Electronic Journal of Statistics
Vol. 19 (2025) 593–655
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2345

Fast and optimal inference for change points in piecewise polynomials via differencing

Shakeel Gavioli-Akilagun[1] and **Piotr Fryzlewicz**[1]

[1] *London School of Economics and Political Science Department of Statistics, Columbia House, Houghton Street, London, WC2A 2AE, UK,*
e-mail: s.a.gavioli-akilagun@lse.ac.uk; p.fryzlewicz@lse.ac.uk

Abstract: 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. Our procedure is implemented in the R package `ChangePointInference` which is available via GitHub.

MSC2020 subject classifications: 62F25, 62F05.
Keywords and phrases: Confidence intervals, uniform coverage, unconditional coverage, structural breaks, piecewise polynomials, extreme value analysis.

Received March 2024.

Contents

arXiv: 2307.03639

1. Introduction

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:

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

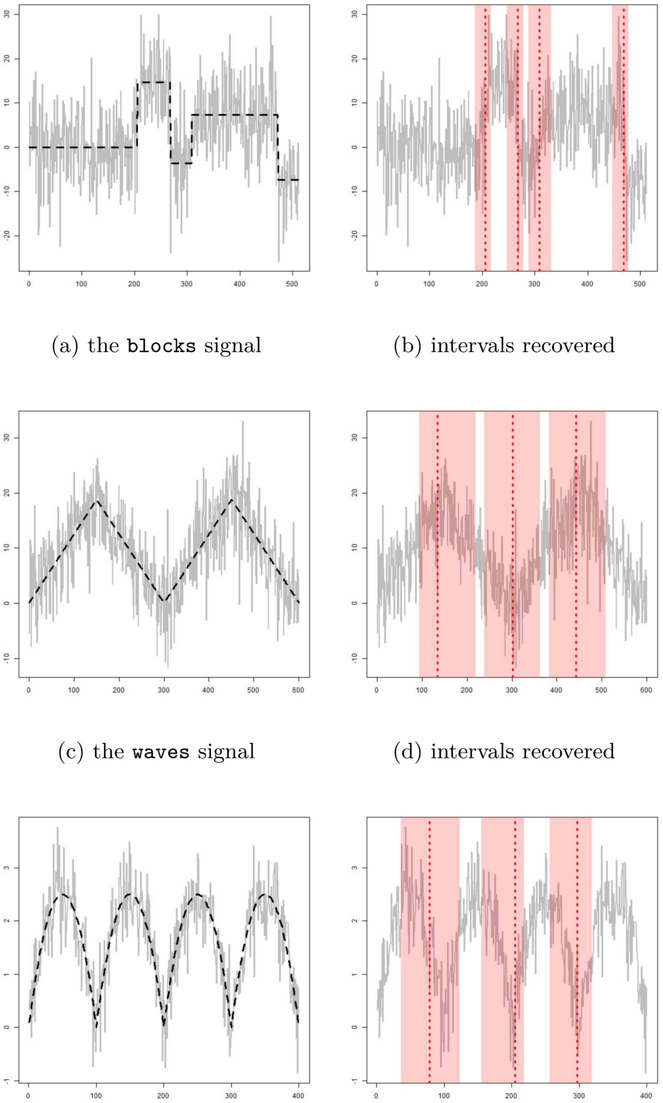
The signal component $f_\circ : [0,1] \mapsto \mathbb{R}$ is known to be a piecewise polynomial function of arbitrary but fixed degree p. That is, associated with $f_\circ(\cdot)$ are N integer valued change point locations $\Theta = \{\eta_1, \ldots, \eta_N\}$, whose number is possibly diverging with n, such that for each $k = 1, \ldots, N$ 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 $[(\eta_k - p)/n, (\eta_k + 1)/n]$. Examples of such signals are shown in the left column of Figure 1. Both N and Θ are unknown. Several algorithms exist for estimating N and Θ in specific instances of model (1), such as when $f_\circ(\cdot)$ is piecewise constant [51, 36, 28, 35] or when $f_\circ(\cdot)$ is piecewise linear [34, 9, 2, 64]. While the piecewise constant and piecewise linear change point regression are well studied, the generic piecewise polynomial model has attracted less attention. Nevertheless, the piecewise polynomial model has practical applications in areas as diverse as finance [76, 61, 66], aerospace engineering [22], protein folding [11], light transmittance [1], climatology [5], and data compression [39].

Our goal in this paper 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: the best rate at which a change point can be localised on the domain $\{1, \ldots, n\}$ is $\mathcal{O}_{\mathbb{P}}(1)$ [85, 86], however this can be as high as $\mathcal{O}_{\mathbb{P}}\left(n^{\varpi_k}\right)$ for each $k = 1, \ldots, N$ where $\varpi_k \in [0, 1)$ depends on the smoothness of $f_\circ(\cdot)$ at each change point location η_k [90, 91]. Moreover, since most algorithms do not quantify 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 shown in the right column of Figure 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 unconditional family-wise error control. An advantage of this approach is that once can post-search each unconditional interval for the best location(s) of the change-point(s) with appropriate statistics without worrying about significance testing. 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 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 [15], 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 2.4 when studying the behaviour of the 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 3.1 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.

We now review existing procedures in the literature for change point inference



(a) the `blocks` signal

(b) intervals recovered

(c) the `waves` signal

(d) intervals recovered

(e) the `hills` signal

(f) intervals recovered

Fig 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.

in specific instances of model (1). If one is able to localise all change points at a fast enough rate it is possible to construct asymptotically valid confidence intervals for the change point locations. This is done by [68, 19] for piecewise constant $f_\circ(\cdot)$ and by [7, 8] for regressions with piecewise constant coefficients. A crucial limitation of these approaches is that confidence intervals are only valid conditional on the number of change points being correctly estimated. Since there is no guarantee this will happen in a finite sample these intervals are problematic to interpret in practice. We further note that the piecewise constant regression model considered by [7, 8] does not actually cover generic the piecewise polynomial model (1), as it is necessary to assume the regressors are stationary or satisfy certain regularity conditions which are necessarily violated by polynomial functions of t. The SMUCE estimator and its many variants [35, 72, 23, 48] estimates a piecewise constant signal subject to the constraint that empirical residuals pass a multi-scale test, and produces a confidence set for the signal from which uniform confidence intervals for the change point locations can be extracted. However, the multi-scale tests have been observed to be poorly calibrated [37]. Moreover, letting the size of the test determine the estimated signal leads to larger nominal coverage actually reducing coverage [16]. In [32, 31] approximations of the tail probability of the supremum of local likelihood ratio tests for constant means and constant slopes calculated at all possible scales and locations are derived, and an algorithm is provided which returns uniform confidence intervals for the change point locations. However, the approach in these papers does not extend to the case of generic piecewise polynomials, and the algorithm propose has cubic time complexity in the worst case. The Narrowest Significance Pursuit algorithm [38, 37] tests for local deviations from a linear model using the multi-resolution norm [69], and bounds the supremum of local tests by the multi-resolution norm of the unobserved noise which in turn can be controlled using the results in [49, 50, 77]. However, computing each local test requires solving a linear program, which makes the procedure slow in practice. Moreover, other than for piecewise constant and continuous piecewise linear signals contaminated with Gaussian noise, it has not been shown that the procedure can detect change points optimally.

We finally review some approaches for problems closely related to the one studied in this paper. The problem of testing for the presence of a single change point in piecewise polynomials has previously been considered by [46, 5, 4] by studying the supremum of likelihood ratio tests, and by [45, 63] by studying partial sums of residuals from a least squares fit. These tests however do not extend to the case of multiple change points, which is the focus of this paper. Given estimates $\widehat{\Theta}$ and $\widehat{N}$ some authors focus on testing whether a change did in fact occur at each estimated change point location. This is a post selection inference problem as it requires conditioning on the estimation of $\widehat{\Theta}$, and has been studied by [44, 47, 12] for piecewise constant $f_\circ(\cdot)$ and by [67] for generic piecewise polynomial $f_\circ(\cdot)$. However, the goal of these methods is to quantify uncertainty about the size of each change, whereas our goal is to simultaneously quantify uncertainty about the existence and the location of the change point. The piecewise polynomial problem is closely related to the problem of detecting changes in the smoothness of the regression function in nonparametric regressions [78, 41]. The distribution of certain estimators for the location of a single change have been derived for instance by [70]; such results allow for inference on the location of the change. Our focus on the piecewise polynomial problem is motivated by practical considerations: a parametric model is often preferable to practitioners due to ease of interpretability. Finally, Bayesian approaches to change point detection provide an alternative approach to uncertainty quantification, via credible intervals derived from the posterior. However, choosing sensible priors and sampling from the posterior remain non-trial tasks. Methods for evaluating the posterior have been studied by [80, 33, 71].

The remainder of the paper is structured as follows. In Section 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 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 4 we compare the performance of our algorithm with that of existing procedures when applied to simulated data. Finally in Section 5 we show the practical use of our algorithm via two real data examples.

2. Difference based tests with family-wise error control

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 $\{l, \ldots, l+w-1\}$ we first compute the following non-overlapping local sums:

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

We then declare a change if the test statistic defined in (2) below, which corresponds to the the $(p+1)$-th differences of the sequence $\bar{Y}_{l,w}^{0}, \ldots, \bar{Y}_{l,w}^{p+1}$ 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.

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

The functional introduced in (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 $\eta\in\{l,\dots,l+w-1\}$.

The first two properties ensure (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 $\{l,\dots,l+w-1\}$ takes the following form:

$$T_{l,w}^{\lambda}\left(\mathbf{Y}\right)=\mathbf{1}\left\{|D_{l,w}^{p}\left(\boldsymbol{Y}\right)|>\lambda\right\}. \tag{3}$$

When $p=0$ the statistic (2) recovers the moving sum filter used for change point detection in the piecewise constant model [28]. 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 Wlad statistics for testing the null of a polynomial mean on the segment [9, 31, 2, 52], which however are hard to stochastically control. We show that simply extending the order of differencing leads to simple and powerful tests.

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) 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

$$\mathcal{T}_{\mathcal{G}}^{\lambda}\left(\mathbf{Y}\right)=\left\{T_{l,w}^{\lambda}\left(\mathbf{Y}\right)\mid(l,w)\in\mathcal{G}\right\} \tag{4}$$

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 (2), and a decay parameter $a>1$, controlling the density of the grid:

$$\begin{aligned} \mathcal{G}\left(W,a\right) &= \left\{(l,w)\in\mathbb{N}^2 \mid w\in\mathcal{W}(W,a), 1\le l\le n-w\right\} \\ \mathcal{W}\left(W,a\right) &= \left\{w=\left\lfloor a^k\right\rfloor \mid \lfloor\log_a(W)\rfloor\le k\le\lfloor\log_a(n/2)\rfloor\right\}. \end{aligned} \tag{5}$$

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

The grid defined by (5) is similar to several grids already proposed for different change point detection problems [56, 14, 74], in that the size of scales decays exponentially. Two key difference are first that for any scale w all possible locations l are considered, and second 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.

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) over

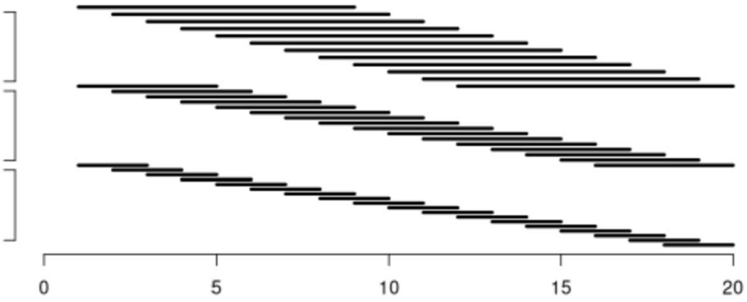


Fig 2: intervals associated with $\mathcal{G}\left(W,a\right)$ when $n=20$, $W=2$, and $a=2$.

the grid (5) when the noise terms are independently distributed and Gaussian. The approach is to tightly bound the maximum of the local test statistics (2) under the null of no change points, and use this bound to select an appropriate threshold λ for (4). We impose the following assumptions on the minimum grid scale and on the noise components.

Assumption 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 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 (2) under the null of no change points.

Theorem 2.1. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be from model (1) with signal component having no change points and grant Assumptions 2.1 - 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 [30, 50]. Therefore, in light of Theorem 2.1 it follows that under Assumptions 2.1 - 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

$$\lambda_\alpha = \sqrt{2\log(n)} + \frac{-\frac{1}{2}\log\log(n) - \log\left(2\sqrt{\pi}/H_{1,2}\right) + \log\left(-2\log^{-1}(1-\alpha)\right)}{\sqrt{2\log(n)}} \quad (6)$$

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 3.2 we given an example of an estimator

$\widehat{\sigma}$ which satisfies condition (ii) in Theorem 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 (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 (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 a change, 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.

$$H_{1,i} = \sum_{j=0}^{\infty} p_{\infty}^2 \left(\frac{2C_p}{a^j b_i d} \right) \qquad i = 1, 2 \tag{7}$$

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

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

The effect of the decay parameter a on $H_{1,1}$ and $H_{1,2}$ can now be understood via (7) using the additional fact that [49, 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 2.1, and why it is sufficient for strong family-wise error control. In Theorem 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 [49, 50], 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. Such fluctuations can arise even for maxima of sequences of i.i.d. random variables [3]; for instance the maximum of n i.i.d. Poisson random variables with fixed rate fluctuates between two integers [53].

2.4. Extension to dependent and non-Gaussian noise

We now extend the result of Theorem 2.1 to dependent and non-Gaussian noise. This is done through the standard approach [43, 54, 28] 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 2.3. *The noise terms are mean zero and weakly stationary, with auto-covariance 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 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 2.5. *With the same ν as in Assumption 2.4, the minimum grid scale W satisfies $n/W \to \infty$ and $n^{\frac{2}{2+\nu}}\log(n)/W \to 0$.*

Assumption 2.4 holds under a wide range of common dependence conditions such as β-mixing, functional dependence, and auto-covaraince decay [10, 73, 58]; these dependence conditions in turn hold for a range of popular time series models such as ARMA, GARCH, and bilinear models [26, 87]. If the noise terms are independently distributed Assumption 2.4 holds as long as their $(2+\nu)$-th moment is bounded [55, 21]. With these assumption in place we have the following result on the behaviour of the maximum of local test statistics (2) under the null of no change points.

Theorem 2.2. *Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ be from model (1) with signal component having no change points and grant Assumptions 2.3 - 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

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

(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 2.3 under assumptions 2.3 - 2.5 Theorem 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

$$\lambda_\alpha = \sqrt{2\log(n/W)} + \frac{\frac{1}{2}\log\log(n/W) - \log\left(\sqrt{\pi}/H_{2,2}\right) + \log\left(-2\log^{-1}(1-\alpha)\right)}{\sqrt{2\log(n/W)}} \tag{8}$$

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 3.2 we give examples of variance and long run variance estimators which satisfy condition (ii) in Theorem 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 2.1 the threshold (8) is adaptive to the chosen grid. The constants $H_{2,1}$ and $H_{2,2}$ in Theorem 2.2 are as shown below, where C_p and b_i are as in Section 2.3.

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

The proofs of Theorems 2.1 and 2.2 reveal that maxima achieved over different scales in the grid (5) will be asymptotically independent. This combined with the tightness of the normalised maximum shows that the thresholds (6) and (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 $\mathcal{O}(W)$ 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. A fast algorithm for change point inference

3.1. The algorithm

We now present an algorithm, based on the tests introduced in Section 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 [38], 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 [38], 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 [38], where each local test requires solving a linear program.
- We compute local tests over the sparse grid defined in (5), whereas [38] uses a two stage procedure where local tests are initially performed over

a coarse grid and intervals flagged in the first stage are exhaustively subsearched. 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 used in our local tests are designed to adapt to the chosen grid, which accounts for the statistical-computational trade off in large scale problems change point problems. However, the threshold used in [38] does not depend on the chosen grid.

Given a grid of (l, w) pairs $\mathcal{G}\left(W, a\right)$ constructed according to (5) our approach is to greedily search for a pair on which the associated local test (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. Pseudo code for the procedure is given below in Algorithm 1. In the pseudo code given integers s and e which satisfy $1 \leq s < e \leq n$ we write $\mathcal{G}_{s,e}\left(W, a\right)$ for the set of (l, w) pairs in $\mathcal{G}\left(W, a\right)$ 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 (6) or (8), depending on whether we are operating under Assumptions 2.1 - 2.2 or Assumptions 2.3 - 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 2.1 or in part (ii) of Theorem 2.2, depending on the set of assumptions we are operating under.

Algorithm 1: 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 $e - s < \min\left(W, p + 1\right)$ then
        STOP
    end
    detection ← False
    for $(l, w)$ in $\mathcal{G}_{s,e}\left(W, a\right)$ do
        if $\left|D^p_{l,w}\left(\boldsymbol{Y}\right)\right| > \widehat{\tau}\lambda_\alpha$ 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 (6) and (8) in Algorithm 1 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 3.1 below.

Corollary 3.1. *Let $\hat{I}_1, \ldots, \hat{I}_{\hat{N}}$ be intervals returned by Algorithm 1. 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 3.1 is asymptotic in nature, in practice we find that Algorithm 1 provides accurate coverage in finite samples, and in fact tends to deliver over coverage; see the simulation results in Section 4.2 and in Section 7. The thresholds proposed for Algorithm 1 rely on extreme value asymptotics. For such results the convergence rate is often slow, and indeed we conjecture that the convergence rate for our procedure is no better than $\mathcal{O}\left(\log^{-1}(n)\right)$. However, it is also known that the extreme value approximation for upper quantiles works well even for moderate sample sizes; confer for instance [20, Figure 1.3.1], [59, Section 2.4], and [4, 5].

The worst case run time of Algorithm 1 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 [85] by [60] for an example in in the piecewise constant setting, and the discussion on the practical selection of tuning parameters in [52] for an example in the piecewise linear setting. Meanwhile, the thresholds used in Algorithm 1 are the sharpest possible, and do not rely on any unknown constants other than the scale of the noise.

3.2. Variance and long run variance estimation

In general the scale of the noise will not be known, and to make Algorithm 1 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 2.1 or Theorem 2.2. In this section we give several strategies for consistently estimating the noise level in the presence of an unknown piecewise polynomial signal.

3.2.1. 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 [42]. 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 $X_{p+2}, \ldots, X_n$ be the $(p+1)$-th difference of the sequence $Y_1, \ldots, Y_n$ the estimator is defined as follows:

$$\widehat{\sigma}_{\mathrm{MAD}} = \frac{\mathrm{median}\left\{\left|X_{p+2}\right|, \ldots, \left|X_n\right|\right\}}{\Phi^{-1}\left(3/4\right)\sqrt{\sum_{j=0}^{p+1}\binom{p+1}{j}^2}}. \tag{9}$$

As shown by the following lemma, when the assumptions of Theorem 2.1 hold the modified MAD estimator satisfies the condition in part (ii) of the Theorem 2.1 as long as the number of change points grows more slowly than $n/\log(n)$.

Lemma 3.1. *If the noise terms are independently distributed and Gaussian with common variance σ^2 it holds that*

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

3.2.2. Variance estimation under non-Gaussian noise

For variance estimation under independently distributed light tailed homoskedastic noise, difference based estimators are often used [27, 79, 40]. 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:

$$\widehat{\sigma}^2_{\mathrm{DIF}} = \frac{1}{n-(p+1)}\sum_{t=p+2}^{n}\left\{\frac{X_t^2}{\sum_{j=0}^{p+1}\binom{p+1}{j}^2}\right\}. \tag{10}$$

As shown by the following lemma, under some mild conditions on signal component the difference based estimator satisfies condition (ii) in Theorem 2.2 as long as the number of change points again grows more slowly than $n/\log(n)$.

Lemma 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*

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

3.2.3. Long-run variance estimation

For estimating the long run variance we extend the estimator proposed in [89], 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 (5), and form the following local sums:

$$\bar{Y}_{t,W'} = Y_{(t-1)W'+1} + \cdots + Y_{tW'}, \qquad t = 1, \ldots, \lfloor n/W' \rfloor \tag{11}$$

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

$$\widehat{\tau}^2_{\mathrm{DIF}} = \frac{1}{\lfloor n/W' \rfloor - (p+1)} \sum_{t=p+2}^{\lfloor n/W' \rfloor} \left\{ \frac{\bar{X}^2_{t,W'}}{W' \sum_{i=0}^{p+1} \binom{p+1}{i}^2} \right\}. \tag{12}$$

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 3.1. *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 $\{1, \ldots, n-h\}$, 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. *If the function $f_\circ(\cdot)$ is bounded and the noise terms satisfy Assumption 2.3 and Assumption 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 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 (12) satisfies the condition in part (ii) of Theorem 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 [89] in setting $W' = n^{1/3}$.

3.3. Consistency of the algorithm

We now investigate the conditions under which algorithm Algorithm 1 is consistent, in the sense that with high probability it is able to detect all change points and returns no spurious intervals. It is useful to parameterise the signal in model (1) between change point locations as follows:

$$f_\circ(t/n) = \begin{cases} \sum_{j=0}^{p} \alpha_{j,k} \left(t/n - \eta_k/n\right)^j & \text{if } \eta_{k-1} < t \leq \eta_k \\ \sum_{j=0}^{p} \beta_{j,k} \left(t/n - \eta_k/n\right)^j & \text{if } \eta_k < t \leq \eta_{k+1} \end{cases} \qquad k = 1, \ldots, N. \tag{13}$$

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:

$$p_k^* \in \underset{0 \leq j \leq p}{\arg\max} \left\{ \Delta_{j,k} \left(\frac{\delta_k}{n} \right)^j \right\} \qquad k = 1, \dots, N. \tag{14}$$

In order to show the consistency of Algorithm 1 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 3.2. *There is a constant $C_\Delta > 0$ such that $|\Delta_{jk}| < C_\Delta$ for each j, k.*

Assumption 3.3. *For each $k = 1, \dots, N$ 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 $\left|\Delta_{j,k}\right| \rho_{k,n}^j \leq C_{p_k^*} \left|\Delta_{p_k^*,k}\right| \rho_{k,n}^{p_k^*}$ for all $j \neq p_k^*$, where $C_{p_k^*} = \frac{1}{2^{p_k^*+2}(p^*+1)p}$.*

For example, Assumption 3.3 would be violated by the piecewise linear signal shown in (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: $\Delta_0 = \Delta_1(\delta/n)$.

$$\mathbf{f} = (-7/8, -6/8, -5/8, -4/8, 3/8, 2/8, 1/8, 0)' \tag{15}$$

In practice, in situations when Assumption 3.3 is violated our procedure is still able to detect the corresponding change point. This is because although signal cancellation such as in (15) may occur on a particular interval considered by Algorithm 1, 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 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 3.1. *Let $\hat{I}_1, \dots, \hat{I}_{\hat{N}}$ be intervals returned by Algorithm 1 run on data $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ from model (1), with parameters $a > 1$, W, and $\alpha \in (0, 1)$. Grant Assumptions 3.2-3.3 and either of Assumptions 2.2-2.1 or 2.3-2.5 hold, and let the threshold λ_α chosen according to (6) or (8) accordingly. If the the* *effective sample size at each change point location satisfies*

$$\delta_k > C_1 \left(W \vee n^{\frac{2p_k^*}{2p_k^*+1}} \left(\frac{\tau^2 \log(n)}{\Delta^2_{p_k^*,k}} \right)^{\frac{1}{2p_k^*+1}} \right) \qquad k = 1, \dots, N \tag{16}$$

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

$$E_3^* = \left\{ \hat{N} = N \right\}$$

$$E_4^* = \left\{ \hat{I}_k \cap \Theta = \{\eta_k\} \mid k = 1, \dots, N \right\}$$

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

Here C_1 and C_2 depend only on α, a and p.

Theorem 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 1 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, which in turn implies a bound on the localisation rate of any change point estimator which lies within a given interval returned by the algorithm.

Theorem 3.1 leads to the following large sample consistency result.

Corollary 3.2. *Let $\hat{I}_1, \dots, \hat{I}_{\hat{N}}$ be intervals returned by Algorithm 1 under the same conditions as Theorem 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 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 3.1 and Corollary 3.2 is that any pointwise estimator $\hat{\eta}_k$ for the k-th change point location which lies in an interval $\hat{I}_k$ will inherit the localisation rate implied by event E_5^*. As explained in Section 3.4 this rate is unimprovable in a minimax sense. This result 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 each $\hat{I}_k$ (see for example Figure 1).

3.4. Optimality of the algorithm

In [91, 90] it was shown that, under independent sub-Gaussian noise with Orlicz-ψ_2 norm bounded from above by some ω^2, the mini-max localisation rate for each change point in the generic piecewsie polynomial model is of the order

$$\mathcal{O}\left(n^{\frac{2p_k^*}{2p_k^*+1}}\left(\frac{\omega^2}{\Delta^2_{p_k^*,k}}\right)^{\frac{1}{2p_k^*+1}}\right), \qquad k=1,\dots,N. \tag{17}$$

Examining the proof of Lemma 2 in [91] one can see that the same rate holds for weakly dependent noise by replacing the sub-Gaussian parameter ω^2 with the long run variance τ^2. Therefore, under Assumptions 2.1- 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 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 2.3-2.5, 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 (17) it is clear the bounds are no longer optimal.

The aforementioned lack of optimality is due to Assumption 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 [91, 90], then Theorem 1 in [55] 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 3.1 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.

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 minimax bound on the sup-norm risk for p^*-smooth Holder regression functions [84, Theorem 2.10]. The error probability α does not appear explicitly in Theorem 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 2.3 and 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 $-2\log^{-1}(1-\alpha) \sim 2/\alpha$.

Finally, the effect of the degree of serial dependence on the lengths of the intervals is explicit, as the long run variance of the noise appears in the upper bound on the interval lengths. The nature of this dependence is similar to that found by [29] in the simpler problem of detecting a bump in the mean function of a stationary Gaussian process.

3.5. On the polynomial order of the signal

We emphasise 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. If p is unknown, it should be considered as an input to our algorithm. However, provided this input is chosen large than or equal to the maximum polynomial order, it only affects the output in terms of constants and not rates. Of course, in a finite sample there is a price to pay: choosing p larger leads to longer intervals through inflating the constant C_2, and changes the change point detection condition through inflating the constant C_1.

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 for the setup in Sections 2.3 and 2.4 respectively.

3.5.1. Estimating p via the strengthened Schwarz Information Criterion

[36] 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 (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 1 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 [36] for some arbitrary but fixed $\alpha > 1$ the sCIC at p' is defined as

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

where in particular

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

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

$$\hat{p} = \underset{\underline{p} \leq p' \leq \overline{p}}{\arg\min}\, \mathrm{sSIC}\left(p'\right). \tag{18}$$

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

Lemma 3.4. *Let $\hat{p}$ be the estimator defined in (18). Grant Assumptions 2.2 and 2.1 as well as condition (16) hold, and moreover assume moreover that: (i) $\underline{p} \le p \le \overline{p}$ and $(\overline{p} - \underline{p}) = \mathcal{O}(1)$, (ii) $N = \mathcal{O}(1)$, and (iii) the coefficients in (13) are all of the order $\mathcal{O}(1)$. Then $\mathbb{P}\left(\hat{p} = p\right) \to 1$ as $n \to \infty$.*

3.5.2. 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

$$D_K^{p'}\left(\boldsymbol{Y}\right) = \left\{ \left\lfloor \frac{|K|}{p'+2} \right\rfloor \sum_{i=0}^{p'+1} \binom{p'+1}{i}^2 \right\}^{-1/2} \sum_{j=0}^{p'+1} (-1)^{p'+1-j} \binom{p'+1}{j} \bar{Y}_K^j \tag{19}$$

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

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

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 (19) will be large in (absolute) expectation, whereas if $p' \ge p$ then (19) will be small.

Using the above intuition, to estimate p we first run Algorithm 1 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 $p' = \overline{p}$. 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 $(p'+1)$. Note that since $\overline{p} \ge p$ by Corollary 3.2 with high probability each $\hat{K}$ corresponds to a stretch of data which contains no change points. Finally we test whether $|D_{\hat{K}}^{p'-1}\left(\boldsymbol{Y}\right)| > (1+\varepsilon)\mathfrak{a}_{W,n}$ for each $\hat{K}$. If any such test is not passed we conclude that $p = p'$. Else, we repeat the procedure with $p'-1$. The procedure automatically ends once $\underline{p}$ is reached, since we assume $p \ge \underline{p}$, and by this point we have concluded that $p < p'$ for all $p' > \underline{p}$. The procedure is sumarized in Algorithm 2. Regarding the large sample consistency of the output of Algorithm 2 we have the following result.

Lemma 3.5. *Let $\hat{p}$ be the output of Algorithm 2. Grant Assumptions 2.3, 2.4, and 2.5 as well as condition (16) hold, and moreover assume moreover that: (i) $\underline{p} \le p \le \overline{p}$ and $(\overline{p} - \underline{p}) = \mathcal{O}(1)$, (ii) $N = \mathcal{O}(1)$, and (iii) the coefficients in (13) are all of the order $\mathcal{O}(1)$. Then $\mathbb{P}\left(\hat{p} = p\right) \to 1$ as $n \to \infty$.*

Algorithm 2: 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 1 using
        threshold $\lambda = (1+\varepsilon)\mathfrak{a}_{W,n}$ and assuming maximum degree $p'$.
        for $\hat{K} \in \hat{\mathbb{K}}$ do
            if $|D^{p'-1}_{\hat{K}}(\boldsymbol{Y})| > (1+\varepsilon)\mathfrak{a}_{W,n}$ then
                Detection $\leftarrow$ True
            end
        end
        if Detection then
            BREAK
        end
        $p' \leftarrow (p'-1)$
    end
return
```

4. Simulation studies

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) algorithm of [38], its self-normalised variant (NSP-SN), and its extension to auto-regressive signals (NSP-AR); the bootstrap confidence intervals for moving sums (MOSUM) of [19] using a single bandwidth (uniscale) and multiple bandwidths (multiscale); the simultaneous multiscale change point estimator (SMUCE) of [35], as well as its extension to heterogeneous noise (H-SMUCE) developed by [72], and its extension to dependent noise (Dep-SMUCE) developed by [23]. We also consider the conditional confidence intervals of [7] (B&P) with significance level Bonferroni-corrected for the estimated number of change-points. For our own procedure we write DIF1 for Algorithm 1 run under the assumptions of Theorem 2.1 and DIF2 for the algorithm run under the assumption of Theorem 2.2. Additionally we write MAD if the scale of the noise is estimated using the median absolute deviation estimator (9), SD if the scale is estimated using the difference based estimator of the standard deviation (10), and LRV if the long run variance is estimated using (12). Each of the methods considered is designed for different noise types and different change point models, and we summarise this information in Table 1 below.

Table 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	✓	✓	✗

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 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 [68]. For our own procedure we set the decay parameter regulating the density of the grid to $a = \sqrt{2}$ as was done in [56] for the grid proposed therein.

4.2. Coverage on null signals

We first investigate empirically the coverage provided by our algorithm and the alternatives introduced in Section 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 100 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`): $\zeta_t \sim \sigma \times \text{Laplace}(0, 1/\sqrt{2})$ i.i.d.
- (`N4`): $\zeta_t = \phi\zeta_{t-t} + \varepsilon_t$ with $\phi = 0.8$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2/(1-\phi^2)$ i.i.d.
- (`N5`): $\zeta_t = \phi\zeta_{t-t} + \varepsilon_t$ with $\phi = 0.8$ and $\varepsilon_t \sim t_5 \times \sigma\sqrt{0.6/(1-\phi^2)}$ i.i.d.
- (`N6`): $\zeta_t = \phi_1\zeta_{t-1} + \phi_2\zeta_{t-2} + \sum_{j=1}^{6} \theta_j\varepsilon_{t-j} + \varepsilon_t$ with $\phi_1 = 0.75$, $\phi_2 = -0.5$, $\theta_j = 0.1 \times (9-j)$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Table 2

Proportion of times out of 100 *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	✓	0.93	0.92	0.95
DIF2-SD	✓	0.98	1.00	1.00
DIF2-LRV	✓	0.97	0.99	0.97
NSP	✓	0.96	0.99	0.99
NSP-SN	✓	1.00	1.00	1.00
NSP-AR	✓	1.00	1.00	0.99
B&P	**c**	0.99	-	-
MOSUM (uniscale)	**c**	0.98	-	-
MOSUM (multiscale)	**c**	0.94	-	-
SMUCE	✓	0.96	-	-
H-SMUCE	✓	0.95	-	-
Dep-SMUCE	✓	0.92	-	-

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

	guarantee	degree 0	degree 1	degree 2
DIF1-MAD	✗	0.46	0.45	0.38
DIF2-SD	✓	0.98	0.97	0.95
DIF2-LRV	✓	0.93	0.92	0.91
NSP	✗	0.05	0.04	0.04
NSP-SN	✓	1.00	1.00	1.00
NSP-AR	✗	0.14	0.10	0.19
B&P	**c**	0.97	-	-
MOSUM (uniscale)	**c**	0.99	-	-
MOSUM (multiscale)	**c**	0.98	-	-
SMUCE	✗	0.21	-	-
H-SMUCE	**e**	1.00	-	-
Dep-SMUCE	✓	0.95	-	-

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

	guarantee	degree 0	degree 1	degree 2
DIF1-MAD	✗	0.36	0.33	0.37
DIF2-SD	✓	0.97	0.99	0.99
DIF2-LRV	✓	0.98	0.98	0.94
NSP	✗	0.02	0.04	0.03
NSP-SN	✓	1.00	1.00	1.00
NSP-AR	✗	0.19	0.23	0.22
B&P	**c**	0.95	-	-
MOSUM (uniscale)	**c**	1.00	-	-
MOSUM (multiscale)	**c**	0.98	-	-
SMUCE	✗	0.14	-	-
H-SMUCE	**e**	1.00	-	-
Dep-SMUCE	✓	0.90	-	-

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

The results of the simulation study are reported in Tables 2 and 3. 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

Table 3

Proportion of times out of 100 *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	✗	0.00	0.00	0.00
DIF2-SD	✗	0.00	0.00	0.00
DIF2-LRV	✓	0.90	0.90	0.89
NSP	✗	0.00	0.00	0.00
NSP-SN	✗	0.00	0.00	0.01
NSP-AR	✓	1.00	0.99	0.98
B&P	✗	0.00	-	-
MOSUM (uniscale)	✗	0.00	-	-
MOSUM (multiscale)	✗	0.00	-	-
SMUCE	✗	0.00	-	-
H-SMUCE	✗	0.00	-	-
Dep-SMUCE	✓	0.41	-	-

(a) Coverage on noise type `N4` 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	✓	0.87	0.91	0.95
NSP	✗	0.00	0.00	0.00
NSP-SN	✗	0.00	0.01	0
NSP-AR	✗	0.17	0.12	0.07
B&P	✗	0.00	-	-
MOSUM (uniscale)	✗	0.00	-	-
MOSUM (multiscale)	✗	0.00	-	-
SMUCE	✗	0.00	-	-
H-SMUCE	✗	0.00	-	-
Dep-SMUCE	✓	0.32	-	-

(b) Coverage on noise type `N5` 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	✓	0.99	0.95	1.00
NSP	✗	0.00	0.00	0.00
NSP-SN	✗	0.03	0.10	0.12
NSP-AR	✗	0.83	0.87	0.93
B&P	✗	0.00	-	-
MOSUM (uniscale)	✗	0.00	-	-
MOSUM (multiscale)	✗	0.00	-	-
SMUCE	✗	0.00	-	-
H-SMUCE	✗	0.00	-	-
Dep-SMUCE	✓	0.94	-	-

(c) Coverage on noise type `N6` with $\sigma = 1$

estimated. The majority of 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 only exception occurs for Dep-SMUCE which delivers significant under-coverage on noise types `N4` and `N5`.

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 2.1 and 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 7 of the appendix, which shows that the same results hold for a range of signal lengths.

4.3. Coverage in the presence of strong serial dependence

Calibrating change point procedures in the presence of serial dependence is a difficult problem, and in practice few available methods work well uniformly; see for instance the numerical comparison in [18]. We remark that in the presence of strong serial dependence the coverage provided by our procedure can break down. To illustrate this, Table 4 reports the proportion of times over 100 replications for which DIF2-LRV reported no intervals on significance on the the signal

$$\zeta_t = \phi_j \zeta_{t-1} + \varepsilon_t \text{ with } \phi_j = 0.8 + j/100 \tag{20}$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d. and $j = 0, \dots, 10$. For large ϕ the procedure no longer delivers the desired coverage. However, on closer inspection this appears to be a failure of the long run variance estimator proposed in (12) which for values of ϕ close to 1 tends to under-estimate the long run variance, rather than the asymptotic theory. This is because scaling each local test by the true time average variance constant (TAVC, [88]), which for a given scale W'' is defined as

$$\mathrm{TAVC}\left(W''\right) = \mathbb{E}\left[\left(\frac{1}{\sqrt{W''}}\sum_{t=1}^{W''}\zeta_t\right)^2\right], \tag{21}$$

at a scale proportional to the W supplied to DIF2-LRV our procedure attains the desired level of coverage. The time average variance constant converges to the long run variance as long as W'' diverges with the sample size. As argued by [65] it is preferable to scale by the time average variance constant, as opposed to the long run variance, as with a properly chosen scale the latter better accounts for the local variation of each test. In fact, close inspection of the proof of

Table 4

Proportion of times out of 100 replications DIF2-LRV returned no intervals of significance when applied to a noise vector of length $n = 750$ form (20) and normalized with estimated long run standard deviation $\hat{\tau}_{DIF}$ calculated according to (12) as well as the TAVC calculated at scale $W'' = \frac{2}{5}\sqrt{n}$, the true long run standard deviation.

ϕ	0.8	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.9
Estimated LRV	0.94	0.86	0.89	0.89	0.86	0.81	0.84	0.83	0.62	0.72	0.51
True TAVC	0.98	0.96	0.96	0.98	0.96	0.94	0.93	0.96	0.90	0.95	0.87

Lemma 3.3 reveals that (12) is consistent for the TAVC calculated at the scale W'. However, the conditions of Lemma 3.3 limit (12) to scales of the order $W' = o\left(\sqrt{n}\right)$.

In light of the above, there are a number of approaches one may take if strong serial dependence is suspected. For instance, one could slightly pre-whiten the data using the heuristic methods suggested in Section 4.1 of [9]. Alternatively one may run DIF2-LRV with a conservative, but nonetheless consistent, estimator of the long run variance. For instance one may scale by $\hat{\tau}^2_{\mathrm{DIF}} + C/W'''$ for some positive constant $C > 0$ and some W''' diverging with n.

4.4. 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 100 sample paths from the change point models `M1`, `M2`, and `M3` listed below, contaminated with each of the four noise types introduced in Section 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 piecewise constant the `blocks` signal from [25], shown in Figure 1a, with $N = 4$ change points at locations $\Theta = \{205, 267, 308, 472\}$
- (`M2`): the first $n = 600$ values of the piecewise linear `waves` signal from [9], shown in Figure 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 6e, with $N = 3$ change points at locations $\Theta = \{100, 200, 300\}$

The results of the simulation study are reported in Tables 5 - 7. 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.

Table 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 constant `blocks` signal contaminated with noise `N1-N4` over 100 *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	N5	N6
DIF1-MAD	no. genuine	3.69	3.75	3.87	3.46	3.45	3.42
	prop. genuine	0.99	0.89	0.85	0.13	0.11	0.09
	length	34.86	27.19	23.89	9.36	8.87	8.30
	coverage	0.97	0.61	0.42	0.00	0.00	0.00
DIF2-SD	no. genuine	3.34	3.36	3.40	3.27	3.41	3.65
	prop. genuine	1.00	1.00	1.00	0.17	0.19	0.17
	length	43.72	43.80	43.41	16.63	16.59	16.15
	coverage	1.00	0.99	0.99	0.00	0.00	0.00
DIF2-LRV	no. genuine	1.98	2.03	1.97	1.35	1.33	1.39
	prop. genuine	0.99	1.00	0.99	0.90	0.91	0.95
	length	61.35	60.67	58.03	69.27	80.50	71.57
	coverage	1.00	1.00	1.00	1.00	1.00	1.00
NSP	no. genuine	3.20	3.48	3.52	3.67	3.74	3.86
	prop. genuine	1.00	0.63	0.60	0.15	0.12	0.10
	length	59.63	36.06	32.45	11.34	9.35	7.85
	coverage	1.00	0.14	0.12	0.00	0.00	0.00
NSP-SN	no. genuine	1.92	1.90	1.93	3.11	3.11	3.00
	prop. genuine	1.00	1.00	1.00	0.83	0.84	0.94
	length	120.41	117.62	113.95	75.43	75.16	73.08
	coverage	1.00	1.00	1.00	0.36	0.42	0.80
NSP-AR	no. genuine	0.12	0.87	0.88	0.68	0.99	1.56
	prop. genuine	0.12	0.54	0.55	0.61	0.51	0.89
	length	24.58	63.29	78.52	40.76	46.77	36.83
	coverage	1.00	0.50	0.54	1.00	0.43	0.98
B&P	no. genuine	3.85	3.88	3.93	3.49	3.73	3.60
	prop. genuine	0.96	0.96	0.98	0.19	0.20	0.25
	length	16.78	17.14	16.32	13.71	13.68	14.84
	coverage	0.83	0.85	0.92	0.00	0.00	0.00
MOSUM (uniscale)	no. genuine	1.96	2.02	2.06	3.54	3.59	3.51
	prop. genuine	0.83	0.83	0.88	0.20	0.21	0.24
	length	14.03	14.21	14.21	15.87	15.82	15.36
	coverage	0.89	0.91	0.94	0.00	0.00	0.00
MOSUM (multiscale)	no. genuine	3.90	3.98	3.97	4.90	4.96	4.69
	prop. genuine	0.96	0.98	0.98	0.23	0.23	0.25
	length	22.01	21.14	20.62	21.30	21.13	22.02
	coverage	0.86	0.93	0.92	0.00	0.00	0.00
SMUCE	no. genuine	3.71	3.61	3.82	2.12	1.87	1.68
	prop. genuine	0.95	0.72	0.73	0.09	0.06	0.05
	length	36.02	24.21	23.99	8.59	7.03	5.91
	coverage	0.89	0.35	0.24	0.00	0.00	0.00
H-SMUCE	no. genuine	3.40	3.08	3.21	2.81	2.91	2.42
	prop. genuine	0.92	0.86	0.89	0.84	0.84	0.77
	length	49.42	44.92	45.63	49.32	52.98	54.19
	coverage	0.80	0.70	0.72	0.63	0.65	0.56
Dep-SMUCE	no. genuine	2.12	2.15	2.30	3.11	2.94	3.42
	prop. genuine	0.80	0.78	0.82	0.49	0.46	0.61
	length	73.66	69.10	72.25	32.73	29.83	33.06
	coverage	0.57	0.59	0.61	0.00	0.01	0.06

Table 6

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	N5	N6
DIF1-MAD	no. genuine	2.98	2.77	2.66	1.51	1.60	2.28
	prop. genuine	0.98	0.83	0.77	0.06	0.06	0.05
	length	81.57	65.57	58.03	12.85	11.91	9.13
	coverage	0.92	0.49	0.39	0.00	0.00	0.00
DIF2-SD	no. genuine	2.99	2.98	2.97	1.66	1.71	2.56
	prop. genuine	1.00	0.99	0.99	0.08	0.09	0.08
	length	94.25	92.87	92.98	16.67	17.14	15.05
	coverage	0.99	0.98	0.96	0.00	0.00	0.00
DIF2-LRV	no. genuine	3.00	2.98	2.98	1.36	1.51	1.54
	prop. genuine	1.00	0.99	0.99	0.96	0.99	0.99
	length	95.78	97.32	98.90	233.37	219.28	239.46
	coverage	0.99	0.98	0.97	0.97	0.97	0.99
NSP	no. genuine	3.00	2.60	2.67	1.73	1.85	1.85
	prop. genuine	1.00	0.65	0.66	0.11	0.09	0.07
	length	93.06	58.00	57.31	20.36	17.06	14.02
	coverage	1.00	0.16	0.19	0.00	0.00	0.00
NSP-SN	no. genuine	2.99	3.00	3.00	2.42	2.34	2.44
	prop. genuine	1.00	1.00	1.00	0.84	0.83	0.96
	length	126.23	124.74	125.23	119.64	116.84	135.12
	coverage	1.00	1.00	1.00	0.58	0.57	0.90
NSP-AR	no. genuine	0.62	1.38	1.63	0.00	0.44	0.09
	prop. genuine	0.51	0.63	0.77	0.00	0.29	0.09
	length	98.93	96.03	113.80	0.00	59.68	22.95
	coverage	1.00	0.39	0.59	1.00	0.39	0.97

5. Real data examples

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 [6]. 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. There is some disagreement over the age at which peak bone mass density is attained in adolescents [57, 83, 62]. One possible solution is to model the data in Figure 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 4, because as the data are strictly positive the assumption of Gaussian noise is unlikely to hold. We additionally estimate change point

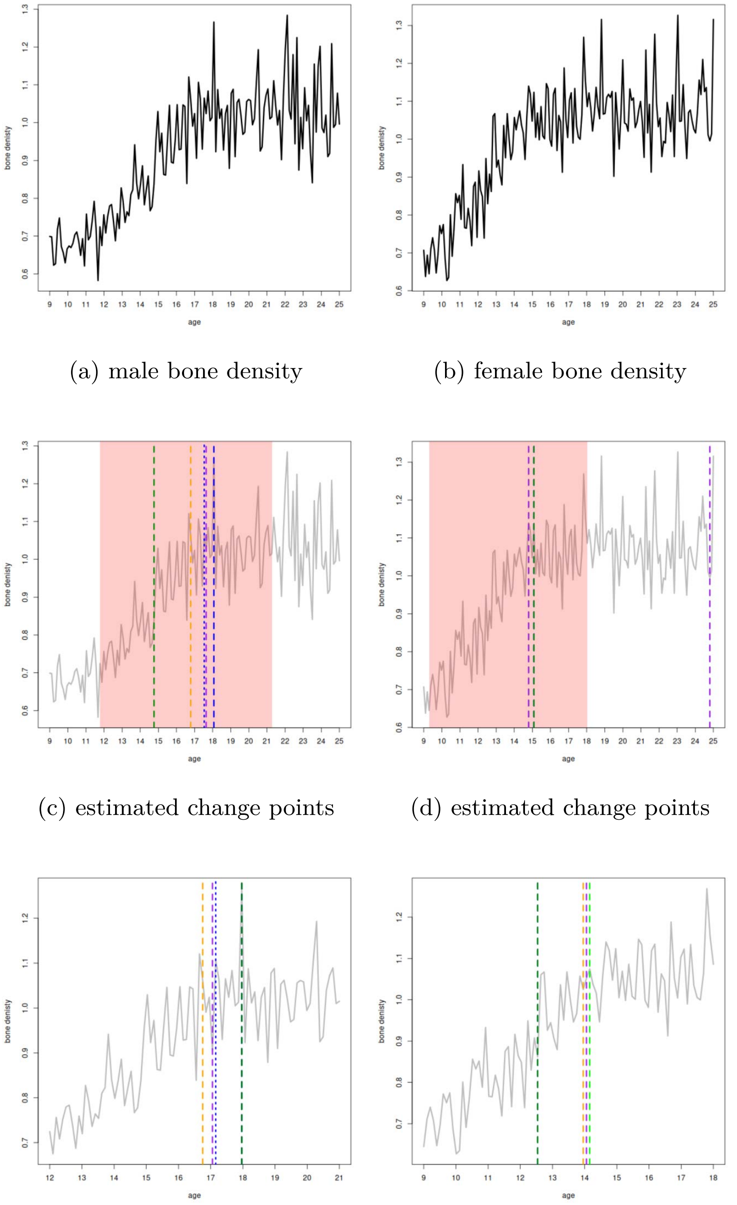
Table 7

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	N5	N6
DIF1-MAD	no. genuine	3.00	2.85	2.90	1.86	1.79	1.99
	prop. genuine	0.99	0.85	0.86	0.14	0.12	0.07
	length	43.32	36.10	35.09	16.66	14.93	8.60
	coverage	0.95	0.51	0.58	0.00	0.00	0.00
DIF2-SD	no. genuine	3.00	2.99	3.00	1.72	1.92	2.60
	prop. genuine	1.00	0.99	1.00	0.16	0.18	0.12
	length	51.96	51.37	51.16	19.77	20.36	16.01
	coverage	1.00	0.97	1.00	0.00	0.00	0.00
DIF2-LRV	no. genuine	3.00	3.00	3.00	1.82	1.72	1.69
	prop. genuine	1.00	1.00	1.00	0.98	0.95	0.98
	length	69.29	68.27	68.84	122.55	121.99	131.37
	coverage	1.00	1.00	1.00	0.99	0.95	1.00
NSP	no. genuine	3.00	2.89	2.97	2.04	2.13	1.99
	prop. genuine	1.00	0.83	0.87	0.25	0.22	0.16
	length	50.55	40.40	39.59	28.13	23.17	19.24
	coverage	1.00	0.44	0.60	0.00	0.00	0.00
NSP-SN	no. genuine	2.96	2.96	2.93	2.77	2.77	2.69
	prop. genuine	1.00	1.00	1.00	1.00	0.98	1.00
	length	83.66	83.21	83.62	92.23	90.86	95.98
	coverage	1.00	1.00	1.00	1.00	0.95	1.00
NSP-AR	no. genuine	0.52	1.40	1.60	0.03	0.55	0.24
	prop. genuine	0.44	0.75	0.85	0.03	0.40	0.24
	length	60.37	64.41	72.31	3.51	52.93	41.64
	coverage	1.00	0.65	0.79	1.00	0.48	0.93

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 [9], and the same algorithm run with the requirement that the estimated signal be continuous (NOT-cont), the Isolate Detect algorithm (ID) of [2], the dynamic programming based algorithm of [7] (BP), and the Continuous-piecewise-linear Pruned Optimal Partitioning algorithm (CPOP) of [34]. 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. 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



(a) male bone density

(b) female bone density

(c) estimated change points

(d) estimated change points

(e) change points on sub-interval

(f) change points on sub-interval

Fig 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 (- - -).

data set save the extraneous change point detected by CPOP. By Corollary 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, 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 4.1 applicable to the data, cover essentially the entire range of the data.

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 [18]. We follow their analysis in [17] by taking the square root transform of the data and removing seasonal and weekly variation. The processed data is plotted in Figure 4. [18] 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 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 of the art algorithms for recovering changes

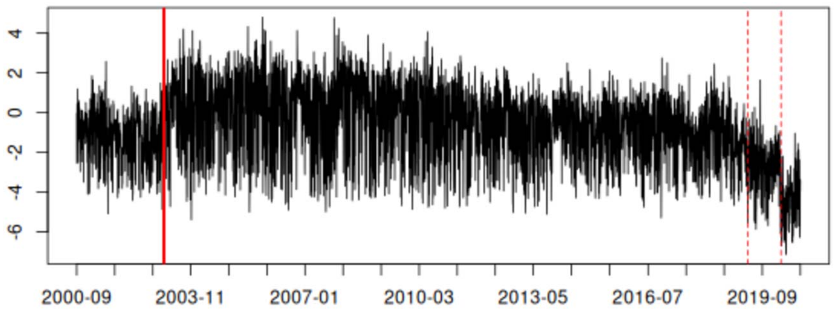


Fig 4: daily average concentrations of NO_2 at Marylebone Road after square root transform and with seasonal variation removed, red dashed lines (- - -) and dark red shaded region (🟥) represent dates of events which are likely to have affected NO_2 concentration levels.

in piecewise constant signals in the presence of serially correlated noise, which however do not come with coverage guarantees. These are: the algorithm of [81] for Detecting Changes in Autocorrelated and Fluctuating Signals (DeCAFS), the algorithm of [13] 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 [18]. 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.

The results of the analysis are shown in Figure 5. DIF2-LRV returns four intervals, among which the first, third, and fourth cover the dates of important events identified by [18]. 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 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.

6. Proofs

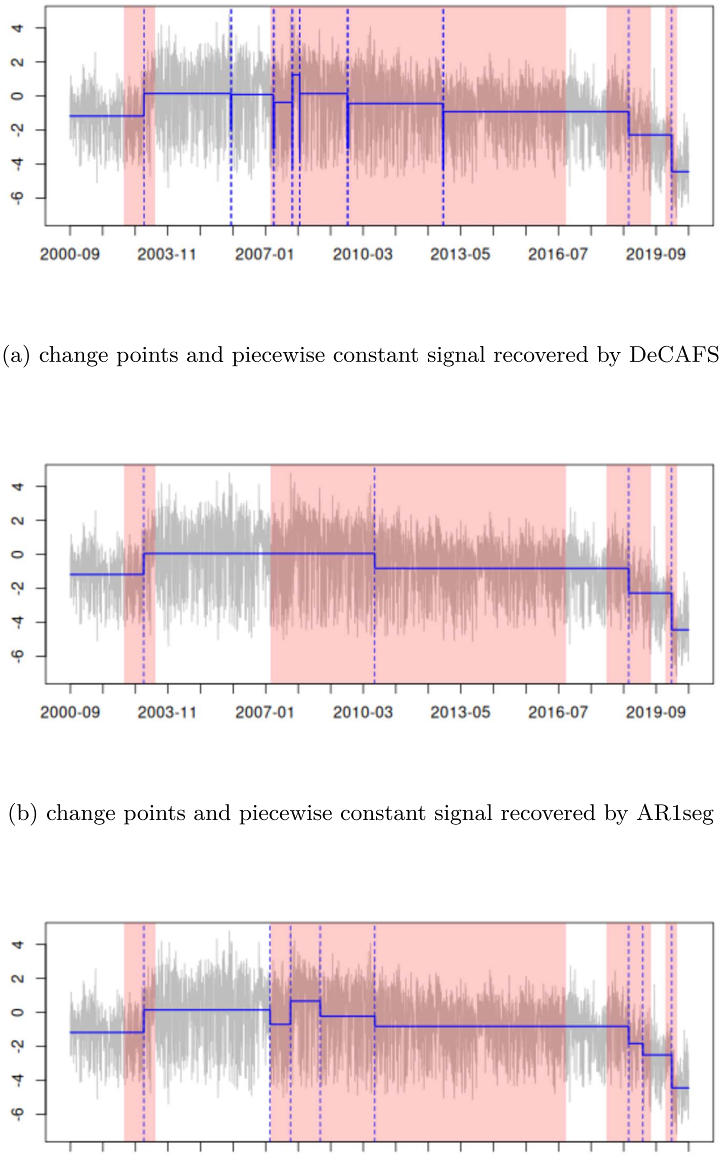
For sequences $\{a_n\}_{n>0}$ and $\{b_n\}_{n>0}$ we write $a_n \lesssim b_n$ if there is a constant $C > 0$ for which $a_n \leq Cb_n$ for every $n > 0$. We write $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. We write $|\mathcal{A}|$ for the cardinality of a set $\mathcal{A}$. The density, cumulative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

6.1. Preparatory results

Definition 6.1. *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.



(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

Fig 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.

Lemma 6.1 (Berman's lemma)**.** *Let $\zeta_1, \ldots, \zeta_n$ and $\tilde{\zeta}_1, \ldots, \tilde{\zeta}_n$ be two sequences of Gaussian 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 $\rho_{ij} = \max\left(|\Lambda_{ij}|, |\tilde{\Lambda}_{ij}|\right)$. For any real numbers $u_1, \ldots, u_n$ the following holds:*

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

Proof. See Theorem 4.2.1 in [59]. □

Lemma 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 [59]. □

Lemma 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:*

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

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

Proof. See Theorem 9.15 in [75], and Remark 12.2.10 in [59] for the values of H_α. □

Lemma 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

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

Proof. See Lemma 12.2.1 in [59]. □

Lemma 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

$$G(y) \sim 1/2y \text{ as } y \to \infty.$$

Proof. See Theorem 7.2 and Corollary 3.18 respectively in [49]. □

6.2. Intermediate results

Lemma 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 $\frac{1}{\sigma}D_{l,w}^p(Y)$ under Assumption 2.1 and the null of no change points, in the sense that for a given scale w the following holds:

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

(ii) According to Definition 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 $\mathbb{E}\left(\xi^2(l)\right) = 1$, so it remains to calculate the covariance between $\xi\left(l\right)$ and $\xi\left(l+s_l\right)$ for $|s_l| \to 0$. First, taking $s_l > 0$ we have the following:

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

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

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

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

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

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

Lemma 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*

$$T_{1,m}^{\lambda} = \mathbf{1}\left\{|D_{1,(p+2)\delta}^{p}\left(\mathbf{Y}\right)| > \lambda\right\}$$

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

$$\left\{L_{\mathcal{G}(W,a)}^{\widehat{\tau}}\left(\boldsymbol{\zeta}\right) \leq \bar{\lambda}\right\} \cap \left\{\widehat{\tau} < 2\tau\right\} \tag{22}$$

as long Assumption 3.3 is satisfied and δ' satisfied the inequality

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

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:

$$\left|D_{1,m}^{p}\left(\boldsymbol{f}\right)\right| > \left|D_{1,m}^{p}\left(\boldsymbol{\zeta}\right)\right| + \lambda. \tag{23}$$

Moreover on (22) 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 $\delta > 1$:

$$\begin{aligned}
\frac{1}{\delta'}\sum_{t=1}^{\delta'}(1-t/\delta')^p &= (\delta')^{-(p+1)}\sum_{s=1}^{\delta'-1}s^p \\
&= \left(\frac{1}{p+1}\right)\left(\frac{\delta'-1}{\delta'}\right)^{p+1}\sum_{k=0}^{p}\binom{p+1}{k}B_k\left(\delta'-1\right)^{-k} \\
&\geq \left(\frac{1}{p+1}\right)\left(\frac{\delta'-1}{\delta'}\right)^{p+1} \\
&\geq \frac{1}{2^{p+1}(p+1)}.
\end{aligned} \tag{24}$$

Using the above along with Assumption 3.3 and the fact that the test statistic (2) is invariant to the addition of arbitrary degree p polynomials we have the following:

$$\begin{aligned}
\left|D_{1,m}^{p}\left(\boldsymbol{f}\right)\right| &= \left\{\delta\sum_{i=0}^{p+1}\binom{p+1}{i}^2\right\}^{-\frac{1}{2}}\left|\sum_{j=0}^{p}\Delta_j\sum_{t=1}^{\delta}\left(\frac{t}{n}-\frac{\delta}{n}\right)^j\right| \\
&\geq \sqrt{\delta}\left|\Delta_{p^*}\right|\left(\frac{\delta}{n}\right)^{p^*}\left[\frac{\frac{1}{\delta}\sum_{t=1}^{\delta}\left(1-\frac{t}{\delta}\right)^{p^*}}{\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}}\right] \\
&\quad - \sum_{\substack{0\leq j\leq p \\ j\neq p^*}}\sqrt{\delta}\left|\Delta_j\right|\left(\frac{\delta}{n}\right)^{j}\left[\frac{\frac{1}{\delta}\sum_{t=1}^{\delta}\left(1-\frac{1}{\delta}\right)^{j}}{\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}}\right] \\
&\geq C_{p,p^*}^{-1}\left|\Delta_{p^*}\right|\delta^{\frac{2p^*+1}{2}}n^{-p^*}.
\end{aligned} \tag{25}$$

Therefore combining (23) and (25) we have that on the event (22) the change will be detected if $C_{p,p^*}^{-1}\left|\Delta_{p^*}\right|\delta^{\frac{2p^*+1}{2}}n^{-p^*} > 4\tau\bar{\lambda}$, and the desired result follows by rearranging. □

Theorem 6.1. *Put $w = \lfloor c\log(n) \rfloor$ for some constant $c > 0$ and introduce maximum of the local test statistics (2) appropriately standardised and restricted to scales w as follows:*

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

Then under Assumption 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 2.1:

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

Proof. Omitting dependence on x introduce the following 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} defined as follows:

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

The proof proceeds in three steps.

STEP 1: we first show that the behaviour 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 $\mathfrak{u}_n^2 = 2\log(n) - \log\log(n) + \mathcal{O}\left(1\right)$ the following holds:

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

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 as shown below:

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

For any l_1, l_2 write $j_{1,2} = |\{l_1, \ldots, l_1 + w - 1\} \cap \{l_2, \ldots, l_2 + w - 1\}|$ and put $\Lambda_{l_1,l_2} = \Lambda_{j_{1,2}}$. Writing $\mathcal{A}_n = \cup_i A_i$ and using Lemma 6.1 we have the following:

$$\begin{aligned}
&\left| \mathbb{P}\left(\max_{l \in \mathcal{A}_n} \left\{ \frac{1}{\sigma} D^p_{l,w}(\mathbf{Y}) \right\} \leq \mathfrak{u}_n \right) - \mathbb{P}\left(\max_{l \in \mathcal{A}_n} \left\{ \frac{1}{\sigma} \tilde{D}^p_{l,w}(\mathbf{Y}) \right\} \leq \mathfrak{u}_n \right) \right| \qquad (26) \\
&\leq \frac{1}{2\pi} \sum_{\substack{l_1 \in A_i, l_2 \in A_j \\ i \neq j}} \left| \Lambda_{l_1,l_2} - \tilde{\Lambda}_{l_1,l_2} \right| \left(1 - \Lambda^2_{l_1,l_2}\right)^{-1/2} \exp\left(- \frac{\mathfrak{u}_n^2}{1 + \Lambda_{l_1,l_2}} \right) \\
&\lesssim \sum_{i=0}^{|\mathcal{A}_n|/|A_0|} \sum_{\substack{l_1 \in A_i \\ l_2 \in A_{i+1}}} \left| \Lambda_{l_1,l_2} - \tilde{\Lambda}_{l_1,l_2} \right| \left(1 - \Lambda^2_{l_1,l_2}\right)^{-1/2} \exp\left(- \frac{\mathfrak{u}_n^2}{1 + \Lambda_{l_1,l_2}} \right) \\
&\lesssim \frac{|\mathcal{A}_n|}{|A_0|} \sum_{l=1}^{|A_0|} \sum_{j=1}^{l} \Lambda_j \left(1 - \Lambda_j^2\right)^{-1/2} \exp\left(- \frac{2\log(n) - \log\log(n)}{1 + \Lambda_j} \right) \\
&\lesssim \log(n) \frac{|\mathcal{A}_n|}{|A_0|} \sum_{l=1}^{|A_0|} \sum_{j=1}^{l} \Lambda_j \left(1 - \Lambda_j^2\right)^{-1/2} \exp\left(- \frac{2\log(n)}{1 + \Lambda_j} \right).
\end{aligned}$$

Note that for some fixed $K > 0$ depending on p it must hold that $\Lambda_j \leq \min(jK, w - w^\rho)/w$. Therefore the first term after the double sum can be bounded as follows:

$$\begin{aligned}
\Lambda_j \left(1 - \Lambda_j^2\right)^{-1/2} &\leq \Lambda_j \left(1 - \Lambda_j\right)^{-1/2} \\
&\leq \min(jK, w - w^\rho) / \sqrt{(w - \min(jK, w - w^\rho))\, w} \\
&\leq \min(jK, w - w^\rho) / \sqrt{w}. \qquad (27)
\end{aligned}$$

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

$$\begin{aligned}
\delta_j &= (1 - \Lambda_j) / (1 + \Lambda_j) \\
&\geq (w - \min(jK, w - w^\rho)) / (w + \min(jK, w - w^\rho)) \\
&\geq (w - \min(jK, w - w^\rho)) / 2w. \qquad (28)
\end{aligned}$$

Therefore substituting (27) and (28) into (26) we obtain the following:

$$\begin{aligned}
(26) &\lesssim \frac{\sqrt{\log(n)}}{n} \frac{|\mathcal{A}_n|}{|A_0|} \sum_{j=1}^{l} \min(jK, w - w^\rho) \left(n^{\frac{1}{2w}} \right)^{-(w - \min(jK, w - w^\rho))} \\
&= \frac{\sqrt{\log(n)}}{n} \frac{|\mathcal{A}_n|}{|A_0|} \left\{ \sum_{l=1}^{\lfloor |A_0|/K \rfloor} \sum_{j=1}^{l} jK \left(n^{\frac{1}{2w}} \right)^{-(w - jK)} + \ldots \right. \\
&\quad \left. \cdots + \sum_{l=\lfloor |A_0|/K \rfloor + 1}^{|A_0|} \sum_{j=1}^{l} (w - w^\rho) \left(n^{\frac{1}{2w}} \right)^{w^\rho} \right\}. \qquad (29)
\end{aligned}$$

The first sum in (29) can be bounded as follows:

$$\begin{aligned}\sum_{l=1}^{\lfloor |A_0|/K\rfloor}\sum_{j=1}^{l} jK\left(n^{\frac{1}{2w}}\right)^{-(w-jK)} &\lesssim n^{-1/2}\int_1^{\lfloor |A_0|/K\rfloor+1}\int_1^{y+1} x\left(n^{\frac{1}{2w}}\right)^{Kx}\mathrm{d}x\mathrm{d}y\\ &\lesssim wn^{-w^{-(1-\rho)}/2}. \qquad (30)\end{aligned}$$

The second sum in (29) can be bounded as follows:

$$\begin{aligned}\sum_{l=\lfloor |A_0|/K\rfloor+1}^{|A_0|}\sum_{j=1}^{l}(w-w^{\rho})\left(n^{\frac{1}{2w}}\right)^{w^{\rho}} &\lesssim wn^{-w^{-(1-\rho)}/2}\sum_{l=\lfloor |A_0|/K\rfloor+1}^{|A_0|}(l)\\ &\lesssim w^3 n^{-w^{-(1-\rho)}/2}. \qquad (31)\end{aligned}$$

Finally plugging (30) and (31) into (26) 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:

$$\begin{aligned}(26) &\lesssim \frac{\sqrt{\log(n)}}{n}\frac{|\mathcal{A}_n|}{|A_0|}\left\{n^{-w^{-(1-\rho)}/2}\left(w+w^3\right)\right\}\\ &\lesssim \log^{5/2}(n)n^{-w^{-(1-\rho)}/2}\\ &\lesssim \exp\left(-C\log^{\rho}(n)\right).\end{aligned}$$

STEP 3: we now prove Theorem 6.1. Using Lemma 6.4 and part (i) of Lemma 6.6 and noting that $\mathfrak{u}_n^2/w \sim 2/c$ gives the following for any $i=0,\dots,|\mathcal{A}_n|-1$

$$\mathbb{P}\left(\max_{l\in A_i}\left\{\frac{1}{\sigma}D_{l,w}^{p}\left(\mathbf{Y}\right)\right\} > \mathfrak{u}_n\right) \sim \left(\frac{w}{n}\right)\left(\frac{2C_p}{c}\right)F\left(\frac{2C_p}{c}\right)e^{-x}. \qquad (32)$$

It is evident that

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

Therefore (32), the results of step 2, and that $|\mathcal{A}_n|/|A_0| \sim n/w$ imply that

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

Going the other way it is also evident that

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

Using (32) and the results of Steps 1 and 2 gives the following:

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

Therefore, the theorem is proved. □

6.3. *Proof of Theorem 2.1*

Proof. Given the result in part (i), part (ii) follows immediately from Lemma 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 (5) to scales no larger than Wa^A:

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

Introduce also the restriction of (5) to scales strictly larger than Wa^A:

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

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 6.3 we have the following:

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

$$
\begin{aligned}
&\leq \sum_{k=k_n+A}^{\lfloor \log_a(n/2)\rfloor} \left(\frac{n}{a^k}\right) \mathbb{P}\left(\sup_{t\in[0,1)} \{\xi(t)\} > \mathfrak{u}_n\right)\\
&\lesssim \sum_{k=k_n+A}^{\lfloor \log_a(n/2)\rfloor} \left(\frac{n}{a^k}\right) \mathfrak{u}_n e^{-\mathfrak{u}_n^2/2}\\
&\lesssim \frac{a^{-A}}{1-a^{-1}}.
\end{aligned}
$$

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

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

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

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

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 $\Lambda_{l_1,w_1,l_2,w_2} \leq \Lambda_a$. Therefore using Lemma 6.1 we have the following:

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

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

$$\left\{ \mathfrak{a}_n M^{\sigma}_{\mathcal{G}(W,a)}(\mathbf{Y}) - \mathfrak{b}_n \mid n \in \mathbb{N} \right\} \tag{33}$$

converges weakly to a Gumbel distribution. On such a sub-sequence for each $j \in \mathbb{N}$ we have that $a^{k_n+j} \sim a^j bd \times \log(n)$. Therefore from Theorem 6.1 we have that

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

The following inequality is evident:

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

Therefore (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 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 (6.3) and the result from steps 1 and 2 imply the following:

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

Therefore, the statement is proved.

STEP 4: we now prove the result in part (i). Since b_n may have any subsequential limit between $1/a$ and 1 it follows from step 4 that the sequence of random variables (33) is tight. Using part (i) of Lemma 6.5 the constants in (7) are easily recognised as the largest and smallest constants which may appear in the extreme value limit. □

6.4. Proof of Theorem 2.2

Proof. With W satisfying Assumption 2.5 and omitting dependence on x introduce the following 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 (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

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

Putting $\mathbf{B} = (B(1), \ldots, B(n))'$, where $\{B(t)\}_{t>0}$ is the process introduced in Assumption 2.4, making use of Assumption 2.4 the following holds:

$$M^{\tau}_{n,W}\left(\mathbf{Y}\right) = \max\left\{D^{p}_{l,w}\left(\mathbf{B}\right) \mid 1 \leq l \leq n-w\right\} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{n^{\frac{2}{2+\nu}}/W}\right). \tag{34}$$

Moreover, using Lemma 6.3 and arguing as in the proof of Theorem 6.1 the following holds:

$$\begin{aligned}\mathbb{P}\left(M^{1}_{cW}\left(\mathbf{B}\right) \leq \mathfrak{u}_{n,W}\right) &\sim \prod_{i=0}^{\lfloor n/w \rfloor} \mathbb{P}\left(\max\left\{D^{p}_{l,w}\left(\mathbf{B}\right) \mid i \times w < l \leq (i+1) \times w\right\} \leq \mathfrak{u}_{n,W}\right)\\ &\sim \left[1 - \mathbb{P}\left(\sup_{l\in[0,1)}\left\{\xi\left(l\right)\right\} > \mathfrak{u}_{n,W}\right)\right]^{\lfloor n/w \rfloor}\\ &\sim \exp\left(-\frac{C_p}{c}e^{-x}\right). \end{aligned} \tag{35}$$

Therefore, combining (34) and (35) and arguing as in the proof of Theorem 2.1, we immediately have that

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

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 2.1, we therefore have under the null of no change points that

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

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 6.2. □

6.5. Proof of Lemma 3.1

Proof. Write $m = (n-p-1)/(p+1)$ and $c_{\Phi} = \Phi\left(2\Phi^{-1}\left(3/4\right)\right) - 3/4$. For some $\varepsilon > 0$ not depending on n put $A_{\varepsilon} = \left(3/c_{\Phi} + \sqrt{9/c_{\Phi}^2 + 2\varepsilon}\right)/2$, and therefore define

$$\delta = \frac{A_{\varepsilon}\sigma}{c_{\Phi}}\left(\frac{1}{\sqrt{m}} \vee \frac{N}{m}\right). \tag{36}$$

We will show that with n sufficiently large

$$\mathbb{P}\left(|\hat{\sigma}_{\mathrm{MAD}} - \sigma| > \delta\right) \leq 2(p+1)e^{-\varepsilon}, \tag{37}$$

which implies the desired result. 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

$$B_t^\delta = \mathbf{1}\left\{ |X_t| > \Phi^{-1}(3/4)\sqrt{\sum_{i=0}^{p+1} \binom{p+1}{i}^2}\,[\sigma+\delta] \right\}, \quad t = p+1, \ldots, n$$

and put $p_\delta = \mathbb{E}\left(B_t^\delta \mid t \notin I_\eta\right)$. The following holds via Hoeffding's inequality:

$$\begin{aligned} &\mathbb{P}\left(\widehat{\sigma}_{\mathrm{MAD}} - \sigma > \delta\right) \\ &\quad = \mathbb{P}\left(\frac{\operatorname{median}\{|X_{p+1}|, \ldots, |X_n|\}}{\Phi^{-1}(3/4)\sqrt{\sum_{i=0}^{p+1}\binom{p+1}{i}^2}} > \sigma + \delta \right) \\ &\quad \leq \sum_{j=0}^{p} \mathbb{P}\left(\sum_{t \in I_{j,1}} B_t^\delta + \sum_{t \in I_{j,2}} B_t^\delta > \frac{n-(p+1)}{2(p+1)} \right) \\ &\quad \leq \sum_{j=0}^{p} \mathbb{P}\left(\sum_{t \in I_{j,1}} \left(B_t^\delta - p_\delta\right) > \frac{n-(p+1)}{2(p+1)} - |I_{j,2}| - p_\delta\,|I_{j,1}| \right) \\ &\quad \leq (p+1)\exp\left(-2m\left[(1/2 - p_\delta)^2 + (N/m)^2 - 2\,(1/2 - p_\delta)\,(N/m) \right]^2 \right). \end{aligned} \tag{38}$$

We now bound p_δ from above and from below. For the lower bound, putting $Z \sim \mathcal{N}(0,1)$ we have that

$$\begin{aligned} p_\delta &= \mathbb{P}\left(|Z| > \Phi^{-1}(3/4)\,[1+\delta/\sigma]\right) \\ &= 2\left(1 - \int_{-\infty}^{\Phi^{-1}(3/4)} \phi\left(x\,[1+\delta/\sigma]\right) \mathrm{d}x\,[1+\delta/\sigma] \right) \\ &\geq 2\left(1 - \Phi\left(\Phi^{-1}(3/4)\right)[1+\delta/\sigma]\right) \\ &= 1/2 - (3/2) \times (\delta/\sigma), \end{aligned} \tag{39}$$

which holds because for all $\alpha > 1$ and any $x \in \mathbb{R}$ it holds that $\phi(\alpha x) \leq \phi(x)$. For the upper bound write $f(x) = \Phi\left(\Phi^{-1}(3/4)\,(1+x)\right)$ for $x \in [0,1]$. Then using the facts that (i) $f(0) = 3/4$, (ii) $f(1) = \Phi\left(2\Phi^{-1}(3/4)\right)$, (iii) $\Phi(\cdot)$ is concave on $[0,1]$, and (iv) $\delta/\sigma \leq 1$ for n sufficiently large, we obtain that

$$p_\delta = 2\left(1 - \Phi\left(\Phi^{-1}(3/4)\left[1 + \delta/\sigma\right]\right)\right) \leq 1/2 - \frac{c_\Phi \delta}{\sigma} \tag{40}$$

for n sufficiently large. Therefore plugging (39), (40) and (36) into (38) we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\widehat{\sigma}_{\mathrm{MAD}} - \sigma > \delta\right)\\
&\leq (p+1)\\
&\quad \times \exp\left(-2m\left[A_\varepsilon^2\left(m^{-1/2} \vee N/m\right)^2 + (N/m)^2 - \frac{3A_\varepsilon}{c_\Phi}\left(m^{-1/2} \vee N/m\right)(N/m)\right]\right)\\
&\leq (p+1)\exp\left(-2m\left[\left(A_\varepsilon^2 - \frac{3A_\varepsilon}{c_\Phi}\right)\left(m^{-1/2} \vee N/m\right)^2\right]\right)\\
&\leq (p+1)\exp\left(-\varepsilon\right).
\end{aligned}$$

Similar arguments give identical bounds on the probability that $\hat{\sigma}_{\mathrm{MAD}} - \sigma$ is smaller that a given δ, which overall establishes (37). □

6.6. *Proof of Lemma 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 [24].

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

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

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

It therefore follows that

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

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

6.7. Proof of Lemma 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 $\hat{\tau}^2_{\text{DIF}} = \frac{1}{mW}\bar{\boldsymbol{Y}}'\boldsymbol{D_p}\bar{\boldsymbol{Y}}$, and the absolute difference between our estimator and the truth can be bounded as follows:

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

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

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

From which it follows that

$$\frac{1}{mW'}\mathbb{E}\left(\bar{\boldsymbol{\zeta}}'\boldsymbol{D_p}\bar{\boldsymbol{\zeta}}\right) = \sum_{j=0}^{p+1}\psi^2_{p,j}\left(\gamma_0 + 2\sum_{h=1}^{W'-1}\left(1 - \frac{h}{W'}\right)\gamma_h\right)$$

$$+\sum_{\substack{k\neq l\\0\le k,l\le p+1}}\psi_{p,k}\psi_{p,l}\left(\gamma_{W'|k-l|}+2\sum_{h=1}^{W'-1}\left(1-\frac{h}{W'}\right)\gamma_{W'|k-l|+h}\right).$$

Using these facts term T_1 can be bounded as follows

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

For the first term we have that

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

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

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

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. □

6.8. *Proof of Theorem 3.1*

Proof. With slight abuse of notation write $I \in \mathcal{G}(W,a)$ if $I = \{l,\dots,l+w-1\}$, where $(l,w) \in \mathcal{G}(W,a)$. For each $k = 1,\dots,N$ introduce the set of intervals

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

Moreover assume that

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

Since $\lambda_\alpha^2 = \mathcal{O}(\log(n))$ for any fixed α and either of threshold (6) or threshold (8), this assumption can be seen to correspond to condition (16) in Theorem 3.1. For ease of reading introduce the notation

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

Due to lemma 6.7, testing for a change point on an interval $I' \in \mathcal{I}_k$ using (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

$$\left\{L_{\mathcal{G}(W,a)}^{\widehat{\tau}}(\boldsymbol{\zeta}) \le \lambda_\alpha\right\} \cap \{\widehat{\tau} < 2\tau\}. \tag{41}$$

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 (41) 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. □

6.9. *Proof of Lemma 3.4*

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

$$\begin{aligned}
&\mathrm{sSIC}(p') - \mathrm{sSIC}(p)\\
&\quad = \frac{n}{2}\log\left(1 - \frac{\hat\sigma_p^2 - \hat\sigma_{p'}^2}{\hat\sigma_p^2}\right)\\
&\quad\quad + \left[\left(\hat N_{p'} + 1\right)(p'+1) - \left(\hat N_p + 1\right)(p+1)\right]\log^\alpha(n)\\
&\quad := T_1 + T_2.
\end{aligned}$$

Observe that by Corollary 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 [82], 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 $p' \ge p$. Consequently

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

Again by Corollary 3.2 we have that with high probability

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

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

$$\begin{aligned}\mathrm{sSIC}(p') - \mathrm{sSIC}(p) & \\ &= \frac{n}{2}\log\left(\frac{\hat{\sigma}^2_{p'}}{\hat{\sigma}^2_{p}}\right) + \left[\left(\hat{N}_{p'}+1\right)(p'+1) - \left(\hat{N}_{p}+1\right)(p+1)\right]\log^{\alpha}(n) \\ &:= T_1 + T_2. \end{aligned} \tag{42}$$

By condition (iii) on a high probability set we must have that T_1 is negative and of the order $\mathcal{O}(n\log(n))$, 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, sSIC$(p') >$ sSIC(p) for all $p' \neq p$. This completes the proof. □

6.10. Remarks on Assumption 3.3

We remark that 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 1 is able to localize all change points at the optimal rate when the assumption is violated, albeit with different leading constants in (16). The reason for the claim is the following: Assumption 3.3 is made to avoid the possibility of signal cancellation will, however examining the proof of Lemma 6.7 it can be seen that there are only p values of δ' for which exact signal cancellation occurs, 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 1 can localize change points at the optimal rate in the absence of Assumption 3.3 the when signal is piecewise linear. Moreover we provide some simulated examples of piecewise polynomial signals which violate Assumption 3.3 and show that the change points are still detected.

6.10.1. Relaxing Assumption 3.3 for piecewise linear signals signals

Here we show that for piecewise linear signals Algorithm 1 is able to localize all changes at the minimax optimal rate when Assumption 3.3 does not hold, provided the remaining assumption in Theorem 3.1 hold. Without loss of generality we consider the case of a single change:

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

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 (14). If $\mathrm{sign}(\alpha_0 - \beta_0) = \mathrm{sign}(\alpha_1 - \beta_1)$ this can be shown precisely as in Lemma 6.7. Therefore, we examine the setting in which $\mathrm{sign}(\alpha_0 - \beta_0) \neq \mathrm{sign}(\alpha_1 - \beta_1)$, 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 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

$$C_1 = 1/\sqrt{\sum_{i=0}^{2}\binom{2}{i}^2}$$

$$g_{\delta'} = \frac{1}{\delta'}\sum_{t=1}^{\delta'}(1 - t/\delta') \text{ for } \delta' \in \mathbb{N}.$$

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

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

and the desired result follows by rearranging (23).

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 (24), we have that

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

Therefore, rearranging (23) and accounting for the facts that we must have $\delta' > 24\delta''$ we obtain that the change will be detected as soon as

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

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

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

and as in the previous cases the desired result follows by rearranging (23).

6.10.2. Examples of higher order polynomials which violate 3.3

Here we give simulated examples of higher order piecewise polynomial signals which violate Assumption 3.3, and show that Algorithm 1 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:

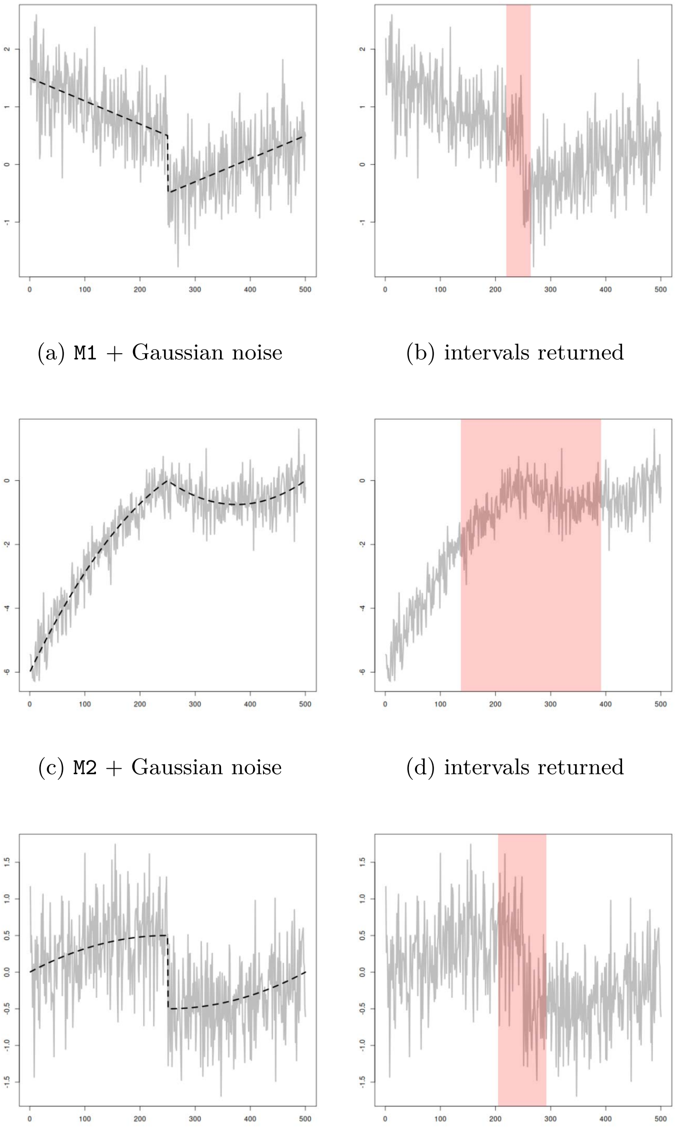
$$f_\circ\left(t/n\right) = \begin{cases} \alpha_0 + \alpha_1\left(t/n - \eta/n\right) + \alpha_2\left(t/n - \eta/n\right)^2 & \text{if } t \leq \eta \\ \beta_0 + \beta_1\left(t/n - \eta/n\right) + \beta_2\left(t/n - \eta/n\right)^2 & \text{if } t > \eta \end{cases} \tag{43}$$

We consider three instances of (43) 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 $\Delta_j\left(\delta/n\right)^j$ 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 8 below.

We contaminate the signals with independent noise having marginal $\mathcal{N}\left(0, 0.5^2\right)$ distribution and apply Algorithm 1 with parameter $\alpha = 0.1$. The results of this experiment, which was run with random seed `42` in `R`, are shown in Figure 6. In all three cases Algorithm 1 returns a single interval which contains the true change point location.

Table 8
Values of α's and β's for three instances of (43) which violate Assumption 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



(a) `M1` + Gaussian noise

(b) intervals returned

(c) `M2` + Gaussian noise

(d) intervals returned

(e) `M3` + Gaussian noise

(f) intervals returned

Fig 6: Piecewise polynomial signals which violate Assumption 3.3 with coefficients specified in Table 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.

7. 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 4.2 for signals of length $n \in \{100, 500, 1000, 2000\}$. The results are shown in Tables 9, and confirm that for a range of signal lengths our procedure continues to provide accurate coverage.

Table 9

Proportion of times out of 100 *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.*

		degree 0	degree 1	degree 2
n = 100	DIF1-MAD	0.93	0.92	0.94
	DIF2-SD	1.00	1.00	1.00
	DIF2-LRV	0.98	0.92	0.95
n = 500	DIF1-MAD	0.93	0.97	0.94
	DIF2-SD	1.00	1.00	1.00
	DIF2-LRV	0.99	0.98	0.96
n = 1000	DIF1-MAD	0.93	0.92	0.93
	DIF2-SD	1.00	1.00	0.99
	DIF2-LRV	0.98	0.98	0.99
n = 2000	DIF1-MAD	0.88	0.93	0.95
	DIF2-SD	1.00	0.99	0.99
	DIF2-LRV	0.98	0.98	0.95

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

		degree 0	degree 1	degree 2
n = 100	DIF1-MAD	0.88	0.64	0.81
	DIF2-SD	0.99	0.98	0.99
	DIF2-LRV	1.00	0.91	0.91
n = 500	DIF1-MAD	0.52	0.42	0.50
	DIF2-SD	0.98	0.97	0.99
	DIF2-LRV	0.99	0.93	0.94
n = 1000	DIF1-MAD	0.43	0.28	0.39
	DIF2-SD	0.94	0.99	0.97
	DIF2-LRV	0.91	0.95	0.90
n = 2000	DIF1-MAD	0.23	0.19	0.18
	DIF2-SD	0.99	0.97	0.96
	DIF2-LRV	0.99	0.97	0.94

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

		degree 0	degree 1	degree 2
n = 100	DIF1-MAD	0.72	0.71	0.65
	DIF2-SD	0.98	1.00	0.99
	DIF2-LRV	0.95	0.94	0.91
n = 500	DIF1-MAD	0.40	0.31	0.37
	DIF2-SD	0.99	0.97	1.00
	DIF2-LRV	0.98	0.94	0.91
n = 1000	DIF1-MAD	0.23	0.26	0.37
	DIF2-SD	0.98	0.96	0.98
	DIF2-LRV	0.98	0.97	0.93
n = 2000	DIF1-MAD	0.17	0.15	0.15
	DIF2-SD	0.96	0.98	0.99
	DIF2-LRV	0.97	0.98	0.97

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

Table 10

Proportion of times out of 100 *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.*

		degree 0	degree 1	degree 2
n = 100	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.74	0.79	0.77
n = 500	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.92	0.87	0.91
n = 1000	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.95	0.89	0.94
n = 2000	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.98	0.98	0.97

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

		degree 0	degree 1	degree 2
n = 100	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.75	0.69	0.72
n = 500	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.89	0.82	0.88
n = 1000	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.98	0.90	0.89
n = 2000	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.94	0.98	0.98

(b) Coverage on noise type `N5` with $\sigma = 1$

		degree 0	degree 1	degree 2
n = 100	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.97	0.91	0.94
n = 500	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.99	1.00	0.99
n = 1000	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.95	0.98	0.95
n = 2000	DIF1-MAD	0.00	0.00	0.00
	DIF2-SD	0.00	0.00	0.00
	DIF2-LRV	0.99	0.97	0.99

(c) Coverage on noise type `N6` with $\sigma = 1$

Acknowledgments

The authors would like to thank Prof. Zakhar Kabluchko for helpful discussions.

References

[1] Abramovich, F., Antoniadis, A. and Pensky, M. (2007), 'Estimation of piecewise-smooth functions by amalgamated bridge regression splines', *Sankhyā: The Indian Journal of Statistics* pp. 1–27. MR2385276

[2] Anastasiou, A. and Fryzlewicz, P. (2022), 'Detecting multiple generalized change-points by isolating single ones', *Metrika* **85**(2), 141–174. MR4371188

[3] Anderson, C. W. (1970), 'Extreme value theory for a class of discrete distributions with applications to some stochastic processes', *Journal of Applied Probability* **7**(1), 99–113. MR0256441

[4] 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. MR2498730

[5] Aue, A., Horvath, L., Husková, M. and Kokoszka, P. (2008), 'Testing for changes in polynomial regression', *Bernoulli* **14**(3), 637–660. MR2537806

[6] 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.

[7] Bai, J. and Perron, P. (1998), 'Estimating and testing linear models with multiple structural changes', *Econometrica* **66**(1), 47–78. MR1616121

[8] Bai, J. and Perron, P. (2003), 'Computation and analysis of multiple structural change models', *Journal of applied econometrics* **18**(1), 1–22.

[9] Baranowski, R., Chen, Y. and Fryzlewicz, P. (2019), 'Narrowest-over-threshold detection of multiple change points and change-point-like features', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **81**(3), 649–672. MR3961502

[10] Berkes, I., Liu, W. and Wu, W. B. (2014), 'Komlós–major–tusnády approximation under dependence', *The Annals of Probability* **42**(2), 794–817. MR3178474

[11] Butt, H.-J., Cappella, B. and Kappl, M. (2005), 'Force measurements with the atomic force microscope: Technique, interpretation and applications', *Surface science reports* **59**(1-6), 1–152.

[12] Carrington, R. and Fearnhead, P. (2023), 'Improving power by conditioning on less in post-selection inference for changepoints', *arXiv preprint arXiv:2301.05636*. MR4835211

[13] 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. MR3606770

[14] Chan, H. P. and Walther, G. (2013), 'Detection with the scan and the average likelihood ratio', *Statistica Sinica* **23**(1), 409–428. MR3076173

[15] 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* pp. 737–744. MR0537642

[16] Chen, Y., Shah, R. and Samworth, R. (2014), 'Discussion of 'multiscale

change point inference'by frick, munk and sieling', *Journal of the Royal Statistical Society: Series B* **76**, 544–546. MR3210728

[17] Cho, H. and Fryzlewicz, P. (2022), 'wcm.gsa', `https://github.com/haeran-cho/wcm.gsa`.

[18] Cho, H. and Fryzlewicz, P. (2024), 'Multiple change point detection under serial dependence: Wild contrast maximisation and gappy schwarz algorithm', *Journal of Time Series Analysis* **45**(3), 479–494. MR4731839

[19] Cho, H. and Kirch, C. (2022), 'Bootstrap confidence intervals for multiple change points based on moving sum procedures', *Computational Statistics & Data Analysis* **175**, 107552. MR4446726

[20] Csörgö, M., Csörgö, M., Horváth, L. et al. (1997), 'Limit theorems in change-point analysis'. MR2743035

[21] Csörgo, M. and Révész, P. (2014), *Strong approximations in probability and statistics*, Academic press. MR0666546

[22] 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.

[23] Dette, H., Eckle, T. and Vetter, M. (2020), 'Multiscale change point detection for dependent data', *Scandinavian Journal of Statistics* **47**(4), 1243–1274. MR4178193

[24] 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. MR1649480

[25] Donoho, D. L. and Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *biometrika* **81**(3), 425–455. MR1311089

[26] Doukhan, P. (2012), *Mixing: properties and examples*, Vol. 85, Springer Science & Business Media. MR1312160

[27] Dumbgen, L. and Spokoiny, V. G. (2001), 'Multiscale testing of qualitative hypotheses', *Annals of Statistics* **29**(1), 124–152. MR1833961

[28] Eichinger, B. and Kirch, C. (2018), 'A mosum procedure for the estimation of multiple random change points', *Bernoulli* **24**(1), 526–564. MR3706768

[29] Enikeeva, F., Munk, A., Pohlmann, M. and Werner, F. (2020), 'Bump detection in the presence of dependency: Does it ease or does it load?', *Bernoulli* (4), 3280–3310. MR4140545

[30] Falk, M. and Reiss, R.-D. (1988), 'Independence of order statistics', *The Annals of Probability* pp. 854–862. MR0929082

[31] 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. MR4124337

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

[33] Fearnhead, P. (2006), 'Exact and efficient bayesian inference for multiple changepoint problems', *Statistics and computing* **16**, 203–213. MR2227396

[34] 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. MR3974878

[35] 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. MR3210728

[36] Fryzlewicz, P. (2014), 'Wild binary segmentation for multiple change-point detection', *Ann. Statist* **42**(6), 2243–2281. MR3269979

[37] Fryzlewicz, P. (2021), 'Robust narrowest significance pursuit: inference for multiple change-points in the median', *arXiv preprint arXiv:2109.02487*. MR4799142

[38] Fryzlewicz, P. (2023), 'Narrowest significance pursuit: inference for multiple change-points in linear models', *Journal of the American Statistical Association* pp. 1–14. MR4766015

[39] Gao, J., Ji, W., Zhang, L., Shao, S., Wang, Y. and Shi, F. (2020), 'Fast piecewise polynomial fitting of time-series data for streaming computing', *IEEE Access* **8**, 43764–43775.

[40] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986), 'Residual variance and residual pattern in nonlinear regression', *Biometrika* **73**(3), 625–633. MR0897854

[41] Hall, P. and Titterington, D. (1992), 'Edge-preserving and peak-preserving smoothing', *Technometrics* **34**(4), 429–440. MR1190262

[42] Hampel, F. R. (1974), 'The influence curve and its role in robust estimation', *Journal of the american statistical association* **69**(346), 383–393. MR0362657

[43] Hušková, M. and Slabỳ, A. (2001), 'Permutation tests for multiple changes', *Kybernetika* **37**(5), 605–622. MR1877077

[44] 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. MR4320676

[45] 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. MR1436560

[46] Jarusková, D. (1999), 'Testing appearance of polynomial trend', *Extremes* **2**, 25–37. MR1772398

[47] 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. MR4494153

[48] Jula Vanegas, L., Behr, M. and Munk, A. (2022), 'Multiscale quantile segmentation', *Journal of the American Statistical Association* **117**(539), 1384–1397. MR4480719

[49] Kabluchko, Z. (2007), 'Extreme-value analysis of standardized gaussian increments', *arXiv preprint arXiv:0706.1849*.

[50] 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. MR3217426

[51] 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. MR3036418

[52] Kim, J., Oh, H.-S. and Cho, H. (2022), 'Moving sum procedure for change point detection under piecewise linearity', *arXiv preprint arXiv:2208.04900*. MR4783223

[53] Kimber, A. (1983), 'A note on poisson maxima', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **63**(4), 551–552. MR0705624

[54] Kirch, C. and Klein, P. (2023), 'Moving sum data segmentation for stochastics processes based on invariance', *Statistica Sinica* **33**, 873–892. MR4575326

[55] 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. MR0375412

[56] 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. MR4565454

[57] 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.

[58] 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. MR0602377

[59] Leadbetter, M. R., Lindgren, G. and Rootzén, H. (2012), *Extremes and related properties of random sequences and processes*, Springer Science & Business Media. MR0691492

[60] Liehrmann, A. and Rigaill, G. (2023), 'Ms. fpop: An exact and fast segmentation algorithm with a multiscale penalty', *arXiv preprint arXiv:2303.08723*.

[61] 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. MR3851984

[62] 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.

[63] 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. MR0474645

[64] Maeng, H. and Fryzlewicz, P. (2023), 'Detecting linear trend changes in data sequences', *Statistical Papers* pp. 1–31. MR4757084

[65] McGonigle, E. T. and Cho, H. (2023), 'Robust multiscale estimation of time-average variance for time series segmentation', *Computational Statistics & Data Analysis* **179**, 107648. MR4503388

[66] McZgee, V. E. and Carleton, W. T. (1970), 'Piecewise regression', *Journal of the American Statistical Association* **65**(331), 1109–1124.

[67] Mehrizi, R. V. and Chenouri, S. (2021), 'Valid post-detection inference for change points identified using trend filtering', *arXiv preprint arXiv:2104.12022*.

[68] 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.

[69] Mildenberger, T. (2008), 'A geometric interpretation of the multiresolution criterion in nonparametric regression', *Journal of Nonparametric Statistics* **20**(7), 599–609. MR2454614

[70] Muller, H.-G. (1992), 'Change-points in nonparametric regression analysis', *The Annals of Statistics* **20**(2), 737–761. MR1165590

[71] 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. MR2969913

[72] 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. MR3689315

[73] Philipp, W., Stout, W. F. and Stout, W. (1975), *Almost sure invariance principles for partial sums of weakly dependent random variables*, Vol. 161, American Mathematical Soc. MR0433597

[74] 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. MR4576243

[75] Piterbarg, V. (2015), 'Twenty lectures about gaussian processes', *Atlantic Financial, London*.

[76] Qiu, P. and Yandell, B. (1998), 'Local polynomial jump-detection algorithm in nonparametric regression', *Technometrics* **40**(2), 141–152. MR1626927

[77] Račkauskas, A. and Suquet, C. (2003), 'Invariance principle under self-normalization for nonidentically distributed random variables', *Acta Applicandae Mathematica* **79**, 83–103. MR2021879

[78] Raimondo, M. (1998), 'Minimax estimation of sharp change points', *Annals of statistics* pp. 1379–1397. MR1647673

[79] Rice, J. (1984), 'Bandwidth choice for nonparametric regression', *The Annals of Statistics* **12**(4), 1215–1230. MR0760684

[80] 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. MR2913792

[81] 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. MR4528495

[82] 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. MR4433268

[83] 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.

[84] Tsybakov, A. B. (2004), 'Introduction to nonparametric estimation, 2009', *URL* *https://doi.org/10.1007/b13794. Revised and extended from the* **9**(10). MR2013911

[85] Verzelen, N., Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2020), 'Optimal change-point detection and localization', *Ann. Statist (to appear)*. MR4658569

[86] 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. MR4091859

[87] Wu, W. B. (2005), 'Nonlinear system theory: Another look at dependence', *Proceedings of the National Academy of Sciences* **102**(40), 14150–14154. MR2172215

[88] Wu, W. B. (2009), 'Recursive estimation of time-average variance constants', *The Annals of Applied Probability* **19**(4), 1529–1552. MR2538079

[89] 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. MR2323759

[90] Yu, Y. (2020), 'A review on minimax rates in change point detection and localisation', *arXiv preprint arXiv:2011.01857*.

[91] 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. MR4396490