.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27:519–533.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151:435–474.

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

Norwood, B. and Killick, R. (2018). Long memory and changepoint models: a spectral classification procedure. *Statistics and Computing*, 28:291–302.

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.

Peligrad, M. and Utev, S. (2006). Invariance principle for stochastic processes with short memory. In *High Dimensional Probability, IMS Lecture Notes Monograph Series*, volume 51, pages 18–32. Institute of Mathematical Statistics.

Perron, P. (2006). Dealing with structural breaks. *Palgrave Handbook of Econometrics*, 1:278–352.

Pešta, M. and Wendler, M. (2020). Nuisance parameters free changepoint detection in non-stationary series. *TEST*, 29(2):379–408.

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.

Robbins, M., Gallagher, C., Lund, R., and Aue, A. (2011). Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32:498–511.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2020a). *DeCAFS: Detecting Changes in Autocorrelated and Fluctuating Signals*. R package version 3.1.5.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2020b). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *arXiv preprint arXiv:2005.01379v1*.

Safikhani, A. and Shojaie, A. (2020). Joint structural break detection and parameter estimation in high-dimensional non-stationary VAR models. *To appear in Journal of the American Statistical Association*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

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

Tecuapetla-Gómez, I. and Munk, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: a difference-based approach. *Scandinavian Journal of Statistics*, 44:346–368.

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.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. (2019). Sub-weibull distributions: generalizing sub-gaussian and sub-exponential properties to heavier-tailed distributions. *arXiv preprint arXiv:1905.04955*.

Wang, D., Yu, Y., Rinaldo, A., and Willett, R. (2019). Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*.

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*.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6:181–189.

Yau, C. Y. and Zhao, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:895–916.

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 (2020a).

We remark that WBS2 as defined in Fryzlewicz (2020a) 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.

Algorithm 1: Wild Binary Segmentation 2

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

Function 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 $\widetilde{R}$ intervals 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** $(s_\circ, k_\circ, e_\circ, |\mathcal{X}_{s_\circ,k_\circ,e_\circ}|) \cup$ wbs2$(\{X_t\}_{t=1}^n, R_n, s, k_\circ) \cup$ wbs2$(\{X_t\}_{t=1}^n, R_n, k_\circ, e)$

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

Output: $\mathcal{P}_0$

A.2 The gSC 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 $l = M$, the gSC algorithm performs the following steps.

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

points detected in $\widehat{\Theta}_l$ such that $(\widehat{\theta}_{l-1,u}, \widehat{\theta}_{l-1,u+1}) \cap (\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}) \neq \emptyset$. 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', $\mathbf{F} \in \{\texttt{TRUE}, \texttt{FALSE}\}^{q_l'}$, initialised as $\mathbf{F} = (\texttt{TRUE}. \ldots, \texttt{TRUE})$.

Step 1.1: Setting $\mathcal{A} = (\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1}) \cap \widehat{\Theta}_l$, obtain $\widehat{p} \in \{0, \ldots, p_{\max}\}$ that returns the smallest SC over the longest local interval defined by $\mathcal{A}$ within $(\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$ as outlined in (12), and the corresponding AR parameter estimator $\widehat{\boldsymbol{\alpha}}(\widehat{p})$ as given in (11).

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

Step 2: If some elements of $\mathbf{F}$ satisfy $F_v = \texttt{TRUE}$ and $l > 1$, update $l \leftarrow l - 1$ and go to Step 1. If $F_v = \texttt{FALSE}$ 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 $F_v = \texttt{FALSE}$ for all $v = 1, \ldots, q_l'$ when the corresponding $\widehat{\Theta}_l = \widehat{\Theta}[q]$, or $F_v = \texttt{TRUE}$ for all v when $\widehat{\Theta}_l$ contains spurious estimators. In implementing the methodology, we take a more 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 $\widehat{\theta}_j$, $1 \leq j \leq q$, is sufficiently close to the corresponding change point θ_j in the sense that $|\widehat{\theta}_j - \theta_j| \leq (f_j')^{-2}\rho_n \leq c\delta_j$ 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.

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

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

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 $1 \leq a_n \leq \min_{1\leq j\leq q}(f'_j)^2\delta_j$ 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

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

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

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

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

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

for some constant $C > 0$ and $\nu > 2$, see Proposition 2.1 (c.ii) of Cho and Kirch (2020b). 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 Fromont et al. (2020)).

C Implementation and the choice of tuning parameters

In line with the condition (7) 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. For simplicity, we directly utilise the OLS estimator $\boldsymbol{\beta}$ obtained from regressing $\mathbf{Y}$ on $\mathbf{X}$ in (10) as $\widehat{\boldsymbol{\beta}}(\mathcal{A}, r)$ in the implementation of the gSC methodology, in place of the segment-wise estimation strategy described in (11); in our numerical experiments, this modification did not alter the results greatly. 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 Simulation studies

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.

(M1) 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 $Z_t = \varepsilon_t$.

(M2) 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 $\sigma_\varepsilon = 1/2.14285$.

(M3) 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 $\sigma_\varepsilon = \sqrt{1 - a_1^2}$.

(M4) 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 $\sigma_\varepsilon = \sqrt{(1 - a_1^2)/(1 + a_1 b_1 + b_1^2)}$.

(M5) 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 $b_1 = 0.3$.

(M6) As in (M5) but with $b_1 = -0.9$.

(M7) f_t undergoes $q = 5$ change points as in (M1) 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: $Z_t = \varepsilon_t + 0.9\varepsilon_{t-1} + 0.8\varepsilon_{t-2} + 0.7\varepsilon_{t-3} + 0.6\varepsilon_{t-4}$.

(M8) f_t undergoes $q = 5$ change points θ_j as in (M1) with $n = 1000$ and $(f_0, f_1', f_2', f_3', f_4', f_5') = (0, 5, -3, 6, -7, -3)$, and $\{Z_t\}$ follows an ARMA(2, 6) model: $Z_t = 0.75Z_{t-1} - 0.5Z_{t-2} + \varepsilon_t + 0.8\varepsilon_{t-1} + 0.7\varepsilon_{t-2} + 0.6\varepsilon_{t-3} + 0.5\varepsilon_{t-4} + 0.4\varepsilon_{t-5} + 0.3\varepsilon_{t-6}$.

(M9) 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_{θ_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 (M3) with $a_1 = 0.5$.

(M10) As in (M9) but with $a_1 = 0.9$.

(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 AR(2, 6) model as in (M8).

(M12) f_t is as in (M1) 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 (M1) 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 (M1), all others model have serial correlations in $\{Z_t\}_{t=1}^n$. Models (M2) (motivated by an example in Wu and Zhou (2020)), (M3) and (M4) consider relatively short time series with $n \in [150, 300]$. Models (M5), (M7) and (M8) are taken from Dette et al. (2020). In (M6), the LRV is close to zero and thus its accurate estimation is difficult. Models (M9)–(M10) have a teeth-like signal containing frequent change points and the underlying $\{Z_t\}$ has strong autocorrelations in (M10), and (M11) considers frequent, heterogeneous changes in the mean. In Models (M12)–(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 WEM.gSC with the sequence of nested models generated as described in Section 2.1, either directly identifying the largest differences ('LD') in the ordered max-CUSUMs as in (5), or by examining the double CUSUM statistics ('DC') as in (6); the rest of the tuning parameters are selected as described in Appendix C except that for Model (M3), we set $p_{\max} = 5$ to account for the relative shortness of the time series.

For comparison, DepSMUCE (Dette et al., 2020), DeCAFS (Romano et al., 2020b) and MACE (Wu and Zhou, 2020) are applied to the same datasets. 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. 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; advised by the authors, we have adapted the main routine of its R implementation (Romano et al., 2020a) to change point analysis under (1). For DepSMUCE and MACE, we consider $\alpha \in \{0.05, 0.2\}$ since, in our experience, higher values of α lead to inadequate performance when there are no change points present. 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. We report the MSE and d_H computed with the refined estimators as described in Appendix B for WEM.gSC.

Overall, across the various scenarios, WEM.gSC 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 WEM.gSC 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 (M3) 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, WEM.gSC 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. Between the two gap identification methods, DC performs marginally better than LD in identifying the correct number of change points except for the case of (M10) and (M11). In particular, (M11) contains heterogeneous shifts in the mean and, as seen in Figure D.1, LD is better suited than DC for identifying the gap between those log-CUSUMs due to smaller jumps towards the end of the signal, and those due to time series fluctuations. The results under (M12)–(M13) show that WEM.gSC is able to handle mild nonstationarities in $\{Z_t\}_{t=1}^n$.

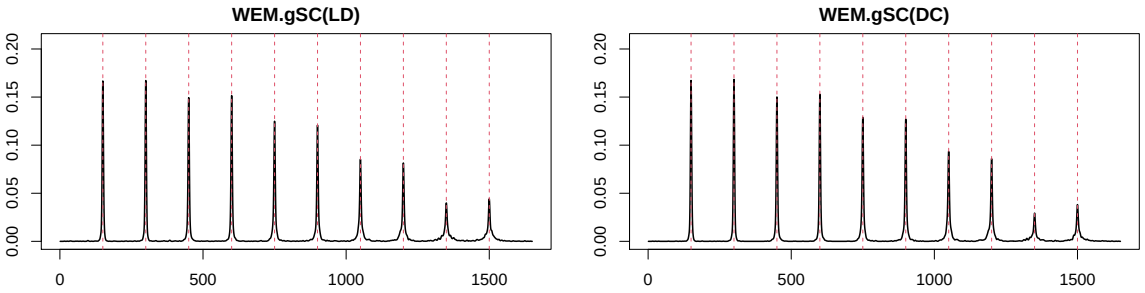

Figure D.1: Density of change point estimators returned by WEM.gSC(LD) (left) and WEM.gSC(DC) (right) over 1000 realisations generated under (M11) with the true locations of θ_j denoted by the vertical broken lines.

DepSMUCE performs well for short series (see (M3)) or in the presence of weak serial correlations as in (M5), 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 (M6), or when there are strong autocorrelations as in (M10), thus incurring a large number of falsely detected change points under H_0.

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 (M10) 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 (M7) and (M8) when compared to WEM.gSC, 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: (M1)–(M13): 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 and MACE, and at 0.05 for WEM.gSC 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)	WEM.gSC(LD)	**0.009**	0.000	0.000	0.000	0.978	0.019	0.003	0.000	4.940	8.866
	WEM.gSC(DC)	**0.000**	0.000	0.000	0.002	**0.994**	0.003	0.001	0.000	4.881	**7.892**
	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	**0.014**	0.000	0.000	0.000	0.979	0.019	0.002	0.000	**4.847**	8.172
	MACE(0.05)	0.573	0.669	0.250	0.065	0.014	0.002	0.000	0.000	96.871	310.478
	MACE(0.2)	0.810	0.376	0.336	0.204	0.069	0.013	0.002	0.000	83.401	255.7
(M2)	WEM.gSC(LD)	0.102	0.000	0.000	0.000	0.862	0.098	0.029	0.011	2.826	5.475
	WEM.gSC(DC)	0.080	0.000	0.000	0.000	0.884	0.086	0.015	0.015	2.753	4.583
	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	0.099	0.000	0.000	0.000	0.885	0.083	0.022	0.010	**1.910**	5.697
	MACE(0.05)	0.457	0.000	0.007	0.126	0.761	0.104	0.002	0.000	25.329	11.473
	MACE(0.2)	0.713	0.000	0.001	0.042	0.759	0.184	0.014	0.000	13.335	10.988
(M3)	WEM.gSC(LD)	0.074	0.000	0.000	0.000	0.865	0.119	0.016	0.000	5.993	4.782
	WEM.gSC(DC)	0.067	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	0.197	0.000	0.004	0.007	0.749	0.116	0.066	0.058	9.137	9.331
	MACE(0.05)	0.611	0.000	0.148	0.446	0.373	0.033	0.000	0.000	43.102	25.958
	MACE(0.2)	0.781	0.000	0.060	0.321	0.554	0.064	0.001	0.000	30.616	21.736
(M4)	WEM.gSC(LD)	**0.033**	0.000	0.096	0.010	0.832	0.035	0.018	0.009	13.588	9.39
	WEM.gSC(DC)	**0.027**	0.000	0.102	0.001	**0.852**	0.025	0.009	0.011	**13.490**	**7.821**
	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	0.066	0.000	0.103	0.041	0.816	0.030	0.007	0.003	13.885	13.094
	MACE(0.05)	0.305	0.000	0.266	0.281	0.425	0.026	0.002	0.000	60.122	33.958
	MACE(0.2)	0.486	0.000	0.135	0.280	0.522	0.057	0.006	0.000	42.749	37.39
(M5)	WEM.gSC(LD)	**0.006**	0.001	0.002	0.013	0.953	0.027	0.004	0.000	5.141	18.921
	WEM.gSC(DC)	**0.000**	0.000	0.000	0.012	**0.972**	0.016	0.000	0.000	5.053	16.36
	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	**0.019**	0.000	0.004	0.004	0.968	0.023	0.000	0.001	**4.956**	**15.819**
	MACE(0.05)	0.588	0.801	0.151	0.038	0.009	0.001	0.000	0.000	63.685	336.515

	MACE(0.2)	0.809	0.519	0.273	0.162	0.035	0.011	0.000	0.000	54.649	286.579
(M6)	WEM.gSC(LD)	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**68.720**	**1.988**
	WEM.gSC(DC)	**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	**0.001**	0.000	0.000	0.000	0.995	0.004	0.001	0.000	68.909	2.002
	MACE(0.05)	0.240	0.000	0.000	0.914	0.086	0.000	0.000	0.000	1729.538	57.587
	MACE(0.2)	0.519	0.000	0.000	0.802	0.190	0.008	0.000	0.000	1723.032	65.264
(M7)	WEM.gSC(LD)	**0.008**	0.041	0.029	0.016	0.899	0.013	0.001	0.001	5.691	46.539
	WEM.gSC(DC)	**0.003**	0.000	0.001	0.003	**0.926**	0.059	0.008	0.003	4.776	**21.35**
	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	0.002	0.059	0.020	0.863	0.055	0.001	0.000	**3.786**	31.441
	MACE(0.05)	0.681	0.768	0.168	0.048	0.015	0.001	0.000	0.000	48.913	331.673
	MACE(0.2)	0.872	0.455	0.279	0.193	0.061	0.011	0.001	0.000	38.975	284.883
(M8)	WEM.gSC(LD)	**0.016**	0.032	0.024	0.062	0.814	0.047	0.012	0.009	5.738	55.998
	WEM.gSC(DC)	**0.001**	0.000	0.000	0.019	**0.873**	0.092	0.014	0.002	4.907	**34.627**
	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	0.010	0.037	0.140	0.766	0.047	0.000	0.000	**4.082**	63.433
	MACE(0.05)	0.694	0.754	0.166	0.065	0.015	0.000	0.000	0.000	40.553	313.2
	MACE(0.2)	*0.875*	0.473	0.275	0.154	0.087	0.009	0.002	0.000	33.796	284.811
(M9)	WEM.gSC(LD)	**0.006**	0.000	0.000	0.000	0.888	0.077	0.022	0.013	2.485	9.01
	WEM.gSC(DC)	**0.000**	0.000	0.000	0.008	**0.982**	0.006	0.003	0.001	2.425	**5.485**
	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	0.978	0.019	0.003	0.000	**2.103**	5.516
	MACE(0.05)	0.905	0.910	0.053	0.028	0.008	0.000	0.001	0.000	61.217	230.091
	MACE(0.2)	0.986	0.620	0.174	0.114	0.052	0.028	0.010	0.002	47.381	175.819
(M10)	WEM.gSC(LD)	**0.009**	0.048	0.027	0.039	0.519	0.186	0.081	0.100	2.805	51.057
	WEM.gSC(DC)	**0.000**	0.087	0.177	0.233	0.319	0.076	0.041	0.067	3.184	86.139
	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	0.567	0.000	0.004	0.039	**0.784**	0.155	0.017	0.001	**1.086**	**16.318**
	MACE(0.05)	1.000	0.052	0.059	0.083	0.126	0.168	0.173	0.339	7.080	125.924
	MACE(0.2)	1.000	0.007	0.006	0.021	0.042	0.089	0.109	0.726	5.741	106.281
(M11)	WEM.gSC(LD)	**0.012**	0.166	0.205	0.032	**0.503**	0.060	0.028	0.006	5.761	183.927
	WEM.gSC(DC)	**0.001**	0.080	0.360	0.252	0.287	0.013	0.006	0.002	5.435	**180.548**
	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	0.076	0.220	0.481	0.060	0.228	0.011	0.000	0.000	**4.829**	285.997
	MACE(0.05)	0.781	0.998	0.002	0.000	0.000	0.000	0.000	0.000	33.582	604.839

	MACE(0.2)	0.942	0.961	0.022	0.012	0.004	0.001	0.000	0.000	30.708	470.998
(M12)	WEM.gSC(LD)	**0.025**	0.045	0.060	0.118	0.686	0.058	0.021	0.012	6.846	88.52
	WEM.gSC(DC)	**0.002**	0.000	0.002	0.061	0.718	0.151	0.048	0.020	5.828	**50.476**
	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	0.079	0.009	0.080	0.081	**0.726**	0.077	0.024	0.003	**5.769**	83.559
	MACE(0.05)	0.685	0.778	0.171	0.045	0.005	0.001	0.000	0.000	33.466	325.439
	MACE(0.2)	0.873	0.537	0.250	0.151	0.049	0.012	0.001	0.000	28.316	304.252
(M13)	WEM.gSC(LD)	**0.018**	0.020	0.028	0.072	0.813	0.052	0.009	0.006	6.135	57.273
	WEM.gSC(DC)	**0.001**	0.000	0.002	0.043	0.831	0.089	0.030	0.005	5.442	**38.565**
	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**	**0.004**	0.043	0.049	**0.840**	0.052	0.012	0.000	**5.105**	51.444
	MACE(0.05)	0.665	0.802	0.143	0.051	0.003	0.001	0.000	0.000	38.599	328.101
	MACE(0.2)	0.855	0.543	0.254	0.142	0.051	0.008	0.002	0.000	32.992	301.252

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 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 1 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

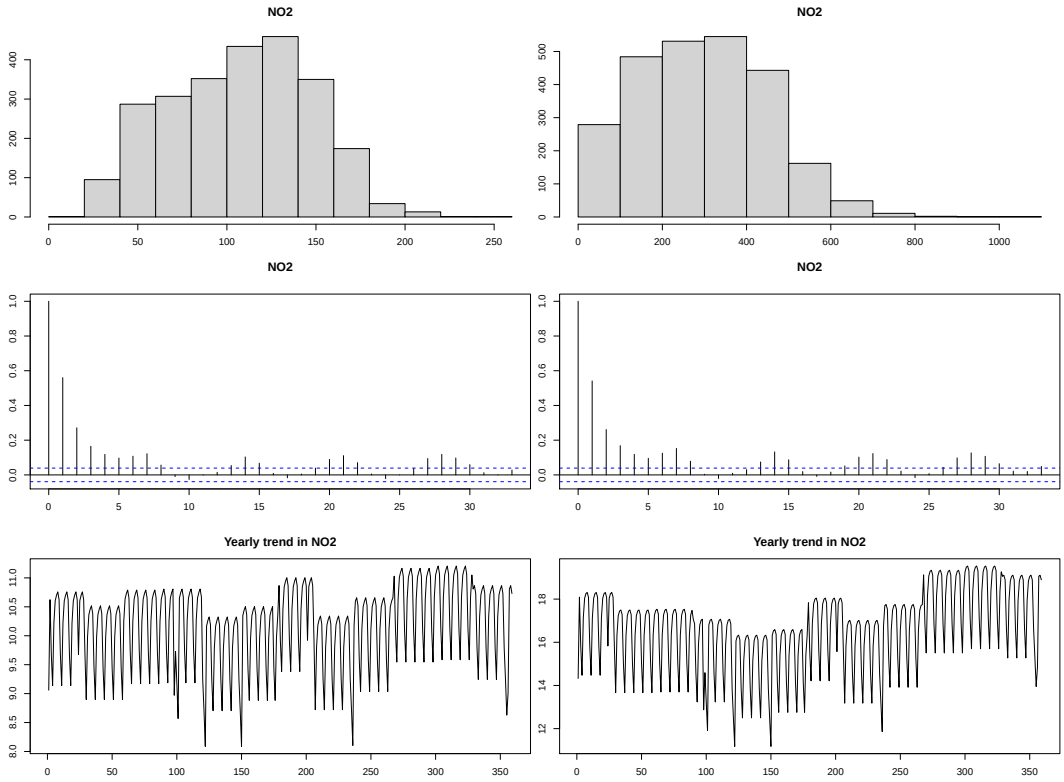

Figure E.1: Various statistical properties of the daily concentrations of NO_2(left) and NO_x (right) measured at Marylebone Road in London between January 2004 and December 2010. Top: histogram of raw concentrations. Middle: autocorrelations after square root transform. Bottom: yearly fitted patterns.

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 with the AR coefficient equal to 0.5 (as suggested by the sample autocorrelation function in Figure E.1), and work with the empirical residuals from this fit.

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 WEM.gSC(LD) method. Out of the 7 locations estimated by IDetect, there is very good agreement with WEM.gSC(LD) for 6 out of these locations. The exception is the

WEM.gSC(LD)-estimated change point at 2010-07-25, which IDetect estimates some 800 days later. However, IDetect also does not estimate the following WEM.gSC(LD)-estimated change point at 2018-10-13, which is a possible reason for IDetect to replace these two WEM.gSC(LD)-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 WEM.gSC(LD) 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 WEM.gSC 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 WEM.gSC 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 Appendix 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 WEM.gSC(DC) and the WEM.gSC(LD) as well as those detected by DepSMUCE and DeCAFS for comparison. We note that the refinement of location estimators described in Appendix B does not alter the results.

On all three datasets, WEM.gSC(LD) and WEM.gSC(DC) return, between them, two nearly identical estimators, and the same change points are detected by DepSMUCE (with $\alpha = 0.2$) and DeCAFS. Figure E.2 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.

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

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
WEM.gSC(DC)	1892, 1988	1892, 1988	1892, 1987
WEM.gSC(LD)	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

matrix $\mathbf{A}$ is denoted by $\mathbf{\Pi}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$, 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

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

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 (7). 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 (2020b))**.** For $\max(s, \theta_{j-1}) < k < \theta_j < \min(e, \theta_{j+1})$, it holds that

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

where $a_+ = a \cdot \mathbb{I}_{a\ge 0}$. Similarly, for $\max(s, \theta_{j-1}) < \theta_j \le k < \min(e, \theta_{j+1})$, 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 $0 \le s < e \le n$ with $e - s > 1$, let $\Theta \cap [s, e] = \{\theta_1^\circ, \dots, \theta_m^\circ\}$ with $m \le q$, and we adopt the notations $\theta_0^\circ = s$ and $\theta_{m+1}^\circ = e$. If the series $\mathcal{F}_{s,k,e}$ is not constantly zero for $\theta_j^\circ + 1 \le k \le \theta_{j+1}^\circ$ for some $j = 0, \dots, m$, one of the following is true:

(i) $j = 0$ and $\mathcal{F}_{s,k,e}$, $\theta_j^\circ + 1 \le k \le \theta_{j+1}^\circ$ does not change sign and has strictly increasing absolute values,

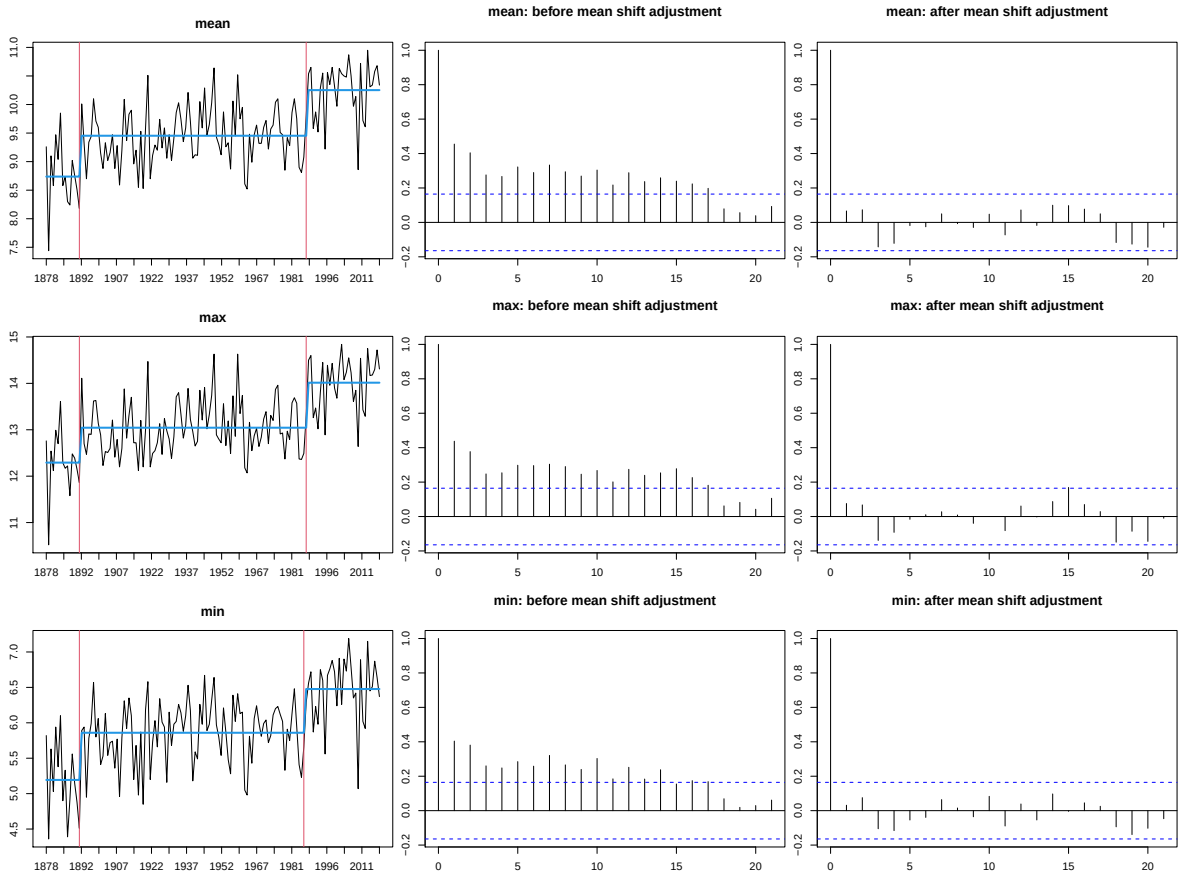

Figure E.2: Left: yearly average of the mean, maximum and minimum monthly temperatures (top to bottom), plotted together with the change points estimated by WEM.gSC (vertical lines) and piecewise constant mean (bold lines). Middle and right: autocorrelation function of the data without and with the time-varying mean adjusted.

(ii) $j = m$ and $\mathcal{F}_{s,k,e}$, $\theta_j^\circ + 1 \le k \le \theta_{j+1}^\circ$ 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.

Lemma F.3. For given $z \in (0,1)$, define

$$g(t) = \sqrt{t(1-t)} \cdot z/t \quad \text{for } t \in [z,1].$$

Then, we have

$$g(z) - g(t) \ge \begin{cases} (t-z)/(4\sqrt{2z}) & \text{for } t \in (z, \min(2z,1)), \\ \sqrt{z(1-z)}/4 & \text{for } t \in [\min(2z,1), 1]. \end{cases}$$

Proof. Suppose that $t \geq z$. Then,

$$\frac{g(z)-g(t)}{\sqrt{z(1-z)}} = 1 - \sqrt{\frac{z(1-t)}{t(1-z)}} = 1 - \sqrt{\frac{1-(t-z)/t}{1+(t-z)/(1-t)}}$$
$$\geq 1 - \sqrt{1-\frac{t-z}{t}} \geq 1 - \left(1 - \frac{t-z}{2t}\right) = \frac{t-z}{2t}.$$

Further, by mean value theorem,

$$g(z) - g(t) = \frac{1}{2}\frac{(t-z)z}{w^{3/2}\sqrt{1-w}}$$

for some $w \in [z,t]$. If $t - z < z$ and thus $w < \min(2z, 1)$, we obtain

$$g(z) - g(t) \geq \frac{t-z}{4\sqrt{2z}}.$$

□

F.1.2 Proof of Theorem 2.1

Throughout the proofs, $C_0, C_1, \ldots$ denote some positive constants.

Proof of (i)–(ii). 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 $R_n \leq \widetilde{K}(\widetilde{K}+1)/2$. 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 $3\lfloor (e-s)/\widetilde{K} \rfloor \leq 2\min_{1\leq j\leq q} \delta_j$. This condition in turn is met under (7). Then, it follows from the proof of Proposition B.1 of Cho and Kirch (2020b) that there exists a permutation $\{\pi(1), \ldots, \pi(q)\}$ of $\{1, \ldots, q\}$ such that on $\mathcal{Z}_n$,

$$\max_{1\leq j\leq q} (f'_{\pi(j)})^2 |k_{(j)} - \theta_{\pi(j)}| \leq \rho_n = c_2 \zeta_n^2, \quad \text{and} \tag{F.2}$$
$$\exp(\mathcal{Y}_{(j)}) = \left|\mathcal{X}_{(j)}\right| \geq C_0 |f'_{\pi(j)}| \sqrt{\delta_{\pi(j)}} \geq C_1 n^{\varphi/2} \zeta_n \tag{F.3}$$

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 $(f_j')^2 \min(\theta_j - s_{(m)}, e_{(m)} - \theta_j) \leq \rho_n$, or

(c) $(s_{(m)}, e_{(m)}) \cap \Theta = \{\theta_j, \theta_{j+1}\}$ and $\max\{(f_j')^2(\theta_j - s_{(m)}), (f_{j+1}')^2(e_{(m)} - \theta_{j+1})\} \leq \rho_n$,

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

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

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

$$\begin{aligned}\exp(\mathcal{Y}_{(m)}) &\leq |\mathcal{F}_{s_{(m)},k_{(m)},e_{(m)}}| + |\mathcal{Z}_{s_{(m)},k_{(m)},e_{(m)}}| \\ &\leq \sqrt{\frac{(k_{(m)} - s_{(m)})(e_{(m)} - k_{(m)})}{e_{(m)} - s_{(m)}}} \frac{(\theta_j - s_{(m)})|d_j|}{k_{(m)} - s_{(m)}} + 2\zeta_n \\ &\leq \sqrt{d_j^2 \min(\theta_j - s_{(m)}, e_{(m)} - \theta_j)} + 2\zeta_n \leq \sqrt{\rho_n} + 2\zeta_n \leq C_2\zeta_n \end{aligned} \tag{F.5}$$

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).

Proof of (iii). Suppose $m \geq q + 1$. We write $\mathbb{Y}_{i,m,Q} = \mathbb{Y}_{i,m,Q}^{(1)} + \mathbb{Y}_{i,m,Q}^{(2)}$, where

$$\begin{aligned}\mathbb{Y}_{i,m,Q}^{(1)} &= \sqrt{\frac{(m-i)(Q-m)}{(Q-i)}} \frac{1}{m-i} \sum_{r=i+1}^{q} \gamma_r \log(n)(1 + o(1)), \quad \text{and} \\ \mathbb{Y}_{i,m,Q}^{(2)} &= \sqrt{\frac{(m-i)(Q-m)}{(Q-i)}} \left\{\frac{1}{m-i} \sum_{r=i+1}^{m} \kappa_r \log(\zeta_n)(1 + o(1)) - \frac{1}{Q-m} \sum_{r=m+1}^{Q} \kappa_r \log(\zeta_n)(1 + o(1))\right\},\end{aligned}$$

where $\kappa_r = 1$ for $r = 1, \ldots, q$. Note that

$$\begin{aligned}\mathbb{Y}_{i,m,Q}^{(2)} - \mathbb{Y}_{i,q,Q}^{(2)} &\leq \left\{\sqrt{\frac{(m-i)(Q-m)}{Q-i}} - \sqrt{\frac{(q-i)(Q-q)}{Q-i}} \frac{Q-m}{Q-q}\right\} \log(\zeta_n)(1 + o(1)) \\ &= \frac{(m-q)\sqrt{q-i}(Q-m)}{2\sqrt{\widetilde{m}-i}(Q-\widetilde{m})^{3/2}} \log(\zeta_n)(1 + o(1)) \end{aligned} \tag{F.7}$$

for some $\widetilde{m} \in [q, m]$, where the inequality holds under the constraints on $\{\kappa_r\}$, while the equality follows from the mean value theorem. Similarly,

$$\mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} \leq \left\{ \sqrt{\frac{(q-i)(Q-q)}{Q-i}} - \sqrt{\frac{(m-i)(Q-m)}{Q-i}} \frac{q-i}{m-i} \right\} \log(\zeta_n)(1+o(1)). \quad \text{(F.8)}$$

Finally, it trivially holds that

$$\left| \mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} \right| \leq \sqrt{Q} \log(\zeta_n)(1+o(1)). \quad \text{(F.9)}$$

Also, by Lemma F.3, we obtain

$$\frac{\mathbb{Y}^{(1)}_{i,q,Q} - \mathbb{Y}^{(1)}_{i,m,Q}}{(q-i)^{-1} \sum_{r=i+1}^{q} \gamma_r \log(n)(1+o(1))} \geq \begin{cases} (m-q)/(4\sqrt{2(q-i)}) & \text{for } q < m \leq 2q-i, \\ \sqrt{(q-i)(Q-q)/(Q-i)}/4 & \text{for } 2q-i < m < q. \end{cases} \quad \text{(F.10)}$$

From (F.7)–(F.8) and (F.10), if $m - i \leq 2(q-i)$, we have

$$\frac{\left| \mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} \right|}{\mathbb{Y}^{(1)}_{i,q,Q} - \mathbb{Y}^{(1)}_{i,m,Q}} \leq \max(1, 4\sqrt{Q}) \frac{\log(\zeta_n)(1+o(1))}{\gamma_q \log(n)} = o(1)$$

for any $i = 0, \ldots, q-1$ under (F.1). Similarly, for $m - i > 2(q-1)$ and any i, we have

$$\frac{\left| \mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} \right|}{\mathbb{Y}^{(1)}_{i,q,Q} - \mathbb{Y}^{(1)}_{i,m,Q}} \leq \frac{4\sqrt{Q} \log(\zeta_n)(1+o(1))}{\gamma_q \log(n)} = o(1)$$

from (F.9) and (F.10), which proves the assertion.

Proof of (iv). Note that for $m \leq q-1$, we have

$$\mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} = \left\{ \sqrt{\frac{(q-i)(Q-q)}{Q-i}} - \sqrt{\frac{(m-i)(Q-m)}{Q-i}} \frac{Q-q}{Q-m} \right\} \times \left(1 - \frac{1}{Q-q} \sum_{r=q+1}^{Q} \kappa_r \right) \log(\zeta_n)(1+o(1))$$

such that

$$0 \leq \mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} \leq \left\{ \sqrt{\frac{(q-i)(Q-q)}{Q-i}} - \sqrt{\frac{(m-i)(q-m)}{Q-i}} \frac{Q-q}{Q-m} \right\} \log(\zeta_n)(1+o(1))$$

under the constraints on $\{\kappa_r\}$, while

$$\mathbb{Y}^{(1)}_{i,q,Q} - \mathbb{Y}^{(1)}_{i,m,Q} = \left\{ \sqrt{\frac{(q-i)(Q-q)}{Q-i}} - \sqrt{\frac{(m-i)(Q-m)}{Q-i}} \frac{Q-q}{Q-m} \right\} \gamma_q \log(n)(1+o(1)).$$

Therefore, we obtain

$$\frac{\left| \mathbb{Y}^{(2)}_{i,q,Q} - \mathbb{Y}^{(2)}_{i,m,Q} \right|}{\mathbb{Y}^{(1)}_{i,q,Q} - \mathbb{Y}^{(1)}_{i,m,Q}} \leq \frac{\log(\zeta_n)(1+o(1))}{\gamma_q \log(n)} = o(1)$$

which, together with (iii), completes the proof.

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 (10). Also we set $\mathbf{X}_{(j)} = [\mathbf{L}_{(j)} : \mathbf{R}_{(j)}] = [\mathbf{L}_{(j)} : \mathbf{1}]$ 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 (8) with the true change point locations θ_j and AR order p by

$$\mathbf{Y} = \mathbf{X}^\circ \boldsymbol{\beta}^\circ + (\boldsymbol{\nu}^\circ - \mathbf{L}^\circ \boldsymbol{\alpha}^\circ) + \boldsymbol{\varepsilon} = \begin{bmatrix} \underbrace{\mathbf{L}^\circ}_{n \times p} & \underbrace{\mathbf{R}^\circ}_{n \times (q+1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^\circ \\ \boldsymbol{\mu}^\circ \end{bmatrix} + (\boldsymbol{\nu}^\circ - \mathbf{L}^\circ \boldsymbol{\alpha}^\circ) + \boldsymbol{\varepsilon} \tag{F.11}$$

with $\boldsymbol{\nu}^\circ = ((1 - a(B)) f_t,\ 1 \leq t \leq n)^\top$ and the rows of $\mathbf{X}^\circ$ are 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.4. 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 (10), 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 $\mathcal{A} \subset \widehat{\Theta}_M$:

$$\mathrm{tr}(\mathbf{L}^\top \mathbf{L}) = O(n), \quad \mathrm{tr}(\mathbf{L}_{(j)}^\top \mathbf{L}_{(j)}) = O(N_j), \tag{F.12}$$

$$\liminf_{n\to\infty} n^{-1}\lambda_{\min}(\mathbf{L}^\top \mathbf{L}) > 0, \quad \liminf_{N_j\to\infty} N_j^{-1}\lambda_{\min}(\mathbf{L}_{(j)}^\top \mathbf{L}_{(j)}) > 0, \tag{F.13}$$

$$\mathrm{tr}(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)}) = O(N_j), \quad \liminf_{N_j\to\infty} N_j^{-1}\lambda_{\min}(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)}) > 0, \tag{F.14}$$

$$(\mathbf{L}^\top \mathbf{L})^{-1}\mathbf{L}^\top \boldsymbol{\varepsilon} = O\left(\sqrt{\frac{\log(n)}{n}}\right), \quad (\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)})^{-1}\mathbf{X}_{(j)}^\top \boldsymbol{\varepsilon}_{(j)} = O\left(\sqrt{\frac{\log(n)}{N_j}}\right). \tag{F.15}$$

When $r = 0$, we still have (F.14) and the second statement of (F.15) hold.

Proof. The results in (F.12)–(F.13) 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.14) follow from that $\mathbf{R}_{(j)}^\top \mathbf{R}_{(j)} = N_j$. By Lemma 1 of Lai and Wei (1982b), we have

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

which, together with (F.12) and (F.14), leads to the rates of convergence in (F.15). □

Lemma F.5 (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.6. 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 $1 \le j \le m$, 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,

$$\|(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{R}_{-j}})\mathbf{U}\|^2 - \|(\mathbf{I} - \boldsymbol{\Pi}_{\mathbf{R}})\mathbf{U}\|^2 = |\mathcal{C}_{k_{j-1},k_j,k_{j+1}}(\mathbf{U})|^2 \tag{F.16}$$

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.4.

We briefly sketch the proof, which proceeds in four steps (i)–(iv). We first suppose that Assumption 3.2 holds with $M=1$, and also that p is known. Then, a single iteration of the gSC algorithm in Appendix 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), under $H_0: q=0$, we show that $\mathbf{R}\widehat{\boldsymbol{\mu}} \approx \mathbf{1}\mu_0^\circ \approx \mathbf{\Pi_1}(\mathbf{Y}-\mathbf{L}\widehat{\boldsymbol{\alpha}})$ with $\mu_0^\circ = (1-\sum_{i=1}^p a_i)f_0$ representing the time-invariant overall level, and therefore $\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \approx \|(\mathbf{I}-\mathbf{\Pi_1})(\mathbf{Y}-\mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2$ 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), under $H_1 : q \geq 1$, we show that

$$\|(\mathbf{I} - \mathbf{\Pi_1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \geq Cq \min_{1 \leq j \leq q} d_j^2 \delta_j \gg q\xi_n$$

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 (14). 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 gSC algorithm: When $l > l^*$, any $\widehat{\theta}_{l,j} \notin \widehat{\Theta}_{l^*}$ are spurious estimators and thus we have `SCalg` return `TRUE`; 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 `SCalg` returns `FALSE`.

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 $\mathbf{X} = \mathbf{X}(\widehat{\Theta}, p) = [\mathbf{L} : \mathbf{R}]$, we define the corresponding $\mathbf{X}_{(j)} = [\mathbf{L}_{(j)} : \mathbf{1}]$, $\mathbf{Y}_{(j)}$, $\boldsymbol{\varepsilon}_{(j)}$ with respect to $\widehat{\Theta} = \{\widehat{\theta}_j,\, j = 1, \ldots, \widehat{q}\}$ and let $N_j = \widehat{\theta}_{j+1} - \widehat{\theta}_j$ and $N = \max_{0 \leq j \leq \widehat{q}} N_j$. Then, it trivially follows that

$$n \leq (\widehat{q} + 1)N. \tag{F.17}$$

Recalling the notations in (F.11), we define $\mathbf{R}^\circ_{(j)}$, a sub-matrix of $\mathbf{R}^\circ$, analogously as $\mathbf{X}_{(j)}$ is defined with $\mathbf{X}$ with respect to $\widehat{\Theta}$. Also defined in (F.11), for $\boldsymbol{\nu}^\circ$, there exists a constant $A > 0$ such that, for $t = \theta_j + 1, \ldots, \theta_j + p$, $1 \leq j \leq q$,

$$|[\boldsymbol{\nu}^\circ - \mathbf{R}^\circ \boldsymbol{\mu}^\circ]_t| \leq |d_j| \max_{1 \leq i \leq p} \left| \sum_{i'=i}^{p} a_{i'} \right| \leq |d_j|, \tag{F.18}$$

while $[\boldsymbol{\nu}^\circ - \mathbf{R}^\circ \boldsymbol{\mu}^\circ]_t = 0$ elsewhere.

(i) Proof under $H_0 : q = 0$. We first note that by Proposition F.4, we have

$$\left\| \widehat{\boldsymbol{\beta}}_{(j)} - (\boldsymbol{\alpha}^{\circ\top}, \mu_0^\circ)^\top \right\| = O\left(\sqrt{\frac{\log(n)}{N_j}} \right) \text{ and } \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O\left(\sqrt{\frac{\log(n)}{N}} \right) \tag{F.19}$$

due to how the parameters are estimated, see (11).

We decompose the residual sum of squares as

$$\begin{aligned}
&\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) + \mathbf{R}(\widehat{\boldsymbol{\mu}} - \mathbf{1}\mu_0^\circ) - \boldsymbol{\varepsilon}\|^2 \\
=&\|\boldsymbol{\varepsilon}\|^2 + \|\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)\|^2 + \|\mathbf{R}(\widehat{\boldsymbol{\mu}} - \mathbf{1}\mu_0^\circ)\|^2 + 2(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)^\top \mathbf{L}^\top \mathbf{R}(\widehat{\boldsymbol{\mu}} - \mathbf{1}\mu_0^\circ) \\
&- 2\boldsymbol{\varepsilon}^\top \mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) - 2\boldsymbol{\varepsilon}^\top \mathbf{R}(\widehat{\boldsymbol{\mu}} - \mathbf{1}\mu_0^\circ) =: \|\boldsymbol{\varepsilon}\|^2 + \mathcal{R}_{11} + \mathcal{R}_{12} + \mathcal{R}_{13} + \mathcal{R}_{14} + \mathcal{R}_{15}.
\end{aligned}$$

Invoking (F.12) and (F.19),

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

due to (F.17). Also by (F.19),

$$\mathcal{R}_{12} = O\left(\sum_{j=0}^{\widehat{q}} N_j \cdot \frac{\log(n)}{N_j}\right) = O(\widehat{q}\log(n))$$

from which we obtain $\mathcal{R}_{13} = O(\widehat{q}\log(n))$. By (F.12), (F.15) and (F.19),

$$|\mathcal{R}_{14}| \leq \|(\mathbf{L}^\top\mathbf{L})^{-1}\mathbf{L}^\top\boldsymbol{\varepsilon}\| \, \|\mathbf{L}^\top\mathbf{L}\| \, \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O\left(\sqrt{n\log(n)} \cdot \sqrt{\frac{\log(n)}{N}}\right) = O\left(\sqrt{\widehat{q}}\log(n)\right).$$

Also, on $\mathcal{E}_n$, we have $|[\mathbf{R}^\top\boldsymbol{\varepsilon}]_{j+1}| \leq \sqrt{N_j}\omega_n$ for $j = 0, \ldots, \widehat{q}$, see Assumption 3.1 (iv). Hence by (F.19),

$$|\mathcal{R}_{15}| = O\left(\sum_{j=0}^{\widehat{q}} \sqrt{N_j}\omega_n \cdot \sqrt{\frac{\log(n)}{N_j}}\right) = O\left(\widehat{q}\omega_n\sqrt{\log(n)}\right).$$

Putting together the bounds on $\mathcal{R}_{11}$–$\mathcal{R}_{15}$, we conclude that

$$\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O\left(\widehat{q}(\log(n) \vee \omega_n^2)\right). \tag{F.20}$$

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.4, we have $|\mathcal{R}_{21}| = O(\log(n))$. Also, $\mathcal{R}_{22} \leq \mathcal{R}_{11} = O(\widehat{q}\log(n))$, and

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

where the first term is handled as $|\mathcal{R}_{14}|$ while the second term is bounded by $\sqrt{|\mathcal{R}_{21}\mathcal{R}_{11}|} = O(\sqrt{\widehat{q}}\log(n))$. Therefore,

$$\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{1})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O(\widehat{q}\log(n)). \tag{F.21}$$

Combining (F.20) and (F.21) with Assumption 3.1 (ii)–(iii), and noting that $\log(1 + x) \leq x$

for all $x \geq 0$,

$$\begin{aligned}\mathrm{SC}_0(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(p)) - \mathrm{SC}(\{X_t\}_{t=1}^n, \widehat{\Theta}, p) &= \frac{n}{2}\log\left(1 + \frac{\|(\mathbf{I}-\mathbf{\Pi_1})(\mathbf{Y}-\mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 - \|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}\right) - \widehat{q}\xi_n \\ &= O\left(\widehat{q}(\log(n) \vee \omega_n^2)\right) - \widehat{q}\xi_n < 0\end{aligned}$$

for n large enough.

(ii) Proof under $H_1 : q \geq 1$. Recall that in $\mathcal{M}_n$, we have $\widehat{q} = q$. We sometimes use that

$$\begin{aligned}\left|\sum_{t=\widehat{\theta}_j}^{\widehat{\theta}_{j+1}-1} X_t\right| &\leq (1-\sum_{i=1}^p a_i)\bar{f}N_j + \left|\sum_{t=\widehat{\theta}_j}^{\widehat{\theta}_{j+1}-1}(X_t - f_t)\right| \\ &= (1-\sum_{i=1}^p a_i)\bar{f}N_j + O\left(\sqrt{N_j}\omega_n\right) = O(N_j) \qquad \text{(F.22)}\end{aligned}$$

where $\bar{f} = \max_{0\leq j\leq q}|f_{\theta_j+1}|$, from Assumptions 3.1 (iv), 3.2 and that $D_n^{-1}\rho_n \to 0$.

We first establish the consistency of $\widehat{\mu}_j$ in estimating μ_j°. Applying Lemma F.5, we write

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

where $\boldsymbol{\nu}_{(j)}^\circ = ((1-a(B))f_t, \widehat{\theta}_j + 1 \leq t \leq \widehat{\theta}_{j+1})^\top$. Since $(\mathbf{1}^\top(\mathbf{I}-\mathbf{\Pi}_{\mathbf{L}_{(j)}})\mathbf{1})^{-1}$ is a sub-matrix of $(\mathbf{X}_{(j)}^\top\mathbf{X}_{(j)})^{-1}$, we have $(\mathbf{1}^\top(\mathbf{I}-\mathbf{\Pi}_{\mathbf{L}_{(j)}})\mathbf{1})^{-1} \leq (\lambda_{\min}(\mathbf{X}_{(j)}^\top\mathbf{X}_{(j)}))^{-1}$ (see e.g. Theorem 4.2.2 of Horn and Johnson (1985)) and thus $\liminf_{N_j\to\infty} N_j^{-1}(\mathbf{1}^\top(\mathbf{I}-\mathbf{\Pi}_{\mathbf{L}_{(j)}})\mathbf{1}) > 0$ by (F.14). Also, since $\mathbf{1}^\top(\mathbf{I}-\mathbf{\Pi}_{\mathbf{L}_{(j)}})\mathbf{1} \leq N_j$ trivially, we obtain

$$|\mathcal{R}_{32}| = O\left(\sqrt{\frac{\log(n)}{N_j}}\right)$$

adopting the same arguments used in the proof of (F.15). Next, by (F.18) and

$$\mathbf{R}_{(j)}^\circ\boldsymbol{\mu}^\circ - \mathbf{1}\mu_j^\circ = (\underbrace{-d_j, \ldots, -d_j}_{\max(0,\theta_j-\widehat{\theta}_j)}, 0, \ldots, 0, \underbrace{d_{j+1}, \ldots, d_{j+1}}_{\max(0,\widehat{\theta}_{j+1}-\theta_{j+1})})^\top,$$

we obtain

$$|\mathcal{R}_{31}|^2 = O\left(\frac{\sum_{l=j}^{j+1} d_l^2 \cdot d_l^{-2}\rho_n}{N_j}\right) = O\left(\frac{\rho_n}{N_j}\right).$$

Putting together the bounds on $\mathcal{R}_{31}$–$\mathcal{R}_{32}$, we obtain

$$|\widehat{\mu}_j - \mu_j^\circ| = O\left(\sqrt{\frac{\log(n) \vee \rho_n}{N_j}}\right). \tag{F.23}$$

Similarly by Lemma F.5,

$$\widehat{\boldsymbol{\alpha}}_{(j)} = \boldsymbol{\alpha}^\circ + (\mathbf{L}_{(j)}^\top \mathbf{L}_{(j)})^{-1} \mathbf{L}_{(j)}^\top \left\{ \boldsymbol{\varepsilon}_{(j)} + (\boldsymbol{\nu}_{(j)}^\circ - \mathbf{1}\mu_j^\circ) - \mathbf{1}(\widehat{\mu}_j - \mu_j^\circ) \right\}$$

and from (F.13), (F.22) and (F.23), we obtain

$$\|\widehat{\boldsymbol{\alpha}}_{(j)} - \boldsymbol{\alpha}^\circ\| = O\left(\sqrt{\frac{\log(n) \vee \rho_n}{N_j}}\right). \tag{F.24}$$

Next, we consider

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

By (F.12) and (F.24),

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

Also, due to (F.23) and the arguments leading up to it,

$$\begin{aligned}
\mathcal{R}_{42} &\leq 2\|\mathbf{R}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\circ)\|^2 + 2\|\mathbf{R}\boldsymbol{\mu}^\circ - \boldsymbol{\nu}^\circ\|^2 \\
&= O\left\{ \sum_{j=1}^{q} \left(N_j \cdot \frac{\log(n) \vee \rho_n}{N_j} + d_j^2 \cdot d_j^{-2} \rho_n \right) \right\} = O\left(q(\log(n) \vee \rho_n)\right)
\end{aligned} \tag{F.25}$$

and we also obtain $\mathcal{R}_{43} = O\left(q(\log(n) \vee \rho_n)\right)$. By (F.15) and (F.24),

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

while with (F.23) and Assumption 3.1 (iv),

$$|\mathcal{R}_{45}| \leq 2|\boldsymbol{\varepsilon}^\top \mathbf{R}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\circ)| + 2|\boldsymbol{\varepsilon}^\top (\mathbf{R}\boldsymbol{\mu}^\circ - \boldsymbol{\nu}^\circ)|$$

$$= O\left(\sum_{j=1}^{q} \sqrt{N_j}\omega_n \cdot \sqrt{\frac{\log(n)}{N_j}}\right) + O\left(\sum_{j=1}^{q} |d_j| \cdot \sqrt{d_j^{-2}\rho_n}\omega_n\right)$$
$$= O\left(q\omega_n(\sqrt{\log(n)} \vee \sqrt{\rho_n})\right).$$

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

$$\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O\left(q\left(\log(n) \vee \omega_n^2 \vee \rho_n\right)\right). \tag{F.26}$$

Next, note that

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

Repeatedly invoking Lemma F.6, we have

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

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 $\widehat{\theta}_j \leq \theta_j$. Analogous arguments apply when $\widehat{\theta}_j > \theta_j$. By Lemma F.1,

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

Under Assumptions 3.2, 3.3 and 3.4, we have $\min(N_{j-1}, N_j)^{-1} d_j^2 |\widehat{\theta}_j - \theta_j| = O(\delta_j^{-1}\rho_n) = o(1)$ and thus

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

since $D_n^{-1}\rho_n \to 0$ as $n \to \infty$ due to Assumption 3.4, 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

$$\min_{1\le j\le q} |\mathcal{C}_j(\mathbf{R}^\circ \boldsymbol{\mu}^\circ)| \ge \sqrt{\frac{D_n}{2}}(1 + o(1)) + o(\sqrt{\rho_n}). \tag{F.27}$$

Similarly, by (F.18), we derive

$$|\mathcal{C}_j(\boldsymbol{\nu}^\circ - \mathbf{R}^\circ \boldsymbol{\mu}^\circ)| \le p\sqrt{\frac{N_{j-1}N_j}{N_{j-1}+N_j}} \left\{ \frac{|d_j| + |d_{j+1}|}{N_j} + \frac{|d_{j-1}|}{N_{j-1}} \right\} = o(1). \tag{F.28}$$

Invoking Assumption 3.1 (iv), it is easily seen that

$$|\mathcal{C}_j(\boldsymbol{\varepsilon})| \le 2\omega_n. \tag{F.29}$$

Finally, by (F.22) and (F.24),

$$\begin{aligned} |\mathcal{C}_j(\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ))| &= \sqrt{\frac{N_{j-1}N_j}{N_{j-1}+N_j}} \left| \frac{1}{N_{j-1}} \mathbf{1}^\top \mathbf{L}_{(j-1)}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) - \frac{1}{N_j} \mathbf{1}^\top \mathbf{L}_{(j)}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ) \right| \\ &= O\left(\sqrt{\min(N_{j-1}, N_j)} \cdot \sqrt{\frac{\log(n) \vee \rho_n}{N}} \right) = O\left(\sqrt{\log(n) \vee \rho_n} \right). \end{aligned} \tag{F.30}$$

By (F.27)–(F.30), under Assumption 3.3, there exists some constant $C_0 > 0$ satisfying

$$\mathcal{R}_{51} \ge C_0 q D_n \qquad \text{for } n \text{ large enough.} \tag{F.31}$$

Next, we note that

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

First, by Assumption 3.1 (iv), $\mathcal{R}_{71} \le \sum_{j=0}^q N_j \omega_n^2 \cdot N_j^{-1} = q\omega_n^2$. As in (F.30), $\mathcal{R}_{72} = O(q(\log(n) \vee \rho_n))$. Note that

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

where the first term is bounded by $O(q(\log(n) \vee \rho_n))$ as in (F.25). From that

$$\boldsymbol{\mu}^\circ - (\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\mathbf{R}^\circ\boldsymbol{\mu}^\circ = \begin{bmatrix} \frac{(\theta_0-\widehat{\theta}_0)_+ d_0 - (\widehat{\theta}_1-\theta_1)_+ d_1}{\widehat{\theta}_1} \\ \frac{(\theta_1-\widehat{\theta}_1)_+ d_1 - (\widehat{\theta}_2-\theta_2)_+ d_2}{\widehat{\theta}_2-\widehat{\theta}_1} \\ \vdots \\ \frac{(\theta_q-\widehat{\theta}_q)_+ d_q - (\widehat{\theta}_{q+1}-\theta_{q+1})_+ d_{q+1}}{n-\widehat{\theta}_q} \end{bmatrix}, \tag{F.32}$$

$$\left|[(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top(\mathbf{R}^\circ\boldsymbol{\mu}^\circ - \boldsymbol{\nu}^\circ)]_j\right| \leq \frac{p(|d_{j-1}| + |d_j|)}{\widehat{\theta}_j - \widehat{\theta}_{j-1}} \tag{F.33}$$

(recall that $\widehat{\theta}_0 = \theta_0 = 0$ and $\widehat{\theta}_{q+1} = \theta_{q+1} = n$ and (F.18)) and Assumptions 3.2 and 3.3, we obtain

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

for some constant $C_1 > 0$, hence $\mathcal{R}_{73} = O(q(\log(n) \vee \rho_n))$. The bounds on $\mathcal{R}_{72}$ and $\mathcal{R}_{73}$ imply the same bound on $\mathcal{R}_{74}$. Next,

$$|\mathcal{R}_{75}| \leq 2|\boldsymbol{\varepsilon}^\top\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)| + 2|\boldsymbol{\varepsilon}^\top\boldsymbol{\Pi}_\mathbf{R}\mathbf{L}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ)|$$

where the first term in the RHS is bounded as $\mathcal{R}_{44} = O(\sqrt{q}(\log(n) \vee \rho_n))$, while the second term is bounded by

$$\sqrt{\boldsymbol{\varepsilon}^\top\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\boldsymbol{\varepsilon}}\,\|\mathbf{L}\|\,\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\circ\| = O\left(\sqrt{n}\omega_n \cdot \sqrt{\frac{\log(n) \vee \rho_n}{N}}\right) = O\left(\sqrt{q}(\omega_n^2 \vee \log(n) \vee \rho_n)\right)$$

by Assumption 3.1 (iv) and (F.12). Hence, $|\mathcal{R}_{75}| = O(\sqrt{q}(\log(n) \vee \rho_n \vee \omega_n^2))$. 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 the first term is bounded by $C_2\sum_{j=1}^{q}\sqrt{d_j^{-2}\rho_n}\omega_n \cdot |d_j| = O(q\omega_n\sqrt{\rho_n})$ due to Assumption 3.1 (iv), while the second term is bounded by

$$C_3\sum_{j=1}^{q}\sqrt{N_j}\omega_n \cdot \frac{\sqrt{d_j^{-2}\rho_n} \cdot |d_j|}{N_j} = O(q\sqrt{\rho_n}),$$

recalling (F.32)–(F.33) and by Assumptions 3.1 (iv), 3.2 and 3.3. Therefore, $\mathcal{R}_{76} = O(q\omega_n\sqrt{\rho_n})$. Collecting the bounds on $\mathcal{R}_{71}$–$\mathcal{R}_{76}$, we obtain

$$\|(\mathbf{I} - \boldsymbol{\Pi}_\mathbf{R})(\mathbf{Y} - \mathbf{L}\widehat{\boldsymbol{\alpha}})\|^2 = \|\boldsymbol{\varepsilon}\|^2 + O\left(q(\log(n) \vee \omega_n^2 \vee \rho_n)\right). \tag{F.34}$$

From (F.26), (F.31) and (F.34),

$$\|(\mathbf{I}-\mathbf{\Pi_1})(\mathbf{Y}-\mathbf{L}^\circ\widehat{\boldsymbol{\alpha}})\|^2-\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2\geq C_0qD_n+O\left(q(\log(n)\vee\omega_n^2\vee\rho_n)\right). \tag{F.35}$$

Note that

$$\begin{aligned}\mathrm{SC}_0(\{X_t\}_{t=1}^n,\widehat{\boldsymbol{\alpha}}(p))-\mathrm{SC}(\{X_t\}_{t=1}^n,\widehat{\Theta},p)&=\frac{n}{2}\log\left(1+\frac{\|(\mathbf{I}-\mathbf{\Pi_1})(\mathbf{Y}-\mathbf{L}^\circ\widehat{\boldsymbol{\alpha}})\|^2-\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}\right)-q\xi_n\\&=:\frac{n}{2}\log(1+\mathcal{R}_8)-q\xi_n.\end{aligned} \tag{F.36}$$

When $\mathcal{R}_8\geq 1$, we have the RHS of (F.36) 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 $C_4>0$ such that $\inf_{1\leq x<2}g(x)\geq C_4$. Therefore,

$$\frac{n}{2}\log(1+\mathcal{R}_8)-q\xi_n\geq C_5qD_n+O\left(q(\log(n)\vee\omega_n^2\vee\rho_n)\right)-q\xi_n>0,$$

invoking Assumption 3.1 (ii)–(iii), (F.26) and (F.35) for some $C_5>0$.

(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 (12). Firstly, suppose that $r>p$ while $r\leq p_{\max}$. Then, by Proposition F.4, we have

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

both under H_0 or H_1, see (F.24). Then, the arguments similar to those leading to (F.20) or (F.26) establish that

$$\|\mathbf{Y}_{(j^*)}-\mathbf{X}_{(j^*)}(r)\widehat{\boldsymbol{\beta}}_{(j^*)}(r)\|^2=\|\boldsymbol{\varepsilon}_{(j^*)}\|^2+O\left(\log(n)\vee\omega^2\vee\rho_n\right)$$

and therefore, we have

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

for n large enough, by Assumption 3.4.

Next, consider $r<p$. For notational convenience, let $\mathbf{\Pi}_{(j^*)}(r)=\mathbf{\Pi}_{\mathbf{X}_{(j^*)}(r)}$, and the sub-matrix of $\mathbf{X}_{(j^*)}(p)$ containing its columns corresponding to the i-th lags for $i=r+1,\ldots,p$ by $\mathbf{X}_{(j^*)}(p|r)$. Then, $[\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)]^{-1}$ is a sub-matrix of $(\mathbf{X}_{(j^*)}(p)^\top\mathbf{X}_{(j^*)}(p))^{-1}$

and thus by Theorem 4.2.2 of Horn and Johnson (1985) and Proposition F.4, we have

$$\begin{aligned}
&\lambda_{\max}\left(\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right) \le \lambda_{\max}\left(\mathbf{X}_{(j^*)}(p)^\top\mathbf{X}_{(j^*)}(p)\right)\\
&\qquad \le \mathsf{tr}\left(\mathbf{X}_{(j^*)}(p)^\top\mathbf{X}_{(j^*)}(p)\right) = O(N) \quad \text{and similarly,} \qquad \text{(F.37)}\\
&\lambda_{\min}\left(\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right) \ge \lambda_{\min}\left(\mathbf{X}_{(j^*)}(p)^\top\mathbf{X}_{(j^*)}(p)\right) \quad \text{and thus}\\
&\liminf_{N\to\infty} N^{-1}\lambda_{\min}\left(\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right) > 0. \qquad \text{(F.38)}
\end{aligned}$$

It then follows that

$$\begin{aligned}
&\|\mathbf{Y}_{(j^*)}-\mathbf{X}_{(j^*)}(r)\widehat{\boldsymbol{\beta}}_{(j^*)}(r)\|^2 - \|\mathbf{Y}_{(j^*)}-\mathbf{X}_{(j^*)}(p)\widehat{\boldsymbol{\beta}}_{(j^*)}(p)\|^2\\
=&\left\|\left[\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right]^{-1/2}\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{Y}_{(j^*)}\right\|^2 \qquad \text{(F.39)}\\
\ge&\,\lambda_{\min}\left(\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right)\left\|\begin{bmatrix}\alpha^\circ_{r+1}\\ \vdots\\ \alpha^\circ_p\end{bmatrix}\right\|^2\\
&-\left\|\left[\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right]^{-1/2}\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\boldsymbol{\varepsilon}_{(j^*)}\right\|^2\\
&-\left\|\left[\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))\mathbf{X}_{(j^*)}(p|r)\right]^{-1/2}\mathbf{X}_{(j^*)}(p|r)^\top(\mathbf{I}-\mathbf{\Pi}_{(j^*)}(r))(\boldsymbol{\nu}_{(j^*)}-\mu^\circ_{j^*}\mathbf{1})\right\|^2\\
\ge&\, C_6 N_{j^*}\sum_{i=r+1}^{p}(\alpha^\circ_i)^2 + O(\log(n)) + O(\rho_n) \qquad \text{(F.40)}
\end{aligned}$$

for n large enough, where the $O(\log(n))$ bound on the RHS of (F.40) is due to (F.37), (F.38) and Lemma 1 of Lai and Wei (1982a), while the $O(\rho_n)$ bound from (F.37) and the arguments adopted in controlling $\mathcal{R}_{31}$, both regardless of whether H_0 or H_1 holds. Therefore, we have

$$\begin{aligned}
&\mathrm{SC}\left(\{X_t\}_{t=k_{j^*}+1}^{k_{j^*+1}},\emptyset,r\right) - \mathrm{SC}\left(\{X_t\}_{t=k_{j^*}+1}^{k_{j^*+1}},\emptyset,p\right)\\
=&\frac{N_{j^*}}{2}\log\left(1+\frac{\|\mathbf{Y}_{(j^*)}-\mathbf{X}_{(j^*)}(r)\widehat{\boldsymbol{\beta}}_{(j^*)}(r)\|^2-\|\mathbf{Y}_{(j^*)}-\mathbf{X}_{(j^*)}(p)\widehat{\boldsymbol{\beta}}_{(j^*)}(p)\|^2}{\|\mathbf{Y}_{(j^*)}-\mathbf{X}_{(j^*)}(p)\widehat{\boldsymbol{\beta}}_{(j^*)}(p)\|^2}\right)-(p-r)\xi_n\\
\ge&\, C_7 N_{j^*}-(p-r)\xi_n > 0
\end{aligned}$$

for n large enough, by Assumption 3.2 on ξ_n. Thus we conclude that the AR order p is consistently estimated by $\widehat{p}=\arg\min_{0\le r\le p_{\max}}\mathrm{SC}(\{X_t\}_{t=k_{j^*}+1}^{k_{j^*+1}},\emptyset,r)$.

The above (i)–(iii) completes the proof in the special case when Assumption 3.2 holds with $M=1$.

(iv) Sequential model selection. In the general case where Assumption 3.2 holds with $M>1$, 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 gSC algorithm, $(\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$, $v = 1, \ldots, q'_l$, correspond to one of the following cases under Assumption 3.2: **null case** with no 'detectable' change points, i.e. either $\Theta \cap (\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1}) = \emptyset$, or all $\theta_j \in \Theta \cap (\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$ 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 $\Theta \cap (\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1}) \neq \emptyset$ and $d_j^2 \min(\theta_j - \widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1} - \theta_j)) > \rho_n$ for at least one $\theta_j \in \Theta \cap (\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$.

In fact, when $l = l^*$, all such intervals $(\widehat{\theta}_{l^*-1,u_v}, \widehat{\theta}_{l^*-1,u_v+1})$ 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 $\mathcal{A} = \widehat{\Theta}_l \cap (\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$ serves the role of the set of spurious estimators, $\widehat{\Theta}$, in the proof of (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 $H_1 : q \geq 1$, by replacing the bound (F.19) derived under H_0 in (i), with (F.23)–(F.24) under H_1 in (ii). Consequently, (F.20) is written as with $O\left(\widehat{q}(\log(n) \vee \omega_n^2 \vee \rho_n)\right)$ and similarly, (F.21) is written with $O\left(\widehat{q}(\log(n) \vee \rho_n)\right)$, which leads to

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

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 $(\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$ as defined in (a). Then, we obtain

$$\begin{aligned} \mathrm{SC}_0\left(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \widehat{\boldsymbol{\alpha}}(p)\right) - \mathrm{SC}\left(\{X_t\}_{t=\widehat{\theta}_{l-1,u_v}+1}^{\widehat{\theta}_{l-1,u_v+1}}, \mathcal{A}, p\right) \\ \geq C_8|\mathcal{A}|D_n + O\left(|\mathcal{A}|(\log(n) \vee \omega_n^2 \vee \rho_n)\right) - |\mathcal{A}|\xi_n > 0 \end{aligned}$$

for n large enough.

(d) The proof on order selection consistency in (iii) holds from Proposition F.4, regardless of whether there are detectable change points in $(\widehat{\theta}_{l-1,u_v}, \widehat{\theta}_{l-1,u_v+1})$ 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

$$\max_{\ell<k<r} |\mathcal{Z}_{\ell,k,r}| \leq \max_{\ell<k<r} \left(\sqrt{\frac{r-k}{r-\ell}} + \sqrt{\frac{k-\ell}{r-\ell}} \right) \zeta_n = \sqrt{2}\zeta_n, \tag{F.41}$$

while by (B.1)–(B.2),

$$|\mathcal{F}_{\ell,\theta,r}| \geq \sqrt{\frac{(f')^2\delta}{4}}. \tag{F.42}$$

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

$$|\mathcal{F}_{\ell,\theta,r} - \mathcal{F}_{\ell,k,r}| \geq \frac{2}{3\sqrt{6}} \frac{|f'|\,|k-\theta|}{\sqrt{\min(\theta-\ell, r-\theta)}} \tag{F.43}$$

for $|k-\theta| \leq \min(\theta-\ell, r-\theta)/2$. Then, from (F.1) and (F.41)–(F.42),

$$|\mathcal{F}_{\ell,\check{\theta},r}| \geq |\mathcal{F}_{\ell,\theta,r}| - 2\max_{\ell<k<r} |\mathcal{Z}_{\ell,k,r}| \geq \sqrt{\frac{(f')^2\delta}{4}} - 2\sqrt{2}\zeta_n > \frac{\sqrt{(f')^2\delta}}{4}, \tag{F.44}$$

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 $\check{\theta} \leq \theta$; the case where $\check{\theta} > \theta$ can be handled analogously. We first establish that

$$\theta - \check{\theta} \leq \min(\theta-\ell, r-\theta)/2. \tag{F.45}$$

If $\theta - \check{\theta} > \min(\theta-\ell, r-\theta)/2 \geq \delta/4$ (due to (B.1)), by Lemma F.2 and (F.43), we have

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

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

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

$$\leq 2\mathsf{P}\left(\max_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}\frac{|A_1(k)|}{D_1(k)}\geq\frac{1}{3}\right)+2\mathsf{P}\left(\max_{\theta-\delta/4\leq k\leq\theta-(f')^{-2}\widetilde{\rho}_n}\frac{|A_2(k)|}{D_2(k)}\geq\frac{1}{3}\right),\quad\text{where}$$

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

Note that

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

For $k<\theta$, we obtain

$$\begin{aligned}&\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}-\sqrt{\frac{r-\ell}{(k-\ell)(r-k)}}=\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\left(1-\sqrt{\frac{(\theta-\ell)(r-\theta)}{(k-\ell)(r-k)}}\right)\\\leq&\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\left(1-\sqrt{1-\frac{\theta-k}{r-k}}\right)\leq\frac{1}{2}\sqrt{\frac{r-\ell}{(\theta-\ell)(r-\theta)}}\frac{\theta-k}{r-k}\end{aligned}$$

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.45),

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

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.43) 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.41),

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

while from (F.42),

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

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.