Detecting multiple generalized change-points by isolating single ones

Andreas Anastasiou and Piotr Fryzlewicz*
Department of Statistics, The London School of Economics and Political Science

March 2, 2018

Abstract

We introduce a new approach, called Isolate-Detect (ID), for the consistent estimation of the number and location of multiple generalized change-points in noisy data sequences. Examples of signal changes that ID can deal with, are changes in the mean of a piecewise-constant signal and changes in the trend, accompanied by discontinuities or not, in the piecewise-linear model. The number of change-points can increase with the sample size. Our method is based on an isolation technique, which prevents the consideration of intervals that contain more than one change-point. This isolation enhances ID's accuracy as it allows for detection in the presence of frequent changes of possibly small magnitudes. Thresholding and model selection through an information criterion are the two stopping rules described in the article. A hybrid of both criteria leads to a general method with very good practical performance and minimal parameter choice. In the scenarios tested, ID is at least as accurate as the state-of-the-art methods; most of the times it outperforms them. The R package (name removed for anonymity) implementing the method from the paper is available from CRAN.

Keywords: Segmentation; symmetric interval expansion; threshold criterion; Schwarz information criterion.

*Work supported by *the Engineering and Physical Sciences Research Council grant no.EP/L014246/1.*

1 Introduction

Change-point detection is an active area of statistical research that has attracted a lot of interest in recent years. According to the National Research Council (2013), detecting changes in a data sequence in order to extract information about the underlying signal will continue to play an essential role in the development of the mathematical sciences. A non-exhaustive list of application areas includes financial econometrics (Bai and Perron, 2003; Schröder and Fryzlewicz, 2013); credit scoring (Bolton and Hand, 2002; Curry et al., 2007); bioinformatics (Futschik et al., 2014; Muggeo and Adelfio, 2011; Olshen et al., 2004) and cyber security (Siris and Papagalou, 2006; Tartakovsky et al., 2006). Our work's focus is on *a posteriori* change-point detection, where the aim is to estimate the number and locations of certain changes in the behavior of the data. We work in the model

$$X_t = f_t + \sigma\epsilon_t, \quad t = 1, 2, \ldots, T, \tag{1.1}$$

where X_t are the observed data and f_t is a one-dimensional, deterministic signal with structural changes at certain points. Two examples are: change-points in the level when f_t is seen as piecewise-constant, and change-points in the first derivative when f_t is piecewise-linear with or without the continuity constraint. We highlight, however, that our methodology and analysis apply to more general scenarios, for instance the detection of knots in a piecewise polynomial signal of order k, where k is not necessarily equal to zero (piecewise-constant mean) or one (piecewise-linear mean). The number N of change-points as well as their locations $r_1, r_2, \ldots, r_N$ are unknown and our aim is to estimate them. In addition, N can grow with T. The random variables ϵ_t in (1.1) have mean zero and variance one; further assumptions will be given later.

When f_t is assumed to be piecewise-constant, the existing change-point detection techniques are mainly split into two categories based on whether the change-points are detected all at once or one at a time. The former category mainly includes optimization-based methods, in which the estimated signal is chosen based on its least squares or log-likelihood fit to the data, penalized by a complexity rule in order to avoid overfitting. The most common example of a penalty function is the Schwarz Information Criterion (SIC); see Yao (1988) for details. To solve the implied penalization problem, dynamic programming approaches, such as the Segment Neighborhood (SN) and Optimal Partitioning (OP) methods of Auger and Lawrence (1989) and Jackson et al. (2005), have been developed. In an attempt to improve on OP's computational cost, Killick et al. (2012) introduce the PELT method, based on a pruning step applied to OP's dynamic programming approach. A non-parametric adaptation of PELT is given in Haynes et al. (2017). Rigaill (2015) introduces an improvement over classical SN algorithms, through a pruning approach called PDPa, while Maidstone et al. (2017b) give two algorithms by combining ideas from PELT and PDPa.

In the latter category, in which change-points are detected one at a time, a popular method is binary segmentation, which performs an iterative binary splitting of the data on intervals determined by the previously obtained splits. Vostrikova (1981) introduces and proves the validity of binary segmentation in the setting of change-point detection for piecewise-constant signals. Among others, binary segmentation is used for change-point detection in Chen and Gupta (1997), Yang and Swartz (2005), Fryzlewicz and Subba Rao (2014), and Badagián et al. (2015). The main advantages of binary segmentation are its conceptual simplicity and low computational cost. However, at each step of the algorithm, binary segmentation looks for a single change-point, which leads to its suboptimality in terms of accuracy, especially for signals with frequent change-points. Some variants of binary segmentation that work towards solving this issue are the Circular Binary Segmentation (CBS) of Olshen et al. (2004), the Wild Binary Segmentation (WBS) of Fryzlewicz (2014) and the Narrowest-Over-Threshold (NOT) method of Baranowski et al. (2018).

CBS searches for at most two change-points at each step of the segmentation algorithm. Instead of initially calculating the contrast value for the whole data sequence, WBS and NOT are based on a random draw of subintervals of the domain of the data, on which an appropriate statistic is tested against a threshold. Apart from binary-segmentation-related approaches, this category also includes methods that control the False Detection Rate (FDR). For instance, in Li et al. (2016), the FDRSeg method is introduced as a combination of FDR-control and global segmentation methods in a multiscale way. The "pseudo-sequential" (PS) procedure of Venkatraman (1992), as well as the CPM method of Ross (2015) are based on a complete embodiment of online detection algorithms to a posteriori situations and work by bounding the Type I error rate of falsely detecting change-points. Some methods do not fall in either category. For example, the tail-greedy algorithm in Fryzlewicz (2018) achieves a multiscale decomposition of the data using Unbalanced Haar wavelets in an agglomerative way.

Going beyond the detection of change-points in piecewise-constant signals, the literature becomes poorer, which is surprising given the importance of change-point detection in general frameworks. For example, the detection of changes in the first derivative of a continuous piecewise-linear signal is used in tracking the health progress of patients for variables such as the heart rate, electroencephalogram, and electrocardiogram (Aminikhanghahi and Cook, 2017). Other examples come from estimating trends for banks' monetary statistics (Bianchi et al., 1999), climate change (Robbins et al., 2011), and time series of AIDS cases (Stasinopoulos and Rigby, 1992). For change-point detection in a continuous piecewise-linear signal, the principle of minimizing the residual sum of squares taking into account a penalty, has been used in Bai and Perron (1998), in the trend filtering (TF) approach (Kim et al., 2009; Tibshirani, 2014), as well as in the dynamic programming algorithm CPOP (Maidstone et al., 2017a). The NOT approach of Baranowski et al. (2018) has been shown to lead to consistent estimation of change-points in different scenarios, such as piecewise-linear mean signals.

Our proposed approach, labeled Isolate-Detect (ID), is a generic technique for generalized change-point detection in various different structures, such as piecewise-constant or piecewise-linear signals with or without the continuity constraint. In the paper we focus on piecewise-constant and continuous piecewise-linear signals. The concept behind ID is simple and is split into two stages; firstly, the attempted isolation of each of the true change-points within subintervals of the domain $[1, 2, \ldots, T]$, and secondly their detection. The terms *subinterval* and *interval* will be used interchangeably. Although a detailed explanation of our methodology is provided in Section 2.1, the basic idea is that for an observed data sequence of length T and with λ_T a suitably chosen positive constant, ID first creates two ordered sets of $K = \lceil T/\lambda_T \rceil$ right- and left-expanding intervals as follows. The j^{th} right-expanding interval is $R_j = [1, j\lambda_T]$, while the j^{th} left-expanding interval is $L_j = [T - j\lambda_T + 1, T]$. We collect these intervals in the ordered set $S_{RL} = \{R_1, L_1, R_2, L_2, \ldots, R_K, L_K\}$. For a suitably chosen contrast function, ID identifies the point with the maximum contrast value in R_1. If its value exceeds a certain threshold, denoted by ζ_T, then it is taken as a change-point. If not, then the process tests the next interval in S_{RL}. Upon detection, the algorithm makes a new start from the end-point (or start-point) of the right- (or left-) expanding interval where the detection occurred. Upon correct choice of ζ_T, ID ensures that we work on intervals with at most one change-point.

The NOT and WBS methods also include localization ideas; however, the nature of localization in ID means that it is of an order of magnitude faster than these two. Furthermore, in NOT and WBS, it is not certain that the intervals drawn will cover the whole data domain without ignoring areas that include change-points. This is an issue of fundamental importance, especially in signals with a large number of change-points, in which

NOT and WBS need to increase the number M of intervals drawn. However, doing this also increases the computational cost. In contrast, due to its interval expansion approach, ID will certainly examine all possible change-point locations. No choice of M is required, which leads to better practical performance with more predictable execution times.

The paper is organized as follows. Section 2 gives a formal explanation of the ID methodology along with two different scenarios of use and the associated theory. In Section 3, we first discuss the computational aspects of ID and the choice of parameter values. ID variants, which lead to improved practical performance, are also explained. In Section 4, we provide a thorough simulation study to compare ID with state-of-the-art methods. Real-life data examples are provided in Section 5.

2 Methodology and Theory

2.1 Methodology

The model is given in (1.1) and the unknown number, N, of change-points can possibly grow with T. Let $r_0 = 0$ and $r_{N+1} = T$ and let $\delta_T = \min_{j=1,2,\ldots,N+1} |r_j - r_{j-1}|$. For clarity of exposition, we start with a simple example before providing a more thorough explanation of how ID works. Figure 2.1 covers a specific case of two change-points, $r_1 = 38$ and $r_2 = 77$. We will be referring to Phases 1 and 2 involving six and four intervals, respectively. These are clearly indicated in the figure and they are only related to this specific example, as for cases with more change-points we would have more such phases. At the beginning, $s = 1$, $e = T = 100$, and we take $\lambda_T = 10$ (how to choose λ_T will be described in Section 3.2). Suppose the threshold ζ_T has been chosen well enough (more details in Section 3.2) so that r_2 gets detected in $\{X_{s^*}, X_{s^*+1}, \ldots, X_e\}$, where $s^* = 71$. After the detection, e is updated as the start-point of the interval where the detection occurred; therefore, $e = 71$. In Phase

2 indicated in the figure, ID is applied in $[s, e] = [1, 71]$. Intervals 1, 3 and 5 of Phase 1 will not be re-examined in Phase 2 and r_1 gets, upon a good choice of ζ_T, detected in $\{X_s, X_{s+1}, \dots, X_{e^*}\}$, where $e^* = 40$. After the detection, s is updated as the end-point of the interval where the detection occurred; therefore, $s = 40$. Our method is then applied in $[s, e] = [40, 71]$; supposing there is no interval $[s^*, e^*] \subseteq [40, 71]$ on which the contrast function value exceeds ζ_T, the process will terminate.

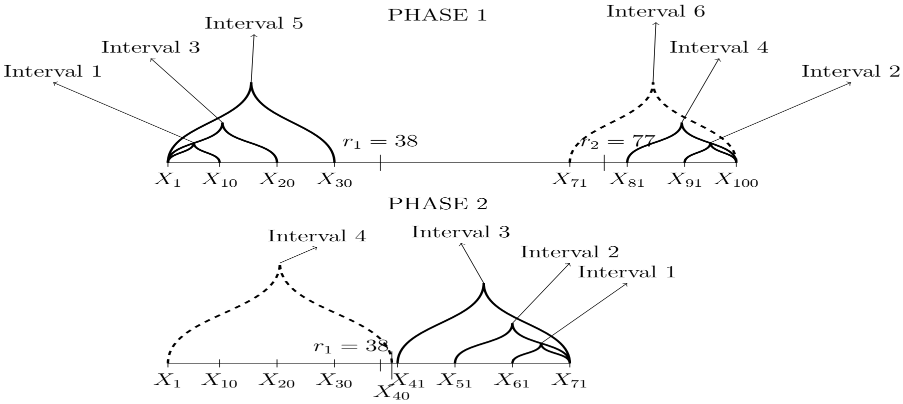

Figure 2.1: An example with two change-points; $r_1 = 38$ and $r_2 = 77$. The dashed line is the interval in which the detection took place in each phase.

We now describe ID more generically. For each change-point, r_j, ID works in two stages: Firstly, isolating r_j in an interval that hopefully contains no other change-point, and secondly detecting r_j through the use of a suitably chosen contrast function, which is denoted by $C^b_{s,e}(\boldsymbol{X})$, for every integer triple (s, e, b), with $1 \leq s \leq b < e \leq T$. Heuristically, the value of $C^b_{s,e}(\boldsymbol{X})$ is relatively small if b is not a change-point. For instance, in piecewise-constant signals, the contrast function reduces to the absolute value of the CUSUM statistic defined in (2.3), while for the case of continuous, piecewise-linear signals, the contrast function is given in Section 2.2.2.

To achieve isolation we employ the idea of interval expansion in the following sense: For

$\lambda_T < \delta_T$ and with $K = \lceil T/\lambda_T \rceil$, let $c_j^r = j\lambda_T$ and $c_j^l = T - j\lambda_T + 1$ for $j = 1, 2, \dots, K-1$, while $c_K^r = T$ and $c_K^l = 1$. For a generic interval $[s, e]$, let us define the sequences

$$\mathrm{I}^r_{s,e} = \left[c^r_{k_1}, c^r_{k_1+1}, \dots, e\right], \quad \mathrm{I}^l_{s,e} = \left[c^l_{k_2}, c^l_{k_2+1}, \dots, s\right], \tag{2.1}$$

where $k_1 := \mathrm{argmin}_{j\in\{1,2\dots,K\}} \{j\lambda_T > s\}$ and $k_2 := \mathrm{argmin}_{j\in\{1,2\dots,K\}} \{T - j\lambda_T + 1 < e\}$. At the beginning, $[s, e] = [1, T]$ and we have that $k_1 = k_2 = 1$. ID starts by looking for change-points interchangeably in *right-* and *left-expanding* intervals, denoted by $[s_j^*, e_j^*]$. For example, the first four intervals are $[s_1^*, e_1^*] = [s, c^r_{k_1}]$, $[s_2^*, e_2^*] = [c^l_{k_2}, e]$, $[s_3^*, e_3^*] = [s, c^r_{k_1+1}]$, $[s_4^*, e_4^*] = [c^l_{k_2+1}, e]$. In each interval $[s_j^*, e_j^*]$, the maximum with respect to b of $C^b_{s,e}(\boldsymbol{X})$ will be tested against a threshold ζ_T. At some point in this interval expansion process, there will be $\left[s^*_{\tilde{k}}, e^*_{\tilde{k}}\right]$, with $\tilde{k} \in \{1, 2, \dots, 2K\}$, which contains only one change-point; this being either r_1 or r_N, depending on whether r_1 is closer to s, or r_N is closer to e. The interval $\left[s^*_{\tilde{k}}, e^*_{\tilde{k}}\right]$ will not contain any other change-points, due to the fact that at each step we expand the intervals by the quantity λ_T that is smaller than the minimum distance between two change-points. The practical choice of λ_T is described in Section 3.2. W.l.o.g. assume that $r_1 - s \leq T - r_N$. In this case, $\left[s^*_{\tilde{k}}, e^*_{\tilde{k}}\right] = [1, c^r_{k^*}]$, where $k^* = k_1 + \left(\tilde{k} - 1\right)/2$ with k_1 as in (2.1). For $b_1 = \mathrm{argmax}_{1\leq t < c^r_{k^*}} C^t_{1,c^r_{k^*}}(\boldsymbol{X})$, if $C^{b_1}_{1,c^r_{k^*}}(\boldsymbol{X}) > \zeta_T$, then b_1 is assigned as the estimate of r_1. After that, ID restarts and looks for change-points in $[s, e] = [c^r_{k^*}, T]$, where now s_j^* and e_j^* are taken from $\mathrm{I}^l_{c^r_{k^*},T}$ and $\mathrm{I}^r_{c^r_{k^*},T}$ as in (2.1), respectively. The algorithm will stop when it is applied in an interval $[s, e]$, such that all expanding intervals $[s_j^*, e_j^*] \subseteq [s, e]$ do not include a point b_j^* with $C^{b_j^*}_{s_j^*,e_j^*}(\boldsymbol{X}) > \zeta_T$.

The idea of a-posteriori change-point detection in which change-points are detected sequentially, has appeared previously in the literature; see for instance the PS and CPM methods of Venkatraman (1992) and Ross (2015), respectively. ID is conceptually and in practice different from these methods in a number of ways related to the threshold choice, the construction of the estimated change-point locations as well as the way PS and CPM restart upon detection. Furthermore, ID's isolation technique does not appear in CPM. A comparison between the performance of ID and that of state-of-the-art methods (including CPM) is given in Section 4.

2.2 Theoretical behavior of ID

We work under the assumption

(A1) The random sequence $\{\epsilon_t\}_{t=1,2,\ldots,T}$ is independent and identically distributed (i.i.d.) from the normal distribution with mean zero and variance one.

The assumption of i.i.d. normal random variables in (A1) is made for technical convenience; Section 3.5 shows how to use ID under non-Gaussianity. The assumption that $\sigma = 1$ is not restrictive. If σ is unknown, then we need to estimate it. In the cases of piecewise-constant and piecewise-linear signals, σ can be estimated via the Median Absolute Deviation method proposed in Hampel (1974). For simplicity purposes, let $\sigma = 1$, and (1.1) becomes

$$X_t = f_t + \epsilon_t, \quad t = 1, 2, \ldots, T. \tag{2.2}$$

With $r_0 = 0$ and $r_{N+1} = T$, and for $j = 1, 2, \ldots, N+1$, we examine the theoretical behavior of ID in the following two illustration cases:

Piecewise-constant signals: $f_t = \mu_j$ for $t = r_{j-1}+1, r_{j-1}+2, \ldots, r_j$, and $f_{r_j} \neq f_{r_j+1}$.

Continuous, piecewise-linear signals: $f_t = \mu_{j,1} + \mu_{j,2}t$, for $t = r_{j-1}+1, r_{j-1}+2, \ldots, r_j$ with the additional constraint of $\mu_{k,1} + \mu_{k,2}r_k = \mu_{k+1,1} + \mu_{k+1,2}r_k$ for $k = 1, 2, \ldots, N$. The change-points, r_k, satisfy $f_{r_k-1} + f_{r_k+1} \neq f_{r_k}$.

The above scenarios are only examples in which the ID methodology can be applied. The isolation aspect of the method allows its application to various different cases, such as the estimation of the number and the position of knots (either continuous or not) in piecewise polynomial functions. The latter problem is considered for example in Tibshirani (2014), where the Trend Filtering (TF) approach of Kim et al. (2009) is used and compared to smoothing and locally adaptive regression splines.

2.2.1 Piecewise-constant mean signals

Under piecewise-constancy, the contrast function used is the absolute value of the CUSUM statistic, the latter being

$$\tilde{X}_{s,e}^{b} = \sqrt{\frac{e-b}{n(b-s+1)}}\sum_{t=s}^{b} X_t - \sqrt{\frac{b-s+1}{n(e-b)}}\sum_{t=b+1}^{e} X_t, \tag{2.3}$$

where $1 \leq s \leq b < e \leq T$ and $n = e - s + 1$. Under assumption (A1), it can be shown that $\text{argmax}_b \left|\tilde{X}_{s,e}^{b}\right| = \text{argmax}_b \mathcal{R}_{s,e}^{b}(\boldsymbol{X})$, where $\mathcal{R}_{s,e}^{b}(\boldsymbol{X})$ is the generalized log-likelihood ratio statistic for all potential single change-points within $[s,e]$. For i.i.d. Gaussian noise, the CUSUM estimator of the change-point locations is the maximum likelihood estimator. For the main result of Theorem 2.1, we also make the assumption

(A2) The minimum distance, δ_T, between two change-points and the minimum magnitude of jumps, $\underline{f}_T$, are connected by $\sqrt{\delta_T}\underline{f}_T \geq \underline{C}\sqrt{\log T}$, for a large enough constant $\underline{C}$.

We explain after Theorem 2.1 that the above $\sqrt{\log T}$ rate for the lower bound of $\sqrt{\delta_T}\underline{f}_T$ is optimal up to a double logarithmic term. The number of change-points, N, is assumed to be neither known nor fixed. It can grow with T and the only indirect assumption on N is due to the minimum distance, δ_T, between two change-points in the sense that $N + 1 \leq T/\delta_T$. Below, we give the main theoretical result for the consistency of the number and location of the estimated change-points. The proof is in Section 4 of the online supplement.

Theorem 2.1. *Let $\{X_t\}_{t=1,2,\ldots,T}$ follow model (2.2), with f_t being a piecewise-constant signal and assume that (A1) and (A2) hold. Let N and $r_j, j = 1,2,\ldots,N$ be the number* *and locations of the change-points, while $\hat{N}$ and $\hat{r}_j, j = 1, 2, \ldots, \hat{N}$ are their estimates, with $\hat{r}_j$ sorted in increasing order. In addition, $\Delta_j^f = \left|f_{r_j+1} - f_{r_j}\right|$, $j = 1, 2, \ldots, N$. Then, there exist positive constants C_1, C_2, C_3, C_4, which do not depend on T, such that for $C_1\sqrt{\log T} \leq \zeta_T < C_2\sqrt{\delta_T}\underline{f}_T$ and for a sufficiently large T, we obtain*

$$\mathbb{P}\left(\hat{N} = N, \max_{j=1,2,\ldots,N}\left(|\hat{r}_j - r_j|\left(\Delta_j^f\right)^2\right) \leq C_3 \log T\right) \geq 1 - \frac{C_4}{T}. \tag{2.4}$$

From (2.4), we notice that in order to be able to match the estimated change-point locations with the true ones, δ_T should be at least as large as $\max_{j=1,2,\ldots,N}|\hat{r}_j - r_j|$, meaning that δ_T must be at least $\mathcal{O}(\log T)$. For this order of δ_T, Chan and Walther (2013) argue that the smallest possible $\sqrt{\delta_T}\underline{f}_T$ that allows change-point detection is $\mathcal{O}\left(\sqrt{\log T - \log(\log T)}\right)$. In our case, assumption (A2) ensures that the $\sqrt{\log T}$ rate is attained, which is optimal up to the double logarithmic term. This provides evidence that ID allows for detection in complex scenarios, such as limited spacings between change-points. The quantity on the right-hand side of (2.4) is $\mathcal{O}(1 - 1/T)$; the same order as in WBS and NOT. However, ID gives a significantly lower constant C_4 for the bound; see the proof in the online supplement for more details. The rate of the lower bound for the threshold is $\mathcal{O}\left(\sqrt{\log T}\right)$ and this is what will be used in practice as the default rate: we use

$$\zeta_T = C\sqrt{2\log T} \tag{2.5}$$

and the choice of the constant C will be explained in Section 3. Furthermore, (2.4) indicates that δ_T does not affect the rate of convergence of the estimated change-point locations; these only depend on Δ_j^f.

2.2.2 Continuous piecewise-linear mean signals

Under Gaussianity and with $R_{s,e}^b(\boldsymbol{X})$ being the generalized log-likelihood ratio for all possible single change-points within $[s, e)$, the idea is to find a contrast function $C_{s,e}^b(\boldsymbol{X})$, which is maximized at the same point as $R^b_{s,e}(\boldsymbol{X})$. The contrast function is constructed by taking inner products of the data with a contrast vector. In the case of continuous piecewise-linear signals, it has been shown in Baranowski et al. (2018) that the appropriate contrast vector is $\boldsymbol{\psi^b_{s,e}} = \left(\psi^b_{s,e}(1), \psi^b_{s,e}(2), \ldots, \psi^b_{s,e}(T)\right)$, where

$$\psi^b_{s,e}(t) = \begin{cases} \alpha^b_{s,e}\beta^b_{s,e}\left[(e+2b-3s+2)t - (be+bs-2s^2+2s)\right], & t = s, s+1, \ldots, b, \\ -\frac{\alpha^b_{s,e}}{\beta^b_{s,e}}\left[(3e-2b-s+2)t - (2e^2+2e-be-bs)\right], & t = b+1, b+2, \ldots, e, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

where $n = e-s+1$, $\alpha^b_{s,e} = (6/[n(n^2-1)(1+(e-b+1)(b-s+1)+(e-b)(b-s))])^{\frac{1}{2}}$ and $\beta^b_{s,e} = ([(e-b+1)(e-b)]/[(b-s+1)(b-s)])^{\frac{1}{2}}$. The contrast function is $C^b_{s,e}(\boldsymbol{X}) = \left|\left\langle \boldsymbol{X}, \boldsymbol{\psi^b_{s,e}}\right\rangle\right|$. In order to explain the reasoning behind the choice of the triangular function $\psi^b_{s,e}(\cdot)$, we define, for the interval $[s,e]$, the linear vector

$$\gamma_{s,e}(t) = \begin{cases} \left(\frac{1}{12}(e-s+1)\left(e^2-2es+2e+s^2-2s\right)\right)^{-\frac{1}{2}}\left(t - \frac{e+s}{2}\right), & t = s, s+1, \ldots, e, \\ 0, & \text{otherwise}, \end{cases}$$

as well as the constant vector

$$1_{s,e}(t) = \begin{cases} (e-s+1)^{-\frac{1}{2}}, & t = s, s+1, \ldots, e, \\ 0, & \text{otherwise}. \end{cases}$$

Now on the vector

$$\tilde{\psi}^b_{s,e}(t) = \begin{cases} (t-b)^{-\frac{1}{2}}, & t = b+1, b+2, \ldots, e, \\ 0, & \text{otherwise}, \end{cases}$$

which is linear with a kink at $b+1$, we apply the Gram-Schmidt orthogonalization with respect to $\gamma_{s,e}(t)$ and $1_{s,e}(t)$. Normalizing the obtained vector such that $\|.\|_2 = 1$, returns the contrast vector $\psi_{s,e}^b(t)$ defined in (2.6). The best approximation, in terms of the Euclidean distance, of X_t in $[s,e]$ is a linear combination of $\gamma_{s,e}(t)$, $1_{s,e}(t)$, and $\psi_{s,e}(t)$, which are mutually orthonormal. This orthonormality leads to $R_{s,e}^b(\boldsymbol{X}) = \left|\left\langle \boldsymbol{X}, \boldsymbol{\psi_{s,e}^b} \right\rangle\right| = C_{s,e}^b(\boldsymbol{X})$ (Baranowski et al., 2018). For the consistency of ID in continuous piecewise-linear signals, we make the following assumption,

(A3) The minimum distance, δ_T, between two change-points and the minimum magnitude of jumps, $\underline{f}_T = \min_{j=1,2,\dots,N} \left|2f_{r_j} - f_{r_j+1} - f_{r_j-1}\right|$, are connected by the requirement $\delta_T^{3/2}\underline{f}_T \geq C^*\sqrt{\log T}$, for a large enough constant C^*.

The term $\delta_T^{3/2}\underline{f}_T$ characterizes the difficulty level of the detection problem and is analogous to $\sqrt{\delta_T}\underline{f}_T$ in the scenario of piecewise-constant signals. The following theorem gives the consistency result under the case of continuous piecewise-linear signals. The proof is in Section 4 of the online supplement.

Theorem 2.2. *Let $\{X_t\}_{t=1,2,\dots,T}$ follow model (2.2) with f_t being a continuous, piecewise-linear signal and assume that (A1) and (A3) hold. We denote by N and $r_j, j = 1,2,\dots,N$ the number and locations of the change-points, while $\hat{N}$ and $\hat{r}_j, j = 1,2,\dots,\hat{N}$ are their estimates, with $\hat{r}_j$ sorted in increasing order. Also, we denote by $\Delta_j^f = \left|2f_{r_j} - f_{r_j+1} - f_{r_j-1}\right|$. Then, there exist positive constants C_1, C_2, C_3, C_4, which do not depend on T, such that for $C_1\sqrt{\log T} \leq \zeta_T < C_2\delta_T^{3/2}\underline{f}_T$ and for sufficiently large T,*

$$\mathbb{P}\left(\hat{N} = N, \max_{j=1,2,\dots,N}\left(|\hat{r}_j - r_j|\left(\Delta_j^f\right)^{2/3}\right) \leq C_3(\log T)^{1/3}\right) \geq 1 - \frac{C_4}{T}. \tag{2.7}$$

The quantity on the right-hand side of (2.7) is $\mathcal{O}\left(1 - 1/T\right)$. In addition, in the common case of $\underline{f}_T \sim T^{-1}$, ID's change-point detection rate is $\mathcal{O}\left(T^{2/3}\left(\log T\right)^{1/3}\right)$, as can be seen from (2.7). This differs from the $\mathcal{O}\left(T^{2/3}\right)$ rate derived in Raimondo (1998) only by a logarithmic factor. The lower bound of the threshold is $\mathcal{O}\left(\sqrt{\log T}\right)$ and therefore,

$$\zeta_T = \tilde{C}\sqrt{2\log T}, \tag{2.8}$$

where $\tilde{C}$ is a constant and we will comment on its choice in Section 3.2.

ID exhibits a great degree of flexibility as it does not depend on the structure of the signal; what changes is the choice of an appropriate contrast function. Adapting a similar approach as the one for the case of continuous piecewise-linear signals, one can construct contrast functions for the detection of other types of features.

2.3 Information Criterion approach

Misspecification of the threshold in the ID algorithm can lead to the misestimation of the number of change-points. To solve this, we develop an approach which starts by possibly overestimating the number of change-points and then creates a solution path, with the estimates ordered according to a certain predefined criterion. The best fit is then chosen, based on the optimization of a model selection criterion.

The solution path algorithm: The estimated number of change-points depends on ζ_T and this allows us to denote $\hat{N} = \hat{N}(\zeta_T)$. For given data, we employ ID using first ζ_T and then $\tilde{\zeta}_T$, where $\tilde{\zeta}_T < \zeta_T$. Let $C_{\tilde{\zeta}_T}$ and $\tilde{C}_{\tilde{\zeta}_T}$ be the $\tilde{\zeta}_T$-associated constants in (2.5) and (2.8), respectively. With $J \geq \hat{N}(\zeta_T)$, we estimate $\tilde{r}_j, j = 1, 2, \ldots, J$, which are sorted in increasing order in $\tilde{S} = [\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_J]$. The algorithm is split into four parts. Although the description of each part is fairly technical, we note that the different parts are very similar and are based on the idea of removing change-points according to their contrast function values as well as their distance to neighboring estimates. In the algorithm we refer to three parameters: $C^*, \tilde{\tilde{C}}$, and α. Although we do not give a recipe for the choice of C^* and $\tilde{\tilde{C}}$, Section 3.4 describes how to circumvent their choice. The default value of α is 1.01. An explanation of this choice is given before Theorem 2.3. We now give the four parts of the solution path algorithm. All events below occur with probability tending to one with T.

Part 1: With C^* a positive constant, the aim is to prune the estimates in $\tilde{S}$, such that, for each true change-point, there are at most four and at least one estimated change-point within a distance of $C^*(\log T)^\alpha$. To achieve this, $\forall j \in \{1,2,\ldots,J\}$, we collect triplets $(\tilde{r}_{j-1}, \tilde{r}_j, \tilde{r}_{j+1})$ and we calculate $CS(\tilde{r}_j) := C_{\tilde{r}_{j-1},\tilde{r}_{j+1}}^{\tilde{r}_j}(\boldsymbol{X})$, with $C_{s,e}^b(\boldsymbol{X})$ being the relevant contrast function. For $m = \mathrm{argmin}_j\{CS(\tilde{r}_j)\}$, firstly we check whether $CS(\tilde{r}_m) \leq \tilde{\tilde{C}}\sqrt{\log T}$, for $\tilde{\tilde{C}} > 0$; in the proofs of Theorems 2.3 and 2.4, $\tilde{\tilde{C}} = 2\sqrt{2}$, but smaller values could be sufficient. If $CS(\tilde{r}_m) \leq \tilde{\tilde{C}}\sqrt{\log T}$ and also $\tilde{r}_{j+1} - \tilde{r}_{j-1} \leq 2C^*(\log T)^\alpha$, we remove $\tilde{r}_m$ from $\tilde{S}$, reduce J by 1, relabel the remaining estimates (in increasing order) in $\tilde{S}$, and repeat this estimate removal process. We proceed to Part 2 when $CS(\tilde{r}_m) > \tilde{\tilde{C}}\sqrt{\log T}$.

Part 2: The aim is to continue the pruning process in a way that at the end of Part 2, there is at least one estimate within a distance of $C^*(\log T)^\alpha$ from each true change-point, but also there are at most two estimates between any pair of consecutive true change-points. For the relabeled estimates in $\tilde{S}$ after the completion of Part 1, if $\tilde{r}_j - \tilde{r}_{j-1} \leq C^*(\log T)^\alpha$, then we remove $\tilde{r}_j$, relabel the remaining estimates, and keep removing the estimates until there is no pair $(\tilde{r}_{j-1}, \tilde{r}_j)$, such that $\tilde{r}_j - \tilde{r}_{j-1} \leq C^*(\log T)^\alpha$. We then calculate $CS(\tilde{r}_j)$ as in Part 1 and for $m = \mathrm{argmin}_j\{CS(\tilde{r}_j)\}$, if $CS(\tilde{r}_m) \leq \tilde{\tilde{C}}\sqrt{\log T}$, then we remove $\tilde{r}_m$ and relabel the remaining elements of $\tilde{S}$. This estimates removal process is repeated and we proceed to Part 3 only when $CS(\tilde{r}_m) > \tilde{\tilde{C}}\sqrt{\log T}$.

Part 3: We need to ensure that once $\tilde{S}$ contains N estimates, then for $j = 1,2,\ldots,N$, each $\tilde{r}_j$ is within a distance of $C^*\left(\log T\right)^\alpha$ from r_j. To achieve this, for the remaining estimated change-points after Part 2, we use triplets $(\tilde{s}_j, \tilde{r}_j, \tilde{e}_j)$, with $\tilde{s}_j = \lfloor(\tilde{r}_{j-1} + \tilde{r}_j)/2\rfloor + 1$ and $\tilde{e}_j = \lceil(\tilde{r}_j + \tilde{r}_{j+1})/2\rceil$. For $m = \mathrm{argmin}_j C_{\tilde{s}_j,\tilde{e}_j}^{\tilde{r}_j}(\boldsymbol{X})$, if $C_{\tilde{s}_m,\tilde{e}_m}^{\tilde{r}_m}(\boldsymbol{X}) \leq \tilde{\tilde{C}}\sqrt{\log T}$, then we remove $\tilde{r}_m$ and relabel the remaining estimates in $\tilde{S}$ in increasing order. We repeat this removal procedure until $C_{\tilde{s}_m,\tilde{e}_m}^{\tilde{r}_m}(\boldsymbol{X}) > \tilde{\tilde{C}}\sqrt{\log T}$, which is when we proceed to Part 4.

Part 4: For the estimated change-points that are in $\tilde{S}$ after Part 3 is completed, we use again the triplets $(\tilde{r}_{j-1}, \tilde{r}_j, \tilde{r}_{j+1})$ in order to find $m = \mathrm{argmin}_j\{CS(\tilde{r}_j)\}$ and then remove

$\tilde{r}_m$ from $\tilde{S}$. This estimates removal approach is repeated until $\tilde{S} = \emptyset$.

At the end of Part 4, we collect the estimates in a vector

$$\boldsymbol{b} = (b_1, b_2, \dots, b_J), \tag{2.9}$$

where b_J is the estimate that was removed first, b_{J-1} is the one that was removed second, and so on. From now on, $\boldsymbol{b}$ is called the solution path and is used to give a range of different fits. We define the collection $\{\mathcal{M}_j\}_{j=0,1,\dots,J}$ where $\mathcal{M}_0 = \emptyset$ and $\mathcal{M}_j = \{b_1, b_2, \dots, b_j\}$. For $j = 2, \dots, J$, let $\tilde{b}_1 < \dots < \tilde{b}_j$ be the sorted elements of $\mathcal{M}_j$. Among the collection of models $\{\mathcal{M}_j\}_{j=0,1,\dots,J}$, we propose to select the one that minimizes the strengthened Schwarz Information Criterion (Liu et al. (1997), Fryzlewicz (2014)), defined as

$$\mathrm{sSIC}(j) = -2\sum_{k=1}^{j+1} \ell\left(X_{\tilde{b}_{k-1}+1}, \dots, X_{\tilde{b}_k}; \hat{\theta}_k\right) + n_j \left(\log T\right)^{\alpha}, \tag{2.10}$$

where $\tilde{b}_0 = 0$ and for each collection $\mathcal{M}_j$, $\tilde{b}_{j+1} = T$ and $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{j+1}$ are the maximum likelihood estimators of the segment parameters for the model (2.2) with change-point locations $b_1, b_2, \dots, b_j$. The quantity n_j is the total number of estimated parameters related to $\mathcal{M}_j$. Taking $\alpha = 1$ in (2.10) gives the standard SIC penalty, but our theory requires $\alpha > 1$. In practice we use $\alpha = 1.01$ in order to stay close to SIC. Theorems 2.3 and 2.4 below give the consistency results for the piecewise-constant and continuous piecewise-linear models, based on the sSIC approach. The proof of Theorem 2.3 is in the supplementary material and the same approach can be followed to prove Theorem 2.4.

Theorem 2.3. *Let $\{X_t\}_{t=1,2,\dots,T}$ follow model (2.2) under piecewise-constancy and let the assumptions of Theorem 2.1 hold. Let N and $r_j, j = 1, 2, \dots, N$ be the number and locations of the change-points. Let $N \leq J$, where J is a constant that does not depend on T. In addition, let $\alpha > 1$ be such that $(\log T)^{\alpha} = o(\delta_T \underline{f}_T^2)$ is satisfied, where δ_T and $\underline{f}_T$ are* *defined in (A2). With $\{\mathcal{M}_j\}_{j=0,1,\dots,J}$ being the set of candidate models obtained by the solution path algorithm, we define $\hat{N} = \operatorname{argmin}_{j=0,1,\dots,J} \mathrm{sSIC}(j)$. Then, there exist positive constants C_1, C_2, which do not depend on T, such that for $\Delta_j^f = \left|f_{r_j+1} - f_{r_j}\right|$*

$$\mathbb{P}\left(\hat{N} = N, \max_{j=1,2,\dots,N}\left(|\hat{r}_j - r_j|\left(\Delta_j^f\right)^2\right) \leq C_1\left(\log T\right)^{\alpha}\right) \geq 1 - \frac{C_2}{T}. \tag{2.11}$$

Theorem 2.4. *Let $\{X_t\}_{t=1,2,\dots,T}$ follow model (2.2) under continuous piecewise-linearity and let the assumptions of Theorem 2.2 hold. Let N and $r_j, j = 1, 2, \dots, N$ be the number and locations of the change-points. Let $N \leq J$, where J is a constant that does not depend on T. In addition, let $\alpha > 1$ be such that $(\log T)^{\alpha} = o(\delta_T^3 \underline{f}_T^2)$ is satisfied, where δ_T and $\underline{f}_T$ are defined in (A3). With $\{\mathcal{M}_j\}_{j=0,1,\dots,J}$ being the set of candidate models obtained by the solution path algorithm, we define $\hat{N} = \operatorname{argmin}_{j=0,1,\dots,J} \mathrm{sSIC}(j)$. Then, there exist positive constants C_1, C_2, which do not depend on T, such that for $\Delta_j^f = \left|2f_{r_j} - f_{r_j+1} - f_{r_j-1}\right|$,*

$$\mathbb{P}\left(\hat{N} = N, \max_{j=1,2,\dots,N}\left(|\hat{r}_j - r_j|\left(\Delta_j^f\right)^{2/3}\right) \leq C_1(\log T)^{\alpha/3}\right) \geq 1 - \frac{C_2}{T}. \tag{2.12}$$

The quantities on the right hand sides of (2.11) and (2.12) are $\mathcal{O}\left(1 - 1/T\right)$; the same order as those in (2.4) and (2.7). The lowest admissible $\delta_T \underline{f}_T^2$ and $\delta_T^3 \underline{f}_T^2$ in Theorems 2.3 and 2.4, respectively, are slightly larger than the same quantities in the thresholding approach. This is expected because SIC-based approaches tend to exhibit better practical behavior for signals that have a moderate number of change-points with large spacings between them, see Yao (1988) for more details. A hybrid that combines the advantages of the thresholding and the SIC-based approach is introduced in Section 3.4.

3 Computational complexity and practicalities

3.1 Computational cost

With δ_T being the minimum distance between two change-points, and λ_T the interval-expansion parameter, we need $\lambda_T < \delta_T$. As $K = \lceil T/\lambda_T \rceil > \lceil T/\delta_T \rceil$ and the total number,

$\tilde{M}$, of distinct intervals required to scan the data is no more than $2K$ (K intervals created from each expanding direction), in the worst case scenario we have

$$\tilde{M} = 2K > 2\left\lceil \frac{T}{\delta_T} \right\rceil. \tag{3.1}$$

As a comparison, in WBS and NOT, one needs to draw at least M intervals where $M \geq (9T^2/\delta_T^2)\log(T^2/\delta_T)$. The lower bound for M in WBS and NOT is $\mathcal{O}(T^2/\delta_T^2)$ up to a logarithmic factor, whereas the lower bound in (3.1) is $\mathcal{O}(T/\delta_T)$. This results in great speed gains of ID over WBS and NOT. In our understanding, the reason behind this significant difference in the computational complexity of the methods is that in WBS and NOT both the start- and end-points of the randomly drawn intervals have to be chosen, whereas in ID, depending on the expanding direction, we keep the start- or end-point fixed.

3.2 Parameter choice

Choice of the threshold constant. In order to decide C and $\tilde{C}$ in (2.5) and (2.8), respectively, we ran a large-scale simulation study involving a wide range of signals. The number of change-points, N, was generated from the Poisson distribution with rate parameter $N_\alpha \in \{4, 8, 12\}$. For $T \in \{100, 200, 500, 1000, 2000, 5000\}$, we uniformly distributed the change-points in $\{1, 2, \dots, T\}$. Then, for piecewise-constant (or continuous piecewise-linear) signals, at each change-point location we introduced a jump (or a slope change) which followed the normal distribution with mean zero and variance $\sigma^2 \in \{1, 3, 5\}$. Standard Gaussian noise was then added onto the simulated signal. For each value of N_α, σ^2 and T we generated 1000 replicates and estimated the number of change-points using ID with threshold ζ_T as in (2.5) and (2.8) for a variety of constant values C and $\tilde{C}$, respectively. The best behavior occurred when, approximately, $C = 1$ and $\tilde{C} = 1.4$, for piecewise-constant and continuous piecewise-linear signals, respectively. These values will be called the default constants. In the SIC-based approach of Section 2.3, we started by detecting change-points using threshold $\tilde{\zeta}_T < \zeta_T$. In practical implementations, we take the constants related to $\tilde{\zeta}_T$, namely $C_{\tilde{\zeta}_T}$ and $\tilde{C}_{\tilde{\zeta}_T}$ as defined in Section 2.3, to be 0.9 and 1.25, respectively.

Choice of the expansion parameter λ_T. Due to the low computational complexity of ID, we take $\lambda = 3$, leading to good accuracy even for signals with very frequent change-points. For an example of execution speeds on a single core of an Intel Xeon 3.60GHz CPU with 16 GB of RAM, see Table 3.1.

T	Time (s)
7×10^3	0.31
7×10^4	2.25
7×10^5	26.41
7×10^6	266.72

Table 3.1: The average computational time of ID based on the threshold stopping rule, for a simulated data sequence of length $7 \times 10^j, j = 3, 4, 5, 6$, from a signal that fluctuates between 0 and 4 every 7 observations. The standard deviation is $\sigma = 0.5$.

3.3 Variants

This section describes a number of different ways to further improve ID's practical performance with respect to both accuracy and speed.

Very long signals: If T is large, we split the given data sequence uniformly into smaller parts (windows), to which ID is then applied. In practical implementations, the length of the window is 3000 and we apply this window structure only when $T > 12000$, because for smaller values of T there were not significant differences in the execution times of ID and its window-based variant. An idea of the computational improvement that this structure offers is explained in Section 1 of the supplement.

Restarting after detection: In practice, instead of starting from the end-point e^* (or start-point s^*) of the right-expanding (or left-expanding) interval where a detection occurred, we could start from the estimated change-point, $\hat{b}$. This alternative, labeled ID_{det}, leads to accuracy improvements without affecting the speed of the method.

Faster solution path algorithm: In practice, we use only Part 4 of the solution path algorithm described in Section 2.3 because it is quicker and conceptually simpler; it requires only the choice of α, and tends not to affect the accuracy of ID.

3.4 A hybrid between thresholding and SIC stopping rules

For signals with a large number of regularly occurring change-points, the threshold-based ID tends to behave better than the SIC-based procedure. As explained after Theorems 2.3 and 2.4, this is unsurprising because SIC-based approaches typically perform better on signals with a moderate number of change-points separated by large spacings. This difference in ID's behavior between the threshold- and SIC-based versions is what motivates us to introduce a hybrid of these two stopping rules with minimal parameter choice, which works as follows. Firstly, we find the estimated change-points using the threshold approach ID_{det} with $\lambda_T^{th} = 3$. If the estimated number of change-points is larger than a constant J^*, then the result is accepted and we stop. Otherwise, the hybrid method proceeds to detect the change-points using the SIC-based approach with $\lambda_T > \lambda_T^{th}$, since the already-applied thresholding rule has not suggested a signal with many change-points. In the simulation results later, we use $J^* = 100$ and $\lambda_T = 10$.

3.5 Extension to different noise structures

This section describes how to use ID when the noise is not Gaussian. We pre-process the data in order to obtain a noise structure that is closer to Gaussianity. For a given scale number s and data $\{X_t\}_{t=1,2,\ldots,T}$, let us denote by $Q = \lceil T/s \rceil$ and $\tilde{X}_q = \frac{1}{s}\sum_{t=(q-1)s+1}^{qs} X_t$, for $q = 1, 2, \ldots, Q-1$, while $\tilde{X}_Q = (T-(Q-1)s)^{-1}\sum_{t=(Q-1)s+1}^{T} X_t$. We apply ID on

$\left\{\tilde{X}_q\right\}_{q=1,2,\dots,Q}$, to obtain the estimated change-points, namely $\tilde{\tilde{r}}_1, \tilde{\tilde{r}}_2, \dots, \tilde{\tilde{r}}_{\hat{N}}$ in increasing order. To estimate the original locations of the change-points we define $\hat{r}_k = \left(\tilde{\tilde{r}}_k - 1\right) s + \left\lfloor \frac{s}{2} + 0.5 \right\rfloor$, $k = 1, 2, \dots, \hat{N}$. There is a trade-off in the choice of the scaling parameter s. The larger the value of s, the closer the distribution of the noise to normal, but the more the amount of pre-processing. In simulations presented in Section 4, we use $s = 3$ for the case of Student-t_5 distributed noise, while if the tails are heavier (Student-t_3), we set $s = 5$. The hybrid version of ID will be employed on $\left\{\tilde{X}_q\right\}_{q=1,2,\dots,Q}$ and to be consistent with the choice of the expansion parameter as defined in Section 3.2, we take $\lambda_T^* = \lfloor \lambda_T / s \rfloor$. In practice, for unknown noise, our recommendation is to set $s = 5$.

4 Simulations

In this section, we provide a comprehensive simulation study of the performance of ID against the currently best methods in the scenarios of piecewise-constant signals and continuous piecewise-linear signals. Table 4.2 below shows the competitors used.

Type of signal	Method notation	Reference	R package
Piecewise-constant	PELT	Killick et al. (2012)	**changepoint**
	NP.PELT	Haynes et al. (2017)	**changepoint.np**
	S3IB	Rigaill (2015)	**Segmentor3IsBack**
	CumSeg	Muggeo and Adelfio (2011)	**cumSeg**
	CPM	Ross (2015)	**cpm**
	WBS	Fryzlewicz (2014)	**wbs**
	NOT	Baranowski et al. (2018)	**not**
	FDR	Li et al. (2016)	**FDRSeg**
	TGUH	Fryzlewicz (2018)	**breakfast**
Continuous piecewise-linear	NOT	Baranowski et al. (2018)	**not**
	TF	Kim et al. (2009)	-
	CPOP	Maidstone et al. (2017a)	-

Table 4.2: The competing methods used in the simulation study.

The CPOP method is implemented in http://www.research.lancs.ac.uk/portal/en/datasets/cpop(56c07868-3fe9-4016-ad99-54439ec03b6c).html, and TF in https://stanford.edu/~boyd/l1_tf. For WBS, we give results based on both the information criterion and the thresholding (for $C = 1$) stopping rules. The notation is WBSIC and WBSC1, respectively. In the **cpm** package, the threshold is decided through the average run length (ARL) until a false positive occurs. In our simulations, we give results for ARL $= 500$ (the default value) and if the signal length, l_s, is greater than 500, results are also given for ARL $= 1000\lceil ls/1000\rceil$. The notation is CPM.l.$A$, with A the value of ARL.

We use the ID version of Section 3.4. Although all signals are fully specified in Section 2 of the online supplement, Figure 4.2 shows examples of the data generated by models (M3) *teeth*, (M4) *stairs*, (W1) *wave 1*, and (W2) *wave 2*.

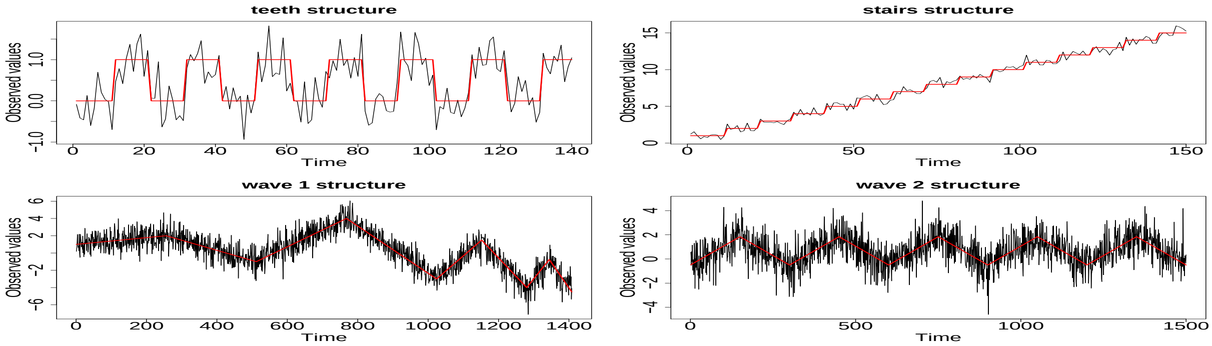

Figure 4.2: Examples of data series, used in simulations. The true signal, f_t, is in red.

Tables 4.3–4.10 summarize the results in the case of i.i.d. Gaussian noise. Table 4.11 presents the behavior of IDHT under the setting of i.i.d. scaled Student-t_d noise, where $d = 3, 5$. More examples are in Section 3 of the supplementary material. We ran 100 replications for each signal and the frequency distribution of $\hat{N} - N$ for each method is presented. The methods with the highest empirical frequency of $\hat{N} - N = 0$ (or in a neighborhood of zero, depending on the example) and those within 10% off the highest are given in bold. As a measure of the accuracy of the detected locations, we provide Monte-Carlo estimates of the mean squared error, MSE $= T^{-1}\sum_{t=1}^{T}\mathbb{E}\left(\hat{f}_t - f_t\right)^2$, where $\hat{f}_t$ is the ordinary least square approximation of f_t between two successive change-points. In continuous piecewise-linear signals, $\hat{f}_t$ is the splines fit obtained using the **splines** package in R. The scaled Hausdorff distance, $d_H = n_s^{-1}\max\left\{\max_j \min_k |r_j - \hat{r}_k|, \max_k \min_j |r_j - \hat{r}_k|\right\}$, where n_s is the length of the largest segment, is also given. The average computational time for all methods, apart from FDR, is also provided. FDR is excluded due to its non-uniform procedure in terms of the execution speed for each signal (if a newly obtained signal has length greater than previously treated signals, then FDR estimates the threshold by 5000 Monte-Carlo simulations, which makes it very slow).

				$\hat{N} - N$							
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	d_H	Time (ms)
PELT		9	29	45	17	0	0	0	3.34	0.13	6.9
NP.PELT		0	0	17	**57**	12	12	2	2.93	0.13	203
S3IB		1	6	36	56	1	0	0	2.54	0.07	282.2
CumSeg		43	26	28	3	0	0	0	6.64	0.20	96.4
CPM.l.500		0	0	0	0	2	7	91	4.71	0.49	7.8
CPM.l.3000		0	1	12	45	23	9	10	3.01	0.17	8.9
WBSC1	(M1)	0	0	14	27	17	24	18	3.05	0.29	162.4
WBSIC		1	3	35	56	5	0	0	2.68	0.06	159.8
NOT		0	3	47	48	1	1	0	2.75	0.08	76.2
FDR		0	0	32	53	9	3	3	2.65	0.10	-
TGUH		0	5	36	49	5	5	0	3.55	0.09	167.2
ID		0	1	34	**63**	2	0	0	2.61	0.06	15.4

		$\hat{N}-N$									
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	d_H	Time (ms)
PELT	(M2)	1	20	3	75	1	0	0	53×10^{-4}	0.17	1.1
NP.PELT		0	0	2	**94**	4	0	0	40×10^{-4}	0.07	24.5
S3IB		0	1	0	**85**	11	2	1	37×10^{-4}	0.08	56.1
CumSeg		0	76	3	20	1	0	0	130×10^{-4}	0.35	22.8
CPM.l.500		0	0	0	40	27	19	14	54×10^{-4}	0.49	0.9
WBSC1		0	0	0	34	19	15	32	54×10^{-4}	0.69	71
WBSIC		0	1	1	**92**	4	2	0	38×10^{-4}	0.09	67.3
NOT		0	0	0	**94**	5	1	0	36×10^{-4}	0.07	33.3
FDR		0	0	0	**85**	13	1	1	36×10^{-4}	0.12	-
TGUH		0	1	0	84	4	9	2	49×10^{-4}	0.16	66.8
ID		0	0	0	**92**	8	0	0	36×10^{-4}	0.10	5.9
PELT	(M3)	82	7	0	11	0	0	0	176×10^{-3}	6.87	1.7
NP.PELT		89	9	2	0	0	0	0	162×10^{-3}	4.32	4.9
S3IB		43	5	2	50	0	0	0	119×10^{-3}	4.15	16.9
CumSeg		100	0	0	0	0	0	0	251×10^{-3}	13	5.2
CPM.l.500		81	5	12	2	0	0	0	153×10^{-3}	4.17	0.4
WBSC1		0	2	8	60	19	10	1	57×10^{-3}	0.39	49.6
WBSIC		8	3	2	64	14	7	2	65×10^{-3}	0.91	45.7
NOT		13	2	7	70	5	2	1	70×10^{-3}	1.13	30.8
FDR		13	10	9	52	11	3	2	76×10^{-3}	0.99	-
TGUH		4	11	3	68	10	4	0	68×10^{-3}	0.50	30.4
ID		3	3	1	**88**	4	1	0	55×10^{-3}	0.62	5.4
PELT	(M4)	1	1	7	**91**	0	0	0	24×10^{-3}	0.19	1.8
NP.PELT		100	0	0	0	0	0	0	771×10^{-3}	1.95	5.6
S3IB		97	2	1	0	0	0	0	213×10^{-3}	1.01	17
CumSeg		0	1	14	74	11	0	0	60×10^{-3}	0.32	7.3
CPM.l.500		0	5	91	4	0	0	0	50×10^{-3}	0.96	0.4
WBSC1		0	0	1	72	21	4	2	23×10^{-3}	0.20	49.8
WBSIC		0	0	0	63	28	8	1	22×10^{-3}	0.20	46.2
NOT		0	0	0	**93**	6	1	0	20×10^{-3}	0.13	115.6
FDR		1	0	0	77	14	6	2	22×10^{-3}	0.17	-
TGUH		0	1	1	**93**	4	1	0	24×10^{-3}	0.16	31.2
ID		0	0	0	**93**	7	0	0	20×10^{-3}	0.14	5.6

Table 4.3: Distribution of $\hat{N}-N$ over 100 simulated data sequences of the piecewise-constant signals (M1)-(M4). The average MSE, d_H and computational time are also given.

	$\hat{N}-N$							
Method	≤ -150	$(-150,-50]$	$(-50,-10)$	$[-10,10]$	>10	MSE	d_H	Time (s)
PELT	100	0	0	0	0	0.55	130.64	0.016
NP.PELT	0	12	88	0	0	0.21	4.95	0.496
S3IB	100	0	0	0	0	0.56	249	2.047
CumSeg	100	0	0	0	0	0.56	249	0.428
CPM.l.500	0	0	0	9	91	0.12	0.59	0.008
CPM.l.10000	0	0	0	**100**	0	0.12	1.21	0.008
WBSC1	0	84	16	0	0	0.27	4.22	0.631
WBSIC	7	0	0	88	5	0.16	15.92	1.051
NOT	100	0	0	0	0	0.56	240.08	0.610
FDR	0	0	0	**99**	1	0.11	0.72	-
TGUH	0	0	2	**98**	0	0.14	1.43	0.580
ID	0	0	0	**100**	0	0.11	0.76	0.139

Table 4.4: Distribution of $\hat{N}-N$ over 100 simulated data sequences from the piecewise-constant signal (M5). The average MSE, d_H and computational time are also given.

	$\hat{N}-N$							
Method	≤ -500	$(-500,-50]$	$(-50,-10)$	$(-10,10]$	>10	MSE	d_H	Time (s)
PELT	100	0	0	0	0	1.97	107.26	0.0311
NP.PELT	100	0	0	0	0	2.25	515.73	10.738
S3IB	100	0	0	0	0	2.25	1999	3.993
CumSeg	100	0	0	0	0	2.25	1999.	0.834
CPM.l.500	0	43	55	2	0	0.19	8.30	0.002
CPM.l.20000	100	0	0	0	0	2.23	1999	0.943
WBSC1	100	0	0	0	0	1.50	34.58	1.262
WBSIC	100	0	0	0	0	2.25	1999	23.236
NOT	100	0	0	0	0	2.25	1999	0.623
FDR	0	0	0	13	87	0.14	0.51	-
TGUH	0	0	0	**100**	0	0.16	0.78	1.468
ID	0	0	0	**100**	0	0.14	0.99	0.479

Table 4.5: Distribution of $\hat{N}-N$ over 100 simulated data sequences from the piecewise-constant signal (M6). The average MSE, d_H and computational time are also given.

	$\hat{N} - N$							
Method	≤ -300	$(-300, -100]$	$(-100, -15)$	$[-15, 15]$	> 15	MSE	d_H	Time (s)
PELT	0	100	0	0	0	0.66	1.02	0.014
NP.PELT	100	0	0	0	0	9657.18	113.77	9.198
S3IB	100	0	0	0	0	687.69	44.76	26.187
CumSeg	100	0	0	0	0	87.11	15.08	0.427
CPM.l.500	0	0	0	**100**	0	0.19	0.74	0.003
CPM.l.10000	0	0	24	76	0	0.22	1	0.004
WBSC1	0	0	36	64	0	0.24	1.01	0.626
WBSIC	0	0	2	**98**	0	0.22	0.99	2.058
NOT	100	0	0	0	0	9.26	5.95	5.598
FDR	0	0	0	**99**	1	0.19	0.82	-
TGUH	0	0	98	2	0	0.29	1.01	0.571
ID	0	0	0	**100**	0	0.20	0.99	0.312

Table 4.6: Distribution of $\hat{N} - N$ over 100 simulated data sequences from the piecewise-constant signal (M7). The average MSE, d_H and computational time are also given.

		$\hat{N} - N$									
Method	Model	≥ -3	-2	-1	0	1	2	≥ 3	MSE	d_H	Time (s)
NOT	(W1)	0	0	0	**100**	0	0	0	0.015	0.057	0.360
TF		0	0	0	0	0	1	99	0.031	0.409	1.599
CPOP		0	0	0	**98**	2	0	0	0.012	0.053	34.672
ID		0	0	0	**95**	5	0	0	0.028	0.093	0.016
NOT	(W2)	0	0	0	**100**	0	0	0	0.016	0.116	0.328
TF		0	0	0	0	0	2	98	0.022	0.261	1.593
CPOP		0	0	0	**96**	4	0	0	0.015	0.123	47.122
ID		0	1	0	**98**	1	0	0	0.028	0.194	0.017

Table 4.7: Distribution of $\hat{N} - N$ over 100 simulated data sequences from the continuous piecewise-linear signals (W1) and (W2). The average MSE, d_H and computational time for each method are also given.

	$\hat{N}-N$									
Method	≤ -90	$(-90,-1)$	-1	0	1	$(1,60]$	> 60	MSE	d_H	Time (s)
NOT	100	0	0	0	0	0	0	4.730	98.990	0.960
TF	0	0	0	0	0	0	100	203.340	0.402	1.357
CPOP	0	0	0	**89**	11	0	0	0.162	0.207	1.664
ID	0	0	0	**97**	3	0	0	0.243	0.286	0.029

Table 4.8: Distribution of $\hat{N}-N$ over 100 simulated data sequences of the continuous piecewise-linear signal (W3). The average MSE, d_H and computational time are also given.

	$\hat{N}-N$									
Method	≤ -100	$(-100,-1)$	-1	0	1	$(1,10]$	> 10	MSE	d_H	Time (s)
NOT	100	0	0	0	0	0	0	1.063	119	0.306
TF	0	0	0	0	0	0	100	210231.1	0.306	0.708
CPOP	0	0	0	**98**	2	0	0	0.027	0.151	0.413
ID	0	0	0	**100**	0	0	0	0.039	0.212	0.024

Table 4.9: Distribution of $\hat{N}-N$ over 100 simulated data sequences of the continuous piecewise-linear signal (W4). The average MSE, d_H and computational time are also given.

		$\hat{N}-N$									
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	d_H	Time (s)
NOT	(W5)	25	11	2	5	8	9	40	7255.568	2.793	0.418
TF		0	0	0	0	0	0	100	$> 34 \times 10^6$	0.444	1.320
CPOP		0	0	0	**99**	1	0	0	0.008	0.002	0.604
ID		0	0	0	**100**	0	0	0	0.007	0	0.024
NOT	(W6)	0	0	0	0	2	2	96	0.068	0.997	1.214
TF		0	0	0	0	0	0	100	647.405	0.468	1.115
CPOP		0	0	0	**95**	5	0	0	0.016	0.095	1.313
ID		0	0	0	**96**	4	0	0	0.037	0.119	0.023

Table 4.10: Distribution of $\hat{N}-N$ over 100 simulated time series of the continuous piecewise-linear signals (W5) and (W6). The average MSE, d_H and computational time are also given.

		$\hat{N} - N$									
d	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	d_H	Time (ms)
5	(M1)	0	2	31	44	8	10	5	3.39	0.12	21.4
	(M2)	0	1	0	78	9	11	1	139×10^{-4}	0.20	5.2
	(M3)	6	2	2	74	9	5	2	60×10^{-3}	0.86	9.7
	(M4)	0	0	0	75	16	5	4	21×10^{-3}	0.16	9.2
	(W1)	0	0	0	80	17	0	3	30×10^{-3}	0.12	29.1
	(W2)	0	0	0	86	12	2	0	31×10^{-3}	0.23	32.8
3	(M1)	0	0	13	28	12	19	28	5.17	0.20	23.3
	(M2)	0	2	4	59	10	12	13	207×10^{-4}	0.32	6.7
	(M3)	7	1	2	52	21	8	9	71×10^{-3}	1.18	8.7
	(M4)	0	1	0	59	20	13	7	26×10^{-3}	0.22	9.8
	(W1)	0	0	0	59	18	11	12	37×10^{-3}	0.16	19.5
	(W2)	0	0	0	62	28	4	6	32×10^{-3}	0.25	22.6

Table 4.11: ID results for the distribution of $\hat{N} - N$ for the models (M1)-(M4) and (W1), (W2), over 100 simulated data sequences where the distribution of the noise is Student-t_d, for $d = 3, 5$. The average MSE, d_H and computational time of IDHT are also given.

ID exhibits excellent performance in all models in both piecewise-constant and continuous piecewise-linear signals. With regards to piecewise-constancy, ID is always in the top 10% of the best methods when considering accuracy in any aspect (estimation of N, MSE, d_H); in most cases it is the best method overall. In continuous piecewise-linear signals, our method is in all cases in the top 10% of the best methods in terms of the accurate estimation of N and it exhibits good performance with respect to the MSE and d_H. We can deduce that ID is consistent in detecting with high accuracy the change-points for various different signal structures, a characteristic which is at least partly absent from its competitors. Furthermore, ID's behavior is particularly impressive in extremely long signals with a large number of frequently occurring change-points; see Tables 4.4, 4.5, 4.6, 4.8, and 4.9. Compared to other well-behaved methods, such as NOT, WBS, FDR, TGUH for piecewise-constancy and NOT, CPOP for continuous piecewise-linear signals, our methodology has by far the lowest computational cost. To conclude, ID is an accurate, trustworthy, and quick method for generalized change-point detection.

The results of Table 4.11 are very good for $d = 5$ and not too different from those under Gaussian noise. For $d = 3$, there is a slight overestimation of the number of change-points. When the tails of the distribution of the noise are significantly heavier than those of the normal distribution, one can obtain better results by increasing the threshold constant. For example, the results in Table 4.11 for $d = 3$ were improved when the threshold constant was slightly increased.

5 Real data examples

5.1 UK House Price Index

We investigate the performance of ID on monthly percentage changes in the UK House price index from January 1995 to August 2017 in two London Boroughs: Tower Hamlets and Hackney. The data are available from http://landregistry.data.gov.uk/app/ukhpi and they were last accessed in January 2018. Figure 5.3 shows the fits of ID, NOT and TGUH. In both data sets, ID behaves similarly to NOT whereas TGUH estimates more change-points. The estimation of two change-points near March 2008 and September 2009 for both boroughs may be related to the financial crisis during that time, which lead to a decrease in the house prices. We highlight that as explained in Section 2.3, our ID methodology returns the solution path defined in (2.9), which can then be used to obtain different fits; see Section 5.2 for more details.

Residual diagnostics have indicated that the behavior of the raw residuals, $\epsilon_t = Y_t - \hat{f}_t$, in relation to normality and independence is good for all methods. This means that there are not significant discrepancies in regards to the goodness of fit between TGUH and the other two methods (which detect the same number of change-points in both data sets).

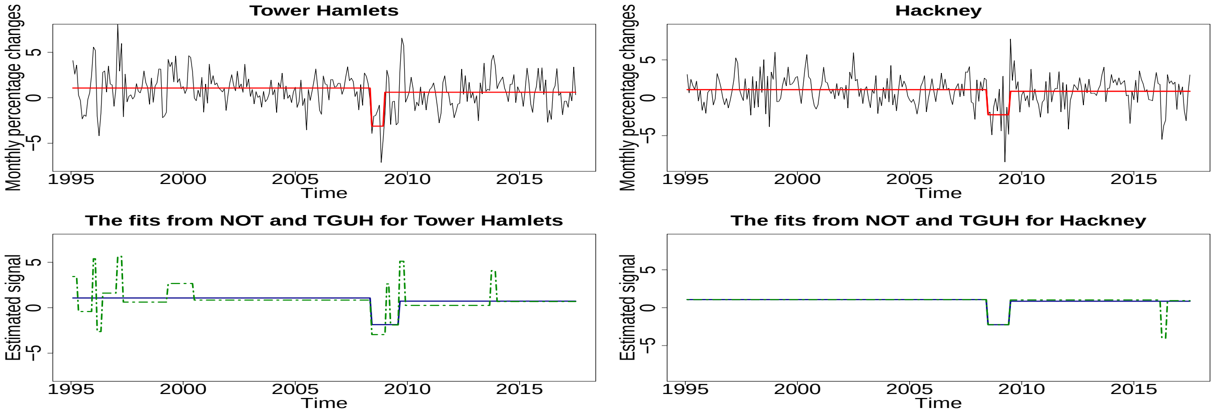

Figure 5.3: **Top row:** The original time series and the fitted piecewise-constant mean signal obtained by ID for both Tower Hamlets and Hackney. **Bottom row:** NOT (solid) and TGUH (dashed) estimates for Tower Hamlets and Newham.

5.2 Samsung Stock Prices

In this section we apply ID to the daily closing stock prices of Samsung Electronics Co. from July 2012 until July 2017. The data are available from https://finance.yahoo.com/quote/005930.KS/history?p=005930.KS and they were last accessed in January 2018. We look for changes in a continuous piecewise-linear mean signal. Figure 5.4 shows the results for the ID, NOT and CPOP methods, which detect 79, 14, and 135 change-points, respectively. From both the fit and the residuals given in Figure 5.4, it is not easy to say which of the three methods gives the “best” number of change-points. ID can return a range of different fits providing users with the flexibility to choose according to their preference. In Figure 5.5, we use the solution path and we obtain the estimated signal and the raw residuals of ID for $\hat{N} = 135$ (the estimated change-point number through CPOP).

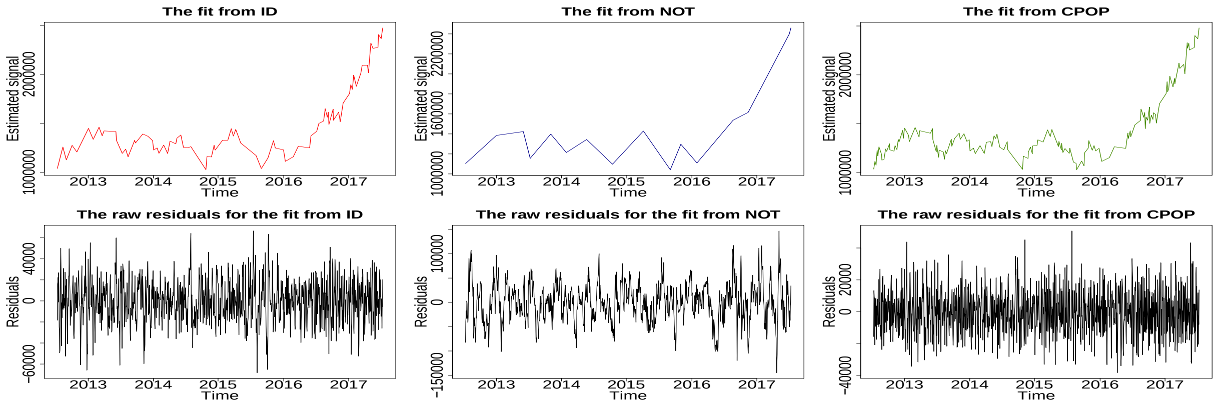

Figure 5.4: **Top row:** From left to right, the fits for ID, NOT and CPOP, respectively. **Bottom row:** The raw residuals $\epsilon_t = Y_t - \hat{f}_t$ for each method.

The fit is similar to the one obtained by CPOP, found in Figure 5.4. However, CPOP is significantly slower than ID; see Tables 4.7-4.10 for a comparison. To conclude, apart from returning the estimated fit, the ID methodology can directly, and without any extra effort, produce a series of estimated signals based on the solution path defined in (2.9).

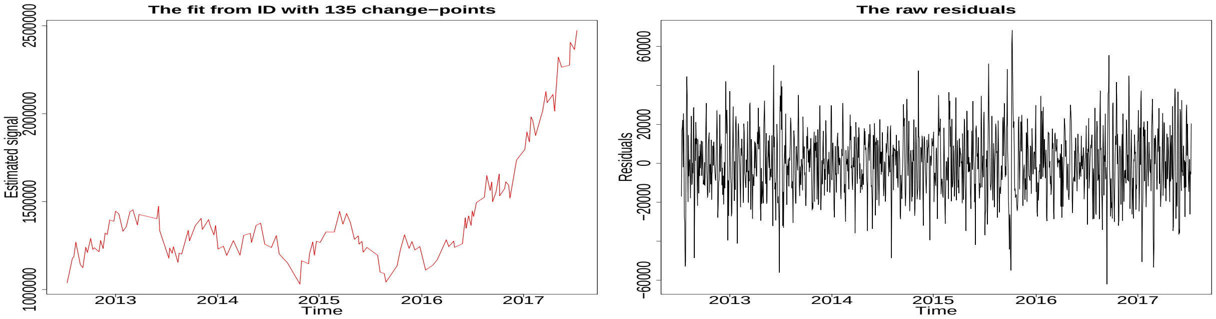

Figure 5.5: The estimated signal and the residuals obtained by ID with 135 change-points.

References

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, **51**, 339–367.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the Optimal Identification of Segment Neighborhoods. *Bulletin of Mathematical Biology*, **51**, 39–54.

Badagián, A. L., Kaiser, R. and Peña, D. (2015). *Time Series Segmentation Procedures to Detect, Locate and Estimate Change-Points. In: Beran J., Feng Y., Hebbel H. (eds) Empirical Economic and Financial Research. Advanced Studies in Theoretical and Applied Econometrics*, Volume **48**. Springer, Cham.

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

Bai, J. and Perron, P. (2003). Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, **18**, 1–22.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2018). Narrowest-Over-Threshold Detection of Multiple Change-points and Change-point-like Features. https://arxiv.org/pdf/1609.00293.pdf.

Bianchi, M., Boyle, M. and Hollingsworth, D. (1999). A comparison of methods for trend estimation. *Applied Economics Letters* **6**, 103–109.

Bolton, R. and Hand, D. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, **17**, 235–255.

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

Chen, J. and Gupta, A. K. (1997). Testing and Locating Variance Changepoints with Application to Stock Prices. *Journal of the American Statistical Association* **92**, 739–747.

Curry, C., Grossman, R. L., Locke, D., Vejcik, S. and Bugajski, J. (2007). Detecting changes in large data sets of payment card data: A case study. *In Proceedings of the* 13^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*, 1018–1022.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, **42**, 2243–2281.

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. http://stats.lse.ac.uk/fryzlewicz/tguh/tguh.pdf.

Fryzlewicz, P. and Subba Rao, S. (2014). Multiple-change-point detection for autoregressive conditional heteroscedastic processes. *Journal of the Royal Statistical Society Series B*, **76**, 903–924.

Futschik, A., Hotz, T., Munk, A. and Sieling, H. (2014). Multiscale DNA partitioning: statistical evidence for segments. *Bioinformatics*, **30**, 2255–2262.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.

Haynes, K., Fearnhead, P. and Eckley, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, **27**, 1293–1305.

Jackson, B., Sargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L. and Tsai, T. T. (2005). An Algorithm for Optimal Partitioning of Data on an Interval. *IEEE Signal Processing Letters*, **12**, 105–108.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, **107**, 1590–1598.

Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D. (2009). ℓ_1 Trend Filtering. *SIAM Review*, **51**, 339–360.

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

Liu, J., Wu, S. and Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica* **7**, 497–526.

Maidstone, R., Fearnhead, P. and Letchford, A. (2017a). Detecting changes in slope with an L_0 penalty. https://arxiv.org/pdf/1701.01672.pdf.

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

Muggeo, V. M. R. and Adelfio, G. (2011). Efficient Change Point Detection for Genomic Sequences of Continuous Measurements. *Bioinformatics*, **27**, 161–166.

National Research Council (2013). *The Mathematical Sciences in 2025.* The National Academies Press.

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

Raimondo, M. (1998). Minimax estimation of sharp change points. *Annals of Statistics* **26**, 1379–1397.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_{max} change-points. *Journal de la Société Française de Statistique*, **156**, 180–205.

Robbins, M. W., Lund, R. B., Gallagher, C. M. and Lu, Q. Q. (2011). Changepoints in the North Atlantic tropical cyclone record. *Journal of the American Statistical Association* **106**, 89–99.

Ross, G. J. (2015). Parametric and Nonparametric Sequential Change Detection in R: The cpm Package. *Journal of Statistical Software*, **66, No.3**, 1–20.

Schröder, A. L. and Fryzlewicz, P. (2013). Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and Its Interface*, **6**, 449–463.

Siris, V. A. and Papagalou, F. (2006). Application of anomaly detection algorithms for detecting syn flooding attacks. *Computer communications*, **29**, 1433–1442.

Stasinopoulos, D. M. and Rigby, R. A. (1992). Detecting break points in generalised linear models. *Computational Statistics and Data Analysis* **13**, 461–471.

Tartakovsky, A. G., Rozovskii, B. L., Blaźek, R. B. and Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, **3**, 252–293.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, **42**, 285–323.

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems.* Ph.D. thesis, Stanford University.

Vostrikova, L. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathematics: Doklady*, **24**, 55–59.

Yang, T. and Swartz T. B. (2005). Applications of Binary Segmentation to the Estimation of Quantal Response Curves and Spatial Intensity. *Biometrical Journal* **47**, 489–501.

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