Annals of the Institute of Statistical Mathematics (2022) 74:653–684
https://doi.org/10.1007/s10463-021-00811-5

Two-stage data segmentation permitting multiscale change points, heavy tails and dependence

Haeran Cho[1] · **Claudia Kirch**[2]

Received: 16 January 2021 / Revised: 29 May 2021 / Accepted: 26 August 2021 /
Published online: 25 September 2021

Abstract

The segmentation of a time series into piecewise stationary segments is an important problem both in time series analysis and signal processing. In the presence of multiscale change points with both large jumps over short intervals and small jumps over long intervals, multiscale methods achieve good adaptivity but require a model selection step for removing false positives and duplicate estimators. We propose a localised application of the Schwarz criterion, which is applicable with any multiscale candidate generating procedure fulfilling mild assumptions, and establish its theoretical consistency in estimating the number and locations of multiple change points under general assumptions permitting heavy tails and dependence. In particular, combined with a MOSUM-based candidate generating procedure, it attains minimax rate optimality in both detection lower bound and localisation for i.i.d. sub-Gaussian errors. Overall competitiveness of the proposed methodology compared to existing methods is shown through its theoretical and numerical performance.

Keywords Change point detection · Data segmentation · Schwarz criterion · Localised pruning · Multiscale procedure

✉ Haeran Cho
haeran.cho@bristol.ac.uk

Claudia Kirch
claudia.kirch@ovgu.de

[1] Institute for Statistical Science, School of Mathematics, Fry Building, University of Bristol, Bristol BS8 1UG, UK

[2] Department of Mathematics, Center for Behavioral Brain Sciences (CBBS), Institute for Mathematical Stochastics, Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany

1 Introduction

Change point analysis has a long tradition in statistics since Page (1954). In recent years, there has been a surge of interest for computationally fast and statistically efficient methods for change point analysis due to its importance in time series analysis, signal processing and many other applications where data are routinely collected over time in naturally nonstationary environments. In particular, many papers address the problem of testing for a change point, either retrospectively or sequentially, when at most one change is expected; see Csörgö and Horváth (1997) and Horváth and Rice (2014) for an overview. Based on such tests, the location of a single change point can be estimated with optimal localisation properties.

However, it is often unknown how many structural changes are present in the data, and allowing for multiple change points, the goal of change point analysis is to estimate both the total number and locations of the change points. Examples where data segmentation is popularly employed include genomics (detecting chromosomal copy number aberrations; see Olshen et al. (2004), Li et al. (2016), Niu and Zhang (2012), Chan and Chen (2017), neurophysiology [modelling the instabilities in the rate at which a neuron fires an action potential, Messer et al. (2014)], astronomy [detecting orbiting planets and their periodicity, Fisch et al. (2018)] and finance [identifying and dating change points in financial time series, Cho and Fryzlewicz (2012)], to name but a few.

Broadly, approaches to retrospective change point analysis in the literature can be categorised into two: one line of research relates to the aforementioned tests, while the other aims at optimising objective functions constructed on the principle of penalised likelihood or minimum description length, via dynamic programming (Killick et al. 2012; Maidstone et al. 2017) or genetic algorithm (Davis and Yau 2013). There are also methods based on hidden Markov models with algorithms for estimating the sequence of hidden states (Titsias et al. 2016). For an overview of the literature on data segmentation methods; see Cho and Kirch (2020).

Recent algorithmic developments include multiscale methodologies which focus on isolating each change point within an interval sufficiently large for its detection, whereby the tests and the estimators designed for the at-most-one-change alternatives are applicable to detect multiple change points. The wild binary segmentation (WBS) algorithm proposed in Fryzlewicz (2014) accomplishes this by drawing a large number of random intervals. Eichinger and Kirch (2018) investigate a moving sum (MOSUM) procedure which systematically tests for at most a single change point over moving windows at a single bandwidth, and briefly discuss its multiscale extension for better adaptivity. On the one hand, such multiscale methods enjoy the near-optimal localisation of change points through scanning the same regions of the data at multiple resolutions. On the other, this may result in conflicting (duplicate) estimators detected for the identical change point, as well as false positives spuriously detected without any change points in their vicinity, which makes a model selection step inevitable.

There exist post-processing and pruning procedures specifically tailored for particular multiscale candidate generating methods and settings to handle false

positives and duplicates, but there is a lack of a unified approach to this task. In this paper, we propose a generic methodology for this purpose, which utilises the Schwarz criterion (Schwarz 1978) and performs an exhaustive search for change point estimators in a *localised* way on a candidate set generated by multiscale methods. Contrary to the common usage of information criteria in change point problems, the proposed localised pruning algorithm does not require the maximum number of change points as an input, nor does it seek for the global minimiser of the criterion which is computationally costly.

We show that as a generic tool, the localised pruning algorithm inherits the properties of the candidate generating method. Therefore, with a suitable, multiscale candidate generating method, it consistently estimates the total number of change points as well as locating the change points with accuracy while being computationally feasible. In this paper, we verify the suitability of two candidate generating multiscale methods based on the MOSUM and cumulative sum (CUSUM) statistics; the implementation of the algorithm combining the localised pruning with the former is available in the R package `mosum` (Meier et al. 2021a), with an accompanying paper detailing its efficient implementation (Meier et al. 2021b).

1.1 Main contributions

Below, we summarise the main contributions made in this paper.

(a) **Two-stage procedure.** We explicitly separate the statistical analysis of the candidate generating method (Stage 1, see Sect. 4) from that of the model selection (pruning) methodology (Stage 2, see Sect. 3). This allows us (i) to easily extend our statistical conclusions to different candidate generating methods, and (ii) to gain insights into the assumptions required for each stage separately.

(b) **Truly multiscale change points.** In contrast to the assumptions commonly found in the literature that require homogeneity on the change point structure, we adopt a truly multiscale setting that accommodates the situation when both large changes over short stretches of stationarity, as well as small changes over long stretches of stationarity are present simultaneously in the signal; see Definition 1.

(c) **Minimax optimality.** We show that the proposed localised pruning, combined with a MOSUM-based multiscale candidate generating mechanism, achieves minimax optimality in change point localisation as well as matching the rate of the minimax detection lower bound when the errors are distributed as i.i.d. sub-Gaussian random variables; see Corollary 2.

(d) **Assumptions on the error distribution.** We provide insights into which stochastic properties of the error distribution affect the detection lower bound and the localisation rate of the proposed methodology, which allow for very general assumptions on the error distribution permitting both serial dependence and heavy tails beyond the i.i.d. (sub-)Gaussianity commonly imposed in the literature; see Assumption 1.

(e) **Universally competitive performance in simulations and data analysis.** For a range of test signals of varying length, frequency of change points and error

distributions, the proposed method performs uniformly well in both model selection consistency and localisation accuracy, and within reasonable computation time (see Sect. 5.1). Applied to real data examples, our procedure is capable of handling the issues often encountered in practice such as heteroscedasticity and low signal-to-noise ratio. We provide its implementation with a MOSUM-based candidate generating procedure in the R package `mosum` available on CRAN (Meier et al. 2021a).

(f) **Computational complexity.** The computational complexity of the localised pruning algorithm with the MOSUM-based candidate generating method is given by $O(n\log(n))$, which is comparable to or much lower than that of most competing methods (see Table 1). With other candidate generating methods, the computational complexity of the combined procedure will effectively be determined by that of the first-stage candidate generation.

The problem of detecting multiple change points in the mean has been extensively studied in the literature, often laying the groundwork for generalisations to more complex and high-dimensional problems. The proposed localised pruning methodology has been constructed with such extensions in view, and we discuss these possibilities in Sect. 6.

The rest of the paper is organised as follows: In Sect. 2, we define a truly multiscale change point problem and introduce the assumptions for theoretical

Table 1 Comparison of change point detection methodologies on the rates of detection lower bound and localisation derived under (sub-)Gaussianity where $\delta_n = \min_{1\le j\le q_n}\delta_j$, and whether they are formulated in a multiscale way according to Definition 1

Methodology	Detection lower bound		Localisation		Computational complexity	Beyond sub-Gaussianity
	Multiscale	Rate	Multiscale	Rate		
MoLP	✓	$\log(\boldsymbol{n})$	✓	$\log(\boldsymbol{q_n})$	$\boldsymbol{O(n\log(n))}$	✓
Chan and Chen (2017)	×	$\log(n/\delta_n)$	×	$\log(n)$	$O(n\log(n))$	×
Single-scale MOSUM	×	$\log(n/\delta_n)$	✓	$\log(q_n)$	$O(n)$	✓
Fromont et al. (2020)	✓	$\log(n/\delta_n)$	✓	$\log(q_n)$	$O(n^2)$	×
Wang et al. (2020b)†	×	$\log(n)$	✓	$\log(n)$	$O(n^2)$	×
Wang et al. (2020b)∗	×	$\log(n)$	✓	$\log(n)$	$O(nR_n)$ with	×
Baranowski et al. (2019)	×	$\log(n)$	✓	$\log(n)$	$(n/\delta_n)^2/R_n \to 0$	×
Frick et al. (2014)	×	$\log(n/\delta_n)$	×	$\log(n)$	$O(n^2)$	✓
Li et al. (2019)	×	$q_n\log(n)$	×	$q_n\log(n)$	–	×
Fryzlewicz (2018)	×	$\log^2(n)$	×	$\log^2(n)$	$O(n\log^2(n))$	✓

We also provide their computational complexity, and whether their theoretical guarantee goes beyond the (sub-)Gaussian setting. Wang et al. (2020b) †refers to their ℓ_0-penalised LSE estimator, while Wang et al. (2020b) ∗refers to their modified WBS

consistency. Also, we present the minimax optimality results available from the literature, and provide a comparative study of our proposed methodology and those shown to be near-minimax optimal. In Sect. 3, we motivate and propose the localised pruning as a generic methodology applicable with a class of candidate generating mechanisms and establish its theoretical consistency. Section 4 discusses a MOSUM-based candidate generating procedure and shows the minimax optimality of the combined two-stage methodology. In Sect. 5, we briefly summarise the simulation studies and apply the proposed methodology to a genomic dataset. Section 6 concludes the paper. The proofs of the theoretical results, discussion of an alternative, CUSUM-based candidate generating procedure related to the WBS (Fryzlewicz 2014), complete simulation results and additional real data example are provided in the Supplementary Appendix.

1.1.1 Notations

Throughout the paper, we adopt ν_n to denote a sequence satisfying $\nu_n \to \infty$ at an arbitrarily slow rate, which may differ from one occasion to another. We adopt the notation $a_n \asymp b_n$ to denote that $a_n = O(b_n)$ and $b_n = O(a_n)$. For convenience, the assumptions are formulated with asymptotic arguments but the proofs work directly with non-asymptotic conditions on the corresponding quantities on the set $\mathcal{M}_n$ defined in Theorem 1 (collected in Eq (C.1) of the supplementary document), making constants traceable in principle.

2 Multiscale change point analysis

2.1 Multiscale change point detection problem

We consider the canonical change point model

$$X_t = f_t + \varepsilon_t = f_0 + \sum_{j=1}^{q_n} d_j \cdot \mathbb{I}_{t \geq \theta_j + 1} + \varepsilon_t, \tag{1}$$

where $\theta_1 < \theta_2 < \ldots < \theta_{q_n}$ with $\theta_j = \theta_{j,n}$ denote the q_n change points (with $\theta_0 = 0$ and $\theta_{q_n+1} = n$), at which the mean of X_t undergoes changes of size $|d_j|$ where, again, $d_j = d_{j,n}$. We denote by $\delta_j = \delta_{j,n} = \min(\theta_j - \theta_{j-1}, \theta_{j+1} - \theta_j)$ the minimum distance of θ_j to its neighbouring change points, and by $\Theta = \Theta_n = \{\theta_1, \ldots, \theta_{q_n}\}$ the set of change points. The sequence of errors $\{\varepsilon_t\}_{t=1}^n$ satisfies $\mathsf{E}(\varepsilon_t) = 0$ and is allowed both serial dependence and heavy-tailedness as specified later. We assume that $\max_{1 \leq j \leq q_n} |d_j| = O(1)$ as well as $\min_{1 \leq j \leq q_n} \delta_j \to \infty$, separating the problem of change point detection under (1) from that of outlier detection; see Cho and Kirch (2020) for further discussion on this point.

Operating under (1), a change point detection methodology is deemed consistent if it returns a set of change point estimators $\widehat{\Theta} = \{\widehat{\theta}_j,\ 1 \leq j \leq \widehat{q} : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}}\}$ which satisfies

$$\mathrm{P}\left\{\widehat{q} = q_n \text{ and } \max_{1\le j\le q_n} w_j|\widehat{\theta}_j - \theta_j| \le \rho_n\right\} \to 1 \quad \text{as} \quad n \to \infty$$

for suitable w_j and ρ_n which fulfil at least $(w_j n)^{-1}\rho_n \to 0$ (a more detailed discussion from the minimax perspective is given in Sect. 2.3). Here, the weight w_j is related to the squared magnitude of the change, d_j^2, and thus signifies the difficulty associated with localising individual change points θ_j. In combination with this weight, ρ_n denotes the rate of localisation. The above consistency is typically established under some conditions on how fast the *detection lower bound* Δ_n, relating the squared magnitude of the change d_j^2 to the minimum distance to adjacent change points δ_j, diverges as $n \to \infty$.

In this paper, our interest lies in studying the performance of the proposed change point detection methodology in a truly multiscale, heterogeneous change point setting, by formulating the associated detection lower bound such that signals containing both frequent large jumps as well as small jumps over long stretches of stationarity are allowed. Definition 1 distinguishes multiscale formulations of the detection lower bound and localisation rate from their non-multiscale counterparts.

Definition 1

(a) **Detection lower bound and separation rate.** We distinguish the following change point scenarios that are linked to different detection lower bounds: Changes are detectable with asymptotic power one as soon as Δ_n defined below diverges faster than the separation rate associated with a given methodology.

(i) **Homogeneous change points:** $\Delta_n = \min_{1\le j\le q_n} d_j^2 \cdot \min_{1\le j\le q_n} \delta_j$.

(ii) **Finite mixture of homogeneous change points:** There are $N < \infty$ disjoint subsets of change points with their indices given by $\mathcal{J}_k$, $k = 1, \dots, N$, such that $\bigcup_{k=1}^N \mathcal{J}_k = \{1, \dots, q_n\}$, whereby change points within each subset are homogeneous as defined in (i) and $\Delta_n = \min_{1\le k\le N}(\min_{j\in\mathcal{J}_k} d_j^2 \cdot \min_{j\in\mathcal{J}_k} \delta_j)$. When there are finitely many changes ($q_n = N$) is a special case.

(iii) **Multiscale change points:** $\Delta_n = \min_{1\le j\le q_n} d_j^2\, \delta_j$.

(b) **Localisation rate:** We distinguish between a **homogeneous localisation rate** where the estimation error in localising the jth change point is weighted globally with $w_j = \min_{1\le j\le q_n} d_j^2$, and a **multiscale localisation rate** where it is weighted locally with $w_j = d_j^2$.

Definition 1 (a) shows different extensions of the assumption $d_1^2 \min(\theta_1, n - \theta_1) \to \infty$ commonly found in the change point testing literature [where $q_n = 1$ at most; see e.g. Csörgö and Horváth (1997)]. Proceeding from (i) to (iii), the associated parameter space becomes more general and only (iii) truly requires multiscale methods that scan the data for change points at diverging number of scales. Nevertheless, most papers in the change point detection literature

formulate the detection lower bound for the homogeneous setting only (see Table 1 and Sect. 2.4). Theoretical guarantees for some methodologies considered therein may be extended to accommodate the multiscale change points in (iii), while some cannot (see Appendix E in the supplementary document for the discussion on the WBS). For our proposed methodology, we adopt the most general setting and impose an assumption on the size of changes correspondingly (see Assumption 2).

The multiscale localisation rate in (b) reflects that the difficulty in accurate localisation of each change point depends on the corresponding jump size only.

2.2 Main assumptions

The mathematical analysis in this paper is based on the following properties of the error distributions only, which makes the results very general permitting e.g. heavy tails, dependence and even non-stationarity.

Assumption 1 (*Error distribution*) We assume that $\{\varepsilon_t\}_{t=1}^n$ is ergodic with $\mathsf{E}(\varepsilon_t) = 0$ and $0 < c \leq \mathsf{Var}(\varepsilon_t) \leq C < \infty$ for some $c, C > 0$. Further:

(a) For some ω_n satisfying $\sqrt{\log(n)} = O(\omega_n)$, let $\mathsf{P}(\mathcal{M}_n^{(11)}) \to 1$ where

$$\mathcal{M}_n^{(11)} = \left\{ \max_{0 \leq s < e \leq n} \frac{1}{\sqrt{e-s}} \left| \sum_{t=s+1}^{e} \varepsilon_t \right| \leq \omega_n \right\}.$$

(b) For any sequences $1 \leq a_n, b_n \leq D_n$ with D_n defined in Assumption 2, let $\mathsf{P}(\mathcal{M}_n^{(12)} \cap \mathcal{M}_n^{(13)}) \to 1$ where

$$\mathcal{M}_n^{(12)} = \left\{ \max_{1 \leq j \leq q_n} \max_{d_j^{-2} a_n \leq \ell \leq \theta_j - \theta_{j-1}} \frac{\sqrt{d_j^{-2} a_n}}{\ell} \left| \sum_{t=\theta_j - \ell + 1}^{\theta_j} \varepsilon_t \right| \leq \omega_n^{(1)} \right\}$$
$$\bigcap \left\{ \max_{1 \leq j \leq q_n} \max_{d_j^{-2} a_n \leq \ell \leq \theta_{j+1} - \theta_j} \frac{\sqrt{d_j^{-2} a_n}}{\ell} \left| \sum_{t=\theta_j + 1}^{\theta_j + \ell} \varepsilon_t \right| \leq \omega_n^{(1)} \right\}, \quad \text{and}$$

$$\mathcal{M}_n^{(13)} = \left\{ \max_{1 \leq j \leq q_n} \max_{1 \leq \ell \leq d_j^{-2} b_n} \frac{1}{\sqrt{d_j^{-2} b_n}} \left| \sum_{t=\theta_j - \ell + 1}^{\theta_j} \varepsilon_t \right| \leq \omega_n^{(2)} \right\}$$
$$\bigcap \left\{ \max_{1 \leq j \leq q_n} \max_{1 \leq \ell \leq d_j^{-2} b_n} \frac{1}{\sqrt{d_j^{-2} b_n}} \left| \sum_{t=\theta_j + 1}^{\theta_j + \ell} \varepsilon_t \right| \leq \omega_n^{(2)} \right\}.$$

Remark 1

(a) The lower bound on ω_n in Assumption 1 (a) is quite natural in light of Theorem 1 of Shao (1995) which derives the corresponding result for i.i.d. random variables whose moment-generating function exists. The bound ω_n is closely linked to the detection lower bound of our proposed methodology as shown in Assumption 2.
(b) The rates $\omega_n^{(1)}$ and $\omega_n^{(2)}$ are closely connected with the localisation rate of the localised pruning method (see Assumption 4) for the precise statement. Also, the bound $\omega_n^{(1)}$ gives the rate of localisation for the multiscale MOSUM procedure considered as one of the candidate generating mechanisms in Sect. 4. Note that $\omega_n^{(1)}$ and $\omega_n^{(2)}$ are always dominated by ω_n and are often much smaller, particularly in the presence of heavy tails and when q_n is bounded (see Proposition 1 for specific examples).
(c) The bounds for the respective second set in $\mathcal{M}_n^{(12)}$ and $\mathcal{M}_n^{(13)}$ follow from the bounds of the first set in the case of i.i.d. errors, but this is not necessarily so for time series errors.

Assumption 2 (*Multiscale lower bound on the size of changes*) For $D_n := \min_{1 \le j \le q_n} d_j^2 \, \delta_j$, we require $D_n^{-1} \omega_n^2 \to 0$ for ω_n as in Assumption 1. In addition, D_n dominates the penalty used in the localised pruning algorithm (see Assumption 3).

The next proposition provides the exact rates for ω_n, $\omega_n^{(1)}$ and $\omega_n^{(2)}$ in Assumption 1 for some special cases.

Proposition 1 *In all follows, $\nu_n \to \infty$ arbitrarily slow.*

(a) ***Sub-Gaussianity.*** *Let $\{\varepsilon_t\}_{t=1}^n$ be a sequence of i.i.d. random variables following a sub-Gaussian distribution as defined e.g. in Section 2.5 of Vershynin (2018). Then, Assumption 1 holds with $\omega_n \asymp \sqrt{\log(n)}$ and $\omega_n^{(1)} = \omega_n^{(2)} \asymp \max(\sqrt{\log(q_n)}, \nu_n)$.*
(b) ***Heavy tails.*** *Let $\{\varepsilon_t\}_{t=1}^n$ be a sequence of i.i.d. regularly varying random variables with index of regular variation $\alpha > 0$ as defined e.g. in Mikosch and Račkauskas (2010). Then Assumption 1 holds with $\omega_n \asymp n^{1/\beta}$ and $\omega_n^{(1)} = \omega_n^{(2)} \asymp \max(q_n^{1/\beta}, \nu_n)$ for any $\beta < \alpha$.*
(c) *In the following situations, Assumption 1 holds with the rates given below which, however, are generally not tight:*

  (i) ***Invariance principle.*** *If there exists (possibly after changing the probability space) a standard Wiener process $W(\cdot)$ such that $\sum_{t=1}^{\ell} \varepsilon_t - W(\ell) = O(\lambda_\ell)$ a.s. with $\lambda_\ell = o(\sqrt{\ell})$, then Assumption 1 (a) holds with $\omega_n \asymp \max(\lambda_n \nu_n, \sqrt{\log(n)})$.*
  (ii) **Moment conditions.** *If $\mathsf{E}|\sum_{t=l+1}^{r} \varepsilon_t|^\gamma \le C(r-l)^{\gamma/2}$ for any $-\infty < l < r < \infty$ and some constants $C > 0$ and $\gamma > 2$, then Assumption 1 (b) holds with $\omega_n^{(1)} = \omega_n^{(2)} \asymp q_n^{1/\gamma} \nu_n$.*

Remark 2

(a) In Proposition 1 (a)–(b), the term ν_n can be ignored in the requirement on $\omega_n^{(1)}$ and $\omega_n^{(2)}$ when $q_n \to \infty$. If q_n is fixed, $\omega_n^{(1)}$ and $\omega_n^{(2)}$ can diverge arbitrarily slowly.

(b) For regularly varying jump size distributions, ω_n in Proposition 1 (b) cannot be improved beyond $\omega_n = n^{1/\alpha} L(n)$ for some slowly varying function L (see Theorem 1.1 of Mikosch and Račkauskas (2010) and Proposition B.1.9 (9) of De Haan and Ferreira (2007)). For dependent errors, similar results are derived in Mikosch and Moser (2013). Furthermore, in the special case of a t-distribution with α degrees of freedom, then Assumption 1 holds with $\omega_n \asymp n^{1/\alpha}$ (Schlüter and Fischer 2009, Section 4.2).

(c) Invariance principles as in Proposition 1 (c.i) have been derived for a variety of situations including dependent data under weak dependency conditions such as mixing (Kuelbs and Philipp 1980, Theorem 4) and functional dependence measure conditions (Berkes et al. 2014), to name but a few. The rate λ_n is typically directly linked to the number of moments that exist, e.g. for i.i.d. errors, $\lambda_\ell = \log(\ell)$ if the moment generating function exists, and $\lambda_\ell = \ell^{1/(2+\Delta)}$ if $\mathsf{E}(\varepsilon_t^{2+\Delta}) < \infty$ (Komlós et al. 1975, 1976). Comparing the rate of ω_n in Proposition 1 (c.i) with the one in 1 shows that the rates from the invariance principle are usually not tight.

Moment conditions as in Proposition 1 (c.ii) have been shown for many time series; see e.g. Appendix B.1 in Kirch (2006).

2.3 Minimax optimality

In this section, we state the benchmark for the minimax optimal separation and localisation rates. The following result is from Proposition 1 of Arias-Castro et al. (2011).

Proposition 2 (Lower bound on the minimax separation rate) Under (1), *let* $H_{0,n} : q_n = 0$ *and* $H_{1,n}$ *describe the setting where* $q_n = 2$, $d_n := d_1 = -d_2$ *and* $\delta_n := \theta_2 - \theta_1$ *with* $n^{-1}\delta_n \to 0$. *Then,* $H_{0,n}$ *and* $H_{1,n}$ *are asymptotically inseparable if* $|d_n|\sqrt{\delta_n} \le \sqrt{2\log(n/\delta_n)} - \nu_n$ *where* $\nu_n \to \infty$.

Proposition 2 provides an instance under (1) where the change points are not detectable by any method. Together with Table 1, which provides a summary of various methodologies for multiple change point detection including their separation rates under the column 'Detection lower bound', the proposition shows that the minimax optimal separation rate is given by $\log(n/\delta_n)$ for the detection lower bounds defined in Definition 1 (a). In the case of sublinear changes, where $\max_{1\le j\le q_n} \delta_j = O(n^{1-\kappa})$ for some $\kappa > 0$, this rate amounts to $\log(n)$.

The next proposition is from Proposition 6 of Fromont et al. (2020) which is stated here with an enlarged parameter space for ready comparison.

Proposition 3 (Lower bound on the minimax localisation rate for possibly an unbounded number of change points) *Under* (1), *let* $|d_j| =: d_n$ *for all* $j = 1, \ldots, q_n$ *with* $q_n \geq 2$, *and denote by* $\Xi = \{(\theta_1, \ldots, \theta_{q_n}) : 0 \equiv \theta_0 < \theta_1 < \ldots < \theta_{q_n} < \theta_{q_n+1} \equiv n$ and $d_n^2 \min_{1 \leq j \leq q_n} (\theta_{j+1} - \theta_{j-1}) > c \log(q_n)\}$ *for some* $c > 0$, *the parameter space for the locations of change points. Then, for some* $C > 0$,

$$\inf_{\mathcal{K} \in \mathbb{N}^{q_n}} \sup_{\Theta \in \Xi} \mathsf{E}_\Theta \{d_H(\mathcal{K}, \Theta)\} \geq C d_n^{-2} \log(q_n)$$

where $d_H(\mathcal{K}, \Theta) = \max\{\max_{k \in \mathcal{K}} \min_{\theta \in \Theta} |k - \theta|, \max_{\theta \in \Theta} \min_{k \in \mathcal{K}} |\theta - k|\}$, *denotes the Hausdorff distance*.

This proposition, together with the results reported in Table 1 (under the column 'Localisation'), establishes that the minimax optimal localisation rate is given by $\log(q_n)$ when the number of change points q_n is permitted to diverge with n.

Both Propositions 2–3 are derived under the special case where $\{\varepsilon_t\}_{t=1}^n$ are i.i.d. random variables following a (sub-)Gaussian distribution. To the best of our knowledge, there do not exist equivalent results on the detection lower bound or the localisation rate (when $q_n \to \infty$) beyond the i.i.d. sub-Gaussianity. We show that under sub-Gaussianity, the two-stage procedure combining a MOSUM-based candidate generating method and the proposed localised pruning algorithm, achieves minimax optimal rates in both localisation and detection lower bound, the latter in the sublinear change point setting where $\log(n/\delta_j) \asymp \log(n)$ for each j; if $\min_{1 \leq j \leq q_n} \delta_j$ is (near-) linear, the rate it attains is greater than the minimax optimal rate by the factor of $\log(n)$ at most. Further, even in the presence of heavy-tailed errors and dependence, we obtain the same localisation rate as in the sub-Gaussian setting when there are finitely many change points (i.e. q_n is finite); see Corollary 2. This rate is then automatically minimax optimal also. We note that once the results equivalent to Propositions 2–3 become available in more general settings permitting heavier tails and serial dependence, the theoretical properties we derive for the proposed methodology under Assumption 1 are general enough to be immediately compared to such a benchmark.

2.4 Comparison with the existing literature

There exist various data segmentation algorithms which are shown to be near-minimax optimal in detecting and locating multiple change points. Here, we concentrate on procedures for univariate time series with changes in the mean, some of which have been extended to e.g. high-dimensional change point detection problems (see Sect. 6). Frick et al. (2014) and Li et al. (2016) propose procedures that are termed as multiscale change point segmentation methods in Li et al. (2019); noting empirical and theoretical limitations of the WBS as proposed in Fryzlewicz (2014), Baranowski et al. (2019) and Wang et al. (2020b) propose modifications of the WBS which require additional tuning parameters such as a threshold or a lower bound on $\delta_n := \min_{1 \leq j \leq q_n} \delta_j$; Boysen et al. (2009), Wang et al. (2020b) and Fromont et al. (2020) investigate an ℓ_0-penalised least squares (LSE) estimator, the former two with the Schwarz criterion-type penalty and

the latter with an adaptive one; Chan and Chen (2017) propose two methods, where one bears some resemblance to a multiscale MOSUM procedure with 'bottom-up' merging [see also Messer et al. (2014)] while the other to the tail-greedy unbalanced Haar (TGUH) method of Fryzlewicz (2018). All the papers discussed above present their theoretical findings under the assumption that $\{\varepsilon_t\}_{t=1}^n$ is a sequence of i.i.d. (sub-) Gaussian random variables, with the exception of Fryzlewicz (2018) providing the results under heavy-tails and serial dependence, and Frick et al. (2014) allowing for i.i.d. errors following exponential family distributions; an extension of their results to dependent error processes is studied in Dette et al. (2020).

Table 1 provides an overview of these methodologies alongside the localised pruning applied with a multiscale MOSUM procedure for candidate generation (referred to as 'MoLP'), on their theoretical performance, computational complexity and generality beyond the sub-Gaussian setting. Boysen et al. (2009) assume that $|d_j|$ and δ_j/n are bounded away from zero, and thus we exclude it from the table. We also note that the separation rate reported in Wang et al. (2020b) is slightly larger than $\log(n)$ by a logarithmic factor, and the requirement on R_n for the computational complexity associated with Wang et al. (2020b)$*$ and Baranowski et al. (2019) is also slightly stronger by a logarithmic factor.

Apart from the current paper and Fromont et al. (2020), all others derive the separation rates only for the case of homogeneous change points according to Definition 1 (a). Most procedures achieve the minimax optimal separation rate for sublinear changes (see Proposition 2 and the discussion below), and Chan and Chen (2017), Fromont et al. (2020) and Frick et al. (2014) slightly improve upon this when δ_n, the minimal distance between change points, is (near-)linear; see also Chan and Walther (2013) where similar observations are made on scan likelihood ratio statistic for a signal detection problem. Extension of our result beyond the sublinear setting would require the adoption of a scale-dependent penalty as in Fromont et al. (2020) for the pruning methodology. To the best of our knowledge, such a choice of penalty is available only for light-tailed errors, and its extension to the general error distribution we consider in this paper has not been investigated in the literature.

The localisation rates are obtained in a multiscale formulation [according to Definition 1 (b)] by most methods but not all, and many achieve only near-minimax optimality in multiple change point localisation (see Proposition 3). In particular, their localisation rates are worse by the factor of $\log(n)$ or more, when there are a finite number of change points. Exceptions are the MoLP, the penalised LSE of Fromont et al. (2020) and the single-scale MOSUM procedure, which achieve the exact minimax optimal localisation rate of $\log(q_n)$. Our proposed method achieves this with the computational complexity of $O(n\log(n))$ rather than $O(n^2)$ required for solving the ℓ_0-penalised least squares estimation problem; we defer a detailed discussion on the computational complexity to Appendix F of the supplementary material. Additionally, theoretical analysis in this paper is conducted under Assumption 1 that permits heavy-tailed and serially correlated errors, which sets our paper apart from the rest.

3 Localised pruning via Schwarz criterion

Our goal is to estimate both the total number q_n and the locations of the change points θ_j, $j = 1, \ldots, q_n$ under (1). For this purpose, we introduce a generic, localised pruning methodology which, applicable to a set of candidate change point estimators returned by multiscale change point procedures, achieves consistent estimation of multiple change points in their total number and locations.

Many multiscale change point procedures are based on the principle of isolating each change point for its detection and estimation, and typically attach extra information to change point estimators about their detection intervals. Such examples include the multiscale extension of the MOSUM procedure (Eichinger and Kirch 2018) and the WBS (Fryzlewicz 2014). The MOSUM procedure scans a series of MOSUM statistics

$$T_{b,n}(G;X) := \sqrt{\frac{G}{2}}\big(\bar{X}_{(b-G+1):b} - \bar{X}_{(b+1):(b+G)}\big) \tag{2}$$

where $\bar{X}_{s:e} = (e-s+1)^{-1}\sum_{t=s}^{e} X_t$, for a given bandwidth G and $G \le b \le n-G$, and marks as change point candidates the locations where $|T_{b,n}(G;X)|$ simultaneously exceeds a critical value and forms local maxima; thus each candidate estimator k is associated with its natural detection interval $\mathcal{I}_N(k) = (k-G, k+G]$. The WBS examines the CUSUM statistics

$$\mathcal{X}_{s,b,e} \equiv \mathcal{X}_{s,b,e}(X) = \sqrt{\frac{(b-s)(e-b)}{e-s}}\big(\bar{X}_{(s+1):b} - \bar{X}_{(b+1):e}\big) \tag{3}$$

for $s+1 \le b \le e-1$ over a large number of randomly drawn intervals $(s, e] \subset [1, n]$. The maximiser of the CUSUM statistics $k = \arg\max_{s<b<e} |\mathcal{X}_{s,b,e}|$ can be regarded as a change point candidate if the test statistic $|\mathcal{X}_{s,k,e}|$ exceeds a certain threshold, and the interval $\mathcal{I}_N(k) = (s, e]$ is readily associated with its detection.

In what follows, we describe the proposed localised pruning methodology assuming that a set of candidate estimators $\mathcal{K}$ is given. Specific candidate generating methods are discussed in Sect. 4 and Appendix E of the supplementary material.

3.1 Methodology

Let $\mathcal{K}$ denote the set of all the candidate change point estimators to be pruned down. For each $k \in \mathcal{K}$, we denote the detection interval of k by $\mathcal{I}(k) \equiv (k - G_L, k + G_R]$, where the left detection distance $G_L = G_L(k)$ is the distance from k to the leftmost point of the interval, and the right detection distance $G_R = G_R(k)$ is defined analogously.

Information criteria are frequently adopted for model selection in change point problems, and we adopt the Schwarz criterion (Schwarz 1978, SC) for this purpose. For a given set of change point candidates $\mathcal{A} = \{\tilde{k}_1 < \ldots < \tilde{k}_m\} \subset \mathcal{K}$, the SC is evaluated as

$$\mathrm{SC}(\mathcal{A}) = \frac{n}{2}\log\left\{\frac{\mathrm{RSS}(\mathcal{A})}{n}\right\} + |\mathcal{A}| \cdot \xi_n, \tag{4}$$

where it balances between the goodness-of-fit measured by the residual sum of squares

$$\mathrm{RSS}(\mathcal{A}) = \sum_{j=0}^{m}\sum_{t=\tilde{k}_j+1}^{\tilde{k}_{j+1}}\left(X_t - \bar{X}_{(\tilde{k}_j+1):\tilde{k}_{j+1}}\right)^2 \quad \text{with} \quad \tilde{k}_0 = 0 \text{ and } \tilde{k}_{m+1} = n,$$

and the penalty imposed on the model complexity $|\mathcal{A}|$.

Assumption 3 (*Penalty*) The penalty parameter ξ_n satisfies

$$\frac{\xi_n}{D_n} \to 0 \quad \text{and} \quad \frac{\omega_n^2}{\xi_n} \to 0,$$

where ω_n and D_n are as in Assumptions 1 (a) and 2, respectively.

The assumption shows the connection between the penalty parameter ξ_n, the noise level ω_n and the detection lower bound D_n. For i.i.d. sub-Gaussian random variables, the rate of ω_n in Proposition 1 (a) cannot be improved (Shao 1995, Theorem 1) and thus the (strengthened) Schwarz penalty of $\xi_n = \log^{1+\Delta}(n)$ with some $\Delta > 0$ can be allowed by Assumption 3 [see e.g. Yao (1988) and Fryzlewicz (2014)]. Proposition 1 (b) and Remark 2 (b) indicate that a penalty stronger than logarithmic in n is required for heavy-tailed errors in order to guarantee consistent estimation of the number of change points by means of the SC, an observation also made by Kühn (2001).

In the literature, exhaustive minimisation of an information criterion over all $\mathcal{A} \subset \mathcal{K}$ for a given candidate set $\mathcal{K}$ has been considered as a model selection method (see e.g. Niu and Zhang (2012), Chan et al. (2014) and Yau and Zhao (2016)). Such an exhaustive approach may result in a computationally inhibitive search space as its size grows exponentially with $|\mathcal{K}|$. Moreover, it does not utilise the information immediately available about the detection intervals of change point estimators. For example, if the detection interval of a candidate k does not overlap with that of any other estimator, there is little to be gained by having k considered alongside other candidates in the evaluation of SC. On the other hand, if $\mathcal{I}(k)$ overlaps with the detection interval of another candidate, say k', it is possible that k and k' are conflicting estimators of the identical change point, which justifies the joint consideration of the two.

Based on these observations, we propose the localised pruning methodology consisting of two nested algorithms, where the outer algorithm iteratively selects the local environment on which the inner algorithm performs the pruning.

3.1.1 Outer algorithm: localisation (LocAlg)

Taking the set of change point candidates $\mathcal{K}$ as an input, the outer algorithm for localisation iteratively selects a subset of candidates to be pruned down by the inner algorithm (PrunAlg) described in Sect. 3.1.2. For this, the algorithm sorts the candidates in $\mathcal{K}$ according to a sorting function h. One possibility is to use the jump size associated with each $k \in \mathcal{K}$, which is calculated within the detection interval $\mathcal{I}(k) = (k - G_L(k), k + G_R(k)]$ as

$$h_{\mathcal{J}}(k) = \left| \bar{X}_{(k-G_L(k)+1):k} - \bar{X}_{(k+1):(k+G_R(k))} \right|. \tag{5}$$

If (asymptotic) null distributions of the test statistics are available, another possibility is to use the inverse of the p-values, say $h_{\mathcal{P}}$, as a sorting function. Either with $h_{\mathcal{J}}$ or $h_{\mathcal{P}}$, additional tie-breaking rules can be employed, e.g. by preferring the candidates associated with the smallest detection interval according to $G_L(k) + G_R(k)$, $G_L(k)$ or $G_R(k)$; if there are still ties, an arbitrary choice can be made. When some candidate k is detected at different scales, the sorting function and the tie-breaking rule will select only a single instance of k, and any other duplicates will be removed in Step 4 of LocAlg given below. Our theoretical results do not depend on the choice of the sorting function or the tie-breaking rule (see Theorem 1).

Denote by $\mathcal{C}$ the candidates for which no decision has been reached yet, and by $\widehat{\Theta}$ the set of already accepted candidates. At the beginning of the algorithm, the active candidate set $\mathcal{C}$ is given by the complete candidate set $\mathcal{K}$ and $\widehat{\Theta}$ is set to be empty. Then, the outer algorithm iteratively processes the candidates in the following way.

Step 1 **Find the most prominent candidate**. According to a sorting function h (and tie-breakers if necessary), find a candidate $k_\circ \in \mathcal{C}$ from the active candidate set that maximises h.
Step 2 **Define the local search environment**. Find k_L that is closest to $k_\circ$ while being strictly left to $k_\circ$ from the candidates which either

- have already been accepted (and belong to $\widehat{\Theta} \cup \{0\}$), or
- are still to be accepted or discarded ($\mathcal{C}$) whose detection intervals do not overlap with that of $k_\circ$, i.e. $\mathcal{I}(k_L) \cap \mathcal{I}(k_\circ) = \emptyset$ or equivalently $|k_\circ - k_L| \geq G_R(k_L) + G_L(k_\circ)$.

Identify k_R strictly to the right of $k_\circ$ from $\widehat{\Theta} \cup \{n\} \cup \mathcal{C}$ with analogous restrictions. Then, any candidates without decision that fall within (k_L, k_R) are considered as candidates competing with $k_\circ$. We denote this set of change point candidates by $\mathcal{D}$, i.e. $\mathcal{D} = \mathcal{C} \cap (k_L, k_R)$.
Step 3 **Pruning Algorithm** (**PrunAlg**, **see Sect.** 3.1.2). Apply the inner algorithm for pruning, PrunAlg, with the arguments $(\mathcal{D}, \mathcal{C}, \widehat{\Theta}, k_L, k_R)$. As an output, we yield a subset $\widehat{\mathcal{A}} \subset \mathcal{D}$ (possibly empty) which contains candidates to be accepted in the next step.
Step 4 **Update the accepted** ($\widehat{\Theta}$) **and active** ($\mathcal{C}$) **candidate sets.** We accept all estimators from the output of PrunAlg, $\widehat{\mathcal{A}}$, but not all of $\mathcal{D} \setminus \widehat{\mathcal{A}}$ are discarded yet. This is because $\mathcal{D}$ may contain acceptable estimators of change points that are too close to

the boundaries k_L or k_R, for which we cannot guarantee their acceptance at the current iteration (see Theorem 1 and Definition 3). However, if k_L (resp. k_R) has already been accepted, we discard any candidates in $\mathcal{D} \setminus \widehat{\mathcal{A}}$ which lie to the left (right) of the leftmost (rightmost) candidate in $\widehat{\mathcal{A}}$. Similarly, unaccepted candidates in $\mathcal{D} \setminus \widehat{\mathcal{A}}$ that lie between any two elements of $\widehat{\mathcal{A}}$ are discarded. In addition, we remove $k_\circ$ identified in Step 1 from the future consideration regardless of whether it has been accepted by `PrunAlg` or not.
In summary, we denote by $\mathcal{R}$ the set of all the candidates for which a decision has been reached, either because it has been accepted or discarded according to the above consideration.
Then, we add $\widehat{\mathcal{A}}$ to $\widehat{\Theta}$ and remove all the candidates in $\mathcal{R}$ from $\mathcal{C}$.
Step 5 **Iteration.** Repeat Steps 1 to 4 until $\mathcal{C}$ is empty. The set $\widehat{\Theta}$ is the final set of estimators and the output of the algorithm.

A pseudo-code of the outer algorithm can be found in Algorithm 1 of Appendix H in the supplementary material.

`LocAlg` is guaranteed to terminate since at each iteration, Step 4 discards at least one candidate $k_\circ$ from the active candidate set. Under a mild condition on $\mathcal{K}$, we show that this yields consistent estimation by guaranteeing that at least one suitable estimators remain in $\mathcal{C}$ for all the undetected change points (see Assumption 5 and the discussion thereafter).

In Step 3 of `LocAlg`, the inner algorithm `PrunAlg` makes a decision between competing candidates using SC, which are evaluated at each $\mathcal{A} \subset \mathcal{D} = \mathcal{C} \cap (k_L, k_R)$ as

$$\mathrm{SC}\,(\mathcal{A}|\mathcal{C}, \widehat{\Theta}, k_L, k_R) = \frac{n}{2} \log \left\{ \frac{\mathrm{RSS}\,(\mathcal{A} \cup \widehat{\Theta} \cup (\mathcal{C} \setminus \mathcal{D}))}{n} \right\} + (|\mathcal{A}| + |\widehat{\Theta}| + |\mathcal{C} \setminus \mathcal{D}|) \cdot \xi_n.$$

By construction, it makes a decision which of the candidates in $\mathcal{D}$ to accept while treating all other currently surviving candidates outside of (k_L, k_R) as given. Therefore, at any iterations of `LocAlg`, all X_t, $1 \le t \le n$, enter in the computation of SC. In other words, `LocAlg` has the interpretation of performing an adaptively selected subset of the exhaustive search over the complete candidate set $\mathcal{K}$ in a localised manner, by utilising the information readily available about the detection intervals of change point candidates.

3.1.2 Inner algorithm: pruning (`PrunAlg`)

The inner pruning algorithm `PrunAlg` in Step 3 of the outer localisation algorithm `LocAlg` takes as its input $(\mathcal{D}, \mathcal{C}, \widehat{\Theta}, k_L, k_R)$ and looks for a subset $\widehat{\mathcal{A}} \subset \mathcal{D}$ to be added to the finally accepted candidates according to the following rules:

Let $\mathcal{F}$ denote the collection of all subsets $\mathcal{A} \subset \mathcal{D}$ for which it holds:

$$\begin{aligned} &\text{adding further change point candidates to } \mathcal{A} \\ &\text{monotonically increases the SC} \end{aligned} \tag{6}$$

and denote by $m^* = \min_{\mathcal{A} \in \mathcal{F}} |\mathcal{A}|$. Then, we select $\widehat{\mathcal{A}}$ as

$$\widehat{\mathcal{A}} = \arg\min \Big\{ \mathcal{A} \subset_R \mathcal{A}' \text{ with } \mathcal{A}' \in \mathcal{F} \text{ and } m^* \le |\mathcal{A}'| \le m^* + 2 : \ \mathrm{SC}\,(\mathcal{A}|\mathcal{C}, \widehat{\Theta}, k_L, k_R) \Big\} \tag{7}$$

where, by $\mathcal{A} \subset_R \mathcal{A}' = \{\tilde{k}_j,\, 1 \le j \le m : \tilde{k}_1 < \tilde{k}_2 < \ldots < \tilde{k}_m\}$, we indicate that $\mathcal{A}' \setminus \mathcal{A} \subset \{\tilde{k}_1, \tilde{k}_m\}$, i.e. $\mathcal{A}$ contains all *inner* elements of $\mathcal{A}'$ (if exist) while the first and the last elements of $\mathcal{A}'$ may or may not be included in $\mathcal{A}$. If there are multiple subsets yielding the minimum SC in (C2), we choose the one with the minimum cardinality. If there are ties in the cardinality as well, we arbitrarily select one.

Remark 3 By performing a top-down search, the condition (C1) typically prunes down the search space quickly: If removing $k \in \mathcal{A}$ from $\mathcal{A}$ leads to an increase in SC, no subset of $\mathcal{A} \setminus \{k\}$ can be an element of $\mathcal{F}$. Efficient application of the pruning rules (C1)–(C2), including the computation of $\mathcal{F}$, involves careful implementation of this search process. For a complete algorithmic description of `PrunAlg`, see Algorithm 2 in Appendix H, and also Meier et al. (2021b) for computational details.

Remark 4 It is possible to apply the search criteria (C1)–(C2) to $\mathcal{K}$ directly, without going through the outer algorithm. In such a case, (C2) is simplified to

$$\widehat{\mathcal{A}} = \arg\min\{\mathcal{A} \in \mathcal{F} \text{ with } |\mathcal{A}| = m^* : \ \mathrm{SC}\,(\mathcal{A}|\mathcal{K}, \emptyset, 0, n)\}. \tag{C2$'$}$$

This approach still gains computationally compared to minimising the SC among all the $2^{|\mathcal{K}|}$ subsets of $\mathcal{K}$ while, as shown in Corollary 1, achieves consistency in multiple change point estimation. However, it is still to be avoided when there are many candidates to be pruned down, and `LocAlg` greatly reduces the computational cost by breaking down the scope of `PrunAlg` at each iteration.

Remark 5 We highlight the key differences between the use of SC in the proposed localised pruning, and the conventional use of information criteria as a model selection tool in the change point literature. Once candidate change points are generated, a commonly adopted pruning strategy is to evaluate and minimise an information criterion along a sequence of nested candidate models with an increasing number of change points. Such an approach requires the ordering of the candidate estimators according to their importance, as well as the maximum allowable number of change points, say $q_{\max}$, as an input parameter. This ordering plays an important role in establishing the consistency of such an approach because a spurious estimator added to the nested model sequence at an early stage cannot be removed. On the other hand, the sorting function h adopted in `LocAlg` does not play any role in the theoretical result presented in the next section (see Theorem 1).

Also, the selection of $q_{\max}$ is not straightforward especially when n is large, without pre-supposing the frequency or the sparsity of the change points, and some approaches require $q_{\max}$ to be finite in their theoretical consideration (Fryzlewicz 2014; Baranowski et al. 2019). In contrast, `PrunAlg` does not need this quantity to be explicitly set in its application, and the theoretical results require only that the candidate set $\mathcal{K}$ is not too large. This is a natural requirement in view of

Proposition 2, which implies an upper bound on the number of change points that any change point detection procedure can handle. In simulation studies, we observe empirical evidence of the sub-optimality of sequential evaluation and minimisation of an information criterion, particularly when there are frequent changes in the signal (see e.g. Table 4 in the supplementary material), which further supports the search criteria (C1)–(C2) adopted by `PrunAlg`.

3.2 Consistency of the localised pruning algorithm

In this section, we show that the localised pruning algorithm combining `LocAlg` and `PrunAlg` consistently estimates the total number of change points when applied to a suitable set of candidates. Furthermore, it 'almost' inherits the rate of convergence of the change point estimators from the candidate generating mechanisms and thus achieves consistency in change point localisation under mild conditions on the set of candidates.

We make the following assumption on candidate generation.

Assumption 4 (*Candidate generating algorithm*) Let $\mathcal{K} = \mathcal{K}_n$ denote the set of candidates obtained from $\{X_t\}_{t=1}^n$ and $Q_n = |\mathcal{K}|$ the total number of candidates. Then, with $\omega_n^{(1)}$, $\omega_n^{(2)}$ and ω_n as in Assumption 1:

(a) With probability approaching one, each change point has at least one candidate in its $(d_j^{-2}\rho_n)$-environment, i.e. as $n \to \infty$,

$$\mathsf{P}(\mathcal{M}_n^{(2)}) \to 1 \quad \text{where} \quad \mathcal{M}_n^{(2)} = \left\{ \max_{1 \le j \le q_n} \min_{k \in \mathcal{K}} d_j^2 \ |k - \theta_j| \le \rho_n \right\}$$

for a sequence ρ_n with $\max(\omega_n^{(1)}, \omega_n^{(2)})^2 = O(\rho_n)$ and $\rho_n = O(\omega_n^2)$.

(b) The total number of candidates Q_n fulfils $n^{-1}\omega_n^2 Q_n \to 0$.

The sequence ρ_n is the precision associated with the candidate generating method. We show that the proposed pruning algorithm almost inherits this rate in the sense made more precise in Theorem 1. We conjecture that typically, $\omega_n^{(1)} \asymp \omega_n^{(2)}$ as in all of the examples in Proposition 1. We further conjecture that, if so, $(\omega_n^{(1)})^2$ (or a related term) gives a lower bound for the minimax optimal localisation rate: This agrees with our observations in Propositions 1 and 3 under sub-Gaussian errors and when there are a finite number of change points, and thus indicates that the lower bound $\max(\omega_n^{(1)}, \omega_n^{(2)})^2$ on ρ_n is a reasonable one. The requirement $\rho_n = O(\omega_n^2)$ is a weak one with ω_n always dominating $\omega_n^{(1)}$ and $\omega_n^{(2)}$ (see Remark 1 (a)). If the precision attained by a particular candidate generating procedure is worse than ω_n^2, the localised pruning can still achieve consistency but with a stronger penalty ξ_n fulfilling $\rho_n/\xi_n \to 0$; see Equation (C.1) in the supplementary document and the discussion underneath.

Assumption 4 (b) on the number of candidates replaces a more stringent condition requiring q_n to be fixed, which is found in the literature adopting the information

criterion for determining the number of change points (Yao 1988; Kühn 2001). In particular, this rules out applying the localised pruning algorithm with every possible point as candidate estimators, i.e. $\mathcal{K} = \{1, \dots, n-1\}$. However, a reasonably good candidate generating method ought not to return too many candidates while meeting Assumption 4 (a), and we show that the MOSUM- and CUSUM-based candidate generating methods fulfils this requirement; see Proposition 5 and Proposition E.1 (b) of the supplementary material, respectively.

The following definitions that categorise the candidate estimators in $\mathcal{K}$ are frequently used throughout the paper.

Definition 2

(a) A candidate $k^* \in \mathcal{K}$ that yields $d_j^2|k^* - \theta_j| \leq \rho_n$ with ρ_n as in Assumption 4 (a) is referred to as a *strictly valid* estimator for θ_j, and the set of such candidates is denoted by $\mathcal{V}_j^*$ for each $j = 1, \dots, q_n$.
(b) For $\nu_n \to \infty$ at an arbitrarily slow rate, a candidate $k' \in \mathcal{K}$ with $d_j^2|k' - \theta_j| \leq \rho_n \nu_n$ is referred to as an *acceptable* estimator for θ_j, and the set of such candidates is denoted by $\mathcal{V}_j'$.
(c) The remaining candidates $k \in \mathcal{K} \setminus \mathcal{V}_j'$ are *unacceptable* for θ_j.

The gap between the best localisation rate ρ_n of the candidate generating procedure and what is acceptable for the localised pruning algorithm is unavoidable: For two very close candidates, the SC evaluated with the one slightly further away from a change point than the other can end up being smaller simply by chance.

We now show that `PrunAlg` described in Sect. 3.1.2, as a generic pruning algorithm, achieves consistent estimation of the number of change points as well as returning acceptable estimators for all θ_j, $j = 1, \dots, q_n$. Although the boundary points (k_L, k_R) supplied as input arguments to `PrunAlg` are always chosen among the change point candidates (including 0 and n) in Step 2 of `LocAlg`, our theory below is applicable to any $(s, e]$ with $0 \leq s < e \leq n$ as the interval of consideration and $\mathcal{D} = \mathcal{K} \cap (s, e)$ as the set of local candidates to be pruned down. In this context, it is understood that $\widehat{\Theta}$ contains candidates lying outside (s, e) only.

It may be the case that some change points are too close to either s or e and thus may or may not be detectable by `PrunAlg` within $(s, e]$, which necessitates the pruning criterion (C2) instead of the simpler (). We define the following sets of local change points with universal constants $0 < c^* < C^* < \infty$ defined in Proposition C.1 of the supplementary document:

$$\Theta^{(s,e)} = \left\{ \theta_j : d_j^2 \min(\theta_j - s, e - \theta_j) \geq C^* \xi_n \right\}, \tag{6}$$

$$\bar{\Theta}^{(s,e)} = \left\{ \theta_j : d_j^2 \min(\theta_j - s, e - \theta_j) \geq c^* \xi_n \right\}. \tag{7}$$

Theorem 1 establishes the connection between the output of `PrunAlg` and the sets defined in (6)–(7).

Theorem 1 *Let Assumptions 1, 2, 3 and 4 hold and denote by* $\widehat{\Theta}^{(s,e)}$ *the output of* `PrunAlg` from applying the criteria (C1)–(C2) to the local candidates $\mathcal{D} = \mathcal{K} \cap (s, e]$ *within an interval* $(s, e]$, and by $\mathcal{P}_n^{(s,e)}$ *the following event: The output set* $\widehat{\Theta}^{(s,e)}$ *contains*

(a) *exactly one acceptable candidate for each* $\theta_j \in \Theta^{(s,e)}$, *i.e.* $|\widehat{\Theta}^{(s,e)} \cap \mathcal{V}_j'| = 1$ *for* $\theta_j \in \Theta^{(s,e)}$,
(b) *at most one acceptable candidate for each* $\theta_j \in \bar{\Theta}^{(s,e)} \setminus \Theta^{(s,e)}$, *i.e.* $|\widehat{\Theta}^{(s,e)} \cap \mathcal{V}_j'| \le 1$ *for* $\theta_j \in \bar{\Theta}^{(s,e)} \setminus \Theta^{(s,e)}$, *and*
(c) *no other candidates, i.e.* $\widehat{\Theta}^{(s,e)} \setminus \bigcup_{j:\, \theta_j \in \bar{\Theta}^{(s,e)}} \mathcal{V}_j' = \emptyset$.

Then, with $\mathcal{M}_n := \mathcal{M}_n^{(11)} \cap \mathcal{M}_n^{(12)} \cap \mathcal{M}_n^{(13)} \cap \mathcal{M}_n^{(2)}$, *we have*

$$\mathsf{P}\left(\bigcap_{0 \le s < e \le n} \mathcal{P}_n^{(s,e)},\ \mathcal{M}_n \right) \to 1 \quad \text{as} \quad n \to \infty.$$

In view of Theorem 1, we categorise the change points according to their detectability within a given interval in the following definition.

Definition 3 For any $0 \le s < e \le n$, we refer to

(a) any change points in $\Theta^{(s,e)}$ as *surely detectable* within $(s, e]$,
(b) any change points in $\bar{\Theta}^{(s,e)}$ as *detectable* within $(s, e]$, and
(c) any change points in $\{\Theta \cap (s, e)\} \setminus \bar{\Theta}^{(s,e)}$ as *undetectable* within $(s, e]$.

The following corollary establishes that `PrunAlg`, when applied to the complete candidate set $\mathcal{K}$ directly, achieves consistency in multiple change point estimation.

Corollary 1 *Under the assumptions of Theorem 1, applying the search criteria (C1) and (C2) to the candidate set* $\mathcal{K}$ within $(0, n]$ yields $\widehat{\Theta}^{(0,n)} = \{\widehat{\theta}_j,\ 1 \le j \le \widehat{q}_n : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}_n}\}$ *which consistently estimates* Θ, *i.e.*,

$$\mathsf{P}\left\{ \widehat{q}_n = q_n;\ \max_{1 \le j \le q_n} d_j^2 |\widehat{\theta}_j - \theta_j| \le \rho_n v_n \right\} \ge \mathsf{P}(\mathcal{M}_n) + o(1) \to 1.$$

As pointed out in Remark 4, pruning down $\mathcal{K}$ according to (C1) and (C2) is computationally more efficient than the exhaustive minimisation of SC over all subsets of $\mathcal{K}$. Nevertheless, the localisation from the outer algorithm `LocAlg` results in a considerable computational advantage when a large set of candidates needs to be pruned down.

Next, we establish that the consistency achieved by `PrunAlg` within local search environments (as in Theorem 1), is carried over to the entire data set via the outer localisation algorithm `LocAlg`.

Assumption 5 Recall that the detection interval of each $k \in \mathcal{K}$ is denoted by $\mathcal{I}(k) = (k - G_L(k), k + G_R(k)]$. Then, for each $j = 1, \ldots, q_n$, there exists at least one acceptable candidate $\check{k}_j \in \mathcal{V}_j'$ which is situated well within its own detection interval by satisfying

$$\frac{\xi_n}{d_j^2 \min\{G_L(\check{k}_j), G_R(\check{k}_j)\}} \to 0. \tag{8}$$

Assumption 5 justifies the removal of $k_\circ$ identified in Step 1 of each iteration from the future consideration, regardless of whether it is accepted by `PrunAlg` or not: If $k_\circ$ is an acceptable estimator for some θ_j while meeting (8), such θ_j is surely detectable within $(k_L, k_R]$ and either $k_\circ$ or some $k \in \mathcal{V}_j'$ is accepted by `PrunAlg` at the current iteration; if not, there still remain at least one acceptable estimators in the active candidate set $\mathcal{C}$ for any undetected change points after removing $k_\circ$. We discuss how Assumption 5 is met by the MOSUM-based candidate generating procedure in Remark 7, and provide a similar discussion for the CUSUM-based procedure in Appendix E of the supplementary material.

Theorem 2 proves that `PrunAlg` combined with the outer algorithm `LocAlg` achieves consistency in multiple change point estimation.

Theorem 2 *Under the assumptions of Theorem 1* and *Assumption 5*, *the localised pruning algorithm* `LocAlg` *outputs* $\widehat{\Theta} = \{\widehat{\theta}_j, 1 \le j \le \widehat{q}_n : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}_n}\}$ *which consistently estimates* Θ, i.e.

$$\mathsf{P}\left\{\widehat{q}_n = q_n;\ \max_{1 \le j \le q_n} d_j^2 |\widehat{\theta}_j - \theta_j| \le \rho_n \nu_n\right\} \ge \mathsf{P}(\mathcal{M}_n) + o(1) \to 1,$$

for some $\nu_n \to \infty$ *at an arbitrarily slow rate*.

Its proof follows from the following two observations:

- When a change point is surely detectable for the first time at some iteration (in the sense of Definition 3 (a)), it gets detected by an acceptable estimator by Theorem 1 and consequently is no longer detectable in the subsequent iterations thanks to how the local environments are defined in Step 2 of `LocAlg`.
- On the other hand, those change points which are yet to be detected have corresponding acceptable estimators in the pool of candidates $\mathcal{C}$ due to how $\mathcal{C}$ is reduced in Step 4 of `LocAlg`.

4 Candidate generation

In this section, we investigate a two-stage procedure combining the localised pruning methodology with a multiscale extension of the MOSUM procedure of Eichinger and Kirch (2018). In Appendix E of the supplementary material, we provide the corresponding results for a CUSUM-based procedure motivated by the WBS

(Fryzlewicz 2014). Our theoretical analysis indicates that both the detection lower bound and the localisation rate achieved with the MOSUM-based candidate generating procedure are always better than those achievable with the CUSUM-based one.

4.1 MOSUM procedure and its multiscale extension

Eichinger and Kirch (2018) analyse the properties of a single-scale MOSUM procedure which, for a bandwidth $G = G_n$, estimates the locations of the change points by the locations of *significant* local maxima of the MOSUM statistic (2) according to two different criteria. For the purpose of generating candidates for the localised pruning, we adopt the method termed η-criterion with a lower false negative rate (see Section 2.2 of Meier et al. (2021b)). Let $\mathcal{K}(G) = \{k_{G,j},\ 1 \le j \le \widehat{q}_G\}$ denote the set of candidate estimators obtained with a bandwidth G. By the η-criterion, each $k_{G,j}$ is the local maximiser of the MOSUM detector (2) within its $\lfloor \eta G \rfloor$-radius for some $\eta \in (0,1)$, and $|T_{k_{G,j},n}(G;X)| > \pi_{n,G}$ with a threshold fulfilling $\pi_{n,G} = O(\sqrt{\log(n)})$. As we can show that $(\log(n))^{-1/2}|T_{\theta_j,n}(G(j);X)| \to \infty$ with a suitable bandwidth $G(j)$ for all $j = 1, \dots, q_n$, we can make sure that suitable estimators are added to the candidate set for all θ_j via this approach (see Proposition 4).

One way of selecting the threshold is to use the asymptotic distribution of $\max_{G \le k \le n-G} \tau^{-1}|T_{k,n}(G;\varepsilon)|$ (with τ^2 denoting the (long-run) variance of the error sequence $\{\varepsilon_t\}_{t=1}^n$), which can be derived under mild assumptions (see Theorem 2.1 of Eichinger and Kirch (2018)). Then, we can set $\pi_{n,G} = \widehat{\tau}_n D_n(G;\alpha)$, where $\widehat{\tau}_n^2$ is an estimator of τ^2 and $D_n(G;\alpha)$ a critical value chosen from this distribution with the significance level $\alpha \in (0,1)$; accordingly, we denote the corresponding candidate set by $\mathcal{K}(G,\alpha)$. This threshold fulfils the $O_P(\sqrt{\log(n)})$ bound for any fixed α provided that $\widehat{\tau}_n^2$ is bounded.

When a single-scale MOSUM procedure is adopted for estimating both the number and the locations of the change points, $\widehat{\tau}_n^2$ and α need to be chosen with care. Specifically, it requires that $|\widehat{\tau}_n^2 - \tau^2| = o_P(\log^{-1}(n))$ and also that α is sufficiently small (for the asymptotic analysis, we need $\alpha = \alpha_n \to 0$) in order not to incur any false positives, at the cost of possibly incurring false negatives.

On the other hand, when the MOSUM procedure is adopted solely for candidate generation, followed by the localised pruning procedure, accurate estimation of τ^2 is of less importance. This is particularly beneficial as consistent estimation of τ^2 in the presence of multiple mean shifts is a difficult task; see e.g. Chan (2020) for a robust estimator of τ^2. More complicated arguments, as given in Section 2.3 of Eichinger and Kirch (2018), are needed when a scale-dependent, local estimator $\widehat{\tau}_{t,G}^2$ is adopted in place of τ^2, but this estimator needs not be uniformly consistent (in $G \le t \le n-G$); we adopt this local estimator in our simulation studies for the threshold selection (see Appendix G for details). Similarly, we can select α generously or even do without thresholding. In practice, it is recommended to apply a mild threshold since setting $\pi_{n,G} = 0$ adds to computational burden and possibly leads to a loss of estimation accuracy. Based on our numerical experiments, we recommend $\alpha = 0.2$ as a generous enough choice balancing between the two requirements on the candidate set in Assumption 4.

For simplicity, in all our theoretical analysis below, we assume that $\pi_{n,G} = \tau D_n(G;\alpha)$ is used with τ known. The following proposition extends Theorem 3.2 of Eichinger and Kirch (2018).

Proposition 4 *Let $\eta \in (0,1)$ for the η-criterion and suppose*:

(a) For each $j = 1, \dots, q_n$, *there exists $G(j)$ such that $2G(j) \le \delta_j$ and $d_j^2 G(j) \ge c_M D_n$ for some constant $c_M > 0$ that does not depend on j.*
(b) $\mathsf{P}(\mathcal{M}_n^{(11)}) \to 1$ *with* $D_n^{-1}\omega_n^2 \to 0$, *where $\mathcal{M}_n^{(11)}$ is as in Assumption 1* (a).
(c) $\mathsf{P}(\mathcal{M}_n^{(12)} \cap \mathcal{M}_n^{(12+)} \cap \mathcal{M}_n^{(12-)}) \to 1$ *with $\mathcal{M}_n^{(12)}$ from Assumption 1* (b), and $\mathcal{M}_n^{(12\pm)}$ *defined analogously as*

$$\mathcal{M}_n^{(12\pm)} = \left\{ \max_{1\le j\le q_n} \max_{d_j^{-2}a_n \le \ell \le \theta_j - \theta_{j-1}} \frac{\sqrt{d_j^{-2}a_n}}{\ell} \left| \sum_{t=\theta_j - \ell \pm G(j)+1}^{\theta_j \pm G(j)} \varepsilon_t \right| \le \omega_n^{(1)} \right\}$$
$$\bigcap \left\{ \max_{1\le j\le q_n} \max_{d_j^{-2}a_n \le \ell \le \theta_{j+1} - \theta_j} \frac{\sqrt{d_j^{-2}a_n}}{\ell} \left| \sum_{t=\theta_j \pm G(j)+1}^{\theta_j \pm G(j)+\ell} \varepsilon_t \right| \le \omega_n^{(1)} \right\}.$$

Then, for a set $\mathcal{S}_n$ (specified in Lemma D.1 of the supplementary document) fulfilling $\mathsf{P}(\mathcal{S}_n) \to 1$, there exists a universal constant $C_M > 0$ (not depending on the signal or the distribution of $\{\varepsilon_t\}_{t=1}^n$) such that

$$\mathsf{P}\left(\max_{1\le j\le q_n} \min_{k\in\mathcal{K}(G(j),\alpha)} d_j^2 |k - \theta_j| \ge C_M (\omega_n^{(1)})^2,\ \mathcal{S}_n \right) \to 0.$$

Remark 6

(a) Condition 4 of Proposition 4 requires that for each change point θ_j, there exists a bandwidth $G(j)$ suitable for its detection.
(b) Condition 4 is also assumed for the consistency of the localised pruning method. Proposition 4 continues to hold under the following weaker condition:

$$\max_{1\le j\le q_n} \frac{1}{|d_j|\sqrt{G(j)}} \max_{|\ell - \theta_j| \le \frac{3}{2}G(j)} \left| \frac{1}{\sqrt{G(j)}} \sum_{t=\lfloor \ell - G(j)/2+1 \rfloor}^{\lfloor \ell + G(j)/2 \rfloor} \varepsilon_t \right| = o_P(1).$$

This assertion follows e.g. when an invariance principle holds as in Proposition 1 (c.i), and there are a finite mixture of homogeneous change points with an appropriate bandwidth for each of the homogeneous subsets (see Definition 1 (a)), in addition to

$$\frac{\lambda_n^2}{\min_{1\le j\le q_n} d_j^2 G(j)^2} = o(1) \quad \text{and} \quad \frac{\log(n)}{\min_{1\le j\le q_n} d_j^2 G(j)} = o(1).$$

(c) The assumptions on $\mathcal{M}_n^{(12\pm)}$ in Condition 4 do not impose additional constraints in the following cases:

- When $\{\varepsilon_t\}_{t=1}^n$ are independent and identically distributed.
- When $\{\varepsilon_t\}_{t=1}^n$ are stationary time series errors and there are a finite mixture of homogeneous change points.

In Corollary D.1 of the supplementary document, we show that the single-scale MOSUM procedure yields consistent estimators with optimal localisation rate, either under sub-Gaussianity or when there are finitely many change points, but only under the assumption that the change points are *homogeneous* as defined in Definition 1 (a). On the other hand, when the change points are heterogeneous, it cannot produce consistent estimators by construction.

As noted in Remark 6 (a), a natural solution to this lack of adaptivity is to apply the MOSUM procedure with a range of bandwidths. At the same time, scanning the same data at multiple scales introduces duplicate estimators and false positives, necessitating the use of a pruning method. Messer et al. (2014) and Messer et al. (2018) propose to prune down the estimators from a multiscale MOSUM procedure in a bottom-up manner, and a similar approach is also taken by Chan and Chen (2017): Accepting all the estimators from the smallest bandwidth, it proceeds to coarser scales and only accepts a change point estimator if its detection interval does not contain any estimators that are already accepted. While the bottom-up approach is applicable with multiple symmetric bandwidths, there is no canonical ordering when asymmetric bandwidths are used. More importantly, this approach rules out the possibility of removing any spurious estimators including those detected from the finest bandwidth and thus requires the finest bandwidth to be large relative to n in order to avoid spurious change point estimators. In Sect. 5.1, we observe on the simulated datasets that indeed, the bottom-up merging tends to incur a large number of false positives.

4.2 Localised pruning with MOSUM-based candidate generation

The localised pruning algorithm proposed in Sect. 3.1 is well-suited for pruning down the candidates generated by the multiscale MOSUM procedure. Let $\mathcal{G}$ denote a set of bandwidths. Each estimator $k \in \mathcal{K}(G, \alpha)$ for $G \in \mathcal{G}$ is associated with the natural detection interval $\mathcal{I}_N(k) = (k - G, k + G]$. Asymmetric bandwidths $\mathbf{G} = (G_\ell, G_r)$ with $(G_\ell, G_r) \in \mathcal{H} \subset \mathcal{G} \times \mathcal{G}$ are readily incorporated into the methodology using the MOSUM statistics defined as an appropriately scaled difference between $\bar{X}_{(b-G_\ell+1):b}$ and $\bar{X}_{(b+1):(b+G_r)}$ for $b = G_\ell, \ldots, n - G_r$, and the corresponding $\mathcal{I}_N(k) = (k - G_\ell, k + G_r]$ for $k \in \mathcal{K}(\mathbf{G}, \alpha)$. Then, the collection of all the estimators from the multiscale MOSUM procedure, $\mathcal{K}(\mathcal{H}, \alpha) = \bigcup_{\mathbf{G} \in \mathcal{H}} \mathcal{K}(\mathbf{G}, \alpha)$, can serve as the set of candidates $\mathcal{K}$. For Step 1 of the outer localisation algorithm `LocAlg`, we can sort the candidate change points either according to the size of associated jumps (see (5)) or using the p-values derived from the asymptotic null distribution defined for each pair of bandwidths, although care should be taken in their interpretation across multiple scales.

Selection of bandwidths. We propose to generate the set of bandwidths $\mathcal{G}$ as follows. Selecting a single parameter G_0, which should be smaller than the minimal distance between adjacent change points, and setting $G_1 = G_0$, we iteratively yield G_m, $m \geq 2$, as a Fibonacci sequence, i.e. $G_m = G_{m-1} + G_{m-2}$. Equivalently, we set $G_m = F_m G_0$ where $F_m = F_{m-1} + F_{m-2}$ with $F_0 = F_1 = 1$ are the Fibonacci numbers. This is repeated until for some $H = H_n$, it holds that $G_H < \lfloor n/\log(n) \rfloor$ while $G_{H+1} \geq \lfloor n/\log(n) \rfloor$. When using asymmetric bandwidths, it is advisable to avoid the pairs of bandwidths which are too unbalanced, both in view of the asymptotic theory and the finite sample performance as is well-known from the two-sample testing literature; a similar requirement is also found in Chan and Chen (2017). For this reason, we only include the pairs of bandwidths $\mathbf{G} = (G_\ell, G_r)$ in $\mathcal{H}$ that satisfy

$$G_\ell, G_r \in \mathcal{G} = \{G_1, \ldots, G_H\} \quad \text{with} \quad \frac{\max(G_\ell, G_r)}{\min(G_\ell, G_r)} \leq C_{\text{asym}} \tag{9}$$

for some constant $C_{\text{asym}} > 0$.

With the thus-constructed set of asymmetric bandwidths $\mathcal{H}$, Assumption 4 (b) is met by $\mathcal{K}(\mathcal{H}, \alpha)$.

Proposition 5 *Suppose that $\omega_n^2/G_0 \to 0$ with ω_n as in Assumption 1 (a). Then, for $\mathcal{H}$ fulfilling (9), we have $n^{-1}\omega_n^2 |\mathcal{K}(\mathcal{H}, \alpha)| \to 0$.*

The assumption $\omega_n^2/G_0 \to 0$ is made solely to obtain a crude deterministic upper bound on the number of possible candidates from the smallest bandwidth. We may replace it by a condition that directly limits the number of candidates detected at each bandwidth, or an assumption on q_n in combination with a stochastic version of Assumption 4. The finiteness of C_{asym} is also required for the bounding of $|\mathcal{K}(\mathcal{H}, \alpha)|$. While the use of asymmetric bandwidths does not improve the asymptotic rates over symmetric bandwidths, it does improve the small sample performance; we find that $C_{\text{asym}} = 4$ works well in practice and have used this choice in all our numerical experiments.

Remark 7

(a) For each $k \in \mathcal{K}(\mathcal{H}, \alpha)$, the natural detection interval $\mathcal{I}_N(k)$ can serve as its detection interval $\mathcal{I}(k) = (k - G_L, k + G_R]$, whereby the detection distances (G_L, G_R) are given by the set of bandwidths (G_ℓ, G_r) with which k has been detected. Then, we have Assumption 5 fulfilled by $\mathcal{K}(\mathcal{H}, \alpha)$ provided that there exists a single bandwidth $G(j) \in \mathcal{G}$ satisfying $d_j^2 G(j)/\xi_n \to \infty$ for each $j = 1, \ldots, q_n$, which is readily met under Condition 4 of Proposition 4 and Assumption 3.

(b) It may be the case that $\mathcal{K}(\mathcal{H}, \alpha)$ contains identical acceptable candidates k of θ_j returned at multiple scales, including some (G_ℓ, G_r) that does not satisfy $d_j^2 \min(G_\ell, G_r)/\xi_n \to \infty$. Against such a contingency, we propose to assign as $\mathcal{I}(k)$ the natural detection interval that returns the smallest p-value for the MOSUM test associated with the detection of k. Because the p-values decrease with the increase of jump size as well as that of bandwidths, this strategy will

recommend a reasonably large natural detection interval as $\mathcal{I}(k)$. In simulation studies, we use an implementation of the algorithm which simply supposes that Assumption 5 is satisfied by the candidate generating mechanism.

The consistency of the localised pruning algorithm in combination with the MOSUM-based candidate generating mechanism follows immediately from Propositions 4, 5 and Theorem 2.

Theorem 3 *Let Assumptions 1, 2, 3 and 5 hold, and suppose that the conditions in Propositions 4 and 5 are satisfied. Then, the localised pruning algorithm* `LocAlg` *applied to* $\mathcal{K}(\mathcal{H}, \alpha)$*, yields* $\widehat{\Theta} = \{\widehat{\theta}_j,\ 1 \le j \le \widehat{q}_n : \widehat{\theta}_1 < \ldots < \widehat{\theta}_{\widehat{q}_n}\}$ *which consistently estimates* Θ*, i.e.*

$$\mathsf{P}\left\{\widehat{q}_n = q_n;\ \max_{1\le j\le q_n} d_j^2|\widehat{\theta}_j - \theta_j| \le \nu_n(\omega_n^{(1)})^2\right\} \to 1$$

for any $\nu_n \to \infty$ *arbitrarily slowly.*

The next corollary provides the consistency of $\widehat{\Theta}$ in specific settings, which follows directly from Proposition 1 and Theorem 3.

Corollary 2 *Let Assumptions 2, 3, 5 and Condition 4 of Proposition 4 hold and* $\omega_n^2/G_0 \to 0$*, with* ω_n *specified below.*

(a) ***Sub-Gaussianity.*** *Let* $\{\varepsilon_t\}_{t=1}^n$ *meet the conditions of Proposition 1 (a). Then, with* $\omega_n \asymp \sqrt{\log(n)}$*, we have*

$$\mathsf{P}\left\{\widehat{q}_n = q_n;\ \max_{1\le j\le q_n} d_j^2|\widehat{\theta}_j - \theta_j| \le \nu_n \log(q_n)\right\} \to 1.$$

(b) ***Heavy tails.*** Let $\{\varepsilon_t\}_{t=1}^n$ *meet the conditions of Proposition 1 (b). Then, with* $\omega_n \asymp n^{1/\beta}$ *for any* $\beta < \alpha$*, we have*

$$\mathsf{P}\left\{\widehat{q}_n = q_n;\ \max_{1\le j\le q_n} d_j^2|\widehat{\theta}_j - \theta_j| \le \nu_n q_n^{2/\beta}\right\} \to 1.$$

(c) ***Invariance principle and moment conditions.*** Let $\{\varepsilon_t\}_{t=1}^n$ *meet the conditions of Propositions 1* (c) and 4 (c) with $\omega_n^{(1)} \asymp \nu_n q_n^{1/\gamma}$*. Then, with* $\omega_n \asymp \max(\lambda_n \nu_n, \sqrt{\log(n)})$*, we have*

$$\mathsf{P}\left\{\widehat{q}_n = q_n;\ \max_{1\le j\le q_n} d_j^2|\widehat{\theta}_j - \theta_j| \le \nu_n q_n^{2/\gamma}\right\} \to 1.$$

In light of Propositions 2 and 3, Corollary 2 shows that under sub-Gaussianity, the localisation pruning applied with the MOSUM-based candidate generating procedure yields minimax optimal rates both in terms of the detection lower bound in

the sublinear change point regime, and the localisation rate. Also, even when $\{\varepsilon_t\}_{t=1}^n$ is heavy-tailed, if the number of change points q_n is finite, the combined methodology achieves the minimax optimal localisation rate.

5 Numerical results

5.1 Simulation results

We conducted an extensive simulation study comparing the performance of the proposed localised pruning algorithm combined with the MOSUM- and CUSUM-based candidate generation, respectively, discussed in Sect. 4 and Appendix E of the supplementary material, against that of a variety of competitors whose implementations are readily available in R. For a complete description of the simulation results, see Appendix G.

We consider the five test signals from Fryzlewicz (2014) and their extensions ($n \geq 2 \times 10^4$) with both frequent and sparse change points, in order to assess the scalability of different methods. As error sequences, we consider i.i.d. random variables following Gaussian and t_5 distributions, and AR(1) processes with both weak and strong autocorrelations. The localised pruning procedure requires the selection of ξ_n for the penalty of SC, for which we consider $\xi_n \in \{\log^{1.01}(n), \log^{1.1}(n)\}$ in the case of independent Gaussian errors, $\xi_n \in \{\log^{1.1}(n), n^{2/4.99}\}$ for independent t_5-distributed errors and $\xi_n \in \{\log^{1.1}(n), \log^2(n)\}$ for AR(1) processes as the 'light' and 'heavy' penalties, in view of Assumption 3. We consider two candidate sorting function for Step 1 of `LocAlg` including (5); as indicated by the theoretical results (Theorem 1), its choice has little influence on the numerical results. We discuss in details the selection of tuning parameters for individual candidate generation methods in Appendix G.1.2. Based on the numerical results, we provide default parameter choices in the implementation of the localised pruning algorithm with the MOSUM-based candidate generation in the R package `mosum` (Meier et al. 2021a).

Overall, the proposed localised pruning performs well according to a variety of criteria, often performing as well as or even better than many competitors both in terms of the total number of estimated change points and their locations. At the same time, the localised pruning is shown to be scalable to long signals with $n \geq 2 \times 10^4$. Most competing methods are specifically tailored for i.i.d. Gaussian errors and thus struggle with heavy tails or serial correlations. In the i.i.d. Gaussian settings, our proposed method is robust to the choice between the light and the heavy penalties. When the errors are heavy-tailed, the heavy penalty chosen in line with Assumption 3 is successful in not causing false alarms, while the light penalty leads to good power at the price of slightly increased false positive rate (which still is much lower than that obtained by other competitors). Similar observations are made under serial dependence. From the above, our conclusion is that the simulation results well-support the theoretical findings relating the behaviour of $\{\varepsilon_t\}_{t=1}^n$ to the choice of penalty, and that the former should be considered in selecting ξ_n, an observation that applies to all change point detection methodologies. Between the two different candidate generating methods, the MOSUM-based method produces estimators of

better localisation accuracy while the CUSUM-based one tends to incur more false positives.

5.2 Real data analysis: array CGH data

In this section, we illustrate the performance of the proposed methodology using array comparative genomic hybridisation (CGH) data that has previously been analysed in the literature. In Appendix A, an application to Kepler light curve data first analysed in Fisch et al. (2018) is provided as a second example.

Microarray-based comparative genomic hybridisation (array CGH) provides a means to quantitatively measure DNA copy number aberrations and to map them directly onto genomic sequences (Snijders et al. 2001). We analyse a dataset obtained from a breast tumour specimen (S0034) described in Snijders et al. (2001) ($n = 2227$). A number of algorithms have been proposed which, regarding any gains or losses in the copy number from the normalised copy number ratios between two DNA samples as change points, identify their total number and locations under the model (1) (see e.g. Olshen et al. (2004), Li et al. (2016) and Niu and Zhang (2012)).

Olshen et al. (2004) proposed to smooth the array CGH data for outlier removal prior to change point analysis. Noticing that such a step may introduce serial correlations, we choose to analyse the raw data and account for possible outliers by adopting the penalty $\xi_n = \log^{1.1}(n)$ (used as the heavy penalty in the case of independent Gaussian errors in the simulation studies, see Sect. 5.1 and also Appendix G.1) for the localised pruning algorithm, with $\alpha = 0.2$ and $\eta = 0.4$ for the MOSUM-based candidate generation (MoLP) and $C_\zeta = 0.5$ (see Appendix G.1.2 for the description of its role) for the CUSUM-based one (CuLP). In addition to the methods included in the comparative simulation study in Sect. 5.1, we consider the circular binary segmentation algorithm of Olshen et al. (2004) (CBS, implemented in Seshan and Olshen (2018)) and the modified screening and ranking algorithm of Xiao et al. (2014) (modSaRa, implemented in Xiao et al. (2016)). It is important to note that the CBS takes all boundary markers between neighbouring chromosomes as an input unlike any other procedures in consideration, and automatically reports all of them as change points.

Figure 1 plots the normalised fluorescence ratios from S0034 and the change point estimators returned by various methods, and Table 2 reports the number of estimated change points. Overall, MoLP and CuLP detect fewer number of change points compared to most of the competitors, and many elements of the two sets of estimators either coincide or lie very close to each other. Also, many change point estimators coincide with the boundary markers although they are detected *without* knowing their positions unlike the CBS.

The data exhibit heteroscedasticity particularly beyond the genome order 2274 where there is a dramatic increase in the variability. Both candidate generating methods return a large number of candidates (MoLP has 137 candidates, CuLP has 82) and our localised approach to pruning manages to reduce the size of the candidate sets reasonably well. On the other hand, WBS.sBIC, WBS2.SDLL, TGUH, PELT, S3IB and FDRSeg are susceptible to returning (possibly) spurious change

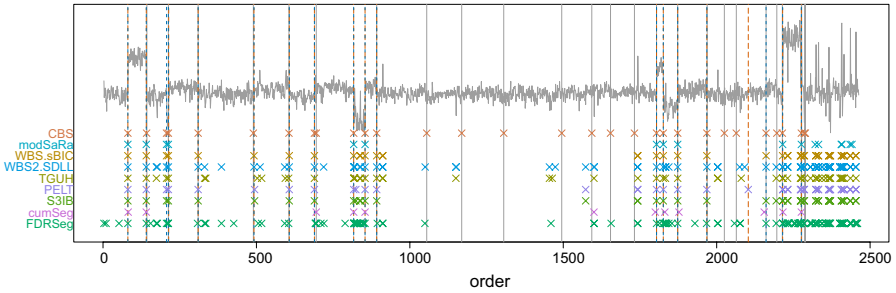

Fig. 1 Normalised copy number ratios of a comparison of DNA from cell strain S0034. Vertical solid lines indicate the boundaries between chromosomes, longdashed lines are change points estimated by MoLP and dashed lines are those estimated by CuLP. Change-point estimators from different methods are also plotted (×)

point estimators particularly in this region of increased volatility. cumSeg misses some of the change points commonly detected by many methods, which is consistent with the findings from the simulation studies reported in Appendix G.

Table 2 Number of change points estimated from the S0034 data set

MoLP	CuLP	CBS	mod-SaRa	WBS.sBIC	WBS2.SDLL	TGUH	PELT	S3IB	cumSeg	FDRSeg
17	18	31	17	46	82	58	46	49	12	126

We node that CuLP, WBS.sBIC, WBS2.SDLL and FDRSeg are affected by the randomness involved in generating either the candidate estimators or the critical values, and yield different results on different runs when applied to this data set. It may be due to that the underlying signal is not exactly piecewise constant, a phenomenon known as genomic waves (Diskin et al. 2008). The results for these methods reported here were obtained by setting the seed of R's random number generator to be one.

6 Conclusions and outlook

In this paper, we propose the localised pruning algorithm which, together with a class of multiscale candidate generating procedures, forms a two-stage methodology for data segmentation. Adopting a truly multiscale framework, we prove the consistency of the proposed methodology in multiple change point estimation under mild conditions, and show that it inherits the localisation property of the candidate generating mechanism. Theoretical properties for the second-stage localised pruning algorithm are discussed independently from the choice of first-stage candidate generating methods, allowing an easy extension of the results to other candidate

generating methods. Two examples for this choice are provided: A multiscale MOSUM procedure and a WBS algorithm. Combined with the former, the localised pruning algorithm achieves minimax rate optimality both in change point localisation and detection lower bound in those settings where such optimality results are available. Importantly, we work with meta-assumptions on the key elements of the change point structure and the error distribution, the latter of which only concern the bounds given in Assumption 1 and thus permit both heavy-tailedness and serial dependence. In doing so, the influence of each element on our theoretical arguments is made transparent and discussed in details, allowing for their ready extension to other error distributions in the future.

A comparison with competitors in terms of (a) theoretical properties such as the detection lower bound and the localisation rate, (b) computational complexity, speed and scalability to large sample sizes, and (c) the performance on a variety of simulations and real data examples shows that our proposed methodology performs universally well, especially when combined with the MOSUM-based candidate generating method, whose implementation is provided in the R package `mosum` available on CRAN (Meier et al. 2021a).

While we focus on the univariate mean change point detection problem in this paper, there are natural ways for extending the proposed methodology to more general change point problems: Via an appropriate transformation of the data, e.g. by adopting an *M*-estimation framework, change points in the stochastic properties of interest can be made detectable as change points in the mean of the transformed time series. With a suitably modified information criterion, our methodology becomes applicable to a variety of more complex change point scenarios, such as the detection of changes in regression parameters (e.g. neural network-based nonparametric (auto-)regression); other distributional parameters (e.g. for integer-valued time series) and robust change point detection (Kirch and Kamgaing 2015a, b; Kirch and Weber 2018). Some first results in this direction based on the current paper have already been obtained in Reckrühm (2019), where the necessity for a model selection strategy in such general change point problems is well-motivated (see Chapter 2.4 therein). Besides, our results can be adapted to detect parameter changes in renewal processes (Kühn 2001; Messer et al. 2014; Kirch and Klein 2021).

Another venue for extensions is high-dimensional data segmentation, where algorithms developed for multiple change point detection in the mean of univariate data have been adopted for that in the mean (Wang and Samworth 2018), covariance (Wang et al. 2020a) and model parameters (Safikhani and Shojaie 2020) in high dimensions. The development of a methodology for multiscale data segmentation is a separate problem from the aggregation of information on change points across components in high-dimensional settings. We refer to Cho and Kirch (2020) for further discussions on extending univariate mean change point detection procedures to general change point problems.

With such extensions—to high dimensions and more complex changes—in mind, attaining deeper understanding into the properties of multiscale data segmentation and proposing a methodology of improved performance is of particular interest even in the univariate mean change problem. The present work can be seen as an

important first step towards an extended methodology for more general data segmentation problems.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00811-5.

Acknowledgements Haeran Cho was supported by the EPSRC grant no. EP/N024435/1. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme 'Statistical scalability' (supported by EPSRC grant number EP/R014604/1) when work on this paper was undertaken.

References

Arias-Castro, E., Candes, E. J., Durand, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics, 39*, 278–304.

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

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

Boysen, L., Kempe, A., Liebscher, V., Munk, A., Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics, 37*, 157–183.

Chan, H. P., Chen, H. (2017). Multi-sequence segmentation via score and higher-criticism tests. arXiv preprint, arXiv:1706.07586.

Chan, H. P., Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica, 23*, 409–428.

Chan, K. W. (2020). Mean-structure and autocorrelation consistent covariance matrix estimation. *Journal of Business & Economic Statistics,* 1–15.

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

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

Cho, H., Kirch, C. (2020). Data segmentation algorithms: Univariate mean change and beyond. arXiv preprint arXiv:2012.12814.

Csörgö, M., Horváth, L. (1997). *Limit theorems in change-point analysis* (Vol. 18). New York: Wiley.

Davis, R. A., Yau, C. Y. (2013). Consistency of minimum description length model selection for piecewise stationary time series models. *Electronic Journal of Statistics, 7*, 381–411.

De Haan, L., Ferreira, A. (2007). *Extreme value theory*: *An introduction*. New York: Springer.

Dette, H., Schüler, T., Vetter, M. (2020). *Multiscale change point detection for dependent data*. *To appear in Scandinavian Journal of Statistics*

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research, 36*, e126–e126.

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

Fisch, A. T. M., Eckley, I. A., Fearnhead, P. (2018). *A linear time method for the detection of point and collective anomalies*. arXiv preprint arXiv:1806.01947.

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

Fromont, M., Lerasle, M., 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. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics,* 3390–3421.

Horváth, L., Rice, G. (2014). Extensions of some classical methods in change point analysis. *TEST, 23*, 1–37.

Killick, R., Fearnhead, P., 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*. Universität zu Köln. PhD thesis.

Kirch, C., Kamgaing, J. T. (2015a). Detection of change points in discrete valued time series. In *Handbook of discrete valued time series* (pp. 219–244).

Kirch, C., Kamgaing, J. T. (2015b). On the use of estimating functions in monitoring time series for change points. *Journal of Statistical Planning and Inference, 161*, 25–49.

Kirch, C., Klein, P. (2021). Moving sum data segmentation for stochastics processes based on invariance. *Statistica Sinica* (**to appear**).

Kirch, C., Weber, S. (2018). Modified sequential change point procedures based on estimating functions. *Electronic Journal of Statistics, 12*, 1579–1613.

Komlós, J., Major, P., Tusnády, G. (1975). An approximation of partial sums of independent RV's, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 32*, 111–131.

Komlós, J., Major, P., Tusnády, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 34*, 33–58.

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

Li, H., Munk, A., Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electronic Journal of Statistics, 10*, 918–959.

Li, H., Guo, Q., Munk, A. (2019). Multiscale change-point segmentation: Beyond step functions. *Electronic Journal of Statistics, 13*(2), 3254–3296.

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

Meier, A., Cho, H., Kirch, C. (2021a). mosum: Moving sum based procedures for changes in the mean. *R package version, 1*(2), 5.

Meier, A., Kirch, C., Cho, H. (2021b). mosum: A package for moving sums in change point analysis. *Journal of Statistical Software, 97*(8), 1–42.

Messer, M., Kirchner, M., Schiemann, J., Roeper, J., Neininger, R., Schneider, G. (2014). A multiple filter test for the detection of rate changes in renewal processes with varying variance. *The Annals of Applied Statistics, 8*, 2027–2067.

Messer, M., Albert, S., Schneider, G. (2018). The multiple filter test for change point detection in time series. *Metrika, 81*, 589–607.

Mikosch, T., Moser, M. (2013). The limit distribution of the maximum increment of a random walk with dependent regularly varying jump sizes. *Probability Theory and Related Fields, 156*, 249–272.

Mikosch, T., Račkauskas, A. (2010). The limit distribution of the maximum increment of a random walk with regularly varying jump size distribution. *Bernoulli, 16*, 1016–1038.

Niu, Y. S., Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *The Annals of Applied Statistics, 6*, 1306–1326.

Olshen, A. B., Venkatraman, E., Lucito, R., Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics, 5*, 557–572.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100–115.

Reckrühm, K. (2019). *Estimating multiple structural breaks in time series-a generalized MOSUM approach based on estimating functions*. Magdeburg, Germany: Otto von Guericke University. PhD thesis.

Safikhani, A., 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*

Schlüter, S., Fischer, M. J. (2009). *A tail quantile approximation formula for the student t and the symmetric generalized hyperbolic distribution*. FAU Discussion Papers in Economics 05/2009, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics.

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

Seshan, V. E., Olshen, A. (2018). *DNAcopy*: *DNA copy number data analysis*. R package version, 1(54).
Shao, Q.-M. (1995). On a conjecture of Révész. *Proceedings of the American Mathematical Society, 123*, 575–582.
Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics, 29*, 263.
Titsias, M. K., Holmes, C. C., Yau, C. (2016). Statistical inference in hidden Markov models using k-segment constraints. *Journal of the American Statistical Association, 111*, 200–215.
Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge: Cambridge University Press.
Wang, D., Yu, Y., Rinaldo, A. (2020a). *Optimal covariance change point localization in high dimension*. To appear in Bernoulli.
Wang, D., Yu, Y., Rinaldo, A. (2020b). Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics, 14*, 1917–1961.
Wang, T., Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society*: *Series B, 80*, 57–83.
Xiao, F., Min, X., Zhang, H. (2014). Modified screening and ranking algorithm for copy number variation detection. *Bioinformatics, 31*, 1341–1348.
Xiao, F., Niu, Y., Hao, N., Xu, Y., Jin, Z., Zhang, H. (2016). modSaRa: modSaRa: a computationally efficient R package for CNV identification. *R package version, 1*.
Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters, 6*, 181–189.
Yau, C. Y., Zhao, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society*: *Series B, 78*, 895–916.