On Inference and Causality in Change Point Regressions

Shakeel Gavioli-Akilagun

Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of Doctor of Philosophy

December 2023

Coming with a light heart
To pick some violets,
I found it difficult to leave,
And slept overnight
Here in this spring field.

Yamabe no Akahito

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of 54,976 words.

I confirm that Chapters 3 and 4 are co-authored with Professor Piotr Fryzlewicz and I contributed 80% to both works. Chapter 3 has been submitted to a peer-reviewed statistical journal, and we plan to submit Chapters 4 and 5 for publication soon.

Shakeel Gavioli-Akilagun
December 2023

Acknowledgements

I would like to thank my supervisor, Professor Piotr Fryzlewicz, for accepting me as one of hist students and for his constant support and encouragement over the years. I have learnt a lot from him, and it has truly been a privilege to work under his guidance. I am thankful the the LSE Department of Statistics for providing an ideal environment in which to work, and even more to my friends and colleagues within the department. Finally I owe an immeasurable debt of gratitude to my mother and father, Ganiat and Stefano, who sacrificed so much for my education and along the way taught me everything there is to know about life.

Abstract

Change point analysis, broadly defined, concerns the setting in which one observes time indexed data whose distribution is liable to change at a certain number of unknown locations in time. These locations are known as change points, and data indexed between change point locations can be understood to be in some sense homogeneous. This thesis studies two relatively neglected problems in change point analysis. Namely, statistical inference and causal structure discovery. For the first problem, we propose two methods for recovering disjoint intervals each contain a change point location uniformly at some significance which may be tuned by the user. We focus principally on the piecewise polynomial change point model, in which the data are modeled as weakly dependent noise fluctuating around a piecewise trend. For the second problem we consider a multivariate time series and model change points across the series as arrival times of a marked point process. We introduce a procedure for recovering a graph which encodes causal information about the process, in the sense that an edge in the graph can be (under some conditions) understood as indicating that change points in one time series cause change points in another time series.

Contents

List of Tables

List of Figures

List of Algorithms

1 Introduction

This thesis studies the problems of statistical inference and causal structure discovery in change point models. We focus principally on the piecewise polynomial change point model, in which the data are modeled as weakly dependent noise fluctuating around a piecewise trend. The piecewise polynomial model is popular in applied work, as piecewise polynomial functions allow to model temporal trends in the data. However, compared to the canonical problem in which the signal is piecewise constant, the piecewise polynomial model has received comparatively little attention on the theoretical front. Besides, the piecewise polynomial problem is not a straightforward extension of the canonical problem: whereas the latter problem is purely local in the sense that the error in estimating each change point location does not depend on the sample size, the former problem is global and the estimation error depends in a precise way on the sample size and the smoothness of the trend. In Chapter 2 we introduce the relevant literature on change point analysis, change point inference, and causal structure discovery. The main original contributions occur in Chapters 3, 4, and 5 and are summarized below. Finally, in Chapter 6 we give a brief summary of our contributions and discuss potential directions for future research.

Chapter 3. Fast and Optimal Inference for Change Points in Piecewise Polynomials via Differencing *We consider the problem of uncertainty quantification in change point regressions, where the signal can be piecewise polynomial of arbitrary but fixed degree. That is we seek disjoint intervals which, uniformly at a given confidence level, must each contain a change point location. We propose a procedure based on performing local tests at a number of scales and locations on a sparse grid, which adapts to the choice of grid in the sense that by choosing a sparser grid one explicitly pays a lower price for multiple testing. The* *procedure is fast as its computational complexity is always of the order $\mathcal{O}(n \log(n))$ where n is the length of the data, and optimal in the sense that under certain mild conditions every change point is detected with high probability and the widths of the intervals returned match the mini-max localisation rates for the associated change point problem up to log factors. A detailed simulation study shows our procedure is competitive against state of the art algorithms for similar problems.*

Chapter 4. Robust Inference for Change Points in Piecewise Polynomials using Confidence Sets *We revisit the problem of uncertainty quantification in the piecewise polynomial model, and consider the setting in which the contaminating noise may be arbitrarily distributed with, for example, atoms in its distribution, an arbitrary number of finite moments, and non-constant variance. We focus on the case where the polynomial degree of the underlying signal is either* 0 *or* 1 *on stationary segments. We present a procedure which, under minimal assumptions, returns localized regions of a data sequence which must contain a change point at some global significance level chosen by the user. The procedure works by performing local tests for the presence of a change point at a variety of scales and locations, and recursively retaining the narrowest region on which a change is detectable. The local tests in turn are based on properties of confidence sets for an underlying regression function obtained by inverting certain robust multi-resolution tests. We work implicitly with signs of the data sequence, and as a result we only require sign symmetry and sign independence of the contaminating noise for accurate inference. Despite this we attain the best possible rates in the minimax sense (up to logarithms) for change point detection and localization in both the piecewise constant and piecewise linear and continuous signal settings.*

Chapter 5. Recovering Dependence Structures in Change Point Regressions with a View to Causality *We propose a method for estimating graphs which encode causal information about change points occurring in a moderate number of data streams. That is, the presence of an edge in the graph signifies that change points in one series cause changes in another. Typically after performing change point analysis relationships between estimated change points can only be described qualitatively, since in general change point locations are held* *to be unknown but non-stochastic. From the perspective of practitioners this is a limitation, since in many settings it is reasonable to believe change points will be causally linked. We model unobserved change point locations as arrival times of a marked point process, for which (non) causality is well understood (Didelez, 2008), and propose a method for consistently estimating the underlying causal graph.*

2 Literature Review

2.1 Change point analysis

This section reviews the statistical literature on offline change point analysis, with a particular emphasis on the piecewise polynomial change point problem which is the focus of the thesis. We begin with a description of the generic change point problem, and introduce the piecewise polynomial problem as a special case. We then introduce the three main problems in change point analysis, namely testing, estimation, and inference, and survey various methods for solving the first two problems for the piecewise polynomial change point model. Methods for change point inference are surveyed in Section 2.2.

2.1.1 Problem statement and motivation

The generic change point problem

In the generic change point problem the analyst observes data $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, where for $t = 1, \ldots, n$ each Y_t has marginal distribution F_t. Given a transformation $\mathcal{H}(\cdot)$ and writing $\theta_t = \mathcal{H}(F_t)$ the interest lies in understanding whether the sequence $\theta_1, \ldots, \theta_n$ is constant or contains changes. An integer η is called a change point location if $\theta_\eta \neq \theta_{\eta+1}$. Throughout the thesis the number of change points in such a sequence will be denoted by N, the set of change point locations will be denoted by $\Theta = \{\eta_1, \ldots, \eta_N\}$, and we follow the convention that $\eta_0 = 0$ and $\eta_{N+1} = n$. Given a distance measure $d(\cdot, \cdot)$ the size of the change at the k-th change point location will be denoted by $\Delta_k = d(\theta_{\eta_k}, \theta_{\eta_k+1})$. It is worth noting that in applications N is typically unknown, and that in theoretical analyses N, Θ, and $\{\Delta_k \mid k = 1, \ldots, N\}$ may depend on the sample size n. Change point analysis is principally concerned with the following three problems.

1. **Testing:** deciding whether at least one change exists in the sequence $\theta_1, \ldots, \theta_n$.
2. **Estimation:** accurately estimating N, the number of changes, and Θ, their locations.
3. **Inference:** quantifying the uncertainty around estimates of N, the number of changes, around estimates of Θ, their locations, and around estimates of $\{\Delta_k \mid k = 1, \ldots, N\}$, the jump sizes.

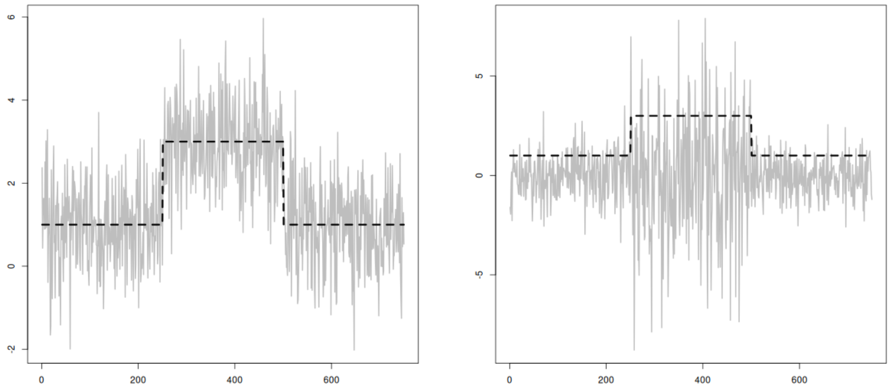
The above formulation is quite general and appears for example in Pilliat et al. (2023). Typically one has in mind a particular transformation, and some restrictions are placed on the sequence of distribution functions. To fix the idea, two examples are given below.

Example 2.1.1. *Consider the process $\{Y_t = \theta_t + \zeta_t \mid t = 1, \ldots, n\}$ for some $n \in \mathbb{N}$ where the ζ's are mutually independent with marginal $\mathcal{N}(0, 1)$ distribution and the sequence $\{\theta_t \mid t = 1, \ldots, n\}$ is piecewise constant. Then, the process has piecewise constant mean and the transformation of interest is: $\mathcal{H}_t(F) = \int xF(\mathrm{d}x)$. A sample path $\boldsymbol{Y} = (Y_t \mid t = 1, \ldots, 750)'$ with $\theta_t = 1 + 2 \times \mathbf{1}_{\{250<t\leq 500\}}$ is shown in Figure 2.1a.*

Example 2.1.2. *Consider the process $\{Y_t = \theta_t \times \zeta_t \mid t = 1, \ldots, n\}$ for some $n \in \mathbb{N}$ where the ζ's and θ's are as in Example 2.1.1. Then, the process has zero mean and piecewise constant variance, and the transformation of interest is: $\mathcal{H}_t(F) = \int x^2 F(\mathrm{d}x)$. A sample path $\boldsymbol{Y} = (Y_t \mid t = 1, \ldots, 750)'$ again with $\theta_t = 1+2\times\mathbf{1}_{\{250<t\leq 500\}}$ is shown in Figure 2.1b.*

The process described in Example 2.1.1, in which the data are Gaussian with common variance and the sequence of means is piecewise constant, is referred to as the canonical change point problem in the statistics literature. For historical reasons this problem has attracted a significant amount of attention (Yao, 1988; Venkatraman, 1992; Fryzlewicz, 2014; Wang et al., 2020; Verzelen et al., 2023); throughout this chapter the canonical problem shall be used as a running example. Other change point problems which have

Figure 2.1: Sample paths of two processes with change points. Black dashed lines (- - -) represent the unobserved sequence of θ's; that is, the parameter in which the change occurs. Light grey lines (—) represent the observed data sequence. See Examples 2.1.1 and 2.1.2 for details.



(a) Piecewise constant mean: θ_t (b) Piecewise constant variance: θ_t^2

beed studied include detecting mean changes in a sequence of high-dimensional vectors (Wang and Samworth, 2018; Cho, 2016; Enikeeva and Harchaoui, 2019), detecting changes in regression coefficients in linear models (Bai and Perron, 1998, 2003), detecting changes in covariance matrices (Wang et al., 2021; Li et al., 2023b; Avanesov and Buzun, 2018), detecting changes in factor loadings in time series factor models (Barigozzi and Trapani, 2020; Barigozzi et al., 2018; Ma and Su, 2018), and detecting changes in sequences of random networks (Padilla et al., 2022; Wang et al., 2021; Chu and Chen, 2019), to name just a few example.

The piecewise polynomial change point problem

This thesis will be principally concerned with the piecewise polynomial change point problem, in which the data can be written as the sum of a signal component and a noise component:

$$Y_t = f_\circ\,(t/n) + \zeta_t \qquad t = 1, \ldots, n.$$

The signal $f_\circ : [0,1] \mapsto \mathbb{R}$ is known to be a piecewise polynomial function of arbitrary but fixed degree p. That is, associated with $f_\circ(\cdot)$ are N integer valued change point locations Θ, such that for each $k = 1, \ldots, N$ we have that $\eta_k - \eta_{k-1} > p$ and the function can be described as a degree p polynomial on the sub-interval $[(\eta_k - p - 1)/n, \eta_k/n]$ but not on $[(\eta_k - p)/n, (\eta_k + 1)/n]$. We emphasis that the setup allows for the polynomial degree to change between segments, and indeed p should be understood as the maximum polynomial degree across all segments. The aim is to detect changes in the derivatives of $f_\circ(\cdot)$ up to order p. For now we do not place any restriction on the ζ's, however the contaminating noise should be thought of as an appropriately centered (e.g. mean centered or median centered) stochastic sequence. To fix the idea we give a concrete example below.

Example 2.1.3. *Consider the process $\{Y_t = f_\circ(t/n) + \zeta_t \mid t = 1, \ldots, n\}$ where $f_\circ(\cdot)$ is piecewise polynomial as described above and the ζ's are mutually independent with common distribution function*

$$F \in \left\{ G \mid \exists C > 0 \ s.t. \ \int \exp\left(x^2/C^2\right) G(\mathrm{d}x) \le 2 \ and \ \int x G(\mathrm{d}x) = 0 \right\}.$$

Then, the process is sub-Gaussian with piecewise polynomial mean, and the transformation of interest is: $\mathcal{H}(F) = \left(\frac{\mathrm{d}^j}{\mathrm{d}(t/n)^j j!} \int x F(\mathrm{d}x) \mid j = 0, \ldots, p \right)'$ for each $t = 1, \ldots, n$.

The piecewise polynomial change point problem generalizes the canonical change point problem in two directions: the data are no longer restricted to be Gaussian, and the sequence of centrality parameters now forms a piecewise polynomial, as opposed to a piecewise constant, sequence. Relative to the canonical change point problem, the question of change point analysis for piecewise polynomials has attracted comparatively little attention in the statistics literature. Nevertheless, in applications piecewise polynomial change point modeling is a particularly appealing choice since:

1. The methodology is flexible enough to capture complex trends in data.

2. Between change point locations the parametric trend is easily interpretable.

3. The change point locations themselves are often of interest to practitioners.

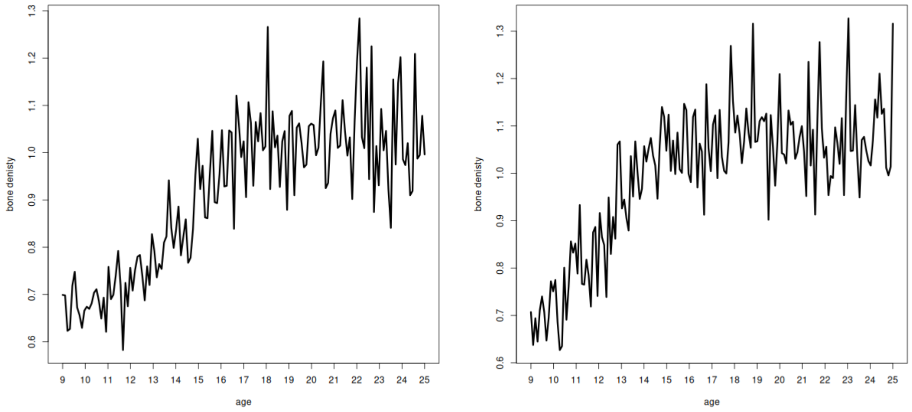
The piecewise polynomial change point model has found practical applications in areas as diverse as finance (Liu et al., 2018; McZgee and Carleton, 1970; Schröder and Fryzlewicz, 2013) where time series of (log) prices can be modeled as fluctuating around a piecewise linear trend, aerospace engineering (Cunis et al., 2019) where coefficients in full-envelope models of flight dynamics can be modeled as piecewise polynomial, light transmittance (Abramovich et al., 2007) where light availability under long-leaf pines can be modeled as a higher order piecewise polynomial function of time with jump discontinuities caused by passing clouds, climatology (Aue et al., 2009) where global near-surface temperatures can be modeled with piecewise quadratics, and epidemiology (Jiang et al., 2020, 2021) where COVID-19 infection curves can be modeled as piecewise linear, to name just a few examples.

In addition to the above, we mention three datasets which are particularly amenable to analysis using the piecewise polynomial change point model and which will be analyzed throughout in this thesis using the novel methods proposed. Each dataset presents a unique statistical challenge, in addition to the problem of change point detection, and we highlight how these challenges can be overcome using the methods proposed in this thesis.

Bone mineral density acquisition curves

Figure 2.2 shows bone mineral acquisition curves for males and females between the ages of 9 and 25. The data are obtained from a longitudinal study of 423 healthy males and females in which four consecutive yearly measurements of bone mass by dual energy x-ray absorptiometry were taken from each subject. The data can be downloaded from `hastie.su.domains`. There is some disagreement over the age at which peak bone mass density is attained in adolescents (Kröger et al., 1993; Theintz et al., 1992; Lu et al., 1996). One possible solution is to model the data in Figure 2.2 as following a piecewise linear trend, and to infer this information from any estimated change point locations. The number of available data points is however quite small after aggregating by gender and age, and as will be discussed in Section 2.1.1 estimates of change points in the piecewise polynomial model are not consistent in the usual sense. That is, the estimation error does not decay zero to with the sample size. A method which additionally quantifies the uncertainty around each estimated change point is therefore particularly useful in this setting, and we propose two such methods in Chapters 3 and 4 of this thesis.

Figure 2.2: Bone density acquisition curves based on data from 423 healthy males and females aged between 9 and 25. Data were obtained from `hastie.su.domains`; see the main text for details.



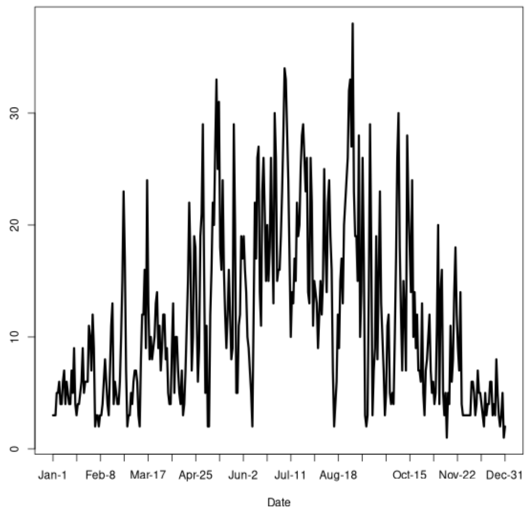
(a) male bone density acquisition

(b) female bone density acquisition

Ozone concentration in the Los Angeles basin

Figure 2.3 shows a time series of daily Ozone concentration levels (maximum of one hour averages) in the Los Angeles basin during 1976. The data is available through the `mlbench` package and was initially studied by Breiman and Friedman (1985). It is well documented that Ozone concentrations in the Northern hemisphere follow a pronounced yearly cycle with the maximum occurring towards the middle of the year (Monks, 2000). In terms of signal estimation, one visually appealing option is to model the data as piecewise linear with a single change point where concentration levels peaks. However, the data exhibit heavy tails and heteroskedasticity, and as a result a non-robust change point detection method is likely to estimate many spurious change points. In Chapter 4 we propose a method for robust change point detection and inference, which acts implicitly on signs of the data and as such is robust to heteroskedasticity and arbitrarily heavy tails, and in adition is able to localize change points at the mini-max optimal rate.

Figure 2.3: Time series of daily Ozone concentration (maximum of one hour averages) in the Los Angeles basin during 1976.



Nitrogen dioxide concentrations in Madrid, Barcelona, Valencia, and Sevilla during COVID-19 lockdowns

Figure 2.4 shows time series of nitrogen dioxide concentrations in the four largest cities in Spain - Madrid, Barcelona, Valencia, and Sevilla - during the year 2020. In this year the World Health Organization declared a Public Health Emergency of International Concern in response to the spread of the COVID-19 virus. The data were obtained via the `European air quality portal`, and consist of daily averages of hourly nitrogen dioxide concentration readings across all monitoring sites available for a particular city. In response to the COVID-19 virus the Spanish government declared a state of alarm on March 14 lasting until June 21, requiring all citizens to remain at home except to buy food and medicine and all non-essential businesses to close. There is evidence supporting improvements in air quality in urban centers during COVID-19 lock downs in various countries (Slezakova and Pereira, 2021; Jephcote et al., 2021). One approach to testing this hypothesis wold be to fit a change point model to the data in Figure 2.3 and check whether estimate change points align with the start and end of the state of emergency; inspecting Figure 2.4 a piecewise linear model seems appropriate. However, it is well know that time series of nitrogen dioxide concentrations are strongly serially correlated, and therefore change point methods which ignore this dependence are likely to estimate many spurious change points. In Sections 3.2.4 and 4.4.1 of the thesis we propose methods for change point detection which exploit strong invaraince results, by which partial sums of a dependent process can be approximated well by increments of a Wiener process, and as such are robust to serial dependence in the data.

2.1.2 Fundamental statistical limits

In modern change point estimation problems the difficulty associated with recovering all N change points is typically quantified via a global signal to noise ratio (Chan and Walther, 2013; Wang et al., 2020; Fryzlewicz, 2014), which can be defined as follows:

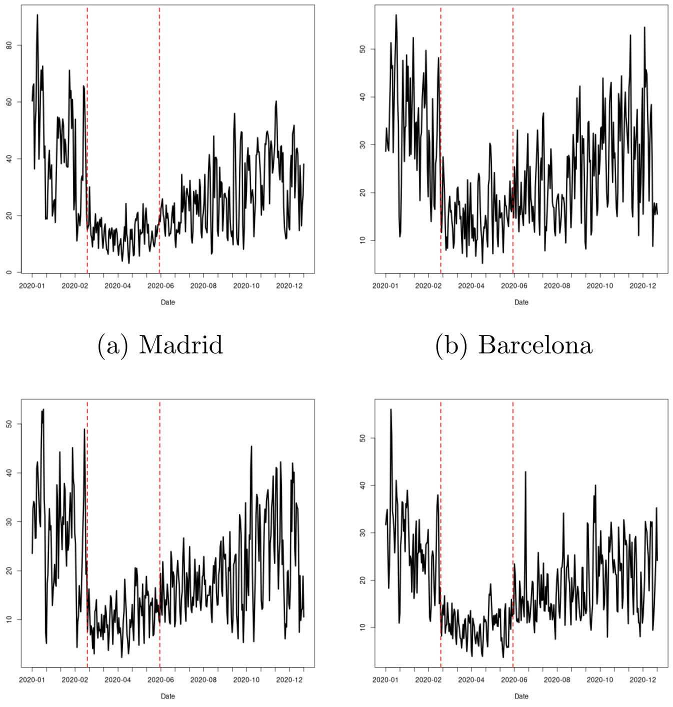
$$\overline{\mathrm{NSR}} = \tau^{-1}\sqrt{\delta}\Delta.$$

Here $\delta = \min_k \delta_k$ where $\delta_k = \min\left(\eta_k - \eta_{k-1}, \eta_{k+1} - \eta_k\right)$ measures the effective sample size associated with the k-th change point, $\Delta = \min_k \Delta_k$, and τ is some measure of dispersion for the data. Some recent papers (Cho and Kirch, 2022b; Verzelen et al., 2023; Fryzlewicz, 2023) characterize the difficulty of recovering each individual change point via a local signal to noise ratio, which can be defined as follows:

$$\mathrm{SNR}\left(k\right) = \tau^{-1}\sqrt{\delta_k}\Delta_k \qquad k = 1, \dots, N.$$

Methods which measure the difficulty of the problem via $\overline{\mathrm{NSR}}$ implicitly require the effective sample sizes and change sizes to be bounded by the same quantities for each change point

Figure 2.4: Daily concentrations of nitrogen dioxide in four Spanish cities during 2020; red dashed lines (- - -) represent the start and end dates of the Spanish national state of alarm imposed due to the COVID-19 pandemic.



(a) Madrid

(b) Barcelona

(c) Valencia

(d) Sevilla

location, and for this reason can be described as homogeneous, by contrast, methods which measure the difficulty of the problem via SNR (k) for $k = 1, \ldots, N$ allow for a combination of small changes over large intervals and large changes over small intervals, and for this reason can be described as multi-scale (Cho and Kirch, 2021). For the problem of change point estimation, there are two fundamental quantities of interest:

1. **Detection lower bound:** a quantity such that, if the signal to noise ratio falls below this level, no algorithm is guaranteed to detect all change points consistently.

2. **Localization lower bound:** the best rate at which any algorithm may localize change points under the worst possible configuration of the data for which the changes are still detectable.

There are some subtlties involved in defining the SNR in the piecewise polynomial change point problem, since at each change point location up to $p+1$ different changes in derivative may occur. It is useful to parameterize the signal between change point locations as follows:

$$f_\circ(t/n) = \begin{cases} \sum_{j=0}^{p} \alpha_{j,k} \left(t/n - \eta_k/n\right)^j & \text{if } \eta_{k-1} < t \leq \eta_k \\ \sum_{j=0}^{p} \beta_{j,k} \left(t/n - \eta_k/n\right)^j & \text{if } \eta_k < t \leq \eta_{k+1} \end{cases} \qquad k = 1, \ldots, N,$$

where moreover it holds that $\beta_{j,k} = \alpha_{j,k+1}$ for all j and each $k < N$. Then, the absolute change in the j-th derivative of $f_\circ(\cdot)$ at the k-th change point location can be written as $\Delta_{j,k} = |\alpha_{j,k} - \beta_{j,k}|$, and at the k-th change point location we have up to $p+1$ non-zero changes $\{\Delta_{0,k}, \ldots, \Delta_{p,k}\}$. To define the SNR it makes intuitive sense to use the largest or "most prominent" change in derivative. The index of most prominent change in derivative at each change point location can be defined as follows:

$$p_k^* \in \underset{0 \leq j \leq p}{\arg\max} \left\{ \Delta_{j,k} \left(\frac{\delta_k}{n} \right)^j \right\} \qquad k = 1, \ldots, N.$$

The above quantity may not be unique; although any element in the set of maximizers could be used to define the SNR, a sensible convention is to use the smallest element. Then, the global and local signal to noise ratios can respectively be defined as follows:

$$\begin{aligned}
\overline{\mathrm{SNR}} &= \tau^{-1}\sqrt{\delta}\left[\min_{1\leq k\leq N}\Delta_{p_k^*,k}\left(\frac{\delta_k}{n}\right)^{p_k^*}\right], \\
\mathrm{SNR}\,(k) &= \tau^{-1}\sqrt{\delta_k}\Delta_{p_k^*,k}\left(\frac{\delta_k}{n}\right)^{p_k^*} \qquad k = 1,\dots,N.
\end{aligned}$$

Fundamental limits in the piecewise polynomial problem have been studied by Yu et al. (2022) when the contaminating noise is sub-Gaussian and independently distributed. The detection lower bound in terms of $\overline{\mathrm{SNR}}$ was found to be of the order $\sqrt{\log(n)}$. The localization lower bound for the same problem was found by Yu et al. (2022) to be of the order

$$n^{\frac{2p_k^*}{2p_k^*+1}}\left(\frac{\tau^2}{\Delta^2_{p_k^*,k}}\right)^{\frac{1}{2p_k^*+1}} \qquad k = 1,\dots,N.$$

The localization lower bounds reveals that, provided the jumps sizes do not diverge with the sample size, estimates of each change point location will not be consistent in the usual sense: that is, the estimation error does not decay zero to with the sample size. The localization lower bounds also reveal an interesting transition between change point detection in piecewise constant signals and in higher order piecewise polynomials: for piecewise constant signals the estimation problem is local and in fact the error in estimating each change point location does not depend on the sample size, whereas for piecewise polynomial signals the problem can be thought of as global as the estimation error does depend on the sample size.

2.1.3 Change point testing

In change point testing problems one typically assumes the data contain at most one change, and the aim is to test apart the following hypotheses:

$$\begin{aligned}
&H_0 : \theta_1 = \cdots = \theta_n \\
&H_1 : \exists\eta \in \{1,\dots,n-1\} \;\text{ s.t. }\; \theta_\eta \neq \theta_{\eta+1}.
\end{aligned}$$

The general approach is to begin with the two sample testing problem in which the change, if it exists, occurs at some known location η. That is:

$$H_{1,\eta} : \theta_1 = \cdots = \theta_\eta \neq \theta_{\eta+1} = \cdots = \theta_n.$$

Then, if $S_\eta\left(\cdot\right)$ is a test statistic for testing H_0 against $H_{1,\eta}$, the problem of testing for an unknown change point location can be resolved with the statistic $S\left(\cdot\right) = \max_{\eta \in I}\left\{S_\eta\left(\cdot\right)\right\}$ where I is some ordered subset of $\{1, \ldots, n\}$. Natural choices for $S_\eta\left(\cdot\right)$ include the likelihood ratio, Wald, and Lagrange statistics, as well as various U-statistics. For establishing the behavior of $S\left(\mathbf{Y}\right)$ under H_0 there are three main approaches.

1. If $I = \{n_0, \ldots, n - n_0\}$ and n_0 diverges with n in such a way that $n_0/n \to C \in [0, 1/2)$, or the local statistics involve weights such that $S_\eta\left(\cdot\right)$'s with η close to the boundaries 1 and n are down-weighted, it is often possible to show that $S\left(\mathbf{Y}\right)$ converges to some functional of a Bridge process Andrews (1993).

2. If $|I| = \mathcal{O}\left(n\right)$ and the data are either independent and Gaussian or satisfy the conditions of a strong approximation theorem, it is often possible to show that after appropriate centering and scaling $S\left(\mathbf{Y}\right)$ converges to an extreme value distribution Csörgö et al. (1997).

3. If I is sufficiently sparse, and / or enough is know about the sequence of distribution functions $F_1, \ldots, F_n$, concentration inequalities may be used to tightly bound $S\left(\mathbf{Y}\right)$ on a high probability set (Enikeeva and Harchaoui, 2019; Liu et al., 2021; Verzelen et al., 2023).

To fix the idea we give a concrete example below.

Example 2.1.4. *Consider a sample* $\mathbf{Y} = \left(Y_t \mid t = 1, \ldots, n\right)'$ *from the canonical change point process introduced in Example 2.1.1, where additionally the sequence of means* $\left\{\theta_t \mid t = 1, \ldots, n\right\}$ *is piecewise constant with at most one change. The square root of the log likelihood ratio statistic, also known as the CUSUM statistic, for testing* H_0

against $H_{1,\eta}$ has the form:

$$S_\eta(\boldsymbol{Y}) = \sqrt{\frac{\eta(n-\eta)}{n}} \left| \frac{1}{\eta} \sum_{t=1}^{\eta} Y_t - \frac{1}{n-\eta} \sum_{t=\eta+1}^{n} Y_t \right|.$$

Moreover putting $\mathfrak{a}_n = \sqrt{2\log\log(n)}$ and $\mathfrak{b}_n = \frac{1}{2}\log\log\log(n) - \log(\sqrt{\pi})$ it can be shown (Csörgö et al., 1997) that as $n \to \infty$, under H_0, for any fixed $z \in \mathbb{R}$

$$\mathbb{P}(\mathfrak{a}_n S(\boldsymbol{Y}) - \mathfrak{b}_n \leq z) \to e^{-2e^{-z}}$$

where in particular $S(\boldsymbol{Y}) = \arg\max_{1\leq\eta<n} S_\eta(\boldsymbol{Y})$.

In the context of change point testing for piecewise polynomials (Jarusková, 1999; Aue et al., 2008, 2009) have studied the supremum of likelihood ratio test statistics for testing each of $\{H_{1,\eta} \mid (p+1) \leq \eta \leq n-(p+1)\}$ against the null of no change points. The latter papers establish the convergence of the maximum of local test statistics to an extreme value distribution, whereas the paper by Jarusková (1999) makes use of an analytic approximation for the tail of the maximum.

If the goal is to test for change points in a sequence of centrality parameters, an alternative approach to change point testing involves applying a transformation to partial sums of the data is such a way that the transformed sequence of partial sums is invariant to changes in the centrality parameter of interest. Then, the facts that the transformed partial sum process is pivotal, and that under H_0 and some technical conditions scaled partial sums of the data will satisfy a functional central limit theorem, can be used to construct a test with asymptotically correct size. We give a concrete example below.

Example 2.1.5. *Consider a sample path $\boldsymbol{Y} = (Y_t \mid t = 1, \ldots, n)'$ from the canonical change point process introduced in Example 2.1.1, and define the partial sum process* *as* $\mathbb{Y}_n = \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{[ns]} (Y_t) \mid 0 \leq s \leq 1 \right\}$. *Then writing*

$$\mathcal{Q}(\mathbb{Y}_n) = \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{[ns]} Y_t - \frac{1}{\sqrt{n}} \frac{[ns]}{n} \sum_{t=1}^{n} Y_t \mid 0 \leq s \leq 1 \right\}$$

it is clear that the distribution of $\mathcal{Q}(\mathbb{Y}_n)$ *is invariant to constant shifts in the mean of the* Y*'s, and under the null of no change points* $\mathcal{Q}(\mathbb{Y}_n) \Rightarrow \{B(s) - sB(1) \mid 0 \leq s \leq 1\}$ *as* $n \to \infty$ *where "*$\Rightarrow$*" denotes weak convergence and* $\{B(s) \mid 0 \leq s \leq 1\}$ *is a standard Wiener process.*

This approach was pioneered by Kuan and Hornik (1995), and has been applied to the specific problem of testing for changes in piecewise polynomials in Kuan (1998). A natural way to eliminate a polynomial trend is to work with residuals from a least squares fit, and the empirical processes obtained from the sequence of partial sums of such residuals has been studied by (Jandhyala and Minogue, 1993; Jandhyala and MacNeill, 1997; MacNeill, 1978) who again establish convergence to a Bridge-like process under some technical condition on the contaminating noise.

2.1.4 Change point estimation

In change point estimation problems the goal is to accurately recover the number of change points and their locations. Methods for change point estimation can be broadly divided it two classes, global segmentation methods and greedy or local segmentation methods.

Global segmentation methods

Global segmentation methods aim to recover Θ, the set of change point locations, and as a consequence also N, the number of change points, in one step by solving a single optimization problem. We review the two most common approaches to global segmentation: algorithms based ℓ_0 penalization and algorithms based on their convex relaxations.

Algorithms based on ℓ_0 penalization

For integers $s < e$ let $\boldsymbol{Y}_{s:e}$ denote the sub-vector of $\boldsymbol{Y}$ consisting of all elements indexed by $t = s+1, \ldots, e$. Moreover let $\mathcal{C}\left(\boldsymbol{Y}_{s:e}; \theta\right)$ denote a loss function measuring the level of agreement between the data on the segment $\{s+1, \ldots, e\}$ and a given θ. Global segmentation methods estimate the number of change points and their locations by solving:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{k=0}^{N} \left\{ \min_{\theta_k} \mathcal{C}\left(\boldsymbol{Y}_{\eta_k:\eta_{k+1}}; \theta_k\right) \right\} + \mathrm{pen}\left(\Theta\right). \tag{2.1}$$

The minimization is over all subsets of the set $\{1, \ldots, n-1\}$, and the dependence of N on Θ is left implicit. Finally, pen $(\cdot)$ stands for a penalty function which penalizes a proposed segmentation of the data in terms of its complexity. As long as the loss function and cost function are additive, the above optimization problem can be solved efficiently via dynamic programming (Killick et al., 2012; Maidstone et al., 2017). To fix the idea, we give a concrete example below.

Example 2.1.6. *Consider a sample path $\boldsymbol{Y} = (Y_t \mid t = 1, \ldots, n)'$ from the canonical change point process introduced in Example 2.1.1. Estimating the number of change points and their locations by minimizing twice the negative log-likelihood with a penalty linear in the number of change points leads to:*

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{k=0}^{N} \left\{ \min_{\theta_k \in \mathbb{R}} \sum_{\eta_k \leq t < \eta_{k+1}} \left(Y_t - \theta_k\right)^2 \right\} + \lambda N$$

where λ is a constant to be tuned.

Putting $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_n\right)'$ for the vector of estimated θ's associated with the solution to the generic optimization problem given in (2.1), it is clear that the estimated number of change points can be written as $\hat{N} = \sum_{t=1}^{n-1} \mathbf{1}_{\left\{\hat{\theta}_t \neq \hat{\theta}_{t+1}\right\}}$. Consequently, penalizing the estimated number of change points is equivalent to penalizing the ℓ_0 norm of the first difference of the vector $\hat{\boldsymbol{\theta}}$.

The most common choice of penalty is the Schwartz penalty which was first studied by Yao (1988) and is given by pen $\left(\hat{\Theta}\right) = \lambda \hat{N} \log(n)$. Here, λ is a parameter which must be tuned by the user, and which generally involves the some measure of the data's dispersion. Schwartz-like penalties are generally optimal for localizing change points in homogeneous problems (using the language of Section 2.1.1), however they can be sub-optimal for multi-scale change point problems. Some authors have proposed penalties which additionally account for the spacing of estimated change points (Zhang and Siegmund, 2007; Davis et al., 2006). Recently Verzelen et al. (2023) proposed only multi-scale which in the piecewise constant mean problem was proven to estimate multi-scale change points optimally, and in certain settings is even able to recover change points when SNR is of the order of a constant.

In terms of change point detection in the piecewise polynomial model, Fearnhead et al. (2019) showed that an estimator basted on optimizing the ℓ_0 penalized least squares function, similar to Example 2.1.6, is optimal for detecting change points in piecewise linear and continuous signals. The generic piecewise polynomial model has been studied by (Yu et al., 2022), who showed that an estimator likewise based on ℓ_0 penalized least squares can detect and localize change points at optimal rates in most regimes. However, the algorithm proposed by Yu et al. (2022) has cubic time complexity in the worst case, making it impractical for larger datasets. Moreover, the optimal penalty discussed in the paper in fact depends on the number of change points, which is a quantity that is not typically known in advance.

Convex relaxations of ℓ_0 constraints

In order to gain in computational efficiency, some authors suggest using a convex relaxation of the original ℓ_0 constraint, and instead penalizing the total variation semi-norm of the vector of estimated θ's. This approach has been taken by (Harchaoui and Lévy-Leduc, 2007, 2010; Lin et al., 2017) among others, and results in the following optimization problem for change point detection in signals with piecewise constant means:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\arg\min} \|\boldsymbol{Y} - \boldsymbol{\theta}\|_2^2 + \lambda \sum_{t=2}^{n} |\theta_t - \theta_{t-1}| \, .$$

Penalizing the total variation semi-norm encourages a piecewise constant structure in $\hat{\boldsymbol{\theta}}$. The number of change points and their locations can be extracted from the estimated $\hat{\boldsymbol{\theta}}$ respectively via $\hat{N} = \sum_{t=1}^{n-1} \mathbf{1}_{\left\{\hat{\theta}_t \neq \hat{\theta}_{t+1}\right\}}$ and $\hat{\Theta} = \left\{\eta \mid \hat{\theta}_\eta \neq \hat{\theta}_{\eta+1}\right\}$. Moreover, the solution to the above can be efficiently computed using the LARS algorithm proposed by Efron et al. (2004).

The above procedure naturally extends to the setting of piecewise polynomial signals of general degrees, by replacing the penalty on the fist difference of the sequence of θ's with a penalty on their $(p+1)$-th difference. This approach is known as ℓ_1 trend filtering, and was studied independently by Kim et al. (2009) and Tibshirani (2014). The optimization problem to be solved becomes:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\arg\min} \, \|\boldsymbol{Y} - \boldsymbol{\theta}\|_2^2 + \lambda \sum_{t=p+2}^{n} \left| \sum_{i=0}^{p+1} (-1)^i \binom{p+1}{i} \theta_{t-i} \right|.$$

We stress that the original ℓ_1 trend filtering algorithm was developed for the purpose of signal estimation, and not change point detection. Nevertheless, estimates for the number of change points and their locations can be extracted from $\hat{\boldsymbol{\theta}}$, and some recent works (Mehrizi and Chenouri, 2021, 2020) have studied the performance of ℓ_1 trend filtering for change point detection in the generic piecewise polynomial model.

While they are attractive from the computational perspective, the algorithms discussed in this section tend to offer sub-optimal statistical guarantees. For example Cho and Fryzlewicz (2011) show that the total variation penalty is sub-optimal for change point detection in piecewise constant signals by appealing to results on the efficiency of various change point tests discussed in Brodsky and Darkhovsky (1993). Moreover, Rojas and Wahlberg (2014) observe that for the same problem the method may be altogether inconsistent if the data contain a so called *staircase pattern*; that is, multiple changes in mean all having the same sign.

Greedy or local segmentation methods

In contrast to global segmentation methods, greedy or local segmentation methods aim to recover the change point locations one at a time by solving a separate optimization problem for each change point estimated. We review the two most common such approaches: binary segmentation algorithms and scanning algorithms.

Binary segmentation

The binary segmentation algorithm was originally studied by (Vostrikova, 1981; Venkatraman, 1992) for change point detection in the canonical setting, but can be generalized to more complex change point problems. Using the language of Section 2.1.3 let $S_{s,\eta,e}\left(\cdot\right)$ be a test statistic for testing apart

$$
\begin{aligned}
&H_0^{s:e} : \theta_s = \cdots = \theta_e \\
&H_{1,\eta}^{s:e} : \theta_s = \cdots = \theta_\eta \neq s_{\eta+1} = \cdots = \theta_e.
\end{aligned}
$$

The main idea behind binary segmentation is that the quantity $\hat{\eta} = \arg\max_{1\leq\eta<n} S_{1,\eta,n}\left(\mathbf{Y}\right)$ will generally be consistent for one of the change point locations in the data. The generic binary segmentation algorithm estimates the first change point in this way, then recursively repeats the search for the next most likely change point location to the left and right of $\hat{\eta}$ until no more statistically significant significant change points can be found. Pseudo code for the generic binary segmentation algorithm is given below in Algorithm 1, where λ is a tuning parameter chosen such that the probability of the following event, under the global null of no change points, is small:

$$
E = \left\{ \max_{1\leq s<e\leq n} \max_{s\leq\eta<e} S_{s,\eta,e}\left(\boldsymbol{Y}\right) > \lambda \right\}. \tag{2.2}
$$

Binary segmentation is computationally efficient and easy to code, but suffers from several drawbacks: the procedure lacks power and in general will not be able to estimate change points at the best rate (Fryzlewicz, 2014), and in some change point models

Algorithm 1: The generic binary segmentation algorithm for change point estimation. Inputting start and end points $\{s, e\}$ and a threshold λ the algorithm recursively estimates the locations of any change points in the data $\boldsymbol{Y}$ in the ordered set $\{s, \ldots, e\}$.

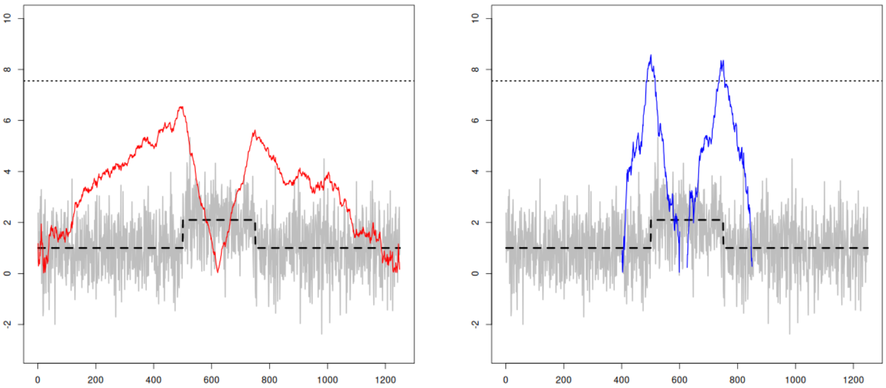
```
function BinarySegmentation($s, e, \lambda$):
    if $e - s < 1$ then
        STOP
    end
    $\hat{\eta} \leftarrow \arg\max_{s \leq \eta < e} S_{s,\eta,e}(\boldsymbol{Y})$
    if $S_{s,\hat{\eta},e}(\boldsymbol{Y}) > \lambda$ then
        RecordChangePoint($\hat{\eta}$)
        BinarySegmentation($s, \hat{\eta}, \lambda$)
        BinarySegmentation($\hat{\eta} + 1, e, \lambda$)
    end
    else
        STOP
    end
return
```

$\hat{\eta} = \arg\max_{1 \leq \eta < n} S_{1,\eta,n}(\mathbf{Y})$ may not be a consistent estimator for any change point location in the data (Baranowski et al., 2019a). These issues arise because the algorithm may inspect a stretch of the data containing more than one change point, but each statistic $S_{s,\eta,e}(\cdot)$ is designed to test against the alternative of exactly one change point at location η. We give a concrete example of this problem below.

Example 2.1.7. *Figure 2.5 shows a sample path* $\boldsymbol{Y} = (Y_t \mid t = 1, \ldots, 1250)'$ *from the canonical change point problem introduced in Example 2.1.1 with* $\theta_t = 1 + 1.1 \times \mathbf{1}_{\{500 < t \leq 750\}}$ *along with the CUSUM statistic from Example 2.1.4 calculated on the entire sample path (2.5a) and on two sub-intervals each containing a single change point location (2.5b). The threshold* $\lambda = \sqrt{8 \log(n)}$ *with* $n = 1250$*, shown in Baranowski et al. (2019b) to control the event (2.2), is also plotted. The CUSUM statistic calculated on the entire sample does not exceed the threshold anywhere, whereas the two statistics calculated on localized intervals do exceed the threshold and their maxima occur close to the true change point locations.*

To correct the drawbacks of binary segmentation, modern variants of the algorithms

Figure 2.5: CUSUM statistic computed on a sample path from the canonical change point problem. Light grey lines (—) represent the observed data sequence, and black dashed lines (- - -) represent the unobserved piecewise constant mean. Coloured lines (— / —) represent the value of the CSUSM statistic, and black dotted lines $(\cdots)$ represent the $\sqrt{8\log(n)}$ threshold. See Example 2.1.7 for more details.



(a) CUSUM statistic computed on the entire data sequence

(b) CUSUM statistic computed on localized intervals

focus on localizing the procedure by applying it to a grid of intervals designed in such a way that for each change point location there is likely to be one interval in the grid which contains only that change point. This is generally possible under some mild assumptions on the distance between change points (Yu et al., 2022). For example (Fryzlewicz, 2014; Baranowski et al., 2019a) propose to use intervals with start and end points drawn uniformly at random from $\{1, \dots, n\}$, Kovács et al. (2023) propose a deterministic grid of intervals with exponentially decaying lengths, and (Anastasiou and Fryzlewicz, 2022; Fang and Siegmund, 2020) propose to use a sequence of gradually expanding intervals. All of the aforementioned variants of binary segmentation are able to estimate change points in the piecewise constant signal setting with optimal rates. Moreover, the works of Baranowski et al., Kovács et al., and Anastasiou and Fryzlewicz are optimal for estimating change points in piecewise linear and continuous signals. However, we are unaware of any variant of binary segmentation which has been proved to estimate change points optimally in the generic piecewise polynomial model.

Scanning algorithms

The main idea behind scanning algorithms is to scan through the data with a window of fixed length say $2W$, where W is referred as the bandwidth, and at each point test whether the window contains a change point location. In the language of Section 2.1.3 let $S_{W,\eta}(\cdot)$ be a test statistic for testing apart:

$$\begin{aligned} H_0^{W,\eta} &: \theta_{\eta-W+1} = \cdots = \theta_{\eta+W} \\ H_{1,\eta}^{W,\eta} &: \theta_{\eta-W+1} = \cdots = \theta_{\eta} \neq \theta_{\eta+1} = \cdots = \theta_{\eta+W}. \end{aligned}$$

Then heuristically the local maxima of the process $\mathbb{S}_W = \{S_{W,\eta}(\boldsymbol{Y}) \mid W \leq \eta \leq n - W\}$ are likely to occur near locations of any change points in the data. Such scanning algorithms have historically been used for change point testing (Hušková and Slabỳ, 2001; Chu et al., 1995; Bauer and Hack1, 1980) when the data may contain multiple change points, however recent papers have applied the same approach to the problem of change point estimation. The typical approach to change point estimation with scanning algorithms is to choose a threshold λ such that under the null of no change points the probability of the event $\{\max_{W \leq \eta \leq n-\eta} S_{W,\eta}(\boldsymbol{Y}) > \lambda\}$ is small. Then, letting $\mathcal{M}_W$ be the set of all integer pairs (L, R) satisfying:

1. $R - L \geq cW$ for some $c \in (0, 1/2)$
2. $S_{W,\eta}(\boldsymbol{Y}) > \lambda$ for $\eta \in \{L, \ldots, R\}$
3. $S_{W,\eta}(\boldsymbol{Y}) \leq \lambda$ for $\eta \in \{L-1, R+1\}$

Then the number of change points in the data can be estimated via $\hat{N} = |\mathcal{M}_W|$ and the change point locations can be estimated via $\hat{\eta}_k = \arg\max_{L_k \leq \eta \leq R_k} S_{w,\eta}(\boldsymbol{Y})$ for each $k = 1, \ldots, \hat{N}$.

When the bandwidth is carefully chosen, the above approach to is generally able to estimate change points at the optimal rate. For example, Eichinger and Kirch (2018) propose a scanning algorithm for change point detection in piecewise constant signals and

Kim et al. (2022) propose an algorithm for change point detection in piecewise linear (though not necessarily continuous) signals, and both attain the respective optimal rates discussed in Section 2.1.1. As with binary segmentation algorithms, we are unaware of any scanning algorithm which has been proved to estimate change points optimally in the generic piecewise polynomial model.

A crucial limitation of scanning algorithms is the need to choose a bandwidth parameter. In order to attain best rates for change point estimation, the bandwidth should be chosen to be as large as possible without exceeding the minimum distance between change points in the data. However, this information is not generally know to the analyst. If the bandwidth is chosen to be too small the procedure will lack power and may not be able to detect all change points, whereas if the bandwidth is chosen to be too large the procedure may not be able to distinguish between adjacent change points which are too close. We illustrate this problem with a concrete example below.

Example 2.1.8. *Figure 2.6 shows a sample path* $\boldsymbol{Y} = (Y_t \mid t = 1, \ldots, 1250)'$ *from the canonical change point problem introduced in Example 2.1.1, with the same sequence of* θ*'s as in Example 2.1.7. The process* $\mathbb{S}_W$ *employing the test statistic from example 2.1.4 (known in the literature as the MOSUM process) with bandwidths* $W = 20$ *and* $W = 50$ *is also plotted. We additionally plot the threshold*

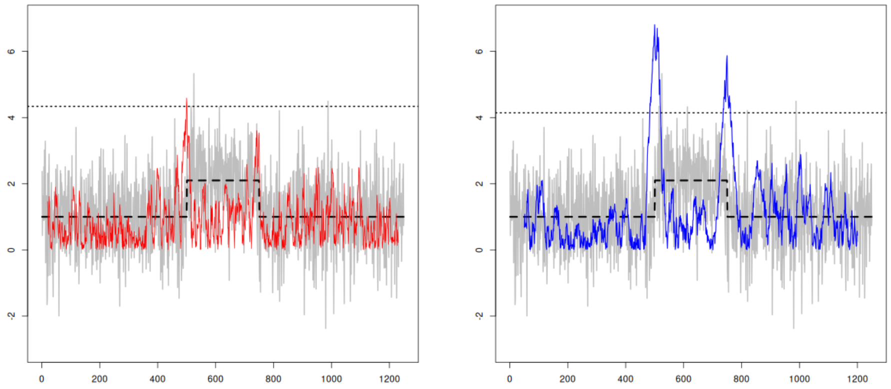
$$\lambda = \sqrt{2\log(n/W)} + \frac{\frac{1}{2}\log\log(n/W) + \log(\frac{3}{2\sqrt{\pi}}) + z_\alpha}{\sqrt{2\log(n/W)}} \tag{2.3}$$

proposed in Eichinger and Kirch (2018), where $z_\alpha = \log\left(-2\log^{-1}(1-\alpha)\right)$ *and* $\alpha = 0.05$*. The MOSUM statistic calculated using the smaller bandwidth only exceeds the threshold in the vicinity of one of the change points, whereas the statistic using the larger bandwidth exceeds the threshold in the vicinity of both change points.*

A handful of recent papers have proposed solutions to the bandwidth selection problem. For example, Cho and Kirch (2022b) propose using a grid of bandwidths and pruning the list of estimated change points via an additional model selection step. The paper by

Levajković and Messer (2023) considers using all possible bandwidths from 1 to $n-1$.

Figure 2.6: MOSUM statistic computed on a sample path from the canonical change point problem. Light grey lines (—) represent the observed data sequence, and black dashed lines (**- - -**) represent the unobserved piecewise constant mean. Coloured lines (— / —) represent the value of the MOSUM statistic, and black dotted lines ($\cdots$) represent the threshold proposed in Eichinger and Kirch (2018) with $\alpha = 0.05$. See Example 2.1.8 for details.



(a) MOSUM statistic with bandwidths $W = 20$

(b) MOSUM statistic with bandwidths $W = 50$

2.2 Change point inference

This section reviews the literature on statistical inference in change point problems, where the aim is to quantify the level of uncertainty around one or more of: the number of changes N; their locations Θ; the jump sizes $\{\Delta_k \mid k = 1, \dots, N\}$. We review the four main approaches to change point inference, namely: post selection inference, simultaneous inference and selection, inference without selection, and inference through Bayesian analysis.

2.2.1 Post selection inference

Having obtained an estimate $\hat{\Theta}$ for the change point locations from data, a naive approach to change point inference would involve applying two sample homogeneity tests to data in a neighborhood of each estimated change point location in order to test whether each estimated change is in fact “real”. However, such tests would not have correct level, since the data would have effectively been used twice: once to decide where to carry out each test, and a second time to carry out the test itself. Therefore, viewing change point estimation as a model selection problem, the question of change point inference translates to post selection inference.

The simplest practical solution is to use sample splitting (Fithian et al., 2014). For example, one could use odd indexed data points to estimate the number of change points and their locations, then perform tests for the significance of each change point on the remaining even indexed data. A modern alternative to sample splitting is data thinning (Rasines and Young, 2023; Neufeld et al., 2023). For random variables from the natural exponential family data thinning takes a random variable Y and decomposes into two independent random variables $Y^{(1)}$ and $Y^{(2)}$ such that: (i) $Y \overset{d}{=} Y^{(1)} + Y^{(2)}$, and (ii) $Y^{(1)}$ and $Y^{(2)}$ follow the same distribution as Y up to a known scaling factor. Given a sample path $\boldsymbol{Y}$ from a change point model, one could therefore use the first half of the thinned path, say $\boldsymbol{Y^{(1)}}$, for change point estimation, and the second half of the thinned path, say $\boldsymbol{Y^{(2)}}$, for testing. An application of data thinning to change point detection is given in Dharamshi et al. (2023).

Both sample splitting and data thinning suffer from a loss of accuracy and power. The loss of accuracy occurs in the first stage, where only part of the sample is used for change point estimation leading to less accurate estimates of the change point locations. In the second stage each test carried out suffers from a loss of power, since again not all of the data is being used. A refinement involves using the entire sample for estimation and testing, but carefully conditioning on the model selection step when each test is carried out. Most research in this direction has focused on the piecewise constant mean model. For example Duy et al. (2020) compute valid conditional p-values for when the change point locations are estimated vai ℓ_0 penalized least squares, Hyun et al. (2021) do the same when the change point locations are estimated via binary segmentation or ℓ_1 trend filtering, and (Jewell et al., 2022; Carrington and Fearnhead, 2023) do the same when the change point are estimated by either (wild) binary segmentation or ℓ_0 penalized least squares. Finally, in the more general piecewise polynomial setting Mehrizi and Chenouri (2021) obtain valid conditional p-values when the change point locations are estimated via ℓ_1 trend filtering.

A different but complementary approach involves the fact that if each change point location can be localized at a fast enough rate then a limit distribution for the change point estimator can be obtained. Based on this, a confidence interval for each estimated change point location can be calculated. This approach goes back to (Antoch, 1999; Antoch et al., 1995), who studied the distribution of the estimated change point location in the piecewise constant mean model with a single change. Bai (1995); Bai and Perron (1998) give the asymptotic distribution of estimated change point locations in the piecewise linear regressions model when N is fixed. Recently (Kaul and Michailidis, 2023; Kaul et al., 2021) studied the asymptotic distribution of change point location estimates in the high dimensional mean shift model. In the piecewise constant mean model Cho and Kirch (2022a) develop bootstrap confidence intervals for the change point locations. We emphasize the difference between this inference problem and the one above: in the paragraph above the goal was to quantify the uncertainty about the *existence* of each change point, whereas here the goal is to quantify uncertainty about the *location*.

The inference methods discussed so far suffer from an important practical limitation: they are only valid conditional the number of change points being correctly estimated, which in a finite sample is not guaranteed to occur. Therefore, confidence statements for change point problems arrived at via post selection inference can be problematic to interpret in practice.

2.2.2 Simultaneous inference and selection

Rather than performing the change point estimation and inference steps sequentially, one may do the two simultaneously. This line of research was initiated with the work of Frick et al. (2014) who studied change point models from the one parameter exponential family, with piecewise constant parameter θ. That is:

- Each $Y_t \sim F_{\theta_t}$ for $t = 1, \ldots, n$.

- The sequence $\{\theta_t\}_{t=1}^n$ is piecewise constant.

- $\{F_\theta\}_{\theta \in \Gamma}$ is a one dimensional exponential family.

Frick et al. introduce the Simultaneous MUltiscale Change-point Estimator (SMUCE), which estimates the change point locations by minimizing the number of jumps in the estimated sequence of θ's subject to the constraint that the empirical residuals pass a multi-scale test having size α. More precisely, they solve the optimization problem

$$\min_{\hat{\boldsymbol{\theta}}} \quad \sum_{t=1}^{n-1} \mathbf{1} \left\{ \hat{\theta}_t \neq \hat{\theta}_{t+1} \right\} \text{ s.t. } T_n \left(\boldsymbol{Y}, \hat{\boldsymbol{\theta}} \right) \leq q \tag{2.4}$$

where the minimization is over all piecewise constant vectors, and the multi-scale test is

$$T_n \left(\boldsymbol{Y}, \hat{\boldsymbol{\theta}} \right) = \max_{\substack{1 \leq s < e \leq n \\ \hat{\theta}_t = \bar{\theta} \text{ for } t \in (s,e]}} \left\{ T_{s:e} \left(\boldsymbol{Y}, \bar{\theta} \right) - \sqrt{2 \log \frac{en}{e-s}} \right\}. \tag{2.5}$$

Finally, $T_{s:e} \left(\boldsymbol{Y}, \bar{\theta} \right)$ is the square root of the twice the log-likelihood ratio statistic for testing apart the local hypotheses

$$\begin{aligned} H_0^{s:e} &: \theta_t = \bar{\theta} \ \forall \ t = s+1, \ldots, e \\ H_1^{s:e} &: \theta_t \neq \bar{\theta} \ \forall \ t = s+1, \ldots, e. \end{aligned}$$

An appealing property of the SMUCE estimator is that solving (2.4) produces a confidence set for all admissible piecewise constant vectors of θ's, from which uniform confidence sets for the change point locations can be extracted. The parameter q can be chosen such that at the true vector of θ's we have $\lim_{n \to \infty} \mathbb{P} \left(T_n \left(\boldsymbol{Y}, \boldsymbol{\theta} \right) \leq q \right) \geq 1 - \alpha$. Therefore, with a proper choice of q the SMUCE estimator can produce asymptotically $1 - \alpha$ level confidence sets for the change point locations.

To fix the idea, below we give a concrete example of the local tests used by the SMUCE estimator for a particular exponential family distribution.

Example 2.2.1. *Let* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *be a sample path from the canonical change point problem introduced in Example 2.1.1. Then the local likelihood ratio test statistics used in the SMUCE algorithm will be of the form*

$$T_{s:e}\left(\boldsymbol{Y}, \bar{\theta}\right) = \frac{1}{\sqrt{e-s}} \left| \sum_{s=s+1}^{e} \left(Y_t - \bar{\theta}\right) \right|. \tag{2.6}$$

We stress that in (2.5) the local test statistics are only calculated on stretches of the data where the proposed vector of θ*'s is constant.*

Many variants of the original SMUCE procedure have been proposed. For example: Pein et al. (2017) proposes a heterogeneous extension of SMUCE, called H-SMUCE, in which the multi-scale test involves a local variance estimator and therefore permits a certain degree of heterogeneity in the data; Jula Vanegas et al. (2021) propose the Multi-scale Quantile Segmentation procedure (MQS), in which the authors look for change points in the quantiles of a data sequence, and use a variant of the SMUCE procedure in which the multi-scale test is based on signs of empirical residuals and therefore requires almost no assumptions on the distribution of the data; Dette et al. (2020) extend the SMUCE estimator to serially dependent data sequences (Dep-SMUCE) scaling the basic SMUCE statistic (2.6) by a consistent estimator of the data's long-run variance and appealing to strong approximation results by which partial sums of the data can be approximated well by a Gaussian process.

Although the family of SMUCE estimators seem to have avoided the post selection inference problem by doing estimation and inference simultaneously, in practice this is not strictly true. In fact, they suffer from the following drawback which was first noted by Chen et al. (2014): letting α, through q, determine both the nominal coverage level and the estimated piecewise constant vector of θ's leads to the counter-intuitive situation in which larger nominal coverage may reduce actual coverage.

2.2.3 Inference without selection

A further approach to change point inference involves skipping the model selection step altogether, and focusing exclusively on inference via detecting intervals which each contain a change point location with high probability. Consider the generic change point problem introduced in Section 2.1.1. Let $T_{s:e} : (Y_s, \ldots, Y_e)' \mapsto \{0, 1\}$ test the local null hypothesis

$$H_0^{s:e} : \theta_s = \cdots = \theta_e$$

and let $\mathcal{T}(\boldsymbol{Y}) = \{T_{s:e}(\boldsymbol{Y}) \mid (s, e) \in \mathcal{G}\}$ be a collection of such tests indexed over some grid $\mathcal{G}$ of (s, e) pairs. If the collection of tests has family-wise error bounded by some $\alpha \in (0, 1)$ it is immediate that uniformly with probability at least $1 - \alpha$ each (s, e) pair for which a local null is rejected must correspond to an interval $\{s, \ldots, e\}$ which contains at least one change point location. Methods which perform inference based on this idea proceed in two steps:

1. A gird $\mathcal{G}$ is specified, and a collection of suitably powerful local tests with bounded family-wise error on the chosen gird is proposed.

2. An algorithm is introduced for turning the collection of local tests into a collection of mutually disjoint intervals with the following properties: (i) the corresponding local null is rejected on each interval, and (ii) each interval is as short as possible.

Intervals recovered in this way can be understood as simultaneously quantifying the uncertainty around the existence and around the location of each putative change point.

Such a general scheme is discussed in Pilliat et al. (2023), who study in detail the problem of detecting changes in the mean of high dimensional sub-Gaussian data. The authors propose local tests based on symmetric differences in means, indexed over a grid of sub-intervals of $\{1, \ldots, n\}$ for which each interval in the grid has dyadic length. They suggest an algorithm for interval recovery based on inspecting intervals in their grid from the smallest to the largest, and merging overlapping intervals which each detect a change. Fang et al. (2020) and Fang and Siegmund (2020) obtain analytic approximations to the supremum of likelihood ratio statistics for testing changes in the mean and changes in the linear trend of a sequence of independent Gaussian random variables at all scales and locations, under the null of no change points. This can be seen to correspond to the complete grid of all sub-interval of the index set $\{1, \ldots, n\}$. The authors propose an algorithm based on testing gradually expanding sub intervals of the index set, which is similar to the Isolate Detect procedure for change point estimation proposed by Anastasiou and Fryzlewicz (2022).

Fryzlewicz (2023) proposed the Narrowest Significance Pursuit (NSP) procedure for inference it the linear regression model with piecewise constant coefficients: $Y_t = \beta_t' X_t + \zeta_t$ for $t = 1, \ldots, n$ with $\{\beta_t \mid t = 1, \ldots, n\}$ being a piecewise constant sequence of vectors. This model includes as a special case the piecewise polynomial regression model. The procedure is based on the following local tests:

$$T_{s:e}^{\lambda}(\boldsymbol{Y}) = \mathbf{1}\left\{\min_{\hat{\beta}} \max_{s \leq i \leq j \leq e} \frac{1}{\sqrt{j-i+1}} \left| \sum_{t=i}^{j} \left(Y_t - \hat{\beta}' X_t\right) \right| > \lambda \right\}, \quad \text{for } 1 \leq s \leq e \leq n \tag{2.7}$$

These tests have the appealing property that, uniformly over all intervals free from change points and for any sequence of design matrices $\{X_t \mid t = 1, \ldots, n\}$, it must hold that

$$\min_{\hat{\beta}} \max_{s \leq i \leq j \leq e} \frac{1}{\sqrt{j-i+1}} \left| \sum_{t=i}^{j} \left(Y_t - \hat{\beta}' X_t\right) \right| \leq \max_{1 \leq i \leq j \leq n} \frac{1}{\sqrt{j-i+1}} \left| \sum_{t=i}^{j} \zeta_t \right| . \tag{2.8}$$

Kabluchko and Wang (2014) study the limiting distribution of standardized increments of partial sum of random variables belonging to a range of distributions. Therefore, with knowledge of the distribution of the ζ's their results can be used to select a λ which asymptotically controls the family-wise error of the collection of tests (2.7) at a desired level.

For recovering intervals with the local tests (2.7) Fryzlewicz proposed the Narrowest Significance Pursuit algorithm, which we describe in detail here as it will be used in the change point inference procedure proposed in Chapter 4. Pseudo code for the algorithm is provided in Algorithm 2 below. The algorithm is defined recursively, and begins by generating a gird of sub-intervals indexing the stretch of data being inspected via the function `subIntervalsGrid`. Then, each sub-interval is tested for a change point using a pre-specified local test. The intervals are tested from shortest to longest, and among intervals with the same lengths the order of testing is determined by the the left endpoint of the interval, with intervals having smaller values being inspected first. If all sub-intervals pass their respective tests the algorithm terminates. Else, on the first sub-interval for which a local test is not passed the algorithm starts a second exhaustive search for the narrowest sub-interval on which a change can be detected. This is done by inspecting intervals generated by the function `allSubIntervals`, which given start and end values which draws all contiguous sub-intervals in the range. The order of testing is the same as previously described. Once such an interval is found it is recorded, and the algorithm recurs to the left and to the right of the afore mentioned interval. The coarse first stage search allows for efficient inspection of the data in the event of no change points being present, whereas the exhaustive second stage search guarantees the intervals ultimately returned by the algorithm are the narrowest possible.

For future reference, we stress the difference between the Narrowest Significance Pursuit algorithm, which can be used in conjunction with any collection of change point tests having bounded family-wise error, and the NSP procedure which refers explicitly to Algorithm 2 used in conjunction with the local tests proposed in Fryzlewicz (2023).

2.2.4 Bayesian inference

Finally, Bayesian approaches to change point detection provide an alternative approach to uncertainty quantification, via credible intervals derived from the posterior distribution of the change point locations recovered. Bayesian approaches to the task of change point detection include Cappello and Padilla (2022) who study data sequences with piecewise constant variance, Cappello et al. (2023) who study the canonical change point problem, Liu et al. (2017) who study the piecewise polynomial change point problem, and Hahn et al. (2020) who study changes in means of high dimensional data. However, choosing sensible priors and sampling from the posterior remain non-trial and in the case of the first problem highly controversial tasks. Posterior distributions can be approximated via MCMC (Chib,

Algorithm 2: The generic narrowest significance pursuit algorithm. Inputting start and end points $\{s, e\}$ and a collection of local tests with family-wise error bounded by some $\alpha \in (0, 1)$ the algorithm returns the narrowest disjoint collection of sub intervals $\{s, \ldots, e\}$ such that with probability at least $1 - \alpha$ each interval returned contains a change point location.

```
function NSP($Y, s, e$):
    if $e - s < 1$ then
        STOP
    end
    $\mathcal{G}_1 \leftarrow$ subIntervalsGrid($s$,$e$)
    for $(t_1, t_2)$ in $\mathcal{G}_1$ do
        if $T_{t_1:t_2}(\boldsymbol{Y}) = 1$ then
            $\mathcal{G}_2 \leftarrow$ allSubIntervals($t_1, t_2$)
            for $(u_1, u_2)$ in $\mathcal{G}_2$ do
                if $T_{u_1,u_2}(\boldsymbol{Y}) = 1$ then
                    recordIntervals($u_1, u_2$)
                    NSP($Y, s, u_1$)
                    NSP($Y, u_2, e$)
                    BREAK
                end
            end
        end
    end
return
```

1998), however this can be computationally demanding. Modern computationally efficient methods for evaluating the posterior have been studied by (Rigaill et al., 2012; Fearnhead, 2006; Nam et al., 2012).

2.3 Dynamic causal structure discovery

2.3.1 Problem statement and motivation

We review selected articles from the causality literature relating specifically to dynamic causal structure discovery. That is, a collection of random variables is observed over time, and the interest lies in understanding whether the behavior of some variables is causing the behavior of others. We will be particularly interested in notions of association which are testable from data. The ideas presented here will become relevant in Chapter 5, where we study a multivariate time series model in which the change point locations are random, and seek to understand whether changes in one time series actively cause change points in another series.

Pearl (2009) draws an important distinction between statistical parameters, which are any parameters which can be estimated in terms of a joint probability distribution over the variables observed (for example expectations), and causal parameters which are parameters from a causal model in which each variable is written as a function of the variables which cause it. In the following sections we introduce two notions of statistical association, which however can be interpreted causally if additional assumptions about the data are made.

2.3.2 Granger causality: causality for time series processes

In his seminal papers Granger (1969, 1980) introduced the concept of Granger causality as a testable notion of causal feedback in econometric models. Let $\mathbb{Y}_t = (Y_{1,t}, \ldots, Y_{d,t})'$ be a multivariate time series process and put $\mathcal{F}_t = \sigma(\mathbb{Y}_s \mid s \leq t)$ for the natural filtration generated by the $\mathbb{Y}$'s up to time t. Without loss of generality we assume $\mathbb{Y}$ is a centered process. Moreover put $\mathbb{Y}_{t\setminus j} = (Y_{i,t} \mid i \neq j)'$ and $\mathcal{F}_{t\setminus j} = \sigma\left(\mathbb{Y}_{s\setminus j} \mid s \leq t\right)$. Then, Y_j is said to

Granger cause Y_k if

$$\mathbb{E}\left[\left(Y_{t,k} - \mathbb{E}\left(Y_{t,k} \mid \mathcal{F}_{(t-1)}\right)\right)^2\right] < \mathbb{E}\left[\left(Y_{t,k} - \mathbb{E}\left(Y_{t,k} \mid \mathcal{F}_{(t-1)\setminus j}\right)\right)^2\right],$$

in which case we write $j \to k$. Else, Y_j does not Granger cause Y_k and we write $j \not\to k$. Intuitively, Y_j does not Granger cause Y_k if knowledge of Y_j's past is not useful for predicting Y_k's future. The notion of Granger causality be extended to collections of time series. Let A, B, and C be disjoint subsets of $\{1, \dots, d\}$ and put $\mathbb{Y}_{t,A} = (Y_{i,t} \mid i \in A)'$ and $\mathcal{F}_{t,A} = \sigma\left(\mathbb{Y}_{s,A} \mid s \leq t\right)$. Then $\mathbb{Y}_A$ is said to Granger cause $\mathbb{Y}_B$ given $\mathbb{Y}_C$ if

$$\mathbb{E}\left[\left(\mathbb{Y}_{t,B} - \mathbb{E}\left(\mathbb{Y}_{t,B} \mid \mathcal{F}_{(t-1),A\cup B\cup C}\right)\right)^2\right] < \mathbb{E}\left[\left(\mathbb{Y}_{t,B} - \mathbb{E}\left(\mathbb{Y}_{t,B} \mid \mathcal{F}_{(t-1),B\cup C}\right)\right)^2\right], \qquad (2.9)$$

in which case we write $A \to B \mid C$. Else, $\mathbb{Y}_A$ does not Granger case $\mathbb{Y}_B$ given $\mathbb{Y}_C$ and we write $A \not\to B \mid C$. For concreteness we give an example below.

Example 2.3.1. *Let $\mathbb{Y}_t$ be generated according to the Vector Auto-Regression*

$$\mathbb{Y}_t = \boldsymbol{A}\mathbb{Y}_{t-1} + \mathcal{E}_t,$$

where $\mathcal{E}$ is a white noise process with variance-covaraince matrix Σ and $\boldsymbol{A}$ is a $d \times d$ matrix. Then Y_j Granger causes Y_k if and only if $\boldsymbol{A}_{k,j} \neq 0$.

Granger causality is asymmetric, since $A \to B \mid C$ does not necessarily imply $B \to A \mid C$. Moreover, it is quite different from the concept of conditional independence, since $A \not\to B \mid C$ does not imply $\mathbb{Y}_B \perp\!\!\!\perp \mathbb{Y}_A \mid \mathbb{Y}_C$. A particular useful concept is the Granger causal graph or network, $\mathcal{G}$, with vertex set $V = \{1, \dots, j\}$ and edge set $E \in \{0,1\}^{d\times d}$ given by

$$E_{j,k} = \begin{cases} 1 & \text{if } j \to k \\ 0 & \text{if } j \not\to k \end{cases}.$$

The presence of an edge between two vertices indicates a potential causal link and the absence of an edge indicates non-causality.

Besides Example 2.3.1, we briefly mention that the notion of Granger causality has been applied to several interesting non-standard time series processes including: categorical time series Tank et al. (2021), non-linear time series (Henderson and Michailidis, 2014; Wu et al., 2014), and mixed-frequency time series (Danks and Plis, 2013; Schorfheide and Song, 2015).

2.3.3 Local independence: causality for marked point processes

Next we review the concept of local independence, which provides a testable notion of dynamic dependence between arrival times of point processes. Before introducing local independence we briefly review point processes on the positive real line, and introduce some concepts which will be of use in Chapter 5.

Point processes on the real line

A point process on the positive real line admits three equivalent definitions (Daley and Vere-Jones, 2003, 2008). It can be defined in terms of:

1. A sequence $\Theta = \{\eta_k \mid k \geq 1\}$ of strictly positive ordered random variables denoting arrival times, which satisfy: (i) $\mathbb{P}\left(0 < \eta_1 \leq \eta_2 \leq \dots\right) = 1$, (ii) $\mathbb{P}\left(\eta_k < \eta_{k+1}, \eta_k < \infty\right) = \mathbb{P}\left(\eta_k < \infty\right)$ for all $k \geq 1$, and (iii) $\mathbb{P}\left(\lim_{k\to\infty} \eta_k = \infty\right) = 1$.

2. A sequence of strictly positive random variables $\{\delta_k \mid k \geq 1\}$ with $\delta_1 = \eta_1$ and $\delta_k = \eta_k - \eta_{k-1}$ for $k > 1$ denoting the inter-arrival times of the η's.

3. A random counting measure $N\left(\cdot\right)$ on the real line, which counts the number of η's in an interval. That is, for any $(a, b] \subset \mathbb{R}_+$ we have $N(a, b] = \sum_{k=1}^{\infty} \mathbf{1}_{\{a < \eta_k \leq b\}}$. For convenience we will occasionally write $N\left(b\right)$ for $N(0, b]$.

A point process is called simple if the number of points in a bounded region is almost surely finite, it is called stationary if the distribution of $N(a, b]$ depends only on the length $b - a$ for any $0 \leq a < b < \infty$, and it is called orderly (Khinchin et al., 1995) if $\mathbb{P}\left(N(0, q] > 1\right) = o(q)$ as $q \downarrow 0$. A highly useful tool when working with point processes is the conditional intensity function. Let $\mathcal{F}_{t-}$ be the natural filtration generated by the η's up to but not including time t. Then, the conditional intensity function is given by

$$\lambda_j^* (t) = \lim_{h \downarrow 0} \frac{\mathbb{E}\left(N[t, t+h) \mid \mathcal{F}_{t-}\right)}{h}. \tag{2.10}$$

The conditional intensity function exists under fairly general conditions, and when it exists it uniquely determines the probability structure of the point process (Daley and Vere-Jones 2003, Proposition 7.2.IV).

For modeling multiple event types occurring on the positive real line, we introduce the multi-type point process $\boldsymbol{N}(\cdot) = (N_1(\cdot), \ldots, N_d(\cdot))'$, associated with which is the sequence of arrival times $\{\eta_{j,k} \mid j = 1, \ldots, d, k \geq 1\}$. Therefore, each component measure $N_j(\cdot)$ counts the number of arrivals of the associated η_j's. The conditional intensities for the component measures are defined as in (2.10), except we condition on the natural filtration of all $\eta_{j,k}$'s up to but not including time t. Multi-type point processes can be equivalently defined as a marked point process (Jacobsen and Gani, 2006), where we have the double sequence $\{(j_k, \eta_k) \mid k \geq 1\}$ taking values in the space $\{1, \ldots, d\} \times \mathbb{R}_+$. Here the η_k's mark the time of the k-th event, and j_k's marks the type of event.

Local independence for marked point processes

We are now in a position to introduce the concept of local independence for marked point processes. The definition we will use is due to Didelez (2008), however as pointed out by Didelez a similar idea was proposed by Schweder (1970) for Markov processes.

Let $\mathcal{F}_t$ be the natural filtration generated by all of the η's up to time t and let $\mathcal{F}_{t \setminus j}$ be the filtration generated by η_k's for $k \in \{1, \ldots, d\} \setminus \{j\}$. The random measure $N_k(\cdot)$ is said to be locally independent of $N_j(\cdot)$ if the conditional intensity function $\lambda_k^*(t)$ is $\mathcal{F}_{t \setminus j}$ measurable for all $t > 0$. In which case we write $j \not\to k$. Else we speak of local dependence, and write $j \to k$. The definition extends naturally to collections of random measures: let A, B, and C be disjoint subsets of $\{1, \ldots, d\}$, put $\boldsymbol{N}_A(\cdot) = (N_j(\cdot) \mid j \in A)'$ and let $\mathcal{F}_{t,A}$ be the filtration generated by the corresponding η's. Then $\boldsymbol{N}_B(\cdot)$ is said to be locally independent of $\boldsymbol{N}_A(\cdot)$ given $\boldsymbol{N}_C(\cdot)$ if $\lambda_j^*(t)$ is $\mathcal{F}_{t,B\cup C}$ measurable for all $j \in B$ and all $t > 0$, in which case we write $A \not\rightarrow B \mid C$. Else we speak of conditional local dependence, and write $A \rightarrow B \mid C$. For concreteness, we give an example below.

Example 2.3.2. *Let point process $\boldsymbol{N}(\cdot) = (N_1(\cdot), \dots, N_d(\cdot))'$ be generated according to a multivariate Hawkes process (Hawkes, 1971) with conditional intensities*

$$\lambda_j^*(t) = \nu_j + \sum_{i=1}^{d} \int_0^{\infty} g_{j,i}(t-u)\,\mathrm{d}N_i(u), \qquad j = 1, \dots, d.$$

Where $\nu_j > 0$ and $g_{j,i} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ for each $i, j \in \{1, \dots, d\}$. Then $N_k(\cdot)$ is locally independent of $N_j(\cdot)$ if and only if $\int_0^{\infty} g_{k,j}(u)\,\mathrm{d}u = 0$. We stress that if $\boldsymbol{N}(\cdot)$ is fully observed then $\int_0^{\infty} g_{k,j}(u)\,\mathrm{d}$ is a causal parameter in the sense of Pearl (2009), since if the quantity is non-zero events in the past of $N_j(\cdot)$ are responsible for future changes in the intensity of $N_k(\cdot)$.

The concept of local independence is closely linked to Granger causality. Observe that under general conditions (Medvegyev, 2007) a point process on the real line will admit Doob-Meyer decomposition as: $N(t) = \int_0^t \lambda^*(u)\,\mathrm{d}u + M(t)$, where $M(t)$ is an $\mathcal{F}_t$ martingale. In light of this decomposition, the statement $A \not\rightarrow B \mid C$ can again be understood in terms the past of $\boldsymbol{N}_A(\cdot)$ being useful for predicting the future of $\boldsymbol{N}_B(\cdot)$ once the past of $\boldsymbol{N}_C(\cdot)$ is observed.

Similarly to Granger causality, local independence is asymmetric, since $A \rightarrow B \mid C$ does not necessarily imply $B \rightarrow A \mid C$, and different from conditional independence, since $A \not\rightarrow B \mid C$ does not imply $\boldsymbol{N}_B(\cdot) \perp\!\!\!\perp \boldsymbol{N}_A(\cdot) \mid \boldsymbol{N}_C(\cdot)$. Didelez (2008) also proposed the concept of a local independence graph, in which the absence of an edge indicates local independence between two component processes of a marked point process, and the presence of an edge may be indicative of causality. We briefly mention that specifically for Hawkes processes as briefly defined in Example 2.3.2, such a graph was also introduced by Embrechts and Kirchner (2016) under the name Hawkes skeleton.

Methods for recovering local independence graphs

There are several approaches in the literature for recovering local independence graphs from data. For the case of Markov processes, procedures based on maximum likelihood estimation have been proposed (Didelez, 2007, 2001). Bayesian methods have been proposed by Nodelman et al. (2012). Most notably, when the point processes are Hawkes processes as in Example 2.3.2 very many estimating procedures exist, and we mention just a few (Achab et al., 2018; Chen et al., 2017; Hansen et al., 2015). Finally we mention the work of Thams and Hansen (2023). The authors point out that Hawkes processes are not closed under marginalization, in the sense that if one of the component processes is not observed the remaining processes processes will necessarily not be Hawkes processes. Instead, the authors test for the presence of an edge in the graph using basis expansions of the (potentially) marginalized intensities.

3 Fast and Optimal Inference for Change Points in Piecewise Polynomials via Differencing

3.1 Introduction and problem statement

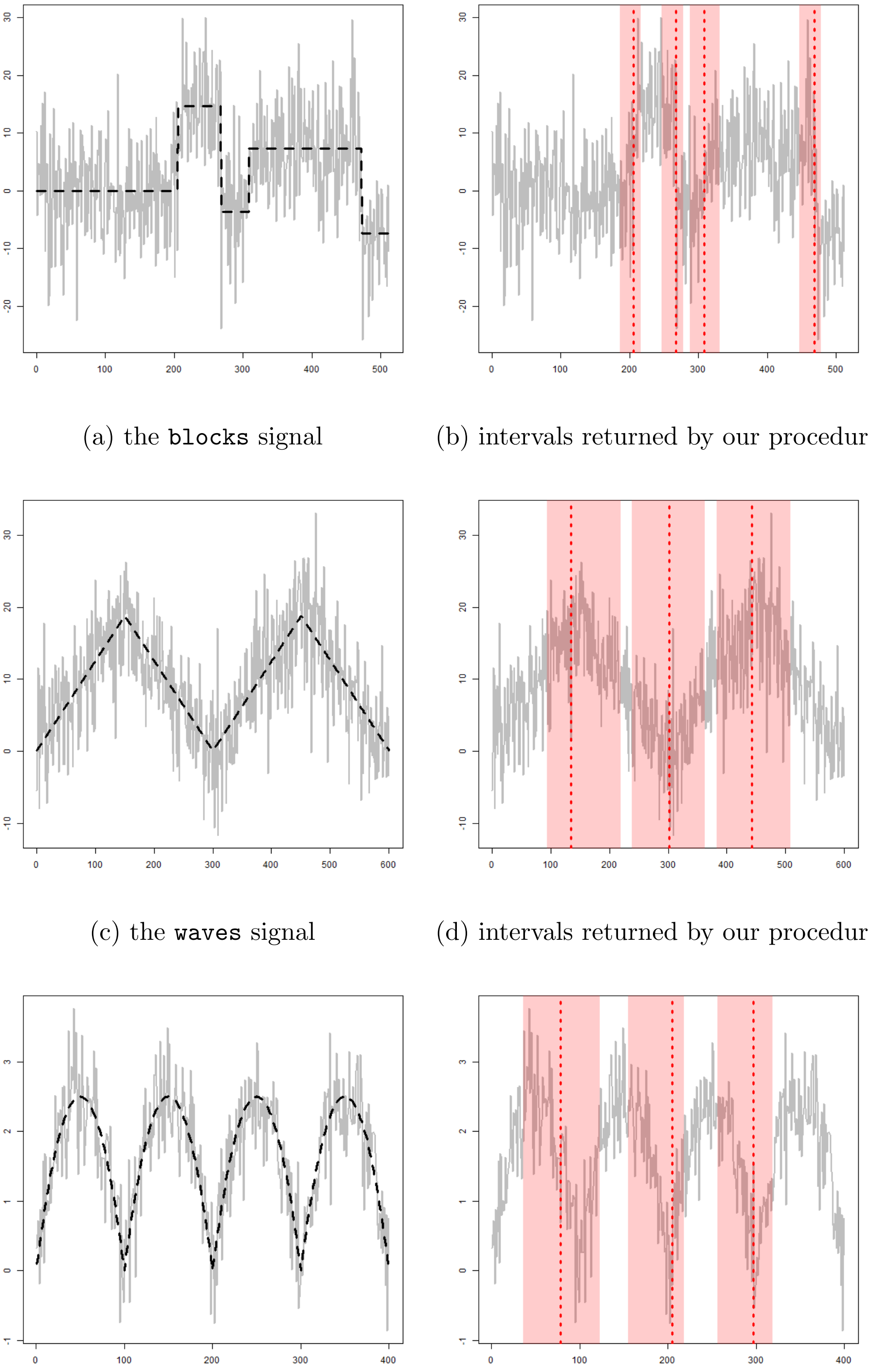
In this chapter we study the setting in which an analyst observes data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ on an equi-spaced grid which can be written as the sum of a signal component and a noise component:

$$Y_t = f_\circ\,(t/n) + \zeta_t \qquad t = 1, \ldots, n. \tag{3.1}$$

The signal component $f_\circ : [0, 1] \mapsto \mathbb{R}$ is known to be a piecewise polynomial function of arbitrary but fixed degree p, and associated with $f_\circ\,(\cdot)$ there are N integer-valued change points at locations $\Theta = \{\eta_1, \ldots, \eta_N\}$. Both Θ and N are unknown. Our goal is to simultaneously quantify the level of uncertainty the around the existence and location of each putative change point in the generic piecewise polynomial model. This is a worthwhile task since estimates of the change point locations are not consistent in the usual sense of the estimation error tending to zero with the sample size. Moreover, since most algorithms for change point estimation do not quantify the uncertainty around the change points they recover, it is difficult to say whether these change points are real or spuriously estimated.

We propose a procedure which aims to return the narrowest possible disjoint sub-intervals of the index set $\{1, \ldots, n\}$ in such a way that each must contain a change point location uniformly at some confidence level chosen by the user. Examples of such intervals are

Figure 3.1: the piecewise constant `blocks` signal, piecewise linear `waves` signal, and piecewise quadratic `hills` signal each contaminated with i.i.d. Gaussian noise (left column). Intervals of significance with uniform 90% coverage returned by our procedure (right column). Black dashed lines (- - -) represent underlying piecewise polynomial signal, light grey lines (—) represent the observed data sequence, red shaded regions (🟥) represent intervals of significance returned by our procedure, red dotted lines (· · ·) represent split points within each interval associated with the piecewsie polynomial fit providing the lowest sum of squared residuals.



(a) the `blocks` signal

(b) intervals returned by our procedure

(c) the `waves` signal

(d) intervals returned by our procedure

(e) the `hills` signal

(f) intervals returned by our procedure

shown in the right column of Figure 3.1. This is done by testing for a change at a range of scales and locations belonging to a sparse grid, and tightly bounding the supremum of local test statistics over the same grid which guarantees sharp family-wise error control. We initially study the setting in which the noise components are independent with marginal $\mathcal{N}\left(0, \sigma^2\right)$ distribution and later in Section 3.2.4 extend our results to dependent and non-Gaussian noise. Motivated by the fact that taking $(p+1)$-th differences will eliminate a degree p polynomial trend (Chan et al., 1977), we consider tests based on differences of (standardised) local sums of the data sequence. There are several advantages to working with tests based on local sums as opposed to for example likelihood ratio or Wald statistics, which we list below.

- Each of our local test can be completed in $\mathcal{O}(1)$ time in a straightforward manner, regardless of the degree of the underlying polynomial or the scale at which the test is performed, leading to a procedure with worst case complexity $\mathcal{O}\left(n \log(n)\right)$ when test are carried out on a sparse grid.
- Local averaging brings the contaminating noise closer to Gaussianity, which is a feature we exploit in Section 3.2.4 when studying the behaviour of the our procedure under non-Gaussian and possibly dependent noise.
- Unlike procedures based on differencing the raw data, which are known to be sub-optimal, as we show in Theorem 3.3.1 the combination of local averaging followed by differencing leads to a procedure which is optimal in a mini-max sense.
- The asymptotic analysis is by design uncomplicated, as it boils down to analysing the high excursion probability of a stationary Gaussian field whose local structure depends on the polynomial degree in a straightforward way.

The remainder of the chapter is structured as follows. In Section 3.2 we introduce local tests for the presence of a change based on differences of local sums of the data, and study their behaviour under the null of no change points in terms of the family-wise error when the test are applied over a sparse grid. In Section 3.3 we introduce a fast algorithm for turning our local tests into a collection of disjoint intervals which each must contain a change at a prescribed significance level, and show the algorithm's consistency and optimality in terms of recovering narrow intervals which each contain a change point location. In Section 3.4 we compare the performance of our algorithm with that of existing procedures when applied to simulated data. Finally in Section 3.5 we show the practical use of our algorithm via two real data examples.

3.2 Difference based tests with family-wise error control

3.2.1 Local tests for a change point

We begin by describing tests for the presence of a change on a localised segment of the data. Motivated by the fact that a polynomial trend will be eliminated by differencing, if it were suspected that a segment of the data contained a change point location one could divide the segment into $p+2$ chunks of roughly equal size and take the $(p+1)$-th difference of the sequence of local sums on each chunk. Since summing boosts the signal from the change point, and differencing eliminates the polynomial trend, one could then declare a change if the resulting quantity coming from the summed and differenced sequence, appropriately scaled, was large in absolute value. By contrast, simply differencing the data on the segment would reduce the signal from the change, and any statistic based on the differenced data only would be sub-optimal for detecting the change.

For each local test we write l for the location of the data segment being inspected for a change point and w for the width of the data segment. Following the reasoning above, to test for the presence of a change point on the interval $\{l, \dots, l+w-1\}$ we first compute the following non-overlapping local sums:

$$\bar{Y}_{l,w}^{j} = Y_{l+j\left\lfloor \frac{w}{p+2} \right\rfloor} + \cdots + Y_{l+(j+1)\left\lfloor \frac{w}{p+2} \right\rfloor - 1}, \qquad j = 0, \dots, p+1.$$

We then declare a change if the test statistic defined below, which corresponds to the the $(p+1)$-th differences of the sequence $\bar{Y}_{l,w}^{0}, \dots, \bar{Y}_{l,w}^{p+1}$ scaled so that its variance is constant independent of l and w when the noise is homoskedastic and independently distributed, is large in absolute value.

$$D_{l,w}^{p}\left(\boldsymbol{Y}\right) = \left\{ \left\lfloor \frac{w}{p+2} \right\rfloor \sum_{i=0}^{p+1} \binom{p+1}{i}^{2} \right\}^{-1/2} \sum_{j=0}^{p+1} (-1)^{p+1-j} \binom{p+1}{j} \bar{Y}_{l,w}^{j} \tag{3.2}$$

The functional introduced in (3.2) enjoys the following properties, which make it well suited for the task of change change point testing on piecewise polynomials:

- Additivity: for any two vectors $\boldsymbol{f}, \boldsymbol{g} \in \mathbb{R}^{n}$ it holds that $D_{l,w}^{p}\left(\boldsymbol{f}+\boldsymbol{g}\right) = D_{l,w}^{p}\left(\boldsymbol{f}\right) + D_{l,w}^{p}\left(\boldsymbol{g}\right)$ for all admissible l's and w's.
- Annihilation of polynomials: if the entries of $\boldsymbol{f} \in \mathbb{R}^{n}$ are from a polynomial of degree no larger than p then $D_{l,w}^{p}\left(\boldsymbol{f}\right) = 0$ for all admissible l's and w's.
- Large for discontinuous functions: if the entries of $\boldsymbol{f} \in \mathbb{R}^{n}$ are from a piecewise monomial with a single discontinuity at location η then $|D_{l,w}^{p}\left(\boldsymbol{f}\right)| > 0$ for all l's and w's such that $\eta \in \{l, \ldots, l+w-1\}$.

The first two properties ensure (3.2) is small under the local null of no change, whereas the third property can be used to show that for some admissible (l, w) pair the the statistic will be large in absolute value in the presence of a change.

Consequently, for some $\lambda > 0$ to be chosen later on, each local test for the presence of a change on an interval $\{l, \ldots, l+w-1\}$ takes the following form:

$$T_{l,w}^{\lambda}\left(\mathbf{Y}\right) = \mathbf{1}\left\{ |D_{l,w}^{p}\left(\boldsymbol{Y}\right)| > \lambda \right\}. \tag{3.3}$$

When $p = 0$ the statistic (3.2) recovers the moving sum filter used for change point detection in the piecewise constant model Eichinger and Kirch (2018). This also corresponds to the (square root of) the likelihood ratio statistic for testing the null of a constant mean on the segment under Gaussian noise, as well as the Wald statistic for the same problem. Typical approaches for generalising to higher order polynomial change point problems involve local likelihood-ratio or Wald statistics for testing the null of a polynomial mean on the segment

(Baranowski et al., 2019a; Fang et al., 2020; Anastasiou and Fryzlewicz, 2022; Kim et al., 2022), which however are hard to stochastically control. We show that simply extending the order of differencing leads to simple and powerful tests.

3.2.2 Local tests on a sparse gird

For the purpose of making inference statements about an unknown number of change point locations, we would like to apply the local tests (3.3) over a grid which is both dense enough to cover all potential change point locations well and sparse enough to allow all local tests to be computed quickly. Given a suitable grid $\mathcal{G}$ of (l, w) pairs, if λ were chosen to control the family-wise error of the collection of tests

$$\mathcal{T}_{\mathcal{G}}^{\lambda}(\mathbf{Y}) = \left\{ T_{l,w}^{\lambda}(\mathbf{Y}) \mid (l, w) \in \mathcal{G} \right\} \tag{3.4}$$

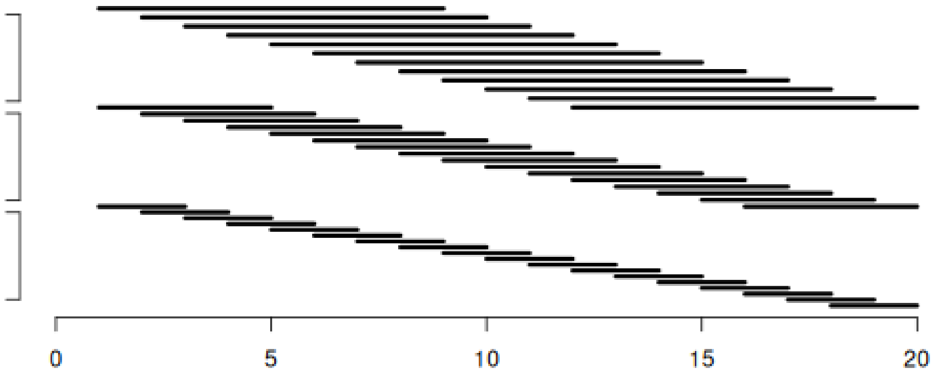
at some level α, we could be sure that with probability $1 - \alpha$ every (l, w) pair on which a test rejects corresponds to a segment of the data containing at least one change point location.

We propose to use the following grid, which is parameterised by a minimum grid scale parameter W, controlling the minimum support of the detection statistic (3.2), and a decay parameter $a > 1$, controlling the density of the grid:

$$\begin{aligned} \mathcal{G}(W, a) &= \left\{ (l, w) \in \mathbb{N}^2 \mid w \in \mathcal{W}(W, a), 1 \leq l \leq n - w \right\} \\ \mathcal{W}(W, a) &= \left\{ w = \left\lfloor a^k \right\rfloor \mid \lfloor \log_a(W) \rfloor \leq k \leq \lfloor \log_a(n/2) \rfloor \right\}. \end{aligned} \tag{3.5}$$

Associated with the grid is the collection of sub-intervals of $\{1, \dots, n\}$ whose length is larger than W and can be written as an integer power of a. For example, the collection of intervals $\{l, \dots, l + w - 1\}$ associated with the (l, w) pairs in the grid obtained when $n = 20$ and setting $W = 2$ and $a = 2$ is shown in Figure 3.2 below. For this configuration of a and W, the associated collection of intervals consists of all contiguous sub-interval of $\{1, \dots, 20\}$ having dyadic length.

Figure 3.2: Intervals associated with the $\mathcal{G}(W, a)$ with minimum grid scale $W = 2$ and decay parameter $a = 2$, for a sample size of $n = 20$.



The grid defined by (3.5) is similar to several grids already proposed for different change point detection problems (Kovács et al., 2023; Chan and Walther, 2013; Pilliat et al., 2023), in that the size of scales decays exponentially. Pilliat et al. mention that “from a purely statistical perspective, it is difficult to appreciate the respective benefits of denser or sparser grids” and this motivated us to study a procedure which is in fact adaptive to the chosen grid through the decay parameter a. Two key differences between our grids and previously proposed grids are that: (i) for any scale w all possible locations l are considered, and (ii) that all scales with $w = o(W)$ are excluded from the grid. Regarding the minimum grid scale, if the noise were known to be independently distributed and Gaussian we could take $W = \mathcal{O}(1)$ and still retain family-wise error control using our proof technique. However, under dependent or non-Gaussian noise letting the minimum grid scale diverge at an appropriate rate with n is necessary for controlling the family-wise error, as this allows local sums of the noise to be treated as approximately uncorrelated and Gaussian.

3.2.3 Family-wise error control under Gaussianity

As a starting point for family-wise error analysis in more general noise settings, we first show how to control the family-wise error of the local tests (3.3) over the grid (3.5) when the noise terms are independently distributed and Gaussian. The approach is to tightly bound the maximum of the local test statistics (3.2) under the null of no change points, and use this bound to select an appropriate threshold λ for (3.4). We impose the following assumptions on the minimum grid scale and on the noise components.

Assumption 3.2.1. *The noise terms $\zeta_1, \ldots, \zeta_n$ are mutually independent with marginal $\mathcal{N}(0, \sigma^2)$ distribution for some $\sigma > 0$.*

Assumption 3.2.2. *The minimum grid scale W satisfies $W/\log(n) \to d$ for some $d \in (0, \infty)$.*

With these assumption in place we have the following result on the behaviour of the maximum of local test statistics (3.2) under the null of no change points.

Theorem 3.2.1. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be from model (3.1) with signal component having no change points and grant Assumptions 3.2.1 - 3.2.2 hold. For fixed $a > 1$ introduce the following quantity:*

$$M^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) = \max_{(l,w)\in\mathcal{G}(W,a)} \left\{ \frac{1}{\sigma} D^{p}_{l,w}(\boldsymbol{Y}) \right\}.$$

(i) Putting $\mathfrak{a}_n = \sqrt{2\log(n)}$ and $\mathfrak{b}_n = 2\log(n) - \frac{1}{2}\log\log(n) - \log(2\sqrt{\pi})$ the sequence of random variables $\left\{ \mathfrak{a}_n M^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) - \mathfrak{b}_n \mid n \in \mathbb{N} \right\}$ is tight, and there are constants $H_{1,1}$ and $H_{1,2}$ depending only on a, p, and d such that for fixed $x \in \mathbb{R}$ the following holds

$$o(1) + \exp\left(-H_{1,1}e^{-x}\right) \leq \mathbb{P}\left(\mathfrak{a}_n M^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) - \mathfrak{b}_n \leq x \right) \leq \exp\left(-H_{1,2}e^{-x}\right) + o(1).$$

(ii) Moreover the result in (i) continues to hold if σ is replaced with any consistent estimator $\widehat{\sigma}$ which satisfies $|\widehat{\sigma}/\sigma - 1| = o_{\mathbb{P}}\left(\log^{-1}(n)\right)$.

Note that for large values of n the quantity

$$L^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) = \max_{(l,w)\in\mathcal{G}(W,a)} \left\{ \frac{1}{\sigma} \left| D^{p}_{l,w}(\boldsymbol{Y}) \right| \right\}$$

behaves asymptotically like the maximum of two independent copies of $M^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y})$. We do not give a formal proof of this statement, however the statement can be understood intuitively by writing $L^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) = M^{\sigma}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) \vee M^{\sigma}_{\mathcal{G}(W,a)}(-\boldsymbol{Y})$ and then using the well known fact that order statistics are asymptotically independent (Falk and Reiss, 1988; Kabluchko and Wang, 2014). Therefore, in light of Theorem 3.2.1 it follows that under Assumptions 3.2.1 - 3.2.2, for any $\alpha \in (0,1)$, choosing $\lambda = \widehat{\sigma}\lambda_{\alpha}$ with $\widehat{\sigma}$ satisfying the condition given in part (ii) and λ_{α} defined as follows

$$\lambda_{\alpha} = \sqrt{2\log(n)} + \frac{-\frac{1}{2}\log\log(n) - \log\left(2\sqrt{\pi}/H_{1,2}\right) + \log\left(-2\log^{-1}(1-\alpha)\right)}{\sqrt{2\log(n)}} \tag{3.6}$$

will result in the collection of tests $\mathcal{T}^{\lambda}_{\mathcal{G}(W,a)}(\mathbf{Y})$ having family-wise error asymptotically no larger than α. In Section 3.3.2 we given an example of an estimator which satisfies condition (ii) in Theorem 3.2.1 above, even if the data contains change points, provided the number of change points does not grow too quickly with the n.

Importantly, the threshold (3.6) explicitly accounts for the grid used, in the sense that if one chooses a coarser gird a lower price is paid for multiple testing. More specifically, if one chooses a coarser grid by increasing a the constant $H_{1,2}$ adjusts which reduces the size of (3.6). As a result, each local test performed will have higher power with the same family-wise error guarantee. Naturally, on a coarser grid the collection of tests may overall have lower power for detecting all change point locations, since fewer tests are carried out in total.

The constants $H_{1,1}$ and $H_{1,2}$ are defined explicitly below, where we put $b_1 = 1/a$ and

$b_2 = 1$, and $\bar{\Phi}(\cdot)$ for the tail function of a standard Gaussian random variable.

$$\begin{aligned} H_{1,i} &= \sum_{j=0}^{\infty} p_{\infty}^{2}\left(\frac{2C_p}{a^j b_i d}\right) \qquad i = 1,2 \qquad (3.7) \\ p_{\infty}(x) &= \exp\left(-\sum_{k=1}^{\infty}\frac{1}{k}\bar{\Phi}\left(\sqrt{kx/4}\right)\right) \\ C_p &= (p+2)\left(1+\sum_{j=1}^{p+1}\binom{p+1}{j}\binom{p+1}{j-1}\Big/\sum_{i=0}^{p+1}\binom{p+1}{i}^2\right) \end{aligned}$$

The effect of the decay parameter a on $H_{1,1}$ and $H_{1,2}$ can now be understood via (3.7) using the additional fact that (Kabluchko, 2007, Corollary 3.18) for any $C > 0$ the quantity $p_{\infty}^{2}(C/x)$ behaves like $C/(2x)$ when x is large.

We now explain the origin of the double inequality in Theorem 3.2.1, and why it is sufficient for strong family-wise error control. In Theorem 3.2.1 we are only able to establish tightness of the normalised maximum, as opposed to convergence to an extreme value distribution, for the following reason: the maximum over standardised increments of a sequence of Gaussian variables will be achieved on scales of the order $\mathcal{O}(\log(n))$ as was shown by Kabluchko (2007) and Kabluchko and Wang (2014), but scales of this order cannot necessarily be expressed as integer powers of a. Consequently the choice of grid introduces small fluctuations in the maximum, which persist in the limit, and correspond to the difference between $\log(n)$ and the closest integer power of a. However for a sub-sequence of n's on which the quantity $b_n = a^{\lfloor \log_a(W) \rfloor}/W$ converges the normalised maximum does converge. The constants $H_{1,1}$ and $H_{1,2}$ therefore correspond to the largest and smallest constants which may appear in the extreme value limit on a sub-sequence of n's on which b_n converges to some constant.

3.2.4 Extension to dependent and non-Gaussian noise

We now extend the result of Theorem 3.2.1 to dependent and non-Gaussian noise. This is done through the standard approach (Hušková and Slabý, 2001; Kirch and Klein, 2023; Eichinger and Kirch, 2018) of computing local tests only on scales large enough such that partial sums of the data can be replaced by increments of a Wiener process without affecting the asymptotics. Therefore, we impose the following assumptions on the minimum grid scale and the noise component.

Assumption 3.2.3. *The noise terms are mean zero and weakly stationary, with autocovariance function* $\gamma_h = Cov(\zeta_0, \zeta_h)$ *and strictly positive long run variance* $\tau^2 = \gamma_0 + 2\sum_{h>0}\gamma_h$.

Assumption 3.2.4. *There exists a Wiener process* $\{B(t)\}_{t>0}$ *such that for some* $\nu > 0$, *possibly after enlarging the probability space, it holds* $\mathbb{P}$*-almost surely that* $\sum_{t=1}^{n}\zeta_t - \tau B(n) = \mathcal{O}\left(n^{\frac{1}{2+\nu}}\right)$.

Assumption 3.2.5. *With the same* ν *as in Assumption 3.2.4, the minimum grid scale* W *satisfies* $n/W \to \infty$ *and* $n^{\frac{2}{2+\nu}}\log(n)/W \to 0$.

Assumption 3.2.4 holds under a wide range of common dependence conditions such as β-mixing, functional dependence, and covaraince decay (Berkes et al., 2014; Philipp et al., 1975; Kuelbs and Philipp, 1980); these dependence conditions in turn hold for a range of popular time series models such as ARMA, GARCH, and bilinear models (Doukhan, 2012; Wu, 2005). If the noise terms are independently distributed Assumption 3.2.4 holds as long as their $(2+\nu)$-th moment is bounded (Komlós et al., 1975; Csörgo and Révész, 2014). With these assumption in place we have the following result on the behaviour of the maximum of local test statistics (3.2) under the null of no change points.

Theorem 3.2.2. *Let* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *be from model (3.1) with signal component having no change points and grant Assumptions 3.2.3 - 3.2.5 hold. For fixed* $a > 1$ *introduce the following quantity:*

$$M^{\tau}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) = \max_{(l,w)\in\mathcal{G}(W,a)} \left\{ \frac{1}{\tau} D^{p}_{l,w}(\boldsymbol{Y}) \right\}.$$

(i) Putting $\mathfrak{a}_{n,W} = \sqrt{2\log(n/W)}$ *and* $\mathfrak{b}_{n,W} = 2\log(n/W) + \frac{1}{2}\log\log(n/W) - \log(\sqrt{\pi})$ *the sequence of random variables* $\left\{\mathfrak{a}_{n,W} M^{\tau}_{\mathcal{G}(W,a)}(\boldsymbol{Y}) - \mathfrak{b}_{n,W} \mid n \in \mathbb{N}\right\}$ *is tight, and there*

are constants $H_{2,1}$ and $H_{2,2}$ depending only on a and p such that for fixed $x \in \mathbb{R}$ the following holds

$$o(1) + \exp\left(-H_{2,1}e^{-x}\right) \leq \mathbb{P}\left(\mathfrak{a}_{n,W} M^{\tau}_{\mathcal{G}(W,a)}\left(\boldsymbol{Y}\right) - \mathfrak{b}_{n,W} \leq x\right) \leq \exp\left(-H_{2,2}e^{-x}\right) + o(1).$$

(ii) Moreover the result in (i) continues to hold if τ is replaced with any consistent estimator $\widehat{\tau}$ which satisfies $|\widehat{\tau}/\tau - 1| = o_{\mathbb{P}}\left(\log^{-1}(n/W)\right)$.

By the same reasoning used in Section 3.2.3 under assumptions 3.2.3 - 3.2.5 Theorem 3.2.2 guarantees that choosing $\lambda = \widehat{\tau}\lambda_{\alpha}$, with $\widehat{\tau}$ satisfying the condition given in part (ii), and with λ_{α} defined as follows

$$\lambda_{\alpha} = \sqrt{2\log(n/W)} + \frac{\frac{1}{2}\log\log(n/W) - \log\left(\sqrt{\pi}/H_{2,2}\right) + \log\left(-2\log^{-1}\left(1-\alpha\right)\right)}{\sqrt{2\log(n/W)}} \tag{3.8}$$

will result in the collection of tests $\mathcal{T}^{\lambda}_{\mathcal{G}(W,a)}\left(\mathbf{Y}\right)$ having family-wise error asymptotically no larger than α. In Section 3.3.2 we give examples of variance and long run variance estimators which satisfy condition (ii) in Theorem 3.2.2, even in the presence of change points, provided the number of change points does not grow too quickly with n.

By the same mechanism as in Theorem 3.2.1 the threshold (3.8) is adaptive to the chosen grid. The constants $H_{2,1}$ and $H_{2,2}$ in Theorem 3.2.2 are as shown below, where C_p and b_i are as in Section 3.2.3:

$$H_{2,i} = \frac{b_i^{-1}C_p}{1-a^{-1}} \qquad i = 1, 2.$$

The proofs of Theorems 3.2.1 and 3.2.2 reveal that maxima achieved over different scales in the grid (3.5) will be asymptotically independent. This combined with the tightness of the normalised maximum shows that the thresholds (3.6) and (3.8) are the sharpest possible for each scale in the grid, under their respective sets of assumptions. That is, if one were to restrict tests to a single scale of the order $\mathcal{O}\left(W\right)$ the threshold needed to control the family-wise error of the collection of tests would be asymptotically equivalent to the thresholds presented for controlling the family-wise error of test conducted on the whole grid.

3.3 A fast algorithm for change point inference

3.3.1 The algorithm

We now present an algorithm, based on the tests introduced in Section 3.2, for efficiently recovering disjoint sub-intervals of the index set $\{1, \ldots n\}$ in such a way that each must contain a change point uniformly at some prescribed significance level α. The algorithm is motivated by the Narrowest Significance Pursuit proposed by Fryzlewicz (2023), in that the focuses is on recovering theses intervals through a series of local tests so that each interval is the narrowest possible. However, there are several important differences between our approach and the approach in Fryzlewicz (2023), which we outline below before presenting the algorithm.

- Each of our local tests can be computed in constant time as a function of the sample size and independently of the scale of the computation. This is not the case for Fryzlewicz (2023), where each local test requires solving a linear program.
- We compute local tests over the sparse grid defined in (3.5), whereas Fryzlewicz (2023) uses a two stage procedure where local tests are initially performed over a coarse grid and intervals flagged in the first stage are exhaustively sub-searched. In the worst case the former leads to $\mathcal{O}\left(n \log(n)\right)$ tests being carried out, whereas the latter may lead to $\mathcal{O}\left(n^2\right)$ test being performed.
- The thresholds used in our local tests are designed to adapt to the chosen grid, which accounts for the statistical-computational trade off in large scale change point problems. However, the threshold used in Fryzlewicz (2023) does not depend on the chosen grid.

Given a grid of (l, w) pairs $\mathcal{G}\left(W, a\right)$ constructed according to (3.5) our approach is to greedily search for a pair on which the associated local test (3.3) declares a change, starting from the finest scale in the grid. When such a pair is found the associated interval

$\{l, \ldots, l+w-1\}$ is recorded and the search is recursively repeat to the left and right of this interval. This approach can be seen as a hybrid between a scanning algorithm where one scans through the data with test statistic having fixed support, and a binary segmentation algorithm.

Pseudo code for the procedure is given below in Algorithm 3. In the pseudo code given integers s and e which satisfy $1 \leq s < e \leq n$ we write $\mathcal{G}_{s,e}(W, a)$ for the set of (l, w) pairs in $\mathcal{G}(W, a)$ which can be associated with an interval satisfying $\{l, \ldots, l+w-1\} \subseteq \{s, \ldots, e\}$. We write λ_α for either of the thresholds (3.6) or (3.8), depending on whether we are operating under Assumptions 3.2.1 - 3.2.2 or Assumptions 3.2.3 - 3.2.5. Finally we write $\widehat{\tau}$ for a generic estimator of the (long run) standard deviation of the noise which satisfies either the of the conditions in of part (ii) of Theorem 3.2.1 or in part (ii) of Theorem 3.2.2, depending on the set of assumptions we are operating under.

Algorithm 3: The greedy interval search algorithm for change point inference in piecewise polynomials. Given an appropriate threshold, the algorithm returns a collection of mutually disjoint intervals which each must contain a change point uniformly with probability at least $1 - \alpha + o(1)$.

```
function greedyIntervalSearch($\boldsymbol{Y}, s, e$):
    if $e - s < \min(W, p+1)$ then
        STOP
    end
    detection ← False
    for $(l, w)$ in $\mathcal{G}_{s,e}(W, a)$ do
        if $\left|D^p_{l,w}(\boldsymbol{Y})\right| > \widehat{\tau}\lambda_\alpha$ then
            RecordInterval($l, w$)
            greedyIntervalSearch($Y, s, l$)
            greedyIntervalSearch($Y, l+w-1, e$)
            detection ← True
        end
        if detection then
            BREAK
        end
    end
return
```

A consequence of using thresholds (3.6) and (3.8) in Algorithm 3 is that with no assumptions on the number of change points in the data or their spacing, with high probability, every interval returned is guaranteed to contain at least one change point. The number of intervals returned therefore functions as an assumption free lower bound on the number of change points in the data. This behaviour is summarised in Corollary 3.3.1 below.

Corollary 3.3.1. *Let $\hat{I}_1, \ldots, \hat{I}_{\hat{N}}$ be intervals returned by Algorithm 3. On a set with probability asymptotically larger than $1 - \alpha$ the following events occur simultaneously:*

$$E_1^* = \left\{ \widehat{N} \leq N \right\},$$
$$E_2^* = \left\{ \hat{I}_k \cap \Theta \neq \emptyset \mid k = 1, \ldots, \hat{N} \right\}.$$

Although the coverage guarantee provided by Corollary 3.3.1 is asymptotic in nature, in practice we find that Algorithm 3 provides accurate coverage in finite samples, and in fact tends to deliver over coverage. The worst case run time of Algorithm 3 is always of the order $\mathcal{O}\left(n \log(n)\right)$, independent of the number of change points in the data, their spacing, and the polynomial degree of the signal. This is because the worst case run time will be attained when a test has to be carried out for every (l, w) pair in the grid $\mathcal{G}\left(W, a\right)$. However, for any fixed $a > 1$ the the grid contains at most of the order $\mathcal{O}\left(n \log(n)\right)$ such pairs, and by first calculating all cumulative sums of the data, which can be done in $\mathcal{O}\left(n\right)$ time, each local test can be carried out in constant time.

We finally remark that many existing procedures for change point detection make use of thresholds which involve unknown constants other than the scale of the noise. In general these constants are either chosen sub-optimally, or calibrated via Monte Carlo. See for instance the implementation of Verzelen et al. (2023) by Liehrmann and Rigaill (2023) for an example in in the piecewise constant setting, and the discussion on the practical selection of tuning parameters in Kim et al. (2022) for an example in the piecewise linear setting. Meanwhile, the thresholds used in Algorithm 3 are the sharpest possible, and do not rely on any unknown constants other than the scale of the noise.

3.3.2 Variance and long run variance estimation

In general the scale of the noise will not be known, and to make Algorithm 3 operational the (long run) standard deviation of the noise will need to be estimated consistently, according to the conditions given in part (ii) of either Theorem 3.2.1 or Theorem 3.2.2. In this section we give several strategies for consistently estimating the noise level in the presence of an unknown piecewise polynomial signal.

Variance estimation under Gaussian noise

In change point problems where the noise is independently distributed, homoskedastic, and Gaussian the standard deviation is commonly estimated using the median absolute deviation (MAD) estimator (Hampel, 1974). To account for the unknown piecewise polynomial signal we propose to use the following generalisation of the MAD estimator based on the $(p+1)$-th difference of the data. Letting $X_{p+2}, \dots, X_n$ be the $(p+1)$-th difference of the sequence $Y_1, \dots, Y_n$ the estimator is defined as follows:

$$\widehat{\sigma}_{\mathrm{MAD}} = \frac{\operatorname{median}\left\{\left|X_{p+2}\right|, \dots, \left|X_n\right|\right\}}{\Phi^{-1}\left(3/4\right)\sqrt{\sum_{j=0}^{p+1}\binom{p+1}{j}^2}}. \tag{3.9}$$

As shown by the following lemma, when the assumptions of Theorem 3.2.1 hold the modified MAD estimator satisfies the condition in part (ii) of the Theorem 3.2.1 as long as the number of change points grows more slowly than $n/\log(n)$.

Lemma 3.3.1. *If the noise terms are independently distributed and Gaussian with common variance σ^2 it holds that*

$$\left|\widehat{\sigma}_{MAD} - \sigma\right| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \vee \frac{N}{n}\right).$$

Variance estimation under non-Gaussian noise

For variance estimation under independently distributed light tailed homoskedastic noise, difference based estimators are often used (Dümbgen and Spokoiny, 2001; Rice, 1984; Gasser et al., 1986). To account for the unknown piecewise polynomial signal we propose to use the following estimator based on the $(p+1)$-th difference of the data sequence. The estimator is defined as follows:

$$\widehat{\sigma}^2_{\mathrm{DIF}} = \frac{1}{n-(p+1)} \sum_{t=p+2}^{n} \left\{ \frac{X_t^2}{\sum_{j=0}^{p+1} \binom{p+1}{j}^2} \right\}. \tag{3.10}$$

As shown by the following lemma, under some mild conditions on signal component the difference based estimator satisfies condition (ii) in Theorem 3.2.2 as long as the number of change points again grows more slowly than $n/\log(n)$.

Lemma 3.3.2. *If the function $f_\circ(\cdot)$ is bounded and the noise terms are independently distributed with common variance σ^2 and bounded fourth moments it holds that*

$$\left|\widehat{\sigma}^2_{DIF} - \sigma^2\right| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \vee \frac{N}{n}\right).$$

Long-run variance estimation

For estimating the long run variance we extend the estimator proposed in Wu and Zhao (2007), based on first order differences of local sums of the data, to $(p+1)$-th differences. To form the estimator we choose a scale W', which is not necessarily related to any of the scales in the grid (3.5), and form the following local sums:

$$\bar{Y}_{t,W'} = Y_{(t-1)W'+1} + \cdots + Y_{tW'}, \qquad t = 1, \ldots, \lfloor n/W' \rfloor \tag{3.11}$$

Then, putting $\bar{X}_{p+2,W'}, \dots, \bar{X}_{\lfloor n/W' \rfloor, W'}$ for the $(p+1)$-th difference of the sequence of $\bar{Y}_{W'}$'s, the estimator is defined as follows:

$$\widehat{\tau}^2_{\mathrm{DIF}} = \frac{1}{\lfloor n/W' \rfloor - (p+1)} \sum_{t=p+2}^{\lfloor n/W' \rfloor} \left\{ \frac{\bar{X}^2_{t,W'}}{W' \sum_{i=0}^{p+1} \binom{p+1}{i}^2} \right\}. \tag{3.12}$$

In order to show consistency of our long run variance estimator we need to impose the following assumption, which states that the sequence of auto-covariances for the noise decay sufficiently fast and can be estimated well from a finite sample.

Assumption 3.3.1. *The auto-covariances decay fast enough that* $\sum_{h>1} h \left|\gamma_h\right| < \infty$*, and for any fixed integer* h *and any ordered subset of* $\{1, \dots, n-h\}$*, say* M*, it holds that* $|M|^{-1} \sum_{t \in M} \zeta_t \zeta_{t+h} = \gamma_h + \mathcal{O}_{\mathbb{P}}\left(1/\sqrt{|M|}\right)$.

With the above assumption in place, we have the following guarantee on the consistency of the estimator.

Lemma 3.3.3. *If the function* $f_\circ(\cdot)$ *is bounded and the noise terms satisfy Assumption 3.2.3 and Assumption 3.3.1 it holds that*

$$\left|\hat{\tau}^2_{DIF} - \tau^2\right| = \mathcal{O}_{\mathbb{P}}\left(\frac{W'}{\sqrt{n}} \vee \frac{1}{W'} \vee \frac{NW'^2}{n}\right).$$

Lemma 3.3.3 shows that if, for example, W' is chosen to be of the order $W' = \mathcal{O}\left(n^{\theta}\right)$ for some $\theta < 1/2$ then (3.12) satisfies the condition in part (ii) of Theorem 3.2.2 as long as the number of change points grows more slowly than $n^{1-2\theta} \log^{-1}(n/W)$. In practice we follow Wu and Zhao (2007) in setting $W' = n^{1/3}$.

3.3.3 Consistency of the algorithm

We now investigate the conditions under which algorithm Algorithm 3 is consistent, in the sense that with high probability it is able to detect all change points and returns no spurious intervals. For ease of reading we re-introduce the parametrization of the signal in model (3.1) between change point locations as follows:

$$f_\circ\left(t/n\right) = \begin{cases} \sum_{j=0}^{p} \alpha_{j,k}\left(t/n - \eta_k/n\right)^j & \text{if } \eta_{k-1} < t \leq \eta_k \\ \sum_{j=0}^{p} \beta_{j,k}\left(t/n - \eta_k/n\right)^j & \text{if } \eta_k < t \leq \eta_{k+1} \end{cases} \qquad k = 1, \ldots, N. \tag{3.13}$$

Therefore, the absolute change in the j-th derivative of $f_\circ(\cdot)$ at the k-th change point location can be written as $\Delta_{j,k} = |\alpha_{j,k} - \beta_{j,k}|$. Putting $\eta_0 = 0$ and $\eta_{N+1} = n$ we write $\delta_k = \min\left(\eta_k - \eta_{k-1}, \eta_{k+1} - \eta_k\right)$ for the effective sample size associated with the k-th change location. The most prominent change in derivative at each change point location can therefore be defined as follows:

$$p_k^* \in \arg\max_{0 \leq j \leq p} \left\{ \Delta_{j,k}\left(\frac{\delta_k}{n}\right)^j \right\} \qquad k = 1, \ldots, N. \tag{3.14}$$

In order to show the consistency of Algorithm 3 we impose two restriction on the signal. The first states that the changes in derivative at each change point location are bounded. The second states that although multiple changes in the derivatives of $f_\circ(\cdot)$ can occur at each change point location, there is always one dominating change. This excludes the possibility of signal cancellation occurring.

Assumption 3.3.2. *There is a constant $C_\Delta > 0$ such that $|\Delta_{jk}| < C_\Delta$ for each j, k.*

Assumption 3.3.3. *For each $k = 1, \ldots, N$ the quantity p_k^* is uniquely defined, and for any sequence $\left(\rho_{k,n}\right)_{n\geq 1}$ with the property $\rho_{k,n} \leq \delta_k/n$ for all $n \geq 1$ it holds that $\Delta_{j,k}\rho_{k,n}^j \leq C_{p_k^*}\Delta_{p_k^*,k}\rho_{k,n}^{p_k^*}$ for all $j \neq p_k^*$, where $C_{p_k^*} = \frac{1}{2^{p_k^*+2}(p^*+1)p}$.*

For example, Assumption 3.3.3 would be violated by the piecewise linear signal shown in (3.15) for which $n = 8$ and $\eta = 4$, and the scaled difference in slopes between the first four entries and the last four had the same magnitude but the opposite sign to the corresponding difference in levels. That is: $\Delta_0 = \Delta_1\left(\delta/n\right)$.

$$\mathbf{f} = \left(-7/8, -6/8, -5/8, -4/8, 3/8, 2/8, 1/8, 0\right)' \tag{3.15}$$

In practice, in situations when Assumption 3.3.3 is violated our procedure is still able to detect the corresponding change point. This is because although signal cancellation such as in (3.15) may occur on a particular interval considered by Algorithm 3, it is unlikely to occur on every interval considered. In the above example, if we were to look at the sub-vector $(-5/8, -4/8, 3/8, 2/8)'$ no cancellation would occur. See also Remark 3.6.10 in the Proofs section, where we show how the assumption can be relaxed for piecewise linear functions, and show good practical performance via simulation on higher order piecewise polynomials which violate the assumption. With these assumptions in place we have the following result.

Theorem 3.3.1. *Let $\hat{I}_1, \dots, \hat{I}_{\hat{N}}$ be intervals returned by Algorithm 3 run on data $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ from model (3.1), with parameters $a > 1$, W, and $\alpha \in (0, 1)$. Grant Assumptions 3.3.2-3.3.3 and either of Assumptions 3.2.2-3.2.1 or 3.2.3-3.2.5 hold, and let the threshold λ_α chosen according to (3.6) or (3.8) accordingly. If the the effective sample size at each change point location satisfies*

$$\delta_k > C_1 \left(W \vee n^{\frac{2p_k^*}{2p_k^*+1}} \left(\frac{\tau^2 \log(n)}{\Delta^2_{p_k^*,k}} \right)^{\frac{1}{2p_k^*+1}} \right) \qquad k = 1, \dots, N \tag{3.16}$$

then on the set with probability $1 - \alpha + o(1)$ the following events occur simultaneously:

$$E_3^* = \left\{ \hat{N} = N \right\},$$
$$E_4^* = \left\{ \hat{I}_k \cap \Theta = \{\eta_k\} \mid k = 1, \dots, N \right\},$$
$$E_5^* = \left\{ \left| \hat{I}_k \right| \le C_2 \left(W \vee n^{\frac{2p_k^*}{2p_k^*+1}} \left(\frac{\tau^2 \log(n)}{\Delta^2_{p_k^*,k}} \right)^{\frac{1}{2p_k^*+1}} \right) \middle| k = 1, \dots, N \right\}.$$

Here C_1 and C_2 depend only on α, a and p (and d in the case of Gaussian noise).

Theorem 3.3.1 states that on a set with probability asymptotically larger than $1 - \alpha$, where α can be tuned by the user, the number of intervals returned by Algorithm 3 coincides with the true number of change points (event E_3^*), and every interval returned contains exactly one change point (event E_4^*). Event E_5^* provides bounds on the widths of intervals returned. In Yu et al. (2022) it was shown that, under independent and light tailed noise, the mini-max localisation rate for each change point in the generic piecewsie polynomial model is of the order

$$\mathcal{O}\left(n^{\frac{2p_k^*}{2p_k^*+1}}\left(\frac{\tau^2}{\Delta^2_{p_k^*,k}}\right)^{\frac{1}{2p_k^*+1}}\right), \qquad k=1,\ldots,N. \tag{3.17}$$

Therefore, under Assumptions 3.2.1- 3.2.2 where W is of the order $\mathcal{O}\left(\log(n)\right)$, the bounds guaranteed by E_5^* can be seen to be optimal up to to log terms. That is, the width of each interval returned matches (up to log terms) the best possible rate at which the corresponding change point can be localised. Under Assumptions 3.2.3-3.2.5, where W grows slightly faster than $n^{2/(2+\nu)}$, the bounds provided by event E_5^* are again optimal as long as $\nu > 1$ and the most prominent change occurs in derivatives of order 1 or higher. However, whenever $p_k^* = 0$ comparing to (3.17) it is clear the bounds are no longer optimal.

The aforementioned lack of optimality is due to Assumption 3.2.5, which requires the minimum support of our detection statistic to be relatively larger. This is needed in order that a strong approximation result may be invoked for a range of noise distributions. However, the requirement that W grows at a polynomial rate with n can be overly conservative. For example, if the noise terms are independently distributed with finite moment generating function in a neighbourhood of zero, which is the setting studied by Yu et al. (2022), then Theorem 1 in Komlós et al. (1975) states that after enlarging the probability space

$$\sum_{t=1}^{n}\zeta_t - \tau B(n) = \mathcal{O}\left(\log(n)\right), \qquad \mathbb{P}\text{-almost surely.}$$

Consequently, in this setting the results of Theorem 3.3.1 continue to hold with W of the order $o\left(\log^3(n)\right)$. In which case, setting λ_α accordingly, the bound provided by event E_5^* again results optimal up to the log factors.

Regarding condition (3.16) in Theorem 3.3.1 above, the requirement is essentially unavoidable for the following reason: up to the W term and the logarithmic terms (3.16) agrees with the mini-max detection lower bound for change points in piecewise polynomials discussed in Yu et al. (2022). Therefore, were the spacing between change points any smaller, no method would be able to detect them. The term multi-scale refers to the fact that the condition is formulated for each change point individually, and allows for a combination of small changes over large intervals and large changes over small intervals.

The width of the k-th interval depends (up to constants) only on the order of the derivative at which the most prominent change occurs, and not on the overall polynomial degree of the signal. This shows the intervals adapt locally to the smoothness of the signal. Interestingly the rate $\mathcal{O}\left(n^{2p^*/(2p^*+1)}\right)$ is the same as the mini-max bound on the sup-norm risk for p^*-smooth Holder regression functions (Tsybakov, 2004, Theorem 2.10). The error probability α does not appear explicitly in Theorem 3.3.1 as it is absorbed into the constants C_1 and C_2. Indeed for different but fixed choices of α all thresholds constructed according to the rules discussed in Sections 3.2.3 and 3.2.4 will be asymptotically equivalent. However in finite samples there is a clear price to pay for requesting higher coverage since as $\alpha \downarrow 0$ we have that $-2\log^{-1}(1-\alpha) \sim 2/\alpha$.

Theorem 3.3.1 leads to the following large sample consistency result.

Corollary 3.3.2. *Let $\hat{I}_1, \ldots, \hat{I}_{\hat{N}}$ be intervals returned by Algorithm 3 under the same conditions as Theorem 3.3.1 but with threshold $\lambda = (1+\varepsilon)\, a_{W,n}$ for some fixed $\varepsilon > 0$, where $a_{W,n}$ is as defined in Theorem 3.2.2. Then on a set with probability $1 - o(1)$ the events E_1^*, E_2^*, and E_3^* occur simultaneously.*

An important consequence of Theorem 3.3.1 and Corollary 3.3.2 is that any point-wise estimator $\hat{\eta}_k$ for the k-th change point location which lies in an interval $\hat{I}_k$ will inherit the optimal localisation rate implied by event E_5^*. This extends to the naive estimator formed by setting $\hat{\eta}_k$ to the midpoint of the interval $\hat{I}_k$. However, more sophisticated estimators can be used; for example one may choose $\hat{\eta}_k$ to be the split point which results in the lowest sum of squared residuals when a piecewise polynomial function is fit over $\hat{I}_k$ (see for example Figure 3.1).

3.3.4 On the polynomial order of the signal

We emphasis that in the problem statement p refers to the maximum polynomial order of the signal on any stationary segment, and that the polynomial order of the signal is permitted to vary between segments. We observe that in applications analysts usually have in mind a reasonable idea of p motivated by knowledge of the problem at hand. However, it may be unreasonable to assume that the maximum polynomial order is known exactly. Therefore, we present two methods for determining p from data given upper and lower bounds $\underline{p}$ and $\overline{p}$ such that $p \in \left\{\underline{p}, \ldots, \overline{p}\right\}$. The methods are designed the setup in Sections 3.2.3 and 3.2.4 respectively.

Estimating p via the strengthened Schwarz Information Criterion

Fryzlewicz (2014) introduced the strengthened Schwarz Information Criterion (sSIC) for consistently estimating the number of change points in the canonical change point model for which the signal is piecewise constant and the contaminating noise is independently distributed and Gaussian. The same approach can be extended to estimating p in the piecewise polynomial model.

Given data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ from model (3.1) and some $p' \in \left\{\underline{p}, \ldots, \overline{p}\right\}$ let $\hat{I}_1, \ldots, \hat{I}_{\hat{N}_{p'}}$ be the output of Algorithm 3 under the assumption that the maximum polynomial degree is p', run with threshold $\lambda = (1+\varepsilon)\mathfrak{a}_{W,n}$ for some fixed $\varepsilon > 0$. Let $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}_{p'}}$ be the split points within each interval associated with the piecewsie polynomial fit providing the lowest sum of squared residuals and let $\hat{f}_{p'}(\cdot)$ be the function estimated via least squares between these knots. Following Section 3.4 in Fryzlewicz (2014) for some arbitrary but fixed $\alpha > 1$ the sCIC at p' is defined as

$$\text{sSIC}\left(p'\right) = \frac{n}{2} \log\left(\hat{\sigma}^2_{p'}\right) + (\hat{N}_{p'} + 1)\left(p' + 1\right) \log^{\alpha}(n),$$

where in particular

$$\hat{\sigma}^2_{p'} = \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - \hat{f}_{p'}(t/n)\right)^2.$$

Consequently, the maximum polynomial degree of the signal can be estimated as

$$\hat{p} = \underset{\underline{p} \leq p' \leq \overline{p}}{\arg\min} \, \mathrm{sSIC}\left(p'\right). \tag{3.18}$$

Regarding the large sample consistency of $\hat{p}$, we have the following result.

Lemma 3.3.4. *Let $\hat{p}$ be the estimator defined in (3.18). Grant Assumptions 3.2.2 and 3.2.1 as well as condition (3.16) hold, and moreover assume moreover that: (i) $\underline{p} \leq p \leq \overline{p}$ and $(\overline{p} - \underline{p}) = \mathcal{O}(1)$, (ii) $N = \mathcal{O}(1)$, and (iii) the coefficients in (3.13) are all of the order $\mathcal{O}(1)$. Then $\mathbb{P}\left(\hat{p} = p\right) \to 1$ as $n \to \infty$.*

Estimating p via recursive testing on null intervals

The finite difference functional which has so far been used to test for the presence of a change point can itself be used to estimate the maximum degree of the signal. For some $p' \in \left\{\underline{p}, \dots, \overline{p}\right\}$ let K be a contiguous subset of $\{1, \dots, n\}$ for which $|K|$ is a multiple of $(p'+2)$. Therefore, introduce the statistic

$$D_K^{p'}\left(\boldsymbol{Y}\right) = \left\{ \left\lfloor \frac{|K|}{p'+2} \right\rfloor \sum_{i=0}^{p'+1} \binom{p'+1}{i}^2 \right\}^{-1/2} \sum_{j=0}^{p'+1} (-1)^{p'+1-j} \binom{p'+1}{j} \bar{Y}_K^j \tag{3.19}$$

where in particular letting K have elements $\left\{k_1, \dots, k_{|K|}\right\}$ we write

$$\bar{Y}_K^j = Y_{k_1 + j\frac{|K|}{p'+2}} + \cdots + Y_{(j+k_1)\frac{|K|}{p'+2}}, \qquad j = 0, \dots, p'+1$$

for non-overlapping sums of the data over the $(p'+2)$ equally sized contiguous partitions of K. Note that if K corresponds to a stretch of data which contains no change points and $p' < p$ then (3.19) will be large in (absolute) expectation, whereas if $p' \geq p$ then (3.19) will be small.

Using the above intuition, to estimate p we first run Algorithm 3 with threshold $\lambda = (1+\varepsilon)\mathfrak{a}_{W,n}$ for some small but fixed $\varepsilon > 0$ assuming the maximum polynomial order of the signal is $\overline{p}$. We then obtain sets $\widehat{\mathbb{K}} = \{\hat{K}_1, \hat{K}_2, \dots\}$ by retaining indices *between* each interval returned, and trimming either the first or last few indices so that the number of elements in each $\hat{K}$ is a multiple of $(\overline{p}+1)$. Note that since $\overline{p} \geq p$ by Corollary 3.3.2 with high probability each $\hat{K}$ corresponds to a stretch of data which contains no change points. Finally we test whether $|D_{\hat{K}}^{\overline{p}-1}(\boldsymbol{Y})| > (1+\varepsilon)\mathfrak{a}_{W,n}$ for each $\hat{K}$. If any such test is not passed we conclude that $p = \overline{p}$. Else, we repeat the procedure with $\overline{p}-1$. The procedure automatically ends once $\underline{p}$ is reached, since we assume $p \geq \underline{p}$, and by this point we have concluded that $p < p'$ for all $p' > \underline{p}$. The procedure is sumarized in Algorithm 4. Regarding the large sample consistency of the output of Algorithm 4 we have the following result.

Algorithm 4: An algorithm for determining the maximum polynomial order of the signal by progressively estimating intervals of significance and testing null intervals for the presence of a change points in a lower degree polynomial.

```
function maxDegreeEstimation($\boldsymbol{Y}$, $\overline{p}$, $\underline{p}$):
    $p' \leftarrow \overline{p}$
    Detection $\leftarrow$ False
    while $p' > \underline{p}$ do
        Obtain intervals $\hat{\mathbb{K}} = \{\hat{K}_1, \hat{K}_2, \dots\}$ from Algorithm 3 using
        threshold $\lambda = (1+\varepsilon)\mathfrak{a}_{W,n}$ and assuming maximum degree $p'$.
        for $\hat{K} \in \hat{\mathbb{K}}$ do
            if $|D_{\hat{K}}^{p'-1}(\boldsymbol{Y})| > (1+\varepsilon)\mathfrak{a}_{W,n}$ then
                Detection $\leftarrow$ True
            end
        end
        if Detection then
            BREAK
        end
        $p' \leftarrow (p'-1)$
    end
return
```

Lemma 3.3.5. *Let $\hat{p}$ be the output of Algorithm 4. Grant Assumptions 3.2.3, 3.2.4, and 3.2.5 as well as condition (3.16) hold, and moreover assume moreover that: (i) $\underline{p} \leq p \leq \overline{p}$ and $(\overline{p} - \underline{p}) = \mathcal{O}(1)$, (ii) $N = \mathcal{O}(1)$, and (iii) the coefficients in (3.13) are all of the order $\mathcal{O}(1)$. Then $\mathbb{P}(\hat{p} = p) \to 1$ as $n \to \infty$.*

3.4 Simulation studies

3.4.1 Alternative methods for change point inference

We will compare our proposed methodology with existing algorithms with publicly available implementations, which each promise to return intervals containing true change point locations uniformly at a significance level chosen by the user. These are: the Narrowest Significance Pursuit (NSP) procedure of Fryzlewicz (2023), its self-normalised variant (NSP-SN), and its extension to auto-regressive signals (NSP-AR); the bootstrap confidence intervals for moving sums (MOSUM) of Cho and Kirch (2022a) using a single bandwidth (uniscale) and multiple bandwidths (multiscale); the simultaneous multiscale change point estimator (SMUCE) of Frick et al. (2014), as well as its extension to heterogeneous noise (H-SMUCE) developed by Pein et al. (2017), and its extension to dependent noise (Dep-SMUCE) developed by Dette et al. (2020). We also consider the conditional confidence intervals of Bai and Perron (1998) (B&P) with significance level Bonferroni-corrected for the estimated number of change-points. For our own procedure we write DIF1 for Algorithm 3 run under the assumptions of Theorem 3.2.1 and DIF2 for the algorithm run under the assumption of Theorem 3.2.2. Additionally we write MAD if the scale of the noise is estimated using the median absolute deviation estimator (3.9), SD if the scale is estimated using the difference based estimator of the standard deviation (3.10), and LRV if the long run variance is estimated using (3.12). Each of the methods considered is designed for different noise types and different change point models, and we summarise this information in Table 3.1 below.

Throughout the simulation studies, whenever a method requires the user to specify a minimum support parameter we set this to $W = 0.5n^{1/2}$. Exceptions occur for Dep-SMUCE for which we follow the authors' recommendation in setting $W = n^{1/3}$, for DIF1-MAD in which we set $W = \log(n)$ following the results of Theorem 3.2.1, and for the multiscale MOSUM procedure for which we generate a grid of bandwidths using the `bandwidths.auto` function in the MOSUM package Meier et al. (2021). For our own procedure we set the decay parameter regulating the density of the grid to $a = \sqrt{2}$ as was done in Kovács et al.

Table 3.1: Suitability of each method to non-Gaussian noise, dependent noise, and change point detection in higher order polynomial signals. The the letter **e** indicates that no theoretical guarantees are given but the authors observe good empirical performance of the method.

Method	non-Gaussian noise	dependent noise	higher order polynomials
DIF1-MAD	✗	✗	✓
DIF2-SD	✓	✗	✓
DIF2-LRV	✓	✓	✓
NSP	✗	✗	✓
NSP-SN	✓	✗	✓
NSP-AR	✗	✓	✓
B&P	✓	✗	✗
MOSUM (uniscale)	✓	✗	✗
MOSUM (multiscale)	✓	✗	✗
SMUCE	✗	✗	✗
H-SMUCE	**e**	✗	✗
Dep-SMUCE	✓	✓	✗

(2023) for the grid proposed therein.

3.4.2 Coverage on null signals

We first investigate empirically the coverage provided by our algorithm and the alternatives introduced in 3.4.1. To investigate coverage we apply each method to a vector of pure noise with length $n = 750$ generated according to each of the noise types listed below, setting the noise level to $\sigma = 1$, and over 500 replications record the proportion of times no intervals of significance are returned. For each procedure we set appropriate tuning parameters in order that the family-wise error is nominally controlled at the level $\alpha = 0.1$. Where applicable we ask each procedure to test for change points in polynomial signals of degrees 0, 1, and 2.

- (`N1`): $\zeta_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d.
- (`N2`): $\zeta_t \sim t_5 \times \sigma\sqrt{0.6}$ i.i.d.
- (`N3`): $\zeta_t = \phi\zeta_{t-t} + \varepsilon_t$ with $\phi = 0.5$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2/(1-\phi^2))$ i.i.d.

- (N4): $\zeta_t = \phi\zeta_{t-t} + \varepsilon_t$ with $\phi = 0.5$ and $\varepsilon_t \sim t_5 \times \sigma\sqrt{0.6/(1-\phi^2)}$ i.i.d.

The results of the simulation study are reported in Tables 3.2. We also highlight whether each method comes with theoretical coverage guarantees for each noise type, where the letter **c** indicates that the method should give correct coverage conditional on the event that the number of change points is correctly estimated. With the exception of Dep-SMUCE, which occasionally provides under coverage, all methods tested keep the nominal size well for noise types consistent with the assumptions under which they were developed and in general tend to provide over coverage. The coverage provided by our procedure is likewise accurate, and in particular under Gaussian noise tends to provide coverage closer to the level requested than that provided by competing methods. This shows that the asymptotic results in Theorems 3.2.1 and 3.2.2 hold well in finite samples, and that that our procedure is generally better calibrated than other available methods; see also the additional simulation study in Section 3.4.4 of the appendix, which shows that the same results hold for a range of signal lengths.

3.4.3 Performance on test signals

Next we investigate the performance of our method and its competitors on test signals containing change points. To investigate performance we apply each method to 500 sample paths from the change point models M1, M2, and M3 listed below, contaminated with each of the four noise types introduced in Section 3.4.2 above. On each iteration we record for each method: the number of intervals which contain at least one change point location (no. genuine), the proportion of intervals returned which contain at least one change point location (prop. genuine), the average length of intervals returned (length), and whether all intervals returned contain at least once change point location (coverage). We report the average of these quantities, and again highlight whether each method comes with theoretical coverage guarantees for each noise type (guarantee).

- (M1): the first $n = 512$ values of the piecewise constant blocks signal from Donoho and Johnstone (1994), shown in Figure 3.1a, with $N = 4$ change points at locations

Table 3.2: Proportion of times out of 500 replications each method returned no intervals of significance when applied to a noise vector of length $n = 750$, as well as whether each method is theoretically guaranteed to provide correct coverage. The letter **c** indicates that the method should give correct coverage conditional on the event that the number of change points is correctly estimated. The the letter **e** indicates that no theoretical guarantees are given but the authors observe good empirical performance of the method.

	guarantee	degree 0	degree 1	degree 2
DIF1-MAD	✓	0.91	0.91	0.93
DIF2-SD	✓	1.00	1.00	1.00
DIF2-LRV	✓	0.99	0.95	0.97
NSP	✓	0.96	0.98	0.98
NSP-SN	✓	1.00	1.00	1.00
NSP-AR	✓	0.99	1.00	1.00
B&P	**c**	0.99	-	-
MOSUM (uniscale)	**c**	0.99	-	-
MOSUM (multiscale)	**c**	0.95	-	-
SMUCE	✓	0.97	-	-
H-SMUCE	✓	0.98	-	-
Dep-SMUCE	✓	0.97	-	-

(a) Coverage on noise type `N1` with $\sigma = 1$.

	guarantee	degree 0	degree 1	degree 2
DIF1-MAD	✗	0.39	0.40	0.39
DIF2-SD	✓	0.96	0.95	0.96
DIF2-LRV	✓	0.95	0.90	0.90
NSP	✗	0.05	0.05	0.07
NSP-SN	✓	1.00	1.00	1.00
NSP-AR	✗	0.15	0.17	0.17
B&P	**c**	0.96	-	-
MOSUM (uniscale)	**c**	1.00	-	-
MOSUM (multiscale)	**c**	0.96	-	-
SMUCE	✗	0.18	-	-
H-SMUCE	**e**	0.98	-	-
Dep-SMUCE	✓	0.89	-	-

(b) Coverage on noise type `N2` with $\sigma = 1$.

	guarantee	degree 0	degree 1	degree 2
DIF1-MAD	✗	0.00	0.00	0.00
DIF2-SD	✗	0.00	0.00	0.00
DIF2-LRV	✓	0.98	0.98	0.98
NSP	✗	0.00	0.00	0.00
NSP-SN	✗	0.55	0.65	0.79
NSP-AR	✓	0.99	1.00	0.99
B&P	✗	0.11	-	-
MOSUM (uniscale)	✗	0.14	-	-
MOSUM (multiscale)	✗	0.00	-	-
SMUCE	✗	0.00	-	-
H-SMUCE	✗	0.24	-	-
Dep-SMUCE	✓	0.90	-	-

(c) Coverage on noise type `N3` with $\sigma = 1$.

	guarantee	degree 0	degree 1	degree 2
DIF1-MAD	✗	0.00	0.00	0.00
DIF2-SD	✗	0.00	0.00	0.00
DIF2-LRV	✓	0.97	0.96	0.96
NSP	✗	0.00	0.00	0.00
NSP-SN	✗	0.54	0.67	0.78
NSP-AR	✗	0.13	0.15	0.15
B&P	✗	0.09	-	-
MOSUM (uniscale)	✗	0.17	-	-
MOSUM (multiscale)	✗	0.01	-	-
SMUCE	✗	0.00	-	-
H-SMUCE	✗	0.30	-	-
Dep-SMUCE	✓	0.87	-	-

(d) Coverage on noise type `N4` with $\sigma = 1$.

$\Theta = \{205, 267, 308, 472\}$

- (M2): the first $n = 600$ values of the piecewise linear waves signal from Baranowski et al. (2019a), shown in Figure 3.6c, with $N = 3$ change points at locations $\Theta = \{150, 300, 450\}$

- (M3): the piecewise quadratic hills signal with length $n = 400$, shown in Figure 3.6e, with $N = 3$ change points at locations $\Theta = \{100, 200, 300\}$

The results of the simulation study are reported in Tables 3.3 - 3.5. On the piecewise constant blocks function, among the methods which provide correct coverage, our algorithm is generally among the top performing methods in terms the number of change points detected and the lengths of intervals recovered. In fact, is only outperformed by the MOSUM procedure with multiscale bandwidth under noise types N1 and N2. The family of SMUCE algorithms, as well as the B&P procedure, all suffer from under coverage on noise types for which they should give accurate coverage. Among the methods compared to only the family of NSP algorithms is applicable to higher order piecewise polynomial signals. On the piecewise polynomial waves and hills signals our methods deliver correct coverage where theoretical guarantees are available and consistently outperform the only competitor, the family of NSP algorithms.

3.4.4 Additional numerical illustrations

To further investigate the coverage provided by our method in finite samples, in this section we reproduce the simulation study in Section 3.4.2 for signals of length $n \in \{100, 500, 1000, 2000\}$. The results are shown in Tables 3.6-3.7, and confirm that for a range of signal lengths our procedure continues to provide accurate coverage.

Table 3.3: Average of the number of intervals which contain at least one change point location (no. genuine), the proportion of intervals returned which contain at least one change point location (prop. genuine), the average length of intervals returned (length), and whether all intervals returned contain at least once change point location (coverage), on the piecewise constant `blocks` signal contaminated with noise `N1-N4` over 500 replications. The noise level was set to $\sigma = 10$ for noise types `N1-2`and to $\sigma = 5$ for noise types `N3-4`. We also report whether each method is theoretically guaranteed to provide correct coverage.

		N1	N2	N3	N4
DIF1-MAD	no. genuine	3.68	3.80	3.63	3.70
	prop. genuine	0.98	0.90	0.27	0.23
	length	34.78	27.50	12.30	10.64
	coverage	0.93	0.59	0.00	0.00
	guarantee	✓	✗	✗	✗
DIF2-SD	no. genuine	3.30	3.28	3.74	3.83
	prop. genuine	1.00	1.00	0.44	0.47
	length	42.92	41.57	19.74	19.44
	coverage	1.00	0.99	0.00	0.00
	guarantee	✓	✓	✗	✗
DIF2-LRV	no. genuine	2.14	2.10	1.87	2.27
	prop. genuine	1.00	1.00	1.00	1.00
	length	61.83	61.40	63.06	58.90
	coverage	1.00	1.00	1.00	1.00
	guarantee	✓	✓	✓	✓
NSP	no. genuine	3.20	3.36	3.86	3.86
	prop. genuine	1.00	0.63	0.41	0.25
	length	62.03	34.32	19.08	13.23
	coverage	1.00	0.17	0.00	0.00
	guarantee	✓	✗	✗	✗
NSP-SN	no. genuine	1.85	1.87	2.91	3.01
	prop. genuine	1.00	1.00	1.00	1.00
	length	119.97	115.08	75.98	69.22
	coverage	1.00	1.00	1.00	0.99
	guarantee	✓	✓	✗	✗
NSP-AR	no. genuine	0.10	0.76	0.13	0.75
	prop. genuine	0.10	0.46	0.13	0.42
	length	15.41	59.56	14.94	47.05
	coverage	1.00	0.44	1.00	0.45
	guarantee	✓	✗	✓	✗
B&P	no. genuine	3.88	3.90	3.77	3.88
	prop. genuine	0.96	0.96	0.54	0.57
	length	16.94	17.43	16.52	15.70
	coverage	0.86	0.85	0.06	0.07
	guarantee	**c**	**c**	✗	✗
MOSUM (uniscale)	no. genuine	1.75	1.94	3.36	3.53
	prop. genuine	0.78	0.84	0.54	0.60
	length	13.48	13.87	13.16	11.43
	coverage	0.90	0.93	0.06	0.09
	guarantee	**c**	**c**	✗	✗
MOSUM (multiscale)	no. genuine	3.93	3.93	4.04	4.11
	prop. genuine	0.97	0.98	0.42	0.48
	length	21.91	21.22	20.98	19.67
	coverage	0.89	0.91	0.01	0.02
	guarantee	**c**	**c**	✗	✗
SMUCE	no. genuine	3.70	3.65	3.31	2.84
	prop. genuine	0.95	0.73	0.34	0.20
	length	38.70	25.61	14.95	11.37
	coverage	0.89	0.32	0.00	0.00
	guarantee	✓	✗	✗	✗
H-SMUCE	no. genuine	3.09	3.07	3.17	3.54
	prop. genuine	0.86	0.87	0.89	0.92
	length	45.67	45.87	47.78	40.72
	coverage	0.69	0.70	0.74	0.82
	guarantee	✓	**e**	✗	✗
Dep-SMUCE	no. genuine	2.18	2.29	3.47	3.75
	prop. genuine	0.82	0.82	0.87	0.91
	length	78.05	73.10	44.60	37.20
	coverage	0.64	0.62	0.61	0.69
	guarantee	✓	✓	✓	✓

Table 3.4: Average of the number of intervals which contain at least one change point location (no. genuine), the proportion of intervals returned which contain at least one change point location (prop. genuine), the average length of intervals returned (length), and whether all intervals returned contain at least once change point location (coverage), on the piecewise linear `waves` signal contaminated with noise types `N1-N4` over 100 replications. The noise level was set to $\sigma = 5$ for all noise types. We also report whether each method is theoretically guaranteed to provide correct coverage.

		N1	N2	N3	N4
DIF1-MAD	no. genuine	2.85	2.66	1.30	1.32
	prop. genuine	0.98	0.83	0.09	0.08
	length	124.94	98.11	17.61	14.52
	coverage	0.93	0.52	0.00	0.00
	guarantee	✓	✗	✗	✗
DIF2-SD	no. genuine	2.68	2.67	1.35	1.31
	prop. genuine	1.00	0.99	0.15	0.16
	length	145.58	144.37	25.30	25.85
	coverage	1.00	0.96	0.00	0.00
	guarantee	✓	✓	✗	✗
DIF2-LRV	no. genuine	2.64	2.61	1.75	1.82
	prop. genuine	0.99	0.98	1.00	0.99
	length	145.92	142.67	216.18	198.82
	coverage	0.98	0.93	1.00	0.98
	guarantee	✓	✓	✓	✓
NSP	no. genuine	2.80	2.37	1.87	1.66
	prop. genuine	1.00	0.66	0.31	0.17
	length	143.84	84.95	43.90	27.58
	coverage	1.00	0.24	0.01	0.00
	guarantee	✓	✗	✗	✗
NSP-SN	no. genuine	2.34	2.37	2.16	2.41
	prop. genuine	1.00	1.00	1.00	1.00
	length	163.03	160.86	163.71	153.61
	coverage	1.00	1.00	1.00	1.00
	guarantee	✓	✓	✗	✗
NSP-AR	no. genuine	0.33	1.08	0.01	0.60
	prop. genuine	0.31	0.60	0.01	0.39
	length	74.63	109.75	2.08	80.46
	coverage	1.00	0.46	1.00	0.43
	guarantee	✓	✗	✓	✗

Table 3.5: Average of the number of intervals which contain at least one change point location (no. genuine), the proportion of intervals returned which contain at least one change point location (prop. genuine), the average length of intervals returned (length), and whether all intervals returned contain at least once change point location (coverage), on the piecewise quadratic `hills` signal contaminated with noise types `N1-N4` over 100 replications. The noise level was set to $\sigma = 1$ for all noise types. We also report whether each method is theoretically guaranteed to provide correct coverage.

		N1	N2	N3	N4
DIF1-MAD	no. genuine	2.37	2.39	1.48	1.46
	prop. genuine	0.98	0.86	0.18	0.15
	length	103.50	83.64	21.90	17.84
	coverage	0.95	0.65	0.00	0.00
	guarantee	✓	✗	✗	✗
DIF2-SD	no. genuine	2.08	2.09	1.61	1.56
	prop. genuine	1.00	0.99	0.29	0.28
	length	120.44	119.57	31.56	30.62
	coverage	1.00	0.98	0.00	0.00
	guarantee	✓	✓	✗	✗
DIF2-LRV	no. genuine	2.06	2.04	0.76	1.02
	prop. genuine	0.99	0.98	0.61	0.78
	length	121.59	120.27	90.69	116.77
	coverage	0.99	0.96	0.99	0.98
	guarantee	✓	✓	✓	✓
NSP	no. genuine	2.11	2.26	2.20	1.99
	prop. genuine	1.00	0.80	0.58	0.36
	length	117.56	81.50	54.89	35.28
	coverage	1.00	0.50	0.14	0.01
	guarantee	✓	✗	✗	✗
NSP-SN	no. genuine	1.50	1.57	1.42	1.68
	prop. genuine	1.00	1.00	0.99	1.00
	length	152.72	150.39	147.75	137.51
	coverage	1.00	1.00	1.00	1.00
	guarantee	✓	✓	✗	✗
NSP-AR	no. genuine	0.07	0.69	0.02	0.50
	prop. genuine	0.07	0.49	0.02	0.36
	length	11.76	65.02	3.64	54.29
	coverage	1.00	0.64	1.00	0.55
	guarantee	✓	✗	✓	✗

Table 3.6: Proportion of times out of 500 replications each method returned no intervals of significance when applied to a noise vector of length $n \in \{100, 500, 1000, 2000\}$, as well as whether each method is theoretically guaranteed to provide correct coverage.

		guarantee	degree 0	degree 1	degree 2
n = 100	DIF1-MAD	✓	0.91	0.89	0.92
	DIF2-SD	✓	1.00	1.00	1.00
	DIF2-LRV	✓	0.98	0.97	0.96
n = 500	DIF1-MAD	✓	0.87	0.92	0.91
	DIF2-SD	✓	1.00	1.00	1.00
	DIF2-LRV	✓	0.97	0.97	0.95
n = 1000	DIF1-MAD	✓	0.93	0.91	0.91
	DIF2-SD	✓	0.99	0.99	1.00
	DIF2-LRV	✓	0.97	0.98	0.95
n = 2000	DIF1-MAD	✓	0.89	0.91	0.90
	DIF2-SD	✓	0.99	1.00	0.99
	DIF2-LRV	✓	0.97	0.97	0.98

(a) Coverage on noise type `N1` with $\sigma = 1$.

		guarantee	degree 0	degree 1	degree 2
n = 100	DIF1-MAD	✗	0.76	0.67	0.78
	DIF2-SD	✓	1.00	0.98	0.99
	DIF2-LRV	✓	0.97	0.89	0.93
n = 500	DIF1-MAD	✗	0.54	0.46	0.53
	DIF2-SD	✓	0.99	0.97	0.96
	DIF2-LRV	✓	0.96	0.95	0.90
n = 1000	DIF1-MAD	✗	0.39	0.32	0.33
	DIF2-SD	✓	0.97	0.97	0.95
	DIF2-LRV	✓	0.95	0.95	0.89
n = 2000	DIF1-MAD	✗	0.26	0.21	0.20
	DIF2-SD	✓	0.99	0.96	0.97
	DIF2-LRV	✓	0.99	0.94	0.92

(b) Coverage on noise type `N2` with $\sigma = 1$.

Table 3.7: Proportion of times out of 500 replications each method returned no intervals of significance when applied to a noise vector of length $n \in \{100, 500, 1000, 2000\}$, as well as whether each method is theoretically guaranteed to provide correct coverage.

		guarantee	degree 0	degree 1	degree 2
n = 100	DIF1-MAD	✗	0.01	0.01	0.04
	DIF2-SD	✗	0.11	0.13	0.18
	DIF2-LRV	✓	0.97	0.96	0.96
n = 500	DIF1-MAD	✗	0.00	0.00	0.00
	DIF2-SD	✗	0.00	0.00	0.00
	DIF2-LRV	✓	0.99	0.98	0.98
n = 1000	DIF1-MAD	✗	0.00	0.00	0.00
	DIF2-SD	✗	0.00	0.00	0.00
	DIF2-LRV	✓	0.99	0.97	0.99
n = 2000	DIF1-MAD	✗	0.00	0.00	0.00
	DIF2-SD	✗	0.00	0.00	0.00
	DIF2-LRV	✓	0.99	0.98	0.99

(a) Coverage on noise type N3 with $\sigma = 1$.

		guarantee	degree 0	degree 1	degree 2
n = 100	DIF1-MAD	✗	0.01	0.00	0.01
	DIF2-SD	✗	0.11	0.10	0.16
	DIF2-LRV	✓	0.95	0.94	0.98
n = 500	DIF1-MAD	✗	0.00	0.00	0.00
	DIF2-SD	✗	0.00	0.00	0.00
	DIF2-LRV	✓	0.97	0.95	0.95
n = 1000	DIF1-MAD	✗	0.00	0.00	0.00
	DIF2-SD	✗	0.00	0.00	0.00
	DIF2-LRV	✓	0.96	0.95	0.96
n = 2000	DIF1-MAD	✗	0.00	0.00	0.00
	DIF2-SD	✗	0.00	0.00	0.00
	DIF2-LRV	✓	0.97	0.96	0.97

(b) Coverage on noise type N4 with $\sigma = 1$.

3.5 Real data examples

3.5.1 Application to bone mineral density acquisition curves

We analyse data on bone mineral acquisition in 423 healthy males and females aged between 9 and 25. The data is available from `hastie.su.domains` and was first analysed in Bachrach et al. (1999). The data was originally collected as part of a longitudinal study where four consecutive yearly measurements of bone mass by dual energy x-ray absorptiometry were taken from each subject. We obtain bone density acquisition curves for males and females by grouping measurements by gender and age and averaging over measurements in each grouping. The processed data are plotted in the first row of Figure 3.3. There is some disagreement over the age at which peak bone mass density is attained in adolescents Kröger et al. (1993); Theintz et al. (1992); Lu et al. (1996). One possible solution is to model the data in Figure 3.3 as following a piecewise linear trend, and to infer this information from any estimated change point locations.

We apply the procedure DIF2-SD to the data, with the tuning parameters specified in Section 3.4, because as the data are strictly positive the assumption of Gaussian noise is unlikely to hold. We additionally estimate change point locations using five state of the art algorithms for recovering changes in piecewise linear signals which however do not come with any coverage guarantees. These are: the Narrowest-Over-Threshold algorithm (NOT) of Baranowski et al. (2019a), and the same algorithm run with the requirement that the estimated signal be continuous (NOT-cont), the Isolate Detect algorithm (ID) of Anastasiou and Fryzlewicz (2022), the dynamic programming based algorithm of Bai and Perron (1998) (BP), and the Continuous-piecewise-linear Pruned Optimal Partitioning algorithm (CPOP) of Fearnhead et al. (2019). When applying each method we use the default parameters in their respective R packages.

The results of the analysis are shown in in the second row of Figure 3.3. On both bone density acquisition curves all methods for change point detection estimate a single change point location, save for CPOP. However, on the male bone density acquisition data there is considerable disagreement among the methods regarding the location of the change point detected. Since the methods do not quantify the uncertainty around each estimated change point, it is difficult to say which estimate is closest to the truth. DIF2-SD also returns a single interval of significance when applied to each data set, and each interval returned contains all change point locations recovered by the other methods on each respective data set save the extraneous change point detected by CPOP. By Corollary 3.3.1 one can be certain each interval contains at least one true change point location with high probability. We therefore re-apply the aforementioned change point detectors to this interval only. The results are shown in the third row of Figure 3.3, where this time there is much greater agreement among the methods. We also note that the corresponding intervals returned by NSP-SN (not shown), which is the only competing method from Section 3.4.1 applicable to the data, cover essentially the entire range of the data.

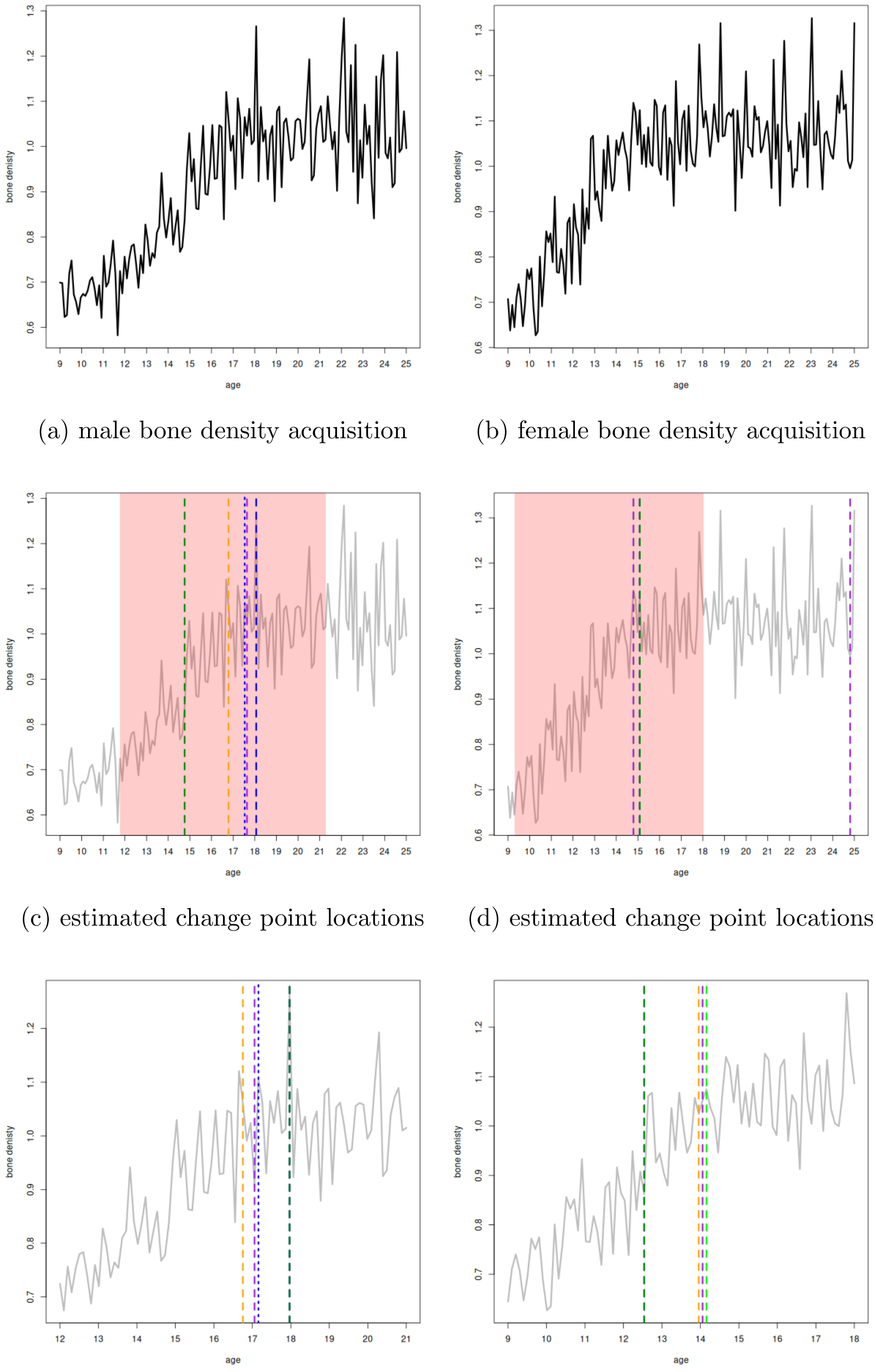
3.5.2 Applications to nitrogen dioxide concentration in London

We analyse daily average concentrations of nitrogen dioxide (NO_2) at Marylebone Road in London between September 2, 2000 and September 30, 2020. The data are available from `uk-air.defra.gov.uk` and were originally analysed from a change point perspective, assuming a piecewsie constant mean, by Cho and Fryzlewicz (2023). We follow their analysis in Cho and Fryzlewicz (2022) by taking the square root transform of the data and removing seasonal and weekly variation. The processed data is plotted in Figure 3.4. Cho and Fryzlewicz (2023) identify three historical events which are likely to have affected NO_2 concentration levels in London during the period in question, which are summarised below.

- February 2003: *installation of particulate traps on most London buses and other heavy duty diesel vehicles.*
- April 8, 2019: *introduction of Ultra Low Emission zones in central London.*
- March 23, 2020: *beginning of the nation-wide COVID-19 lockdown.*

We apply the procedure DIF2-LRV to the data with tuning parameters specified in Section 3.4, since time series of NO_2 concentrations are known to be strongly serially correlated. For comparison we additionally estimate change point locations using three state

Figure 3.3: black / grey solid lines (— / —) represents bone density acquisition curves for males and females between the ages of 9 and 25, red shaded regions (🟥) represent intervals of significance returned by DIF2-SD, dashed coloured lines represent change point locations recovered by NOT (- - -), NOT-cont (···), ID (- - -), BP (- - -), and CPOP (- - -)



(a) male bone density acquisition

(b) female bone density acquisition

(c) estimated change point locations

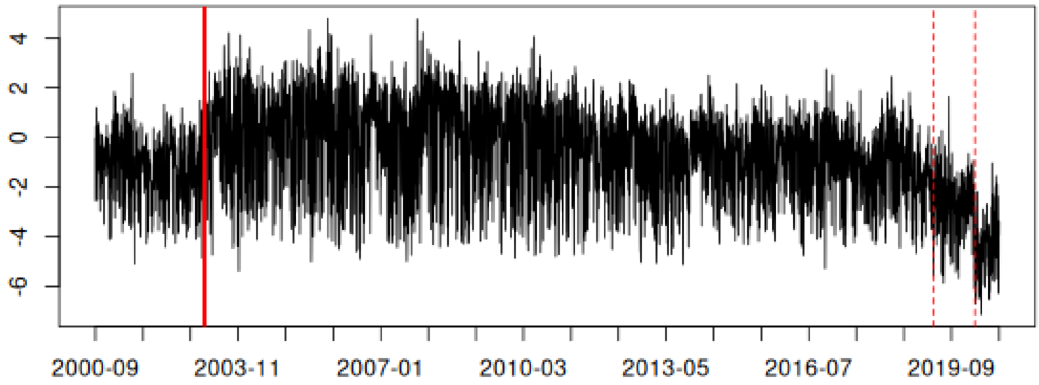
(d) estimated change point locations

(e) change points on sub-interval

(f) change points on sub-interval

of the art algorithms for recovering changes in piecewise constant signals in the presence of serially correlated noise, which however do not come with coverage guarantees. These are: the algorithm of Romano et al. (2022) for Detecting Changes in Autocorrelated and Fluctuating Signals (DeCAFS), the algorithm of Chakar et al. (2017) for estimating multiple change-points in the mean of a Gaussian AR(1) process (AR1seg), and the Wild Contrast Maximisation and gappy Schwarz algorithm (WCM.gSa) of Cho and Fryzlewicz (2023). When applying each method we use default parameters in their respective R packages save for the DeCAFS algorithm for which our choice of tuning parameters is guided by the `guidedModelSelection` function in the DeCAFS R package.

Figure 3.4: daily average concentrations of NO_2 at Marylebone Road after square root transform and with seasonal variation removed, red dashed lines (- - -) and dark red shaded region (🟥) represent dates of events which are likely to have affected NO_2 concentration levels



The results of the analysis are shown in Figure 3.5. DIF2-LRV returns four intervals, among which the first, third, and fourth cover the dates of important events identified by Cho and Fryzlewicz (2023). Within each of these three intervals AR1seg, DeCAFS, and WCM.gSa each identify one change point, with the exception of WCM.gSa which identifies two change points in the third interval returned. However, when we re-apply WCM.gSa over the third interval only one change point is detected, suggesting the second change point in this interval was spuriously estimated. DeCAFS detects a change point between the first and second intervals returned by DIF2-LRV. However, re-applying the algorithms to data between the two intervals no change points are detected suggesting the original change points were also spuriously estimated. We finally note that the data analysed consists of $n = 7139$ observations, and running DIF2-LRV on a desktop computer with a 3.20GHz Intel (R) Core (TM) i7-8700 CPU took 4.1 seconds. Running Dep-SMUCE and NSP-AR, which are the only competing methods from Section 3.4.1 applicable to the data, on the same machine took 15.1 seconds and 145.8 seconds respectively. Dep-SMUCE returns similar intervals to DIF2-LRV, whereas NSP-AR does not detect any change points in the data.

3.6 Proofs

For sequences $\{a_n\}_{n>0}$ and $\{b_n\}_{n>0}$ we write $a_n \lesssim b_n$ if there is a constant $C > 0$ for which $a_n \leq Cb_n$ for every $n > 0$. We write $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. We write $|\mathcal{A}|$ for the cardinality of a set $\mathcal{A}$. The density, cumulative density, and tail functions of a standard Gaussian random variable are written respectively as $\phi(\cdot)$, $\Phi(\cdot)$, and $\bar{\Phi}(\cdot)$.

3.6.1 Preparatory results

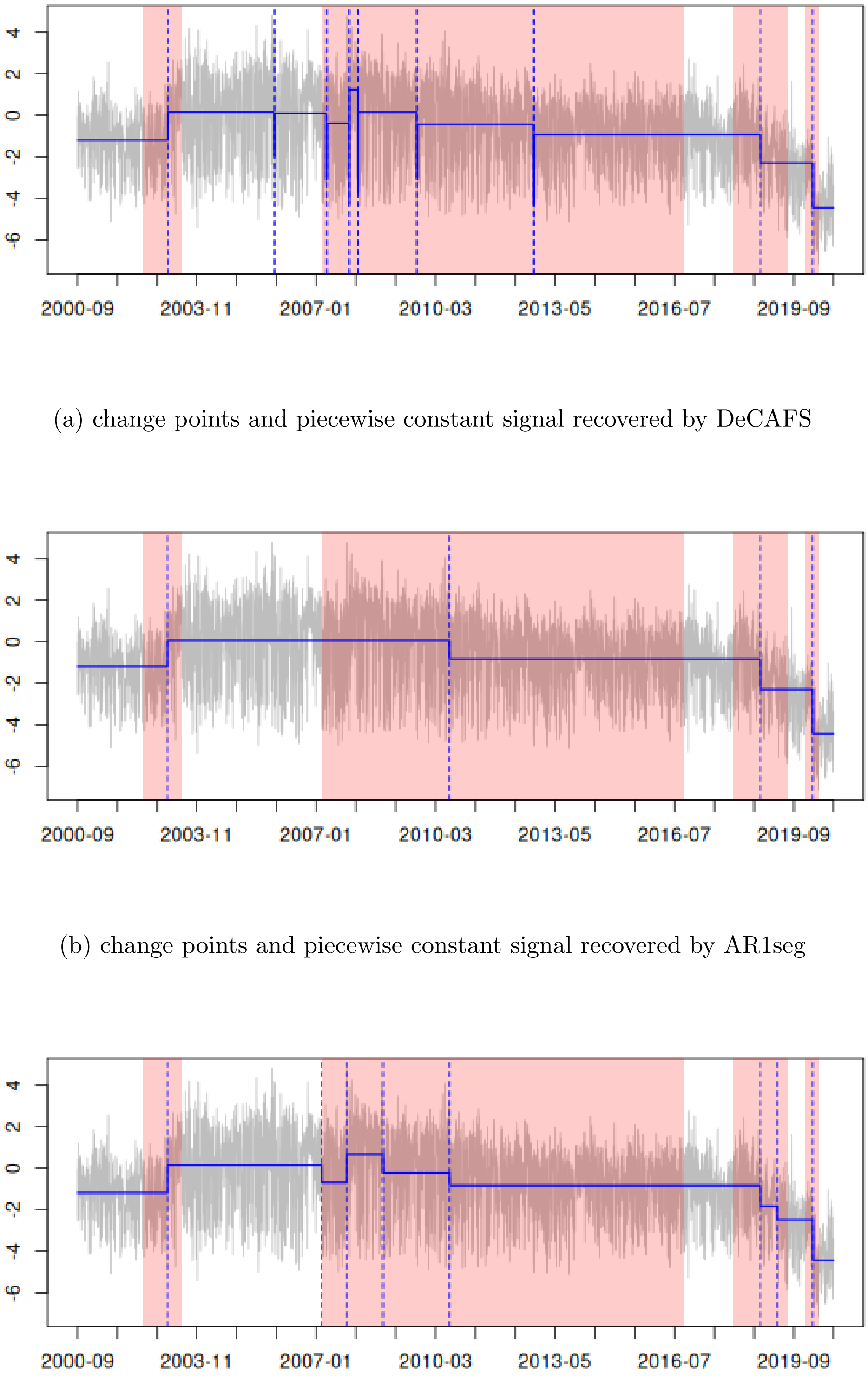
Definition 3.6.1 (Leadbetter et al. 2012, Section 12.1)**.** *Let $\{\xi(t)\}_{t>0}$ be a centred Gaussian process with unit variance, then if there are constants $C_\xi > 0$ and $\alpha \in (0, 2]$ such that for all $t > 0$ the following holds*

$$Cov(\xi(t), \xi(t+s)) = 1 - C_\xi |s|^\alpha + o(|s|^\alpha), \qquad |s| \to 0,$$

the process is called stationary with index α and local structure C_ξ. Moreover, the process has almost surely continuous sample paths and for any compact $K \subset \mathbb{R}^+$ the quantity $M_K = \sup_{t \in K} \{\xi(t)\}$ is well defined.

Lemma 3.6.1 (Berman's lemma)**.** *Let $\zeta_1, \ldots, \zeta_n$ and $\tilde{\zeta}_1, \ldots, \tilde{\zeta}_n$ be two sequences of Gaus-*

Figure 3.5: grey lines (—) represent daily average concentrations of NO_2 at Marylebone Road after square root transform and with seasonal variation removed, red shaded regions (🟥) represent intervals of significance returned by DIF2-LRV, blue dashed lines (- - -) represent change points recovered by a given algorithm, blue solid lines (—) represent the corresponding fitted piecewise constant signal.



(a) change points and piecewise constant signal recovered by DeCAFS

(b) change points and piecewise constant signal recovered by AR1seg

(c) change points and piecewise constant signal recovered by WCM.gSa *sian random variables with marginal* $\mathcal{N}(0,1)$ *distribution and covariances* $Cov\,(\zeta_i,\zeta_j)=\Lambda_{ij}$ *and* $Cov\left(\tilde{\zeta}_i,\tilde{\zeta}_j\right)=\tilde{\Lambda}_{ij}$*. Define* $\rho_{ij}=\max\left(|\Lambda_{ij}|,\left|\tilde{\Lambda}_{ij}\right|\right)$*. For any real numbers* $u_1,\ldots,u_n$ *the following holds:*

$$\begin{aligned}&\left|\mathbb{P}\left(\zeta_j\le u_j\mid 1\le j\le n\right)-\mathbb{P}\left(\tilde{\zeta}_j\le u_j\mid 1\le j\le n\right)\right|\\&\le\frac{1}{2\pi}\sum_{1\le i<j\le n}\left|\Lambda_{ij}-\tilde{\Lambda}_{ij}\right|\left(1-\rho_{ij}^2\right)^{-1/2}\exp\left(-\frac{\frac{1}{2}\left(u_i^2+u_j^2\right)}{1+\rho_{ij}}\right).\end{aligned}$$

Proof. See Theorem 4.2.1 in Leadbetter et al. (2012). □

Lemma 3.6.2 (Khintchine's lemma)**.** *Let* $\{M_n\}_{n>0}$ *be a sequence of random variables and let* G *be a non-degenerate distribution. If* $\{(c_n,d_n)\}_{n>0}$ *are scaling and centring sequences such that* $(M_n-c_n)/d_n\to G$ *then for any alternative sequences* $\{(c_n',d_n')\}_{n>0}$ *satisfying* $d_n/d_n'\sim 1$ *and* $(c_n-c_n')/d_n=o(1)$ *we also have that* $(M_n-c_d')/d_n'\to G$*.*

Proof. See Theorem 1.2.3 in Leadbetter et al. (2012). □

Lemma 3.6.3 (Pickand's lemma, continuous version)**.** *Let* $\{\xi(t)\}_{t>0}$ *be a stationary Gaussian process with index* $\alpha\in(0,2]$ *and local structure* $C_\xi>0$*. There is a constant* $H_\alpha>0$ *such that for any compact* $K\subset\mathbb{R}^+$ *the following holds:*

$$\mathbb{P}\left(\sup_{t\in K}\{\xi\left(t\right)\}>u\right)\sim H_\alpha C_\xi^{1/\alpha}\left|K\right|u^{2/\alpha-1}\phi\left(u\right).$$

Moreover the values $H_1=1$ *and* $H_2=1/\sqrt{\pi}$ *are known explicitly.*

Proof. See Theorem 9.15 in Piterbarg (2015), and Remark 12.2.10 in Leadbetter et al. (2012) for the values of H_α. □

Lemma 3.6.4 (Pickand's lemma, discrete version)**.** *Let* $\{\xi(t)\}_{t>0}$ *be a stationary Gaussian process with index* $\alpha\in(0,2]$ *and local structure* $C_\xi>0$*. If* $q\to 0$ *and* $u\to\infty$ *in such a* *way that $u^{2/\alpha}q \to a > 0$ the following holds for any compact $K \subset \mathbb{R}^+$:*

$$\mathbb{P}\left(\sup_{t\in K\cap\mathbb{Z}q}\{\xi\,(t)\} > u\right) \sim F_\xi\,(a)\,|K|\,u^{2/\alpha-1}\phi\,(u)\,.$$

The function $F_\xi\,(\cdot)$ is defined as follows

$$F_\xi\,(a) = \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\exp\left(\sup_{s\in[0,T]\cap a\mathbb{Z}} Z\,(s)\right)\right].$$

Where $\{Z(s)\}_{s>0}$ is a stationary Gaussian process with first and second moments as follows

$$\mathbb{E}\,(Z(s)) = -C_\xi\,|s|^\alpha\,,$$
$$Cov(Z(s_1), Z(s_2)) = C_\xi\,|s_1|^\alpha + C_\xi\,|s_2|^\alpha - C_\xi\,|s_1 - s_2|^\alpha\,.$$

Proof. See Lemma 12.2.1 in Leadbetter et al. (2012). □

Lemma 3.6.5. *Let $\{B\,(t)\}_{t>0}$ be standard Brownian motion and define the function $F\,(\cdot)$ as follows:*

$$F\,(x) = \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\exp\left(\sup_{s\in[0,T]\cap x\mathbb{Z}}\{B(s) - s/2\}\right)\right].$$

(i) For $x > 0$ it holds that $F\,(x) = p_\infty^2\,(x)\,/x$ where $p_\infty\,(\cdot)$ is defined as follows.

$$p_\infty\,(x) = \exp\left(-\sum_{k=1}^{\infty}\frac{1}{k}\bar{\Phi}\left(\sqrt{kx/4}\right)\right)$$

(ii) Putting $G\,(y) = (1/y)\,F\,(C/y)$ for any fixed $C > 0$ it holds that $G\,(y) \sim 1/2y$ as $y \to \infty$.

Proof. See Theorem 7.2 and Corollary 3.18 respectively in Kabluchko (2007). □

3.6.2 Intermediate results

Lemma 3.6.6. *Let $\{B(t)\}_{t>0}$ be standard Brownian motion and define the process $\{\xi(t)\}_{t>0}$ as follows:*

$$\xi(l) = \left\{ \left(\frac{1}{p+2}\right) \sum_{i=0}^{p+1} \binom{p+1}{i}^2 \right\}^{-1/2} \sum_{j=0}^{p+1} (-1)^{p+1-j} \binom{p+1}{j} \mathcal{Y}_{l,j},$$
$$\mathcal{Y}_{l,j} = \left[B\left(l + \frac{j+1}{p+2}\right) - B\left(l + \frac{j}{p+2}\right) \right].$$

(i) The process $\{\xi(l)\}_{l>0}$ is the continuous time analogue of $\frac{1}{\sigma} D^p_{l,w}(Y)$ under Assumption 3.2.1 and the null of no change points, in the sense that for a given scale w the following holds:

$$\left\{ \frac{1}{\sigma} D^p_{l,w}(Y) \mid 1 \leq l \leq n-w \right\} \overset{d}{=} \{\xi(l/w) \mid 1 \leq l \leq n-w\}.$$

(ii) According to Definition 3.6.1 the process is locally stationary with index $\alpha = 1$ and local structure C_p defined as follows:

$$C_p = (p+2) \left(1 + \frac{\sum_{j=1}^{p+1} \binom{p+1}{j} \binom{p+1}{j-1}}{\sum_{j=0}^{p+1} \binom{p+1}{j}^2} \right).$$

Proof. Part (i) can be verified by inspection. To show part (ii) note that for all $l > 0$ we have $\mathbb{E}(\xi(l)) = 0$ and $\mathbb{E}(\xi^2(l)) = 1$, so it remains to calculate the covariance between $\xi(l)$ and $\xi(l + s_l)$ for $|s_l| \to 0$. First, taking $s_l > 0$ we have the following:

$$\begin{aligned}
\operatorname{Cov}(\xi(l), \xi(l+s_l)) &= \left(\left(\frac{1}{p+2}\right) \sum_{i=0}^{p+1} \binom{p+1}{i}^2 \right)^{-1} \sum_{j=0}^{p+1} \sum_{k=0}^{p+1} (-1)^{j+k} \operatorname{Cov}(\mathcal{Y}_{l,j}, \mathcal{Y}_{l+s_l,k}) \\
&= \left(\left(\frac{1}{p+2}\right) \sum_{i=0}^{p+1} \binom{p+1}{i}^2 \right)^{-1} \left\{ \sum_{j=0}^{p+1} \binom{p+1}{j}^2 \operatorname{Cov}(\mathcal{Y}_{l,j}, \mathcal{Y}_{l+s_l,j}) \right. \\
&\quad \left. + \sum_{j=1}^{p+1} (-1) \binom{p+1}{j} \binom{p+1}{j-1} \operatorname{Cov}(\mathcal{Y}_{l,j}, \mathcal{Y}_{l+s_l,j-1}) \right\}.
\end{aligned}$$

Using the fact that $\mathrm{Cov}\left(B(l_1), B(l_2)\right) = \min\left(l_1, l_2\right)$ gives the following:

$$\mathrm{Cov}\left(\mathcal{Y}_{l,j}, \mathcal{Y}_{l+s_l,j}\right) = \frac{1}{p+2} - s_l,$$
$$\mathrm{Cov}\left(\mathcal{Y}_{l,j}, \mathcal{Y}_{l+s_l,j-1}\right) = s_l.$$

Therefore for $s_l \to 0$ with $s_l > 0$ we have the following:

$$\mathrm{Cov}\left(\xi(l), \xi(l+s_l)\right) = 1 - (p+2)\left(1 + \frac{\sum_{j=1}^{p+1}\binom{p+1}{j}\binom{p+1}{j-1}}{\sum_{j=0}^{p+1}\binom{p+1}{j}^2}\right) s_l.$$

The same calculations can be repeated for the case $s_l < 0$ and so ultimately we have that $\mathrm{Cov}\left(\xi(l), \xi(l+s_l)\right) = 1 - C_p|s_l|$ as $|s_l| \to 0$.

□

Lemma 3.6.7. *Consider the problem of testing for the presence of a change point on the interval $I = \{1, \dots, m\}$ where m satisfies $(p+2)\delta' \leq m < (p+2)(\delta'+1)$ for some integer $\delta' > 1$. If the interval contains a single change point at location δ' with change sizes $\Delta_0, \dots, \Delta_p$ then the test*

$$T_{1,m}^{\lambda} = \mathbf{1}\left\{|D_{1,m}^{p}\left(\boldsymbol{Y}\right)| > \lambda\right\}$$

with threshold $\lambda = \widehat{\tau} \times \bar{\lambda}$, for some $\bar{\lambda} > 0$, will detect the change on the event

$$\left\{L_{\mathcal{G}(W,a)}^{\widehat{\tau}}\left(\boldsymbol{\zeta}\right) \leq \bar{\lambda}\right\} \cap \left\{\widehat{\tau} < 2\tau\right\} \tag{3.20}$$

as long as it holds that

$$\delta' > n^{\frac{2p^*}{2p^*+1}}\left(\frac{16C_{p,p^*}^2\tau^2\bar{\lambda}^2}{\Delta_{p^*}^2}\right)^{\frac{1}{2p^*+1}},$$

where

$$C_{p,p^*} = 2^{p^*+2}(p^*+2)\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}.$$

Proof. By the linearity of the difference operator and the triangle inequality the change will be detected if the following occurs:

$$\left|D_{1,m}^{p}\left(\boldsymbol{f}\right)\right| > \left|D_{1,m}^{p}\left(\boldsymbol{\zeta}\right)\right| + \lambda. \tag{3.21}$$

Moreover on (3.20) we must have that $\left|D_{1,m}^{p}\left(\boldsymbol{\zeta}\right)\right| + \lambda < 4\tau\bar{\lambda}$. Writing B_k for the k-th Bernoulli number we have the following by Faulhaber's formula for any integers $p > 0$ and $\delta' > 1$:

$$\begin{aligned}
\frac{1}{\delta'}\sum_{t=1}^{\delta'}\left(1 - t/\delta'\right)^{p} &= \left(\delta'\right)^{-(p+1)}\sum_{s=1}^{\delta'-1} s^{p} \\
&= \left(\frac{1}{p+1}\right)\left(\frac{\delta'-1}{\delta'}\right)^{p+1}\sum_{k=0}^{p}\binom{p+1}{k}B_k\left(\delta'-1\right)^{-k} \\
&\geq \left(\frac{1}{p+1}\right)\left(\frac{\delta'-1}{\delta'}\right)^{p+1} \\
&\geq \frac{1}{2^{p+1}(p+1)}.
\end{aligned} \tag{3.22}$$

Using the above along with Assumption 3.3.3 and the fact that the test statistic (3.2) is invariant to the addition of arbitrary degree p polynomials we have the following:

$$\begin{aligned}
&\left|D_{1,m}^{p}\left(\boldsymbol{f}\right)\right| \\
&= \left\{\delta'\sum_{i=0}^{p+1}\binom{p+1}{i}^{2}\right\}^{-\frac{1}{2}}\left|\sum_{j=0}^{p}\operatorname{sign}\left(\alpha_j - \beta_j\right)\Delta_j\sum_{t=1}^{\delta'}\left(\frac{t}{n} - \frac{\delta'}{n}\right)^{j}\right| \\
&\geq \sqrt{\delta'}\Delta_{p^*}\left(\frac{\delta'}{n}\right)^{p^*}\left[\frac{\frac{1}{\delta'}\sum_{t=1}^{\delta'}\left(1-\frac{t}{\delta'}\right)^{p^*}}{\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^{2}}}\right] - \sum_{\substack{0\leq j\leq p \\ j\neq p^*}}\sqrt{\delta'}\Delta_j\left(\frac{\delta'}{n}\right)^{j}\left[\frac{\frac{1}{\delta'}\sum_{t=1}^{\delta'}\left(1-\frac{1}{\delta'}\right)^{j}}{\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^{2}}}\right] \qquad (3.23) \\
&\geq C_{p,p^*}^{-1}\Delta_{p^*}\left(\delta'\right)^{\frac{2p^*+1}{2}}n^{-p^*}. \qquad (3.24)
\end{aligned}$$

Therefore combining (3.21) and (3.24) we have that on the event (3.20) the change will be detected if $C_{p,p^*}^{-1}\Delta_{p^*}\left(\delta'\right)^{\frac{2p^*+1}{2}}n^{-p^*} > 4\tau\bar{\lambda}$, and the desired result follows by rearranging. □

Theorem 3.6.1. *Put $w = \lfloor c\log(n)\rfloor$ for some constant $c > 0$ and introduce maximum of the local test statistics (3.2) appropriately standardised and restricted to scales w as follows:*

$$M^{\sigma}_{c\log(n)}(\mathbf{Y}) = \max\left\{\frac{1}{\sigma}D^{p}_{l,w}(\mathbf{Y}) \mid 1 \leq l \leq n-w\right\}.$$

Then under Assumption 3.2.1 and the null of no change points for any fixed $x \in \mathbb{R}$ the following holds, where $\mathfrak{a}_n$ and $\mathfrak{b}_n$ are defined as in Theorem 3.2.1:

$$\mathbb{P}\left(\mathfrak{a}_n M^{\sigma}_{c\log(n)}(\mathbf{Y}) - \mathfrak{b}_n \leq x\right) \sim \exp\left(-\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}\right).$$

Proof. Omitting dependence on x introduce notation

$$\mathfrak{u}_n = \sqrt{2\log(n)} + \left(-\frac{1}{2}\log\log(n) - \log\left(2\sqrt{\pi}\right) + x\right)/\sqrt{2\log(n)}.$$

For some $\rho \in (0,1)$ we decompose the index set $\{1,\ldots,n\}$ into disjoint blocks $A_0, B_0, A_1, B_1, \ldots$ respectively of size w and w^{ρ} as

$$\begin{aligned} A_i &= \{l \mid i\left(w + w^{\rho}\right) < l \leq (i+1)w + iw^{\rho}\}, \\ B_i &= \{l \mid (i+1)w + iw^{\rho} < l \leq (i+1)(w + w^{\rho})\}. \end{aligned}$$

The proof proceeds in three steps.

STEP 1: we first show that the behavior of small blocks is asymptotically unimportant for the maximum. Putting $\mathcal{B}_n = \cup_i B_i$ and using the fact $|\mathcal{B}_n| \sim nw^{\rho}/(w+w^{\rho})$ and $\mathfrak{u}_n^2 = 2\log(n) - \log\log(n) + \mathcal{O}(1)$ it follows that

$$\begin{aligned} \mathbb{P}\left(\max_{l\in\mathcal{B}_n}\left\{\frac{1}{\sigma}D^{p}_{l,w}(\mathbf{Y})\right\} > \mathfrak{u}_n\right) &\leq \sum_{l\in\mathcal{B}_n}\mathbb{P}\left(\frac{1}{\sigma}D^{p}_{l,w}(\mathbf{Y}) > \mathfrak{u}_n\right) \\ &= |\mathcal{B}_n|\,\bar{\Phi}\left(\mathfrak{u}_n\right) \\ &\lesssim \frac{w^{\rho}}{w+w^{\rho}}. \end{aligned}$$

STEP 2: next we show that the any dependence between larger blocks is asymptotically unimportant for the the maximum. Write

$$\Lambda_{l_1,l_2} = \operatorname{Cov}\left(\frac{1}{\sigma}D^p_{l_1,w}\left(\mathbf{Y}\right), \frac{1}{\sigma}D^p_{l_2,w}\left(\mathbf{Y}\right)\right),$$

and let $\sigma^{-1}\tilde{D}^p_{l,w}\left(\mathbf{Y}\right)$ be random variables with the same marginal distributions as $\sigma^{-1}D^p_{l,w}\left(\mathbf{Y}\right)$ and covariances given by

$$\tilde{\Lambda}_{l_1,l_2} = \begin{cases} \Lambda_{l_1,l_2} & l_1 \in A_{i_1}, l_2 \in A_{i_2} \text{ with } i_1 = i_2 \\ 0 & \text{else} \end{cases}.$$

For any l_1, l_2 write $j_{1,2} = \left|\{l_1, \ldots, l_1 + w - 1\} \cap \{l_2, \ldots, l_2 + w - 1\}\right|$ and put $\Lambda_{l_1,l_2} = \Lambda_{j_{1,2}}$. Writing $\mathcal{A}_n = \cup_i A_i$ and using Lemma 3.6.1 we have the following:

$$\begin{aligned}
&\left|\mathbb{P}\left(\max_{l\in\mathcal{A}_n}\left\{\frac{1}{\sigma}D^p_{l,w}\left(\mathbf{Y}\right)\right\} \leq \mathfrak{u}_n\right) - \mathbb{P}\left(\max_{l\in\mathcal{A}_n}\left\{\frac{1}{\sigma}\tilde{D}^p_{l,w}\left(\mathbf{Y}\right)\right\} \leq \mathfrak{u}_n\right)\right| \\
&\leq \frac{1}{2\pi}\sum_{\substack{l_1\in A_i, l_2\in A_j \\ i\neq j}} \left|\Lambda_{l_1,l_2} - \tilde{\Lambda}_{l_1,l_2}\right| \left(1 - \Lambda^2_{l_1,l_2}\right)^{-1/2} \exp\left(-\frac{\mathfrak{u}_n^2}{1 + \Lambda_{l_1,l_2}}\right) \\
&\lesssim \sum_{i=0}^{|\mathcal{A}_n|/|A_0|} \sum_{\substack{l_1\in A_i \\ l_2\in A_{i+1}}} \left|\Lambda_{l_1,l_2} - \tilde{\Lambda}_{l_1,l_2}\right| \left(1 - \Lambda^2_{l_1,l_2}\right)^{-1/2} \exp\left(-\frac{\mathfrak{u}_n^2}{1 + \Lambda_{l_1,l_2}}\right) \\
&\lesssim \frac{|\mathcal{A}_n|}{|A_0|} \sum_{l=1}^{|A_0|}\sum_{j=1}^{l} \Lambda_j \left(1 - \Lambda_j^2\right)^{-1/2} \exp\left(-\frac{2\log(n) - \log\log(n)}{1 + \Lambda_j}\right) \\
&\lesssim \log(n)\frac{|\mathcal{A}_n|}{|A_0|} \sum_{l=1}^{|A_0|}\sum_{j=1}^{l} \Lambda_j \left(1 - \Lambda_j^2\right)^{-1/2} \exp\left(-\frac{2\log(n)}{1 + \Lambda_j}\right).
\end{aligned} \tag{3.25}$$

Note that for some fixed $K > 0$ depending on p it must hold that $\Lambda_j \leq \min\left(jK, w - w^{\rho}\right)/w$.

Therefore the first term after the double sum can be bounded as follows:

$$\begin{aligned}\Lambda_j\left(1-\Lambda_j^2\right)^{-1/2} &\leq \Lambda_j\left(1-\Lambda_j\right)^{-1/2}\\ &\leq \min\left(jK, w-w^\rho\right)/\sqrt{\left(w-\min\left(jK,w-w^\rho\right)\right)w}\\ &\leq \min\left(jK,w-w^\rho\right)/\sqrt{w}. \end{aligned} \tag{3.26}$$

For the exponential term put $2/(1+\Lambda_j) = 1+\delta_j$. The following holds:

$$\begin{aligned}\delta_j &= \left(1-\Lambda_j\right)/\left(1+\Lambda_j\right)\\ &\geq \left(w-\min\left(jK,w-w^\rho\right)\right)/\left(w+\min\left(jK,w-w^\rho\right)\right)\\ &\geq \left(w-\min\left(jK,w-w^\rho\right)\right)/2w. \end{aligned} \tag{3.27}$$

Therefore substituting (3.26) and (3.27) into (3.25) we obtain

$$\begin{aligned}(3.25) &\lesssim \frac{\sqrt{\log(n)}}{n}\frac{|\mathcal{A}_n|}{|A_0|}\sum_{j=1}^{l}\min\left(jK,w-w^\rho\right)\left(n^{\frac{1}{2w}}\right)^{-(w-\min(jK,w-w^\rho))},\\ &= \frac{\sqrt{\log(n)}}{n}\frac{|\mathcal{A}_n|}{|A_0|}\left\{\sum_{l=1}^{\lfloor|A_0|/K\rfloor}\sum_{j=1}^{l}jK\left(n^{\frac{1}{2w}}\right)^{-(w-jK)}+\sum_{l=\lfloor|A_0|/K\rfloor+1}^{|A_0|}\sum_{j=1}^{l}\left(w-w^\rho\right)\left(n^{\frac{1}{2w}}\right)^{w^\rho}\right\}. \end{aligned} \tag{3.28}$$

The first sum in (3.28) can be bounded as follows:

$$\begin{aligned}\sum_{l=1}^{\lfloor|A_0|/K\rfloor}\sum_{j=1}^{l}jK\left(n^{\frac{1}{2w}}\right)^{-(w-jK)} &\lesssim n^{-1/2}\int_1^{\lfloor|A_0|/K\rfloor+1}\int_1^{y+1}x\left(n^{\frac{1}{2w}}\right)^{Kx}\mathrm{d}x\mathrm{d}y\\ &\lesssim wn^{-w^{-(1-\rho)}/2}. \end{aligned} \tag{3.29}$$

The second sum in (3.28) can be bounded as follows:

$$\begin{aligned}\sum_{l=\lfloor |A_0|/K\rfloor+1}^{|A_0|} \sum_{j=1}^{l} (w-w^{\rho})\left(n^{\frac{1}{2w}}\right)^{w^{\rho}} &\lesssim wn^{-w^{-(1-\rho)}/2} \sum_{l=\lfloor |A_0|/K\rfloor+1}^{|A_0|} (l) \\ &\lesssim w^3 n^{-w^{-(1-\rho)}/2}. \end{aligned} \tag{3.30}$$

Finally plugging (3.29) and (3.30) into (3.25) and using the fact that $|\mathcal{A}_n|/|A_0| \sim n/(w+w^{\rho})$ we obtain the following for some $C>0$ depending on ρ as long as n is sufficiently large:

$$(3.25) \lesssim \frac{\sqrt{\log(n)}}{n}\frac{|\mathcal{A}_n|}{|A_0|}\left\{n^{-w^{-(1-\rho)}/2}\left(w+w^3\right)\right\} \lesssim \log^{5/2}(n)n^{-w^{-(1-\rho)}/2} \lesssim \exp\left(-C\log^{\rho}(n)\right).$$

STEP 3: we now prove Theorem 3.6.1. Using Lemma 3.6.4 and part (i) of Lemma 3.6.6 and noting that $\mathfrak{u}_n^2/w \sim 2/c$ gives the following:

$$\mathbb{P}\left(\max_{l\in A_i}\left\{\frac{1}{\sigma}D_{l,w}^{p}\left(\mathbf{Y}\right)\right\} > \mathfrak{u}_n\right) \sim \left(\frac{w}{n}\right)\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}, \qquad i=0,\dots,|\mathcal{A}_n|-1. \tag{3.31}$$

The following inequality is evident:

$$\mathbb{P}\left(M_{c\log(n)}^{\sigma}\left(\mathbf{Y}\right)\leq \mathfrak{u}_n\right) \leq \mathbb{P}\left(\max_{l\in\mathcal{A}_n}\left\{\frac{1}{\sigma}D_{l,w}^{p}\left(\mathbf{Y}\right)\right\}\leq \mathfrak{u}_n\right).$$

Therefore (3.31), the results of step 2, and that $|\mathcal{A}_n|/|A_0| \sim n/w$ imply that

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{P}\left(M_{c\log(n)}^{\sigma}\left(\mathbf{Y}\right)\leq \mathfrak{u}_n\right) &\leq \lim_{n\to\infty}\left\{\mathbb{P}\left(\max_{l\in\mathcal{A}_n}\left\{\frac{1}{\sigma}\tilde{D}_{l,w}^{p}\left(\mathbf{Y}\right)\right\}\leq \mathfrak{u}_n\right)+\mathcal{O}\left(\exp\left(-C\log^{\rho}(n)\right)\right)\right\} \\ &= \lim_{n\to\infty}\left(1-\left(\frac{w}{n}\right)\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}\right)^{|\mathcal{A}_n|/|A_0|} \\ &= \exp\left(-\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}\right).\end{aligned}$$

Going the other way the following inequality is also evident:

$$\mathbb{P}\left(M_{c\log(n)}^{\sigma}\left(\mathbf{Y}\right)\leq \mathfrak{u}_n\right) \geq \mathbb{P}\left(\max_{l\in\mathcal{A}_n}\left\{\frac{1}{\sigma}D_{l,w}^{p}\left(\mathbf{Y}\right)\right\}\leq \mathfrak{u}_n\right) - \mathbb{P}\left(\max_{l\in\mathcal{B}_n}\left\{\frac{1}{\sigma}D_{l,w}^{p}\left(Y\right)\right\} > \mathfrak{u}_n\right).$$

Using (3.31) and the results of Steps 1 and 2 gives that

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}\left(M^{\sigma}_{c\log(n)}(\mathbf{Y}) \leq \mathfrak{u}_n\right) \\
&\geq \lim_{n\to\infty}\left\{\mathbb{P}\left(\max_{l\in\mathcal{A}_n}\left\{\frac{1}{\sigma}\tilde{D}^p_{l,w}(Y)\right\} \leq \mathfrak{u}_n\right) - \mathcal{O}\left(\exp\left(-C\log^{\rho}(n)\right)\right) - \mathcal{O}\left(\frac{w^{\rho}}{w+w^{\rho}}\right)\right\} \\
&= \lim_{n\to\infty}\left(1-\left(\frac{w}{n}\right)\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}\right)^{|\mathcal{A}_n|/|A_0|} \\
&= \exp\left(-\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}\right).
\end{aligned}$$

Therefore, the theorem is proved. □

3.6.3 Proof of Theorem 3.2.1

Proof. Given the result in part (i), part (ii) follows immediately from Lemma 3.6.2. For the proof of part (i) write $k_n = \lfloor \log_a(W) \rfloor$ and for some $A > 0$ introduce the restrictions of the a-adic grid defined in (3.5) to scales no larger than Wa^A:

$$\begin{aligned}
\mathcal{G}_-(A) &= \left\{(l,w)\in\mathbb{N}^2 \mid w\in\mathcal{W}_-(A), 1\leq l\leq n-w\right\}, \\
\mathcal{W}_-(A) &= \left\{w = \left\lfloor a^k \right\rfloor \mid k_n \leq k \leq k_n + A\right\}.
\end{aligned}$$

Introduce also the restriction of (3.5) to scales strictly larger than Wa^A:

$$\begin{aligned}
\mathcal{G}_+(A) &= \left\{(l,w)\in\mathbb{N}^2 \mid w\in\mathcal{W}_+(A), 1\leq l\leq n-w\right\}, \\
\mathcal{W}_+(A) &= \left\{w = \left\lfloor a^k \right\rfloor \mid k_n + A < k \leq \lfloor \log_a(n/2) \rfloor\right\}.
\end{aligned}$$

The proof proceeds in four steps.

STEP 1: we first show that the behaviour of the tests statistic on large scales is asymptotically unimportant for the maximum. Making use of lemma 3.6.3 we have that

$$\begin{aligned}
&\mathbb{P}\left(\max_{(l,w)\in\mathcal{G}_+(A)}\left\{\frac{1}{\sigma}D^p_{l,w}\left(\mathbf{Y}\right)\right\}>\mathfrak{u}_n\right)\\
&\le\sum_{k=k_n+A}^{\lfloor\log_a(n/2)\rfloor}\sum_{i=0}^{\lfloor n/a^k\rfloor-1}\mathbb{P}\left(\max\left\{\frac{1}{\sigma}D^p_{l,\lfloor a^k\rfloor}\left(\mathbf{Y}\right)\mid i\times\left\lfloor a^k\right\rfloor<l\le(i+1)\times\left\lfloor a^k\right\rfloor\right\}>\mathfrak{u}_n\right)\\
&\le\sum_{k=k_n+A}^{\lfloor\log_a(n/2)\rfloor}\left(\frac{n}{a^k}\right)\mathbb{P}\left(\sup_{t\in[0,1)}\left\{\xi\left(t\right)\right\}>\mathfrak{u}_n\right)\\
&\lesssim\sum_{k=k_n+A}^{\lfloor\log_a(n/2)\rfloor}\left(\frac{n}{a^k}\right)\mathfrak{u}_ne^{-\mathfrak{u}_n^2/2}\\
&\lesssim\frac{a^{-A}}{1-a^{-1}}.
\end{aligned}$$

Finally, sending $A\to\infty$ the claim is proved.

STEP 2: next we show that for any fixed A the dependence between maxima occurring over different scales in $\mathcal{W}_-(A)$ is asymptotically unimportant for the overall maximum. Write

$$\Lambda_{l_1,w_1,l_2,w_2}=\mathrm{Cov}\left(\frac{1}{\sigma}D^p_{l_1,w_1}(\mathbf{Y}),\frac{1}{\sigma}D^p_{l_2,w_2}(\mathbf{Y})\right),$$

and let $\sigma^{-1}\tilde{D}^{(p)}_{l,w}\left(\mathbf{Y}\right)$ be random variables with the same marginal distribution as $\sigma^{-1}D^p_{l,w}\left(\mathbf{Y}\right)$ and covariance given by

$$\tilde{\Lambda}_{l_1,l_2,w_1,w_2}=\begin{cases}\Lambda_{l_1,l_2,w_1,w_2} & \text{if } w_1=w_2\\ 0 & \text{else}\end{cases}.$$

Note that for each $a>1$ there will be a $\Lambda_a\in(0,1)$ depending only on a such that for any $w_1\neq w_2$ and all permissible l_1,l_2 it holds that $\Lambda_{l_1,w_1,l_2,w_2}\le\Lambda_a$. Therefore using Lemma

3.6.1 we have the following:

$$
\begin{aligned}
&\left|\mathbb{P}\left(\max_{(l,w)\in\mathcal{G}_-(A)}\left\{\frac{1}{\sigma}D^p_{l,w}\left(\mathbf{Y}\right)\right\}\leq\mathfrak{u}_n\right)-\mathbb{P}\left(\max_{(l,w)\in\mathcal{G}_-(A)}\left\{\frac{1}{\sigma}\tilde{D}^p_{l,w}\left(\mathbf{Y}\right)\right\}\leq\mathfrak{u}_n\right)\right|\\
&\leq\frac{1}{2\pi}\sum_{\substack{w_1,w_2\in\mathcal{W}_-(A)\\ w_1\neq w_2}}\ \sum_{\substack{1\leq l_1\leq n-w_1\\ 1\leq l_2\leq n-w_2}}\left|\Lambda_{\substack{l_1,w_1\\ l_2,w_2}}-\tilde{\Lambda}_{\substack{l_1,w_1\\ l_2,w_2}}\right|\left(1-\Lambda^2_{\substack{l_1,w_1\\ l_2,w_2}}\right)^{-1/2}\exp\left(-\frac{\mathfrak{u}_n^2}{1+\Lambda^2_{\substack{l_1,w_1\\ l_2,w_2}}}\right)\\
&\lesssim\sum_{\substack{w_1,w_2\in\mathcal{W}_-(A)\\ w_1\neq w_2}}\ \sum_{\substack{1\leq l_1\leq n-w_1\\ 1\leq l_2\leq n-w_2\\ |l_1-l_2|<\max(w_1,w_2)}}\Lambda_{\substack{l_1,w_1\\ l_2,w_2}}\left(1-\Lambda^2_{\substack{l_1,w_1\\ l_2,w_2}}\right)^{-1/2}\exp\left(-\frac{\mathfrak{u}_n^2}{1+\Lambda^2_{\substack{l_1,w_1\\ l_2,w_2}}}\right)\\
&\lesssim\log(n)\sum_{\substack{w_1,w_2\in\mathcal{W}_-(A)\\ w_1\neq w_2}}\ \sum_{\substack{1\leq l_1\leq n-w_1\\ 1\leq l_2\leq n-w_2\\ |l_1-l_2|<\max(w_1,w_2)}}\left(\frac{\Lambda_a}{\sqrt{1-\Lambda_a^2}}\right)\exp\left(-\frac{2\log(n)}{1+\Lambda_a}\right)\\
&\lesssim(1+A)^2\,a^A\log^2(n)\times n^{-\frac{1-\Lambda_a}{1+\Lambda_a}}.
\end{aligned}
$$

Since $\Lambda_a<1$ the statement is proved.

STEP 3: we now show that if we pass to a sub-sequence of n's on which the quantity $b_n=a^{\lfloor\log_a(W)\rfloor}/W$ converges to some constant b the sequence of normalised maxima

$$
\left\{\mathfrak{a}_nM^{\sigma}_{\mathcal{G}(W,a)}\left(\mathbf{Y}\right)-\mathfrak{b}_n\mid n\in\mathbb{N}\right\}\tag{3.32}
$$

converges weakly to a Gumbel distribution. On such a sub-sequence for each $j\in\mathbb{N}$ we have that $a^{k_n+j}\sim a^jbd\times\log(n)$. Therefore from Theorem 3.6.1 we have the following:

$$
\mathbb{P}\left(\max_{1\leq l\leq n-\lfloor a^{k_n+j}\rfloor}\left\{\frac{1}{\sigma}D^p_{l,\lfloor a^{k_n+j}\rfloor}\left(\mathbf{Y}\right)\right\}\leq\mathfrak{u}_n\right)\sim\exp\left(-\left(\frac{2C_p}{a^jbd}\right)F\left(\frac{2C_p}{a^jbd}\right)e^{-\tau}\right).
$$

The following inequality is evident:

$$
\mathbb{P}\left(M^{\sigma}_{\mathcal{G}(W,a)}\left(\mathbf{Y}\right)\leq\mathfrak{u}_n\right)\leq\mathbb{P}\left(\max_{(l,w)\in\mathcal{G}_-(A)}\left\{\frac{1}{\sigma}D^p_{l,w}\left(\mathbf{Y}\right)\right\}\leq\mathfrak{u}_n\right).
$$

Therefore (3.6.3) and the result from step 2 imply that

$$\limsup_{n\to\infty} \mathbb{P}\left(M^{\sigma}_{\mathcal{G}(W,a)}(\mathbf{Y}) \leq \mathfrak{u}_n\right) \leq \exp\left(-\sum_{j=0}^{\infty}\left(\frac{2C_p}{a^j bd}\right) F\left(\frac{2C_p}{a^j bd}\right) e^{-x}\right).$$

Note that because $a > 1$ by part (ii) of Lemma 3.6.5 the above sum converges. Going the other way the following inequality is also evident:

$$\mathbb{P}\left(M^{\sigma}_{\mathcal{G}(W,a)}(\mathbf{Y}) \leq \mathfrak{u}_n\right) \geq \mathbb{P}\left(\max_{(l,w)\in\mathcal{G}_-(A)}\left\{\frac{1}{\sigma}D^p_{l,w}(\mathbf{Y})\right\} \leq \mathfrak{u}_n\right) - \mathbb{P}\left(\max_{(l,w)\in\mathcal{G}_+(A)}\left\{\frac{1}{\sigma}D^p_{l,w}(\mathbf{Y})\right\} > \mathfrak{u}_n\right).$$

Therefore (3.6.3) and the result from steps 1 and 2 imply that

$$\liminf_{n\to\infty} \mathbb{P}\left(M^{\sigma}_{\mathcal{G}(W,a)}(\mathbf{Y}) \leq \mathfrak{u}_n\right) \geq \exp\left(-\sum_{j=0}^{\infty}\left(\frac{2C_p}{a^j bd}\right) F\left(\frac{2C_p}{a^j bd}\right) e^{-x}\right).$$

Therefore, the statement is proved.

STEP 4: we now prove the result in part (i). Since b_n may have any sub-sequential limit between $1/a$ and 1 it follows from step 4 that the sequence of random variables (3.32) is tight. Using part (i) of Lemma 3.6.5 the constants in (3.7) are easily recognised as the largest and smallest constants which may appear in the extreme value limit.

□

3.6.4 Proof of Theorem 3.2.2

Proof. With W satisfying Assumption 3.2.5 and omitting dependence on x introduce the notation

$$\mathfrak{u}_{n,W} = \sqrt{2\log(n/W)} + \left(\frac{1}{2}\log\log(n/W) - \log(\sqrt{\pi}) + x\right)/\sqrt{2\log(n/W)}.$$

We first investigate the be behaviour of local test statistics (3.2) restricted to a particular scale of the order $\mathcal{O}(W)$ under the null of no change points. For some $c > 0$ put $w = \lfloor cW \rfloor$, and write

$$M^{\tau}_{cW}\left(\mathbf{Y}\right) = \max\left\{\frac{1}{\tau} D^{p}_{l,w}\left(\mathbf{Y}\right) \mid 1 \le l \le n-w\right\}.$$

Putting $\mathbf{B} = (B(1), \ldots, B(n))'$, where $\{B(t)\}_{t>0}$ is the process introduced in Assumption 3.2.4, making use of Assumption 3.2.4 the following holds:

$$M^{\tau}_{n,W}\left(\mathbf{Y}\right) = \max\left\{D^{p}_{l,w}\left(\mathbf{B}\right) \mid 1 \le l \le n-w\right\} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{n^{\frac{2}{2+\nu}}/W}\right). \tag{3.33}$$

Moreover, using Lemma 3.6.3 and arguing as in the proof of Theorem 3.6.1, or equivalently directly applying Theorem 12.3.5 in Leadbetter et al. (2012), it holds that

$$\begin{aligned} \mathbb{P}\left(M^{1}_{cW}\left(\mathbf{B}\right) \le \mathfrak{u}_{n,W}\right) &\sim \prod_{i=0}^{\lfloor n/w \rfloor} \mathbb{P}\left(\max\left\{D^{p}_{l,w}\left(\mathbf{B}\right) \mid i \times w < l \le (i+1) \times w\right\} \le \mathfrak{u}_{n,W}\right) \\ &\sim \left[1 - \mathbb{P}\left(\sup_{l \in [0,1)} \left\{\xi\left(l\right)\right\} > \mathfrak{u}_{n,W}\right)\right]^{\lfloor n/w \rfloor} \\ &\sim \exp\left(-\frac{C_p}{c} e^{-x}\right). \end{aligned} \tag{3.34}$$

Therefore, combining (3.33) and (3.34) and arguing as in the proof of Theorem 3.2.1, we immediately have that

$$\mathbb{P}\left(M^{\tau}_{cW}\left(\mathbf{Y}\right) \le \mathfrak{u}_{n,W}\right) \sim \exp\left(-\frac{C_p}{c} e^{-x}\right).$$

On a sub-sequence of n's for which the quantity $b_n = a^{\lfloor \log_a(W) \rfloor}/W$ converges to some constant b, arguing as in the proof of Theorem 3.2.1, we therefore have under the null of no change points that

$$\mathbb{P}\left(M^{\tau}_{\mathcal{G}(W,a)}\left(\mathbf{Y}\right) \le \mathfrak{u}_{n,W}\right) \to \exp\left(-\left(\frac{b^{-1}C_p}{1-a^{-1}}\right) e^{-x}\right).$$

However, it is again clear that b_n can have any sub-sequential limit between a^{-1} and 1, so part (i) of the theorem is proved. Part (ii) again follows from Lemma 3.6.2. □

3.6.5 Proof of Lemma 3.3.1

Proof. Putting $m = (n-p-1)\,/\,(p+1)$, we will prove that

$$\mathbb{P}\left(|\widehat{\sigma}_{\mathrm{MAD}} - \sigma| > \delta\right) \leq 2(p+1)\exp\left(-2m\left[(3/2)\times(\delta/\sigma) - N/m\right]^2\right)$$

from which the lemma is evident. We only show the upper bound for the above inequality as the lower bound can be derived analogously. For simplicity assume $n-(p+1)$ is a multiple of $(p+1)$ and introduce the following sets:

$$\begin{aligned}
&I_j = \{p+1 \leq t \leq n \mid (t+j) \bmod (p+1) = 0\}\,,\\
&I_\eta = \cup_{k=1}^N \{\eta_k, \ldots, \eta_k + (p+1)\}\,,\\
&I_{j,1} = I_j \setminus I_\eta,\\
&I_{j,2} = I_j \cap I_\eta.
\end{aligned}$$

Introducing also the random variables $B_t^\delta = \mathbf{1}\left\{|X_t| > \Phi^{-1}(3/4)\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}\,[\sigma+\delta]\right\}$ and put $p_\delta = \mathbb{E}\left(B_t^\delta \mid t \notin I_\eta\right)$. The following holds via Hoeffding's inequality:

$$\begin{aligned}
\mathbb{P}\left(\widehat{\sigma}_{\mathrm{MAD}} - \sigma > \delta\right) &= \mathbb{P}\left(\frac{\mathrm{median}\left\{|X_{p+1}|, \ldots, |X_n|\right\}}{\Phi^{-1}\left(3/4\right)\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}} > \sigma + \delta\right)\\
&\leq \sum_{j=0}^{p}\mathbb{P}\left(\sum_{t\in I_{j,1}} B_t^\delta + \sum_{t\in I_{j,2}} B_t^\delta > \frac{n-(p+1)}{2(p+1)}\right)\\
&\leq \sum_{j=0}^{p}\mathbb{P}\left(\sum_{t\in I_{j,1}}\left(B_t^\delta - p_\delta\right) > \frac{n-(p+1)}{2(p+1)} - |I_{j,2}| - p_\delta\,|I_{j,1}|\right)\\
&\leq (p+1)\exp\left(-\frac{2\left[m\left(1/2 - p_\delta\right) - N\right]^2}{m-N}\right)\\
&\leq (p+1)\exp\left(-2m\left[1/2 - p_\delta - N/m\right]^2\right). \qquad (3.35)
\end{aligned}$$

Turning to p_δ we have the following bound where we put $Z \sim \mathcal{N}(0,1)$:

$$\begin{aligned}
p_\delta &= \mathbb{P}\left(|Z| > \Phi^{-1}(3/4)\left[1+\delta/\sigma\right]\right) \\
&= 2\left(1 - \int_{-\infty}^{\Phi^{-1}(3/4)} \phi\left(x\left[1+\delta/\sigma\right]\right) \mathrm{d}x \left[1+\delta/\sigma\right]\right) \\
&\geq 2\left(1 - \Phi\left(\Phi^{-1}(3/4)\right)\left[1+\delta/\sigma\right]\right). \\
&= 1/2 - (3/2) \times (\delta/\sigma) \qquad (3.36)
\end{aligned}$$

Substituting (3.36) into (3.35) we obtain the desired result.

□

3.6.6 Proof of Lemma 3.3.2

Proof. Write $\gamma_i = \max_{1\leq t\leq n} \mathbb{E}\left|\zeta_t/\sigma\right|^i$ for each $i = 2, 3$ and put $\boldsymbol{D}_p = \tilde{\boldsymbol{D}}_p' \tilde{\boldsymbol{D}}_p$ where $\tilde{\boldsymbol{D}}_p$ is the $n \times n$ difference matrix such that each entry in the vector $\boldsymbol{D}_p \boldsymbol{x}$ is the $(p+1)$-th difference of the corresponding entry in the n-vector $\boldsymbol{x}$ scaled by

$$1/\sqrt{\sum_{i=0}^{p+1} \binom{p+1}{i}^2}.$$

Writing $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\zeta}$ the equation below follows directly from equation (6) in Dette et al. (1998).

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{\sigma}^2_{\mathrm{DIF}} - \sigma^2\right|^2\right] \leq \Big[&\left(\boldsymbol{f}'\boldsymbol{D}_p\boldsymbol{f}\right)^2 + 4\sigma^2\boldsymbol{f}'\boldsymbol{D}_p^2\boldsymbol{f} + 4\boldsymbol{f}'\left(\boldsymbol{D}_p\mathrm{diag}\left(\boldsymbol{D}_p\right)\mathbf{1}\right)\sigma^3\gamma_3 \\
&+ \sigma^4\mathrm{trace}\left\{\mathrm{diag}\left(\boldsymbol{D}_p\right)^2\right\}(\gamma_4 - 3) + 2\sigma^4\mathrm{trace}\left(\boldsymbol{D}_p^2\right)\Big] / (n-p-1)^2.
\end{aligned}$$

Since the noise terms have bounded fourth moment and function $f_\circ(\cdot)$ is assumed to be bounded the it must hold that

$$\begin{aligned}
\sigma^4\mathrm{trace}\left\{\mathrm{diag}\left(\boldsymbol{D}_p\right)^2\right\}(\gamma_4 - 3) + 2\sigma^4\mathrm{trace}\left(\boldsymbol{D}_p^2\right) = \mathcal{O}\left(n\right), \\
\left(\boldsymbol{f}'\boldsymbol{D}_p\boldsymbol{f}\right)^2 + 4\sigma^2\boldsymbol{f}'\boldsymbol{D}_p^2\boldsymbol{f} + 4\boldsymbol{f}'\left(\boldsymbol{D}_p\mathrm{diag}\left(\boldsymbol{D}_p\right)\mathbf{1}\right)\sigma^3\gamma_3 = \mathcal{O}\left(N^2\right).
\end{aligned}$$

It therefore follows that

$$\mathbb{E}\left[\left|\widehat{\sigma}^2_{\mathrm{DIF}} - \sigma^2\right|^2\right] \leq \mathcal{O}\left(\frac{1}{n} \vee \frac{N^2}{n^2}\right),$$

and as such the desired result follows by Chebyshev's inequality. □

3.6.7 Proof of Lemma 3.3.3

Proof. Write $\bar{\boldsymbol{Y}} = \left(\bar{Y}_{1,W'}, \ldots, \bar{Y}_{\lfloor n/W' \rfloor, W'}\right)'$ and let $\bar{\boldsymbol{f}}$ and $\bar{\boldsymbol{\zeta}}$ be defined analogously. Let $\boldsymbol{D_p}$ be as defined in the proof of the last lemma, with its dimensions suitably adjusted. Finally put $m = \lfloor n/W' \rfloor - (p+1)$. We can therefore write $\hat{\tau}^2_{\mathrm{DIF}} = \frac{1}{mW}\bar{\boldsymbol{Y}}'\boldsymbol{D_p}\bar{\boldsymbol{Y}}$, and the absolute difference between our estimator and the truth can be bounded as follows:

$$\begin{aligned}
\left|\hat{\tau}^2_{\mathrm{DIF}} - \tau^2\right| &= \left|\frac{1}{mW'}\left(\bar{\boldsymbol{f}} + \bar{\boldsymbol{\zeta}}\right)' \boldsymbol{D_p}\left(\bar{\boldsymbol{f}} + \bar{\boldsymbol{\zeta}}\right) - \tau^2\right| \\
&\lesssim \left|\frac{1}{mW'}\bar{\boldsymbol{\zeta}}'\boldsymbol{D_p}\bar{\boldsymbol{\zeta}} - \frac{1}{mW'}\mathbb{E}\left(\bar{\boldsymbol{\zeta}}'\boldsymbol{D_p}\bar{\boldsymbol{\zeta}}\right)\right| + \left|\frac{1}{mW'}\mathbb{E}\left(\bar{\boldsymbol{\zeta}}'\boldsymbol{D_p}\bar{\boldsymbol{\zeta}}\right) - \tau^2\right| \\
&\quad + \frac{1}{mW'}\left|\bar{\boldsymbol{f}}'\boldsymbol{D_p}\bar{\boldsymbol{f}}\right| + \frac{1}{mW'}\left|\bar{\boldsymbol{f}}'\boldsymbol{D_p}\bar{\boldsymbol{\zeta}}\right| \\
&= T_1 + T_2 + T_3 + T4.
\end{aligned}$$

We now bound each of the terms in turn. Introducing the notation

$$\psi_{p,j} = (-1)^{p+1-j}\binom{p+1}{j} \Big/ \sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}.$$

We can therefore write

$$\begin{aligned}
\frac{1}{mW'}\bar{\boldsymbol{\zeta}}'\boldsymbol{D_p}\bar{\boldsymbol{\zeta}} &= \frac{1}{m}\sum_{s=p+2}^{\lfloor n/W' \rfloor}\left(\sum_{j=0}^{p+1}\psi_{p,j}\left(\bar{\zeta}_{s-j,W'}/\sqrt{W'}\right)\right)^2 \\
&= \frac{1}{m}\sum_{s=p+2}^{\lfloor n/W' \rfloor}\left[\sum_{j=0}^{p+1}\psi^2_{p,j}\left(\bar{\zeta}_{s-j,W'}/\sqrt{W'}\right)^2 + \sum_{\substack{k \neq l \\ 0 \leq k,l \leq p+1}}\psi_{p,k}\psi_{p,l}\left(\bar{\zeta}_{s-k,W'}/\sqrt{W'}\right)\left(\bar{\zeta}_{s-l,W'}/\sqrt{W'}\right)\right].
\end{aligned}$$

From which it follows that

$$\begin{aligned}\frac{1}{mW'}\mathbb{E}\left(\boldsymbol{\bar{\zeta}'}\boldsymbol{D_p}\boldsymbol{\bar{\zeta}}\right) &= \sum_{j=0}^{p+1}\psi_{p,j}^2\left(\gamma_0 + 2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_h\right)\\ &\qquad + \sum_{\substack{k\neq l\\ 0\leq k,l\leq p+1}}\psi_{p,k}\psi_{p,l}\left(\gamma_{W'|k-l|} + 2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_{W'|k-l|+h}\right).\end{aligned}$$

Using these facts term T_1 can be bounded as follows

$$\begin{aligned}T_1 &\leq \left|\frac{1}{m}\sum_{s=p+2}^{\lfloor n/W'\rfloor}\sum_{j=0}^{p+1}\psi_{p,j}^2\left(\left(\bar{\zeta}_{s-j,W'}/\sqrt{W'}\right)^2 - \gamma_0 - 2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_h\right)\right|\\ &+\left|\frac{1}{m}\sum_{s=p+2}^{\lfloor n/W'\rfloor}\sum_{\substack{k\neq l\\ 0\leq k,l\leq p+1}}\psi_{p,k}\psi_{p,l}\left(\left(\bar{\zeta}_{s-k,W}/\sqrt{W'}\right)\left(\bar{\zeta}_{s-l,W'}/\sqrt{W'}\right)\right.\right.\\ &\qquad\left.\left. -\ \gamma_{W'|k-l|} - 2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_{W'|k-l|+h}\right)\right|\\ &= T_{1,1} + T_{1,2}.\end{aligned}$$

For the first term we have that

$$\begin{aligned}
T_{1,1} &= \left| \frac{1}{m} \sum_{s=p+2}^{\lfloor n/W' \rfloor} \sum_{j=0}^{p+1} \psi_{p,j}^2 \left(\frac{1}{W} \sum_{t=W'(s-j-1)+1}^{W'(s-j)} \zeta_t^2 \right.\right. \\
&\qquad \left.\left. + \frac{2}{W'} \sum_{h=1}^{W'-1} \sum_{t=W'(s-j-1)+1}^{W'(s-j)-h} \zeta_t \zeta_{t+h} - \gamma_0 - 2 \sum_{h=1}^{W'-1} \left(1 - \frac{h}{W'}\right) \gamma_h \right) \right| \\
&= \left| \frac{1}{m} \sum_{s=p+2}^{\lfloor n/W' \rfloor} \sum_{j=0}^{p+1} \psi_{p,j}^2 \left(\frac{1}{W'} \sum_{t=W'(s-j-1)+1}^{W'(s-j)} \left(\zeta_t^2 - \gamma_0\right) \right.\right. \\
&\qquad \left.\left. + \sum_{h=1}^{W'-1} \frac{1}{(W'-h)} \sum_{t=W'(s-j-1)+1}^{W'(s-j)-h} \left(1 - \frac{h}{W'}\right) \left(\zeta_t \zeta_{t+h} - \gamma_h\right) \right) \right| \\
&\leq \sum_{j=0}^{p+1} \psi_{p,j}^2 \left\{ \left| \frac{1}{mW'} \sum_{s=p+2}^{\lfloor n/W' \rfloor} \sum_{t=W'(s-j-1)+1}^{W'(s-j)} \left(\zeta_t^2 - \gamma_0\right) \right| \right. \\
&\qquad \left. + \sum_{h=1}^{W'-1} \left| \frac{1}{m\left(W'-h\right)} \sum_{s=p+2}^{\lfloor n/W' \rfloor} \sum_{t=W'(s-j-1)+1}^{W'(s-j)-h} \left(\zeta_t \zeta_{t+h} - \gamma_h\right) \right| \right\} \\
&= \sum_{j=0}^{p+1} \psi_{p,j}^2 \left\{ \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{mW'}} \right) + \sum_{h=1}^{W'} \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{m\left(W'-h\right)}} \right) \right\} \\
&\leq \mathcal{O}_{\mathbb{P}}\left(\frac{W'}{\sqrt{n}} \right).
\end{aligned}$$

Where in the last line we have used the fact that $m \sim n/W'$ along with the fact that

$$\sum_{h=1}^{W'-1} \frac{1}{\sqrt{n\left(1 - \frac{h}{W'}\right)}} < \frac{1}{\sqrt{n}} \left(\int_1^{W'-1} \frac{1}{\sqrt{1 - \frac{x}{W'}}} \mathrm{d}x + \sqrt{W'} \right) = \frac{2W'}{\sqrt{n}} \left(1 + o(1)\right).$$

Arguing analogously we likewise have that $T_{1,2} \leq \mathcal{O}_{\mathbb{P}}\left(\frac{W'}{\sqrt{n}}\right)$. For the second term we have that

$$\begin{aligned}
T_2 &= \Bigg|\gamma_0 + 2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_h \\
&\quad + \sum_{\substack{k\neq l\\ 0\leq k,l\leq p+1}} \psi_{p,k}\psi_{p,l}\left(\gamma_{W'|k-l|} + 2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_{W'|k-l|+h}\right) - \gamma_0 - 2\sum_{h=1}^{\infty}\gamma_h\Bigg| \\
&\leq 2\left|\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_h - \left\{\sum_{h=1}^{W'-1} + \sum_{h=W'}^{\infty}\right\}\gamma_h\right| + 2\sum_{\substack{k\neq l\\ 0\leq k,l\leq p+1}} \psi_{p,k}\psi_{p,l}\left|\sum_{h=0}^{W'-1}\gamma_{W'|k-l|+h}\right| \\
&\leq 2\sum_{h=1}^{W'-1}\frac{h}{W'}\left|\gamma_h\right| + 2\sum_{h=W'}^{\infty}\left|\gamma_h\right| + 2\sum_{\substack{k\neq l\\ 0\leq k,l\leq p+1}} \psi_{p,k}\psi_{p,l}\sum_{h=0}^{W'-1}\left|\gamma_{W'|k-l|+h}\right| \\
&< \frac{2}{W'}\left(\sum_{h=1}^{\infty}h\left|\gamma_h\right| + \sum_{\substack{k\neq l\\ 0\leq k,l\leq p+1}} \psi_{p,k}\psi_{p,l}\sum_{h=0}^{W'-1}\left(W'|k-l|+h\right)\left|\gamma_{W'|k-l|+h}\right|\right) \\
&= \mathcal{O}\left(W'^{-1}\right).
\end{aligned}$$

For the third term we have that $T_3 \leq \mathcal{O}\left(\frac{NW'^2}{n}\right)$ and for the fourth term we likewise have that $T_4 \leq \mathcal{O}\left(\frac{NW'^2}{n}\right)$. Combining the bounds on terms T_1, T_2, T_3, and T_4 the stated result follows. □

3.6.8 Proof of Theorem 3.3.1

Proof. With slight abuse of notation write $I \in \mathcal{G}(W,a)$ if $I = \{l,\ldots,l+w-1\}$, where $(l,w) \in \mathcal{G}(W,a)$. For each $k = 1,\ldots,N$ introduce the set of all intervals containing the k-th change point η_k, and with $1/(p+1)$ of the points in the interval tying to the left of η_k and the remaining $(p+1)/(p+2)$ points lying to the right of η_k:

$$\mathcal{I}_k = \left\{I \in \mathcal{G}(W,a) \mid \eta_k \in I, \left\lfloor\frac{|I\cap\{1,\ldots,\eta_k\}|}{p+2}\right\rfloor = (p+1)\left\lfloor\frac{|I\cap\{\eta_k+1,\ldots,n\}|}{p+2}\right\rfloor\right\}.$$

Moreover assume that

$$\delta_k > 2a\,(p+2)\left(W \vee n^{\frac{2p_k^*}{2p_k^*+1}}\left(\frac{16C_{p,p_k^*}^2\tau^2\lambda_\alpha^2}{\Delta_{p_k^*,k}^2}\right)^{\frac{1}{2p_k^*+1}}\right), \qquad k = 1, \ldots, N.$$

Since $\lambda_\alpha = \mathcal{O}\left(\sqrt{\log(n)}\right)$ for any fixed α and either of threshold (3.6) or threshold (3.8), this assumption can be seen to correspond to condition (3.16) in Theorem 3.3.1. For ease of reading introduce the notation

$$V_k^\alpha\,(n) = n^{\frac{2p_k^*}{2p_k^*+1}}\left(16C_{p,p^*}^2\tau^2\lambda_\alpha^2/\Delta_{p_k^*,k}^2\right)^{\frac{1}{2p_k^*+1}}, \qquad k = 1, \ldots, N.$$

Due to lemma 3.6.7, testing for a change point on an interval $I' \in \mathcal{I}_k$ using (3.3) with threshold λ_α the k-th change point will be detected as long as $|I'| > (p+1)V_k^\alpha\,(n)$ on the event

$$\left\{L_{\mathcal{G}(W,a)}^{\widehat{\tau}}\,(\boldsymbol{\zeta}) \leq \lambda_\alpha\right\} \cap \{\widehat{\tau} < 2\tau\}\,. \tag{3.37}$$

Therefore, there must be an interval $I'' \in \mathcal{I}_k$ with $|I''| < a\,(p+2)\,(W \vee V_k^\alpha\,(n))$ on which the k-th change can be detected. By the assumption on the δ's and the above discussion, the shortest interval in $\mathcal{G}\,(W,a)$ on which the k-th chaneg point can be detected will not overlap with the shortest intervals on which the $(k-1)$-th and $(k+1)$-th changes will be detected. Finally, on the event (3.37) no test carried out on a sub-interval which are free from change points will spuriously reject. Therefore, events E_3^*, E_4^*, and E_5^* are verified. □

3.6.9 Proof of Lemma 3.3.4

Proof. We must show that $\mathrm{sSIC}(p') > \mathrm{sSIC}(p)$ for all $p' \neq p$ in the set $\{\underline{p}, \dots, \overline{p}\}$. We begin with the case $p' > p$ for which we have that

$$\begin{aligned}
&\mathrm{sSIC}(p') - \mathrm{sSIC}(p)\\
&\qquad = \frac{n}{2}\log\left(1 - \frac{\hat{\sigma}_p^2 - \hat{\sigma}_{p'}^2}{\hat{\sigma}_p^2}\right) + \left[\left(\hat{N}_{p'} + 1\right)(p'+1) - \left(\hat{N}_p + 1\right)(p+1)\right]\log^{\alpha}(n)\\
&\qquad := T_1 + T_2.
\end{aligned}$$

Observe that by Corollary 3.3.2 on a set with probability $1 + o(1)$ we will have that $\hat{N}_{p'} = \hat{N}_p = N$. Therefore, the fact that the ζ's are Gaussian combined with the ℓ_2 risk of constrained least squares spline estimators, which can be found for example in Shen et al. (2022), guarantee that on a set with probability $1 + o(1)$ we will have that $|\hat{\sigma}_{p'}^2 - \sigma^2| \lesssim n^{-1}\log(n)$ for each $p' \geq p$. Consequently

$$T_1 \gtrsim -\frac{n}{2}\left(\hat{\sigma}_p^2 - \hat{\sigma}_{p'}^2\right)/\hat{\sigma}_p^2 \geq -\frac{n}{2}\left(\left|\hat{\sigma}_p^2 - \sigma^2\right| + \left|\hat{\sigma}_{p'}^2 - \sigma^2\right|\right)/\hat{\sigma}_p^2 \gtrsim -\log(n).$$

Again by Corollary 3.3.2 we have that with high probability

$$T_2 = (N+1)\left(p' - p\right)\log^{\alpha}(n) \gg \log(n).$$

Consequently, for n sufficiently larger we have that with high probability $\mathrm{sSIC}(p') - \mathrm{sSIC}(p) > 0$ for $p' > p$. Next we consider the case $p' < p$ for which we have that

$$\begin{aligned}
&\mathrm{sSIC}(p') - \mathrm{sSIC}(p)\\
&\qquad = \frac{n}{2}\log\left(\frac{\hat{\sigma}_{p'}^2}{\hat{\sigma}_p^2}\right) + \left[\left(\hat{N}_{p'} + 1\right)(p'+1) - \left(\hat{N}_p + 1\right)(p+1)\right]\log^{\alpha}(n)\\
&\qquad := T_1 + T_2. \qquad (3.38)
\end{aligned}$$

By condition (iii) on a high probability set we must have that T_1 is negative and of the order $\mathcal{O}\left(n\log(n)\right)$, while $\hat{N}_{p'}$ will be of the order $\mathcal{O}(n/\log(n))$. Therefore, since $\alpha > 1$ we are done. Since $(\overline{p} - \underline{p}) = \mathcal{O}(1)$ a union bound argument is sufficient to establish that with n sufficiently large, on a high probability set, $\text{sSIC}(p') > \text{sSIC}(p)$ for all $p' \neq p$. This completes the proof. □

3.6.10 Remark on Assumption 3.3.3

We remark that 3.3.3 was made for ease of technical exposition, and although it does not seem straightforward to relax the assumption in full generality we conjecture that Algorithm 3 is able to localize all change points at the optimal rate when the assumption is violated, albeit with different leading constants in (3.16). The reason for the claim is the following: Assumption 3.3.3 is made to avoid the possibility of signal cancellation, however examining (3.23) in the proof of Lemma 3.6.7 it can be seen that there are only p values of δ' for which exact signal cancellation, and for any such δ' increasing or decreasing δ' by a constant will result in an interval of the same order for which no signal cancellation occurs.

Here we show that Algorithm 3 can localize change points at the optimal rate in the absence of Assumption 3.3.3 the when signal is piecewise linear. Moreover we provide some simulated examples of piecewise polynomial signals which violate Assumption 3.3.3 and show that the change points are still detected.

Relaxing Assumption 3.3.3 for piecewise linear signals signals

Here we show that for piecewise linear signals Algorithm 3 is able to localize all changes at the minimax optimal rate when Assumption 3.3.3 does not hold, provided the remaining assumption in Theorem 3.3.1 hold. Without loss of generality we consider the case of a single change:

$$f_\circ\left(t/n\right) = \begin{cases} \alpha_0 + \alpha_1\left(t/n - \eta/n\right) & \text{if } t \leq \eta \\ \beta_0 + \beta_1\left(t/n - \eta/n\right) & \text{else} \end{cases}.$$

Therefore we will show that using the threshold $\lambda = \hat{\tau}\bar{\lambda}$, for some $\bar{\lambda} > 0$, on a high probability set the change can be detected on an interval of length at most

$$Cn^{\frac{2p^*}{2p^*+1}} \left(16\tau^2\bar{\lambda}^2/\Delta_{p^*}^2\right)^{\frac{1}{2p^*+1}},$$

where C is a sufficiently large constant and $p^* \in \{0,1\}$ is defined as in (3.14). If $\operatorname{sign}(\alpha_0 - \beta_0) = \operatorname{sign}(\alpha_1 - \beta_1)$ this can be shown precisely as in Lemma 3.6.7. Therefore, we examine the setting in which $\operatorname{sign}(\alpha_0 - \beta_0) \neq \operatorname{sign}(\alpha_1 - \beta_1)$, for which there are three possible cases of interest:

- Case I: $\Delta_0 = \Delta_1 (\delta/n)$
- Case II: $\Delta_0 > \Delta_1 (\delta/n)$
- Case III: $\Delta_0 < \Delta_1 (\delta/n)$

Similar to Lemma 3.6.7, without loss of generality we let δ' be an integer such that the change occurs at location δ' and put $m = (p+2)\delta'$. We therefore need to show that the statistic $|D_{1,m}^1(\boldsymbol{Y})|$ can detect the change point with high probability for an appropriately chosen δ'. For ease of reading introduce the notation

$$\begin{aligned} C_1 &= 1/\sqrt{\sum_{i=0}^{2}\binom{2}{i}^2} \\ g_{\delta'} &= \frac{1}{\delta'}\sum_{t=1}^{\delta'}\left(1 - t/\delta'\right) \text{ for } \delta' \in \mathbb{N}. \end{aligned}$$

Case I: let δ' be an integer for which $\delta' < \delta/2$. Using the facts that $\Delta_1/\Delta_0 = n/\delta$ and $g_{\delta'} < 1/2$ for all δ' we have that

$$\left|D_{1,m}^1(\boldsymbol{f})\right| \geq C_1\sqrt{\delta'}\left(\Delta_0 - g_{\delta'}\Delta_1\left(\delta'/n\right)\right) = C_1\sqrt{\delta'}\left(\Delta_0 - g_{\delta'}\Delta_0\left(\delta'/\delta\right)\right) \geq \frac{3C_1}{4}\sqrt{\delta'}\Delta_0$$

and the desired result follows by rearranging (3.21).

Case II: this can be treated similarly to Case I.

Case III: note that there is a δ'' for which $\Delta_0 = \Delta_1(\delta''/n)$. We first consider the setting where $\delta'' < (2/C_1)^2 \left(16\tau^2\bar{\lambda}^2/\Delta_1^2\right)^{1/3}$, in which case letting δ' be such that $\delta' > 24\delta''$, and using the fact that $g_{\delta'} \geq 1/12$ for all $\delta' > 1$ by (3.22), we have that

$$\left|D^1_{1,m}(\boldsymbol{f})\right| \geq C_1\sqrt{\delta'}\left(g_{\delta'}\Delta_1(\delta'/n) - \Delta_0\right) \geq \frac{C_1}{12}\sqrt{\delta'}\left(\Delta_1(\delta'/n) - 12\Delta_0\right) \geq \frac{C_1}{24}\sqrt{\delta'}\Delta_1(\delta'/n).$$

Therefore, rearranging (3.21) and accounting fort the facts that we must have $\delta' > 24\delta''$, along with the fact that $(2/C_1)^2 > (24/C_2)^{2/3}$, we obtain that the change will be detected as soon as

$$\delta' \geq 24\,(2/C_1)^2\left(16\tau^2\bar{\lambda}^2/\Delta_1^2\right)^{1/3}.$$

Finally we consider the case $\delta'' \geq (2/C_1)^2\left(16\tau^2\bar{\lambda}^2/\Delta_1^2\right)^{1/3}$. In this case, letting $\delta' \leq \delta''$ and using the fact that $\Delta_0 \geq \Delta_1(\delta'/n)$ for all such δ' we obtain that

$$\left|D^1_{1,m}(\boldsymbol{f})\right| \geq C_1\sqrt{\delta'}\left(\Delta_0 - g_{\delta'}\Delta_1(\delta'/n)\right) \geq \frac{C_1}{2}\sqrt{\delta'}\Delta_0 \geq \frac{C_1}{2}\sqrt{\delta'}\Delta_1(\delta'/n),$$

and as in the previous cases the desired result follows by rearranging (3.21).

Examples of higher order polynomials which violate 3.3.3

Here we give simulated examples of higher order piecewise polynomial signals which violate Assumption 3.3.3, and show that Algorithm 3 is still able to detect the change points in practice. Specifically we consider three piecewise quadratic signals with a single change point at location η:

$$f_\circ(t/n) = \begin{cases} \alpha_0 + \alpha_1(t/n - \eta/n) + \alpha_2(t/n - \eta/n)^2 & \text{if } t \leq \eta \\ \beta_0 + \beta_1(t/n - \eta/n) + \beta_2(t/n - \eta/n)^2 & \text{else} \end{cases} \tag{3.39}$$

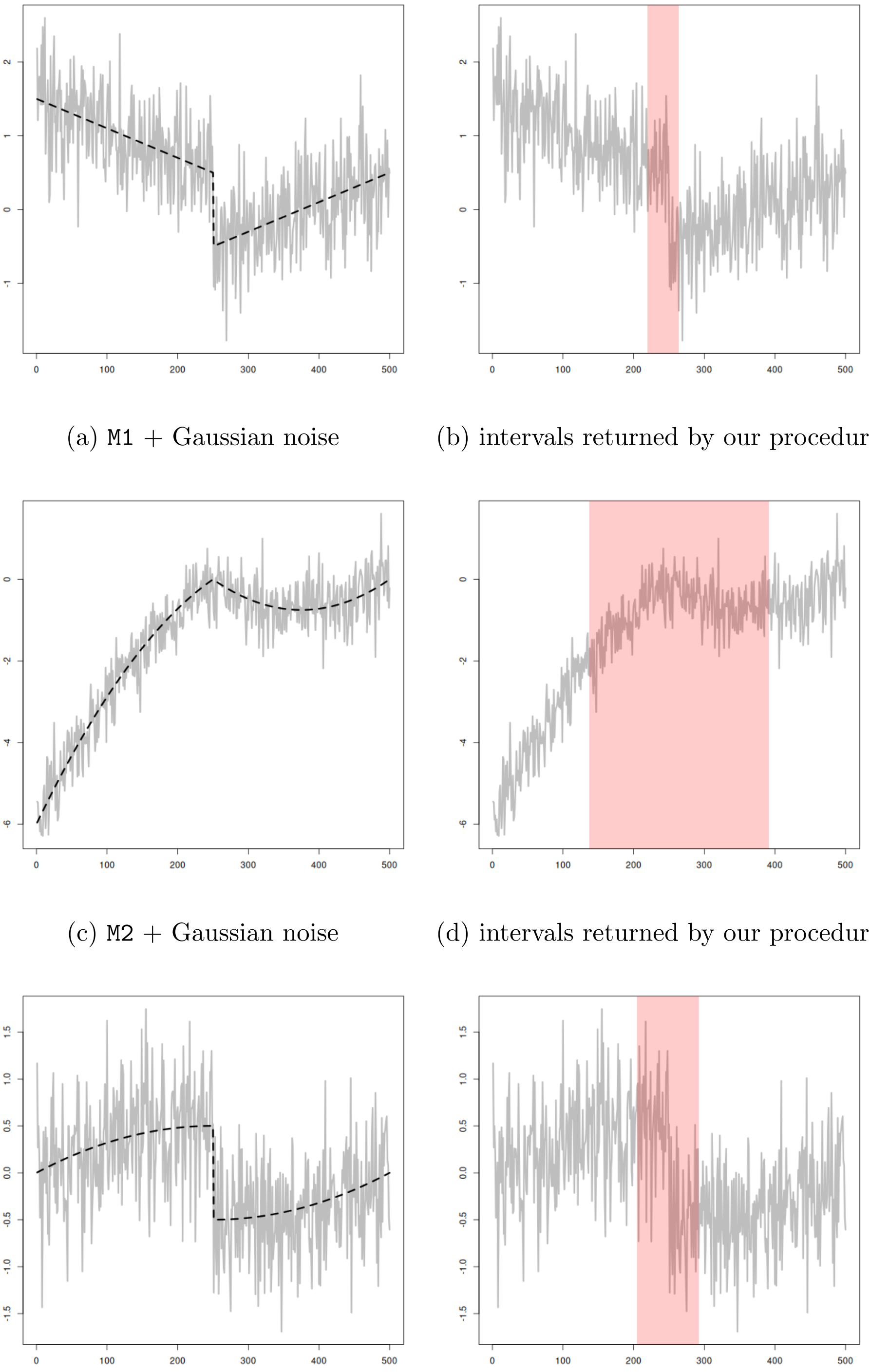
We consider three instances of (3.39) where in each case the sample size is $n = 500$, the change point occurs at location $\eta = n/2$, and changes occur in two derivatives of different order in such a way that the changes work against each other in the sense that they have different signs and the signal strengths as measured by $\Delta_j(\delta/n)^j$ for $j = 0, 1, 2$ exactly match. The three models are denoted by `M1`, `M2`, and `M3` and the values of the α's and β's are given in Table 3.8.

Table 3.8: Values of α's and β's for three instances of (3.39) which violate Assumption 3.3.3 when the sample size is $n = 500$ and the change point occurs at location $\eta = n/2$.

	α_0	β_0	α_1	β_1	α_2	β_2
`M1`	$-1/2$	$1/2$	-2	2	0	0
`M2`	0	0	6	-6	-12	12
`M3`	$1/2$	$-1/2$	0	0	-2	2

We contaminate the signals with independent noise having marginal $\mathcal{N}\left(0, 0.5^2\right)$ distribution and apply Algorithm 3 with parameter $\alpha = 0.1$. The results of this experiment, which was run with random seed `42` in `R`, are shown in Figure 3.6. In all three cases Algorithm 3 returns a single interval which contains the true change point location.

Figure 3.6: Piecewise polynomial signals which violate Assumption 3.3.3 with coefficients specified in Table 3.8, contaminated with i.i.d. Gaussian noise having standard deviation $\sigma = 0.5$ (left column). Intervals of significance with uniform 90% coverage returned by our procedure (right column). Black dashed lines (- - -) represent underlying piecewise polynomial signal, light grey lines (—) represent the observed data sequence, red shaded regions (🟥) represent intervals of significance returned by our procedure.



(a) M1 + Gaussian noise

(b) intervals returned by our procedure

(c) M2 + Gaussian noise

(d) intervals returned by our procedure

(e) M3 + Gaussian noise

(f) intervals returned by our procedure

4 Robust Inference for Change Points using Confidence Sets

4.1 Introduction and problem statement

In this chapter we return to the problem of performing inference on the unobserved change point locations in the piecewise polynomial change point model. However, we restrict our attention to signal which are either piecewise constant or piecewise linear. Motivated by the fact that real data are often messy, exhibiting heavy tails, heteroskedasticity, and distributions with arbitrary atoms, we develop a robust inference procedure which requires almost no assumptions on the distribution of the contaminating noise entering the model. On a high level we model the data's median as piecewise polynomial, and recover disjoint intervals which must each contain a change point location at a prescribed confidence level by performing a larger number of local homogeneity tests. The tests in turn are based on approximations of confidence sets for the underlying regression function obtained by inverting certain multi-scale tests which act on the signs of the data. By working implicitly with signs of the contaminating noise, which are automatically bounded independent of the data's distribution, we are able to develop a procedure which gives accurate inference on the unobserved change point locations with almost no assumptions the data's distribution.

In Section 4.1.1 below we motivate our procedure with a real data example. The remainder of the chapter is structured as follows. In Section 4.2 we introduce the change point model, and present our method for recovering disjoint intervals which must each contain a change point location at a prescribed confidence level. In Section 4.3 we present theoretical properties of the procedure, including coverage guarantees and large sample guarantees on the recovery of all change point locations. In Section 4.4 we extend the procedure to settings in which the contaminating noise is serially dependent, and discuss the procedure's behavior in settings when the data have non-unique median. Finally, in Section 4.5 we show the the practical value of our procedure compared to the state of the art via two simulation studies and two real data examples.

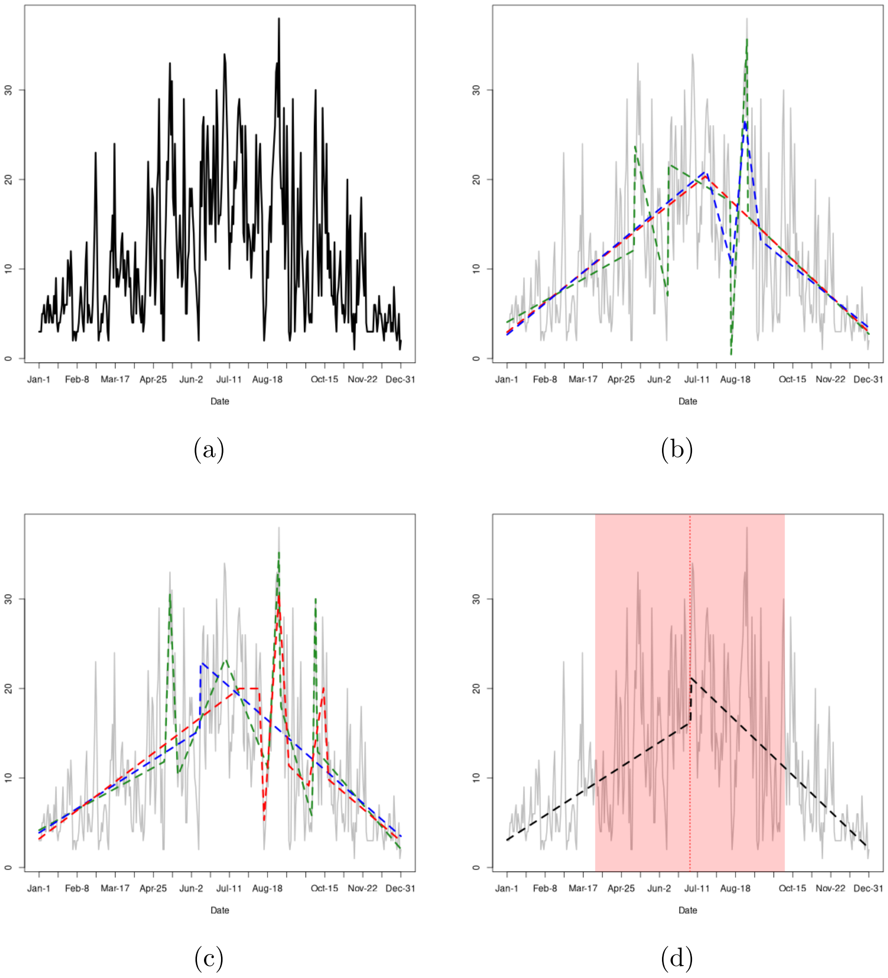
4.1.1 A motivating example: the yearly Ozone concentration cycle

We motivate then need for robust uncertainty quantification in change point problems via the following data example. Consider the time series of daily Ozone concentration (maximum of one hour averages) in the Los Angeles basin during 1976; the data is available through the `mlbench` package (Leisch and Dimitriadou, 2021) and was initially studied by Breiman and Friedman (1985). The data is plotted in Figure 4.1a and exhibits heavy tails and heteroskedasticity, as well as a visually obvious trend. It is well documented that Ozone concentrations in the Northern hemisphere follow a pronounced yearly cycle with the maximum occurring towards the middle of the year (Monks, 2000). In terms of signal estimation using change point algorithms a piecewise constant fit is clearly not appropriate, however one may consider modeling the data as piecewise linear with a single change point where concentration level peaks.

We estimate the underlying signal using six state of the art algorithms for recovering piecewise linear trends, which however do not not quantify the uncertainty around the number of change points they recover or their locations, and show the results in Figures 4.1b and 4.1c. In particular we consider the narrowest-over-threshold (NOT) algorithm of Baranowski et al. (2019a) with and without imposing continuity of the underlying signal (+cont), the Isolate-Detect (ID) algorithm of Anastasiou and Fryzlewicz (2022), the Wald-type test for structural change (SC) of Bai and Perron (2003), free knot splines (FKS) proposed by Spiriti et al. (2013), and finally multivariate adaptive regression splines (MARS) proposed by Friedman (1991).

The plots show uncertainty among the methods regarding the locations and even the number of change points present in the data, most likely due to the fact that the typical

Figure 4.1: (a) Daily Ozone concentration in the Los Angeles basin during 1976; (b) estimated piecewise linear signals using NOT-cont (- - -), NOT (- - -), and ID (- - -); (c) estimated piecewise linear signals using MARS (- - -), FKS (- - -), and SC (- - -); (d) 90% interval of significance obtained with our procedure together with the estimated change point location taken to be the midpoint of the interval and the estimated piecewise linear signal recovered via quantile regression.



(a)

(b)

(c)

(d)

assumptions of light tailed homoskedastic noise are violated. Since none of the methods considered quantify the uncertainty around the objects they recover, it is difficult to judge which is closest to the truth. Figure 4.1d shows regions obtained by running our procedure searching for change points in a piecewise linear parametric description of the median, with a nominal coverage level of %90. In this case a single interval is returned, and we estimate the change point location as the centre of this interval then estimate the signal using a median regression to the left and right of this point. This is justified since, under certain mild conditions set out in Section 4.3, every interval returned by the algorithm contains exactly one change point. The fitted signal agrees with the stylized facts regarding Ozone concentration cycles, and unlike the other methods no change points are estimated at locations which seem to correspond to local extremes in the contaminating noise.

4.2 Methodology

4.2.1 Model set-up

We first describe the data model, and our inference task, more formally. We work in the setting where the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ are observed on an equi-spaced grid, and can be written as the sum of a signal component and an noise component:

$$Y_t = f_\circ\left(t/n\right) + \zeta_t \qquad ,t = 1, \ldots, n. \tag{4.1}$$

The function $f_\circ : [0,1] \to \mathbb{R}$ is piecewise polynomial with known degree $p \in \{0, 1\}$ and unknown break locations. That is to say, associated with $f_\circ(\cdot)$ are N, possibly diverging with n, integer valued change point locations $\Theta = \{\eta_1, \ldots, \eta_N\}$ such that for each $k = 1, \ldots, N$ the function $f_\circ(\cdot)$ can be described as a degree p polynomial on $[(\eta_k - p - 1)/n, \eta_k/n]$ but not on $[(\eta_k - p)/n, (\eta_k + 1)/n]$. The aim is to recover sub-intervals of the index set $\{1, \ldots, n\}$ such that, uniformly at some prescribed significance level, every interval contains at least one change point location.

In order to ensure our inference procedure is robust to the distribution of the contaminating noise we will only impose the two assumptions shown below. We do not impose any restrictions regarding existence of moments or homogeneity of distribution, and stress that the distribution of the contaminating noise does not need to be known.

Assumption 4.2.1. $\mathbb{P}(\zeta_t > 0) = \mathbb{P}(\zeta_t < 0)$ *for each* t.

Assumption 4.2.2. $sign(\zeta_t) \perp\!\!\!\perp sign(\zeta_s)$ *for all* $s \neq t$.

The above assumptions are very mild. Assumption 4.2.1 requires the noise to be sign symmetric, which automatically holds for all continuous (median centered) distributions. Sign symmetry implies the median is a set containing zero, therefore we look for change points in the piecewise polynomial parametrization of the data's median. Issues concerning possible non-uniqueness of the median are dealt with in Section 4.4.2. Assumption 4.2.2 requires the signs of the contaminating noise to be independent, which in fact allows for a certain degree of serial dependence in the raw noise; for instance, if $\zeta_t = \sigma_t \times \varepsilon_t$ with $\varepsilon_1, \ldots, \varepsilon_n$ being a sequence of symmetric and mutually independent random variables and each σ_t being an $\mathcal{F}_{t-1}$ measurable function, then the signs of the ζ's are serially independent whereas the ζ's themselves are not. In Section 4.4.1 we relax Assumption 4.2.2 at the cost of introducing some additional assumptions on the serial dependence in the sequence of signs.

4.2.2 Main idea for change point inference

The main idea for change point inference follows from the idea of inference without selection introduced in Section 2.2.3. Let $\{T_{s:e}\}_{1 \leq s \leq e \leq n}$ be a series of local tests, where each $T_{s:e} : (Y_s, \ldots, Y_e)' \mapsto \{0, 1\}$ tests the local null hypothesis

$$H_0^{s:e} : f_\circ(\cdot) \text{ is a degree } p \text{ polynomial on } [s/n, e/n]. \tag{4.2}$$

If the family-wise error of the tests tests is bounded by a given $\alpha \in (0, 1)$ it is clear that with probability at least $1 - \alpha$ the only (s, e) pairs on which a local null is rejected will correspond to stretches of the data which contain at least one change point location. In Section 4.2.3 below we construct a series of such tests subject to the requirements that:

- The tests are robust to the distribution of the contaminating noise, in the sense of their family-wise error being bounded by a given α, as long as the noise distribution satisfies Assumption 4.2.1 and Assumption 4.2.2.

- The tests can be computed efficiently for any polynomial degree $p \in \{0, 1\}$, and on any local stretch of the data.

In order to efficiently recover the narrowest sub-intervals of the index set which each contain a change point location, we embedding our local tests in the Narrowest Significance Pursuit algorithm which was described in detail in Section 2.2.3 and Algorithm 2. In numerical experiments we use the following method for constructing a grid of sub-intervals, which corresponds to the function `subIntervalsGrid` in Algorithm 2. We first select an integer M which determines how many intervals will be in the grid, then draw all sub-intervals from a restricted index set, and finally re-scale the intervals so they cover the original index set. Pseudo-code is provided in Algorithm 5. In practice we always set $M = 1000$.

Algorithm 5: Algorithm for drawing a coarse grid of contiguous sub-intervals from the interval $\{s, \dots, e\}$.

```
function subIntervalsGrid($s, e, M$):
    $n_M \leftarrow \lfloor (1 + \sqrt{1 + 8M}) / 2 \rfloor \vee (e - s + 1)$
    $\delta_M \leftarrow \lfloor (e - s) / (n_M - 1) \rfloor$
    for $(s, e)$ in allSubIntervals($1, n_M$) do
        $(s, e) \leftarrow ((s - 1)\delta_M + 1, (e - 1)\delta_M + 1)$
    end
return
```

4.2.3 Construction of local tests

Our starting point for a local test is the robustified test of Fryzlewicz (2023), which corresponds to locally fitting the degree p polynomial which produces empirical residuals whose signs have the smallest (absolute) standardized partial sums, then checking whether the same quantity exceeds a given threshold:

$$T_{s:e}^{\lambda}\left(\boldsymbol{Y}\right)=\mathbf{1}\left\{\min_{\hat{f}}\max_{s\leq i\leq j\leq e}\frac{1}{\sqrt{j-i+1}}\left|\sum_{t=i}^{j}\operatorname{sign}\left(Y_t-\hat{f}\left(t/n\right)\right)\right|>\lambda\right\},\qquad 1\leq s\leq e\leq n. \tag{4.3}$$

The minimisation is over all polynomials of degree p. This test has the appealing property that, for any sub-interval on which a change point does not occur, the following inequality must hold:

$$\begin{aligned}\min_{\hat{f}}\max_{s\leq i\leq j\leq e}\frac{1}{\sqrt{j-i+1}}&\left|\sum_{t=i}^{j}\operatorname{sign}\left(Y_t-\hat{f}\left(t/n\right)\right)\right|\\ &\leq\max_{1\leq i\leq j\leq n}\frac{1}{\sqrt{j-i+1}}\left|\sum_{t=i}^{j}\operatorname{sign}\left(\zeta_t\right)\right|.\end{aligned} \tag{4.4}$$

Under Assumptions 4.2.1 and 4.2.2 partial sums of $\operatorname{sign}\left(\zeta_t\right)$ will always be stochastically bounded by partial sums of Rademacher random variable regardless of the distribution of the ζ's. It is therefore straightforward to stochastically bound (4.4) and so obtain a λ which controls the family-wise error of (4.3).

Unfortunately due to the non-linearity of the sign function this local test cannot be computed efficiently. An exception occurs for the case $p=0$, which is studied in Fryzlewicz (2021), where it is shown that all $\mathcal{O}\left(n^2\right)$ possible local tests can be computed in $\mathcal{O}\left(n^3\right)$ time. However, for general p computing (4.3) as described in Fryzlewicz (2023) has computational complexity $\mathcal{O}\left((e-s+1)^{p+3}\right)$ for ever (s,e) pair. This is because for every candidate $\hat{f}$ which produces a unique sequence of residual signs, of which there are at least of the order $\binom{e-s+1}{p+1}$, it is necessary to compute all partial sums of the same residual signs, which is an $\mathcal{O}\left((e-s+1)^2\right)$ operation.

To avoid these computational challenges we relax the parametric assumption on (4.3) and invert the resulting test to obtain a local confidence set for $f_\circ\left(\cdot\right)$. By the duality of hypothesis tests and confidence regions we can equivalently test whether the resulting confidence set contains a polynomial of degree p. Conceptually, a local confidence set can be built as follows. For a candidate regression function $f:[0,1]\to\mathbb{R}$ write the empirical residuals as $\hat{\zeta}_t^f = Y_t - f(t/n)$. In light of (4.3) we say the empirical residuals look locally like noise if they pass the test

$$\psi_{s:e}^f(\boldsymbol{Y}) = \mathbf{1}\left\{\max_{s \le i \le j \le e} \frac{1}{\sqrt{j-i+1}} \left| \sum_{t=i}^{j} \operatorname{sign}\left(\hat{\zeta}_t^f\right) \right| > \lambda \right\}. \tag{4.5}$$

Then, if λ is chosen using (4.4) to control the family-wise error of (4.3) at some level α, a local $1-\alpha$ level confidence set for $f_\circ(\cdot)$ is immediately given by $\mathcal{C}_{s:e}(\boldsymbol{Y},\alpha) = \left\{f \mid \psi_{s:e}^f(\boldsymbol{Y}) \neq 1\right\}$. The following test is therefore equivalent to the computationally infeasible test (4.3):

$$T_{s:e}(\boldsymbol{Y}) = \mathbf{1}\left\{\mathcal{C}_{s:e}(\boldsymbol{Y},\alpha) \text{ does not contain degree } p \text{ polynomials}\right\}, \tag{4.6}$$

We propose to calculate conservative point-wise upper and lower bounds on the set $\mathcal{C}_{s:e}(\boldsymbol{Y},\alpha)$ using existing methods from the shape constrained function estimation literature (Davies and Kovac, 2001; Dümbgen and Johns, 2004), then approximate the local test (4.6) by checking whether a degree p polynomial may pass between these bounds. This procedure is described in detail below.

Point-wise bounds on confidence sets

We describe how conservative point-wise bounds on the set $\mathcal{C}_{s:e}(\boldsymbol{Y},\alpha) = \left\{f \mid \psi_{s:e}^f(\boldsymbol{Y}) \neq 1\right\}$ can be efficiently computed, using a procedure introduced by Dümbgen and Johns (2004). Since the polynomial degree of the underlying signal is either 0 or 1 on stationary segments, when inverting (4.5) we can naturally restrict attention to candidate $f(\cdot)$'s which are either non-increasing or non-decreasing. To show the main idea assume that on the interval $[s/n, e/n]$ it is known that $f_\circ(\cdot)$ belongs to $\mathcal{F}_\uparrow$, the set of all non-decreasing functions. Obvious modifications will give bounds for non-increasing functions. With the additional information that the regression function is non-decreasing, for each $k \in \{s,\dots,e\}$ we care about computing the following point-wise upper and lower bounds:

$$\begin{aligned} L_k^{\uparrow} &= \inf\left\{f\left(k/n\right) \mid f \in \mathcal{C}_{s:e}\left(Y,\alpha\right) \cap \mathcal{F}_{\uparrow}\right\} \\ U_k^{\uparrow} &= \sup\left\{f\left(k/n\right) \mid f \in \mathcal{C}_{s:e}\left(Y,\alpha\right) \cap \mathcal{F}_{\uparrow}\right\}. \end{aligned} \tag{4.7}$$

The presence of the absolute value in (4.5) means that the bounds (4.7) cannot be computed with a single pass through the data. Notice however that the test (4.5) can be thought of as performing a two sided test at all scales and locations on the vector of empirical residuals $\left(\hat{\zeta}_s^f, \ldots, \hat{\zeta}_e^f\right)'$. For computing the lower bound we only care about testing the right tail, and for computing the upper bound we only care about the left tail. This naturally leads to the following bounds based on a one sided version of (4.5):

$$\begin{aligned} \check{L}_k^{\uparrow} &= \inf\left\{f\left(k/n\right) \mid f \in \mathcal{F}_{\uparrow}, \max_{s\leq i\leq j\leq e} \frac{1}{\sqrt{j-i+1}} \sum_{t=i}^{j} \operatorname{sign}\left(Y_t - f(t/n)\right) \leq \lambda\right\}, \\ \check{U}_k^{\uparrow} &= \sup\left\{f\left(k/n\right) \mid f \in \mathcal{F}_{\uparrow}, \max_{s\leq i\leq j\leq e} \frac{1}{\sqrt{j-i+1}} \sum_{t=i}^{j} \operatorname{sign}\left(f(t/n) - Y_t\right) \leq \lambda\right\}. \end{aligned} \tag{4.8}$$

For each $k \in \{s, \ldots, e\}$ it necessarily holds that $\check{L}_k^{\uparrow} \leq L_k^{\uparrow}$ and $\check{U}_k^{\uparrow} \geq U_k^{\uparrow}$, therefore any coverage guarantees for a test based on (4.7) will hold for the same test based on (4.8). Importantly, as explained in detail in in Dümbgen and Johns (2004), these new bounds can be computed with a single pass through the data using the recursions:

$$\begin{aligned} \check{L}_k^{\uparrow} &= \min\left\{\check{f} \in \{-\infty, Y_s, Y_{s+1}, \ldots, Y_k\} \mid \check{f} \geq \check{L}_{k-1}^{\uparrow}, \max_{s\leq i\leq j\leq k} \frac{1}{\sqrt{j-i+1}} \sum_{t=i}^{j} \operatorname{sign}(Y_t - \check{f}) \leq \lambda\right\}, \\ \check{U}_k^{\uparrow} &= \max\left\{\check{f} \in \{\infty, Y_e, Y_{e-1}, \ldots, Y_k\} \mid \check{f} \leq \check{U}_{k+1}^{\uparrow}, \max_{k\leq i\leq j\leq e} \frac{1}{\sqrt{j-i+1}} \sum_{t=i}^{j} \operatorname{sign}\left(\check{f} - Y_t\right) \leq \lambda\right\}. \end{aligned} \tag{4.9}$$

We remark that using (4.9) the vectors $\check{\mathbb{L}}_{s:e}^{\uparrow} = \left(\check{L}_s^{\uparrow}, \ldots, \check{L}_e^{\uparrow}\right)'$ and $\check{\mathbb{U}}_{s:e}^{\uparrow} = \left(\check{U}_s^{\uparrow}, \ldots, \check{U}_e^{\uparrow}\right)'$ can be computed in quadratic time by pre-sorting the candidate set of $\check{f}$'s at each step k and using the fact that it is only necessary to check the value of partial sums with index pairs (i, j) that were not previously considered at any step $k' < k$. Algorithm 6 below provides pseudo code for the computation of $\check{\mathbb{L}}^{\uparrow}_{s:e}$ from data; the vectors $\check{\mathbb{U}}^{\uparrow}_{s:e}, \check{\mathbb{L}}^{\downarrow}_{s:e}$ and $\check{\mathbb{U}}^{\downarrow}_{s:e}$ can be computed in a similar fashion with obvious modifications. We stress however that the algorithm is due to Dümbgen and Johns (2004), and that the main technical innovation in this Chapter is the application of (4.9) to the problem of change point inference.

Algorithm 6: Algorithm for constructing a uniform lower bound on an unknown isotonic function. Given data $Y_s, \ldots, Y_e$ and a threshold λ the algorithm returns a sequence of point-wise lower bounds $\check{L}^{\uparrow}_s, \ldots, \check{L}^{\uparrow}_e$ constructed according to (4.9).

```
function lowerBound($\{Y_s, \ldots, Y_e\}, \lambda$):
    $\check{L}^{\uparrow}_s \leftarrow -\infty$
    for $k \in \{s+1, \ldots, e\}$ do
        $\check{L}^{\uparrow}_k \leftarrow \check{L}^{\uparrow}_{l-1}$
        $j \leftarrow k$
    end
    while $j > s$ do
        if $j = k$ then
            $\texttt{S} \leftarrow 0$
            $\texttt{rNew} \leftarrow \infty$
        end
        if $\min(Y_t \mid j+1 \leq t \leq e) > \check{L}^{\downarrow}_k$ then
            $\texttt{S} \leftarrow \texttt{S} + 1$
            $\texttt{rNew} \leftarrow \min(\texttt{rNew}, \min(Y_t \mid j+1 \leq t \leq e))$
        end
        else
            $\texttt{S} \leftarrow \texttt{S} + 1$
        end
        $d \leftarrow k - j + 1$
        if $S > \lambda\sqrt{d}$ then
            $j \leftarrow j - 1$
        end
        else
            $\check{L}^{\uparrow}_k \leftarrow \texttt{rNew}$
            $j \leftarrow k$
        end
    end
return
```

Explicit local tests based on confidence sets

We are finally in a position to describe our local test. If we knew the (initial) monotonicity of $f_\circ(\cdot)$ on the interval under consideration we would pick the appropriate test from

$$T^{\uparrow}_{s:e}(\boldsymbol{Y}) = \mathbf{1}\left\{\check{\mathbb{L}}^{\uparrow}_{s:e} \text{ and } \check{\mathbb{U}}^{\uparrow}_{s:e} \text{ seperable by degree p polynomial}\right\},$$
$$T^{\downarrow}_{s:e}(\boldsymbol{Y}) = \mathbf{1}\left\{\check{\mathbb{L}}^{\downarrow}_{s:e} \text{ and } \check{\mathbb{U}}^{\downarrow}_{s:e} \text{ seperable by degree p polynomial}\right\}.$$

Since the monotonicity of $f_\circ(\cdot)$ is not known, when $p = 0$ we reject the local null (4.2) if both tests of the above tests reject. However, when $p = 0$ the monotonicity of $f_\circ(\cdot)$ under the local null is known. In this case it is enough to take the maximum, as opposed to the minimum, of the two test in (4.10). Finally, our local tests have the form:

$$T_{s:e}(\boldsymbol{Y}) = \begin{cases} T^{\uparrow}_{s:e}(\boldsymbol{Y}) \vee T^{\downarrow}_{s:e}(\boldsymbol{Y}) & \text{if } p = 0 \\ T^{\uparrow}_{s:e}(\boldsymbol{Y}) \wedge T^{\downarrow}_{s:e}(\boldsymbol{Y}) & \text{if } p = 1 \end{cases}. \tag{4.10}$$

In practice having computed appropriate point-wise bounds, each local tests can be evaluated very quickly. For example, with $p = 0$ we simply check whether the minimum upper bound exceeds the maximum lower bound, which can be done in constant time. With $p = 1$ we check whether the convex hulls of the lower and upper bounds intersect, which can be done in logarithmic time using for example the algorithm in Barba and Langerman 2014. Consequently, the computational complexity of testing an interval $\{s, \ldots, e\}$ for a change point in the underlying picewise polynomial signal of degree $p \in \{0, 1\}$ is of the order $\mathcal{O}(e - s + 1)^2$. This compares favourably to the time complexity of the naive procedure described in Section 4.2.3, which is of the order $\mathcal{O}\left((e - s + 1)^{p+3}\right)$ for generic p's.

4.3 Theoretical properties

4.3.1 Coverage guarantees

We first investigate the choice of threshold λ which guarantees the local tests (4.10) will have correct coverage when embedded in the generic Narrowest Significance Pursuit algorithm. Achieving correct coverage boils down to stochastically bounding all standardized partial sums of the sign of the noise terms in model (4.1), since the proof of Theorem 4.3.1 reveals that on the event that all such partial sums are smaller than the chosen λ the function $f_\circ(\cdot)$ will be contained within the lower and upper bounds constructed according to 4.2.3 on any stretch of the data free from change points. As such, on this event no test of the form (4.10) will wrongly reject.

Kabluchko and Wang (2014) study the limiting distribution of the maximum standardized partial sum of a sequence of independently and identically distributed random variables, for a range of distributions. From their Theorem 1.1 we have that under Assumptions 4.2.1 and 4.2.2, and the additional requirement that $\mathbb{P}(\zeta_t = 0) = 0$, for any fixed constant z it holds that

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{1\leq i\leq j\leq n} \frac{1}{\sqrt{j-i+1}} \left| \sum_{t=i}^{j} \operatorname{sign}(\zeta_t) \right| > \mathfrak{a}_n + z/\mathfrak{a}_n \right) \to 1 - \exp\left(-2\Lambda e^{-z}\right) \qquad (4.11)$$

as $n \to \infty$. Here Λ is a numeric constant and $\mathfrak{a}_n = \sqrt{2\log\left(n \log^{-\frac{1}{2}} n\right)}$. Therefore, for a given $\alpha \in (0,1)$ we propose to use the following threshold in our local tests:

$$\lambda = \sqrt{2\log\left(n\log^{-\frac{1}{2}} n\right)} + \log\left(\frac{1}{\log\left(\frac{1}{1-\alpha}\right)/2\Lambda}\right) / \sqrt{2\log\left(n\log^{-\frac{1}{2}} n\right)}. \qquad (4.12)$$

A value for Λ is not explicitly given by Kabluchko and Wang, and we follow Fryzlewicz (2021) in setting $\Lambda = 0.274$ which is a numeric approximation to the unknown constant obtained via simulation. The threshold (4.12) can be used even if the contaminating noise has an atom at zero. This is because putting $\widetilde{\zeta}_t = \zeta_t \mid \zeta_t \neq 0$ for the sequence of ζ's with ζ's taking value zero removed, putting $\mathcal{I} = \{t \mid \zeta_t \neq 0\}$, and letting $\mathcal{P}(\mathcal{I})$ stand for all contiguous partitions of $\mathcal{I}$, the following double inequality must hold

$$\max_{1\le i\le j\le n} \frac{1}{\sqrt{j-i+1}} \left|\sum_{t=i}^{j} \operatorname{sign}(\zeta_t)\right| \le \max_{I\in\mathcal{P}(\mathcal{I})} \frac{1}{\sqrt{|I|}} \left|\sum_{t\in I} \operatorname{sign}(\zeta_t)\right| \le \max_{1\le i\le j\le n} \frac{1}{\sqrt{j-i+1}} \left|\sum_{t=i}^{j} \operatorname{sign}\left(\widetilde{\zeta}_t\right)\right|. \quad (4.13)$$

The first inequality is due to the fact that each partial sum on the left hand side of (4.13) has a corresponding larger or equal in magnitude partial sum in the right hand side of (4.13) constructed by removing the zeros from its numerator and decreasing (or not increasing) its denominator. The second inequality holds trivially since the maximum on the right hand side is taken over a set which contains the set over which the maximum on the left hand side is taken. Consequently, (4.11) can be applied to the quantity appearing on the right hand side of (4.13) and via (4.12) an asymptotically conservative threshold can be obtained.

Under very mild assumptions we therefore have the following result, which states that with high probability every interval returned by embedding the local tests (4.10) into Algorithm 2 contains at least one change point location.

Theorem 4.3.1. *Let assumption* $(Y_1, \ldots, Y_n)'$ *be a data vector from (4.1) and grant Assumptions 4.2.1 and 4.2.2 hold. Let* $\left\{\hat{I}_1, \ldots \hat{I}_{\hat{N}}\right\}$ *be intervals returned by Algorithm 2 using any sub-sampling scheme, local test (4.10), and threshold (4.12). Then on a set with probability* $1-\alpha+o(1)$ *every interval returned contains at least one change point location.*

In light of Theorem 4.3.1 above, we observe that the number of intervals returned can be treated as an assumption free lower bound on the number of change points in the data.

4.3.2 Detection and localization guarantees

This section provides conditions under which our method is consistent, in the sense that with high probability every interval returned contains exactly one change point and the number of intervals returned matches the number of change points. Since we aim to detect changes in the median we need some control over the behaviour of the noise terms around their medians. This is provided by the following assumption.

Assumption 4.3.1. *The noise terms $\zeta_1, \dots, \zeta_t$ all have a unique median value of zero, and there is a non-decreasing function $H : [0,1] \to \mathbb{R}^+$ with $H(u) = \infty$ if $u = 1$ and a constant $c_H \in (0,1)$ for which H is convex on $[0, c_H]$ such that for any $u \in [0,1]$ it holds that $\mathbb{P}(\zeta_t \leq H(u)) \wedge \mathbb{P}(\zeta_t \geq -H(u)) \geq (1+u)/2$.*

If the noise terms are identically distributed with quantile function $Q(\cdot)$ we may simply take $H(u) = Q\left(\frac{1+u}{2}\right)$. To simplify the exposition we consider a variant of the generic Narrowest Significance Pursuit which acts on all contiguous sub-intervals of the index set in its first stage search (lines 5-8). In practice such an algorithm would be prohibitively slow, however no generality is lost. Indeed the same results (with different constants) could have been obtained by considering an algorithm acting on any grid which is multi-scale in the sense of Nemirovskii (1985); see also *Definition 1* in Li et al. (2019).

Piecewise constant signals

We begin with the canonical setting in which the signal is piecewise constant. Here the signal is a step function and can be written as

$$f_\circ(t/n) = \sum_{k=1}^{N+1} \mu_k \mathbf{1}\left\{\eta_{k-1} < t \leq \eta_k\right\}, \qquad \mu_k \neq \mu_{k+1}.$$

Denote the size of each jump in the signal by $\Delta_k = |\mu_k - \mu_{k+1}|$. Following the discussion in Section 4.2.3 our local test has the following explicit form in the piecewise constant setting:

$$T_{s:e}(\boldsymbol{Y}) = \mathbf{1}\left\{\min_{s \leq t \leq e} \check{U}_t^{\uparrow} < \max_{s \leq t \leq e} \check{L}_t^{\uparrow}\right\} \vee \mathbf{1}\left\{\min_{s \leq t \leq e} \check{U}_t^{\downarrow} < \max_{s \leq t \leq e} \check{L}_t^{\downarrow}\right\}. \tag{4.14}$$

We have the following result, stating that under certain conditions our procedure detects all change points and isolates them to their own interval with high probability.

Theorem 4.3.2. *Let $(Y_1, \ldots, Y_n)'$ be a data vector from (4.1) with piecewise constant signal component, and grant assumptions 4.2.1 - 4.2.2 and 4.3.1 hold. Let $\left\{\hat{I}_1, \ldots, \hat{I}_{\hat{N}}\right\}$ be the intervals returned by a version of Algorithm 2 which acts on all sub intervals of the index set, using threshold (4.12) local test (4.14). Assume the following condition holds*

$$\delta_k > C_1 \left(\log(n) \vee \frac{\log(n)}{\Delta_k^2}\right), \qquad k = 1, \ldots, N. \tag{4.15}$$

Then with probability $1 - \alpha + o(1)$ the following events occurs simultaneously:

$$E_2^* = \left\{\hat{N} = N\right\},$$

$$E_3^* = \left\{\forall k = 1, \ldots, N \ \hat{I}_k \cap \Theta = \eta_k\right\},$$

$$E_4^* = \left\{\left|\hat{I}_k\right| \leq C_2 \left(\log(n) \vee \frac{\log(n)}{\Delta_k^2}\right) |1 \leq k \leq N\right\}.$$

Here C_1, C_2 satisfy $C_1 > 2C_2$.

The event E_4^* gives the asymptotic rate at which our detection intervals expand, and this matches the minimax localization rate for change point detection in the canonical piecewise constant mean model. We note that *Theorem 3.1* can easily be turned into standard a large sample consistency result by choosing a threshold $\lambda = (1 + \varepsilon)\, \mathfrak{a}_n$ for some small but fixed $\varepsilon > 0$. In this case, the events $\{E_2^*, E_3^*, E_4^*\}$ would hold on a set with probability $1 - o\,(1)$.

Condition (4.15) gives the minimum signal to noise ratio which allows our method to detect all change points with high probability. As shown by Wang et al. (2020) in the light tailed noise setting no algorithm can consistently detect all change points when the signal strength $\sqrt{\delta_k}\Delta_k$ grows slower than the rate $\sqrt{\log(n)}$ and in this sense up to the $\mathcal{O}\,(\log\,(n))$ term condition (4.15) is unavoidable. The scale of the noise does not appear explicitly in Theorem 4.3.2 since we work with the signs of the data. We note however that the leading constants C_1 and C_2 do depend on the distribution of the noise through the function $H(\cdot)$. The price to pay for this is precisely that even arbitrarily large jumps will be undetectable if the minimum distance between change points is smaller that $\mathcal{O}\,(\log\,(n))$.

Piecewise linear and continuous signals

Next we consider the setting in which the signal is piecewise linear, and continuous at each change point location. Between adjacent change points the signal can be written as follows:

$$f_\circ\left(t/n\right) = \begin{cases} \mu_k + \alpha_k\left(t/n - \eta_k/n\right) & \text{if } \eta_{k-1} < t \leq \eta_k \\ \mu_k + \beta_k\left(t/n - \eta_k/n\right) & \text{if } \eta_k < t \leq \eta_{k+1} \end{cases}. \tag{4.16}$$

The effect of the change is now measured in terms of the difference in slopes before and after the change point, that is $\Delta_k = |\alpha_k - \beta_k|$. Write $\mathcal{H}^{L,\uparrow}_{s:e}\left(\boldsymbol{Y}\right)$ for the convex hull of the points $\left(s/n, \check{L}^{\uparrow}_s\right), \ldots, \left(e/n, \check{L}^{\uparrow}_e\right)$. Following the discussion in Section 4.2.3 the local test in this setting is as shown below:

$$T_{s:e}\left(\boldsymbol{Y}\right) = \mathbf{1}\left\{\text{vol}\left(\mathcal{H}^{U,\uparrow}_{s:e} \cap \mathcal{H}^{L,\uparrow}_{s:e}\right) > 0\right\} \wedge \mathbf{1}\left\{\text{vol}\left(\mathcal{H}^{U,\downarrow}_{s:e} \cap \mathcal{H}^{L,\downarrow}_{s:e}\right) > 0\right\}. \tag{4.17}$$

We have the following result, stating that under certain conditions our procedure detects all change points and isolates them to their own interval with high probability.

Theorem 4.3.3. *Let $(Y_1, \ldots, Y_n)'$ be a data vector from (4.1) with piecewise linear and continuous signal component, and grant assumptions 4.2.1 - 4.2.2 and 4.3.1 hold. Let $\left\{\hat{I}_1, \ldots, \hat{I}_{\hat{N}}\right\}$ be the intervals returned by a version of Algorithm 2 which acts on all sub intervals of the index set, using threshold (4.12) local test (4.17). Assume the following condition holds*

$$\delta_k > C_1\left(\log(n) \vee n^{2/3}\left(\frac{\log(n)}{\Delta_k^2}\right)^{1/3}\right), \qquad k = 1, \ldots, N. \tag{4.18}$$

Then with probability $1-\alpha+o(1)$ *the following events occurs simultaneously:*

$$E_5^* = \left\{ \hat{N} = N \right\},$$

$$E_6^* = \left\{ \forall k = 1, \ldots, N \ \hat{I}_k \cap \Theta = \eta_k \right\},$$

$$E_7^* = \left\{ \left| \hat{I}_k \right| \leq C_2 \left(\log(n) \vee n^{2/3} \left(\frac{\log(n)}{\Delta_k^2} \right)^{1/3} \right) \middle| 1 \leq k \leq N \right\}.$$

Here C_1, C_2 *satisfy* $C_1 > 2C_2$.

Assuming $\Delta_k = \mathcal{O}(1)$ we obtain intervals with width of order $\mathcal{O}\left(n^{2/3}\log^{1/3}(n)\right)$ which matches the localization rate in Baranowski et al. (2019a) and trails the minimax rate established by Raimondo (1998) by a logarithmic factor. This again reveals that the intervals returned are the narrowest possible (up to log factors). For the same reason as in Theorem 4.3.2 the scale of the noise does not appear in the result.

4.4 Extensions

4.4.1 Serially dependent noise

The local tests described so far are designed with (sign) independent noise in mind, and may break down in the presence of serially dependent noise. We therefore propose a variant of the local tests (4.10) which come with theoretical family-wise error guarantees in the presence of serially dependent noise. We additionally provide some practical approaches to inference in the presence of serially dependent noise, which work well but do not come with any theoretical guarantees.

Construction of local tests under serially dependent noise

The main idea is to invert a non-parametric relaxation of the local test (4.3) where, for some W which diverges with n, partial sums of the signs of empirical residuals are taken only over scales of size W or larger. That is, we replace the point-wise bounds in (4.10) with the new bounds:

$$\begin{aligned}\check{L}^{\uparrow}_{W,k} &= \inf\left\{ f\left(k/n\right) \mid f \in \mathcal{F}_{\uparrow},\ \max_{\substack{s\leq i\leq j\leq e\\ j-i>W}} \frac{1}{\sqrt{j-i+1}} \sum_{t=i}^{j} \operatorname{sign}\left(Y_t - f(t/n)\right) \leq \lambda \right\},\\ \check{U}^{\uparrow}_{W,k} &= \sup\left\{ f\left(k/n\right) \mid f \in \mathcal{F}_{\uparrow},\ \max_{\substack{s\leq i\leq j\leq e\\ j-i>W}} \frac{1}{\sqrt{j-i+1}} \sum_{t=i}^{j} \operatorname{sign}\left(f(t/n) - Y_t\right) \leq \lambda \right\}. \end{aligned} \tag{4.19}$$

In order to control the family-wise error of this new collection of tests, we need an analogue of the result (4.11) for standardized partial sum of signs of the noise calculated at scales larger than or equal to W. In order to obtain such a result, we impose the following additional assumptions.

Assumption 4.4.1. *The signs of the noise terms constitute a weakly stationary process with auto-covariance function $\gamma_h = Cov\left(sign\left(\zeta_t\right), sign\left(\zeta_{t+h}\right)\right)$ and strictly positive long run variance $\tau^2 = \gamma_0 + 2\sum_{h>0}\gamma_h$.*

Assumption 4.4.2. *There exists a Wiener process $\left\{B(t)\right\}_{t>0}$ such that for some $\nu > 0$, possibly after enlarging the probability space, it holds $\mathbb{P}$-almost surely that $\sum_{t=1}^{n} sign\left(\zeta_t\right) - \tau B(n) = \mathcal{O}\left(n^{\frac{1}{2+\nu}}\right)$.*

Assumption 4.4.3. *With the same ν as in Assumption 4.4.2 the quantity W satisfies (i) $n/W \to \infty$ and (ii) $n^{\frac{2}{2+\nu}}\log(n)/W \to 0$.*

Crucially, Assumption 4.4.2 allows partial sums of the sign process taken over scales of size W or larger to be replaced by increments of a Wiener process without affecting the asymptotics. Therefore, the serial dependence in the noise can be safely ignored. With these assumptions in place we have the following result.

Theorem 4.4.1. *Let Assumption 4.2.1 and Assumptions 4.4.1 -4.4.3 hold, and introduce the quantity*

$$L_{n,W}\left(\boldsymbol{\zeta}\right) = \max_{\substack{1\leq i\leq j\leq n\\ j-i>w}} \frac{1}{\sqrt{j-i+1}} \left| \sum_{t=i}^{j} sign\left(\zeta_t\right) \right|.$$

(i) For any fixed $z \in \mathbb{R}$ *it holds that* $\mathbb{P}\left(a_{n,W}\tau^{-1}L_{n,W}\left(\boldsymbol{\zeta}\right) - \mathfrak{b}_{n,W} \leq z\right) \to \exp\left(-2e^{-z}\right)$ *as* $n \to \infty$*, where the scaling and centring sequences are given by* $\mathfrak{a}_{n,W} = \sqrt{2\log\left(n/W\right)}$ *and* $\mathfrak{b}_{n,W} = 2\log\left(n/W\right) + \frac{3}{2}\log\log\left(n/W\right) - \log\left(2\sqrt{\pi}\right)$*. (ii) Moreover, the result in (i) continues to hold if* τ *is replaced by a consistent estimator satisfying* $\left|\hat{\tau}/\tau - 1\right| = o_{\mathbb{P}}\left(\log^{-1}(n/W)\right)$*.*

In light of Theorem 4.4.1 above, for a given α we propose to use the following threshold:

$$\lambda = \hat{\tau}\left(\sqrt{2\log(n/W)} + \frac{\frac{3}{2}\log\log(n/W) - \log\left(2\sqrt{\pi}\right) + \log\left(-2\log^{-1}\left(1-\alpha\right)\right)}{\sqrt{2\log(n/W)}}\right). \tag{4.20}$$

Arguing as in the proof of Theorem 4.3.1 it is clear that local tests based on (4.10) built the using bounds (4.19) and threshold (4.20) will have family-wise error no larger than $1 - \alpha + o(1)$. In numerical experiences we find that using a scale W of the order $\mathcal{O}\left(n^{1/3}\right)$ results in our test having good practical performance, in terms of detection power and the tests maintaining the desired level, across a range of noise types and test signals.

Long run variance estimation

In order to make the threshold defined in (4.20) operational, it is necessary to estimate the long run variance of the sequence of signs of the noise at the rate specified in part (ii) of Theorem 4.4.1. We propose to estimate the time average variance constant (TAVC, Wu 2009) at a particular scale W'. The TAVC at a given scale W' is defined as follows:

$$\text{TAVC}\left(W'\right) = \mathbb{E}\left[\left(\frac{1}{\sqrt{W'}}\sum_{t=1}^{W'}\text{sign}\left(\zeta_t\right)\right)^2\right].$$

As long as W' diverges the TAVC is consistent for the long run variance. Moreover, as argued by McGonigle and Cho (2023) scaling by the (square root of) the TAVC, in place of the long run variance, can improve the performance of change point tests since the TAVC at an appropriately chosen scale will better accounting for the local variability of the test statistic used. Our estimator for the TAVC is based on local block-wise estimates of $f_\circ\left(\cdot\right)$ and its construction proceeds in three steps:

1. Divide the index set $\{1,\ldots,n\}$ into mutually disjoint blocks $I_1,\ldots,I_{\lfloor n/W'\rfloor}$ and likewise mutually disjoint blocks $J_1,\ldots,J_{\lfloor n/W'^2\rfloor}$ where $|I_j| = W'$ for each $j = 1,\ldots,\lfloor n/W'\rfloor$ and $|J_k| = W'^2$ for each $k = 1,\ldots,\lfloor W'^2\rfloor$.

2. On each of the J_k's form an estimator $\hat{f}_{J_k}(\cdot)$ for $f_\circ(\cdot)$ based on the observations $\{Y_t \mid t \in J_k\}$ via some base method, such as for example quantile regression.

3. On each of the I_j's form the sequence of empirical residuals according to $\{\hat{\zeta}_{t,j} = Y_t - \hat{f}_{J_{\sigma(j)}}(t/n) \mid t \in I_j\}$, where in particular $\sigma(j) = \inf\{k \mid 2 \times |I_j \cap J_k| \geq |I_j|\}$,

Finally, our estimator for the TAVC at scale W' is defined as follows:

$$\widehat{\mathrm{TAVC}}\left(W'\right) = \lfloor n/W'\rfloor^{-1} \sum_{j=1}^{\lfloor n/W'\rfloor} \left(\frac{1}{\sqrt{W'}} \sum_{t\in I_j} \mathrm{sign}\left(\hat{\zeta}_{t,j}\right)\right)^2 . \tag{4.21}$$

In order to show consistency of (4.21) for the long run variance of the sign process we need the following assumption on the noise and on the local estimator of the regression function.

Assumption 4.4.4. *The ζ's have absolutely continuous and bounded density functions.*

Assumption 4.4.5. *The auto-covariances decay fast enough that $\sum_{h>1} h\,|\gamma_h| < \infty$, and for any integer h and any ordered subset of $\{1,\ldots,n-h\}$, say M, it holds that*

$$|M|^{-1} \sum_{t\in M} sign\,(\zeta_t)\, sign\,(\zeta_{t+h}) = \gamma_h + \mathcal{O}_{\mathbb{P}}\left(1/\sqrt{|M|}\right).$$

Assumption 4.4.6. *There is an absolute constant C_f such that if $\hat{f}_J(\cdot)$ is obtained from the sample $\{Y_t \mid t \in J\}$ which is free from change points then it holds with probability at least $1 - |J|^{-1}$ that*

$$\left\|\hat{f}_J - f_\circ\right\|_{J,\infty} < C_f\sqrt{\log\left(|J|\right)/|J|},$$

where $\|f\|_{J,\infty} = \sup_{t\in J} |f\,(t/n)|$.

Assumption 4.4.4 is purely technical, and in fact is relaxed in the simulation study presented in Section 4.5. Assumption 4.4.5 requires that the auto-covariance for the sequence of signs decays sufficiently fast, and can be estimated well from a finite sample. Finally Assumption 4.4.6 requires that $f_\circ(\cdot)$ can be estimated accurately on a stretch of the data free from change points. With these assumption in place we have the following result.

Lemma 4.4.1. *Grant Assumption 4.2.1 and Assumptions 4.4.4-4.4.6 hold, then the estimator for the TAVC at the scale W' satisfies*

$$\widehat{TAVC}\left(W'\right) = \tau^2 + \mathcal{O}_{\mathbb{P}}\left(\frac{NW'^3}{n} \vee \sqrt{\frac{\log(W')}{W'}}\right).$$

Lemma 4.4.1 reveals that if W' is chosen to be of the order $\mathcal{O}\left(n^\theta\right)$ for some $\theta < 1/3$ then (4.21) will satisfy part (ii) of Theorem 4.4.1 as long as the number of change points grows more slowly than $n^{1-3\theta}\log\left(n/W\right)$. In practice, in light of our preferred choice for W, we choose W' to be of the order $\mathcal{O}\left(n^{1/3-\varepsilon}\right)$ for some small fixed $\varepsilon > 0$.

Empirical alternatives to inference under dependent noise

Finally, we discuss two practical approaches to handling serially dependent noise. These approaches do not come with any theoretical guarantees, however when applied in the real data example presented in Section 4.5.2 we find they compare favorably to the theoretically justified solution proposed above.

- If one has access to a stretch of data known to be free from change points, it is often possible to fit a parametric model to the change point free data then use the model to pre-white the data being screened for change points prior to applying the original local tests.

- One may pre-average the data over non-overlapping blocks of size h, obtaining a new data set of length $\lfloor n/h \rfloor$, then apply the original tests for sign independent noise to the new data set. The motivation for this being that local averaging will reduce the degree of serial dependence in the data.

4.4.2 Noise with non-unique median

It is desirable to have change point detection guarantees when the contaminating noise terms have medians which are not unique. Removing Assumption 4.3.1 and keeping only Assumptions 4.2.1 and 4.2.2 we can at best claim that the medians are sets which contains zero. That is, there are constants $\underline{\theta}_t \leq 0$ and $\overline{\theta}_t \geq 0$ such that

$$\operatorname{median}(\zeta_t) = \left\{\theta \in \mathbb{R} \mid \mathbb{P}(\zeta_t \geq \theta) \wedge (\zeta_t \leq \theta) \geq \frac{1}{2}\right\} = \left[\underline{\theta}_t, \overline{\theta}_t\right], \qquad t = 1, \ldots, n. \tag{4.22}$$

Our proofs of the theorems in Section 4.3.2 rely on the observation that once the monotonicity of the signal has been correctly established under Assumption 4.3.1 there is always a point where the upper and lower bounds respectively are “not too far” from the signal. Under (4.22) it is clear that with probability $1 - \alpha + o(1)$ uniformly on any interval I the upper bound will never be closer than $\min\left(\overline{\theta}_t \mid t \in I\right)$ and the lower bound will never be closer than $-\max\left(\underline{\theta}_t \mid t \in I\right)$. To allow for a non-unique median we introduce the following generalization of Assumption 4.3.1.

Assumption 4.4.7. *The noise terms $\zeta_1, \ldots, \zeta_t$ have medians given by (4.22) with $\underline{\theta}_t \leq 0 \leq \overline{\theta}_t$, and there is a non-decreasing function $H : [0,1] \to \mathbb{R}^+$ with $H(u) = \infty$ if $u = 1$ and a constant $c_H \in (0,1)$ for which H is convex on $[0, c_H]$ such that for any $u \in [0,1]$ it holds that $\mathbb{P}\left(\zeta_t \leq \overline{\theta}_t + H(u)\right) \wedge \mathbb{P}\left(\zeta_t + \geq -\left(\underline{\theta}_t + H(u)\right)\right) \geq \frac{1+u}{2}$.*

With this assumption in place we can prove consistency of our method for any sign symmetric noise. However, the weakest detectable jump will now also depend on the the width of the median intervals. For example, we have to following counterpart to Theorem 4.3.2.

Corollary 4.4.1. *Let Assumptions 4.2.1, 4.2.2 and 4.4.7 hold. Define the the widest median interval around the j-th change point location as: $\Xi_k = \max_{\eta_{k-1} < t \leq \eta_{k+1}} \left(\overline{\theta}_t - \underline{\theta}_t\right)$.*

Then Theorem 4.3.2 still holds by replacing the definition of the jump size by $\Delta'_k = \Delta_k - \Xi_k$.

A similar analogue to Theorem 4.3.3 can likewise be obtained. In Corollary 4.4.1 above one can even allow for $\Delta'_k = 0$ as long as $\Delta_k > 0$. This is because with discrete valued noise it is possible to obtain a stretch of data around a change point location on which the upper and lower bounds perfectly interpolate the noise. That is, we may have $L_t = f_\circ(t/n) + \underline{\theta}_t$ and $U_t = f_\circ(t/n) + \overline{\theta}_t$ for $t \in \{\dots, \eta_k - 1, \eta_k, \eta_k + 1, \dots\}$. Supplied with these bounds the local test (4.14) will not declare a change point while $\Delta'_k = 0$. However, if perfect interpolation occurs and we test for a deviation from a polynomial of degree $p \in \{0, 1, \}$ it is clear that there is exactly one such polynomial which separates the upper and lower bound. Therefore, we can perform a second stage test for whether this unique polynomial, say $\hat{f}(\cdot)$, produces empirical residuals which pass test (4.5). If the test is not passed, we declare a change point on the interval being inspected.

The second stage test gives our procedure non-trivial power even when $\Delta' = 0$ and moreover does not affect the coverage guarantees. This is because on null intervals, uniformly with probability $1 - \alpha + o(1)$, the upper and lower bounds contain all functions which produce empirical residual that pass test (4.5). However if $\hat{f}(\cdot)$ is the only degree p polynomial which separates the upper and lower bounds then we must have that $\hat{f}(\cdot) \equiv f_\circ(\cdot)$ on the interval being tested.

4.5 Numerical illustrations

4.5.1 Simulation studies

Since existing methods for robust change point inference are mostly limited to the piecewise constant setting, for the sake of comparison we present simulations in this setting. The main take away message from this section is that despite being designed with generality in mind our method is competitive with respect to state of the art methods designed specifically for the piecewise constant setting. We compare our method to the robust Narrowest Significance Pursuit (RNSP) algorithm of Fryzlewicz (2021), the non-robust

Narrowest Significance Pursuit algorithm of Fryzlewicz (2023) with local tests based on self-normalized partial sums (NSP-SN), the multi-scale quantile segmentation algorithm (MQS) of Jula Vanegas et al. (2021), and finally the heterogeneous simultaneous multiscale change point estimator (HSMUCE) of Pein et al. (2017). All of these methods return intervals whose nominal coverage can be specified by the user. MQS can be tuned to detect changes in any quantile, and throughout we select tuning parameters such that it looks for change points in the median. We write SET when describing the performance our main procedure, and SET-DEP when describing the performance of the local tests introduced in Section 4.4.1 designed to deal with serially dependent noise.

Coverage

We first investigate the finite sample coverage for our method and the competing methods described above. This is done by applying each method to a vector of pure noise 100 times, and recording the number of times no intervals of significance are returned. For each method we select tuning parameters so that the output is a set of intervals with nominal 90% coverage. We investigate six noise types, described below, which conform to Assumptions 4.2.1 and 4.2.2, and each time simulate noise vectors having length $n = 512$.

- Gaussian: $\zeta_t \sim \mathcal{N}(0, 10)$ i.i.d.
- Cauchy: $\zeta_t \sim \text{Cauchy}(0, 1)$ i.i.d.
- Sym. Poisson: $\zeta_t = r_t \times P_t$ with $r_t \sim$ Rademacher i.i.d. and $P_t \sim \text{Poisson}(7)$ i.i.d.
- GARCH: $\zeta_t = \varepsilon_t \sigma_t$ with $\varepsilon_t \sim \mathcal{N}(0, 1)$ i.i.d. and $\sigma_t^2 = 10 + 0.45\zeta_{t-1}^2 + 0.45\sigma_{t-1}^2$
- TV-Variance: $\zeta_t = 6\left(1 + \sin(t\pi/n)\right)\varepsilon_t$ and $\varepsilon_t \sim t_3$ i.i.d.
- Mix: $\zeta_t \sim$ Rademacher i.i.d. if $0 < t \leq \lfloor n/3 \rfloor$; $\zeta_t \sim$ Sym. Poisson i.i.d. if $\lfloor n/3 \rfloor < t \leq \lfloor 2n/3 \rfloor$; $\zeta_t \sim t_3$ i.i.d. else

The results of the simulation study are presented in Table 4.1. Both of our proposed methods, along with RNSP and MQS, keep their coverage guarantees across all noise types and in fact tend to deliver over coverage. NSP-SN maintains desired coverage for all noise types apart from ‘Cauchy’ which makes sense since it only promises statistical size control when the noise is in the basin of attraction of the Gaussian distribution. H-SMUCE maintains coverage when the noise is continuous, but breaks down in the presence of discrete noise. This again makes sense since it was designed with Gaussian noise in mind, and with discrete noise the local variance estimator can become unstable when calculated on short intervals.

Table 4.1: Proportion of times out of 100 replications each method returned no intervals of significance when applied to a pure noise vector, satisfying Assumptions 4.2.1 and 4.2.2, with length $n = 512$.

Method	Gauss	Cauchy	Sym. Poisson	GARCH	TV-Variance	Mix
SET	0.97	0.98	0.98	0.98	1.00	0.89
SET-DEP	1.00	1.00	0.98	1.00	0.99	0.99
NSP-SN	1.00	0.53	1.00	1.00	1.00	0.94
RNSP	1.00	0.99	0.99	1.00	1.00	0.93
HSMUCE	0.98	1.00	0.17	0.98	1.00	0.22
MQS	0.97	0.98	1.00	0.99	0.98	0.90

We additionally investigate the finite sample coverage of the above methods in the presence of serially dependent noise. We repeat the experiment from Table 4.1 with five serially dependent noise types described below. Among these INAR stands for the Integer Valued Auto-regressive model proposed by Al-Osh and Alzaid (1987, 1988), which we briefly describe. INAR processes replace the operation of regressing on past values, as is done in classical auto-regressive processes, with thinning. The thinning operator $\circ$ is in turn defined as follows: if X is a non-negative integer-valued random variable and $\phi \in [0, 1]$, then $\phi \circ X = \sum_{i=1}^{X} Y_i$ where the Y_i's are independently distributed Bernoulli distributed random variables with success probability ϕ.

- AR(1)-A: $\zeta_t = 0.25\zeta_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, 1)$ i.i.d.
- AR(1)-B: $\zeta_t = 0.5\zeta_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, 1)$ i.i.d.
- ARMA(2,6): $\zeta_t = 0.75\zeta_{t-1} - 0.5\zeta_{t-2} + \varepsilon_t + 0.8\varepsilon_{t-1} + 0.7\varepsilon_{t-2} + 0.6\varepsilon_{t-3} + 0.5\varepsilon_{t-4} + 0.4\varepsilon_{t-5} + 0.3\varepsilon_{t-6}$ with $\varepsilon_t \sim \mathcal{N}(0, 1)$ i.i.d.

- INAR(1)-A: $\zeta_t = 0.25 \circ \zeta_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \text{Poisson}\,(1)$ i.i.d.
- INAR(1)-B: $\zeta_t = 0.5 \circ \zeta_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \text{Poisson}\,(1)$ i.i.d.

The results of the simulation study are given in Table 4.2. As guaranteed by Theorem 4.4.1 SET-DEP maintains the desired coverage level across each of the noise types. The remaining methods break down in the presence of serially dependent noise, which of course is to be expected as they are all designed to work in the setting of (sign) independent noise.

Table 4.2: Proportion of times out of 100 replications each method returned no intervals of significance when applied to a vector of serially dependent noise having length $n = 512$.

Method	AR(1)-A	AR(1)-B	ARMA(2,6)	INAR(1)-A	INAR(1)-B
SET	0.77	0.21	0.00	0.50	0.09
SET-DEP	1.00	1.00	0.99	0.90	0.92
NSP-SN	0.94	0.62	0.07	0.33	0.23
RNSP	0.90	0.51	0.03	0.89	0.35
H-SMUCE	0.92	0.44	0.00	0.38	0.31
MQS	0.83	0.23	0.00	0.96	0.52

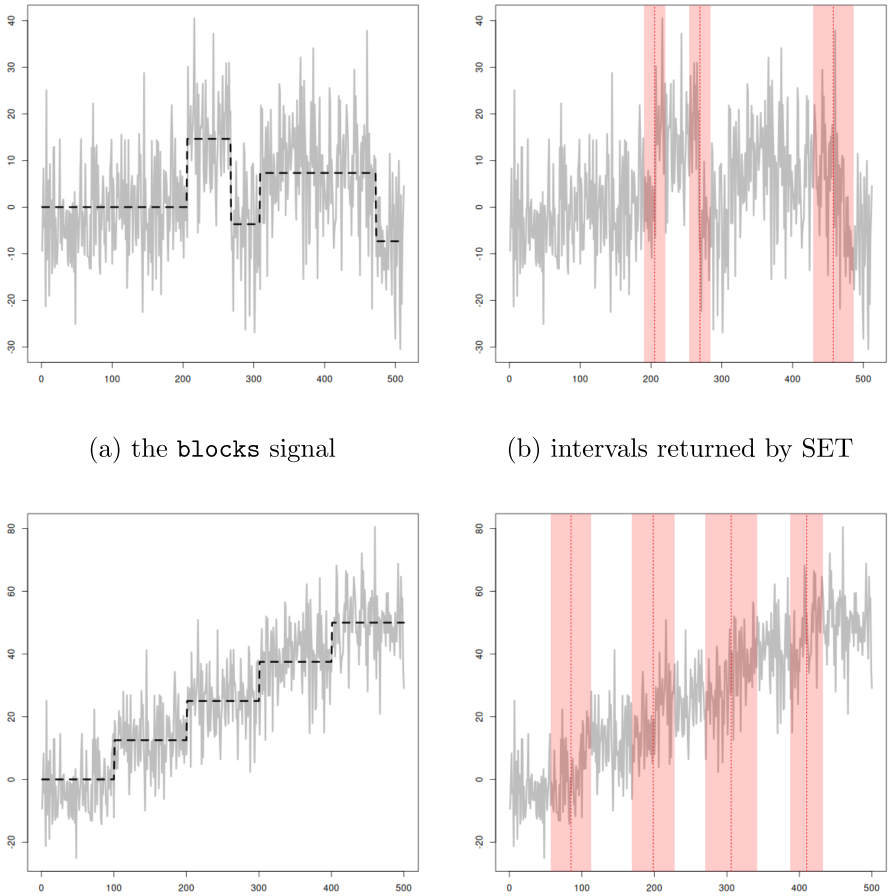
Detection power

Next we test the performance of our methods and their competitors on two test signals contaminated with each of the noise types described above. The signals are described in detail below, and examples of the two signals contaminated with the independent Gaussian noise with standard deviation $\sigma = 10$ are shown in Figure 4.2.

- `blocks`: the first $n = 512$ values of the blocks signal from Donoho and Johnstone (1994), as shown in Figure 4.2a, with $N = 4$ change points at locations $\Theta = \{205, 267, 308, 472\}$
- `staircase`: a signal with length $n = 500$ as shown in Figure 4.2c, with initial value zero and $N = 4$ jumps each of size $\Delta = 12.5$ at locations $\Theta = \{100, 200, 300, 400\}$

We again run 100 simulations, and choose tuning parameters so that each method returns intervals with nominal 90% coverage. On each iteration we record for each method the total length of intervals returned (length), the number of intervals which contain at least one change point location (no. genuine), the proportion of intervals returned which contain at least one change point location (prop. genuine), and whether all intervals returned contain at least once change point location (coverage).

Figure 4.2: Sample paths of the `blocks` and `staircase` signals contaminated with independent Gaussian noise with standard deviation $\sigma = 10$ (left column) and intervals of significance returned by running SET with $\alpha = 0.1$.



(a) the `blocks` signal

(b) intervals returned by SET

(c) the `staircase` signal

(d) intervals returned by SET

Tables 4.3 and 4.4 present the results of the simulation study for each of the signals contaminated with noise conforming to Assumptions 4.2.1 and 4.2.2 introduced in the previous section. In terms of the length of intervals returned and number of genuine intervals returned our method is frequently among the top two performing methods. Its performance is similar to to that of RNSP, which is not surprising given the our local test is essentially a non-parametric relaxation of the test used in Fryzlewicz (2021). We finally remark that NSP-SN and H-SMUCE, the two methods based on self-normalization, generally performed poorly; the extra flexibility provided by the local variance estimator leads to a loss of power, and compared to other methods these methods consistently detected fewer change points and returned longer intervals of significance.

Table 4.3: Average length of intervals returned (length), average of the number of intervals returned which contain at least one change point location (no. genuine), proportion of intervals returned which contain at least one change point location (prop. genuine), and whether all intervals returned contain at least once change point location (coverage), on the `blocks` signal contaminated with noise satisfying Assumptions 4.2.1 and Assumption 4.2.2, over 100 replications.

Method	Metrics	Gaussian	Cauchy	Sym. Poisson	GARCH	TV-Variance	Mix
SET	length	65.35	30.80	51.85	44.86	77.58	44.76
	no. genuine	2.93	3.99	3.03	3.66	2.33	3.93
	prop. genuine	1.00	1.00	1.00	1.00	0.99	0.99
	coverage	1.00	1.00	1.00	1.00	0.98	0.96
SET-DEP	length	118.07	61.49	92.25	87.89	121.14	72.09
	no. genuine	1.47	2.93	2.10	2.27	1.35	2.60
	prop. genuine	0.99	1.00	1.00	1.00	0.88	1.00
	coverage	0.99	1.00	1.00	1.00	0.88	1.00
NSP-SN	length	117.94	135.58	76.21	95.86	149.55	37.04
	no. genuine	1.85	1.50	2.46	2.40	0.92	2.90
	prop. genuine	1.00	0.94	0.83	1.00	0.81	0.50
	coverage	1.00	0.91	0.57	1.00	0.81	0.12
RNSP	length	72.11	33.23	58.04	49.24	85.82	47.49
	no. genuine	2.73	3.99	3.01	3.58	2.16	3.89
	prop. genuine	1.00	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	1.00	1.00
H-SMUCE	length	83.87	26.60	33.88	50.52	144.07	8.94
	no. genuine	1.56	1.92	0.04	1.72	1.49	0.00
	prop. genuine	0.80	0.96	0.04	0.86	0.89	0.00
	coverage	0.63	0.92	0.04	0.73	0.85	0.00
MQS	length	105.38	40.68	87.60	66.74	128.08	49.75
	no. genuine	2.83	3.99	3.03	3.64	2.28	3.93
	prop. genuine	1.00	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	0.99	1.00

Table 4.4: Average length of intervals returned (length), average of the number of intervals returned which contain at least one change point location (no. genuine), proportion of intervals returned which contain at least one change point location (prop. genuine), and whether all intervals returned contain at least once change point location (coverage), on the `staircase` signal contaminated with noise satisfying Assumptions 4.2.1 and Assumption 4.2.2, over 100 replications.

Method	Metrics	Gaussian	Cauchy	Sym. Poisson	GARCH	TV-Variance	Mix
SET	length	76.44	31.28	83.91	50.47	87.31	42.16
	no. genuine	3.56	4.00	3.29	3.98	3.05	3.97
	prop. genuine	1.00	1.00	1.00	1.00	1.00	0.98
	coverage	1.00	1.00	0.99	1.00	1.00	0.89
SET-DEP	length	136.49	66.20	143.99	104.17	151.17	80.43
	no. genuine	2.19	4.00	1.86	2.87	2.09	3.87
	prop. genuine	1.00	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	1.00	1.00
NSP-SN	length	125.61	137.06	76.14	104.15	163.21	17.63
	no. genuine	2.51	2.32	2.95	3.03	1.80	2.81
	prop. genuine	1.00	1.00	0.76	1.00	1.00	0.22
	coverage	1.00	1.00	0.37	1.00	1.00	0.00
RNSP	length	82.69	33.60	91.26	54.83	95.81	45.80
	no. genuine	3.39	4.00	3.00	3.97	2.85	4.00
	prop. genuine	1.00	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	1.00	1.00
H-SMUCE	length	78.32	22.32	28.63	54.02	87.03	9.76
	no. genuine	1.86	2.00	0.23	1.98	1.68	0.00
	prop. genuine	0.93	1.00	0.16	0.99	0.84	0.00
	coverage	0.87	1.00	0.09	0.98	0.76	0.00
MQS	length	99.65	39.32	109.78	66.23	118.86	49.62
	no. genuine	3.53	4.00	3.19	4.00	2.99	3.99
	prop. genuine	1.00	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	1.00	0.99

Tables 4.5 and 4.6 present the results of the simulation study for each of the signals contaminated with serially dependent noise introduced above. Although some methods maintain the desired coverage level under weakly dependent noise, such as AR(1)-A and INAR(1)-A, all of the methods designed with (sign) independent noise in mind break down in the presence of strongly dependent noise. Meanwhile, SET-DEP maintains the desired coverage level across all noise types.

Table 4.5: Average length of intervals returned (length), average of the number of intervals returned which contain at least one change point location (no. genuine), proportion of intervals returned which contain at least one change point location (prop. genuine), and whether all intervals returned contain at least once change point location (coverage), on the `blocks` signal contaminated with serially dependent noise, over 100 replications.

Method	Metrics	AR(1)-A	AR(1)-B	ARMA(2,6)	INAR(1)-A	INAR(1)-B
SET	length	28.59	32.29	32.87	30.28	33.19
	no. genuine	4.00	4.00	3.84	4.00	3.94
	prop. genuine	0.99	0.90	0.61	0.94	0.78
	coverage	0.94	0.54	0.04	0.74	0.17
SET-DEP	length	66.26	84.03	110.36	60.98	78.54
	no. genuine	2.90	1.94	1.02	2.98	2.21
	prop. genuine	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	1.00
NSP-SN	length	41.41	42.51	56.06	7.01	7.58
	no. genuine	4.00	4.00	3.92	0.54	0.58
	prop. genuine	1.00	1.00	0.98	0.02	0.02
	coverage	1.00	1.00	0.89	0.00	0.00
RNSP	length	30.31	33.06	35.38	30.50	33.31
	no. genuine	4.00	4.00	3.96	4.00	4.00
	prop. genuine	1.00	0.94	0.69	0.99	0.93
	coverage	0.98	0.70	0.10	0.96	0.66
H-SMUCE	length	16.66	20.66	34.84	14.47	12.11
	no. genuine	1.97	1.78	0.72	0.00	0.00
	prop. genuine	0.98	0.89	0.36	0.00	0.00
	coverage	0.97	0.79	0.13	0.00	0.00
MQS	length	38.09	43.63	39.80	35.57	38.51
	no. genuine	3.98	3.96	3.65	3.99	3.96
	prop. genuine	0.99	0.92	0.59	1.00	0.97
	coverage	0.96	0.64	0.03	0.99	0.85

Table 4.6: Average length of intervals returned (length), average of the number of intervals returned which contain at least one change point location (no. genuine), proportion of intervals returned which contain at least one change point location (prop. genuine), and whether all intervals returned contain at least once change point location (coverage), on the `staircase` signal contaminated with serially dependent noise, over 100 replications.

Method	Metrics	AR(1)-A	AR(1)-B	ARMA(2,6)	INAR(1)-A	INAR(1)-B
SET	length	28.07	29.09	30.58	28.52	30.34
	no. genuine	4.00	3.98	3.87	3.99	3.98
	prop. genuine	1.00	0.95	0.66	0.97	0.86
	coverage	1.00	0.75	0.07	0.83	0.45
SET-DEP	length	67.91	83.42	110.03	64.75	79.80
	no. genuine	4.00	3.99	2.36	4.00	4.00
	prop. genuine	1.00	1.00	1.00	1.00	1.00
	coverage	1.00	1.00	1.00	1.00	1.00
NSP-SN	length	34.94	37.16	57.09	6.96	7.67
	no. genuine	4.00	4.00	4.00	0.58	0.55
	prop. genuine	1.00	1.00	1.00	0.02	0.02
	coverage	1.00	1.00	1.00	0.00	0.00
RNSP	length	30.08	30.61	32.64	30.07	30.78
	no. genuine	4.00	4.00	3.92	4.00	3.99
	prop. genuine	1.00	0.98	0.73	1.00	0.97
	coverage	1.00	0.88	0.13	1.00	0.83
H-SMUCE	length	10.56	13.16	21.75	14.14	14.06
	no. genuine	1.98	1.83	1.05	0.02	0.04
	prop. genuine	0.99	0.92	0.53	0.02	0.03
	coverage	0.98	0.84	0.18	0.02	0.02
MQS	length	36.58	37.11	38.16	33.12	34.86
	no. genuine	4.00	3.97	3.62	4.00	3.96
	prop. genuine	1.00	0.95	0.61	1.00	0.96
	coverage	1.00	0.78	0.03	1.00	0.83

4.5.2 Real data analysis: air quality during COVID-19 lock downs

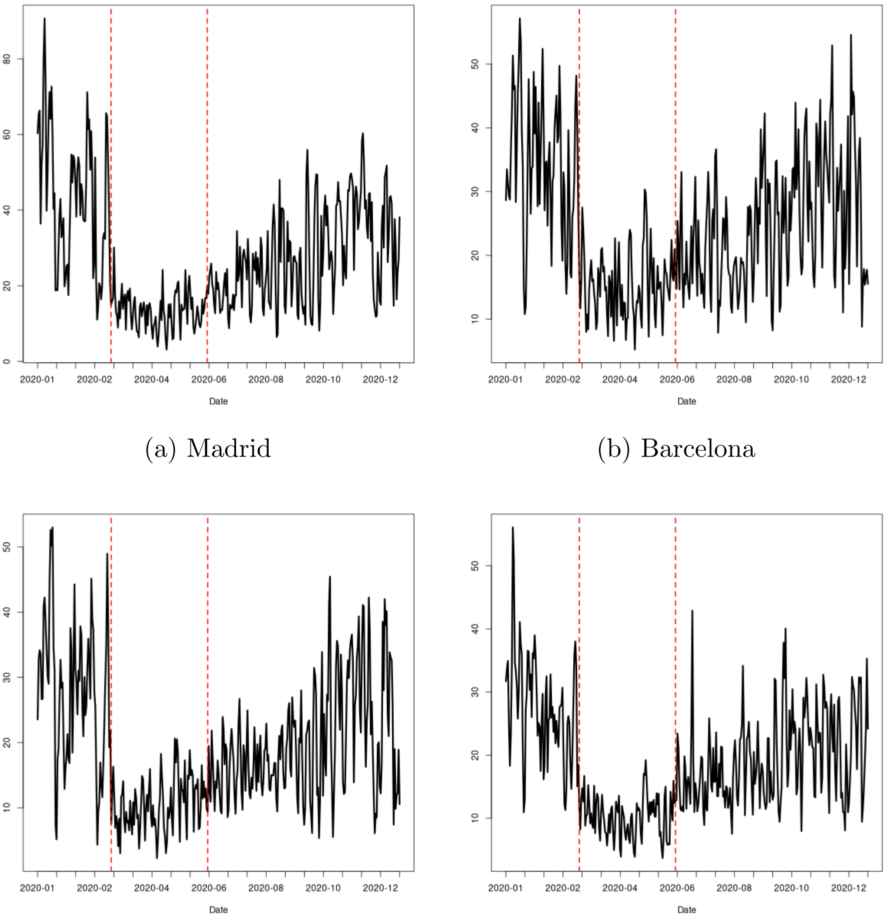
We analyze daily concentrations of nitrogen dioxide as a proxy for air quality for the four largest cities in Spain - Madrid, Barcelona, Valencia, and Sevilla - during the year 2020. The data were obtained from EuroAir (2022), and consist of daily averages of hourly nitrogen dioxide concentration readings across all monitoring sites available for a particular city. There is significant evidence supporting improvements in air quality in urban centres following COVID-19 lock downs in various countries (Slezakova and Pereira 2021, Jephcote et al. 2021). We choose data from Spain as there is strong a-priori evidence for two change point locations: the Spanish government declared a state of alarm on March 14 lasting until June 21, requiring all citizens to remain at home except to buy food and medicine and all non-essential businesses to close. Figure 4.3 shows a plot of the data, and indeed an abrupt change is clearly visible for all four cities around March 14 with a gradual return to base line levels starting in early June. The data exhibits heavy tails, heteroskdasticity, and some degree of auto-correlation, and is therefore a suitable test for the robustness of our method.

Analysis with piecewise constant signal

We initially model the time series as following model (4.1) with a piecewise constant signal component. Air quality during COVID-19 lock downs has been modeled in a similar fashion by Cho and Fryzlewicz (2023) and Grange et al. (2021) among others. We apply SET and SET-DEP with local test (4.14) and parameter $\alpha = 0.1$. We also apply the four competing methods discussed in Section 4.5.1, and select corresponding tuning parameters so that each produces intervals with 90% coverage. We apply SET-DEP to the raw data. For the remaining methods, which may fail in the presence of serially dependent noise, we first remove the bulk of the auto-correlation in the data by working with residuals from an AR(1) fit with parameter 0.5; this particular choice is justified by inspecting the robust ACF of suitably de-trended data from the year 2018, which we assume to be free from change points.

The intervals of significance returned by each method, and their widths, are shown in

Figure 4.3: Black solid lines (—) represent daily concentrations of nitrogen dioxide in four Spanish cities during 2020; red dashed lines (- - -) represent the start and end dates of the Spanish national state of alarm.



(a) Madrid

(b) Barcelona

(c) Valencia

(d) Sevilla

tables 4.7a and 4.7d. For all four cities both SET and SET-DEP return exactly two intervals of significance. This is in contrast to RNSP which fails to detect the end of the national state of alarm in Barcelona returning only one interval. In general, pre-whitening the data and applying SET results in slightly shorter intervals than applying SET-DEP to the raw data. We note however that SET frequently returns intervals which are either the shortest or very close in length to the shortest.

Table 4.7: Intervals of significance and their lengths (in brackets) returned by each method when applied to daily nitrogen dioxide concentration level data for four Spanish cities during the year 2020. The method SET-DEP was applied to the raw data, while all other methods were applied to pre-whitened data; see the main text for details.

SET	SET-DEP	NSP-SN	RNSP	HSMUCE	MQS
Feb 14 - Mar 31 (47) Jul 07 - Nov 25 (142)	Jan 24 - Apr 30 (97) May 22 - Dec 04 (196)	Feb 06 - Apr 17 (72) Jun 02 - Nov 27 (179)	Feb 09 - Mar 31 (52) Jul 01 - Nov 25 (148)	Feb 04 - Apr 06 (63) Jun 11 - Oct 31 (143)	Feb 09 - Mar 29 (49) Apr 10 - Oct 22 (195)

(a) Madrid

SET	SEt-DEP	NSP-SN	RNSP	H-SMUCE	MQS
Jan 13 - Apr 01 (80) Apr 01 - Dec 20 (264)	Jan 06 - Apr 21 (106) May 07 - Dec 20 (227)	Jan 06 - Apr 15 (101) Apr 21 - Dec 03 (227)	Jan 13 - Apr 01 (80)	Feb 04 - Apr 06 (63) Apr 08 - Nov 16 (223)	Jan 13 - Mar 30 (77) Apr 01 - Dec 03 (246)

(b) Barcelona

SET	SET-DEP	NSP-SN	RNSP	HSMUCE	MQS
Feb 17 - Apr 04 (48) May 26 - Nov 25 (184)	Jan 30 - Apr 13 (74) Apr 19 - Dec 01 (226)	Jan 29 - Apr 15 (78) May 02 - Dec 02 (215)	Feb 12 - Apr 05 (54) May 11 - Nov 25 (199)	Feb 04 - Apr 06 (63) May 16 - Sep 13 (121)	Feb 11 - Apr 04 (53) Apr 17 - Nov 28 (225)

(c) Valencia

SET	SET-DEP	NSP-SN	RNSP	H-SMUCE	MQS
Feb 16 - Apr 10 (55) May 25 - Nov 05 (165)	Jan 25 - Apr 22 (88) Apr 26 - Nov 05 (193)	Jan 27 - Apr 17 (82) Apr 17 - Oct 07 (174)	Jan 29 - Apr 09 (72) Jun 01 - Nov 20 (173)	Feb 20 - Mar 29 (39) May 10 - Sep 13 (127)	Jan 07 - Apr 04 (88) Apr 09 - Oct 08 (182)

(d) Sevilla

Analysis with piecewise linear signal

Next we model the data as coming from model (4.1) with a piecewise linear signal component. There is some evidence that climate data is best modeled as piecewise linear (Banesh et al., 2019) and besides a piecewise linear function seems visually to better fit the data in Figure 4.3. We first attempt to locate any change points using methods for piecewise linear signals which however do not provide any coverage guarantees for the change point location. In particular we consider the narrowest-over-threshold (NOT) algorithm of Baranowski et al. (2019a) with and without imposing continuity of the underlying signal

(+cont), isolate-detect (ID) of Anastasiou and Fryzlewicz (2022), the Wald-type test for structural change (SC) of Bai and Perron (2003), free knot splines (FKS) of Spiriti et al. (2013), multivariate adaptive regression splines (MARS) of Friedman (1991), trend filtering (TF) of Tibshirani (2014) and Kim et al. (2009)), and finally the ℓ_0 penalization based algorithm (CPOP) of Maidstone et al. (2017).

The data exhibits heavy tails and is therefore not appropriate for use by any of the above methods. We bring the data back to the light tailed domain by applying the Anscombe transform, and again work with the empirical residuals from an AR(1) fit. The change point locations obtained from the thus transformed data are reported in Table 4.8. Even in this relatively simple setting where the change point is visible by eye the need for uncertainty quantification is clear: there is considerable disagreement among the methods about the location and even the number of change points present.

Next we apply our methods, SET and SET-DEP, again setting $\alpha = 0.1$ but this time testing for changes in a piecewise polynomial signal of degree 1. We apply SET-DEP to the raw data. Prior to applying SET to the data we attempt to remove the bulk of the serial correlation by pre-average the data using non-overlapping blocks of length 3. The intervals obtained are shown in the final row to Table 4.8. The intervals cover the state of alarm on March 14 as well as its relaxation on June 21. Importantly, the intervals appear to be centered at regions of the data where a large proportion of the methods described above agree a change point occurred. As a proportion of the total sample size the intervals are indeed wide, however this reflects the observed uncertainty over change point locations and is consistent with the larger asymptotic width for piecewise linear signals given in Theorem 4.3.3.

4.6 Proofs

4.6.1 Preparatory results

Proposition 4.6.1. *For λ as in (4.12) with $\alpha \in (0,1)$ fixed, for any constant $C_\alpha > \sqrt{2}$ there is an integer n_0 such that for all $n > n_0$ it holds that $\lambda < C_\alpha \sqrt{\log n}$.*

Table 4.8: Estimated change point locations in pre-whitened and Anscombe transformed time series of daily nitrogen dioxide concentrations in four Spanish cities in 2020, obtained by change point detection and signal estimation methods suitable for data having a piecewise linear trend. The final two rows represent intervals of significance obtained by applying SET and SET-DEP to the data, testing for change points in a piecewise polynomial signals of degree 1; see the main text for details.

Method	Madrid	Barcelona	Valencia	Sevilla
NOT	Mar 12	Mar 13	Mar 12	Mar 12
NOT (+cont)	Apr 19	Apr 16	Apr 07	Apr 15
ID	Jan 28 Mar 10 14	Apr 18	Apr 12	Apr 19
SC	Mar 13	Mar 13	Mar 12	Mar 12
MARS	Apr 17 Nov 03,23 Dec 03	Apr 17 Dec 18	Apr 04 Dec 21	Apr 12
FKS	Jan 30,31 Feb 01,02 Apr 07,08	Apr 17,18 Dec 16,17	Jan 14,15,16,17 Feb 24,25,28,29 Mar 10,11,12,13 Dec 20,21	Apr 12,13
CPOP	Jan 21 Feb 28,29 Mar 11,14 Sep 27,28 Oct 02,09,10,13 Dec 31	Apr 19 Dec 17,31	Mar 12,13 Dec 31	Apr 13, Dec 31
TF	Jan 20 27 Jan 20,27 Mar 21 Apr 09,29 Jul 26 Aug 31 Sep 23 Nov 22 Dec 05	Feb 19 Mar 31 Apr 12,17 Aug 21 Nov 17,22	Jan 19 Feb 20 Mar 30 Apr 29 Jul 13 Oct 01 Nov 18	Feb 23 Apr 10,13 Jun 05 Jul 06 Oct 28 Dec 06
SET	Jan 30 - Aug 06	Jan 30 - Jul 16	Jan 30 - Aug 06	Feb 14 - Jul 31
SET-DEP	Jan 25 - Nov 12	Jan 31 - Nov 12	Jan 25 - Aug 10	Jan 01 - Jul 30

Proof. For α fixed it is clear that $\lambda \sim \sqrt{2\log(n)}$. □

Proposition 4.6.2. *For any $\epsilon > 0$ but not necessarily fixed and for C_α chosen as in 4.6.1 the following event holds with probability $1 - o(1)$:*

$$E_2(\epsilon) = \left\{ \max_{1\leq s\leq e\leq n} \frac{1}{\sqrt{e-s+1}} \left| \sum_{t=s}^{e} \left(\mathbf{1}_{\{\zeta_t \leq \epsilon\}} - \mathbb{P}(\zeta_t \leq \epsilon) \right) \right| \leq C_\alpha \sqrt{\log n} \right\}.$$

Proof. See Theorem 1 in Shao (1995). □

Theorem 4.6.1. *Let $\{B(t)\}_{t>0}$ be standard Brownian motion, and introduce the field of its standardised incremetns as follows:*

$$\mathfrak{X}(x,y) = \frac{B(x+y) - B(x)}{\sqrt{y}} \quad x, y > 0. \tag{4.23}$$

Introduce also the triangualr region $\mathbb{H}(n) = \left\{(x,y) \in \mathbb{R}^2 \mid x \in [0,n], y \in [1, n-x]\right\}$. Then putting $\mathfrak{a}_n = \sqrt{2\log(n)}$ and $\mathfrak{b}_n = 2\log(n) + \frac{3}{2}\log\log(n) - \log(2\sqrt{\pi})$, for any $z \in \mathbb{R}$ it holds that:

$$\lim_{n\to\infty} \mathbb{P}\left(\mathfrak{a}_n \left[\sup_{(x,y)\in\mathbb{H}(n)} \mathfrak{X}(x,y) \right] - \mathfrak{b}_n \leq z \right) = \exp\left(-e^{-z}\right).$$

Proof. See Theorem 4.2 in (Kabluchko, 2007). □

4.6.2 Intermediate results

Lemma 4.6.1. *For any underlying $f_\circ(\cdot)$ uniformly over all sub intervals of $I \subset \{1, \dots, n\}$ containing $|I| > \max\left((4C_\alpha/c_H)^2 \log(n), \lambda^2 \right)$ points the bounds constructed according to (4.9) satisfy*

$$E_3(I) = \left\{ \min_{t\in I} \left(f_\circ(t/n) - \check{L}_t^{\uparrow} \right) \vee \left(\check{U}_t^{\uparrow} - f_\circ(t/n) \right) \leq C_H \sqrt{\frac{\log(n)}{|I|}} \right\}$$

with probability $1 - \alpha + o(1)$. Here $C_H > 0$ is finite but may depend on the distribution of $\zeta_1, \dots, \zeta_n$.

Note: the proof of the lemma is almost identical to the proof of Proposition 3.1 (b) in Dümbgen (1998). However, it is not clear that this should be the case since Dümbgen's paper is based on the inversion of a different test. We therefore show a proof below.

Proof. We only show the part of the statement involving the lower bound. Assume there is a quantity $v_n(I) = H\left(u_n/\sqrt{|I|}\right)$ such that for all $t \in I$ we have $\left(f_\circ(t/n) - \check{L}_t^{\uparrow}\right) > v_n(I)$. By proposition 4.6.2 it follows that with high probability

$$\begin{aligned}
&\frac{1}{\sqrt{|I|}} \sum_{t\in I} \operatorname{sign}\left(Y_t - \check{L}_t^{\uparrow}\right) \\
&\geq \frac{1}{\sqrt{|I|}} \sum_{t\in I} \left(2\mathbf{1}_{\{\zeta_t \geq -v_n(I)\}} - 1 \pm 2\mathbb{P}\left(\zeta_t \geq -v_n(I)\right)\right) \\
&\geq \frac{1}{\sqrt{|I|}} \sum_{t\in I} \left(2\mathbb{P}\left(\zeta_t \geq -v_n(I)\right) - 1\right) - 2\left(\max_{1\leq s\leq e\leq n} \frac{1}{\sqrt{e-s+1}} \left|\sum_{t=s}^{e} \left(\mathbf{1}_{\{\zeta_t \geq -v_n(I)\}} - \mathbb{P}\left(\zeta_t \geq -v_n(I)\right)\right)\right|\right) \\
&\geq u_n - 2C_\alpha\sqrt{\log n}.
\end{aligned}$$

However this will not be consistent with the construction of $\check{\mathbb{L}}^{\uparrow}$ as specified in (4.9) unless we also have $u_n - 2C_\alpha\sqrt{\log n} \leq \lambda$ which combined with 4.6.1 implies that $u_n \leq 3C_\alpha\sqrt{\log n}$. Combining with Assumption 4.3.1 and the fact that $|I| > (4C_\alpha/c_H)^2 \log(n)$ we finally obtain that

$$\min_{t\in I}\left(f_\circ(t/n) - \check{L}_t^{\uparrow}\right) \leq H\left(4C_\alpha\sqrt{\frac{\log(n)}{|I|}}\right) \leq \left\{\frac{4H(c_H)\,C_\alpha}{c_H}\right\}\sqrt{\frac{\log(n)}{|I|}} := C_H\sqrt{\frac{\log(n)}{|I|}}.$$

□

Lemma 4.6.2. *Uniformly with probability* $1-\alpha+o(1)$ *for any interval* $I \subseteq \{1,\ldots,n\}$ *on which* $f_\circ(\cdot)$ *is non-decreasing it holds that* $f_\circ(\cdot)$ *must lie between the bounds* $\check{\mathbb{L}}_I^{\uparrow}$ *and* $\check{\mathbb{U}}_I^{\uparrow}$*. Likewise for any interval* I *on which* $f_\circ(\cdot)$ *is non-increasing it holds that* $f_\circ(\cdot)$ *must lie between the bounds* $\check{\mathbb{L}}_I^{\downarrow}$ *and* $\check{\mathbb{U}}_I^{\downarrow}$*.*

Proof. It is immediate from the construction in (4.8) that on any such I the function $\check{L}^{\uparrow} : I \to \mathbb{R}$ is point-wise smaller than the smallest non-decreasing function which produces empirical residuals which pass the test (4.5) and $\check{U}^{\uparrow} : I \to \mathbb{R}$ is point-wise larger than the largest non-decreasing function which produces empirical residuals which pass the same test. Finally by (4.11) with probability $1 - \alpha + o(1)$ the function $f_\circ(\cdot)$ produces residuals which pass the test (4.5). Therefore, on this event the true regression function must lie between the bounds. The same argument for $f_\circ(\cdot)$ non-increasing on I. □

Lemma 4.6.3. *Let $f_\circ(\cdot)$ be continuous, differentiable, and monotone on some interval $I \subseteq \{1, \ldots, n\}$ containing at least*

$$|I| \geq \lambda^2 \vee C\,(n/\psi)^{2/3} \log^{1/3}(n)$$

contiguous points, where $\psi = \inf_{t \in I} |f_\circ'(t/n)|$ and C depends on the constant C_H from Lemma 4.6.1. Then with probability $1 - \alpha + o(1)$ using the bounds (4.9) the monotonicity of $f_\circ(\cdot)$ can be correctly established on I.

Proof. Let $f_\circ(\cdot)$ be non-increasing on such an interval I containing $|I| := \upsilon$ points; for simplicity we assume $\upsilon/3$ is an integer. On this interval construct $\check{\mathbb{L}}_I^{\uparrow}$ and $\check{\mathbb{U}}_I^{\uparrow}$ using (4.9). Partition $I = I_1 \cup I_2 \cup I_3$ with $|I_1| = |I_2| = |I_3| = \upsilon/3$. By Lemma 4.6.1 with probability $1 - \alpha + o(1)$ we must have

$$\begin{aligned} \min_{t \in I_1} \left(f_\circ(t/n) - \check{L}_t^{\uparrow} \right) &\leq C_H \sqrt{3 \log(n)/\upsilon}, \\ \min_{t \in I_3} \left(\check{U}_t^{\uparrow} - f_\circ(t/n) \right) &\leq C_H \sqrt{3 \log(n)/\upsilon}. \end{aligned}$$

By their monotonicity the bounds will cross if for any $t_1, t_2 \in I$ with $t_1 > t_2$ we have $\check{L}_{t_1}^{\uparrow} > \check{U}_{t_2}^{\uparrow}$. Using the above we will surely have that $\check{L}_{\upsilon/3}^{\uparrow} > \check{U}_{2\upsilon/3}^{\uparrow}$ whenever

$$\begin{aligned} & f_\circ(\upsilon/3n) - C_H \sqrt{3 \log(n)/\upsilon} > f_\circ(2\upsilon/3n) + C_H \sqrt{3 \log(n)/\upsilon} \\ & \Rightarrow -(\upsilon/3n) \left[\frac{f_\circ(2\upsilon/3n) - f_\circ(\upsilon/3n)}{\upsilon/3n} \right] > 2C_H \sqrt{3 \log(n)/\upsilon} \\ & \Rightarrow (\upsilon/3n) \inf_{t \in I} \left| f_\circ'(t/n) \right| > 2C_H \sqrt{3 \log(n)/\upsilon} \\ & \Rightarrow \upsilon > C\,(n/\psi)^{2/3} \log^{1/3}(n)\,. \end{aligned}$$

Next by Lemma 4.6.2 on the same interval $\check{\mathbb{U}}^{\downarrow}$ and $\check{U}^{\downarrow}$ will not cross. Therefore on I we we establish $f_{\circ}\left(\cdot\right)$ is non-increasing. □

4.6.3 Proof of Theorem 4.3.1

Proof. This follow immediately from Lemma 4.6.2. □

4.6.4 Proof of Theorem 4.3.2

Proof. Consider testing for the k-th change point, with jump size Δ_k, on an interval I symmetric about the change point location containing $|I| = \upsilon > \lambda^2$ points and w.l.o.g. assume $f_{\circ}\left(\cdot\right)$ is non-decreasing on I. Partition $I = I_1 \cup I_2$ where $|I_1| = |I_2| = \upsilon/2$. See also Figure 4.4 for graphical intuition. By Lemma 4.6.1 with probability $1 - \alpha + o(1)$ for all $t \in I$ we have $\check{L}_t^{\uparrow} \leq f_{\circ}\left(t/n\right) \leq \check{U}_t^{\uparrow}$ and moreover

$$\begin{aligned}
&\min_{t \in I_1} \left(\check{U}_t^{\uparrow} - f_{\circ}(t/n)\right) \leq C_H \sqrt{2\log(n)/\upsilon}, \\
&\min_{t \in I_2} \left(f_{\circ}(t/n) - \check{L}_t^{\uparrow}\right) \leq C_H \sqrt{2\log(n)/\upsilon}.
\end{aligned}$$

Therefore using (4.11) and Proposition 4.6.2, with high probability the change point is detected as long as $\Delta_k > C_H \sqrt{8\log(n)/\upsilon}$ or equivalently as long as $\upsilon > 8C_H \log(n)/\Delta_k^2$. By the condition on the constants in Theorem 4.3.2 such an interval will not interfere with the intervals needed to detect the $(k-1)$-th and $(k+1)$-th change points which verifies event E_4^*. Finally by Theorem 4.3.1 with high probability no spurious change points are detected. □

4.6.5 Proof of Theorem 4.3.3

Proof. Consider testing for the k-th change point, with jump size Δ_k, on an interval I symmetric around the change point location and containing $|I| = \upsilon > \lambda^2$ points. We can assume $f_{\circ}\left(\cdot\right)$ is non-decreasing on I, since if the change point induces a local maximum / minimum Lemma 4.6.3 reveals this would be detected under the same conditions as

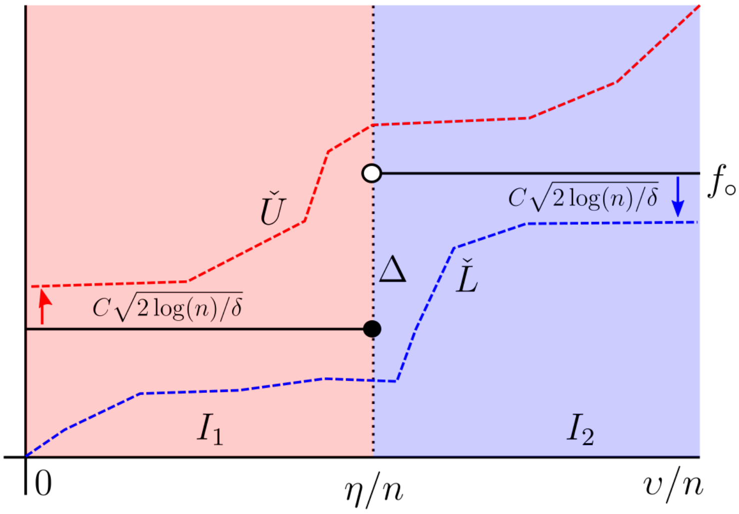


Figure 4.4: graphical intuition for the proof of Theorem 4.3.2.

stated in Theorem 4.3.3. Note that if the change point is detected using $\left(\check{\mathbb{L}}^{\uparrow}, \check{\mathbb{U}}^{\uparrow}\right)$ it will automatically be detected using $\left(\check{\mathbb{L}}^{\downarrow}, \check{\mathbb{U}}^{\downarrow}\right)$.

Without loss of generality let the change point occur in re-scaled time on the interval $[0, \upsilon/n]$ and hence $\eta_k = \lfloor \upsilon/2 \rfloor$. Consider partitioning $I = \bigcup_{j=1}^{5} I_j$ with I_3 symmetric about the change point location and $|I_1| = \cdots = |I_5| = \upsilon/5$. See also Figure 4.5 for graphical intuition. Lemma 4.6.1 again gives that with high probability we will have

$$\min_{t \in I_1} \left(f_\circ(t/n) - \check{L}_t^{\uparrow} \right) \leq C_H \sqrt{5\log(n)/\upsilon} \equiv \epsilon_{n,\upsilon},$$
$$\min_{t \in I_5} \left(f_\circ(t/n) - \check{L}_t^{\uparrow} \right) \leq C_H \sqrt{5\log(n)/\upsilon} \equiv \epsilon_{n,\upsilon}.$$

Next, introduce the strip $\mathcal{S}_{x_1,x_2} = \left\{(x, y) \in \mathbb{R}^2 \mid x_1 \leq x \leq x_2\right\}$ and $\mathbb{H}_{x_1,x_2}$ the open half plane consisting of all points which lie below the line interpolating $(x_1, f_\circ(x_1) - \epsilon_{n,\upsilon})$ and $(x_2, f_\circ(x_2) - \epsilon_{n,\upsilon})$. To fix the argument say the points in I_1 and I_5 at which $\check{\mathbb{L}}^{\uparrow}$ is closest to $f_\circ$ are t_1 and t_5 and put $x_i = t_i/n$. Then if at any point on I_3 the upper bound enters the trapezoid $\mathcal{T}_{x_1,x_5} = \mathbb{H}_{x_1,x_5} \cap \mathcal{S}_{x_1,x_5}$ the change point will be detected. Notice however that for $x_1' < x_1$ and / or $x_5' > x_5$ we must have that $\mathcal{T}_{x_1,x_5} \subset \mathcal{T}_{x_1',x_5'}$. Therefore, the least favourable points at which $\check{\mathbb{L}}^{\uparrow}$ can be close to $f_\circ(\cdot)$ are $t_1^* = \max(t \mid t \in I_1)$ and $t_5^* = \min(t \mid t \in I_5)$ and for this combination, recalling the parametrization (4.16), we have that

$$\mathbb{H}_{x_1^*, x_5^*} = \left\{ (x, y) \in \mathbb{R}^2 \mid y < \mu_k + \frac{1}{2}(\beta_k + \alpha_k)\, x - \left(\frac{1}{2} \cdot \frac{3}{5} \cdot \frac{\upsilon}{n}\right) \alpha_k - C_H \sqrt{5 \log n/\upsilon} \right\}.$$

For the interval I_3 which contains the change point Lemma 4.6.1 again gives that

$$\min_{t \in I_3} \left(\check{U}_t^{\uparrow} - f_\circ(t/n) \right) \leq C_H \sqrt{5 \log(n)/\upsilon}.$$

One of the two least favorable points at which this may occur is $t_3^* = \min(t \mid t \in I_3)$. Substituting we find that the point $(x_3^*, f_\circ(x_3^*) + \epsilon_{n,\upsilon})$ will lie inside $\mathcal{T}_{x_1^*, x_5^*}$ if for some C depending on C_H and powers of 5 we have that $(\beta_k - \alpha_k)\, \upsilon^{3/2} = \Delta_k \upsilon^{3/2} > Cn\sqrt{\log n}$. By the conditions on the constants such an interval will not interfere with the intervals needed to detect the $(k-1)$-th and $(k+1)$-th change points and on by Theorem 4.3.1 no spurious change points are detected. Therefore, the events E_5^*, E_6^*, E_7^* are verified. □

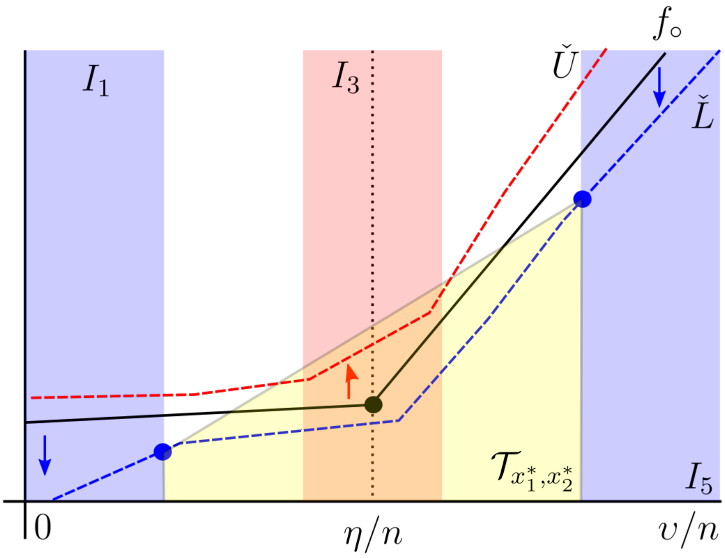


Figure 4.5: graphical intuition for the proof of Theorem 4.3.3.

4.6.6 Proof of Theorem 4.4.1

Proof. For any fixed $z \in \mathbb{R}$, omitting dependence on z, write $\mathfrak{u}_{n,W} = \left(\mathfrak{b}_{n,W} + z\right)/\mathfrak{a}_{n,W}$. Writing also $S_x = \sum_{t=1}^{x} \operatorname{sign}\left(\zeta_t\right)$ with $S_0 \equiv 0$ for the random walk generated by the sign process. We have by Assumption 4.4.2 that

$$M_{n,W}\left(\boldsymbol{\zeta}\right) \equiv \max_{\substack{0 \leq x \leq n-W \\ W \leq y \leq n-x}} \frac{S_{x+y} - S_x}{\sqrt{y}} = \tau \left[\max_{\substack{0 \leq x \leq n-W \\ W \leq y \leq n-x}} \mathfrak{X}\left(x, y\right) \right] + o_{\mathbb{P}}\left(\sqrt{\log\left(n/W\right)}\right) \tag{4.24}$$

where τ is as defined in Assumption 4.4.1. Without loss of generality we may take $\tau = 1$. We first consider the probability that first term on the left hand side of (4.24) is smaller than $\mathfrak{u}_{n,W}$. We have that:

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{\substack{0 \leq x \leq n-W \\ W \leq y \leq n-x}} \mathfrak{X}\left(x, y\right) \leq \mathfrak{u}_{n,W} \right) \geq \lim_{n\to\infty} \mathbb{P}\left(\sup_{(x,y)\in\mathbb{H}(n/W)} \mathfrak{X}\left(x, y\right) \leq \mathfrak{u}_{n,W} \right) = \exp\left(-e^{-z}\right). \tag{4.25}$$

Where we have used the scaling property of Brownian motion after the first inequality, and Theorem 4.6.1 for the first equality. We additionally have that:

$$\mathbb{P}\left(\max_{\substack{0 \leq x \leq n-W \\ W \leq y \leq n-x}} \frac{B\left(x+y\right) - B\left(x\right)}{\sqrt{y}} \leq \mathfrak{u}_{n,W} \right) \leq \mathbb{P}\left(\sup_{\substack{x\in[0,n/W] \\ y\in[1,n/W-x]}} \mathfrak{X}\left(x, y\right) \leq \mathfrak{u}_{n,W} \right) - R. \tag{4.26}$$

Where R is defined as follows:

$$R = \mathbb{P}\left(\sup_{\substack{x\in[0,n/W-1]\cap\mathbb{Z}\lfloor W\rfloor^{-1} \\ y\in[1,n/W-x]\cap\mathbb{Z}\lfloor W\rfloor^{-1}}} \mathfrak{X}\left(x, y\right) \leq \mathfrak{u}_{n,W} \right) - \mathbb{P}\left(\sup_{\substack{x\in[0,n/W-1] \\ y\in[1,n/W-x]}} \mathfrak{X}\left(x, y\right) \leq \mathfrak{u}_{n,W} \right).$$

Further, for some $A > 1$ and $a > 2$ the term can be further bounded as follows:

$$\begin{aligned}
|R| \leq & \sum_{i=0}^{\lfloor n/W \rfloor - 2} \left[\mathbb{P} \left(\sup_{\substack{x \in [i,i+1) \setminus \mathbb{Z} \lfloor W \rfloor^{-1} \\ y \in [1,A] \setminus \mathbb{Z} \lfloor W \rfloor^{-1}}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W}, \sup_{\substack{x \in [i,i+1) \cap \mathbb{Z} \lfloor W \rfloor^{-1} \\ y \in [1,A] \cap \mathbb{Z} \lfloor W \rfloor^{-1}}} \mathfrak{X}(x,y) \leq \mathfrak{u}_{n,W} \right) \right] \\
& + \mathbb{P} \left(\sup_{\substack{x \in (n/W-1, n/W] \\ y \in [1,A]}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W} \right) + \mathbb{P} \left(\sup_{\substack{x \in [0,n/W] \\ y \in (A, n/W - x]}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W} \right) \\
\leq & \sum_{i=0}^{\lfloor n/W \rfloor - 2} \left[\mathbb{P} \left(\sup_{\substack{x \in [i,i+1) \\ y \in [1,A]}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W} \right) - \mathbb{P} \left(\sup_{\substack{x \in [i,i+1) \cap \mathbb{Z} \lfloor a \log(n/W) \rfloor^{-1} \\ y \in [1,A] \cap \mathbb{Z} \lfloor a \log(n/W) \rfloor^{-1}}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W} \right) \right] \\
& + \mathbb{P} \left(\sup_{\substack{x \in (n/W-1, n/W] \\ y \in [1,A]}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W} \right) + \mathbb{P} \left(\sup_{\substack{x \in [0,n/W] \\ y \in (A, n/W - x]}} \mathfrak{X}(x,y) > \mathfrak{u}_{n,W} \right) \\
= & R_1 + R_2 + R_3.
\end{aligned}$$

We now bound each term in turn. Arguing as in the proof of Lemma 4.7 in Kabluchko (2007) we have that $\lim_{a \to \infty} \lim_{n \to \infty} R_1 = 0$. By Example 2.2 in Chan and Lai (2006), or equivalently Lemma 3.15 in Kabluchko (2007), we have that

$$R_2 = \left(\frac{W}{n} \right) \left(\frac{A-1}{A} \right) e^{-\tau} (1 + o(1)) \to 0 \quad \text{as} \quad n \to \infty.$$

Arguing as in the proof Lemma 4.4 in Kabluchko (2007) we have that $\lim_{A \to \infty} \lim_{n \to \infty} R_3 = 0$. By the above arguments bounding R, and Theorem 4.6.1, we therefore have that:

$$\limsup_{n \to \infty} \mathbb{P} \left(\max_{\substack{0 \leq x \leq n-W \\ W \leq y \leq n-x}} \mathfrak{X}(x,y) \leq \mathfrak{u}_{n,W} \right) \leq \exp\left(-e^{-z}\right). \tag{4.27}$$

Then (4.25) and (4.27) together imply that the first term in (4.24), appropriately scaled, converges to an extreme value distribution. By Khintchine's lemma, see Lemma 3.6.2 in Chapter 3, we have that (4.24) converges to an extreme value distribution. By the discussion in Section 3.2.3, $L_{n,W}(\boldsymbol{\zeta})$ behaves asymptotically live two independent copies of

$M_{n,W}(\boldsymbol{\zeta})$, which proves part (i) of Theorem 4.4.1. A further application of Khintchine's lemma proves part (ii) of the theorem. □

4.6.7 Proof of Lemma 4.4.1

Proof. Since the noise terms are assumed to have absolutely continuous densities we have the following for the signs of the residuals on each I_j block:

$$\begin{aligned}
\operatorname{sign}\left(\hat{\zeta}_{t,j}\right) &= 2 \times \mathbf{1}_{\left\{\zeta_t > \hat{f}_{J_{\sigma(j)}}(t/n) - f_\circ(t/n)\right\}} - 1 \\
&= 2 \times \mathbf{1}_{\{\zeta_t > 0\}} - 1 + 2\left[\mathbf{1}_{\left\{\zeta_t > \hat{f}_{J_{\sigma(j)}}(t/n) - f_\circ(t/n)\right\}} - \mathbf{1}_{\{\zeta_t > 0\}}\right] \\
&= \operatorname{sign}(\zeta_t) + 2\left[\mathbf{1}_{\left\{\zeta_t > \hat{f}_{J_{\sigma(j)}}(t/n) - f_\circ(t/n)\right\}} - \mathbf{1}_{\{\zeta_t > 0\}}\right] \\
&= \operatorname{sign}(\zeta_t) + V_{t,j}.
\end{aligned}$$

We also have the following bound for each $V_{t,j}$:

$$|V_{t,j}| \leq 2 \times \mathbf{1}_{\left\{|\zeta_t| \leq \left|\hat{f}_{J_{\sigma(j)}}(t/n) - f_\circ(t/n)\right|\right\}} \leq 2 \times \mathbf{1}_{\left\{|\zeta_t| \leq \left\|\hat{f}_{J_{\sigma(i)}} - f_\circ\right\|_{J,\infty}\right\}}. \tag{4.28}$$

For ease of notation introduce the following sets:

$$\begin{aligned}
\mathcal{J} &= \left\{j \in \mathbb{N} \mid 1 \leq j \leq \lfloor n/W' \rfloor\right\}, \\
\mathcal{K} &= \left\{k \in \mathbb{N} \mid 1 \leq k \leq \lfloor n/W'^2 \rfloor\right\}, \\
\mathcal{J}_\Theta &= \left\{j \in \mathcal{J} \mid \left|\Theta \cap J_{\sigma(j)}\right| > 0\right\}.
\end{aligned}$$

Based on the inequality (4.28) we have that for any $j \in \mathcal{J} \setminus \mathcal{J}_\Theta$

$$\frac{1}{\sqrt{W'}} \sum_{t \in I_j} V_{t,j} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log(W')}{W'}}\right). \tag{4.29}$$

To show this we define the events

$$\begin{aligned}
E_{t,j} &= \left\{ |\zeta_t| \leq \left\| \hat{f}_{J_{\sigma(j)}} - f_\circ \right\|_{J,\infty} \right\}, \\
A_t &= \left\{ |\zeta_t| \leq C_f \sqrt{2\log(W')}/W' \right\}, \\
B_{t,j} &= \left\{ \left\| \hat{f}_{J_{\sigma(j)}} - f_\circ \right\|_{J,\infty} \leq C_f \sqrt{2\log(W')}/W' \right\}.
\end{aligned}$$

Then, for any $\delta > 0$ and any $j \in \mathcal{J} \setminus \mathcal{J}_\Theta$ we have that

$$\begin{aligned}
\mathbb{P}\left(\left| \frac{1}{\sqrt{W'}} \sum_{t \in I_j} V_{t,j} \right| > \delta \right) &\leq 2\delta^{-1} \frac{1}{\sqrt{W'}} \sum_{t \in I_j} \mathbb{P}\left(E_{t,j}\right) \\
&= 2\delta^{-1} \frac{1}{\sqrt{W'}} \sum_{t \in I_j} \left[\mathbb{P}\left(E_{t,j} \mid B_{t,j}\right) \mathbb{P}\left(B_{t,j}\right) + \mathbb{P}\left(E_{t,j} \mid B_{t,j}^c\right) \mathbb{P}\left(B_{t,j}^c\right) \right] \\
&\leq 2\delta^{-1} \frac{1}{\sqrt{W'}} \sum_{t \in I_j} \left[\frac{\mathbb{P}\left(A_t \cap B_{t,j}\right) \mathbb{P}\left(B_{t,j}\right)}{\mathbb{P}\left(B_{t,j}\right)} + \mathbb{P}\left(B_{t,j}^c\right) \right] \\
&\leq 2\delta^{-1} \frac{1}{\sqrt{W'}} \sum_{t \in I_j} \left[\mathbb{P}\left(A_t\right) + \mathbb{P}\left(B_{t,j}^c\right) \right] \\
&= 2\delta^{-1} \frac{1}{\sqrt{W'}} \sum_{t \in I_j} \left[\mathcal{O}\left(\frac{\sqrt{\log(W')}}{W'} \right) + \mathcal{O}\left(W'^{-1}\right) \right] \\
&= \mathcal{O}\left(\delta^{-1} \sqrt{\frac{\log(W')}{W'}} \right),
\end{aligned}$$

which proves the statement. In the penultimate line we have used the fact that since the ζ's have bounded and continuous densities for any $x > 0$ we must have that $\mathbb{P}\left(|\zeta_t| \leq x\right) \leq \mathcal{O}\left(x\right)$. To prove the main result we simply bound the absolute difference between our estimator for the TAVC and the long run variance:

$$\begin{aligned}
\left|\widehat{\mathrm{TAVC}}\left(W'\right)-\tau^2\right| &= \left|\lfloor n/W' \rfloor^{-1}\left[\sum_{j\in\mathcal{J}\setminus\mathcal{J}_\Theta}+\sum_{j\in\mathcal{J}_\Theta}\right]\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}\mathrm{sign}\left(\hat{\zeta}_{t,j}\right)\right)^2-\tau^2\right| \\
&\le \left|\lfloor n/W' \rfloor^{-1}\sum_{j\in\mathcal{J}\setminus\mathcal{J}_\Theta}\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}\mathrm{sign}\left(\hat{\zeta}_{t,j}\right)\right)^2-\tau^2\right| \\
&+ \left|\lfloor n/W' \rfloor^{-1}\sum_{j\in\mathcal{J}_\Theta}\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}\mathrm{sign}\left(\hat{\zeta}_{t,j}\right)\right)^2\right| \\
&= I_1+I_2.
\end{aligned}$$

For the second term we have that $I_2 \le \mathcal{O}\left(\frac{NW'^3}{n}\right)$ since $\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}\hat{\zeta}_{t,j}\right)^2 \le W'$ for each j and $|\mathcal{J}_\Theta| \le NW'$. For the second term we have that

$$\begin{aligned}
I_1 &\le \left|\mathrm{TAVC}\left(W'\right)-\tau^2\right| \\
&+ \left|\lfloor n/W' \rfloor^{-1}\sum_{j\in\mathcal{J}\setminus\mathcal{J}_\Theta}\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}\mathrm{sign}\left(\zeta_t\right)\right)^2-\mathrm{TAVC}\left(W'\right)\right| \\
&+ \left|\lfloor n/W' \rfloor^{-1}\sum_{j\in\mathcal{J}\setminus\mathcal{J}_\Theta}\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}V_{t,j}\right)^2\right| \\
&+ \left|\lfloor n/W' \rfloor^{-1}\sum_{j\in\mathcal{J}\setminus\mathcal{J}_\Theta}2\left(\frac{1}{\sqrt{W'}}\sum_{t\in I_j}\mathrm{sign}\left(\zeta_t\right)\right)\left(\frac{1}{\sqrt{W}}\sum_{t\in I_j}V_{t,j}\right)\right| \\
&= I_{1,1}+I_{1,2}+I_{1,3}+I_{1,4}.
\end{aligned}$$

For the first term we have that $I_{1,1} \le \mathcal{O}\left(W'^{-1}\right)$, which holds because:

$$\begin{aligned}
I_{1,1} &= 2\left|\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_h-\left\{\sum_{h=1}^{W'-1}+\sum_{h=W'}^{\infty}\right\}\gamma_h\right| \\
&\le \frac{2}{W'}\sum_{h=1}^{W'-1}h\left|\gamma_h\right|+2\sum_{h=W}^{\infty}\left|\gamma_h\right| \le \frac{2}{W'}\sum_{h=1}^{\infty}h\left|\gamma_h\right| = \mathcal{O}\left(W'^{-1}\right).
\end{aligned}$$

The passage to the final line follows from Assumption 4.4.5. For the second term we have that $I_{1,2} \leq \mathcal{O}_{\mathbb{P}}\left(\frac{W'}{\sqrt{n}} \vee \frac{NW'^2}{n}\right)$ since:

$$
\begin{aligned}
I_{1,2} &= \left| \lfloor n/W' \rfloor^{-1} \sum_{j \in \mathcal{J} \setminus \mathcal{J}_\Theta} \left(\frac{1}{\sqrt{W'}} \sum_{t \in I_j} \operatorname{sign}(\zeta_t) \right)^2 - \left[1 + 2 \sum_{h=1}^{W'-1} \left(1 - \frac{h}{W'} \right) \gamma_h \right] \right| \\
&\leq \left| \lfloor n/W' \rfloor^{-1} \sum_{j \in \mathcal{J}_\Theta} \left(\frac{1}{\sqrt{W'}} \sum_{t \in I_j} \operatorname{sign}(\zeta_t) \right)^2 \right| \\
&+ \left| \lfloor n/W' \rfloor^{-1} \sum_{j \in \mathcal{J}} \frac{1}{W'} \sum_{\substack{t,s \in I_j \\ t \neq s}} \left(\operatorname{sign}(\zeta_t) \operatorname{sign}(\zeta_s) \right) - 2 \sum_{h=1}^{W'-1} \left(1 - \frac{h}{W'} \right) \gamma_h \right| \\
&= \mathcal{O}_{\mathbb{P}}\left(\frac{W' \left| \mathcal{J}_\Theta \right|}{n} \right) + \left| \lfloor n/W' \rfloor^{-1} \sum_{j=1}^{\lfloor n/W' \rfloor} \frac{2}{W'} \sum_{h=1}^{W'-1} \sum_{t=1+(j-1)W'}^{jW'-h} \left(\operatorname{sign}(\zeta_t) \operatorname{sign}(\zeta_{t+h}) \right) - \frac{2}{W'} \sum_{h=1}^{W'-1} (W'-h) \gamma_h \right| \\
&= \mathcal{O}_{\mathbb{P}}\left(\frac{NW'^2}{n} \right) + \left| \frac{2}{\lfloor n/W' \rfloor (W'-h)} \sum_{h=1}^{W'-1} \left(1 - \frac{h}{W'} \right) \sum_{j=1}^{\lfloor n/W' \rfloor} \sum_{t=1+(j-1)W'}^{jW-h} \left(\operatorname{sign}(\zeta_t) \operatorname{sign}(\zeta_{t+h}) - \gamma_h \right) \right| \\
&\leq \mathcal{O}_{\mathbb{P}}\left(\frac{NW'^2}{n} \right) + 2 \sum_{h=1}^{W'-1} \left| \frac{1}{\lfloor n/W' \rfloor (W'-h)} \sum_{j=1}^{\lfloor n/W' \rfloor} \sum_{t=1+(j-1)W'}^{jW-h} \left(\operatorname{sign}(\zeta_t) \operatorname{sign}(\zeta_{t+h}) - \gamma_h \right) \right| \\
&= \mathcal{O}_{\mathbb{P}}\left(\frac{NW'^2}{n} \right) + 2 \sum_{h=1}^{W'-1} \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\frac{h}{W'})}} \right) \\
&= \mathcal{O}_{\mathbb{P}}\left(\frac{NW'^2}{n} \right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \right) \sum_{h=1}^{W'-1} \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{1-\frac{h}{W'}}} \right) \\
&\leq \mathcal{O}_{\mathbb{P}}\left(\frac{NW'^2}{n} \right) + \mathcal{O}_{\mathbb{P}}\left(\frac{W'}{\sqrt{n}} \right).
\end{aligned}
$$

In the last line we have used the fact that

$$
\sum_{h=1}^{W'-1} \frac{1}{\sqrt{\left(1-\frac{h}{W'}\right)}} < \int_1^{W'-1} \frac{1}{\sqrt{1-\frac{x}{W'}}} \mathrm{d}x + \sqrt{W'} = 2W' \left(1 + o(1)\right).
$$

For the third term we have that $I_{1,3} = \mathcal{O}_{\mathbb{P}}\left(\frac{\log(W')}{W'}\right)$. This follows directly from (4.29). For the fourth term we have that $I_{1,4} \leq \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log(W')}{W'}}\right)$. This holds because

$$I_{1,4} = \lfloor n/W' \rfloor^{-1} \sum_{j \in \mathcal{J} \setminus \mathcal{J}_\Theta} \mathcal{O}_{\mathbb{P}}\left(1\right) \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log(W')}{W'}}\right) \leq \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log(W')}{W'}}\right).$$

Combining the bounds on each of the terms the desired result follows. □

5 Recovering Dependence Structures in Change Point Regressions with a View to Causality

5.1 Introduction and problem statement

In this chapter we a study a change point problem in which an analyst observes multivariate data where the change points themselves are random variables, and change points in one series may cause change points in another series. We introduce an algorithm for estimating graphs which encode causal information about the change points. Typically after performing change point analysis relationships between estimated change points can only be described qualitatively, since in general change point locations are held to be unknown but non-stochastic. From the perspective of practitioners this is a limitation, since in many settings it is reasonable to believe change points will be causally linked. For example: in climatology changes in concentrations levels of certain gasses actively cause changes in the environment (Schmittner et al., 2018), in finance (Schröder and Fryzlewicz, 2013) changes in the behavior of financial instruments, such as equities and commodities, in one market can cause changes in the behavior of other such assets in related markets, and in epidemiology (Kartal et al., 2021; Mastakouri and Schölkopf, 2020) changes in the behavior of a population can cause changes in the spread of a virus.

Previous works (Eichinger and Kirch, 2018) have studied change point problems in which the change point locations are random variables. Moreover, some authors (Hallgren et al., 2023; Fotoohinasab et al., 2020) have considered multivariate time series time series indexed over a graph, where series belonging to the same community are likely to undergo changes at similar times. However, to the best of our knowledge the problem of recovering causal relations among change points has not been studied in the literature.

The remainder of the chapter is structured as follows. In Section 5.2 we introduce the change point model, and precisely define the causal graph we aim to recover. In Section 5.3 we introduce a procedure for recovering the causal graph from data, and in Section 5.4 we give conditions under which our procedure can consistently recover the graph. Finally, in Section 5.5 we illustrate the performance of the procedure via a simulation study and a real data example.

5.2 Model set-up

We introduce a random change point model, in which at each time step a d-dimensional vector is observed, and each element of the vector is liable to experience a change driven by an unobserved counting process. To that end briefly review the multi-type Hawkes process in Section 5.2.1 before presenting in full the data generating mechanism in Section 5.2.2 and formally defining the object we aim to recover in Section 5.2.3. Our goal will be to recover a graph where the presence of an edge $i \to j$ signifies that change points in the i-th component of the observed vector cause change points in j-th component. On a high level, our approach to recovering this graph relies on properties of three separate time scales: the scale at which the data are observed, the scale at which the underlying counting process generates events, and the scale at which an appropriate algorithm can localize the unobserved change points. As we show formally in Section 5.4, as soon as these time scales are sufficiently different the desired graph can be consistently recovered from data.

5.2.1 Review of multi-type Hawkes processes

In the seminal paper Hawkes (1971), Hawkes introduced the mutually exciting point process, now commonly referred to as the marked Hawkes process, in which past events are able to trigger occurrences of future events of different types. Consider a d-dimensional counting process $\boldsymbol{N}(\cdot)$ whose j-th component generates events η_j on $\mathbb{R}_+$. Let $\mathcal{F}_{t-}$ be the natural filtration generated by the events from each component process up to but not including time t. The conditional intensity for the j-th component process is given by

$$\lambda_j^*(t) = \lim_{q \downarrow 0} \frac{\mathbb{E}\left(N_j[t, t+q) \mid \mathcal{F}_{t-}\right)}{q}. \tag{5.1}$$

Under quite general conditions the conditional intensity function exists and completely characterizes the distribution of the random counting measure $N_j(\cdot)$. The entire process $\boldsymbol{N}(\cdot)$ is called a multi-type Hawkes process if the conditional intensities of each component process are of the form

$$\lambda_j^*(t) = \nu_j + \sum_{i=1}^{d} \int_0^{\infty} g_{j,i}(t-u)\,\mathrm{d}N(u), \qquad j = 1, \ldots, d. \tag{5.2}$$

In (5.2) the constants $\nu_j > 0$ are called the base-line intensities, and the functions $g_{j,i} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ are are called the Hawkes kernels. Each $g_{j,i}(\cdot)$ captures the effect of a past event of type i the intensity of a future event of type j; in this sense, the marked Hawkes process can be seen as the counting process analogue of the vector auto-regression (see Example 2.3.1). Restricting the Hawkes kernels to be positive is necessary, as otherwise the conditional intensity function may be negative. Of great importance are the quantities $m_{j,i} = \int_0^\infty g_{j,i}(u)\,\mathrm{d}u$, which are often referred to as the branching coefficients due to the Hawkes process's interpretation as a cluster process in which cluster centers produce offspring distributed according to a multi-type branching process; see Definition 5.6.2, as well as Example 6.3(c) in Daley and Vere-Jones (2003). Provided the largest eigenvalue of the matrix $\boldsymbol{M} = (m_{j,i})_{i,j=1}^d$ is less than one the number of points in any bounded sub-interval of $\mathbb{R}_+$ is almost surely finite, in which case the Hawkes process is said to be simple.

5.2.2 A random change point model

We now introduce the data generating process proper. We consider a multivariate time series $\{\mathbb{Y}_t = (Y_{1,t}, \ldots, Y_{d,t})' \mid t = 1, \ldots, n\}$ where each of the component processes is generated by a change point model with piecewise constant mean. That is, associated with each Y_j is a sequence of ordered change point locations $\Theta_j = \left\{\eta_{j,1}, \ldots, \eta_{j,N_j}\right\}$, and of the Y_j's is generated according to

$$Y_{j,t} = \mu_j + \sum_{k=1}^{N_j} \Delta_{j,k} \mathbf{1}_{\left\{t>\eta_{j,k}\right\}} + \zeta_{j,t}, \qquad t = 1, \ldots, n. \tag{5.3}$$

Following the usual convention in change point problems, we additionally have that $\eta_{j,0} = 0$ and $\eta_{j,N_j+1} = n$ for all j. The μ's constitute a sequence of finite constants, and we let each $\{\Delta_{j,k} \mid k = 1, \ldots, N_k\}$ be a sequence of random variables where different sequences are not necessarily independent.

We model the change point locations in (5.3) as arrival times of events associated with a multi-type Hawkes process on the positive real line: $\boldsymbol{N}(\cdot) = (N_1(\cdot), \ldots, N_d(\cdot))'$. The standard Hawkes process generates events which are $\mathcal{O}(1)$ apart with high probability. However, it is well known that consistent change point detection is impossible if the distance between change points is fixed. We therefore allow the conditional intensities to depend on the sample size n, and to emphasise the temporal scaling of $\boldsymbol{N}(\cdot)$ we write

$$\lambda_j^*(t) = \nu_j + \sum_{i=1}^{d} \int_0^\infty g_{j,i}^{(n)}(t-u)\,\mathrm{d}N(u), \qquad j = 1, \ldots, d. \tag{5.4}$$

where for each (i,j) pair we have that $g_{j,i}^{(n)}(\cdot) = \xi_n \times g_{j,i}(\cdot \times \xi_n)$ for some $\xi_n = \mathcal{O}(n^\Psi)$, where $\Psi > 0$ and $g_{i,j}(\cdot)$ is a function which does not depend on n. Therefore, the change points in each component series are the marginal arrival times of an unobserved process $\boldsymbol{N}(\cdot)$, and dependence between change points in different series is captured by the branching coefficients $\{m_{i,j} = \int g_{i,j}^{(n)} \mid i,j \in \{1, \ldots, d\}\}$.

5.2.3 Causal graphs for the Hawkes process

We now formally introduce the object we wish to recover. Didelez (2008) introduced local independence as a dynamic concept of dependence for marked point processes. For a multi-type process $\boldsymbol{N}(\cdot) = (N_1(\cdot), \ldots, N_d(\cdot))'$ we say that $N_j(\cdot)$ is locally independent of $N_i(\cdot)$

if $\lambda_j^*(t)$ is $\mathcal{F}_{t\setminus i}$ measurable for all $t > 0$. In which case we write $i \not\to j$, else we speak of local dependence and write $i \to j$. See Section 2.3.3 for a detailed discussion. In case the point process is completely observed, in the sense of there being no unobserved confounders, following the reasoning in Granger (1969) the statement $i \to j$ can be interpreted causally in the sense of the process i causing process j. From (5.2) it is clear that $m_{j,i} = 0$ is a necessary and sufficient condition for $i \not\to j$ in the marked Hawkes process. That is, the corresponding branching coefficient should be zero. Revisiting the goal stated in Section 5.2.2, we aim more specifically to recover the graph $\mathcal{G}$ with vertex set $V = \{1, \ldots, d\}$ and edge set

$$E_{i,j} = \begin{cases} 1 & \text{if } m_{j,i} > 0 \\ 0 & \text{if } m_{j,i} = 0 \end{cases} \tag{5.5}$$

from observations (5.3). In Section 5.3 below we introduce an algorithm for this task.

5.3 Causal structure recovery

If the Hawkes process generating the change point locations were observable, we would simply apply any consistent estimation procedure for the parameters of the process to the observed arrival times then estimate (5.5) according to whether each branching coefficient obtained from the estimates was significantly different from zero. Since we only have access to the Hawkes process through (5.3), we first estimate the change point locations via some generic consistent method, then estimate the branching coefficients by treating the estimated change point locations as the true unobserved arrival times of the Hawkes process. A generic algorithm for estimating $\mathcal{G}$ in this way is given below in Algorithm 7.

In order to make Algorithm 7 operational we should specify: a change point estimation procedure $\hat{\Theta}(\cdot)$, a branching ratio estimation procedure $\hat{\boldsymbol{M}}(\cdot)$, and a set of thresholds $\boldsymbol{\lambda}$. We defer discussion of $\hat{\Theta}(\cdot)$ and $\boldsymbol{\lambda}$ to Section 5.4, and conclude this section by presenting a procedure for estimating the branching ratios. Given the estimated change point locations, we choose to estimate the branching ratios non-parametrically using a procedure based on conditional least squares which was independently proposed by Kirchner (2017) and Eichler

Algorithm 7: A generic algorithm for estimating the local independence graph for random change point locations in model (5.3), where change point locations are arrival times of a Marked Hawkes process.

```
Input: The data $\{\mathbb{Y}_t \mid t = 1, \dots, n\}$; a procedure for estimating change points
       from data $\hat{\Theta}(\cdot)$; a procedure for estimating branching ratios $\hat{\boldsymbol{M}}(\cdot)$; a set
       of thresholds $\boldsymbol{\lambda} = \{\lambda_{j,i} \mid j, i = 1, \dots, d\}$.
Output: An estimator $\hat{\mathcal{G}}^{\lambda} = (V, \hat{E})$ for the local independence graph.
begin
    $\hat{\boldsymbol{\Theta}} \leftarrow \hat{\boldsymbol{\Theta}}(\mathbb{Y})$
    $\{\hat{m}_{j,i}\}_{j,i=1}^{d} \leftarrow \hat{\boldsymbol{M}}(\hat{\boldsymbol{\Theta}})$
    for $i, j = 1, \dots, d$ do
        $\hat{E}_{i,j} \leftarrow \mathbf{1}_{\{\hat{m}_{j,i} > \lambda_{j,i}\}}$
    end
end
```

et al. (2017). The procedure is described in detail in the following section.

5.3.1 Hawkes process parameter estimation via conditional least squares

In this section only we consider the multi-type Hawkes process from (5.2) observed on the interval $[0, n]$, whose base line intensities and kernels do not depend on n. The method proposed by Kirchner (2017) and Eichler et al. (2017) for estimating the parameters of (5.2) is based on binning the interval $[0, n]$ and counting the number of events in each bin. Recall that a stationary point process $N(\cdot)$ is said to be orderly if $\mathbb{P}(N(q) > 1) = o(q)$ as $q \downarrow 0$. Let $\boldsymbol{N}(\cdot) = (N_1(\cdot), \dots, N_d(\cdot))'$ be a simple, stationary, and orderly marked Hawkes process. Writing $X_{j,t}^{q} = N_j(tq) - N_j((t-1)q)$ for q we will have that

$$\begin{aligned}\mathbb{E}\left(X_{j,t+1}^{q} \mid \mathcal{F}_{tq}\right) &= q\nu_j + q\sum_{i=1}^{d}\int_0^{\infty} g_{j,i}(qt-u)\,\mathrm{d}N_i(u) + o(q)\\ &= q\nu_j + q\sum_{i=1}^{d}\sum_{s=1}^{\infty}\int_0^{u} g_{j,i}(q(t-s)-u)\,\mathrm{d}N_i(u) + o(q)\\ &\approx q\nu_j + q\sum_{i=1}^{d}\sum_{s=1}^{\infty} g_{j,i}(sq)\,X_{j,t-s}^{q}. \qquad (5.6)\end{aligned}$$

Kirchner (2017) and Eichler et al. (2017) therefore truncate the infinite sum in (5.6) at some value k which diverges with n and estimate the parameters of the Hawkes process via conditional least squares. More precisely, for some small q which goes to zero with n write $\mathbb{X}_t^q = \left(X_{1,t}^q, \ldots, X_{d,t}^q\right)'$ for $t = 1, \ldots \lfloor n/q \rfloor$. Then the parameters $\boldsymbol{\nu}^q = (q\nu_1, \ldots, q\nu_d)'$ and $\boldsymbol{G}^q = (\boldsymbol{G}_s^q \mid 1 \leq s \leq k)'$, where each $\boldsymbol{G}_s^q$ is a $d \times d$ matrix with (j,i)-th entry $\boldsymbol{G}_{q,j,i}^q = q \times g_{j,i}(sq)$, can be estimated via

$$\left\{\hat{\boldsymbol{\nu}}^q, \hat{\boldsymbol{B}}^q\right\} = \underset{\boldsymbol{\nu}^q, \boldsymbol{B}^q}{\arg\min} \sum_{t=k+1}^{\lfloor T/q \rfloor} \left\| \mathbb{X}_t^q - \boldsymbol{\nu}^q - \sum_{s=1}^{k} \boldsymbol{G}_s^q \mathbb{X}_{t-s}^q \right\|^2 . \tag{5.7}$$

Intuitively, provided the Hawkes kernels do not decay to zero too slowly, k diverges at a fast enough rate, and q is sufficiently small, the step function

$$\hat{g}_{j,i}^q(u) = q^{-1} \sum_{s=1}^{k} \hat{\boldsymbol{G}}_{s,j,i}^q \mathbf{1}_{\{(s-1)q \leq u < sq\}}$$

will approximate the corresponding Hawkes kernel well. Therefore, the branching ratios can be estimated via $\hat{m}_{j,i} = \sum_{s=1}^{k} \hat{\boldsymbol{G}}_{s,j,i}^q$.

5.3.2 Parameter estimation for Hawkes processes scaling with n

The procedure described above can be adapted to estimate the branching coefficients for a Hawkes process whose kernels scale with n, as in (5.4). The parameters q and k however must be chosen differently. For instance, one can no longer allow q to decay to zero. Such considerations are addressed in the next section. An important consequence of using the procedure from Section 5.3.1 to estimate the parameters of (5.4) is that, since we work with binned observations, that provided the localization rate of $\hat{\Theta}(\cdot)$ smaller than the bin width used, on a high probability set working with the estimated change point locations is the same as working with the unobserved arrival times of the underlying Hawkes process.

5.4 Theoretical results

We now study the conditions under which the algorithm proposed is able to recover the local independence graph for the underlying Hawkes process. In terms of the Hawkes process generating the change point locations, we require the following assumptions to hold.

Assumption 5.4.1. *The number of component series satisfies $d = \mathcal{O}(1)$ and in each series that the ζ's are mutually independent sub-Gaussian.*

Assumption 5.4.2. *The base-line intensities for each component process in (5.2) are given by $\nu_j = C_j n^{-(1/2+\phi)}$ where $\phi \in (0, 1/2)$ and each C_j is a finite constant. Additionally, there are constants $\underline{m}$ and $\overline{m}$, with $0 < \underline{m} < \overline{m} < 1$ if $E_{i,j} = 1$ the associated branching coefficient satisfies $m_{j,i} \in (\underline{m}, \overline{m})$.*

Assumption 5.4.3. *Put $h_{j,i}(\cdot) = g_{j,i}^{(n)}(\cdot)/m_{j,i}$ for the (j,i)-th Hawkes kernel standardized by the associated branching ratio so that it becomes a density function, $\bar{H}_{j,i}(u) = \int_u^\infty h_{j,i}(z)\,\mathrm{d}z$, and $\breve{h}_{j,i}(u) = \sup_{y\geq 0} \int_{0\vee(y-u)}^{y+u} h_{j,i}(z)\,\mathrm{d}z$. There are quantities Ψ_1, Ψ_2, and Ψ_3, each diverging with n, which satisfy the following conditions:*

(i) $\Psi_1 = o\left(n^{2\phi}\right)$ *and* $\Psi_3 = o\left(n^{\phi}\right)$,

(ii) $\Psi_2 = o(\Psi_1^{1/2})$ *and* $m_{j,i}^{\Psi_2} \vee \bar{H}_{j,i}(\Psi_2) = o\left(n^{\phi-1/2}\right)$,

(iii) $\Psi_3 = o\left(\Psi_2^2\right)$ *and* $\breve{h}_{j,i}(\Psi_3) = o\left(n^{\phi-1/2}\right)$.

Assumption 5.4.4. *The Hawkes kernes are Lipschitz continuous with $|g_{j,i}(u)| = \mathcal{O}(u^{-1})$ as $u \to \infty$, recalling that $g_{j,i}^{(n)}(\cdot) = \xi_n \times g_{j,i}(\cdot \times \xi_n)$. Moreover for k and q chosen according to Assumption 5.4.5 below we have that $\bar{H}_{ji}(qk) = o(1)$.*

The first part of Assumption 5.4.2 lets the expected number of change points in the data diverge with the sample size, but not too quickly. The second part of the assumption guarantees that the underlying Hawkes process is simple. As can be seen in the proof of Theorem 5.4.1, Assumption 5.4.3 guarantees that with high probability the distance between any two change points will not be smaller than n^{ϕ}.

We also need the following assumption on the tuning parameters k and q, the distribution of the jump sizes in (5.3), and the generic algorithm used to recover the change points.

Assumption 5.4.5. *The truncation level and bin width are chosen such the* $k = \mathcal{O}\left(n^{\theta_1}\right)$ *and* $q = \mathcal{O}\left(n^{\theta_2}\right)$*, where* θ_1 *and* θ_2 *satisfy: (i)* $\theta_2 \leq \phi$*, (ii)* $\theta_1 + \theta_2 > \phi + 1/2$*, and (iii)* $\theta_1/2 + \theta_2 < \phi + 1/2$.

Assumption 5.4.6. *There is a constant* C_Θ *such that, when applied to data where the shortest distance between two change points (*δ*) and the smallest change size (*Δ*) satisfy* $\sqrt{\delta}\Delta > C_\Theta\sqrt{\log(n)}$*,* $\hat{\Theta}\left(\cdot\right)$ *is able to detect all change points and the k-th change point in the j-th series is localized at least at the rate* $\mathcal{O}\left(\log(n)/\Delta_{j,k}^2\right)$*, on a set with probability* $1 - o(1)$.

Assumption 5.4.7. *Each of the* $\Delta_{j,k}$*'s has marginal distribution* $F_{j,k}$ *with support* $(-\infty, -S_{j,k}) \cup (S_{j,k}, \infty)$*, where* $S_{j,k}$ *is at least of the order* $\mathcal{O}\left(\sqrt{\log(n)/n^{\theta_3}}\right)$ *for some* $\theta_3 < \theta_2$.

Given that according to Assumption 5.4.1 the number of component series is fixed and the contaminating noise is sub-Gaussian, Assumption 5.4.6 is quite mild, and would hold if any rate optimal change point detection algorithm for univariate data was applied componentwise to the data. Assumption 5.4.7 implies that $\hat{\Theta}\left(\cdot\right)$ will be able to localize each change point at least at the rate $\mathcal{O}\left(n^{\theta_3}\right)$. Following the discussion around Assumption 5.4.3 we will therefore have, on a high probability set, that: change point localization rate $\ll$ bin width $\ll$ minimum change point spacing. With these assumptions in place, we have the following result.

Theorem 5.4.1. *Let* $\{\mathbb{Y}_t \mid t = 1, \ldots, n\}$ *be data from model (5.3) and let* $\hat{\mathcal{G}}^{\boldsymbol{\lambda}}$ *be the output from Algorithm 7 run with parameters* $\boldsymbol{\lambda}$*, k, and q, where the change points are estimates via some generic method* $\hat{\Theta}\left(\cdot\right)$*. Grant Assumptions 5.4.2-5.4.7 holds. Then if* $\boldsymbol{\lambda}$ *is chosen such the smallest element is the set is larger than some* $\underline{\lambda}$*, where* $\sqrt{kq/n} = o\left(\underline{\lambda}\right)$*, it holds that* $\mathbb{P}\left(\hat{\mathcal{G}}^{\boldsymbol{\lambda}} = \mathcal{G}\right) \to 1$ *as* $n \to \infty$.

Regarding practical choices for k and q, if ϕ were known and we were to set $q = \mathcal{O}\left(n^{\phi}\right)$ Assumption 5.4.5 would be satisfied if we set $k = \lfloor n^{0.5+\varepsilon} \rfloor$, where ε is some small positive constant. In practice we can directly choose k in this way, then set $q = \left\lfloor n^{\hat{\phi}} \right\rfloor$ where $\hat{\phi}$ is a rough estimator of ϕ based on the minimum space between any two estimated change point locations. Regarding the choice of $\boldsymbol{\lambda}$, in Remark 1 Kirchner (2017) points out that the coefficients estimated according to (5.7) will be approximately normally distributed. Therefore, we can take $\lambda_{j,i} = z_\alpha \times \mathrm{se}\left(\hat{m}_{j,i}\right)$, where $\mathrm{se}\left(\hat{m}_{j,i}\right)$ stands for the standard error of the estimated branching coefficient. We follow this approach in the numerical experiments below, using a value of $\alpha = 0.1$.

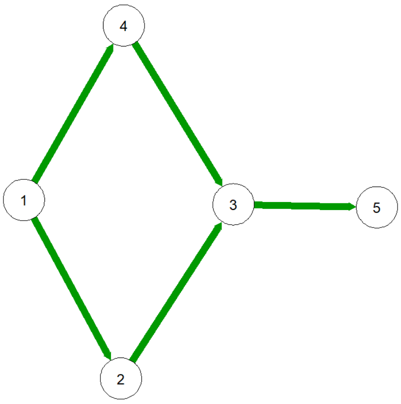
5.5 Numerical illustrations

5.5.1 Simulation study

We investigate the performance of Algorithm 7 via a small simulation study. We simulate $d = 5$ dimensional time series data from model (5.3), where we let the ζ's be mutually independent with marginal $\mathcal{N}(0, 1)$ distribution. Regarding the Hawkes process generating the change point locations, we let its local independence graph be as shown in Figure 5.1 below. We set all of the base line intensities equal to $n^{-0.6}$ and when non-zero we set each of the Hawkes kernels to $g(u) = 0.6n^{-0.3} \exp\left(-0.8n^{-0.3}\right)$. Finally, the Δ's are generated as mutually independent uniform random variables with supports $\left[C_\Delta/2\sqrt{\log(n)/\delta_{j,k}}, C_\Delta\sqrt{\log(n)/\delta_{j,k}}\right]$, where C_Δ is a constant representing the signal strength of the change points to be chosen later on.

We simulate the data for sample sizes $n \in \{500, 1000, 2000\}$ and for signal strengths $C_\Delta \in \{2, 5, 7\}$. Over 100 replications we estimate the underlying local independence graph using bid width $q = \left\lfloor n^{0.1} \right\rfloor$ and truncation level $k = \left\lfloor n^{0.501} \right\rfloor$, and apply the thresholding rule discussed at the end of Section 5.4. The results of the simulation study are shown in Figure 5.2, where each sub-plot displays a graph with edges weighted by the number of times the corresponding edge appeared in an estimate of the local independence graph. From the plots we see that, provided the signal strength is large, the local independence graph can be recovered reasonably well.

Figure 5.1: Local independence graph for the Hawkes process generating change point locations the simulation study presented in Section 5.5.1, with edges weighted by the value of the associated branching coefficients.



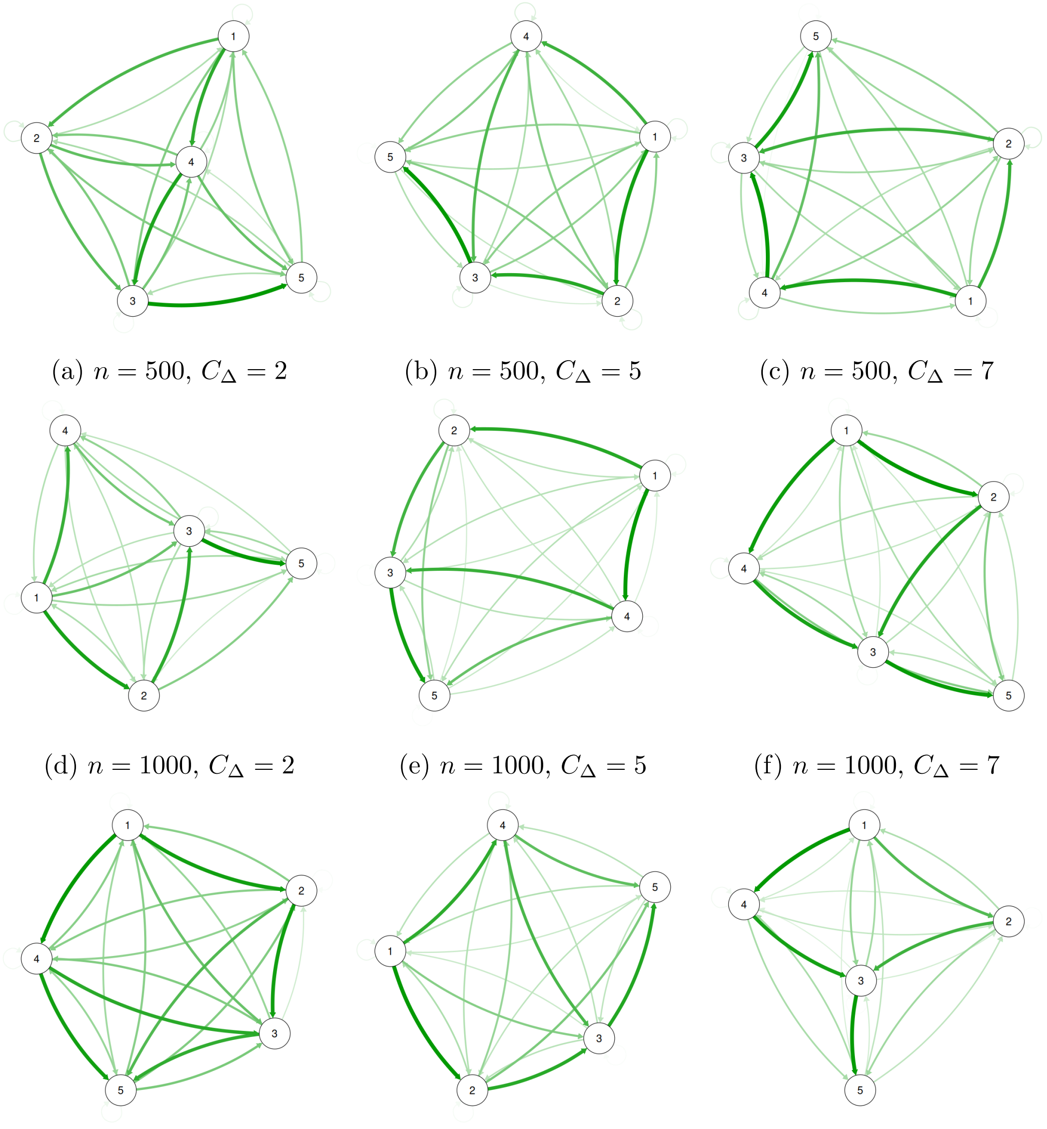
5.5.2 Real data analysis: COVID-19 infection trajectories

We study change points in the number of daily deaths, daily hospitalizations, and daily infections due to the COVID-19 virus between the years 2020 and 2022, across all London boroughs. The data we analyze can be obtained from the Office for National Statistics. We model the data as following a piecewise linear trend, and after applying the Anscombe transform to correct for the fact that we are working with integer-valued data, estimate the trend and the change points using the Narrowest Over Threshold algorithm of Baranowski et al. (2019a). The data, along with the estimated trends, is plotted in Figure 5.3.

It is tempting to think of the change points in the data as causally linked, where perhaps we would have that: Cases → Hospitalizations → Deaths. In Figure 5.4 we plot the times of the change point locations recovered from the data. Based on this plot alone, no such pattern is discernible.

We finally apply Algorithm 7 to the data. We choose the thresholds and the truncation level as discussed at the end of Section 5.4, and set the bin width to $q = 1$. The estimated graph is shown in Figure 5.5, and seems to agree with our previous intuition about potential causal pathways.

Figure 5.2: Graphs representing the number of times a given edge was present in an estimate of the local independence graph for data from model (5.3) simulated as described in Section 5.5.1, over 100 replications, for sample sizes $n \in \{500, 1000, 2000\}$ and signal strengths $C_\Delta \in \{2, 5, 7\}$.



(a) $n = 500$, $C_\Delta = 2$

(b) $n = 500$, $C_\Delta = 5$

(c) $n = 500$, $C_\Delta = 7$

(d) $n = 1000$, $C_\Delta = 2$

(e) $n = 1000$, $C_\Delta = 5$

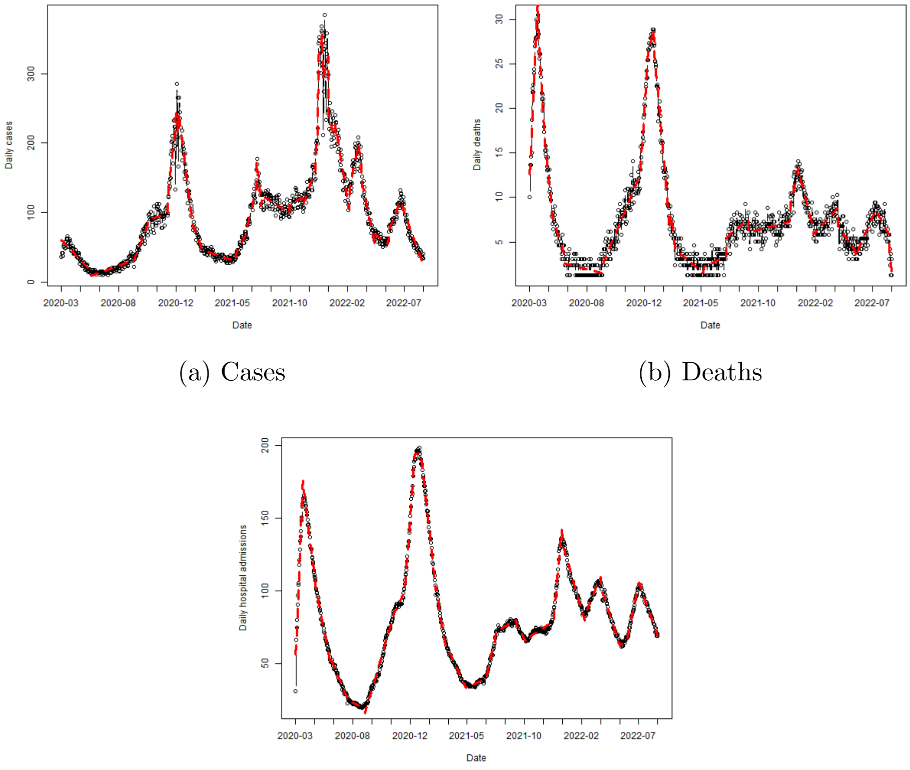
(f) $n = 1000$, $C_\Delta = 7$

(g) $n = 2000$, $C_\Delta = 2$

(h) $n = 2000$, $C_\Delta = 5$

(i) $n = 2000$, $C_\Delta = 7$

Figure 5.3: Anscombe transform of daily COVID-19 trajectories across all London boroughs along with piecewise linear trend (- - -) recovered using the Narrowest Over Threshold algorithm (Baranowski et al., 2019a).



(a) Cases

(b) Deaths

(c) Hospitalizations

Figure 5.4: Times of estimated change point locations recovered by the Narrowest Over Threshold algorithm from (Anscombe transformed) time series of daily deaths, daily hospitalizations, and daily cases of COVID-19 across all London boroughs.

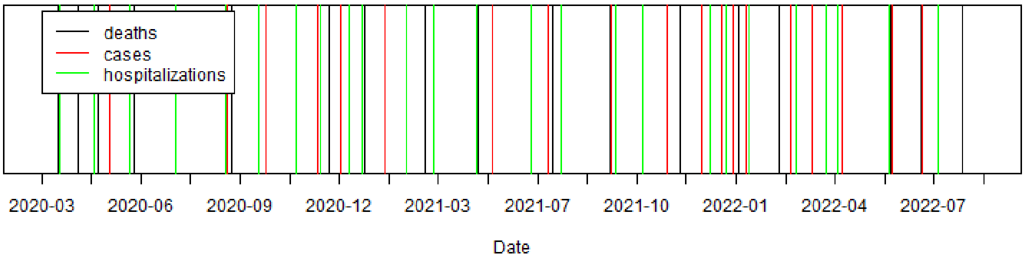
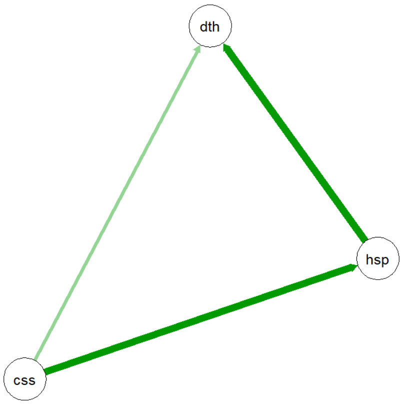


Figure 5.5: Local independence graph for change points in daily deaths, daily cases, and daily hospitalizations from COVID-19. In the plot 'css' stands for 'cases', 'hps' stands for 'hospitalizations', and 'dth' stands for 'deaths'. Where non-zero, edges are weighted by estimates of the associated branching ratio.



5.6 Proofs

5.6.1 Preparatory results

Definition 5.6.1. *Consider an age dependent pure birth process, originated by a single particle at time zero, with kernel $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that for an individual of age t_1 alive at time t_2 the probability of a birth in the interval $(t_2, t_2 + \mathrm{d}t_2)$ is $g(t_1)\,\mathrm{d}t_2$. Let $\int g = m \in (0,1)$ so that the process is almost surely finite. Let X_n be the number of offspring in the n-th generation of a standard Galton Watson process, where: $X_0 = 1$, $X_n = \sum_{j=1}^{X_{n-1}} \xi_j^{(n)}$, and $\xi_j^{(n)} \sim$ Poisson(m). It holds that:*

(i) The total number of offspring in the k-th generation of the pure birth process, say G_k, is distributed as the number of offspring in the k-th generation of the Galton Watson process. That is: $G_k \overset{d}{=} X_k$.

(ii) The total number of generations in the pure birth process, say τ, is distributed as the total number of generations in the Galton Watson process. That is: $\tau \overset{d}{=} \inf\{n \mid X_n = 0\}$.

Definition 5.6.2. *Let $\boldsymbol{N}(\cdot) = (N_1(\cdot), \ldots, N_d(\cdot))'$ be a simple, stationary, and orderly* *multi-type Hakes process with conditional intensities given by (5.2). One way of constructing this process is to take d independent homogeneous Poisson processes $N_1^c(\cdot), \ldots, N_d^c(\cdot)$ with rates $\nu_1, \ldots, \nu_d$ and let their arrival times stand for cluster centers. At each each cluster center we place a multi-type age-dependent branching process, where the probability of a particle of type i of age t_1 alive at time t_2 producing a particle of type j in the interval $(t_2, t_2 + \mathrm{d}t_2)$ is $g_{j,i}(t_1)\,\mathrm{d}t_2$. Then, the superposition of all such points constitutes the arrival times of the Hawkes process.*

Theorem 5.6.1. *Let $\boldsymbol{N}(\cdot)$ be a d-dimensional Hawkes process satisfying Assumption 5.4.4 with d fixed, and with constant base-line intensities. Let $\hat{\boldsymbol{G}}^q = \left(\hat{\boldsymbol{G}}_s^q \mid 1 \leq s \leq k\right)'$ be estimates of the Hawkes kernels obtained vai conditional least squares and let $\hat{m}_{j,i} = \sum_{s=1}^k \hat{\boldsymbol{G}}_{s,j,i}^q$, based on events observed in the interval $[0, \tilde{n}]$, using bin width $\tilde{q}$ and truncation parameter k satisfying: (i) $k\tilde{q} \to \infty$, (ii) $k\tilde{q}^2 \to 0$, and (iii) $k^2/\tilde{n} \to 0$, all as $\tilde{n} \to \infty$. It holds that $\hat{m}_{j,i} - m_{j,i} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{k\tilde{q}/\tilde{n}}\right)$ uniformly in $j, i = 1, \ldots, n$.*

Proof. This follows from the proof of Theorem 4.2 in Eichler et al. (2017), and can also be deduced from from Remark 1 in Kirchner (2017). □

5.6.2 Intermediate results

Lemma 5.6.1. *Let L_∞ be the time between the first birth and the last birth in the pure birth process described in Definition 5.6.1. Putting $h = g/m$ and $\bar{H}(x) = \int_x^\infty h(u)\,\mathrm{d}u$, for any $x > 0$ it holds that:*

$$\mathbb{P}\left(L_\infty > x^2\right) \leq \frac{m}{1-m}\bar{H}(x) + m^x.$$

Proof. Writing L_k for the maximum of 0 and the time elapsed between the last birth in the $(k-1)$-th generation and the last birth in the k-th generation we have that:

$$\mathbb{P}\left(L_\infty > x^2\right) = \mathbb{P}\left(\sum_{k=1}^{\infty} L_k > x^2\right) \leq \mathbb{P}(\tau > x) + \mathbb{P}\left(\sum_{k=1}^{x} L_k > x^2\right).$$

For the first term, by Markov's inequality, we have that:

$$\mathbb{P}(\tau > x) = \mathbb{P}(X_x > 1) \leq \mathbb{E}(X_x) \leq m^x.$$

Put $L_{j,k}$ for the time elapsed between the birth of the j-th offspring in the k-th generation and the time of birth of the particle which generated it from the $(k-1)$-th. For the second term we have that:

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^{x} L_k > x^2\right) &\leq \sum_{k=1}^{x} \mathbb{P}\left(\cup_{j=1}^{G_k} L_{j,k} > x\right) \\
&= \sum_{k=1}^{x} \sum_{g_k=1}^{\infty} \sum_{j=1}^{g_k} \mathbb{P}(L_{j,k} > x)\, \mathbb{P}(G_k = g_k) \\
&\leq \bar{H}(x) \sum_{k=1}^{x} \sum_{g_k=1}^{\infty} g \times \mathbb{P}(G_k = g_k) \\
&\leq \bar{H}(x) \sum_{k=1}^{\infty} \mathbb{E}(G_k) \\
&= \frac{m}{1-m} \bar{H}(x).
\end{aligned}$$

This completes the proof. □

Lemma 5.6.2. *Let T_∞ be the shortest time between any two births, including the birth of the initial particle, in the pure birth process described in Definition 5.6.1. Putting $\check{h}(x) = \sup_{y \geq 0} \int_{0 \vee y-x}^{y+x} h(u)\,\mathrm{d}u$, for any $x > 0$ it holds that:*

$$\mathbb{P}(T_\infty \leq x) \leq \check{h}(x) \left[\frac{m}{1-m} + \frac{2}{(1-\sqrt{m})^2 \log(1/m)}\right].$$

Proof. Put t_0 for the time of bith of the first particle and $t_{j,k}$ for the time of birth of the

k-th offspring in the j-th generation. Introduce the following events:

$$\begin{aligned}
\mathcal{A}_x &= \cup_{i=1}^{\infty} \cup_{j=1}^{\infty} \cup_{k=1}^{G_i} \cup_{l=1}^{G_j} \left\{ |t_{i,k} - t_{j,l}| \leq x, (i,k) \neq (j,l) \right\}, \\
\mathcal{B}_x &= \cup_{i=1}^{\infty} \cup_{j=1}^{G_i} \left\{ t_{i,j} \leq x \right\}.
\end{aligned}$$

We therefore have that $\mathbb{P}\left(T_\infty \leq x\right) \leq \mathbb{P}\left(\mathcal{A}_x\right) + \mathbb{P}\left(\mathcal{B}_x\right)$ and we bound each of the probabilities in turn. For the first we have the following:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{A}_x\right) &\leq \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{g_i=1}^{\infty} \sum_{g_j=1}^{\infty} \mathbb{P}\left(\cup_{k=1}^{g_i} \cup_{l=1}^{g_j} \left\{ |t_{i,k} - t_{j,l}| \leq x, (i,k) \neq (j,l) \right\} \right) \mathbb{P}\left(G_i = g_i, G_j = g_j\right) \\
&\leq \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{g_i=1}^{\infty} \sum_{g_j=1}^{\infty} \sum_{k=1}^{g_i} \sum_{l=1}^{g_j} \mathbb{P}\left(\left\{ |t_{i,k} - t_{j,l}| \leq x, (i,k) \neq (j,l) \right\} \right) \mathbb{P}\left(G_i = g_i, G_j = g_j\right) \\
&\leq \breve{h}\left(x\right) \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{g_i=1}^{\infty} \sum_{g_j=1}^{\infty} g_i g_j \mathbb{P}\left(G_i = g_i, G_j = g_j\right) \\
&< \breve{h}\left(x\right) \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mathbb{E}\left(G_i G_j\right) \\
&\leq \breve{h}\left(x\right) \sum_{i=1}^{\infty} \sqrt{\mathbb{E}\left(G_i^2\right)} \sum_{j=1}^{\infty} \sqrt{\mathbb{E}\left(G_j^2\right)}. \qquad (5.8)
\end{aligned}$$

We now turn our attention to the two expectations. Using the observation from Defintion 5.6.1 that $G_n \overset{d}{=} X_n$ we have the following for any $n \geq 0$:

$$\begin{aligned}
\mathbb{E}\left(G_n^2\right) &= \mathbb{E}\left(X_n^2\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{X_{n-1}} {\xi_i^{(n)}}^2 + \sum_{j \neq k}^{X_{n-1}} \xi_j^{(n)} \xi_k^{(n)} \mid X_{n-1} \right) \right) \\
&= \mathbb{E}\left(X_{n-1}\left(m + m^2\right) + X_{n-1}\left(X_{n-1} - 1\right) m^2 \right) \\
&= m^n + m^2 \mathbb{E}\left(G_{n-1}^2\right).
\end{aligned}$$

Hence we have a first order reccurence recurence relation with varying coefficients: $\mathbb{E}\left(G_n^2\right) = m^n + m^2 \mathbb{E}\left(G_{n-1}^2\right)$. This can be solved with standard techniques (Gross, 2016), then bounded, as follows:

$$\mathbb{E}\left(G_n^2\right) = m^{2n}\left(1+\sum_{l=0}^{n-1} m^{-(l+1)}\right) \leq 2m^{2n-1}\int_0^n \left(\frac{1}{m}\right)^u \mathrm{d}u < \frac{2m^{n-1}}{\log\left(1/m\right)}. \tag{5.9}$$

Plugging (5.9) into (5.8) we finally have that

$$\mathbb{P}\left(\mathcal{A}_x\right) < \breve{h}\left(x\right)\sum_{i=1}^{\infty}\sqrt{\frac{2m^{i-1}}{\log(1/m)}}\sum_{j=1}^{\infty}\sqrt{\frac{2m^{j-1}}{\log(1/m)}} = \frac{2\breve{h}\left(x\right)}{\left(1-\sqrt{m}\right)^2\log\left(1/m\right)}.$$

Turning our attention to the second probability we have

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}_x\right) &\leq \sum_{i=1}^{\infty}\sum_{j=1}^{g_i}\mathbb{P}\left(t_{i,j}\leq x\right)\mathbb{P}\left(G_i = g_i\right)\\ &< \breve{h}\left(x\right)\sum_{i=1}^{\infty}\mathbb{E}\left(G_i\right)\\ &= \breve{h}\left(x\right)\frac{m}{1-m}.\end{aligned}$$

This completes the proof. □

5.6.3 Proof of Theorem 5.4.1

Proof. We consider the cluster process representation of the marked Hawkes process generating the change point locations, as given in Definition 5.6.2, and argue the proof in five steps.

STEP 1: *No two cluster centers are less than Ψ_1 apart.* By Assumption 5.4.2 and standard results on Poisson processes the superposition of the processes generating the cluster centers, say $N^{\overline{c}}(\cdot)$, is a Poisson process with rate $\nu^c = \left(\sum_{j=1}^{d} C_j\right) n^{-(1/2+\phi)}$. The inter-arrival times of this process are mutually independent exponential random variables, say Z_k^c, with rate $\nu^Z = \left(\sum_{j=1}^{d} C_j\right) n^{-(1/2+\phi)}$. Therefore, the probability of two cluster centers being less than Ψ_1 apart can be bounded as

$$\begin{aligned}\mathbb{P}\left(\cup_{k=1}^{N^{\bar{c}}(n)} Z_k^c \leq \Psi_1\right) &\leq \mathbb{P}\left(\cup_{k=1}^{2\nu^c} Z_k^c \leq \Psi_1\right) + \mathbb{P}\left(N^{\bar{c}}(n) > 2\nu^c\right)\\ &\leq 2\nu^c \int_0^{\Psi_1} \nu^Z e^{-\nu^Z u} \mathrm{d}u + o(1)\\ &\leq 2\nu^c \left[1 - e^{-\nu^Z \Psi_1}\right] + o(1)\\ &\leq 2\nu^c \nu^Z \Psi_1 + o(1)\\ &= o(1).\end{aligned}$$

SETP 2: *The time between the first birth and the last birth in any cluster and is at most* Ψ_2^2. Arguing as in Step 1 on a set with probability $1 - o(1)$ there are fewer than $2\nu^c$ cluster centres. By construction the clusters are independent. The time between the first birth and the last birth in a multi-type branching process with kernels $\{g_{j,i}(\cdot)\}_{j,i=1}^d$ will be shorter than the time between the first birth and the last birth in regular age dependent branching using the kernel with the heaviest tail out out each of the $g_{j,i}(\cdot)$'s. Therefore, Lemma 5.6.1 and Assumption 5.4.3 give that the probability that the first birth in any cluster and is greater than Ψ_2^2 is of the order $o(1)$.

STEP 3: *No two events in any cluster are less that* Ψ_3 *distance apart.* Using Lemma 5.6.2 and Assumption 5.4.3 the probability of any two events in a cluster being less than Ψ_3 apart has probability of the order $o(n^{\phi-1/2})$, but again with high probability there are of the order $\mathcal{O}(n^{1/2-\phi})$ clusters, so the statement is proved.

STEP 4: *All change points are detected, and the branching ratios are consistently estimated.* By Steps 1-3 the distance between any two change points will not be smaller than n^{ϕ}, and by Assumption 5.4.6 all of the change points will be detected and each change point will be localized at least at the rate $o(n^{\phi})$. By Assumption 5.4.5, when we bin the interval $[0, n]$ every bin contains either zero or one change points. Therefore, the $\mathbb{X}$'s used to obtain the branching coefficients are, on a high probability set, the same as those we would have if we had observed a Hawkes process with constant base-line intensities $C_1, \ldots, C_d$ on the interval $[0, n^{1/2-\phi}]$, used bin widths $\tilde{q} = q/n^{1/2+\phi}$ where q is as specified in Assumption

5.4.5, and used the same truncation level k. Plugging in $\tilde{n} = n^{1/2-\phi}$ and $\tilde{q} = q/n^{1/2+\phi}$ Theorem 5.6.1 gives that the branching rations can be estimated at the rate $\mathcal{O}\left(\sqrt{\frac{kq}{n}}\right)$, from which the result in Theorem 5.4.1 follows. □

6 Concluding Remarks and Future Work

In Chapters 3, 4, and 5 we introduced procedures for solving the problems of statistical inference and causal structure recovery which often arise in change point analysis. In this chapter we review our main methodological contributions, and highlight some avenues for future work.

6.1 Remarks on Chapter 3

In Chapter 3 we studied the problem of statistical inference on the unobserved change point locations in the piecewise polynomial change point model. Our goal was to recover disjoint intervals which each contain a change point location, uniformly with some probability chosen by the user. We introduced two procedures, DIF1 and DIF2, delaing respectively with the problems of inference in the presence of independently distributed and weakly dependent light tailed noise. Our procedures followed the principle of inference without selection (2.2.3), and recovered the aforementioned intervals by performing local tests for the presence of a change over an exponentially decaying grid. We used tests based on local averages of the data, which we argued provide an attractive alternative to the more commonly used likelihood ratio and Wald tests, both from the perspective of computational efficiency and simpler theoretical analysis. Moreover, the tests themselves were adaptive to the density of the gid used, thereby providing a solution to the open question (Pilliat et al., 2023) of the the statistical benefits of using a sparser or denser grid. Finally, our theoretical results show the lengths of of the intervals returned match the minimax localization rates (Yu et al., 2022) for change points in the associated change point problem, meaning any integer within an interval returned by the procedure can be used as an optimal estimator for the associated change point location.

A natural avenue for future research would involve extending the procedure to the multivariate or high dimensional setting, where each observation constitutes a vector with each entry having piecewise polynomial mean. The minimax detection and localization bounds for high dimensional mean shift problems have been studied in the literature (Pilliat et al., 2023; Liu et al., 2021) however these quantities for the high dimensional piecewise polynomial problem have not been studied. Certainly there are several regimes, depending not only on the sparsity of the change but also on the smoothness of the regression functions at each change point location. Recently Li et al. (2023a) proposed a generic change point detection method based on training neural network classifiers. The authors argued that likelihood ratio test for the presence of a change can be represented as simple neural networks. However, as we showed in Gavioli-Akilagun (2023), the likelihood ratio test for a change point in the piecewise polynomial model cannot be represented in this way. Nevertheless, the local tests proposed in Chapter 3 can be represented as neural networks, and using the techniques from the proof of Theorem 3.3.1 it can be shown that their procedure is near-optimal for localizing change points in this setting. It may therefore be of interest to further investigate the connection between neural network classifiers and difference based tests.

6.2 Remarks on Chapter 4

In Chapter 4 we revisited the problem of change point inference in the piecewise polynomial model, but relaxed the assumption of light tailed contaminating noise. Formally, we looked for change points in the piecewise polynomial parametrization of the data's median. We introduced two procedures, SET and SET-DEP, respectively for the settings of (sign) independent and weakly dependent sign symmetric noise. The local tests employed were based on non-parametric confidence sets for the underlying regression function (Dümbgen, 1998; Dümbgen and Johns, 2004) obtained by inverting certain multi-scale tests (4.3) acting on signs of empirical residuals. By working implicitly with signs of the data, which are automatically bounded regardless of the data's distribution, our procedures provide correct inferential statement assuming only sign symmetric noise. Nevertheless, the theoretical results show the lengths of of the intervals returned match the minimax localization rates for change points in the piecewise polynomial model under light tailed independently distributed noise. This result, though appealing, is non necessarily surprising in light of the fact that robust non-parametric change point estimators (Bhattacharyya and Johnson, 1968; Dumbgen, 1991) can achieve essentially the same localization rates as their non-robust counterparts.

There has been some recent interest in detecting change points in data with quntiles other than the median parametrized as piecewise polynomial (Jiang et al., 2021; Jula Vanegas et al., 2021; Brantley et al., 2020). A natural question is whether the procedures from Chapter 4 can likewise be extended to quantiles other than the median. This indeed seems possible. Consider the data generating process (4.1) where this time the ζ's are continuously distributed with γ-th quantile equal to zero. Introduce the following local tests:

$$T_{s:e}^{\lambda}\left(\boldsymbol{Y}\right) = \mathbf{1}\left\{\min_{\hat{f}}\max_{s\leq i\leq j\leq e}\frac{1}{\sqrt{j-i+1}}\left|\sum_{t=i}^{j}\frac{1}{\sqrt{\gamma(1-\gamma)}}\left(\mathbf{1}_{\left\{Y_t-\hat{f}(t/n)>0\right\}}-1+\gamma\right)\right|>\lambda\right\}. \tag{6.1}$$

Notice that by taking $\gamma=\frac{1}{2}$ the test (4.3) is recovered. The following inequality again holds uniformly over all sub-intervals free from change points:

$$\begin{aligned}\min_{\hat{f}}\max_{s\leq i\leq j\leq e}&\frac{1}{\sqrt{j-i+1}}\left|\sum_{t=i}^{j}\frac{1}{\sqrt{\gamma(1-\gamma)}}\left(\mathbf{1}_{\left\{Y_t-\hat{f}(t/n)>0\right\}}-1+\gamma\right)\right|\\ &\leq\max_{1\leq i\leq j\leq n}\frac{1}{\sqrt{j-i+1}}\left|\sum_{t=i}^{j}\frac{1}{\sqrt{\gamma(1-\gamma)}}\left(\mathbf{1}_{\{\zeta_t>0\}}-1+\gamma\right)\right|.\end{aligned}$$

The random variables $\frac{1}{\sqrt{\gamma(1-\gamma)}}\left(\mathbf{1}_{\{\zeta_t>0\}}-1+\gamma\right)$ are independently distributed centered and scaled Bernoulli random variables, and Theorem 1.1 in Kabluchko and Wang (2014) can again be used to obtain the limiting distribution of the maximum of their scaled partial sums. Such a result can be used to obtain a λ which asymptotically controls the family-wise error of the local tests (6.1). It remain to invert the local tests, subject to shape constraints, as was done in Chapter 4. This can be done efficiently using, for example, the algorithm proposed in Duembgen and Luethi (2022).

6.3 Remarks on Chapter 5

In Chapter 5 we considered a multi-variate time series in which each of the component processes undergoes a random number of changes, and studied the problem of recovering a graph which encodes dependence among the changes in the sense that the presence of an edge in the graph signifies that changes in one series cause changes in another. Formally, we modeled the change points in each series as arrival times of a marked Hawkes process (Hawkes, 1971) and sought to recover the local independence graph (Didelez, 2008) for the underlying unobserved process. We proposed an algorithm for recovering the graph, based on first estimating the change points via some base method then applying existing estimation procedures (Kirchner, 2017; Eichler et al., 2017) treating the estimated change point locations as the true arrival times of the process, and proved that as the sample size n tends to infinity the graph can be consistently recovered.

There are a number of directions in which the proposed methodology can be extended. To begin with, we modeled the change point locations as as arrival times of a marked point process with mark space $\{1, \ldots, d\}$. However, associated with each change point location we also have the magnitude and sign of the change, and these quantities can be consistently estimated. Therefore, in reality we have a marked process with mark space $\{1, \ldots, d\} \times \mathbb{R}_+ \times \{-1, +1\}$. Including this information in the estimation procedure may be beneficial. For example, in the real data example presented in Section 5.5.2 positive changes in the infection rate are more likely to lead to positive changes in the hospitalization rate. Second, our proof for the consistency of the estimated graph relied on the high probability event that all changes are detected. However, in practice this may not happen. Therefore, there may be some advantage in applying a robust estimation procedure which is not sensitive to some events being unobserved. In many change point problems the number of change points can be quite small relative to the sample size. There may therefore be some advantage in using estimation procedures, such as the one proposed by Salehi et al. (2019), which are designed with small sample sizes in mind. Finally, would be worthwhile to extend the theoretical results to the case of random change points in the generic piecewise polynomial model. This indeed seems possible, although the rates governing the time scales presented in Section 5.4 will likely change.

Bibliography

Abramovich, F., Antoniadis, A., and Pensky, M. (2007). Estimation of piecewise-smooth functions by amalgamated bridge regression splines. *Sankhyā: The Indian Journal of Statistics*, pages 1–27. 31

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I., and Muzy, J.-F. (2018). Uncovering causality from multivariate hawkes integrated cumulants. *Journal of Machine Learning Research*, 18(192):1–28. 63

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (inar (1)) process. *Journal of Time Series Analysis*, 8(3):261–275. 157

Al-Osh, M. A. and Alzaid, A. A. (1988). First-order integer-valued autoregressive (inar (1)) process: distributional and regression properties. *Statistica Neerlandica*, 42(1):53–61. 157

Anastasiou, A. and Fryzlewicz, P. (2022). Detecting multiple generalized change-points by isolating single ones. *Metrika*, 85(2):141–174. 46, 55, 70, 100, 134, 167

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856. 38

Antoch, J. (1999). Estimators of changes. *Asymptotics, nonparametrics, and time series.* 51

Antoch, J., Hušková, M., and Veraverbeke, N. (1995). Change-point problem and bootstrap. *Journaltitle of Nonparametric Statistics*, 5(2):123–144. 51

Aue, A., Horváth, L., and Husková, M. (2009). Extreme value theory for stochastic integrals of legendre polynomials. *Journal of Multivariate Analysis*, 100(5):1029–1043. 31, 39

Aue, A., Horvath, L., Husková, M., and Kokoszka, P. (2008). Testing for changes in polynomial regression. *Bernoulli*, 14(3):637–660. 39

Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. *Electron. J. Statist*, 12(2). 29

Bachrach, L. K., Hastie, T., Wang, M.-C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy asian, hispanic, black, and caucasian youth: a longitudinal study. *The journal of clinical endocrinology & metabolism*, 84(12):4702–4712. 100

Bai, J. (1995). Least absolute deviation estimation of a shift. *Econometric Theory*, 11(3):403–436. 51

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78. 29, 51, 90, 100

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22. 29, 134, 167

Banesh, D., Petersen, M., Wendelberger, J., Ahrens, J., and Hamann, B. (2019). Comparison of piecewise linear change point detection with traditional analytical methods for ocean and climate data. *Environmental Earth Sciences*, 78(21):1–16. 166

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019a). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):649–672. 20, 45, 46, 70, 94, 100, 134, 149, 166, 193, 195

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019b). Online supplementary materials for "narrowest-over-threshold detection of multiple change-points and change-point-like features". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):649–672. 45

Barba, L. and Langerman, S. (2014). Optimal detection of intersections between convex polyhedra. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1641–1654. SIAM. 143

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1):187–225. 29

Barigozzi, M. and Trapani, L. (2020). Sequential testing for structural stability in approximate factor models. *Stochastic Processes and their Applications*, 130(8):5149–5187. 29

Bauer, P. and Hack1, P. (1980). An extension of the mosum technique for quality control. *Technometrics*, 22(1):1–7. 47

Berkes, I., Liu, W., and Wu, W. B. (2014). Komlós–major–tusnády approximation under dependence. *The Annals of Probability*, 42(2):794–817. 75

Bhattacharyya, G. K. and Johnson, R. A. (1968). Nonparametric tests for shift at an unknown time point. *The Annals of Mathematical Statistics*, pages 1731–1743. 207

Brantley, H. L., Guinness, J., and Chi, E. C. (2020). Baseline drift estimation for air quality data using quantile trend filtering. 14(2):585–604. 207

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598. 32, 134

Brodsky, E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media. 43

Cappello, L., Madrid Padilla, O. H., and Palacios, J. A. (2023). Bayesian change point detection with spike-and-slab priors. *Journal of Computational and Graphical Statistics*, pages 1–13. 56

Cappello, L. and Padilla, O. H. M. (2022). Variance change point detection with credible sets. *arXiv preprint arXiv:2211.14097*. 56

Carrington, R. and Fearnhead, P. (2023). Improving power by conditioning on less in post-selection inference for changepoints. *arXiv preprint arXiv:2301.05636*. 50

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2017). A robust approach for estimating change-points in the mean of an ar(1) process. *Bernoulli*, 23(2):1408–1447. 103

Chan, H. P. and Lai, T. L. (2006). Maxima of asymptotically gaussian random fields and moderate deviation approximations to boundary crossing probabilities of sums of random variables with multidimensional indices. *Ann. Probab*, 34(1):80–121. 176

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 23(1):409–428. 34, 71

Chan, K. H., Hayya, J. C., and Ord, J. K. (1977). A note on trend removal methods: the case of polynomial regression versus variate differencing. *Econometrica: Journal of the Econometric Society*, pages 737–744. 67

Chen, S., Witten, D., and Shojaie, A. (2017). Nearly assumptionless screening for the mutually-exciting multivariate hawkes process. *Electronic journal of statistics*, 11(1):1207. 63

Chen, Y., Shah, R. D., and Samworth, R. J. (2014). Discussion of multiscale change point inference by frick, munk and sieling. *Journal of Royal Statistical Society: Series B*, pages 544–546. 53

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of econometrics*, 86(2):221–241. 56

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electron. J. Statist.*, 10(2):2000–2038. 29

Cho, H. and Fryzlewicz, P. (2011). Multiscale interpretation of taut string estimation and its connection to unbalanced haar wavelets. *Statistics and computing*, 21:671–681. 43

Cho, H. and Fryzlewicz, P. (2022). wcm.gsa. `https://github.com/haeran-cho/wcm.gsa`. 101

Cho, H. and Fryzlewicz, P. (2023). Multiple change point detection under serial dependence: Wild contrast maximisation and gappy schwarz algorithm. *Journal of Time Series Analysis*. 101, 103, 164

Cho, H. and Kirch, C. (2021). Data segmentation algorithms: Univariate mean change and beyond. *Econometrics and Statistics*. 36

Cho, H. and Kirch, C. (2022a). Bootstrap confidence intervals for multiple change points based on moving sum procedures. *Computational Statistics & Data Analysis*, 175:107552. 51, 90

Cho, H. and Kirch, C. (2022b). Two-stage data segmentation permitting multiscale change points, heavy tails and dependence. *Annals of the Institute of Statistical Mathematics*, pages 1–32. 34, 48

Chu, C.-S. J., Hornik, K., and Kaun, C.-M. (1995). Mosum tests for parameter constancy. *Biometrika*, 82(3):603–617. 47

Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *Ann. Statis*, 47(1):382–414. 29

Csörgö, M., Csörgö, M., Horváth, L., et al. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons. 38, 39

Csörgo, M. and Révész, P. (2014). *Strong approximations in probability and statistics*. Academic press. 75

Cunis, T., Burlion, L., and Condomines, J.-P. (2019). Piecewise polynomial modeling for control and analysis of aircraft dynamics beyond stall. *Journal of guidance, control, and dynamics*, 42(4):949–957. 31

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer. 60, 61, 185

Daley, D. J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer. 60

Danks, D. and Plis, S. (2013). Learning causal structure from undersampled time series, 2013. *URL http://repository. cmu. edu/cgi/viewcontent. cgi*. 60

Davies, P. L. (1995). Data features. *Statistica Neerlandica*, 49(2):185–245.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *The Annals of Statistics*, 29(1):1–65. 140

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239. 42

Dette, H., Eckle, T., and Vetter, M. (2020). Multiscale change point detection for dependent data. *Scandinavian Journal of Statistics*, 47(4):1243–1274. 53, 90

Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):751–764. 121

Dharamshi, A., Neufeld, A., Motwani, K., Gao, L. L., Witten, D., and Bien, J. (2023). Generalized data thinning using sufficient statistics. *arXiv preprint arXiv:2303.12931*. 50

Didelez, V. (2001). *Graphical models for event history analysis based on local independence*. Logos-Verlag. 63

Didelez, V. (2007). Graphical models for composable finite markov processes. *Scandinavian Journal of Statistics*, 34(1):169–185. 63

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264. 25, 61, 62, 186, 208

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455. 92, 158

Doukhan, P. (2012). *Mixing: properties and examples*, volume 85. Springer Science & Business Media. 75

Duembgen, L. and Luethi, L. (2022). Honest confidence bands for isotonic quantile curves. *arXiv preprint arXiv:2206.13069*. 208

Dumbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, pages 1471–1495. 207

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Annals of statistics*, pages 288–314. 170, 206

Dümbgen, L. and Johns, R. B. (2004). Confidence bands for isotonic median curves using sign tests. *Journal of Computational and graphical statistics*, 13(2):519–533. 140, 141, 142, 206

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, 29(1):124–152. 81

Duy, V. N. L., Toda, H., Sugiyama, R., and Takeuchi, I. (2020). Computing valid p-value for optimal changepoint by selective inference using dynamic programming. *Advances in Neural Information Processing Systems*, 33:11356–11367. 50

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist*, 32(2):407–499. 43

Eichinger, B. and Kirch, C. (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564. 18, 47, 48, 49, 69, 74, 183

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242. 187, 188, 189, 197, 208

Embrechts, P. and Kirchner, M. (2016). Hawkes graphs. *arXiv preprint arXiv:1601.01879*. 62

Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *Ann. Statist*, 47(4):2051–2079. 29, 38

EuroAir, x. (2022). Download of air quality data: download service for e1a and e2a data. `https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm`. Accessed: 2022-03-12. 164

Falk, M. and Reiss, R.-D. (1988). Independence of order statistics. *The Annals of Probability*, pages 854–862. 73

Fang, X., Li, J., and Siegmund, D. (2020). Segmentation and estimation of change-point models: false positive control and confidence regions. *Ann. Statist*, 48(3):1615–1647. 54, 70

Fang, X. and Siegmund, D. (2020). Detection and estimation of local signals. *arXiv preprint arXiv:2004.08159*. 46, 54

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16:203–213. 58

Fearnhead, P., Maidstone, R., and Letchford, A. (2019). Detecting changes in slope with an l 0 penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275. 42, 100

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*. 50

Fotoohinasab, A., Hocking, T., and Afghah, F. (2020). A graph-constrained changepoint detection approach for ecg segmentation. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 332–336. IEEE. 183

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 76(3):495–580. 51, 52, 90

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67. 134, 167

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist*, 42(6):2243–2281. 28, 34, 44, 46, 87

Fryzlewicz, P. (2021). Robust narrowest significance pursuit: inference for multiple change-points in the median. *arXiv preprint arXiv:2109.02487*. 139, 144, 155, 160

Fryzlewicz, P. (2023). Narrowest significance pursuit: inference for multiple change-points in linear models. *Journal of the American Statistical Association*, pages 1–14. 34, 55, 56, 77, 90, 138, 139, 156

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625–633. 81

Gavioli-Akilagun, S. (2023). Invited discussion of “automatic change-point detection in time series via deep learning” by li, fearnhead, fryzlewicz, and wang. *Journal of the Royal Statistical Society Series B (to appear)*. 206

Grange, S. K., Lee, J. D., Drysdale, W. S., Lewis, A. C., Hueglin, C., Emmenegger, L., and Carslaw, D. C. (2021). Covid-19 lockdowns highlight a risk of increasing ozone pollution in european urban areas. *Atmospheric Chemistry and Physics*, 21(5):4169–4185. 164

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438. 58, 187

Granger, C. W. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352. 58

Gross, J. L. (2016). *Combinatorial methods with computer applications.* CRC Press. 199

Hahn, G., Fearnhead, P., and Eckley, I. A. (2020). Bayesproject: Fast computation of a projection direction for multivariate changepoint detection. *Statistics and Computing*, 30:1691–1705. 56

Hallgren, K. L., Heard, N. A., and Turcotte, M. J. (2023). Changepoint detection on a graph of time series. *Bayesian Analysis*, 1(1):1–28. 183

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393. 80

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 12(1):83–143. 63

Harchaoui, Z. and Lévy-Leduc, C. (2007). Catching change-points with lasso. In *NIPS*, volume 617, page 624. 42

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493. 42

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90. 62, 184, 208

Henderson, J. and Michailidis, G. (2014). Network reconstruction using nonparametric additive ode models. *PloS one*, 9(4):e94003. 60

Hušková, M. and Slabỳ, A. (2001). Permutation tests for multiple changes. *Kybernetika*, 37(5):605–622. 47, 74

Hyun, S., Lin, K. Z., G'Sell, M., and Tibshirani, R. J. (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049. 50

Jacobsen, M. and Gani, J. (2006). Point process theory and applications: marked point and piecewise deterministic processes. 61

Jandhyala, V. and Minogue, C. (1993). Distributions of bayes-type change-point statistics under polynomial regression. *Journal of statistical planning and inference*, 37(3):271–290. 40

Jandhyala, V. K. and MacNeill, I. B. (1997). Iterated partial sum sequences of regression residuals and tests for changepoints with continuity constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):147–156. 40

Jarusková, D. (1999). Testing appearance of polynomial trend. *Extremes*, 2:25–37. 39

Jephcote, C., Hansell, A. L., Adams, K., and Gulliver, J. (2021). Changes in air quality during covid-19 'lockdown'in the united kingdom. *Environmental Pollution*, 272:116011. 34, 164

Jewell, S., Fearnhead, P., and Witten, D. (2022). Testing for a change in mean after changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1082–1104. 50

Jiang, F., Zhao, Z., and Shao, X. (2020). Time series analysis of covid-19 infection curve: A change-point perspective. *Journal of econometrics*. 31

Jiang, F., Zhao, Z., and Shao, X. (2021). Modelling the covid-19 infection trajectory: A piecewise linear quantile trend model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 31, 207

Jula Vanegas, L., Behr, M., and Munk, A. (2021). Multiscale quantile segmentation. *Journal of the American Statistical Association*, pages 1–14. 53, 156, 207

Kabluchko, Z. (2007). Extreme-value analysis of standardized gaussian increments. *arXiv preprint arXiv:0706.1849*. 74, 107, 169, 176

Kabluchko, Z. and Wang, Y. (2014). Limiting distribution for the maximal standardized increment of a random walk. *Stochastic Processes and their Applications*, 124(9):2824–2867. 55, 73, 74, 144, 207

Kartal, M. T., Depren, Ö., and Depren, S. K. (2021). The relationship between mobility and covid-19 pandemic: Daily evidence from an emerging country by causality analysis. *Transportation Research Interdisciplinary Perspectives*, 10:100366. 183

Kaul, A., Fotopoulos, S. B., Jandhyala, V. K., and Safikhani, A. (2021). Inference on the change point under a high dimensional sparse mean shift. *Electron. J. Statist.*, 15(1):71–134. 51

Kaul, A. and Michailidis, G. (2023). Inference for change points in high dimensional mean shift models. *Statistica Sinica (to appear)*. 51

Khinchin, A. Y., Andrews, D., and Quenouille, M. H. (1995). *Mathematical methods in the theory of queuing*. Courier Corporation. 60

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598. 41

Kim, J., Oh, H.-S., and Cho, H. (2022). Moving sum procedure for change point detection under piecewise linearity. *arXiv preprint arXiv:2208.04900*. 48, 70, 79

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). \ell_1 trend filtering. *SIAM review*, 51(2):339–360. 43, 167

Kirch, C. and Klein, P. (2023). Moving sum data segmentation for stochastics processes based on invariance. *Statistica Sinica*, 33:873–892. 74

Kirchner, M. (2017). An estimation procedure for the hawkes process. *Quantitative Finance*, 17(4):571–595. 187, 188, 192, 197, 208

Komlós, J., Major, P., and 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. 75, 85

Kovács, S., Li, H., Bühlmann, P., and Munk, A. (2023). Seeded binary segmentation: A general methodology for fast and optimal change point detection. *Biometrika*, 110(1):249–256. 46, 71, 90

Kröger, H., Kotaniemi, A., Kröger, L., and Alhava, E. (1993). Development of bone mass and bone density of the spine and femoral neck—a prospective study of 65 children and adolescents. *Bone and mineral*, 23(3):171–182. 31, 100

Kuan, C.-M. (1998). Tests for changes in models with a polynomial trend. *Journal of Econometrics*, 84(1):75–91. 40

Kuan, C.-M. and Hornik, K. (1995). The generalized fluctuation test: A unifying view. *Econometric Reviews*, 14(2):135–161. 40

Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *The Annals of Probability*, 8(6):1003–1036. 75

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (2012). *Extremes and related properties of random sequences and processes.* Springer Science & Business Media. 104, 106, 107, 119

Leisch, F. and Dimitriadou, E. (2021). *mlbench: Machine Learning Benchmark Problems.* R package version 2.1. 134

Levajković, T. and Messer, M. (2023). Multiscale change point detection via gradual bandwidth adjustment in moving sum processes. *Electronic Journal of Statistics*, 17(1):70–101. 49

Li, H., Guo, Q., and Munk, A. (2019). Multiscale change-point segmentation: Beyond step functions. *Electron. J. Statist.*, 13(2):3254–3296. 146

Li, J., Fearnhead, P., Fryzlewicz, P., and Wang, T. (2023a). Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology (to appear).* 206

Li, Y.-N., Li, D., and Fryzlewicz, P. (2023b). Detection of multiple structural breaks in large covariance matrices. *Journal of Business & Economic Statistics*, 41(3):846–861. 29

Liehrmann, A. and Rigaill, G. (2023). Ms. fpop: An exact and fast segmentation algorithm with a multiscale penalty. *arXiv preprint arXiv:2303.08723.* 79

Lin, K., Sharpnack, J. L., Rinaldo, A., and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. *Advances in neural information processing systems*, 30. 42

Liu, C., Martin, R., and Shen, W. (2017). Empirical priors and posterior concentration in a piecewise polynomial sequence model. *arXiv preprint arXiv:1712.03848.* 56

Liu, G.-X., Wang, M.-M., Du, X.-L., Lin, J.-G., and Gao, Q.-B. (2018). Jump-detection and curve estimation methods for discontinuous regression functions based on the piecewise b-spline function. *Communications in Statistics-Theory and Methods*, 47(23):5729–5749. 31

Liu, H., Gao, C., and Samworth, R. J. (2021). Minimax rates in sparse, high dimensional change point detection. *Ann. Statist*, 49(2):1081–1112. 38, 206

Lu, P., Cowell, C. T., LLoyd-Jones, S. A., Briody, J. N., and Howman-Giles, R. (1996). Volumetric bone mineral density in normal subjects, aged 5-27 years. *The Journal of Clinical Endocrinology & Metabolism*, 81(4):1586–1590. 31, 100

Ma, S. and Su, L. (2018). Estimation of large dimensional factor models with an unknown number of breaks. *Journal of econometrics*, 207(1):1–29. 29

MacNeill, I. B. (1978). Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. *The Annals of Statistics*, 6(2):422–433. 40

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and computing*, 27:519–533. 41, 167

Mastakouri, A. and Schölkopf, B. (2020). Causal analysis of covid-19 spread in germany. *Advances in Neural Information Processing Systems*, 33:3153–3163. 183

McGonigle, E. T. and Cho, H. (2023). Robust multiscale estimation of time-average variance for time series segmentation. *Computational Statistics & Data Analysis*, 179:107648. 151

McZgee, V. E. and Carleton, W. T. (1970). Piecewise regression. *Journal of the American Statistical Association*, 65(331):1109–1124. 31

Medvegyev, P. (2007). *Stochastic integration theory*, volume 14. OUP Oxford. 62

Mehrizi, R. V. and Chenouri, S. (2020). Detection of change points in piecewise polynomial signals using trend filtering. *arXiv preprint arXiv:2009.08573*. 43

Mehrizi, R. V. and Chenouri, S. (2021). Valid post-detection inference for change points identified using trend filtering. *arXiv preprint arXiv:2104.12022*. 43, 51

Meier, A., Kirch, C., and Cho, H. (2021). mosum: A package for moving sums in change-point analysis. *Journal of Statistical Software*, 97:1–42. 90

Milne-Thomson, L. M. (2000). *The calculus of finite differences*. American Mathematical Soc.

Monks, P. S. (2000). A review of the observations and origins of the spring ozone maximum. *Atmospheric environment*, 34(21):3545–3561. 32, 134

Nam, C. F., Aston, J. A., and Johansen, A. M. (2012). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823. 58

Nemirovskii, A. (1985). Nonparametric estimation of smooth regression functions. *Soviet Journal of Computer and Systems Sciences*, 23(6):1–11. 146

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2023). Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*. 50

Nodelman, U., Shelton, C. R., and Koller, D. (2012). Continuous time bayesian networks. *arXiv preprint arXiv:1301.0591*. 63

Padilla, O. H. M., Yu, Y., and Priebe, C. E. (2022). Change point localization in dependent dynamic nonparametric random dot product graphs. *The Journal of Machine Learning Research*, 23(1):10661–10719. 29

Pearl, J. (2009). *Causality*. Cambridge university press. 58, 62

Pein, F., Sieling, H., and Munk, A. (2017). Heterogeneous change point inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4):1207–1227. 53, 90, 156

Philipp, W., Stout, W. F., and Stout, W. (1975). *Almost sure invariance principles for partial sums of weakly dependent random variables*, volume 161. American Mathematical Soc. 75

Pilliat, E., Carpentier, A., and Verzelen, N. (2023). Optimal multiple change-point detection for high-dimensional data. *Electronic Journal of Statistics*, 17(1):1240–1315. 28, 54, 71, 205, 206

Piterbarg, V. (2015). Twenty lectures about gaussian processes. *Atlantic Financial, London*. 106

Raimondo, M. (1998). Minimax estimation of sharp change points. *Annals of statistics*, pages 1379–1397. 149

Rasines, D. G. and Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614. 50

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215–1230. 81

Rigaill, G., Lebarbier, E., and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing*, 22(4):917–929. 58

Rojas, C. R. and Wahlberg, B. (2014). On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*. 43

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2022). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *Journal of the American Statistical Association*, 117(540):2147–2162. 103

Salehi, F., Trouleau, W., Grossglauser, M., and Thiran, P. (2019). Learning hawkes processes from a handful of events. *Advances in neural information processing systems*, 32. 209

Schmittner, A., Fisher, D., Houston, L., Moore, K. D., and Capalbo, S. (2018). Introduction to climate science, an online textbook. In *AGU Fall Meeting Abstracts*, volume 2018, pages ED51I–0739. 183

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency var. *Journal of Business & Economic Statistics*, 33(3):366–380. 60

Schröder, A. L. and Fryzlewicz, P. (2013). Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and Its Interface*, 6:449–461. 31, 183

Schweder, T. (1970). Composable markov processes. *Journal of applied probability*, 7(2):400–410. 61

Shao, Q.-M. (1995). On a conjecture of révész. *Proceedings of the American Mathematical Society*, pages 575–582. 169

Shen, Y., Han, Q., and Han, F. (2022). On a phase transition in general order spline regression. *IEEE Transactions on Information Theory*, 68(6):4043–4069. 127

Slezakova, K. and Pereira, M. C. (2021). 2020 covid-19 lockdown and the impacts on air quality with emphasis on urban, suburban and rural zones. *Nature Scientific reports*, 11(1):1–11. 34, 164

Spiriti, S., Eubank, R., Smith, P. W., and Young, D. (2013). Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation*, 83(6):1020–1036. 134, 167

Tank, A., Li, X., Fox, E. B., and Shojaie, A. (2021). The convex mixture distribution: Granger causality for categorical time series. *SIAM Journal on Mathematics of Data Science*, 3(1):83–112. 60

Thams, N. and Hansen, N. R. (2023). Local independence testing for point processes. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12. 63

Theintz, G., Buchs, B., Rizzoli, R., Slosman, D., Clavien, H., Sizonenko, P., and Bonjour, J.-P. (1992). Longitudinal monitoring of bone mass accumulation in healthy adolescents: evidence for a marked reduction after 16 years of age at the levels of lumbar spine and femoral neck in female subjects. *The Journal of Clinical Endocrinology & Metabolism*, 75(4):1060–1065. 31, 100

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.*, 42(1):285–323. 43, 167

Tsybakov, A. B. (2004). Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10). 86

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems.* Stanford University. 28, 44

Verzelen, N., Fromont, M., Lerasle, M., and Reynaud-Bouret, P. (2023). Optimal change-point detection and localization. *Ann. Statist (to appear).* 28, 34, 38, 42, 79

Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. In *Doklady akademii nauk*, volume 259, pages 270–274. Russian Academy of Sciences. 44

Wang, D., Yu, Y., and Rinaldo, A. (2020). Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961. 28, 34, 147

Wang, D., Yu, Y., and Rinaldo, A. (2021). Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist*, 49(1):203–232. 29

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):57–83. 29

Wu, H., Lu, T., Xue, H., and Liang, H. (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716. 60

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154. 75

Wu, W. B. (2009). Recursive estimation of time-average variance constants. *The Annals of Applied Probability*, 19(4):1529–1552. 151

Wu, W. B. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):391–410. 81, 82

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189. 28, 41

Yu, Y., Chatterjee, S., and Xu, H. (2022). Localising change points in piecewise polynomials of general degrees. *Electronic Journal of Statistics*, 16(1):1855–1890. 37, 42, 46, 85, 86, 205

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32. 42