Multiple change point detection under serial dependence: Wild contrast maximisation and gappy Schwarz algorithm

Haeran Cho*
School of Mathematics, University of Bristol
and
Piotr Fryzlewicz
Department of Statistics, London School of Economics

June 27, 2022

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 contrast maximisation principle, and an information criterion-based model selection strategy termed gappy Schwarz algorithm. 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 WCM.gSa, achieves consistency in estimating both the total number and the locations of the change points. The good performance of WCM.gSa 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

*The author gratefully acknowledges the support of the Leverhulme Trust (RPG-2019-390).

1 Introduction

This paper proposes a new methodology for detecting possibly multiple change points in the piecewise constant mean of an otherwise stationary, linear time series. This is a well-known difficult problem in multiple change point analysis, whose challenge stems from the fact that change points can mask as natural fluctuations in a serially dependent process and vice versa. We briefly review the existing literature on multiple change point detection in the presence of serial dependence and situate our new proposed methodology 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 long-run variance (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. 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) and Zhao et al. (2021), both of which adopt local window-based procedures. Consistency of the methods utilising penalised least squares estimation (Lavielle and Moulines, 2000) or Schwarz criterion (Cho and Kirch, 2021b) 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) model, 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. (2021) propose to minimise the penalised cost function for detection of multiple change points in the mean of AR(1) processes via dynamic programming, and Fang and Siegmund (2020) study a pseudo-sequential approach to change point detection in the level or slope of the data. Lu et al. (2010) investigate the problem of climate time series modelling by allowing for multiple mean shifts and periodic AR noise. Gallagher et al. (2021) propose to estimate AR parameters from differenced data in the presence of multiple mean shifts, and investigate the performance of change point detection methods developed for i.i.d. noise setting to the residuals. Fryzlewicz (2020b) proposes to circumvent the need for accurate estimation of AR parameters 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. We also mention that 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 Cho and Korkas (2021) under high-dimensional time series models.

We now describe our proposed methodology against this literature background and summarise its novelty and main contributions of this paper.

1. The first step of the proposed methodology constructs a sequence of candidate change point models by adopting the Wild Contrast Maximisation (WCM) principle: it iteratively locates the next most likely change point in the data between the previously proposed change point estimators, as the one maximising a given contrast (in our case, the absolute CUSUM statistic) in the data sections over a collection of intervals of varying lengths and locations. It produces a complete solution path to the change point detection problem as a decreasing sequence of max-CUSUMs corresponding to

the successively proposed change point candidates. The WCM principle has successfully been applied to the problem of multiple change point detection in the presence of i.i.d. noise (Fryzlewicz, 2014, 2020a). We show that it is particularly useful under serial dependence by generating a large gap between the max-CUSUMs attributed to change points and those attributed to the fluctuations due to serial correlations. This motivates a new, 'gappy' model sequence generation procedure which, by considering only some of the candidate models along the solution path that correspond to large drops in the decreasing sequence of max-CUSUMs as serious contenders, systematically selects a small subset of model candidates. We justify this gappy model sequence generation theoretically and further demonstrate numerically how it substantially facilitates the subsequent model selection step.

2. The second step performs model selection on the sequence of candidate change point models generated in the first step. To this end, we propose a backward elimination strategy termed gappy Schwarz algorithm (gSa), a new application of Schwarz criterion (Schwarz, 1978) constructed under a parametric, AR model assumption on the noise. Information criteria have been widely adopted for model selection in change point problems (Yao, 1988; Kühn, 2001). However, through its application on the gappy model sequence, our proposal differs from the conventional use of an information criterion in the change point literature which involve its global (Davis et al., 2006; Killick et al., 2012; Romano et al., 2021) or local (Chan et al., 2014; Fryzlewicz, 2014) minimisation. Rather than setting out to minimise Schwarz criterion, the Schwarz algorithm starts from the largest model in consideration and iteratively compares a pair of consecutive models by evaluating the reduction of the cost due to newly introduced change point estimators, offset by the increase of model complexity as measured by Schwarz criterion. This has the advantage over the direct minimisation of the information criterion on a solution path as it avoids the substantial technical challenges linked to dealing with under-specified models in the presence of serial dependence.

The two ingredients, WCM-based gappy model sequence generation and model selection

via Schwarz algorithm, make up the WCM.gSa methodology. Throughout the paper, we highlight the important roles played by these two components and argue that WCM.gSa offers state-of-the-art performance in the problem of multiple change point detection under serially dependent noise. WCM.gSa is modular in the sense that each ingredient can be combined with alternative model selection or model sequence generation procedures, respectively. We provide separate theoretical analyses of the two steps so that they can readily be fed into the analysis of such modifications, as well as showing that the combined methodology, WCM.gSa, achieves consistency in estimating the total number and the locations of multiple change points.

The paper is organised as follows. In Sections 2 and 3, we introduce the two ingredients of WCM.gSa individually, and show its consistency in multiple change point detection in the presence of serial dependence. Section 4 summarises our numerical results and applies WCM.gSa to London air quality datasets. The Supplementary Appendix contains comprehensive simulation studies, an additional data application to central England temperature data, and the proofs of the theoretical results. The R software implementing WCM.gSa is available from `https://github.com/haeran-cho/wcm.gsa`.

2 Candidate model sequence generation via WCM principle

2.1 WCM principle and solution path generation

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 $\Theta := \{\theta_1, \ldots, \theta_q\}$ 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'. 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 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 each θ_j, $j = 1, \ldots q$, whose location can be estimated as the maximiser of the series of CUSUM statistics computed on this interval. Methods in this category include the Wild Binary Segmentation (WBS, Fryzlewicz; 2014), the Seeded Binary Segmentation (Kovács et al., 2020) and the WBS2 (Fryzlewicz, 2020a). All of the above are based on the WCM principle, i.e. the recursive maximisation of the contrast between the means of the data to the left and right of each putative change point as measured by the CUSUM statistic, over a large number of intervals. (We propose the term Wild Contrast Maximisation rather than, say, 'wild CUSUM maximisation' since, in other change point detection problems, the WCM principle can be applied with statistics other than CUSUM, e.g. generalised likelihood ratio tests.) Their theoretical properties have been established assuming i.i.d. (sub-)Gaussianity on $\{Z_t\}_{t=1}^n$.

In the remainder of this paper, we focus on WBS2, whose key feature is that for any given $0 \leq s < e \leq n$, we identify the sub-interval $\{s_\circ + 1, \ldots, e_\circ\} \subset \{s + 1, \ldots, e\}$ and its inner point $k_\circ \in \{s_\circ + 1, \ldots, e_\circ - 1\}$, which obtains a local split of the data that yields the maximum CUSUM statistic. More specifically, let $\mathcal{R}_{s,e}$ denote a subset of $\mathcal{A}_{s,e} := \{(\ell, r) \in \mathbb{Z}^2 : s \leq \ell < r \leq e \text{ and } r - \ell > 1\}$, selected 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

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

Starting with $(s, e) = (0, n)$, recursively repeating the above operation over the segments defined by the thus-identified $k_\circ$, i.e. $\{s + 1, \ldots, k_\circ\}$ and $\{k_\circ + 1, \ldots, e\}$, 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 to select $\mathcal{R}_{s,e}$ from $\mathcal{A}_{s,e}$ via deterministic sampling.

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 contrast (measured as in (2)) at a particular iteration, and the corresponding max-CUSUM statistic. 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 WCM 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 contrast, which helps separate the large max-CUSUMs attributed to mean shifts, from those which are not. In the next section, we propose how to utilise this property of the solution path $\mathcal{P}$ generated according to the WCM principle.

2.2 Gappy model sequence generation

The solution path $\mathcal{P}$ consists of a sequence of candidate change point models $\mathcal{K}_1 \subset \mathcal{K}_2 \subset \ldots$ with $\mathcal{K}_l := \{k_{(1)}, \ldots, k_{(l)}\}$. In this section, we propose a 'gappy' model sequence generation step which selects a subset of the above model sequence by discarding candidate models that are not likely to be the final model. More specifically, by the construction of WBS2, which iteratively identifies the local split of the data with the most contrast (max-CUSUM), 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 some

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 the large gap in $\mathcal{Y}_{(1)} \geq \ldots \geq \mathcal{Y}_{(P)}$, the simplest approach is to look for the largest difference $\mathcal{Y}_{(m)} - \mathcal{Y}_{(m+1)}$. However, this 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 identify the M largest gaps from $\mathcal{Y}_{(m)} - \mathcal{Y}_{(m+1)}$, $1 \leq m \leq P-1$, and denote the corresponding indices by $g_1 < \ldots < g_M$ such that

$$\mathcal{Y}_{(g_l)} - \mathcal{Y}_{(g_l+1)} > \mathcal{Y}_{(m)} - \mathcal{Y}_{(m+1)} \quad \text{for all} \quad m \neq g_l,\, 1 \leq l \leq M.$$

This returns a sequence of nested models

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

with $\widehat{\Theta}_l = \widehat{\Theta}_{l-1} \cup \{k_{(g_{l-1}+1)}, \ldots, k_{(g_l)}\}$. Theorem 2.1 below shows that the model sequence in (4) contains one which consistently detects all q change points with high probability. Typically, this gappy model sequence is much sparser than the sequence of all possible models from the solution path and therefore, intuitively, makes our model selection task easier than if we worked with the entire solution path of all nested models. We confirm this point numerically in the simulation studies reported in Appendix D.

2.3 Theoretical properties

In this section, we establish the theoretical properties of the sequence of nested change point models obtained from combining WBS2 with the gappy model sequence generation outlined in Sections 2.1–2.2. 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

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

Remark 2.1. Assumption 2.1 permits $\{Z_t\}_{t=1}^n$ to have heavier tails than sub-Gaussian such as sub-exponential or sub-Weibull (Vladimirova et al., 2020). Appendix G shows that linear time series with short-range dependence and sub-exponential innovations satisfy the assumption, using the Nagaev-type inequality derived in Zhang and Wu (2017). Similar arguments can be made with the concentration inequalities shown in Doukhan and Neumann (2007) for weakly dependent time series fulfilling $\mathsf{E}(|Z_t|^k) \le (k!)^\nu C^k$ for all $k \ge 1$ and some $\nu \ge 0$ and $C > 0$, or in Merlevède et al. (2011) for 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 $\kappa' \ge 1$, 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. As remarked in Proposition 2.1 (c.i) of Cho and Kirch (2021b), the thus-derived ζ_n usually does not provide the tightest upper bound, but it suits our purpose in controlling the level of noise.

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

Under Assumption 2.2, we assume that there are finitely many change points with the spacing between the change points increasing linearly in n. A similar condition can be found in the literature addressing the problems of change point detection in the presence of serial correlations, see e.g. in Zhao et al. (2021). We may relax this assumption at the price of increased rate of localisation in Theorem 2.1 (i) below. The upper bound on $|f'_j|$ is a technical assumption made to distinguish the problem of detecting change points from that of outlier detection, see Cho and Kirch (2021a) for further discussions.

Theorem 2.1. Let Assumptions 2.1 and 2.2 hold. Suppose that R_n, the number of intervals at each iteration of WBS2, satisfies

$$R_n \geq \frac{9}{8}\left(\frac{n}{\min_{1\leq j\leq q}\delta_j}\right)^2 + 1. \tag{5}$$

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 $0 \leq \kappa_m \leq 1$.

Theorem 2.1 (i) establishes that for the solution path $\mathcal{P}$ obtained according to the WCM principle, the entries corresponding to the q largest max-CUSUMs contain the estimators of all q change points θ_j and further, the localisation rate attained by $\widehat{\theta}_j$ is minimax optimal up to a logarithmic factor (see e.g. Verzelen et al. (2020)). Statement (ii) shows that the q largest log-CUSUMs are of order $\log(n)$ and are thus distinguished from the rest of the log-CUSUMs bounded as $O(\log\log(n))$. In summary, Theorem 2.1 establishes that the sequence of nested change point models (4) contains the consistent model $\widehat{\Theta}[q]$ as a candidate model provided that M is sufficiently large. We emphasise that Theorem 2.1 is not (yet) a full consistency result for our complete change point estimation procedure – this will be the objective of Section 3. Theorem 2.1 merely indicates that the solution path we obtain contains the correctly estimated model, hence it is in principle possible to extract it with the right model selection tool. Section 3 proposes such a tool.

3 Model selection with gSa

In this section, we 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). We propose a new backward elimination-type procedure, referred to as 'gappy

Schwarz algorithm' (gSa), which makes use of the Schwarz information criterion constructed under a parametric assumption imposing an AR structure on $\{Z_t\}_{t=1}^n$. The novelty of gSa is in the new way in which it applies Schwarz criterion, rather than in the formulation of the information criterion itself. We show the usefulness of gSa when change point model selection is performed simultaneously with the estimation of the serial dependence.

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.

$$Z_t = \sum_{i=1}^{p} a_i Z_{t-i} + \varepsilon_t \quad \text{such that} \quad X_t = (1 - a(B))f_t + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t, \tag{6}$$

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 $d_j = \mu_j^\circ - \mu_{j-1}^\circ$ 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, and its data-driven choice is incorporated into the model selection methodology as described later. For now, suppose that it is set to be some integer $r \ge 0$, and that a change point model is 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, Schwarz criterion (Schwarz, 1978) is defined as

$$\mathrm{SC}\left(\{X_t\}_{t=1}^n, \mathcal{A}, r\right) = \frac{n}{2} \log\left(\widehat{\sigma}_n^2\left(\{X_t\}_{t=1}^n, \mathcal{A}, r\right)\right) + (|\mathcal{A}| + r)\xi_n, \tag{7}$$

where $\widehat{\sigma}_n^2(\{X_t\}_{t=1}^n, \mathcal{A}, r)$ denotes a measure of goodness-of-fit (its precise definition is given below), and a penalty is imposed on the model complexity determined by the AR order and the number of change points; the requirement on the penalty parameter ξ_n in relation to the distribution of $\{\varepsilon_t\}$ is discussed in Assumption 3.4.

We adopt the residual sum of squares as $\widehat{\sigma}_n^2(\{X_t\}_{t=1}^n, \mathcal{A}, r)$, i.e.

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

$$\mathbf{X} = \mathbf{X}(\mathcal{A}, r) = \left[\underbrace{\mathbf{L}(r)}_{n\times r} \quad \underbrace{\mathbf{R}(\mathcal{A})}_{n\times(m+1)}\right] = \begin{bmatrix} X_0 & \cdots & X_{1-r} & 1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & \\ X_{k_1-1} & \cdots & X_{k_1-r} & 1 & 0 & 0 & \cdots & 0 \\ X_{k_1} & \cdots & X_{k_1-r+1} & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & & & & \\ X_{n-1} & \cdots & X_{n-r} & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{8}$$

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 $t \leq 0$; in practice, we can simply omit the first $p_{\max}$ observations when constructing $\mathbf{Y}$ and $\mathbf{X}$ above, where $p_{\max}$ denotes a pre-specified upper bound on the AR order. The matrix $\mathbf{X}$ is divided into the AR part contained in $\mathbf{L}(r)$ and the deterministic part in $\mathbf{R}(\mathcal{A})$ under (6). We propose to obtain the estimator of regression parameters denoted by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathcal{A}, r) = (\widehat{\boldsymbol{\alpha}}(r)^\top, \widehat{\boldsymbol{\mu}}(\mathcal{A})^\top)^\top$ via least squares estimation, where $\widehat{\boldsymbol{\alpha}}(r) \in \mathbb{R}^r$ denotes the estimator of the AR parameters and $\widehat{\boldsymbol{\mu}}(\mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|+1}$ that of the segment-specific levels.

We select the typitcally unknown AR order p as follows: AR models of varying orders $r \in \{0, \ldots, p_{\max}\}$, are fitted to the data from which we estimate p by

$$\widehat{p} = \widehat{p}(\mathcal{A}) = \arg\min_{r\in\{0,\ldots,p_{\max}\}} \mathrm{SC}\left(\{X_t\}_{t=1}^n, \mathcal{A}, r\right). \tag{9}$$

In our theoretical analysis, we fully address that the estimator $\widehat{p}(\mathcal{A})$ is used rather than the true AR order p.

3.2 gSa: sequential model selection

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.

Suppose that the number and locations of mean shifts are consistently estimated by (a subset of) $\mathcal{A}$ in the sense made clear in Assumption 3.2 below, which includes the

case of no change point ($q = 0$) with the trivial subset $\emptyset \subset \mathcal{A}$. Then, the estimator $\widehat{\boldsymbol{\beta}}(\mathcal{A}, \widehat{p}) = (\widehat{\boldsymbol{\alpha}}(\widehat{p})^\top, \widehat{\boldsymbol{\mu}}(\mathcal{A})^\top)^\top$ can be shown to estimate the AR parameters sufficiently well with $\widehat{p} = \widehat{p}(\mathcal{A})$ returned by (9), and the 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 change point is ignored in fitting an AR model, the resultant AR parameter estimators over-compensate for the under-specification of mean shifts. In our numerical experiments (reported in Appendix D.3), this often leads to $\mathrm{SC}(\{X_t\}_{t=1}^n, \emptyset, \widehat{p}(\emptyset))$ having a smaller value than $\mathrm{SC}(\{X_t\}_{t=1}^n, \mathcal{A}, \widehat{p})$ such that their direct comparison returns the null model even though there are multiple change points present and detected by $\mathcal{A}$.

Instead, we propose to compare $\mathrm{SC}(\{X_t\}_{t=1}^n, \mathcal{A}, \widehat{p})$ against

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

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 $\widehat{\boldsymbol{\alpha}}(\widehat{p})$ from $\widehat{\boldsymbol{\beta}}(\mathcal{A}, \widehat{p})$ in its definition, SC_0 avoids the above-mentioned difficulty arising when evaluating Schwarz criterion at a model that under-specifies the change points. We conclude that the data is better described by the change point model $\mathcal{A}$ if

$$\mathrm{SC}_0\left(\{X_t\}_{t=1}^n, \widehat{\boldsymbol{\alpha}}(\widehat{p})\right) > \mathrm{SC}(\{X_t\}_{t=1}^n, \mathcal{A}, \widehat{p}), \tag{10}$$

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

This Schwarz criterion-based model selection strategy is extended to be applicable with a sequence of nested change point models $\emptyset = \widehat{\Theta}_0 \subset \widehat{\Theta}_1 \subset \ldots \subset \widehat{\Theta}_M$ as in (4) even when $M > 1$. Referred to as the gappy Schwarz algorithm (gSa) 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 $\{s+1, \ldots, e-1\} \cap \widehat{\Theta}_{l-1} = \emptyset$, and suppose that $\mathcal{A} = \{s+1, \ldots, e-1\} \cap (\widehat{\Theta}_l \backslash \widehat{\Theta}_{l-1})$ 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

$\{s+1, \ldots, e-1\}$, which appear in $\widehat{\Theta}_l$ but do not appear in the 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 (10), with the least squares estimator of the AR parameters $\widehat{\boldsymbol{\alpha}}_{s:e}(\widehat{p}_{s:e})$ and its dimension $\widehat{p}_{s:e}$ obtained locally by minimising $\mathrm{SC}(\{X_t\}_{t=s+1}^e, \mathcal{A}, r)$ over r (see (9)). 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 as not being spurious; 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.

In our theoretical analysis, when $q \geq 1$, we assume that there exists some $1 \leq l^* \leq M$ such that $\widehat{\Theta}_{l^*}$ correctly detects all change points and nothing else (see Assumption 3.2 below), which is guaranteed by the model sequence generation method described in Section 2. Then with high probability, we have $\mathrm{SC}(\{X_t\}_{t=s+1}^e, \mathcal{A}, \widehat{p}_{s:e}) < \mathrm{SC}_0(\{X_t\}_{t=s+1}^e, \widehat{\boldsymbol{\alpha}}_{s:e}(\widehat{p}_{s:e}))$ simultaneously in all local regions $\{s+1, \ldots, e\}$ overlapping with $\widehat{\Theta}_{l^*} \setminus \widehat{\Theta}_{l^*-1}$ while when $l > l^*$, we have $\mathrm{SC}(\{X_t\}_{t=s+1}^e, \mathcal{A}, \widehat{p}_{s:e}) \geq \mathrm{SC}_0(\{X_t\}_{t=s+1}^e, \widehat{\boldsymbol{\alpha}}_{s:e}(\widehat{p}_{s:e}))$ in all such regions. Therefore, sequentially examining the nested change point models from the largest model $\widehat{\Theta}_M$, gSa is expected to return $\widehat{\Theta}_{l^*}$ as the final model. In its implementation, in the unlikely event of disagreement across the regions containing $\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}$, we take a conservative approach and conclude that $\widehat{\Theta}_l$ contains spurious estimators, and update $l \to l-1$ to 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 gSa is provided in Appendix A.2.

In summary, gSa does not directly minimise Schwarz criterion but starting from the largest model, searches for the first largest model $\widehat{\Theta}_l$ in which all candidate estimators in $\widehat{\Theta}_l \setminus \widehat{\Theta}_{l-1}$ are deemed important as described above. By adopting SC_0 for model comparison, it avoids evaluating Schwarz criterion at a model that under-estimates the number of change points (which may lead to loss of power) and achieves model selection consistency as shown in the next section.

3.3 Theoretical properties

For the theoretical analysis of gSa, we make a set of assumptions and remark on their relationship to those made in Section 2.3. Assumption 3.1 is imposed on the stochastic part of model (6).

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 Schwarz criterion, 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 (2021b)); Appendix G considers i.i.d. sub-exponential $\{\varepsilon_t\}$ for which $\omega_n \asymp \log(n)$.

Remark 3.1 (Links between Assumptions 2.1, 2.2 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. Then, as discussed in Remark 2.1, 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 $\xi_n(\min_{0\le j\le \widehat{q}_M}(\widehat{\theta}_{M,j+1} - \widehat{\theta}_{M,j}))^{-1} = o(1)$ and $\widehat{q}_M = |\widehat{\Theta}_M|$ is fixed for all n. Additionally, under $H_1 : q \ge 1$, there exists $l^* \in \{1, \ldots, M\}$ such that

$$\widehat{q}_{l^*} = q \quad \text{and} \quad \max_{1\le j\le q} d_j^2 \left|\widehat{\theta}_{l^*,j} - \theta_j\right| \le \rho_n \tag{11}$$

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

By Theorem 2.1, we have the condition (11) satisfied by the gappy model sequence generated as in (4) with $\rho_n \asymp \zeta_n^2$. We state this result as an assumption so that if gSa 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 can be hard-wired into the solution path generation step.

Assumption 3.3 is on the effective size of changes under (6), 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. $\max_{1\le j\le q} |d_j| = O(1)$ and $D_n := \min_{1\le j\le q} d_j^2\,\delta_j \to \infty$ 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, 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 $\mathcal{E}_n \cap \mathcal{M}_n$, gSa returns $\widehat{\Theta} = \{\widehat{\theta}_j,\, 1 \le j \le \widehat{q} :\, \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}}\}$ 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 gSa achieves model selection consistency. Together, Theorems 2.1–3.1 lead to the consistency of WCM.gSa, the methodology combining WCM-based gappy model sequence generation and Schwarz criterion-based model selection steps. 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 Numerical results

4.1 Simulation results

Appendix C discusses in detail the choice of the tuning parameters for WCM.gSa. We investigate the performance of WCM.gSa on simulated datasets, in comparison with DeCAFS (Romano et al., 2021), DepSMUCE (Dette et al., 2020) and SNCP (Zhao et al., 2021) (the latter two applied with significance level $\alpha = 0.05$). Here, we present the results from three representative settings and defer the descriptions of the full simulation results (from thirteen scenarios with varying n, change point and serial dependence structures) and the competing methodologies to Appendix D, where we include DepSMUCE and SNCP applied with different choices of α as well as MACE proposed in Wu and Zhou (2020).

We generate 1000 realisations under each setting where $\varepsilon_t \sim_{\text{iid}} \mathcal{N}(0, 1)$. In addition to when f_t undergoes mean shifts as described below, we also consider the case where $f_t = 0$ to evaluate the size control performance.

(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 $\{Z_t\}$ follows an MA(1) model $Z_t = \varepsilon_t + b_1 \varepsilon_{t-1}$ with $b_1 = -0.9$.

(M2) 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: $Z_t = 0.75 Z_{t-1} -$

$0.5Z_{t-2} + \varepsilon_t + 0.8\varepsilon_{t-1} + 0.7\varepsilon_{t-2} + 0.6\varepsilon_{t-3} + 0.5\varepsilon_{t-4} + 0.4\varepsilon_{t-5} + 0.3\varepsilon_{t-6}$.

(M3) 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: $Z_t = a_1 Z_{t-1} + \sqrt{1 - a_1^2}\varepsilon_t$ with $a_1 = 0.9$.

Table 1 summarises the simulation results; see Table D.1 in Appendix for the full results where the exact definitions of RMSE and d_H can be found. Overall, across the various scenarios, WCM.gSa performs well both when $q = 0$ and $q \geq 1$. In particular, the proportion of the realisations where WCM.gSa detects spurious estimators in the absence of any mean shift is close to 0. Controlling for the size, especially in the presence of serial correlations, is a difficult task and as shown below, competing methods fail to do so by a large margin in some scenarios. When $q \geq 1$, WCM.gSa performs well in most scenarios according to a variety of criteria, such as model selection accuracy measured by $|\widehat{q} - q|$ or the localisation accuracy measured by d_H. We highlight the importance of the gappy model sequence generation step of Section 2.2: see the results reported under 'no gap' which refers to a procedure that omits this step from WCM.gSa and applies the Schwarz criterion-based model selection procedure directly to the model sequence consisting of consecutive entries from the WBS2-generated solution path. It suffers from having to perform a large number of model comparison steps and tends to over-estimate the number of change points in some scenarios.

DepSMUCE occasionally suffers from a calibration issue; in order not to detect spurious change points, it requires α to be set conservatively but for improved detection power, a larger α is better. In addition, the estimator of the LRV proposed therein tends to under-estimate the LRV when it is close to zero as in (M1), or when there are strong autocorrelations as in (M3), thus incurring a large number of falsely detected change points. Similar sensitivity to the choice of α is observable from SNCP. In addition, it tends to return spurious change point estimators when $q = 0$ in the presence of strong autocorrelations as in (M3), while under-detecting change points generally when $q \geq 1$ with the exception of (M1).

DeCAFS operates under the assumption that $\{Z_t\}_{t=1}^n$ is an AR(1) process. Therefore, it is applied under model mis-specification in some scenarios, but still performs reasonably well in not returning false positives. The exception is (M3) where, in the presence of strong autocorrelations, it returns spurious estimators over 50% of realisations even though the model is correctly specified in this scenario. Its detection accuracy suffers under model mis-specification in some scenarios such as (M1) and (M2) when compared to WCM.gSa, but DeCAFS tends to attain good MSE.

4.2 Nitrogen oxides concentrations in London

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 were 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 trained 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 WCM.gSa 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 WCM.gSa are plotted in Figure 1. For comparison,

Table 1: (M1)–(M11): We report the proportion of returning $\widehat{q} \geq 1$ when $q = 0$ (size) and the summary of estimated change points when $q > 1$ according to the distribution of $\widehat{q} - q$, relative MSE (RMSE) and the Hausdorff distance (d_H) over 1000 realisations. Methods that control the size at 0.05, and that achieve the best performance when $q > 1$ according to different criteria, are highlighted in **bold** for each scenario.

			$\widehat{q} - q$								
Model	Method	Size	≥ -3	-2	-1	0	1	2	$3 \leq$	RMSE	d_H
(M1)	WCM.gSa	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	68.720	1.988
	no gap	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	68.720	1.988
	DepSMUCE	1.000	0.000	0.000	0.000	0.485	0.167	0.163	0.185	219.196	48.359
	DeCAFS	0.064	0.000	0.006	0.029	0.742	0.148	0.053	0.022	304.694	26.274
	SNCP	**0.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**35.512**	**1.06**
(M2)	WCM.gSa	**0.001**	0.000	0.000	0.019	**0.873**	0.092	0.014	0.002	4.907	**34.627**
	no gap	**0.020**	0.002	0.002	0.012	0.178	0.024	0.037	0.745	11.030	148.765
	DepSMUCE	**0.031**	0.052	0.385	0.429	0.134	0.000	0.000	0.000	18.567	145.406
	DeCAFS	0.099	0.006	0.035	0.137	0.773	0.049	0.000	0.000	**3.891**	61.517
	SNCP	0.084	0.117	0.293	0.372	0.215	0.002	0.001	0.000	15.428	166.724
(M3)	WCM.gSa	**0.000**	0.087	0.177	0.233	0.319	0.076	0.041	0.067	3.184	86.139
	no gap	0.058	0.000	0.000	0.000	0.000	0.000	0.000	1.000	4.498	92.759
	DepSMUCE	0.936	0.767	0.153	0.070	0.010	0.000	0.000	0.000	8.655	139.298
	DeCAFS	0.565	0.000	0.004	0.019	**0.755**	0.203	0.017	0.002	**1.065**	**19.751**
	SNCP	0.258	0.956	0.034	0.007	0.003	0.000	0.000	0.000	11.698	290.266

we also report the change points estimated by DepSMUCE and DeCAFS, see Table 2.

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. This supports the hypothesis that the (de-trended and transformed) NO_2 and NO_x concentrations over the period in consideration, can plausibly be accounted for by a model with short-range dependence and multiple mean shifts; we refer to Mikosch and Stărică (2004), Berkes et al. (2006) Yau and Davis (2012) and Norwood and Killick (2018) for discussions on how weakly dependent time series with mean shifts may appear as a long-range dependent time series. In Appendix E.2, we further validate the set of change point estimators detected by WCM.gSa 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.

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

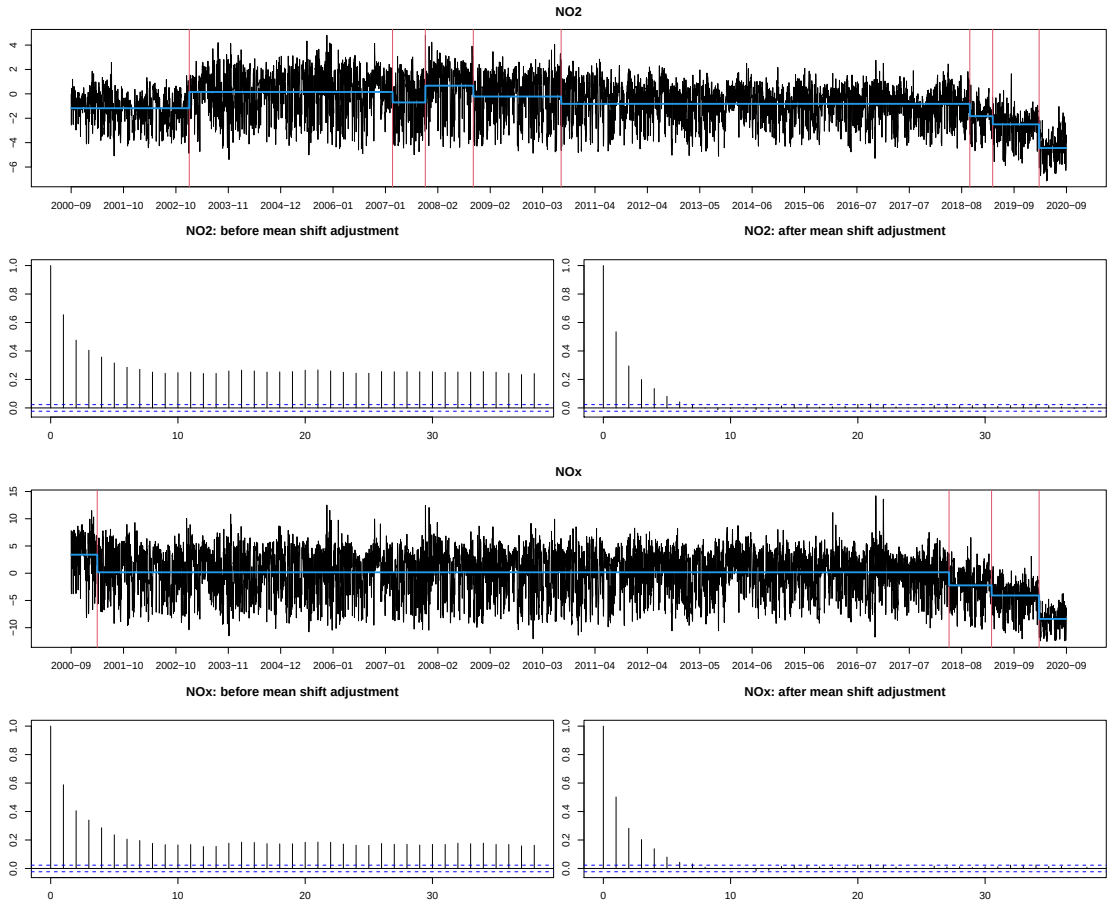

Figure 1: First (third) panel: daily average concentrations of NO_2 (NO_x) after transformation and de-trending, plotted together with the change points detected by WCM.gSa (vertical lines) and estimated piecewise constant mean (bold lines). Second (fourth) panel: autocorrelation function of transformed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

Table 2: 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) by each method 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
WCM.gSa	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**

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.

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

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.

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.

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. 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. (2021a). Data segmentation algorithms: Univariate mean change and beyond. *Econometrics and Statistics (in press)*.

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

Cho, H. and Korkas, K. K. (2021). High-dimensional GARCH process segmentation with an application to Value-at-Risk. *Econometrics and Statistics (in press)*.

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 non-stationary 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*.

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*.

Gallagher, C., Killick, R., Lund, R., and Shi, X. (2021). Autocovariance estimation in the

presence of changepoints. *arXiv preprint arXiv:2102.10669*.

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.

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.

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.

Lu, Q., Lund, R., and Lee, T. C. (2010). An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1):299–319.

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.

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.

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

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. (2021). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *Journal of the American Statistical Association (in press)*.

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

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.

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

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. (2020). Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 9(1):e318.

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 Davis, R. A. (2012). Likelihood inference for discriminating between long-memory and change-point models. *Journal of Time Series Analysis*, 33(4):649–664.

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.

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

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