Narrowest-Over-Threshold Detection of Multiple Change-points and Change-point-like Features

Rafal Baranowski, Yining Chen and Piotr Fryzlewicz
London School of Economics and Political Science

Summary. We propose a new, generic and flexible methodology for nonparametric function estimation, in which we first estimate the number and locations of any features that may be present in the function, and then estimate the function parametrically between each pair of neighbouring detected features. Examples of features handled by our methodology include change-points in the piecewise-constant signal model, kinks in the piecewise-linear signal model, and other similar irregularities, which we also refer to as generalised change-points. Our methodology works with only minor modifications across a range of generalised change-point scenarios, and we achieve such a high degree of generality by proposing and using a new multiple generalised change-point detection device, termed Narrowest-Over-Threshold (NOT). The key ingredient of NOT is its focus on the smallest local sections of the data on which the existence of a feature is suspected. For selected scenarios, we show the consistency and near-optimality of NOT in detecting the number and locations of generalised change-points.
The NOT estimators are easy to implement and rapid to compute. Importantly, the NOT approach is easy to extend by the user to tailor to their own needs. Our methodology is implemented in the R package **not**.

Keywords: Break-point detection, knots, piecewise-polynomial, segmentation, splines.

1. Introduction

This paper considers the canonical univariate statistical model

$$Y_t = f_t + \varepsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where the deterministic and unknown signal f_t is believed to display some regularity across the index t, and the stochastic noise ε_t is exactly or approximately centred at zero. Despite the simplicity of model (1), inferring information about f_t remains a task of fundamental importance in modern applied statistics and data science. When the interest is in the detection of "features" in f_t such as jumps or kinks, then non-linear techniques are usually required.

If f_t is modelled as piecewise-constant and it is of interest to detect its change-points, several techniques are available, and we only mention a selection. For Gaussian noise ε_t, both non-penalised and penalised least squares approaches are considered by Yao and Au (1989). For specific choices of penalty functions, see e.g. Yao (1988), Lavielle (2005) and Davis *et al.* (2006). The Gaussianity assumption on ε_t is relaxed to exponential family

Address for correspondence: Yining Chen, Department of Statistics, Columbia House, Houghton Street, London, WC2A 2AE, U.K.
E-mail: y.chen101@lse.ac.uk

distributions in Lee (1997), Hawkins (2001) and Frick *et al.* (2014). In particular, Frick *et al.* (2014) also provide confidence intervals for the location of the estimated change-points. Often this penalty-type approach requires a computational cost of at least $O(T^2)$. However, there are exceptions, such as the Pruned Exact Linear Time method (PELT; Killick *et al.* (2012a)), which achieves a linear computational cost, but requires the further assumption that change-points are separated by time intervals drawn independently from some probability distribution, a scenario in which considerations of statistical consistency are not generally possible. A nonparametric version of PELT is investigated by Haynes *et al.* (2017). Another general approach is based on the idea of Binary Segmentation (BS; Vostrikova, 1981), which can be viewed as a greedy approach with a limited computational cost. Its popular variants include the Circular Binary Segmentation (CBS; Olshen *et al.*, 2004) and the Wild Binary Segmentation (WBS; Fryzlewicz, 2014). A selection of publications and software can be found in the online repository *changepoint.info* maintained by Killick *et al.* (2012b).

More general change-point problems, in which f_t is modelled as piecewise-parametric (not necessarily piecewise-constant) between "knots", the number and locations of which are unknown and need to be estimated, have attracted less interest in the literature and overwhelmingly focus on linear trend detection. Among them, we mention the approach based on the least squares principle and Wald-type tests by Bai and Perron (1998), dynamic programming using the L_0 penalty (Maidstone *et al.*, 2017), and trend filtering (Tibshirani, 2014; Lin *et al.*, 2017). Finally, we mention a related problem of jump regression, where the aim is to estimate the points of sharp cusps or discontinuities of a regression function. As investigated in, e.g., Wang (1995) and Xia and Qiu (2015), it proceeds by estimating the locations of features nonparametrically via wavelets or local kernel smoothing.

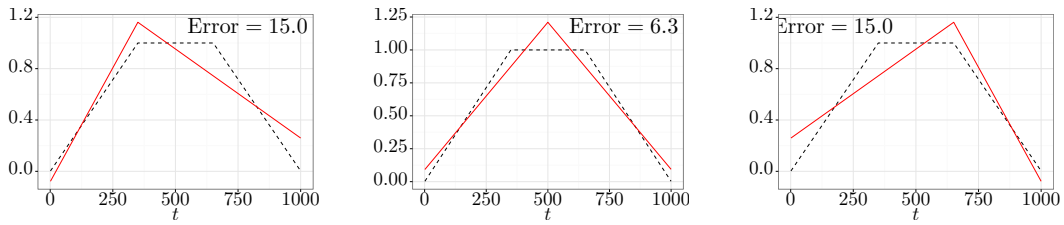
The aim of this work is to propose a new, generic approach to the problem of detecting an unknown number of "features" occurring at unknown locations in f_t. By a feature, we mean a characteristic of f_t, occurring at a location t_0, that is detectable by considering a sufficiently large subsample of data Y_t around t_0. Examples include: change-points in f_t when it is modelled as piecewise-constant, change-points in the first derivative when f_t is modelled as piecewise-linear and continuous, and discontinuities in f_t or its first derivative when f_t is modelled as piecewise-linear but without the continuity constraint. We will provide a precise description of the type of features we are interested in later. Moving beyond f_t only, our approach will also permit the detection of similar features present in some distributional aspects of ε_t, for example in its variance. Since all types of features we consider describe changes in a parametric description of f_t, we use the terms "feature detection" and "change-point detection" interchangeably throughout the paper. Occasionally, for precision, we will be referring to change-point detection in the piecewise-constant model as the "canonical" change-point problem, while our general feature detection problem will sometimes be referred to as a "generalised" change-point problem.

Core to our approach is a particular blend of "global" and "local" treatment of the data Y_t in the search for the multiple features that may be present in f_t, a combination that gives our method a multiscale character. At the first "global" stage, we randomly draw a number of subsamples $(Y_{s+1}, \dots, Y_e)'$, where $0 \leq s < e \leq T$. On each subsample, we assume, possibly erroneously, that *only one* feature is present and use a tailor-made contrast function derived (according to a universal recipe we provide later) from the likelihood theory to find the most likely location of the feature. We retain those subsamples for which the contrast *exceeds a certain user-specified threshold*, and discard the others. Amongst the retained subsamples, we search for the one drawn on the *narrowest* interval, i.e. one

for which $e - s$ is the smallest: it is this step that gives rise to the name *Narrowest-Over-Threshold* (NOT) for our methodology. The focus on the narrowest interval constitutes the "local" part of the method, and is a key ingredient of our approach which ensures that with high probability, at most one feature is present in the selected interval. This key observation gives our methodology a general character and allows it to be used, only with minor modifications, in a wide range of scenarios, including those described in the previous paragraph. Having detected the first feature, the algorithm then proceeds recursively to the left and to the right of it, and stops, on any current interval, if no contrasts can be found that exceed the threshold.

Besides its generic character, other benefits of the proposed methodology include low computational complexity, ease of implementation, accuracy in the detection of the feature locations, and the fact that it enables parametric estimation of the signal on each section delimited by a pair of neighbouring estimated features. Regarding the computational complexity, the fact that typical contrasts are computable in linear time leads to a computational complexity of $O(MT)$ for the entire procedure; typically, only a limited number of data subsamples, M, need to be drawn (we provide precise bounds later; with finitely many change-points, one can take $M = O(\log T)$ in general). Moreover, the entire threshold-indexed solution path can also be computed efficiently, in typically close-to-linear time, as observed from our numerical experiments. Regarding the estimation accuracy, in the scenarios we consider theoretically, our procedure yields near-optimal rates of convergence for the estimators of feature locations.

On a broader level, our methodology promotes the idea of *fitting simple models on subsets of the data (the local aspect), and then aggregating the results to obtain the overall fit (the global aspect)*, an idea also present in the Wild Binary Segmentation method of Fryzlewicz (2014). However, we emphasise that the way the simple models (here: models containing *at most one* change-point or feature) are fitted in the NOT and WBS methods are entirely different and have different aims. Unlike the WBS, the NOT methodology focuses on the *narrowest* intervals of the data on which it is possible to locate the feature of interest. It is this focus that enables NOT to extend beyond change-point detection for a piecewise-constant f_t, the latter being the sole focus of the WBS method. The lack of the narrowest-interval focus in the WBS and BS methods means that they are not applicable to more general feature detection, and we explain the mechanics of this important phenomenon briefly in the following simple example.



(a) $\tau = 350$ (b) $\tau = 500$ (c) $\tau = 651$

Fig. 1. Best ℓ_2 approximation of the true signal (dashed) via a triangular signal with a single change-point, the location of which is fixed at the left change-point (left panel), halfway between the true change-points (middle panel) and at the right change-point (right panel). Approximation errors (in terms of squared ℓ_2 distance) are given in the top-right corners of the corresponding panels.

Consider a continuous piecewise-linear signal that has two change-points:

$$f_t = \begin{cases} \frac{1}{350}t, & t = 1, \ldots, 350, \\ 1, & t = 351, \ldots, 650, \\ \frac{1001}{350} - \frac{1}{350}t, & t = 651, \ldots, 1000. \end{cases} \tag{2}$$

If we approximate f_t using a piecewise-linear signal with only one change-point in its derivative, then the best approximation (in terms of minimising the ℓ_2 distance) will result in an estimated change-point at $t = 500$, which is away from the true ones at $t = 350$ and $t = 650$, as is illustrated in Figure 1. Therefore, taking the entire sample of data and searching for one of its multiple change-points by fitting, via least squares, a triangular signal with a single change-point, does not make sense. It is this issue that leads to the failure of the BS and WBS methods for signals that are not piecewise constant. On the other hand, NOT avoids this issue because of its unique feature of picking the *narrowest* intervals, which are likely to contain only one change-point. To understand the mechanics of this key feature, imagine that now f_t is observed with noise. Through its pursuit of the narrowest intervals, NOT will ensure that, with high probability, some suitably narrow intervals around the change-points $t = 350$ and $t = 650$ are considered. More precisely, by construction, they will be *narrow enough to contain only one change-point each*, but wide enough for the designed contrast (see Section 2.3 for more on contrasts) to indicate the existence of the change-point within both of them. The designed contrast function will indicate the correct location of the change-point (modulo the estimation error) if only one change-point is present in the data subsample considered, unlike in the situation described earlier in which multiple change-points were included in the chosen interval. More details on this example are presented in Section C.3 of the online supplementary materials.

Note that this example is different from the canonical change-point detection problem (i.e. piecewise-constant signal with multiple change-points), where if we approximate the signal using a piecewise-constant function with only one change-point, the change-point of the fitted signal will always be among the true ones (Venkatraman, 1992). Since the latter property does not hold in most generalised change-point detection problems, this highlights the need for new methods with better localisation of the feature of interest, such as our NOT algorithm. Fang *et al.* (2016) independently consider a related shortest-interval idea in the context of the canonical change-point detection problem. However, they did not consider it as a springboard to more general feature detection problems, which is the key motivation behind NOT and its most valuable contribution.

The remainder of this paper is organised as follows. In Section 2, we give a mathematical description of NOT. In particular, we consider NOT in four scenarios, each with a different form of structural change in the mean and/or variance. For the development of both theory and computation, in selected scenarios, we introduce the tailor-made contrast function derived from the generalised likelihood ratio (GLR). Theoretical properties of NOT, such as its consistency and convergence rates are also provided. In Section 3, we propose to use NOT with the strengthened Schwarz Information Criterion (sSIC) and discuss its computational aspects and theoretical properties. Section 4 discusses possible extensions of NOT. A comprehensive simulation study is carried out in Section 5, where we compare NOT with the state-of-art change-point detection tools. In Section 6, we consider data examples of global temperature anomalies and London housing data. All proofs, together with details on the construction of the contrast functions, the computational aspects and extension of NOT, further discussion on model misspecification, as well as additional simulations and real data example can be found in the online supplementary materials.

2. The framework of NOT

2.1. Setup

To describe the main framework of NOT, we consider a simplified version of (1), where $\mathbf{Y} = (Y_1, \ldots, Y_T)'$ is modelled through

$$Y_t = f_t + \sigma_t \varepsilon_t, \quad t = 1, \ldots, T, \tag{3}$$

where f_t is the signal, and where σ_t is the noise's standard deviation at time t. To facilitate the technical presentation of our results, in Sections 2 and 3, we assume that $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. In Section 4, we extend our framework to other noise types.

We assume that (f_t, σ_t) can be partitioned into $q+1$ segments, with q unknown distinct change-points $0 = \tau_0 < \tau_1 < \ldots < \tau_q < \tau_{q+1} = T$. Here the value of q is not pre-specified and can grow with T. For each $j = 1, \ldots, q+1$ and for $t = \tau_{j-1}+1, \ldots, \tau_j$, the structure of (f_t, σ_t) is is modelled parametrically by a local (i.e. depending on j) real-valued d-dimensional parameter vector $\boldsymbol{\Theta}_j$ (with $\boldsymbol{\Theta}_j \neq \boldsymbol{\Theta}_{j-1}$), where d is known and typically small. To fix ideas, in the following, we assume that each segment of f_t and σ_t follows a polynomial. In addition, we require the minimum distance between consecutive change-points to be $\geq d$ for the purpose of identifiability. (Otherwise, e.g. take f_t to be piecewise-linear with a known constant σ_t, in which case $d = 2$; if we had a segment of length 1, then we would not be able to define a line based on a single point.) In other words, (f_t, σ_t) can be divided into q different segments, each from the same parametric family of much simpler structure. Some commonly-encountered scenarios are listed below, where the following holds inside the j-th segment for each $j = 1, \ldots, q+1$:

(S1) **Constant variance, piecewise-constant mean**:
$\sigma_t = \sigma_0$ and $f_t = \theta_j$ for $t = \tau_{j-1}+1, \ldots, \tau_j$.

(S2) **Constant variance, continuous and piecewise-linear mean**:
$\sigma_t = \sigma_0$ and $f_t = \theta_{j,1} + \theta_{j,2}\, t$ for $t = \tau_{j-1}+1, \ldots, \tau_j$, with the additional constraint of

$$\theta_{j,1} + \theta_{j,2}\, \tau_j = \theta_{j+1,1} + \theta_{j+1,2}\, \tau_j$$

for $j = 1, \ldots, q$.

(S3) **Constant variance, piecewise-linear (but not necessarily continuous) mean**:
$\sigma_t = \sigma_0$ and $f_t = \theta_{j,1} + \theta_{j,2}\, t$ for $t = \tau_{j-1}+1, \ldots, \tau_j$. In addition, $f_{\tau_j} + \theta_{j,2} \neq f_{\tau_j+1}$ for $j = 1, \ldots, q$.

(S4) **Piecewise-constant variance, piecewise-constant mean**:
$f_t = \theta_{j,1}$ and $\sigma_t = \theta_{j,2} > 0$ for $t = \tau_{j-1}+1, \ldots, \tau_j$.

Since σ_0 in (S1)–(S3) acts as a nuisance parameter, in the rest of this manuscript, for simplicity we assume that its value is known. If it is unknown, then it can be estimated accurately using the Median Absolute Deviation (MAD) method (Hampel, 1974). More specifically, with i.i.d. Gaussian errors, the MAD estimator of σ_0 is defined as $\hat{\sigma} = \text{Median}\{|Y_2 - Y_1|, \ldots, |Y_T - Y_{T-1}|\}/\{\Phi^{-1}(3/4)\sqrt{2}\}$ in Scenario (S1), and as $\hat{\sigma} = \text{Median}\{|Y_1 - 2Y_2 + Y_3|, \ldots, |Y_{T-2} - 2Y_{T-1} + Y_T|\}/\{\Phi^{-1}(3/4)\sqrt{6}\}$ in Scenarios (S2) and (S3). Here $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. Note that the MAD estimator is robust to any change-points present in the underlying signal f_t, due to its combination of working with the differenced data, and its use of the median. Finally, we note that a different procedure is proposed to estimate σ_0 with dependent errors; see Section 4.1 for more details.

2.2. Main idea

We now describe the main idea of NOT formally; more details can be found in Section 2.4 where the pseudo-code of the NOT algorithm is given.

In the first step, instead of directly using the entire data sample, we randomly extract subsamples, i.e. vectors $(Y_{s+1},\dots,Y_e)'$, where (s,e) is drawn uniformly from the set of pairs of indices in $\{0,\dots,T-1\}\times\{1,\dots,T\}$ that satisfy $0\le s<e\le T$. Let $\ell(Y_{s+1},\dots,Y_e;\boldsymbol{\Theta})$ be the likelihood of $\boldsymbol{\Theta}$ given $(Y_{s+1},\dots,Y_e)'$. We then compute the generalised log-likelihood ratio (GLR) statistic for all potential single change-points within the subsample and pick the maximum, that is,

$$\mathcal{R}^b_{(s,e]}(\mathbf{Y}) = 2\log\left[\frac{\sup_{\boldsymbol{\Theta}^1,\boldsymbol{\Theta}^2}\left\{\ell(Y_{s+1},\dots,Y_b;\boldsymbol{\Theta}^1)\ell(Y_{b+1},\dots,Y_e;\boldsymbol{\Theta}^2)\right\}}{\sup_{\boldsymbol{\Theta}}\ell(Y_{s+1},\dots,Y_e;\boldsymbol{\Theta})}\right]; \tag{4}$$
$$\mathcal{R}_{(s,e]}(\mathbf{Y}) = \max_{b\in\{s+d,\dots,e-d\}}\mathcal{R}^b_{(s,e]}(\mathbf{Y}).$$

Note that here we also implicitly require $e-s\ge 2d$, which comes from the identifiability condition, because typically we need at least d observations to determine $\boldsymbol{\Theta}^1$, and another d observations to determine $\boldsymbol{\Theta}^2$.

If constraints are in place between $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Theta}_{j+1}$ for any $j=1,\dots,q$ (e.g. as in (S2)), the supremum in the numerator of (4) is taken over the set that only contains elements of form $\boldsymbol{\Theta}^1\times\boldsymbol{\Theta}^2$ satisfying these constraints. Otherwise, as in (S1), (S3) and (S4), (4) can be simplified to

$$\mathcal{R}^b_{(s,e]}(\mathbf{Y}) = 2\log\left\{\frac{\sup_{\boldsymbol{\Theta}}\ell(Y_{s+1},\dots,Y_b;\boldsymbol{\Theta})\sup_{\boldsymbol{\Theta}}\ell(Y_{b+1},\dots,Y_e;\boldsymbol{\Theta})}{\sup_{\boldsymbol{\Theta}}\ell(Y_{s+1},\dots,Y_e;\boldsymbol{\Theta})}\right\}.$$

The above procedure is repeated on M randomly drawn pairs of integers $(s_1,e_1),\dots,(s_M,e_M)$.

In the second step, we test all $\mathcal{R}_{(s_m,e_m]}(\mathbf{Y})$ for $m=1,\dots,M$ against a given threshold ζ_T. Among those significant ones, we pick the one corresponding to the interval $(s_{m^*},e_{m^*}]$ that has the smallest length. Once a change-point is found in $(s_{m^*},e_{m^*}]$ (i.e. b^* that maximises $\mathcal{R}^b_{(s_{m^*},e_{m^*}]}(\mathbf{Y})$, a function of b), the same procedure is then repeated recursively to the left and to the right of it, until no further significant GLRs can be found. Note that in each recursive step, one could reuse the previously drawn intervals, provided that they fall entirely within each current subsegment considered.

After the process of estimating the change-points is completed, one can estimate the signals within each segment using standard methods such as least squares or maximum likelihood. Note that the estimation of knot locations in spline regression can be viewed as a multiple change-point detection problem set in the context of polynomial segments that are continuously differentiable but have discontinuous higher order derivatives at the change-points between these segments; NOT can be used for this purpose.

Admittedly, in our framework, one could also use a deterministic scheme (e.g. that in Rufibach and Walther (2010)) to pick a sufficiently rich family of intervals for multiscale inference. However, one advantage of our approach is that through the use of randomness in drawing the intervals, we avoid having to make a subjective choice of a particular fixed design. Nevertheless, with a very large number of intervals drawn, the difference in performance between the random and deterministic designs is likely to be minimal, an observation also made in Fryzlewicz (2014).

2.3. Log-likelihood ratios and contrast functions

In many applications, the GLR (4) in NOT can be simplified with the help of "contrast functions" under the setting of Gaussian noise. In particular, these constructions mainly involve taking inner products between the data and other deterministic vectors, which greatly facilitates the development of both theory and computation, especially if these deterministic vectors are mutually orthonormal. In fact, the form of these contrast functions is crucial in our theoretical development.

More precisely, for every integer triple (s, e, b) with $0 \leq s < e \leq T$, our aim is to find $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ such that:

(a) $\operatorname{argmax}_b \mathcal{C}^b_{(s,e]}(\mathbf{Y}) = \operatorname{argmax}_b \mathcal{R}^b_{(s,e]}(\mathbf{Y})$,
(b) heuristically speaking, the value of $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ is relatively small if there is no change-point in $(s, e]$,
(c) the formulation of $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ mainly consists of taking inner products between the data and certain contrast vectors.

In the following, we give the contrast functions corresponding to scenarios (S1) and (S2), where the aforementioned properties are satisfied. Their details under scenarios (S3) and (S4), as well as a comprehesive discussion on the construction, can be found in Section B of the online supplementary materials. We note that this approach recovers the CUSUM statistic in (S1), which is popular in this canonical change-point detection setting. One can view the resulting statistics as generalisations of CUSUM under other scenarios.

2.3.1. Scenario (S1)

Here f_t is piecewise-constant. For any integer triple (s, e, b) with $0 \leq s < e \leq T$ and $s < b < e$, we define the contrast vector $\boldsymbol{\psi}^b_{(s,e]} = \big(\psi^b_{(s,e]}(1), \ldots, \psi^b_{(s,e]}(T)\big)'$ as

$$\psi^b_{(s,e]}(t) = \begin{cases} \sqrt{\frac{e-b}{(e-s)(b-s)}}, & t = s+1, \ldots, b \\ -\sqrt{\frac{b-s}{(e-s)(e-b)}}, & t = b+1, \ldots, e \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Also, if $b \notin \{s+1, \ldots, e-1\}$, then we set $\psi^b_{(s,e]}(t) = 0$ for all t. As an illustration, plots of $\boldsymbol{\psi}^b_{(s,e]}$ with different (s, e, b) are shown in Figure 2a.

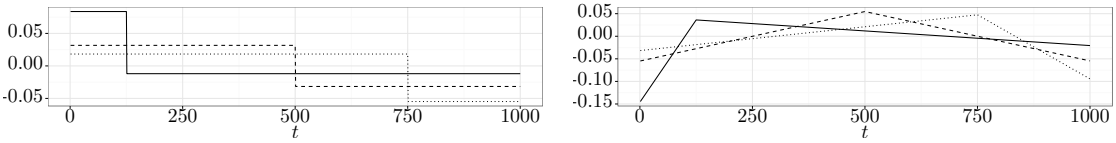
For any vector $\mathbf{v} = (v_1, \ldots, v_T)'$, we define the contrast function as

$$\mathcal{C}^b_{(s,e]}(\mathbf{v}) = \left| \left\langle \mathbf{v}, \boldsymbol{\psi}^b_{(s,e]} \right\rangle \right| \tag{6}$$

2.3.2. Scenario (S2)

Here f_t is piecewise-linear and continuous. For any triple (s, e, b) with $0 \leq s < e \leq T$ and $s + 1 < b < e$, consider the contrast vector $\boldsymbol{\phi}^b_{(s,e]} = \big(\phi^b_{(s,e]}(1), \ldots, \phi^b_{(s,e]}(T)\big)'$ with

$$\phi^b_{(s,e]}(t) = \begin{cases} \alpha^b_{(s,e]}\beta^b_{(s,e]}\Big[\big\{3(b-s) + (e-b) - 1\big\}t - \big\{b(e-s-1) + 2(s+1)(b-s)\big\}\Big], & t = s+1, \ldots, b \\ -\frac{\alpha^b_{(s,e]}}{\beta^b_{(s,e]}}\Big[\big\{3(e-b) + (b-s) + 1\big\}t - \big\{b(e-s-1) + 2e(e-b+1)\big\}\Big], & t = b+1, \ldots, e, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$



(a) $\boldsymbol{\psi}^b_{(s,e]}$ (b) $\boldsymbol{\phi}^b_{(s,e]}$

Fig. 2. Plots of $\boldsymbol{\psi}^b_{(s,e]}$ and $\boldsymbol{\phi}^b_{(s,e]}$ given by, respectively, (5) and (7) for $s = 0$, $e = 1000$ and several values of b. Solid line: $b = 125$; dashed line: $b = 500$; dotted line: $b = 750$.

where $\alpha^b_{s,e} = \left(\frac{6}{l(l^2-1)\left(1+(e-b+1)(b-s)+(e-b)(b-s-1)\right)}\right)^{1/2}$, $\beta^b_{s,e} = \left(\frac{(e-b+1)(e-b)}{(b-s-1)(b-s)}\right)^{1/2}$ and $l = e-s$. If $b \notin \{s+2, \ldots, e-1\}$, then we set $\phi^b_{(s,e]}(t) = 0$ for all t. We illustrate the structure of $\boldsymbol{\phi}^b_{(s,e]}$ in Figure 2b. The contrast function is then defined as

$$\mathcal{C}^b_{(s,e]}(\mathbf{v}) = \left|\left\langle \mathbf{v}, \boldsymbol{\phi}^b_{(s,e]} \right\rangle\right|. \tag{8}$$

2.4. The NOT algorithm

Here we present the pseudo-code of a generic version of the NOT algorithm. The main ingredient of the NOT procedure is a contrast function $\mathcal{C}^b_{(s,e]}(\cdot)$, chosen by the user, depending on the assumed nature of change-points in the data, e.g. as exemplified by our scenarios (S1) and (S2) above, and scenarios (S3) and (S4) in Section B of the online supplementary materials. In addition, some tuning parameters are needed: $\zeta_T > 0$ is the threshold with respect to which the contrast should be tested, while M is the number of the intervals drawn in the procedure. Guidance on the choice of ζ_T and M is given in Section 3. In particular, there we advocate an automatic choice of ζ_T by combining NOT with an information-based criterion, thus making our procedure threshold-free.

To sum up, the input include the data vector $\mathbf{Y}$, the set of F^M_T that contains all randomly drawn sub-intervals for testing, and the global variable $\mathcal{S}$ for the set of estimated change-points initialised with $\mathcal{S} = \emptyset$. Then NOT is started recursively with $(s, e] = (0, T]$ and a given ζ_T. Here the entire set of F^M_T that contains all random intervals is generated before we start running Algorithm 1. In this way, we are better able to control the computational complexity of the entire procedure.

2.5. Theoretical properties of NOT

In this section, we analyse the theoretical behaviour of the NOT algorithm in Scenarios (S1) and (S2). We use infill asymptotics, which is standard in the literature on a posteriori change-point detection. An attractive feature of our methodology is that proofs for other scenarios can in principle be constructed "at home" by the user, by following the same generic proof strategy as the one we use for these two scenarios.

First, we revisit the canonical change-point detection problem, (S1), where the signal vector $\mathbf{f} = (f_1, \ldots, f_T)'$ is piecewise-constant. Here σ_0 is assumed to be known. Otherwise, one can plug in the MAD estimator, described in Section 2.1, without affecting the validity of our theory. For notational convenience, we set $\sigma_0 = 1$. For other values of σ_0, our theorems are still valid with only minor adjustments to the constants therein. Explicit

Algorithm 1 NOT

Input: Data vector $\mathbf{Y} = (Y_1, \ldots, Y_T)'$, F_T^M being a set of M left-open and right-closed intervals, with each pair of start- and end- points drawn independently and uniformly from the set of pairs of indices in $\{0, \ldots, T-1\} \times \{1, \ldots, T\}$ that satisfy the conditions outlined at the beginning of Section 2.2, $\mathcal{S} = \emptyset$.

Output: Set of estimated change-points $\mathcal{S} \subset \{1, \ldots, T\}$.

To start the algorithm: Call NOT$((0, T], \zeta_T)$

procedure NOT$((s, e], \zeta_T)$
  if $e - s \leq 1$ **then** STOP
  else
    $\mathcal{M}_{(s,e]} := \left\{m : (s_m, e_m] \in F_T^M, (s_m, e_m] \subset (s, e]\right\}$
    if $\mathcal{M}_{(s,e]} = \emptyset$ **then** STOP
    else
      $\mathcal{O}_{(s,e]} := \left\{m \in \mathcal{M}_{(s,e]} : \max_{s_m < b \leq e_m} \mathcal{C}^b_{(s_m, e_m]}(\mathbf{Y}) > \zeta_T\right\}$
      if $\mathcal{O}_{(s,e]} = \emptyset$ **then** STOP
      else
        $m^* :\in \operatorname{argmin}_{m \in \mathcal{O}_{(s,e]}} |e_m - s_m|$
        $b^* := \operatorname{argmax}_{s_{m^*} < b \leq e_{m^*}} \mathcal{C}^b_{(s^*_m, e^*_m]}(\mathbf{Y})$
        $\mathcal{S} := \mathcal{S} \cup \{b^*\}$
        NOT$((s, b^*], \zeta_T)$
        NOT$((b^*, e], \zeta_T)$
      end if
    end if
  end if
end procedure

expressions for all the constants (i.e. $\underline{C}, C_1, C_2, C_3$) are given in Section I.2 of the online supplementary materials.

Theorem 1. *Suppose Y_t follow model (3) in Scenario (S1). Let $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$, $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^{\mathbf{f}}$. Let $\hat{q}$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}$ denote, respectively, the number and locations of change-points, sorted in increasing order, estimated by Algorithm 1 with the contrast function given by* (6)*. Then there exist constants $\underline{C}$, $C_1, C_2, C_3 > 0$ (not depending on T) such that given $\delta_T^{1/2} \underline{f}_T \geq \underline{C}\sqrt{\log T}$, $C_1\sqrt{\log T} \leq \zeta_T < C_2 \delta_T^{1/2} \underline{f}_T$ and $M \geq 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1})$, as $T \to \infty$,*

$$\mathbb{P}\left(\hat{q} = q, \max_{j=1,\ldots,q}\left(|\hat{\tau}_j - \tau_j|(\Delta_j^{\mathbf{f}})^2\right) \leq C_3 \log T\right) \to 1. \tag{9}$$

Given two sequences $\{A_T\}_{T=1}^{\infty}$ and $\{B_T\}_{T=1}^{\infty}$, we write $A_T \sim B_T$ when $A_T = O(B_T)$ and $B_T = O(A_T)$. In the simplest canonical case where we have finitely many change-points with $\delta_T \sim T$ and $\underline{f}_T \sim 1$, so the condition $\delta_T^{1/2}\underline{f}_T \geq \underline{C}\sqrt{\log T}$ is always satisfied for a sufficiently large T. Theorem 1 indicates that the NOT procedure requires $M = O(\log T)$ many random intervals for consistent detection of all the change-points, which leads to a total computational cost of $O(T \log T)$ for the entire procedure. Furthermore,

$\max_{j=1,\dots,q}\left(|\hat{\tau}_j - \tau_j|\right) = O_p(\log T)$, which trails the minimax rate of $O_p(1)$ by only a logarithmic factor. In addition, we note that the NOT procedure allows for $\delta_T^{1/2}\underline{f}_T$, a quantity that characterises the difficulty level of the problem, to be of order $\sqrt{\log T}$. As argued in Chan and Walther (2013), this is the smallest rate that permits change-point detection for any method from a minimax perspective.

Next, we revisit Scenario (S2), in which the signal is piecewise-linear and continuous. Again, we set $\sigma_0 = 1$ for notational convenience. Explicit expressions of the constants in the following theorem (i.e. $\underline{C}, C_1, C_2, C_3$) can be found in Section I.3 of the online supplementary materials.

Theorem 2. *Suppose Y_t follow model (3) in Scenario (S2). Let $\delta_T = \min_{j=1,\dots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$, $\underline{f}_T = \min_{j=1,\dots,q}\Delta_j^{\mathbf{f}}$. Let $\hat{q}$ and $\hat{\tau}_1,\dots,\hat{\tau}_{\hat{q}}$ denote, respectively, the number and locations of change-points, sorted in increasing order, estimated by Algorithm 1 with the contrast function given by* (8)*. Then there exist constants $\underline{C}, C_1, C_2, C_3 > 0$ (not depending on T) such that given $\delta_T^{3/2}\underline{f}_T \geq \underline{C}\sqrt{\log T}$, $C_1\sqrt{\log T} \leq \zeta_T < C_2\delta_T^{3/2}\underline{f}_T$ and $M \geq 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1})$, as $T \to \infty$,*

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\dots,q}\left(|\hat{\tau}_j - \tau_j|(\Delta_j^{\mathbf{f}})^{2/3}\right) \leq C_3(\log T)^{1/3}\right) \to 1. \tag{10}$$

In the case in which we have finitely many change-points with $\delta_T \sim T$, we again need $M = O(\log T)$ random intervals for consistent estimation of all the change-points, leading to the total computational cost of $O(T\log T)$. In addition, when $\underline{f}_T \sim T^{-1}$ (a case in which f_t is bounded), our theory indicates that the resulting change-point detection rate of NOT is $O_p(T^{2/3}(\log T)^{1/3})$, which is different from the rate of $O_p(T^{2/3})$ derived by Raimondo (1998) by only a logarithmic factor; moreover, under additional assumptions and with a more careful but restrictive choice of ζ_T, this rate can be further improved to $O_p(T^{1/2}(\log T)^{1/2})$; see Section 3.4 and Lemma 9 in the online supplementary materials for more details. Furthermore, we remark that in more general cases (i.e. number of change-points increasing with T) in Scenario (S2), the difficulty level of the problem in Scenario (S2) can be charaterised by $\delta_T^{3/2}\underline{f}_T$, a quantity analogous to $\delta_T^{1/2}\underline{f}_T$ in the setting of (S1).

Both Theorem 1 and Theorem 2 imply that there exists an admissible range of thresholds that *would* ensure consistent change-point detection. They pave the way for establishing Theorem 3 and Theorem 4 in Section 3, which promote the automatic selection of the threshold via an information criterion.

Finally, we emphasise again that the WBS will fail to estimate change-point consistently in Scenario (S2), for reasons described in Section 1.

3. NOT with the strengthened Schwarz Information Criterion (sSIC)

3.1. Motivation

The success of Algorithm 1 depends on the choice of the threshold ζ_T. Although Theorem 1 and Theorem 2 state that there exists ζ_T that guarantee consistent estimation of the change-points, this choice still typically depends on some unobserved quantities; furthermore, there are many more general scenarios where a theoretically optimal threshold might be difficult to derive.

Note that for a given $\mathbf{Y}$ and F_T^M, each threshold ζ_T corresponds to a candidate model produced by NOT. Therefore, if we could produce a "solution path" of candidate models obtained from NOT along all possible thresholds, we could then try to select the best model along the solution path via minimising an information-based criterion. In this sense, the task of selecting the best threshold is equivalent to selecting the best model on the solution path.

3.2. Algorithm 2: the NOT solution path algorithm

Denote by $\mathcal{T}(\zeta_T) = \{\hat{\tau}_1(\zeta_T), \ldots, \hat{\tau}_{\hat{q}(\zeta_T)}(\zeta_T)\}$ the locations of change-points estimated by Algorithm 1 with threshold ζ_T and define the threshold-indexed solution path as the family of sets $\{\mathcal{T}(\zeta_T)\}_{\zeta_T \geq 0}$. Note that this threshold-indexed solution path has the following important properties. First, as a function $\zeta_T \mapsto \mathcal{T}(\zeta_T)$, it changes its value only at discrete points, i.e. there exist $0 = \zeta_T^{(0)} < \zeta_T^{(1)} < \ldots < \zeta_T^{(N)}$, such that $\mathcal{T}(\zeta_T^{(i)}) \neq \mathcal{T}(\zeta_T^{(i+1)})$ for any $i = 0, 1, \ldots, N-1$, and $\mathcal{T}(\zeta_T) = \mathcal{T}(\zeta_T^{(i)})$ for any $\zeta_T \in [\zeta_T^{(i)}, \zeta_T^{(i+1)})$; and second, $T(\zeta_T) = \emptyset$ for any $\zeta_T \geq \zeta_T^{(N)}$.

However, the thresholds $\zeta_T^{(i)}$ are unknown and depend on the data, therefore naively applying Algorithm 1 on a range of pre-specified thresholds typically does not recover the entire solution path. Moreover, from the computational point of view, repeated application of Algorithm 1 to find the solution path is not optimal either, because intuitively one would expect the solutions for $\zeta_T^{(i+1)}$ and $\zeta_T^{(i)}$ to be similar for most i. These issues are circumvented by Algorithm 2, which is able to compute the entire threshold-indexed solution path quickly, thus facilitating the study of a data-driven approach to the choice of ζ_T in Section 3.3. The key idea of Algorithm 2 is to make use of information from $\mathcal{T}(\zeta_T^{(i)})$ to compute both $\zeta_T^{(i+1)}$ and $\mathcal{T}(\zeta_T^{(i+1)})$ iteratively for every $i = 0, \ldots, N-1$. The pseudo-code of Algorithm 2, as well as other relevant details, can be found in Section C.2 of the online supplementary materials.

3.3. Choice of ζ_T via the strengthened Schwarz Information Criterion (sSIC)

Suppose we have $\mathcal{T}(\zeta^{(1)}), \ldots, \mathcal{T}(\zeta^{(N)})$ that form the NOT solution path, i.e. the collection of candidate models produced by Algorithm 2. We propose to select $\mathcal{T}(\zeta^{(k)})$ that minimises the strengthened Schwarz Information Criterion (sSIC; Liu *et al.* (1997), Fryzlewicz (2014)) defined as follows. Let $k = 1, \ldots, N$, $\hat{q}_k = |\mathcal{T}(\zeta_T^{(k)})|$ and $\hat{\mathbf{\Theta}}_1, \ldots, \hat{\mathbf{\Theta}}_{\hat{q}_k+1}$ be the maximum likelihood estimators of the segment parameters in model (3) with the estimated change-points $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k} \in \mathcal{T}(\zeta_T^{(k)})$. Here for notational convenience, we have suppressed the dependence of $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}_k}$ on $\zeta_T^{(k)}$. Further, denote by n_k the total number of estimated parameters, including the locations of the change-points and free parameters in $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_{\hat{q}_k+1}$ (N.B. the total number of the latter can be different from the dimensionality of each $\mathbf{\Theta}_j$ multiplied by the number of segments, as e.g. in (S2)). Then the strengthened Schwarz Information Criterion (sSIC) is

$$\mathrm{sSIC}(k) = -2 \sum_{j=1}^{\hat{q}_k+1} \log \ell(Y_{\hat{\tau}_{j-1}+1}, \ldots, Y_{\hat{\tau}_j}; \hat{\mathbf{\Theta}}_j) + n_k \log^{\alpha}(T), \tag{11}$$

for some pre-given $\alpha \geq 1$, with $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{q}_k+1} = T$. When $\alpha = 1$, we recover the well-known Schwarz Information Criterion (SIC).

One reason we use the sSIC here is to facilitate our theoretical development below. In fact, once we obtain the NOT solution path via Algorithm 2, other criteria, such as MBIC (Zhang and Siegmund, 2007), Minimum Description Length (MDL; Davis *et al.* (2006)) or Steepest Drop to Low Levels(SDLL; Fryzlewicz (2018b)) could conceivably be used for model (or equivalently, threshold) selection.

3.4. Theoretical properties of NOT with the sSIC

In this section, we analyse the theoretical behaviour of NOT with the sSIC in Scenarios (S1) and (S2). Here we focus on the situation where the number of change-points q is fixed (i.e. does not increase with T). This is typical for the theoretical development of information-criteron-based approaches, and reflects the fact that such approaches tend to work better in practice for signals with at most a moderate number of change-points. See also Yao (1988). Again, for notational convenience, we set $\sigma_0 = 1$. Our results below provide theoretical justifications for using NOT with the sSIC. Crucially, in contrast to Algorithm 1, here one does not need to supply a threshold.

Theorem 3. *Suppose Y_t follow model (3) in Scenario (S1). Let $\delta_T = \min_{j=1,\dots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$ and $\underline{f}_T = \min_{j=1,\dots,q} \Delta_j^{\mathbf{f}}$. Furthermore, assume that q does not increase with T, $\delta_T/(\log T)^{\alpha'} \geq \underline{C}_1$, $\underline{f}_T \geq \underline{C}_2$ and $\max_{t=1,\dots,T} |f_t| \leq \bar{C}$ for some $\underline{C}_1, \underline{C}_2, \bar{C} > 0$ and $\alpha' > 1$. Let $\hat{q}$ and $\hat{\tau}_1, \dots, \hat{\tau}_{\hat{q}}$ denote, respectively, the number and locations of change-points, sorted in increasing order, estimated by NOT (via Algorithm 2) with the contrast function given by (6) and ζ_T picked via the sSIC using $\alpha \in (1, \alpha')$. Then there exists a constant C (not depending on T) such that given $M \geq 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1})$, as $T \to \infty$,*

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\dots,q} |\hat{\tau}_j - \tau_j| \leq C \log T\right) \to 1.$$

Theorem 4. *Suppose Y_t follow model (3) in Scenario (S2). Let $\delta_T = \min_{j=1,\dots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$, $\underline{f}_T = \min_{j=1,\dots,q} \Delta_j^{\mathbf{f}}$. Furthermore, assume that q does not increase with T, $\delta_T/T \geq \underline{C}_1$, $\underline{f}_T T \geq \underline{C}_2$ and $\max_{t=1,\dots,T} |f_t| \leq \bar{C}$ for some $\underline{C}_1, \underline{C}_2, \bar{C} > 0$. Let $\hat{q}$ and $\hat{\tau}_1, \dots, \hat{\tau}_{\hat{q}}$ denote, respectively, the number and locations of change-points, sorted in increasing order, estimated by NOT (via Algorithm 2) with the contrast function given by (8) and ζ_T picked via the sSIC using $\alpha > 1$. Then there exists a constant C (not depending on T) such that given $M \geq 36\underline{C}_1^{-2}\log(\underline{C}_1^{-1}T)$, as $T \to \infty$,*

$$\mathbb{P}\left(\hat{q} = q,\ \max_{j=1,\dots,q} |\hat{\tau}_j - \tau_j| \leq C\sqrt{T \log T}\right) \to 1.$$

For a discussion of the optimality of the rates obtained in Theorems 3 and 4 regarding the accuracy of the estimated change-point locations, see Section 2.5.

3.5. Computational complexity

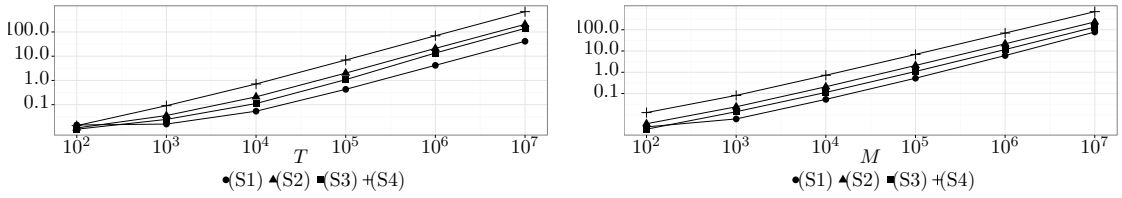
Here we elaborate on the computational complexity of Algorithm 1 (see Section 2.4) and Algorithm 2 (see Section 3.2 and Section C.2 of the online supplementary materials). For both algorithms, the task of computation can be divided into two main parts. First, we need to evaluate a chosen contrast function for all points in the M randomly picked left-open and right-closed intervals with their start- and end-points in $\{0, \dots, T-1\}$ and

$\{1, \ldots, T\}$ respectively. In the second part, we find potential locations of the change-points for a single threshold ζ_T in the case of Algorithm 1 and for all possible thresholds in the case of Algorithm 2.

Naturally, the computational complexity of the first part depends on the cost of computing the contrast function for a single interval. In all scenarios studied in this paper, this cost is linear in the length of the interval, i.e. the cost of computing $\{\mathcal{C}^b_{(s,e]}(\mathbf{Y})\}_{b=s+1}^{e-1}$ is $O(e-s)$. This is explained in detail in Section C.1 of the online supplementary materials. The intervals drawn in the procedures have approximately $O(T)$ points on average, therefore the computational complexity of the first part of the computations is $O(MT)$ in a typical application. Importantly, as the calculations for one interval are completely independent of the calculations for another, it is straightforward to run these computations in an "embarrassingly parallel" manner. In addition, for the second part, as mentioned in detail in the Section C.2 of online supplementary materials, its computational complexity is typically less than $O(MT)$, thus bringing the total computational complexity of both Algorithm 1 and Algorithm 2 to $O(MT)$.

Figure 3 shows execution times for the implementation of Algorithm 2, the NOT solution path algorithm, implemented in the R package **not**, with the data $\{Y_t\}_{t=1}^T$ being i.i.d. $\mathcal{N}(0,1)$. The running times appears to scale linearly both in T (Figure 3a) and in M (Figure 3b), which provides evidence that the computational complexity of Algorithm 2 in this particular example is practically of order $O(MT)$.

Finally, we remark that the memory complexity of Algorithm 2 is also $O(MT)$, which combined with its low computational complexity implies that our approach can handle problems of size T in the range of millions.



(a) fixed $M = 10000$ (b) fixed $T = 10000$

Fig. 3. Execution times (in seconds) for the implementation of Algorithm 2 available in R package **not** (Baranowski *et al.*, 2016b), for various feature detection problems with the data Y_t, $t = 1, \ldots, T$ being i.i.d. $\mathcal{N}(0,1)$. In a single run, computations for the input of the algorithm are performed in parallel, using 8 cores of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM. The computation times are averaged over 10 runs in each case.

3.6. Other practical considerations

3.6.1. Choice of M

As can be seen in Theorem 1 and Theorem 2, the minimum required value for M grows with T (i.e. at $O(\log T)$, for a fixed number of well-spaced change-points). In practice, when the number of observations is of the order of thousands, we would recommend setting $M = 10000$. With this value of M, the implementation of Algorithm 1 provided in the R **not** package (Baranowski *et al.*, 2016b) achieves the average computation time not longer than 2 seconds in all examples in Section 5 using a single core of an Intel Xeon 3.6 GHz

CPU. This can be accelerated further, as the **not** package allows for computing the contrast function over the intervals drawn in parallel using all available CPU cores.

However, caution must be exercised for signals with a large expected number of change-points, for which M may need to be increased. For example, Maidstone *et al.* (2017) found that NOT with $M = 10^5$ offered better practical performance on the change-point-rich signals they considered. In the most extreme scenario where one expects change-points to occur very frequently with a large T, we would recommend picking M as large as possible to match the available computational power and applying a penalty less stringent than the sSIC. See Section F of the online supplementary materials.

3.6.2. Early stopping for NOT with the sSIC

If the number of change-points in the data is expected to be rather moderate, then it may not be necessary to calculate sSIC for all k. In practice, solutions on the path corresponding to very small values of ζ_T contain many estimated change-points. Such solutions are unlikely to minimise (11). By considering $|\mathcal{T}(\zeta_T^{(k)})| \leq q_{max}$, we could achieve some computational gains without adversely impacting the overall performance of the methodology. As such, in all applications presented in this work we compute sSIC only for k such that $|\mathcal{T}(\zeta_T^{(k)})| \leq q_{max}$ with $q_{max} = 25$.

4. NOT under different noise types

In this section, we discuss how NOT can be extended to handle different noise types. Section 4.1 deals with dependent noise, while Section 4.2 covers heavy-tailed noise. In addition, we investigate the case of noise with slow-varying variance in Section D of the online supplementary materials.

4.1. NOT under dependent noise

When the errors ε_t in model (3) are dependent with $\mathbb{E}\varepsilon_t = 0$ and $\mathrm{Var}(\varepsilon_t) = 1$, the aforementioned NOT procedure can still be applied as a quasi-likelihood-type procedure. Conceivably, using NOT here would incur information loss. As is shown in Corollaries 1 and 2 in Scenarios (S1) and (S2), NOT is still consistent if we replace the noise's i.i.d. assumption in Theorems 1 and 2 by stationarity with short-memory. This new dependence assumption is satisfied by a large class of stationary time series models, including autoregressive moving average (ARMA) models. See also numerical examples in Section E of the online supplementary materials, where we again select the thresholds automatically via sSIC. Here we assume that $\sigma_0 = 1$. However, if not, MAD-type estimators based on the simple differencing are no longer appropriate for dependent data. We comment on this issue later. The following corollaries give guidelines on the choice of the threshold, as well as guarantee on the performance of NOT from a theoretical perspective.

Corollary 1. *Suppose Y_t follow model (3) in Scenario (S1), but with $\{\varepsilon_t\}$ being a stationary short-memory Gaussian process, i.e. the auto-correlation function of $\{\varepsilon_t\}$, denoted by ρ_k for any lag $k \in \mathbb{Z}$, satisfies $\sum_{k=-\infty}^{\infty} |\rho_k| < \infty$. Then, the conclusion of Theorem 1 still holds (with different constants).*

Corollary 2. *Suppose Y_t follow model (3) in Scenario (S2), but with $\{\varepsilon_t\}$ being a stationary short-memory Gaussian process. The conclusion of Theorem 2 holds (with different constants).*

In our theoretical development for the dependent noise setting, the smallest permitted threshold to be used in the NOT algorithm depends linearly on $\sigma_0(\sum_{k=-\infty}^{\infty}|\rho_k|)^{1/2}$. This quantity can also be viewed as a generalisation to the independent noise setting, where the threshold is proportional to σ_0 (since $\sum_{k=-\infty}^{\infty}|\rho_k| = 1$). More details of its derivation is provided in Section I.6 of the online supplementary materials.

This poses a few challenges in the practical application of NOT to signals with dependent noise: (i) the (pre-)estimation of the residuals ε_t in preparation for the estimation of their long-run variance; (ii) the estimation σ_0; and (iii) the estimation of $\sigma_0(\sum_{k=-\infty}^{\infty}|\rho_k|)^{1/2}$. These problems are known to be difficult in time series analysis in general. A possible solution is outlined below.

For (i), we have had some success with the wavelet-based method of Johnstone and Silverman (1997), which was implemented in R package **wavethresh** (Nason, 2016); its advantages are that it is specifically designed for dependent noise and that, being based on nonlinear wavelet shrinkage, it is particularly suited for signals with irregularities, such as (generalised) change-points. Here the Haar wavelet transform of the data is appropriate in Scenario (S1), while a transform with respect to any wavelet that annihilates linear functions is appropriate in Scenarios (S2) and (S3). Once the empirical residuals are obtained from (i), we could then estimate σ_0 in (ii) by its sample version, and estimate $\sigma_0(\sum_{k=-\infty}^{\infty}|\rho_k|)^{1/2}$ in (iii) in a model-based way (e.g. using the autoregressive model with its order p chosen by an information criterion).

Another possibility to estimate change-points under dependent noise is to use self-normalising based statistics. See, for instance, Shao and Zhang (2010), Betken (2016), Pešta and Wendler (2018) and Zhang and Lavitas (2018). These statistics could potentially be fed into our NOT approach as well.

Finally, we mention two practical ways of reducing the dependence and making the series closer to Gaussian, before applying NOT: (A) pre-average the data over non-overlapping moving windows of size h, creating a new dataset of length $\lfloor T/h \rfloor$; the hope is that by the law of large numbers, the pre-averaged noise will be closer to Gaussian and also less serially dependent than the original noise; (B) add additional i.i.d. Gaussian noise to the data, with mean zero and suitably chosen standard deviation; this will have a similar effect as previously, i.e. it will bring the distribution of the data closer to Gaussian and reduce the serial dependence within the data.

4.2. Extension of NOT under heavy-tailed noise

NOT appears to be relatively robust under noise misspecification. As is demonstrated later in Section 5, it offers reasonable estimates when the noise is non-Gaussian but the Gaussian contrast functions are used. We now discuss how its performance can be improved further in the presence of heavy-tailed noise.

In Scenario (S1), we propose to apply the following new contrast function, defined for $\mathbf{Y}$ and $0 < s < b < e < T$ as

$$\tilde{\mathcal{C}}^b_{(s,e]}(\mathbf{Y}) = \left\langle \mathcal{S}_{(s,e]}(\mathbf{Y}), \boldsymbol{\psi}^b_{(s,e]} \right\rangle \tag{12}$$

in our NOT procedure. Here for any vector $\mathbf{v} = (v_1, \ldots, v_T)'$, the i-component of $\mathcal{S}_{(s,e]}(\mathbf{v})$ is given by $\mathcal{S}_{(s,e]}(\mathbf{v})_i = \operatorname{sign}\left(v_i - (e-s)^{-1}\sum_{t=s+1}^{e} v_t\right)$ and $\boldsymbol{\psi}^b_{(s,e]}$ is defined by (5). (For certain noise distributions, subtracting the sample median of $\mathbf{v}$ instead of the sample mean would appear more appropriate.) The rationale behind (12) is to assign $Y_{s+1} -$

$\frac{1}{e-s}\sum_{t=s+1}^{e} Y_t, \ldots, Y_e - \frac{1}{e-s}\sum_{t=s+1}^{e} Y_t$ (i.e. residuals for fitting a curve with no change-point on a given interval) into two classes (± 1, i.e. a two-point distribution, thus with light tails) and apply the contrast function to their ± 1 labels. Empirical performance of NOT (via Algorithm 2) combined with (12) and sSIC is also illustrated in Section E of the online supplementary materials.

5. Simulation study

5.1. Settings

We consider examples following (S1)–(S4) introduced in Section 2.3, as well as an extra example satisfying

(S5) $\sigma_t = \sigma_0$ and f_t is a piecewise-quadratic function of t.

We simulate data according to Equation (3) using the test signals (M1) `teeth`, (M2) `blocks`, (M3) `wave1`, (M4) `wave2`, (M5) `mix`, (M6) `vol` and (M7) `quad`, with the noise following

(a) i.i.d. $\mathcal{N}(0,1)$;

(b) i.i.d. $\mathcal{N}(0,2)$;

(c) i.i.d. scaled Laplace distribution with zero-mean and unit-variance;

(d) i.i.d. scaled Student-t_5 distribution with unit-variance;

(e) a stationary Gaussian AR(1) process of $\varphi = 0.3$, with zero-mean and unit-variance.

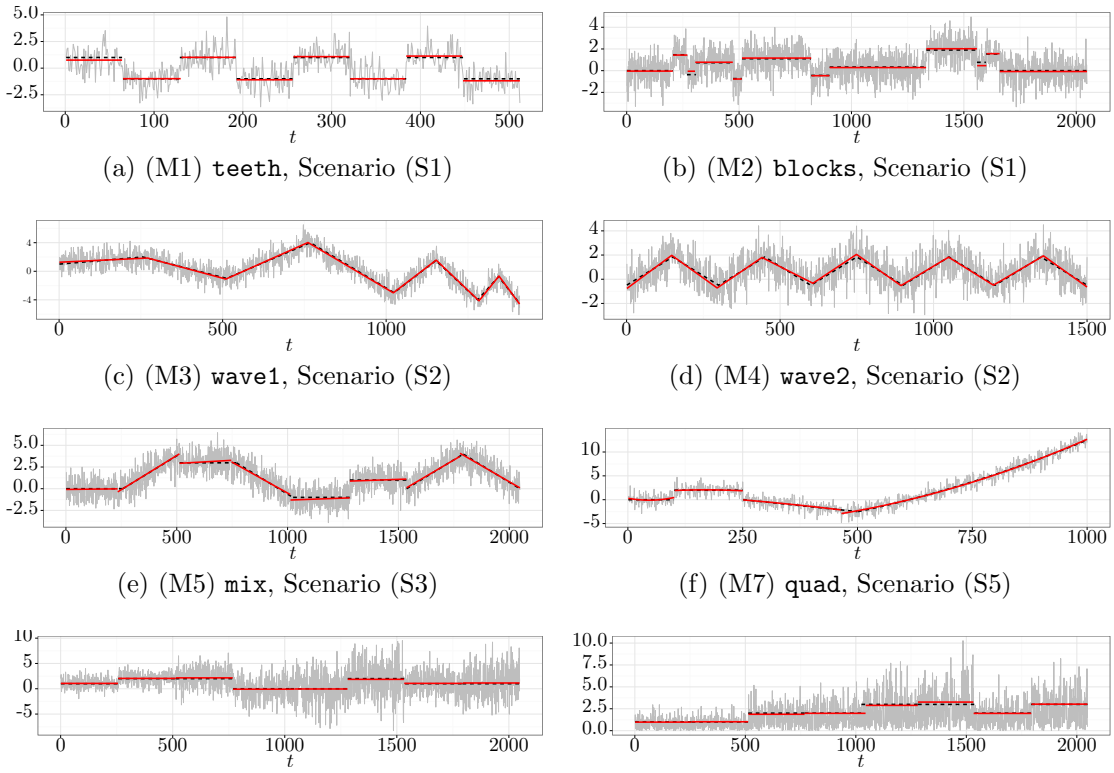
A detailed specification of our test models can be found in Section A of the online supplementary materials. Figure 4 shows the examples of the data generated from models (M1)–(M7), as well as the estimates produced by NOT in a typical run.

5.2. Estimators

We apply Algorithm 2 to compute the NOT solution path and pick the solution minimising the sSIC introduced in Section 3.3 with $\alpha = 1$ (which is equivalent to the SIC). In each simulated example, we use the contrast function designed to detect change-points in the scenario that the example follows, given in Section 2.3 and Section B of the online supplementary materials under the assumption that ε_t is i.i.d. Gaussian. The resulting method is referred to simply as 'NOT'. In addition, for Scenario (S1) only, we also apply Algorithm 2 combined with (12) and the SIC, which we call 'NOT HT'. Here 'HT' stands for 'heavy tails'. The number of intervals drawn in the procedure and the maximum number of change-points for the SIC are set to $M = 10000$ and $q_{max} = 25$, respectively.

We then compare the performance of NOT and NOT HT against the best competitors available on CRAN. To the best of our knowledge, none of the competing packages can be applied in all of Scenarios (S1)–(S5).

For change-point detection in the mean, the selected competitors from CRAN are: **changepoint** (Killick and Eckley, 2014; Killick *et al.*, 2016) implementing the PELT methodology proposed by Killick *et al.* (2012a), **changepoint.np** (Haynes *et al.*, 2016) implementing a nonparametric extension of the PELT methodology studied in Haynes *et al.* (2017), **wbs** (Baranowski and Fryzlewicz, 2015) implementing the Wild Binary Segmentation proposed by Fryzlewicz (2014), **ecp** (James and Matteson, 2014) implementing the e.cp3o method



(a) (M1) `teeth`, Scenario (S1)

(b) (M2) `blocks`, Scenario (S1)

(c) (M3) `wave1`, Scenario (S2)

(d) (M4) `wave2`, Scenario (S2)

(e) (M5) `mix`, Scenario (S3)

(f) (M7) `quad`, Scenario (S5)

(g) (M6) `vol`: f_t, Scenario (S4)

(h) (M6) `vol`: σ_t, Scenario (S4)

Fig. 4. Examples of data generated from simulation models outlined in Section A. Figure 4a– 4g: data series Y_t (thin grey), true signal f_t (dashed black), $\hat{f}_t$ being the least squares (LS) estimate of f_t with the change-points estimated by NOT (thick red). Figure 4h: centered data $|Y_t - \hat{f}_t|$ (thick grey), true standard deviation σ_t (dashed black) and the estimated standard deviation $\hat{\sigma}_t$ between the change-points detected by NOT (thick red).

proposed by James and Matteson (2015), **strucchange** (Zeileis *et al.*, 2002) implementing the methodology of Bai and Perron (2003), **Segmentor3IsBack** (Cleynen *et al.*, 2013) implementing the technique proposed by Rigaill (2015), **nmcdr** (Zou and Lancezhange, 2014) implementing the NMCD methodology of Zou *et al.* (2014), **stepR** (Pein *et al.*, 2018) implementing the SMUCE method proposed by Frick *et al.* (2014), and **FDRSeg** (Li *et al.*, 2017) implementing the FDRSeg method proposed by Li *et al.* (2016). We refer to the corresponding methods as, respectively, PELT, NP-PELT, WBS, e.cp3o, B&P, S3IB, NMCD, SMUCE and FDRSeg.

Note that e-cp3o, NMCD, NOT, PELT and NP-PELT can be also used for change-point detection in Scenario (S4), where change-points occur in the mean and variance of the data. In addition, for Scenario (S4), we also include the Heterogeneous SMUCE method (Pein *et al.*, 2017) implemented in **stepR** (Pein *et al.*, 2018), and the Segment Neighbourhoods method (Auger and Lawrence, 1989) implemented in **changepoint** (Killick and Eckley, 2014; Killick *et al.*, 2016). We refer to them as HSMUCE and SegNeigh respectively.

Only the B&P method allows for change-point detection in piecewise-linear and piecewise-quadratic signals (in particular, the WBS is not suitable for these settings as described in Sections 1 and 2.5), hence we also study the performance of the trend filtering methodology

of Kim *et al.* (2009) termed as TF hereafter, using the implementation available from the R package **genlasso** (Taylor and Tibshirani, 2014), to have a broader comparison. See also Lin *et al.* (2017). The TF method aims to estimate a piecewise-polynomial signal from the data, not focusing on the change-point detection problem directly. Let $\hat{f}_t^{(TF)}$ denote the TF estimate of the true signal f_t, then the TF estimates of the change-points in Scenario (S2) are defined as those τ for which $|2\hat{f}_\tau^{(TF)} - \hat{f}_{\tau-1}^{(TF)} - \hat{f}_{\tau+1}^{(TF)}| > \epsilon$, where $\epsilon > 0$ is a very small number being the numerical tolerance level (more precisely, we set $\epsilon = 1.11 \times 10^{-15}$ in our study). In the piecewise-quadratic case, the change-points are defined as those τ for which the third order differences $|\hat{f}_{\tau+2}^{(TF)} - 3\hat{f}_{\tau+1}^{(TF)} + 3\hat{f}_\tau^{(TF)} - \hat{f}_{\tau-1}^{(TF)}| > \epsilon$. We note that both B&P and TF require a substantial amount of computational resources in this study.

Finally, we remark that the tuning parameters for the competing methods are set to the values recommended by the corresponding R packages, and the R code for all simulations can be downloaded from our GitHub repository (Baranowski *et al.*, 2016a).

5.3. Results

Here we only present the results under the setting where the noise is (a) i.i.d. standard normal in Table 1. Additional results under the other above-mentioned noise settings can be found in Section E of the online supplementary materials.

For each method, we show a frequency table for the distribution of $\hat{q} - q$, where $\hat{q}$ is the number of the estimated change-points and q denotes the true number of change-points. We also report Monte-Carlo estimates of the Mean Squared Error of the estimated signal, given by $\mathrm{MSE} = \mathbb{E}\Big\{\frac{1}{T}\sum_{t=1}^{T}(f_t - \hat{f}_t)^2\Big\}$. For all methods but TF, $\hat{f}_t$ is calculated by finding the least squares (LS) approximation of the signal of the appropriate type depending on the true f_t, between each consecutive pair of estimated change-points. For TF, $\hat{f}_t$ used in the definition of the MSE is the penalised least squares estimate of f_t returned by the TF algorithm.

To assess the performance of each method in terms of the accuracy of the estimated locations of the change-points, we report estimates of the (scaled) Hausdorff distance

$$d_H = T^{-1}\,\mathbb{E}\max\left\{\max_{j=0,\ldots,q+1}\min_{k=0,\ldots,\hat{q}+1}|\tau_j - \hat{\tau}_k|, \max_{k=0,\ldots,\hat{q}+1}\min_{j=0,\ldots,q+1}|\hat{\tau}_k - \tau_j|\right\},$$

where $0 = \tau_0 < \tau_1 < \ldots \tau_q < \tau_{q+1} = T$ and $0 = \hat{\tau}_0 < \hat{\tau}_1 < \ldots \hat{\tau}_q < \hat{\tau}_{q+1} = T$ denote, respectively, true and estimated locations of the change-points. From the definition above, it follows that $0 \leq d_H \leq 1$. An estimator is regarded as performing well when its d_H is close to 0. However, d_H would be large when the number of change-points is under-estimated or some of the estimated change-points are far away from the real ones. In addition, we also report estimates of the inverse V-measure d_V defined as

$$d_V = 1 - \mathbb{E}\,V\Big(\{\hat{\tau}_k\}_{t=0}^{\hat{q}+1}, \{\tau_k\}_{t=0}^{q+1}\Big),$$

where $V(\cdot,\cdot)$ is the V-measure (with $\beta = 1$) proposed by Rosenberg and Hirschberg (2007) for the evaluation of segmentation. An estimator is regarded as performing well when its d_V is close to 0. More specifically, $0 \leq d_V \leq 1$, and a perfect estimator has $d_V = 0$, while $d_V = 1$ means none of the features are detected (i.e. $\hat{q} = 0$).

We find that in most of the simulated scenarios, NOT is among the most competitive methods in terms of the estimation of the number of change-points, their locations, as

Table 1. Distribution of $\hat{q}-q$ for data generated according to (3) with the noise term ε_t being i.i.d. $\mathcal{N}(0,1)$ for various choices of f_t and σ_t given in Section A of the online supplementary materials and competing methods listed in Section 5. Also, the average Mean-Square Error of the resulting estimate of the signal f_t, average Hausdorff distance d_H, average inverse V-measure d_V and average computation time in seconds using a single core of an Intel Xeon 3.6 GHz CPU with 16 GB of RAM, all calculated over 100 simulated data sets. Bold: methods with the largest empirical frequency of $\hat{q}-q=0$ or smallest average of d_H or d_V, and those within 10% of the highest or lowest accordingly.

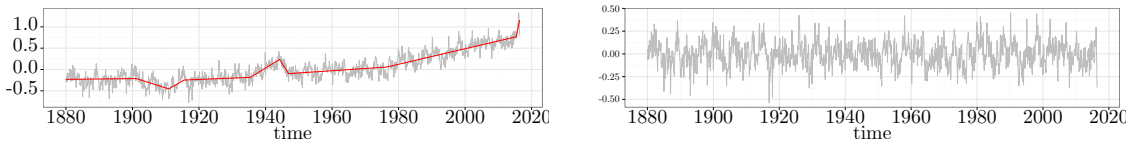
		$\hat{q}-q$										
Method	Model	≤ -3	-2	-1	0	1	2	≥ 3	MSE	$d_H \times 10^2$	d_V	time
B&P		0	0	0	**100**	0	0	0	0.053	**0.52**	**0.02**	1.367
e-cp3o		0	0	0	**100**	0	0	0	0.089	0.65	0.041	0.12
FDRSeg		0	0	0	79	17	2	2	0.089	1.26	0.044	0.092
NMCD		0	0	0	**98**	2	0	0	0.095	0.77	0.044	1.14
NOT		0	0	0	**97**	3	0	0	0.055	0.6	**0.021**	0.047
NOT HT	(M1)	0	0	0	**97**	3	0	0	0.057	0.67	**0.022**	0.06
NP-PELT		0	0	0	82	18	0	0	0.071	0.91	0.029	0.017
PELT		0	0	0	**100**	0	0	0	0.054	**0.52**	**0.02**	0.002
S3IB		0	0	0	88	10	1	1	0.057	0.8	**0.022**	0.092
SMUCE		0	0	0	**100**	0	0	0	0.085	0.58	0.039	0.047
WBS		0	0	0	**93**	7	0	0	0.057	0.69	**0.021**	0.077
B&P		100	0	0	0	0	0	0	0.127	5.85	0.128	29.897
e-cp3o		100	0	0	0	0	0	0	0.197	7.1	0.132	2.057
FDRSeg		0	1	30	**59**	6	3	1	0.029	**1.31**	0.031	1.784
NMCD		0	15	53	32	0	0	0	0.034	2.07	0.036	4.313
NOT		0	4	43	50	3	0	0	0.025	**1.44**	**0.025**	0.079
NOT HT	(M2)	2	8	44	40	6	0	0	0.031	2.05	0.033	0.141
NP-PELT		0	2	13	**58**	17	7	3	0.028	1.58	0.031	0.219
PELT		5	36	48	11	0	0	0	0.032	3	0.035	0.004
S3IB		0	5	34	**59**	1	1	0	0.024	**1.31**	**0.024**	0.318
SMUCE		55	41	4	0	0	0	0	0.069	3.38	0.061	0.018
WBS		0	4	35	53	8	0	0	0.026	**1.35**	**0.025**	0.14
B&P		0	0	100	0	0	0	0	0.218	3.7	0.133	53.978
NOT	(M3)	0	0	0	**99**	1	0	0	0.015	**0.98**	**0.053**	0.38
TF		0	0	0	0	0	0	100	0.017	8.38	0.211	46.489
B&P		0	0	4	**96**	0	0	0	0.063	2.53	0.132	61.631
NOT	(M4)	0	0	0	**100**	0	0	0	0.016	**1.15**	**0.07**	0.399
TF		0	0	0	0	0	0	100	0.016	4.49	0.146	47.794
B&P		0	0	0	**100**	0	0	0	0.021	**2.44**	**0.088**	117.894
NOT	(M5)	0	0	0	**99**	1	0	0	0.021	**2.49**	**0.088**	0.352
TF		0	0	0	0	0	0	100	0.027	6	0.26	58.816
e-cp3o		11	12	12	33	20	5	7	0.145	6.91	0.164	1.707
HSMUCE		97	3	0	0	0	0	0	0.091	12.77	0.209	0.049
NMCD		0	0	18	70	12	0	0	0.06	4.04	0.068	4.206
NOT	(M6)	0	0	13	**85**	2	0	0	0.047	2.6	**0.048**	0.454
NP-PELT		0	0	1	19	26	24	30	0.126	3.17	0.068	0.276
PELT		9	18	31	37	5	0	0	0.069	8.17	0.087	0.011
SegNeigh		0	0	3	49	36	10	2	0.053	**1.98**	**0.048**	17.237
B&P		0	0	0	**100**	0	0	0	0.024	**1.98**	**0.068**	28.93
NOT	(M7)	0	0	0	**100**	0	0	0	0.023	**1.87**	**0.065**	0.245
TF		0	0	0	0	0	0	100	0.052	23.29	0.442	42.717

well as the true signal. Importantly, it is very fast to compute, which gives it a particular advantage over its competitors in Scenarios (S2), (S3) and (S5). Finally, NOT with the contrast function derived under the assumption that the noise is i.i.d. Gaussian is relatively robust against the misspecification in ε_t, when the truth is either correlated or heavy-tailed.

6. Real data analysis

6.1. Temperature anomalies

We analyse the GISS Surface Temperature anomalies data set available from GISTEMP Team (2016), consisting of monthly global surface temperature anomalies recorded from January 1880 to June 2016. The anomaly here is defined as the difference between the average global temperature in a given month and the baseline value, being the average calculated for that time of the year over the 30-year period from 1951 to 1980; for more details see Hansen *et al.* (2010). This and similar anomalies series are frequently studied in the literature with a particular focus on identifying change-points in the data, see e.g. Ruggieri (2013) or James and Matteson (2015).



(a) data (b) residuals

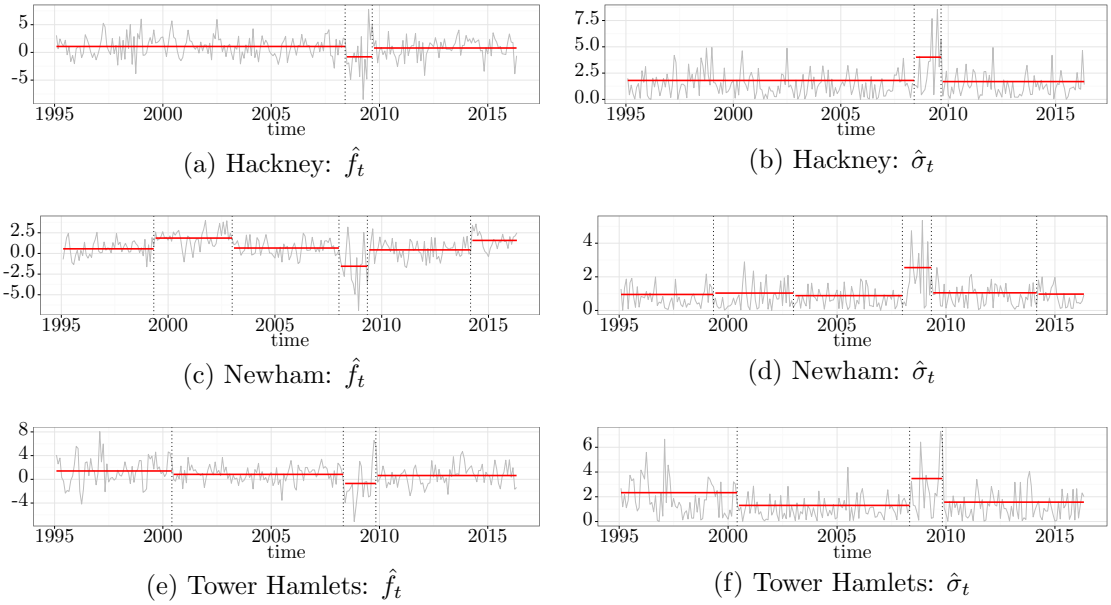
Fig. 5. Change-point analysis for the GISSTEMP data set introduced in Section 6.1. Figure 5a: the data series Y_t (thin grey) and $\hat{f}_t$ estimated using change-points returned by NOT (thick red). Figure 5b: residuals $\hat{\varepsilon}_t = Y_t - \hat{f}_t$.

The plot of the data (Figure 5a) indicates the presence of a linear trend with several change-points in the temperature anomalies series. The corresponding changes are not abrupt, therefore we believe that Scenario (S2) with change-points in the slope of the trend is the most appropriate here. To detect the locations of the change-points, we apply NOT (via Algorithm 2) with the contrast given by (8), combined with the SIC to determine the best model on the solution path.

The NOT estimate of the piecewise-linear trend and the corresponding empirical residuals are shown in Figure 5. We identify 8 change-points located at the following dates: March 1901, December 1910, July 1915, June 1935, April 1944, December 1946, June 1976 and May 2015. Previous studies conducted on similar temperature anomalies series (observed at a yearly frequency and obtained from a different source), report change-points around 1910, 1945 and 1976 (see Ruggieri (2013) for an overview of a number of related analyses). In addition to the change-points around these dates, NOT identifies two periods, 1901–1915 and 1935–1946, with local deviations from the baseline. We also observe a long-lasting upward trend in the anomalies series starting in December 1946. Finally, NOT indicates that the slope of the trend is increasing, with the most recent change-point in May 2015.

6.2. UK House Price Index

We analyse monthly percentage changes in the UK House Price Index (HPI), which provides an overall estimate of the changes in house prices across the UK. The data and a detailed description of how the index is calculated are available online from UK Land Registry (2016). Fryzlewicz (2018a), who proposes a method for signal estimation and change-point detection in Scenario (S1), used this data set to illustrate the performance of his methodology. We perform a similar analysis, assuming the more flexible Scenario (S4), allowing for changes both in the mean and the variance, which, we argue, leads to additional insights and better-interpretable estimates for this dataset.



(a) Hackney: $\hat{f}_t$ (b) Hackney: $\hat{\sigma}_t$

(c) Newham: $\hat{f}_t$ (d) Newham: $\hat{\sigma}_t$

(e) Tower Hamlets: $\hat{f}_t$ (f) Tower Hamlets: $\hat{\sigma}_t$

Fig. 6. Change-point analysis for the monthly percentage changes in the UK House Price Index from January 1995 to May 2016. Figure 6a, 6c and 6e: the monthly percentage changes Y_t and the fitted piecewise-constant mean $\hat{f}_t$, between the change-points estimated with NOT. Figure 6b, 6d and 6f: $|Y_t - \hat{f}_t|$ and the fitted piecewise-constant standard deviation $\hat{\sigma}_t$, between the change-points estimated with NOT.

As in Fryzlewicz (2018a), we analyse the percentage changes in the HPI for three London boroughs, namely Hackney, Newham and Tower Hamlets, all of which are located in East London. Hackney and Tower of Hamlets border on the City of London, a major business and financial district, with the latter being home to Canary Wharf, another important financial centre. On the other hand, Newham, located to the east of Hackney and Tower Hamlets, hosted the London 2012 Olympic Games which involved large-scale investment in that borough.

Figure 6 shows monthly percentage changes in HPI for the analysed boroughs and the corresponding NOT estimates, obtained using the contrast function for Scenario (S4). As recommended in Section 3.3, we set the number of intervals drawn in the procedure to $M = 10000$ and choose the threshold that minimises the SIC. For better comparability, NOT is applied with the same random seed for each data series.

In contrast to Fryzlewicz (2018a), whose TGUH method estimates at least 10 change-points in each HPI series, we detect just a few change-points in the data, facilitating

the interpretation of the results. Furthermore, for all three boroughs, NOT estimates two change-points (one around March 2008 and one around September 2009) that could possibly be linked to the 2008–2009 financial crisis and its impact on the housing market. Estimated standard deviations for that period are much larger than the estimates corresponding to the other segments of piecewise-constancy, suggesting that the market is more volatile during 2008–2009, and thus in this example Scenario (S4) may be more relevant than (S1) considered in Fryzlewicz (2018a).

Acknowledgements

We thank Paul Fearnhead for his helpful comments on an earlier draft, and on the implementation of our R package. We also thank the associate editor and four anonymous referees for their comments and suggestions. Piotr Fryzlewicz's work was supported by the Engineering and Physical Sciences Research Council grant No. EP/L014246/1.

References

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, **51**, 39–54.

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. and Fryzlewicz, P. (2015). wbs: Wild binary segmentation for multiple change-point detection. URL `https://CRAN.R-project.org/package=wbs`. R package v1.3.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016a). Narrowest-over-threshold detection of multiple change-points and change-point-like features: Simulation code. `https://github.com/rbaranowski/not-num-ex`.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016b). not: Narrowest-over-threshold change-point detection. URL `https://cran.r-project.org/web/packages/not`. R package v1.0.

Betken, A. (2016). Testing for changePoints in longrange dependent time series by means of a selfnormalized Wilcoxon test. *Journal of Time Series Analysis*, **37**, 785–809.

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

Cleynen, A., Rigaill, G. and Koskas, M. (2013). Segmentor3isback: A fast segmentation algorithm. URL `https://CRAN.R-project.org/package=Segmentor3IsBack`. R package v1.8.

Davis, R. A., Lee, T. C. M and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101**, 223–239.

Fang, X., Li, J. and Siegmund, D. (2016). Segmentation and estimation of change-point models. *arXiv preprint arXiv:1608.03032*.

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

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

Fryzlewicz, P. (2018a). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Annals of Statistics*, **46**, 3390–3421.

Fryzlewicz, P. (2018b). Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Preprint*, URL `http://stats.lse.ac.uk/fryzlewicz/wbs2/wbs2.pdf`.

GISTEMP Team. GISS Surface Temperature Analysis (GISTEMP). (2016). `http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/`.

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

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, **48**, 1–29.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics and Data Analysis*, **37**, 323–341.

Haynes, K., Fearnhead, P. and Eckley, I. A. (2016). changepoint.np: Methods for nonparametric changepoint detection. URL `https://CRAN.R-project.org/package=changepoint.np`. R package v0.0.2.

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

James, N. A. and Matteson, D. S. (2014). ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, **62**, 1–25.

James, N. A. and Matteson, D. S. (2015). Change points via probabilistically pruned objectives. *arXiv preprint arXiv:1505.04302*.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, **59**, 319–359.

Killick, R. and Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, **58**, 1–19.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012a). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, 1590–1598.

Killick, R., Haynes, K. and Eckley, I. A. (2016). changepoint: Methods for changepoint detection. URL `http://CRAN.R-project.org/package=changepoint`. R package v2.2.2

Killick, R., Nam, C., Aston, J. and Eckley, I. A. (2012b). The changepoint repository. URL `http://changepoint.info/`.

Kim, S.-J., Koh, K. Boyd, S. and Gorinevsky, D. (2009). L1 trend filtering. *SIAM Review*, **51**, 339–360.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Procesing*, **85**, 1501–1510.

Lee, C.-B. (1997). Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics*, **24**, 201–210.

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

Li, H., Sieling, H. and Aspelmeier, T. (2017). FDRSeg: FDR-Control in Multiscale Change-Point Segmentation URL `https://CRAN.R-project.org/package=FDRSeg`. R package v1.0-3.

Lin, K., Sharpnack, J., Rinaldo, A. and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. *Neural Information Processing Systems*.

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. (2017). Detecting changes in slope with an L_0 penalty *arXiv preprint arXiv:1701.01672*.

Nason, G. (2016). wavethresh: wavelet statistics and transforms. URL `http://CRAN.R-project.org/package=wavethresh`. R package v4.6.8.

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

Pein, F., Hotz, T., Sieling, H. and Aspelmeier, T. (2018). stepR: Fitting step-functions. URL `http://CRAN.R-project.org/package=stepR`. R package v2.0-2.

Pein, F., Sieling, H. and Munk, A. (2017). Heterogeneous change point inference. *Journal of the Royal Statistical Society, Series B*, **79**, 1207–1227.

Pešta, M and Wendler, M. (2018). Nuisance parameters free changepoint detection in non-stationary series. *arXiv preprint arXiv:1808.01905*.

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.

Rosenberg, A. and Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 410–420.

Ruggieri, E. (2013). A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, **33**, 520–528.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics*, **19**, 175–190.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association*, **105**, 1228–1240.

Taylor, A. B. and Tibshirani, R. J. (2014). genlasso: Path algorithm for generalized lasso problems. URL `https://CRAN.R-project.org/package=genlasso`. R package v1.3.

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

UK Land Registry. UK house price index. (2016). URL `http://landregistry.data.gov.uk/app/ukhpi`.

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

Vostrikova, L. (1981). Detection of the disorder in multidimensional random processes. *Soviet Mathematics - Doklady*, **259**, 270–274.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, **82**, 385-397.

Xia, Z. and Qiu, P. (2015). Jump information criterion for statistical inference in estimating discontinuous curves. *Biometrika*, **102**, 397–408.

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

Yao, Y.-C. and Au, S. T. (1989) Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics*, **51**, 370–381.

Zeileis, A., Leisch, F., Hornik, K. and Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, **7**, 1–38.

Zhang, N.-R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.

Zhang, T. and Lavitas, L. (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association*, **113**, 637–648.

Zou, C. and Lancezhange. (2014). nmcdr: Non-parametric multiple change-points detection. URL `https://CRAN.R-project.org/package=nmcdr`. R package v0.3.0.

Zou, C., Yin, G., Feng, L. and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics*, **42**, 970–1002.