ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 3, *pp.* 649–672

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, UK

[Received January 2017. Final revision April 2019]

Summary. We propose a new, generic and flexible methodology for non-parametric 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 generalized change points. Our methodology works with only minor modifications across a range of generalized change point scenarios, and we achieve such a high degree of generality by proposing and using a new multiple generalized change point detection device, termed narrowest-over-threshold (NOT) detection. The key ingredient of the NOT method 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 the NOT algorithm in detecting the number and locations of generalized 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

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

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 mention only a selection. For Gaussian noise ε_t, both non-penalized and penalized least squares approaches were considered by Yao and Au (1989). For specific choices of penalty functions, see for example Yao (1988), Lavielle (2005) and Davis *et al.* (2006). The Gaussianity assumption on ε_t was relaxed to exponential family distributions in Lee (1997), Hawkins (2001) and Frick *et al.* (2014). In particular, Frick *et al.* (2014) also provided 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,

Address for correspondence: Yining Chen, Department of Statistics, London School of Economics and Political Science, Columbia House, Houghton Street, London, WC2A 2AE, UK.
E-mail: y.chen101@lse.ac.uk

1369–7412/19/81649

such as the pruned exact linear time (PELT) method (Killick, Fearnhead and Eckley, 2012), 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 non-parametric version of the PELT method was 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 circular binary segmentation (CBS) (Olshen *et al.*, 2004) and wild binary segmentation (WBS) (Fryzlewicz, 2014). A selection of publications and software can be found in the on-line repository *changepoint.info* maintained by Killick, Nam, Aston and Eckley (2012).

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, for example, Wang (1995) and Xia and Qiu (2015), it proceeds by estimating the locations of features non-parametrically 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 shall provide a precise description of the type of features that we are interested in later. Moving beyond f_t only, our approach will also permit the detection of similar features in some distributional aspects of ε_t, e.g. in its variance. Since all types of features that 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 shall be referring to change point detection in the piecewise constant model as the 'canonical' change point problem, whereas our general feature detection problem will sometimes be referred to as a 'generalized' 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 $(Y_{s+1}, \ldots, Y_e)'$, where $0 \leqslant s < e \leqslant T$. 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 that 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. Among the subsamples retained, we search for the subsample that is 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 interval selected. This key observation gives our methodology a general character and enables 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 methodology proposed 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, we 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 that we consider theoretically, our procedure yields nearly 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 that is also present in the WBS method of Fryzlewicz (2014). However, we emphasize that the way that 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 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 detection 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.

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

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

If we approximate f_t by using a piecewise linear signal with only one change point in its derivative, then the best approximation (in terms of minimizing the l_2-distance) will result in an estimated change point at $t = 500$, which is away from the true change points at $t = 350$ and $t = 650$, as is illustrated in Fig. 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. In contrast, NOT detection 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 detection 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 *sufficiently narrow to contain only one change point each*, but sufficiently wide 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

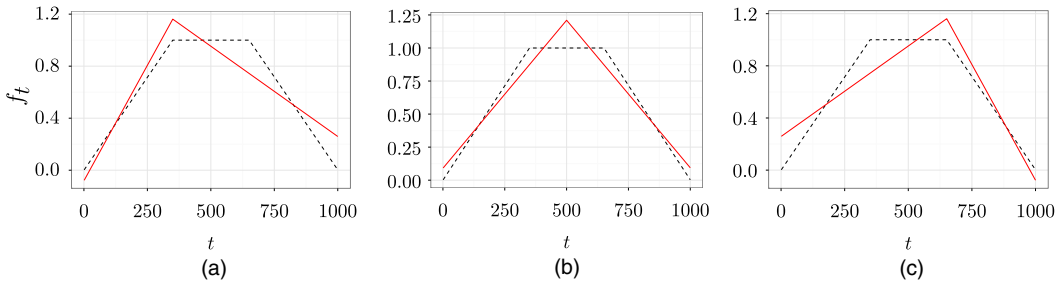

Fig. 1. Best l_2-approximation (———) of the true signal (- - - - -) via a triangular signal with a single change point, the location of which is fixed at (a) the left change point, (b) halfway between the true change points and (c) at the right change point (approximation errors are given in terms of squared l_2-distance): (a) $\tau = 350$, error = 15.0; (b) $\tau = 500$, error = 6.3; (c) $\tau = 651$, error = 15.0

error) if only one change point is present in the data subsample that is considered, unlike in the situation that was 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 on-line supplementary materials.

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 by using a piecewise constant function with only one change point, the change point of the fitted signal will always be among the true change points (Venkatraman, 1992). Since the latter property does not hold in most generalized change point detection problems, this highlights the need for new methods with better localization of the feature of interest, such as our NOT algorithm. Fang *et al.* (2019) independently considered 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 detection and its most valuable contribution.

The remainder of this paper is organized as follows. In Section 2, we give a mathematical description of the NOT algorithm. In particular, we consider the NOT approach 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 that is derived from the generalized likelihood ratio (GLR). Theoretical properties of the NOT algorithm, such as its consistency and convergence rates are also provided. In Section 3, we propose to use the NOT method with the strengthened Schwarz information criterion sSIC and discuss its computational aspects and theoretical properties. Section 4 discusses possible extensions of the NOT method. A comprehensive simulation study is carried out in Section 5, where we compare NOT with the state of the 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 the NOT method and further discussion on model misspecification, as well as additional simulations and a real data example, can be found in the on-line supplementary materials.

2. The narrowest-over-threshold framework

2.1. Set-up

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

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

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 $\varepsilon_t \sim^{\text{IID}} \mathcal{N}(0,1)$. In Section 4, we extend our framework to other types of noise.

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 prespecified 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 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 what follows, 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 d or greater for the purpose of identifiability. (Otherwise, for example, 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 assumptions hold inside the jth segment for each $j = 1, \ldots, q+1$.

(a) *Constant variance, piecewise constant mean* (scenario 1): $\sigma_t = \sigma_0$ and $f_t = \theta_j$ for $t = \tau_{j-1}+1, \ldots, \tau_j$.
(b) *Constant variance, continuous and piecewise linear mean* (scenario 2): $\sigma_t = \sigma_0$ 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 $j = 1, \ldots, q$.
(c) *Constant variance, piecewise linear (but not necessarily continuous) mean* (scenario 3): $\sigma_t = \sigma_0$ 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 $j = 1, \ldots, q$.
(d) *Piecewise constant variance, piecewise constant mean* (scenario 4): $f_t = \theta_{j,1}$ and $\sigma_t = \theta_{j,2} > 0$ for $t = \tau_{j-1}+1, \ldots, \tau_j$.

Since σ_0 in scenarios 1–3 acts as a nuisance parameter, in the rest of this paper, for simplicity we assume that its value is known. If it is unknown, then it can be estimated accurately by using the median absolute deviation (MAD) method (Hampel, 1974). More specifically, with independent and identically distributed (IID) Gaussian errors, the MAD estimator of σ_0 is defined as $\hat{\sigma} = \text{median}(|Y_2 - Y_1|, \ldots, |Y_T - Y_{T-1}|)/\{\Phi^{-1}(\frac{3}{4})\sqrt{2}\}$ in scenario 1, and as $\hat{\sigma} = \text{median}(|Y_1 - 2Y_2 + Y_3|, \ldots, |Y_{T-2} - 2Y_{T-1} + Y_T|)/\{\Phi^{-1}(\frac{3}{4})\sqrt{6}\}$ in scenarios 2 and 3. Here $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. Note that the MAD estimator is robust to any change points in the underlying signal f_t, because of 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 the NOT method formally; more details can be found in Section 2.4, where the pseudocode 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 $(Y_{s+1}, \ldots, Y_e)'$, where (s, e) is drawn uniformly from the set of pairs of indices

in $\{0,\dots,T-1\}\times\{1,\dots,T\}$ that satisfy $0\leqslant s<e\leqslant T$. Let $l(Y_{s+1},\dots,Y_e;\boldsymbol{\Theta})$ be the likelihood of $\boldsymbol{\Theta}$ given $(Y_{s+1},\dots,Y_e)'$. We then compute the GLR statistic for all potential single change points within the subsample and pick the maximum, i.e.

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

$$\mathcal{R}_{(s,e]}(\mathbf{Y})=\max_{b\in\{s+d,\dots,e-d\}}\mathcal{R}^b_{(s,e]}(\mathbf{Y}).$$

Here we also implicitly require $e-s\geqslant 2d$, 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 scenario 2), the supremum in the numerator of equation (4) is taken over the set that contains only elements of form $\boldsymbol{\Theta}^1\times\boldsymbol{\Theta}^2$ satisfying these constraints. Otherwise, as in scenarios 1, 3 and 4, equation (4) can be simplified to

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

This 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 $\mathcal{R}_{(s_m,e_m]}(\mathbf{Y})$s, we pick the one corresponding to the interval $(s_{m^*},e_{m^*}]$ that has the smallest length. Once a change point has been found in $(s_{m^*},e_{m^*}]$ (i.e. b^* that maximizes $\mathcal{R}^b_{(s_{m^*},e_{m^*}]}(\mathbf{Y})$: 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. In each recursive step, we could reuse the previously drawn intervals, provided that they fall entirely within each current subsegment considered.

After the process of estimating the change points has been completed, we can estimate the signals within each segment by 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 detection 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 that was also made in Fryzlewicz (2014).

2.3. Log-likelihood ratios and contrast functions

In many applications, the GLR (4) in NOT detection 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 $0 \leqslant s < e \leqslant T$, our aim is to find $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ such that

(a) $\arg\max_b \mathcal{C}^b_{(s,e]}(\mathbf{Y}) = \arg\max_b \mathcal{R}^b_{(s,e]}(\mathbf{Y})$,
(b) heuristically speaking, the value of $\mathcal{C}^b_{(s,e]}(\mathbf{Y})$ is relatively small if there is no change point in $(s, e]$ and
(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 what follows, we give the contrast functions corresponding to scenarios 1 and 2, where the aforementioned properties are satisfied. Their details under scenarios 3 and 4, as well as a comprehesive discussion on the construction, can be found in section B of the on-line supplementary materials. We note that this approach recovers the cumulative sum statistic in scenario 1, which is popular in this canonical change point detection setting. One can view the resulting statistics as generalizations of cumulative sum statistics under other scenarios.

2.3.1. Scenario 1

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

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

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 various (s, e, b) are shown in Fig. 2(a).

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

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

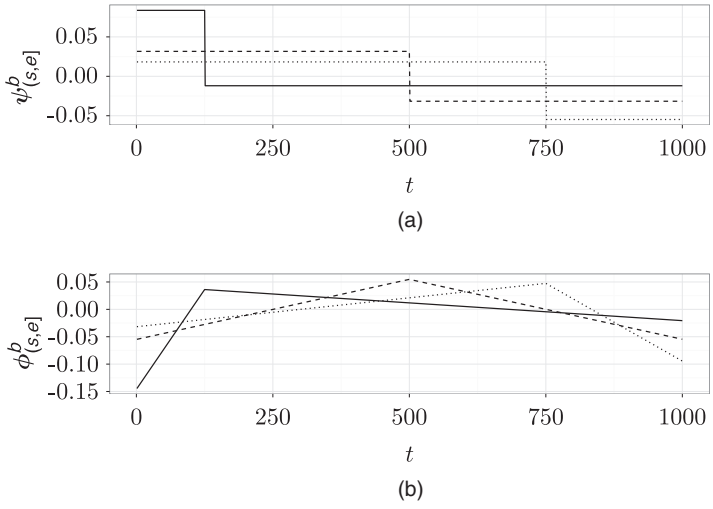

Fig. 2. Plots of (a) $\psi^b_{(s,e]}$ and (b) $\phi^b_{(s,e]}$ given by respectively equation (5) and equation (7) for $s = 0$, $e = 1000$ and several values of b: ———, $b = 125$; - - - - -, $b = 500$; · · · · · · ·, $b = 750$

2.3.2. Scenario 2

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

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

where

$$\alpha^b_{s,e} = \left[\frac{6}{l(l^2-1)\{1+(e-b+1)(b-s)+(e-b)(b-s-1)\}}\right]^{1/2},$$

$$\beta^b_{s,e} = \left\{\frac{(e-b+1)(e-b)}{(b-s-1)(b-s)}\right\}^{1/2}$$

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 Fig. 2(b). The contrast function is then defined as

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

2.4. The narrowest-over-threshold algorithm

Here we present the pseudocode of a generic version of the NOT algorithm. The main ingredient of the NOT procedure is a contrast function $C^b_{(s,e]}(\cdot)$, which is chosen by the user, depending on the assumed nature of change points in the data, e.g. as exemplified by our scenarios 1 and 2 above, and scenarios 3 and 4 in section B of the on-line supplementary materials. In addition, some tuning parameters are needed: $\zeta_T > 0$ is the threshold with respect to which the contrast should be tested, whereas M is the number of the intervals that are 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 the NOT algorithm with an information-based criterion, thus making our procedure threshold free.

To sum up, the input includes the data vector $\mathbf{Y}$, the set of F^M_T that contains all randomly drawn subintervals for testing and the global variable $\mathcal{S}$ for the set of estimated change points initialized with $\mathcal{S} = \emptyset$. Then the NOT algorithm 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 (Table 1). In this way, we are better able to control the computational complexity of the entire procedure.

2.5. Theoretical properties of narrowest-over-threshold method

In this section, we analyse the theoretical behaviour of the NOT algorithm in scenarios 1 and 2. We use infill asymptotics, which are 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 strategy that we use for these two scenarios.

First, we revisit the canonical change point detection problem, scenario 1, 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, which was described in Section 2.1, without affecting the validity

Table 1. Algorithm 1—NOT algorithm

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

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

Theorem 1. Suppose that Y_t follow model (3) in scenario 1. 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 equation (6). Then there are constants $\underline{C}$, $C_1, C_2, C_3 > 0$ (not depending on T) such that, given $\delta_T^{1/2} \underline{f}_T \geqslant \underline{C}\sqrt{\log(T)}$, $C_1\sqrt{\log(T)} \leqslant \zeta_T < C_2 \delta_T^{1/2} \underline{f}_T$ and $M \geqslant 36 T^2 \delta_T^{-2} \log(T^2 \delta_T^{-1})$, as $T \to \infty$,

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

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 $\delta_T^{1/2} \underline{f}_T \geqslant \underline{C}\sqrt{\log(T)}$ 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, $\max_{j=1,\ldots,q}(|\hat{\tau}_j - \tau_j|) = O_p\{\log(T)\}$, which trails the minimax rate of $O_p(1)$ by only a logarithmic factor. In addition, we note that the NOT procedure allows for $\delta_T^{1/2} \underline{f}_T$, which is a quantity that characterizes the level of difficulty 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 2, 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 and C_3) can be found in section I.3 of the on-line supplementary materials.

Theorem 2. Suppose that Y_t follow model (3) in scenario 2. Let $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j - 1} - f_{\tau_j + 1}|$, $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^{\mathbf{f}}$. 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 equation (8). Then there are constants $\underline{C}$, $C_1, C_2, C_3 > 0$ (not depending on T) such that, given $\delta_T^{3/2} \underline{f}_T \geqslant \underline{C}\sqrt{\log(T)}$, $C_1\sqrt{\log(T)} \leqslant \zeta_T < C_2 \delta_T^{3/2} \underline{f}_T$ and $M \geqslant 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1})$, as $T \to \infty$,

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

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 the NOT algorithm is $O_p\{T^{2/3}\log(T)^{1/3}\}$, which is different from the rate of $O_p(T^{2/3})$ that was 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 on-line supplementary materials for more details. Furthermore, we remark that, in more general cases (i.e. the number of change points increasing with T) in scenario 2, the level of difficulty of the problem in scenario 2 can be characterized by $\delta_T^{3/2}\underline{f}_T$, which is a quantity that is analogous to $\delta_T^{1/2}\underline{f}_T$ in the setting of scenario 1.

Both theorem 1 and theorem 2 imply that there is 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 emphasize again that WBS will fail to estimate change points consistently in scenario 2, for reasons that were described in Section 1.

3. Narrowest-over-threshold method with the strengthened Schwarz information criterion

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 are 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.

For a given $\mathbf{Y}$ and F_T^M, each threshold ζ_T corresponds to a candidate model produced by the NOT algorithm. Therefore, if we could produce a 'solution path' of candidate models obtained from the NOT algorithm along all possible thresholds, we could then try to select the best model along the solution path via minimizing 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 narrowest-over-threshold 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 $\{\mathcal{T}(\zeta_T)\}_{\zeta_T \geqslant 0}$. 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 are $0 = \zeta_T^{(0)} < \zeta_T^{(1)} < \ldots < \zeta_T^{(N)}$, 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 $\zeta_T \in [\zeta_T^{(i)}, \zeta_T^{(i+1)})$; second, $\mathcal{T}(\zeta_T) = \emptyset$ for any $\zeta_T \geqslant \zeta_T^{(N)}$.

However, the thresholds $\zeta_T^{(i)}$ are unknown and depend on the data; therefore naively applying algorithm 1 on a range of prespecified 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 we 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 can 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 pseudocode of algorithm 2, as well as other relevant details, can be found in section C.2 of the on-line supplementary materials.

3.3. Choice of ζ_T via the strengthened Schwarz information criterion

Suppose that we have $\mathcal{T}(\zeta^{(1)}), \ldots, \mathcal{T}(\zeta^{(N)})$ that form the NOT solution path, i.e. the collection of candidate models that is produced by algorithm 2. We propose to select $\mathcal{T}(\zeta^{(k)})$ that minimizes 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 $\hat{\boldsymbol{\Theta}}_1, \ldots, \hat{\boldsymbol{\Theta}}_{\hat{q}_k+1}$ 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 $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{\hat{q}_k+1}$ (note that the total number of the latter can be different from the dimensionality of each $\boldsymbol{\Theta}_j$ multiplied by the number of segments, as for example in scenario 2). Then the strengthened Schwarz information criterion is

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

for some pregiven $\alpha \geqslant 1$, with $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{q}_k+1} = T$. When $\alpha = 1$, we recover the well-known Schwarz information criterion.

One reason why we use sSIC here is to facilitate our theoretical development below. In fact, once we have obtained the NOT solution path via algorithm 2, other criteria, such as the modified Bayes information criterion (Zhang and Siegmund, 2007), the minimum description length (Davis *et al.*, 2006) or the steepest drop to low levels (Fryzlewicz, 2018a), could conceivably be used for model (or, equivalently, threshold) selection.

3.4. Theoretical properties of narrowest-over-threshold method with the strengthened Schwarz information criterion

In this section, we analyse the theoretical behaviour of the NOT algorithm with sSIC in scenarios 1 and 2. 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-criterion-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 the NOT

algorithm with sSIC. Crucially, in contrast with algorithm 1, here we do not need to supply a threshold.

Theorem 3. Suppose that Y_t follow model (3) in scenario 1. Let $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |f_{\tau_j+1} - f_{\tau_j}|$ and $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^{\mathbf{f}}$. Furthermore, assume that q does not increase with T, $\delta_T / \log(T)^{\alpha'} \geqslant \underline{C}_1$, $\underline{f}_T \geqslant \underline{C}_2$ and $\max_{t=1,\ldots,T} |f_t| \leqslant \bar{C}$ for some $\underline{C}_1, \underline{C}_2, \bar{C} > 0$ and $\alpha' > 1$. 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 the NOT algorithm (via algorithm 2) with the contrast function given by equation (6) and ζ_T picked via sSIC using $\alpha \in (1, \alpha')$. Then there is a constant C (not depending on T) such that, given $M \geqslant 36T^2\delta_T^{-2}\log(T^2\delta_T^{-1})$, as $T \to \infty$,

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

Theorem 4. Suppose that Y_t follow model (3) in scenario 2. Let $\delta_T = \min_{j=1,\ldots,q+1}(\tau_j - \tau_{j-1})$, $\Delta_j^{\mathbf{f}} = |2f_{\tau_j} - f_{\tau_j-1} - f_{\tau_j+1}|$, $\underline{f}_T = \min_{j=1,\ldots,q} \Delta_j^{\mathbf{f}}$. Furthermore, assume that q does not increase with T, $\delta_T / T \geqslant \underline{C}_1$, $\underline{f}_T T \geqslant \underline{C}_2$ and $\max_{t=1,\ldots,T} |f_t| \leqslant \bar{C}$ for some $\underline{C}_1, \underline{C}_2, \bar{C} > 0$. 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 the NOT algorithm (via algorithm 2) with the contrast function given by equation (8) and ζ_T picked via sSIC using $\alpha > 1$. Then there is a constant C (not depending on T) such that, given $M \geqslant 36\underline{C}_1^{-2}\log(\underline{C}_1^{-1}T)$, as $T \to \infty$,

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

For a discussion of the optimality of the rates that are 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 on-line 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 $\{0, \ldots, T-1\}$ 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 the scenarios that are studied in this paper, this cost is linear in the length of the interval, i.e. the cost of computing $\{C^b_{(s,e]}(\mathbf{Y})\}_{b=s+1}^{e-1}$ is $O(e-s)$. This is explained in detail in section C.1 of the on-line 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 the on-line 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 $O(MT)$.

Fig. 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 IID $\mathcal{N}(0,1)$. The

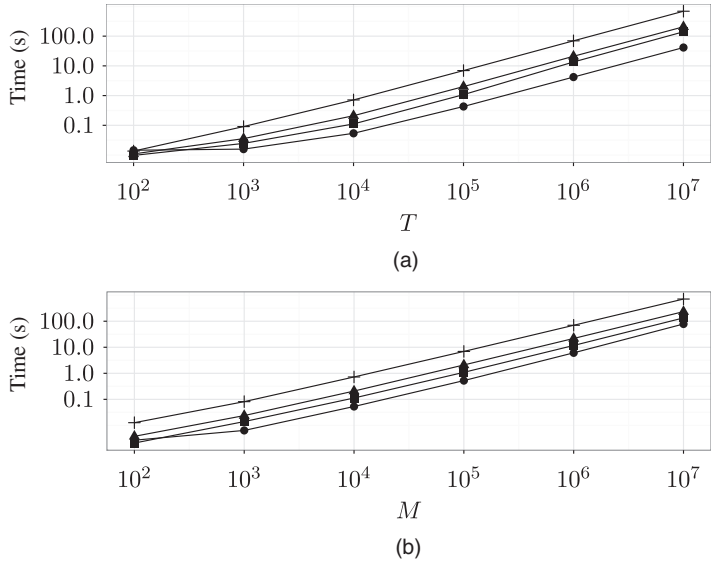

Fig. 3. Execution times for the implementation of algorithm 2 available in R package not (Baranowski *et al.*, 2016a), for various feature detection problems with the data Y_t, $t = 1, \ldots, T$, IID $\mathcal{N}(0, 1)$ (in a single run, computations for the input of the algorithm are performed in parallel, using eight cores of an Intel Xeon 3.6-GHz central processor unit with 16 Gbytes of random-access memory; the computation times are averaged over 10 runs in each case) (●, scenario 1; ▲, scenario 2; ■, scenario 3; +, scenario 4): (a) fixed $M = 10000$; (b) fixed $T = 10000$

running times appear to scale linearly both in T (Fig. 3(a)) and in M (Fig. 3(b)), 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.

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.*, 2016a) achieves an average computation time not longer than 2 s in all the examples in Section 5 by using a single core of an Intel Xeon 3.6-GHz central processor unit. This can be accelerated further, as the `not` package allows for computing the contrast function over the intervals drawn in parallel by using all available central processor unit 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 the NOT algorithm with $M = 10^5$ offered better practical performance on the change point rich signals that they considered. In the most extreme scenario where we expect 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 that is less stringent than sSIC. See section F of the on-line supplementary materials.

3.6.2. Early stopping for narrowest-over-threshold method with the strengthened Schwarz information criterion

If the number of change points in the data is expected to be quite 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 minimize equation (11). By considering $|\mathcal{T}(\zeta_T^{(k)})| \leqslant q_{\max}$, we could achieve some computational gains without adversely impacting the overall performance of the methodology. As such, in all applications that are presented in this work we compute sSIC only for k such that $|\mathcal{T}(\zeta_T^{(k)})| \leqslant q_{\max}$ with $q_{\max} = 25$.

4. Narrowest-over-threshold method under different noise types

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

4.1. Narrowest-over-threshold method 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 the NOT algorithm here would incur information loss. As is shown in corollaries 1 and 2 in scenarios 1 and 2, the NOT method is still consistent if we replace the noise's assumption of IID data 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 auto-regressive moving average models. See also the numerical examples in section E of the on-line 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 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 a guarantee on the performance of the NOT algorithm from a theoretical perspective.

Corollary 1. Suppose that Y_t follow model (3) in scenario 1, 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 $\Sigma_{k=-\infty}^{\infty} |\rho_k| < \infty$. Then, the conclusion of theorem 1 still holds (with different constants).

Corollary 2. Suppose that Y_t follow model (3) in scenario 2, 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 $\sigma_0(\Sigma_{k=-\infty}^{\infty} |\rho_k|)^{1/2}$. This quantity can also be viewed as a generalization of the independent noise setting, where the threshold is proportional to σ_0 (since $\Sigma_{k=-\infty}^{\infty} |\rho_k| = 1$). More details of its derivation are provided in section 1.6 of the on-line supplementary materials.

This poses a few challenges in the practical application of NOT detection to signals with dependent noise:

(a) the (pre-)estimation of the residuals ε_t in preparation for the estimation of their long-run variance;
(b) the estimation of σ_0;

(c) the estimation of $\sigma_0(\Sigma_{k=-\infty}^{\infty}|\rho_k|)^{1/2}$.

These problems are known to be difficult in time series analysis in general. A possible solution is outlined below.

For problem (a), we have had some success with the wavelet-based method of Johnstone and Silverman (1997), which was implemented in the R package `wavethresh` (Nason, 2016); its advantages are that it is specifically designed for dependent noise and that, being based on non-linear wavelet shrinkage, it is particularly suited for signals with irregularities, such as (generalized) change points. Here the Haar wavelet transform of the data is appropriate in scenario 1, whereas a transform with respect to any wavelet that annihilates linear functions is appropriate in scenarios 2 and 3. Once the empirical residuals have been obtained from problem (a) we could then estimate σ_0 in problem (b) by its sample version and estimate $\sigma_0(\Sigma_{k=-\infty}^{\infty}|\rho_k|)^{1/2}$ in problem (c) in a model-based way (e.g. using the auto-regressive model with its order p chosen by an information criterion).

Another possibility to estimate change points under dependent noise is to use self-normalizing-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 detection:

(a) preaverage the data over non-overlapping moving windows of size h, creating a new data set of length $\lfloor T/h \rfloor$; the hope is that, by the law of large numbers, the preaveraged noise will be closer to Gaussian and also less serially dependent than the original noise;
(b) add additional IID Gaussian noise to the data, with mean 0 and suitably chosen standard deviation; this will have a similar effect to that 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 narrowest-over-threshold method under heavy-tailed noise

NOT detection 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 1, we propose to apply the following new contrast function, which is defined for $\mathbf{Y}$ and $0 \leqslant s < b < e \leqslant T$ as

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

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 $\mathcal{S}_{(s,e]}(\mathbf{v})_i = \operatorname{sgn}\{v_i - (e-s)^{-1}\Sigma_{t=s+1}^{e} v_t\}$ and $\psi^b_{(s,e]}$ is defined by equation (5). (For certain noise distributions, subtracting the sample median of $\mathbf{v}$ instead of the sample mean would appear more appropriate.) The rationale behind function (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. The empirical performance of the NOT approach (via algorithm 2) combined with equation (12) and sSIC is also illustrated in section E of the on-line supplementary materials.

5. Simulation study

5.1. Settings

We consider examples following scenarios 1–4 that were introduced in Section 2.3, as well as an extra example satisfying $\sigma_t = \sigma_0$ and f_t is a piecewise quadratic function of t (scenario 5).

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

(a) IID $\mathcal{N}(0, 1)$,
(b) IID $\mathcal{N}(0, 2)$,
(c) IID scaled Laplace with zero mean and unit variance,
(d) IID scaled Student t_5-distribution with unit variance and
(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 on-line supplementary materials. Fig. 4 shows the examples of the data generated from models M1–M7, as well as the estimates produced by the NOT algorithm in a typical run.

5.2. Estimators

We apply algorithm 2 to compute the NOT solution path and pick the solution minimizing sSIC introduced in Section 3.3 with $\alpha = 1$ (which is equivalent to the Schwarz information criterion). In each simulated example, we use the contrast function that was designed to detect change points in the scenario that the example follows, given in Section 2.3 and section B of the on-line supplementary materials under the assumption that ε_t is IID Gaussian. The resulting method is referred to simply as 'NOT'. In addition, for scenario 1 only, we also apply algorithm 2 combined with equation (12) and the Schwarz information criterion, 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 Schwarz information criterion 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 in the Comprehensive R Archive Network. To the best of our knowledge, none of the competing packages can be applied in all of scenarios 1–5.

For change point detection in the mean, the selected competitors from the Comprehensive R Archive Network are `changepoint` (Killick and Eckley, 2014; Killick *et al.*, 2016) implementing the PELT methodology that was proposed by Killick, Fearnhead and Eckley (2012), `changepoint.np` (Haynes *et al.*, 2016) implementing a non-parametric extension of the PELT methodology that was studied in Haynes *et al.* (2017), `wbs` (Baranowski and Fryzlewicz, 2015) implementing WBS proposed by Fryzlewicz (2014), `ecp` (James and Matteson, 2014) implementing the e.cp3o method that was 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 that was proposed by Rigaill (2015), `nmcdr` (Zou and Lancezhange, 2014) implementing NMCD, the non-parametric multiple change point detection methodology of Zou *et al.* (2014), `stepR` (Pein *et al.*, 2018) implementing the simultaneous multiscale change point estimator SMUCE that was proposed by Frick *et al.* (2014) and `FDRSeg` (Li *et al.*, 2017) implementing the method called FDRSeg proposed by Li *et al.* (2016). We refer to the corresponding methods as PELT, NP-PELT, WBS, e.cp3o, B&P, S3IB, NMCD, SMUCE and FDRSeg respectively.

Note that e-cp3o, NMCD, NOT, PELT and NP-PELT can be used also for change point detection in scenario 4, where change points occur in the mean and variance of the data. In

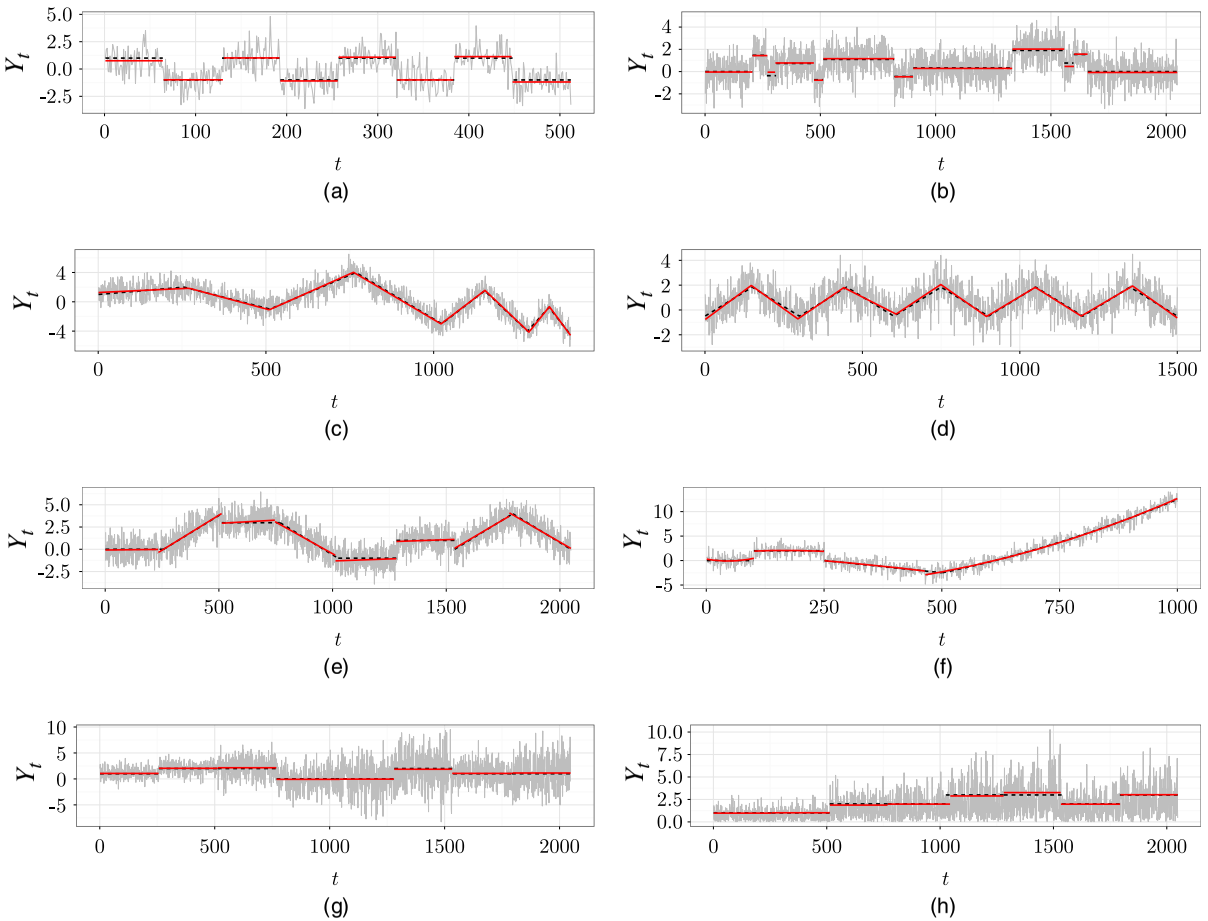

Fig. 4. Examples of data generated from simulation models outlined in section A of the on-line supplementary materials: (a)–(g) data series Y_t (———), true signal f_t (- - - - -), $\hat{f}_t$ being the least squares estimate of f_t with the change points estimated by the NOT algorithm (———); (h) centred data $|Y_t - \hat{f}_t|$ (———), true standard deviation σ_t (- - - - -) and the estimated standard deviation $\hat{\sigma}_t$ between the change points detected by the NOT algorithm (———); (a) model M1 teeth, scenario 1; (b) model M2 blocks, scenario 1; (c) model M3 wave1, scenario 2; (d) model M4 wave2, scenario 2; (e) model M5 mix, scenario 3; (f) model M7 quad, scenario 5; (g) model M6 vol, f_t, scenario 4; (h) model M6 vol, σ_t, scenario 4

addition, for scenario 4, 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 B&P allows for change point detection in piecewise linear and piecewise quadratic signals (in particular, 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 TF hereafter, using the implementation that is available from the R package `genlasso` (Taylor and Tibshirani, 2014), to have a broader comparison. See also Lin *et al.* (2017). TF aims to estimate a piecewise polynomial signal from the data, not focusing on the change point detection problem directly. Let $\hat{f}_t^{(\mathrm{TF})}$ denote the TF estimate of the true signal f_t; then the TF estimates of the change points in scenario 2 are defined as those τ for which $|2\hat{f}_\tau^{(\mathrm{TF})} - \hat{f}_{\tau-1}^{(\mathrm{TF})} - \hat{f}_{\tau+1}^{(\mathrm{TF})}| > \epsilon$, where $\epsilon > 0$ is a very small number being the numerical level of tolerance (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 $|\hat{f}_{\tau+2}^{(\mathrm{TF})} - 3\hat{f}_{\tau+1}^{(\mathrm{TF})} + 3\hat{f}_\tau^{(\mathrm{TF})} - \hat{f}_{\tau-1}^{(\mathrm{TF})}| > \epsilon$. 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 that were recommended by the corresponding R packages, and the R code for all simulations can be downloaded from our GitHub repository (Baranowski *et al.*, 2016b).

5.3. Results

Here we present only the results under the setting where the noise is (a) IID standard normal in Table 2. Additional results under the other above-mentioned noise settings can be found in section E of the on-line 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

$$\mathrm{MSE} = \mathbb{E}\left\{ \frac{1}{T} \sum_{t=1}^{T} (f_t - \hat{f}_t)^2 \right\}.$$

For all methods except TF, $\hat{f}_t$ is calculated by finding the least squares 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 mean-squared error is the penalized 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

$$d_{\mathrm{H}} = T^{-1}\mathbb{E}[\max\{ \max_{j=0,\ldots,q+1} \min_{k=0,\ldots,\hat{q}+1} |\tau_j - \hat{\tau}_k|, \max_{k=0,\ldots,\hat{q}+1} \min_{j=0,\ldots,q+1} |\hat{\tau}_k - \tau_j| \}],$$

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 $0 \leqslant d_{\mathrm{H}} \leqslant 1$. 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 underestimated or some of the estimated change points are far from the real change points. In addition, we also report estimates of the inverse V-measure d_V defined as

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

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, $0 \leqslant d_V \leqslant 1$, and a perfect estimator has $d_V = 0$, whereas $d_V = 1$ means that none of the features are detected (i.e. $\hat{q} = 0$).

We find that, in most of the simulated scenarios, the NOT method is among the most competitive methods in terms of the estimation of the number of change points and their locations, as well as the true signal. Importantly, it is very fast to compute, which gives it a particular advantage over its competitors in scenarios 2, 3 and 5. Finally, the NOT algorithm with the contrast function derived under the assumption that the noise is IID 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 Goddard Institute for Space Studies surface temperature anomalies data set that is available from GISTEMP Team (2016) (`http://data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts+dSST.csv`), 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 for example Ruggieri (2013) or James and Matteson (2015).

The plot of the data (Fig. 5(a)) 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 2 with change points in the slope of the trend is the most appropriate here. To detect the locations of the change points, we apply the NOT algorithm (via algorithm 2) with the contrast given by equation (8), combined with the Schwarz information criterion to determine the best model on the solution path.

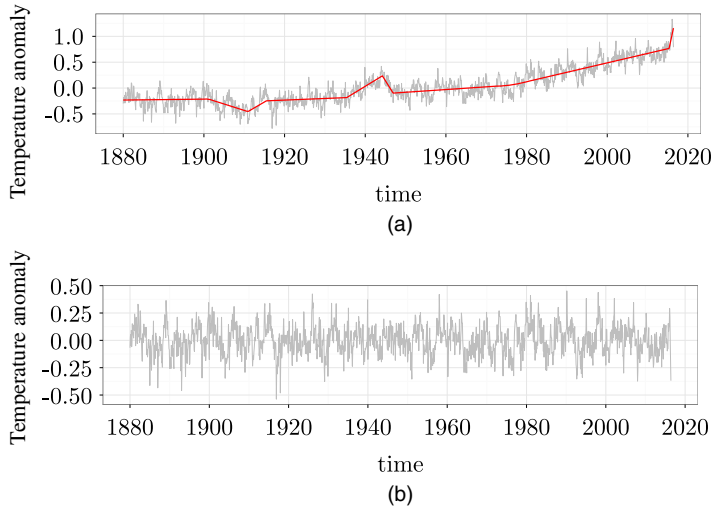

Fig. 5. Change point analysis for the GISTEMP data set introduced in Section 6.1: (a) data series Y_t (——) and $\hat{f}_t$ estimated by using change points returned by the NOT algorithm (——); (b) residuals $\hat{\varepsilon}_t = Y_t - \hat{f}_t$

Table 2. Distribution of $\hat{q} - q$ for data generated according to model (3) with the noise term ε_t IID $\mathcal{N}(0,1)$ for various choices of f_t and σ_t given in section A of the on-line supplementary materials and competing methods listed in Section 5†

Model	*Method*	*Results for the following values of* $\hat{q} - q$*:*							*MSE*	$d_H \times 10^2$	d_V	*Time (s)*
		$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$				
M1	B&P	0	0	0	*97*	3	0	0	0.051	0.59	*0.019*	15.335
	e-cp3o	0	0	0	*100*	0	0	0	0.088	0.62	0.041	0.124
	FDRSeg	0	0	0	83	14	1	2	0.089	1.18	0.044	0.035
	NMCD	0	0	0	*96*	4	0	0	0.098	0.9	0.046	1.098
	NOT	0	0	0	*99*	1	0	0	0.05	*0.54*	*0.019*	0.046
	NOT HT	0	0	0	*97*	3	0	0	0.055	0.62	0.021	0.059
	NP-PELT	0	0	0	83	12	4	1	0.067	0.99	0.028	0.018
	PELT	0	0	0	*100*	0	0	0	0.05	*0.51*	*0.019*	0.002
	S3IB	0	0	0	*92*	5	2	1	0.052	0.67	*0.02*	0.075
	SMUCE	0	0	0	*100*	0	0	0	0.085	0.59	0.04	0.046
	WBS	0	0	0	*96*	4	0	0	0.052	0.59	*0.02*	0.072
M2	B&P	0	4	34	*62*	0	0	0	0.021	*1.27*	*0.022*	382.524
	e-cp3o	100	0	0	0	0	0	0	0.177	6.35	0.127	2.403
	FDRSeg	0	1	30	54	10	5	0	0.029	1.58	0.032	1.189
	NMCD	1	13	61	24	1	0	0	0.036	2.21	0.039	4.807
	NOT	0	3	49	44	3	1	0	0.026	1.66	0.026	0.082
	NOT HT	3	8	54	27	7	0	1	0.034	2.52	0.038	0.149
	NP-PELT	0	3	16	53	23	5	0	0.028	1.64	0.03	0.226
	PELT	7	34	47	12	0	0	0	0.033	3.01	0.036	0.002
	S3IB	0	4	37	*56*	2	1	0	0.024	1.43	*0.024*	0.348
	SMUCE	58	35	7	0	0	0	0	0.071	3.4	0.061	0.019
	WBS	1	3	32	*56*	6	2	0	0.026	1.5	0.027	0.15
M3	B&P	0	0	0	*98*	2	0	0	0.068	2.46	0.117	87.917
	NOT	0	0	0	*100*	0	0	0	0.015	*0.89*	*0.051*	0.621
	TF	0	0	0	0	0	0	100	0.017	8.31	0.219	49.933
M4	B&P	0	0	1	*99*	0	0	0	0.074	2.98	0.156	102.579
	NOT	0	0	0	*100*	0	0	0	0.016	*1.25*	*0.07*	0.609
	TF	0	0	0	0	0	0	100	0.016	4.31	0.147	49.876
M5	B&P	0	0	0	*100*	0	0	0	0.021	*2.53*	*0.089*	201.256
	NOT	0	0	0	*100*	0	0	0	0.02	*2.46*	*0.086*	0.372
	TF	0	0	0	0	0	0	100	0.027	6.03	0.26	60.866
M6	e-cp3o	15	6	8	29	14	17	11	0.156	6.72	0.17	1.857
	HSMUCE	98	2	0	0	0	0	0	0.097	12.66	0.216	0.123
	NMCD	0	0	17	73	9	1	0	0.06	3.75	0.068	4.403
	NOT	0	0	16	*82*	2	0	0	0.049	3.15	*0.051*	0.474
	NP-PELT	0	0	0	20	27	27	26	0.127	3.45	0.072	0.29
	PELT	9	16	30	42	3	0	0	0.071	7.62	0.083	0.008
	SegNeigh	0	0	7	59	26	5	3	0.05	*2.45*	*0.048*	18.452
M7	B&P	0	0	1	*98*	1	0	0	0.021	*2.47*	*0.073*	48.711
	NOT	0	0	1	*98*	1	0	0	0.022	*2.33*	*0.07*	0.468
	TF	0	0	0	0	0	0	100	0.05	23.37	0.442	45.981

†Also tabulated are 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 by using a single core of an Intel Xeon 3.6-GHz central processor unit with 16 Gbytes of random-access memory, all calculated over 100 simulated data sets. 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, are given in italics.

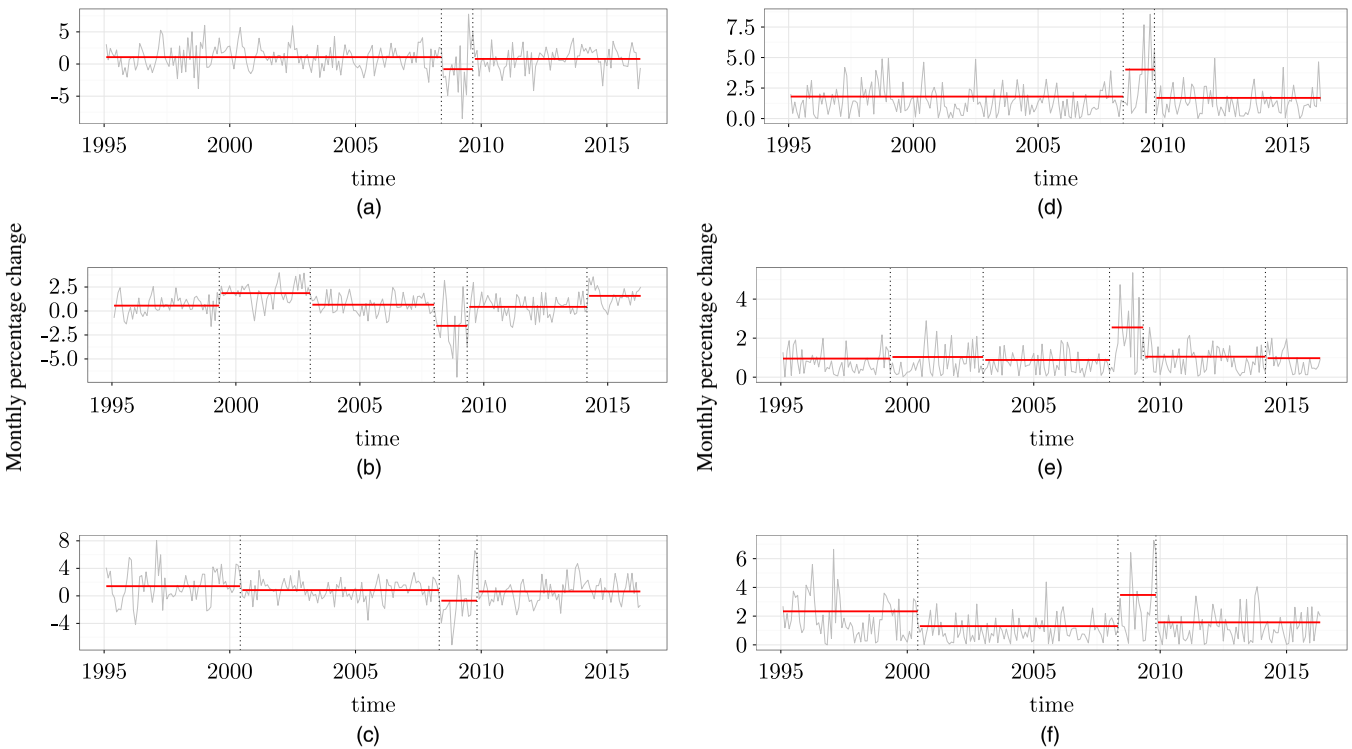

Fig. 6. Change point analysis for the monthly percentage changes in the UK HPI from January 1995 to May 2016: (a)–(c) monthly percentage changes Y_t and the fitted piecewise constant mean $\hat{f}_t$, between the change points estimated with the NOT method; (d)–(f) $|Y_t - \hat{f}_t|$ and the fitted piecewise constant standard deviation $\hat{\sigma}_t$, between the change points estimated with the NOT method; (a), (d) Hackney; (b), (e) Newham; (c), (f) Tower Hamlets

The NOT estimate of the piecewise linear trend and the corresponding empirical residuals are shown in Fig. 5. We identify eight change points 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 some related analyses). In addition to the change points around these dates, the NOT algorithm 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 detection 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) (`https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-january-2017`), 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 on line from UK Land Registry (2016). Fryzlewicz (2018b), who proposed a method for signal estimation and change point detection in scenario 1, used this data set to illustrate the performance of his methodology. We perform a similar analysis, assuming the more flexible scenario 4, allowing for changes both in the mean and in the variance, which, we argue, leads to additional insights and better interpretable estimates for this data set.

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

Fig. 6 shows monthly percentage changes in the HPI for the boroughs analysed and the corresponding NOT estimates, obtained by using the contrast function for scenario 4. As recommended in Section 3.3, we set the number of intervals drawn in the procedure to $M = 10000$ and choose the threshold that minimizes the Schwarz information criterion. For better comparability, the NOT algorithm is applied with the same random seed for each data series.

In contrast with Fryzlewicz (2018b), whose tail greedy unbalanced Haar 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, the NOT algorithm 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 effect 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 4 may be more relevant than scenario 1 considered in Fryzlewicz (2018b).

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 referees for their comments and suggestions. Piotr Fryzlewicz's work was supported by Engineering and Physical Sciences Research Council grant EP/L014246/1.

References

Auger, I. E. and Lawrence, C. E. (1989) Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, **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. *J. Appl. Econmetr.*, **18**, 1–22.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016a) not: narrowest-over-threshold change-point detection. *R Package v1.0*. Department of Statistics, London School of Economics and Political Science. (Available from `https://cran.r-project.org/web/packages/not`.)

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016b) Narrowest-over-threshold detection of multiple change-points and change-point-like features: simulation code. Department of Statistics, London School of Economics and Political Science. (Available from `https://github.com/rbaranowski/not-num-ex`.)

Baranowski, R. and Fryzlewicz, P. (2015) wbs: wild binary segmentation for multiple change-point detection. *R Package v1.3*. Department of Statistics, London School of Economics and Political Science. (Available from `https://CRAN.R-project.org/package=wbs`.)

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

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

Cleynen, A., Rigaill, G. and Koskas, M. (2013) Segmentor3isback: a fast segmentation algorithm. *R Package v1.8*. Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, Montpellier. (Available from `https://CRAN.R-project.org/package=Segmentor3IsBack`.)

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006) Structural break estimation for nonstationary time series models. *J. Am. Statist. Ass.*, **101**, 223–239.

Fang, X., Li, J. and Siegmund, D. (2019) Segmentation and estimation of change-point models: false positive control and confidence regions. *Ann. Statist.*, to be published.

Frick, K., Munk, A. and Sieling, H. (2014) Multiscale change point inference (with discussion). *J. R. Statist. Soc.* B, **76**, 495–580.

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

Fryzlewicz, P. (2018a) Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Preprint*. Department of Statistics, London School of Economics and Political Science. (Available from `http://stats.lse.ac.uk/fryzlewicz/wbs2/wbs2.pdf`.)

Fryzlewicz, P. (2018b) Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Ann. Statist.*, **46**, 3390–3421.

GISTEMP Team (2016) GISS surface temperature analysis (GISTEMP). GISTEMP Team, New York. (Available from `http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/`.)

Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J. Am. Statist. Ass.*, **69**, 383–393.

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

Hawkins, D. M. (2001) Fitting multiple change-point models to data. *Computnl Statist. Data Anal.*, **37**, 323–341.

Haynes, K., Fearnhead, P. and Eckley, I. A. (2016) changepoint.np: methods for nonparametric changepoint detection. *R Package v0.0.2*. Department of Mathematics and Statistics, Lancaster University, Lancaster. (Available from `https://CRAN.R-project.org/package=changepoint.np`.)

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

James, N. A. and Matteson, D. S. (2014) ecp: an R package for nonparametric multiple change point analysis of multivariate data. *J. Statist. Softwr.*, **62**, 1–25.

James, N. A. and Matteson, D. S. (2015) Change points via probabilistically pruned objectives. *Preprint arXiv:1505.04302*. Department of Statistical Science, Cornell University, Ithaca.

Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc.* B, **59**, 319–351.

Killick, R. and Eckley, I. A. (2014) changepoint: an R package for changepoint analysis. *J. Statist. Softwr.*, **58**, 1–19.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *J. Am. Statist. Ass.*, **107**, 1590–1598.

Killick, R., Haynes, K. and Eckley, I. A. (2016) changepoint: methods for changepoint detection. *R Package v2.2.2*. Department of Mathematics and Statistics, Lancaster University, Lancaster. (Available from `http://CRAN.R-project.org/package=changepoint`.)

Killick, R., Nam, C., Aston, J. and Eckley, I. A. (2012) The changepoint repository. Department of Mathematics and Statistics, Lancaster University, Lancaster. (Available from `http://changepoint.info/`.)

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

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

Lee, C.-B. (1997) Estimating the number of change points in exponential families distributions. *Scand. J. Statist.*, **24**, 201–210.
Li, H., Munk, A. and Sieling, H. (2016) FDR-control in multiscale change-point segmentation. *Electron. J. Statist.*, **10**, 918–959.
Li, H., Sieling, H. and Aspelmeier, T. (2017) FDRSeg: FDR-control in multiscale change-point segmentation. *R Package v1.0-3*. (Available from `https://CRAN.R-project.org/package=FDRSeg`.)
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. In *Advances in Neural Information Processing Systems 30*.
Liu, J., Wu, S. and Zidek, J. V. (1997) On segmented multivariate regression. *Statist. Sin.*, **7**, 497–526.
Maidstone, R., Fearnhead, P. and Letchford, A. (2017) Detecting changes in slope with an L_0 penalty. *Preprint arXiv:1701.01672*. Department of Mathematics and Statistics, Lancaster University, Lancaster.
Nason, G. (2016) wavethresh: wavelet statistics and transforms. *R Package v4.6.8*. Department of Mathematics, University of Bristol, Bristol. (Available from `http://CRAN.R-project.org/package=wavethresh`.)
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. *R Package v2.0-2*. (Available from `http://CRAN.R-project.org/package=stepR`.)
Pein, F., Sieling, H. and Munk, A. (2017) Heterogeneous change point inference. *J. R. Statist. Soc.* B, **79**, 1207–1227.
Pešta, M. and Wendler, M. (2018) Nuisance parameters free changepoint detection in non-stationary series. *Preprint arXiv:1808.01905*. Department of Probability and Mathematical Statistics, Charles University, Prague.
Raimondo, M. (1998) Minimax estimation of sharp change points. *Ann. Statist.*, **26**, 1379–1397.
Rigaill, G. (2015) A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_{max} change-points. *J. Soc. Fr. Statist.*, **156**, 180–205.
Rosenberg, A. and Hirschberg, J. (2007) V-Measure: a conditional entropy-based external cluster evaluation measure. In *Proc. Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 410–420. Madison: Omnipress.
Rufibach, K. and Walther, G. (2010) The block criterion for multiscale inference about a density, with applications to other multiscale problems. *J. Computnl Graph. Statist.*, **19**, 175–190.
Ruggieri, E. (2013) A Bayesian approach to detecting change points in climatic records. *Int. J. Climtol.*, **33**, 520–528.
Shao, X. and Zhang, X. (2010) Testing for change points in time series. *J. Am. Statist. Ass.*, **105**, 1228–1240.
Taylor, A. B. and Tibshirani, R. J. (2014) genlasso: path algorithm for generalized lasso problems. *R Package v1.3*. University of Richmond, Richmond. (Available from `https://CRAN.R-project.org/package=genlasso`.)
Tibshirani, R. J. (2014) Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.*, **42**, 285–323.
UK Land Registry (2016) UK house price index. UK Land Registry. (Available from `http://landregistry.data.gov.uk/app/ukhpi`.)
Venkatraman, E. S. (1992) Consistency results in multiple change-point problems. *PhD Thesis*. Stanford University, Stanford.
Vostrikova, L. (1981) Detection of the disorder in multidimensional random processes. *Sov. Math. Dokl.*, **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. *Statist. Probab. Lett.*, **6**, 181–189.
Yao, Y.-C. and Au, S. T. (1989) Least-squares estimation of a step function. *Sankhya* A, **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. *J. Statist. Softwr.*, **7**, 1–38.
Zhang, T. and Lavitas, L. (2018) Unsupervised self-normalized change-point testing for time series. *J. Am. Statist. Ass.*, **113**, 637–648.
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.
Zou, C. and Lancezhange (2014) nmcdr: non-parametric multiple change-points detection. *R Package v0.3.0*. Nankai University, Tianjin. (Available from `https://CRAN.R-project.org/package=nmcdr`.)
Zou, C., Yin, G., Feng, L. and Wang, Z. (2014) Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, **42**, 970–1002.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Online supplementary materials for "Narrowest-over-threshold detection of multiple change-points and change-point-like features"'.