Wild binary segmentation for multiple change-point detection

The Annals of Statistics
2014, Vol. 42, No. 6, 2243–2281
DOI: 10.1214/14-AOS1245

WILD BINARY SEGMENTATION FOR MULTIPLE CHANGE-POINT DETECTION

BY PIOTR FRYZLEWICZ

London School of Economics

We propose a new technique, called wild binary segmentation (WBS), for consistent estimation of the number and locations of multiple change-points in data. We assume that the number of change-points can increase to infinity with the sample size. Due to a certain random localisation mechanism, WBS works even for very short spacings between the change-points and/or very small jump magnitudes, unlike standard binary segmentation. On the other hand, despite its use of localisation, WBS does not require the choice of a window or span parameter, and does not lead to a significant increase in computational complexity. WBS is also easy to code. We propose two stopping criteria for WBS: one based on thresholding and the other based on what we term the 'strengthened Schwarz information criterion'. We provide default recommended values of the parameters of the procedure and show that it offers very good practical performance in comparison with the state of the art. The WBS methodology is implemented in the R package `wbs`, available on CRAN.

In addition, we provide a new proof of consistency of binary segmentation with improved rates of convergence, as well as a corresponding result for WBS.

1. Introduction. A posteriori change-point detection problems have been of interest to statisticians for many decades. Although, naturally, details vary, a theme common to many of them is as follows: a time-evolving quantity follows a certain stochastic model whose parameters are, exactly or approximately, piecewise constant. In such a model, it is of interest to detect the number of changes in the parameter values and the locations of the changes in time. Such piecewise-stationary modelling can be appealing for a number of reasons: the resulting model is usually much more flexible than the corresponding stationary model but still parametric if the number of change-points is fixed; the estimated change-points are often 'interpretable' in the sense that their locations can be linked to the behaviour of some exogenous quantities of interest; the last estimated segment can be viewed as the 'current' regime of stationarity, which can be useful in, for example, forecasting future values of the observed process. Finally, a posteriori segmentation can be a useful exploratory step in the construction of more complex models in which

Received November 2013; revised April 2014.
MSC2010 subject classifications. 62G05.
Key words and phrases. Multiple change-points, change-point detection, binary segmentation, randomised algorithms, thresholding, Bayesian information criterion.

the piecewise constant variables are themselves treated as random and evolving according to a certain, perhaps Markovian, mechanism.

Arguably the simplest, 'canonical' model with change-points is that of the form

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

where f_t is a deterministic, one-dimensional, piecewise-constant signal with change-points whose number N and locations $\eta_1, \ldots, \eta_N$ are unknown. The sequence ε_t is random and such that $\mathbb{E}(\varepsilon_t)$ is exactly or approximately zero. In the simplest case ε_t are modelled as i.i.d., but can also follow more complex time series models. The task is to estimate N and $\eta_1, \ldots, \eta_N$ under various assumptions on N, the magnitudes of the jumps and the minimum permitted distance between the change-point locations. Being univariate, model (1) excludes, for example, many interesting time series segmentation problems in which the process at hand is typically parameterised by more than one parameter in each segment. However, it still provides a useful training ground for change-point detection techniques in the sense that if a given method fails to perform in the simple model (1), it should not typically be expected to perform well in more complex settings.

There is considerable literature on a posteriori multiple change-point detection in different variants of model (1). Yao and Au (1989) consider least-squares estimation of f_t in the case of a fixed N (either known or unknown), under the assumption of ε_t being i.i.d. In the case of a known N, they show the consistency of the estimated change-point locations with the rate of $O_P(1)$. They also propose a penalised least-squares estimator of N in the case when it is unknown but bounded. In the Gaussian case, the Schwarz criterion is used to estimate an unknown but bounded N in Yao (1988), and a more general criterion that is also linear in the number of change-points appears in Lee (1995). For an unknown but bounded N, Lavielle and Moulines (2000) consider penalised least-squares estimation, with a penalty linear in the number of change-points, and show its consistency for the number and locations of change-points for dependent ε_t's, including the cases of strong mixing and long-range dependence; see also Lavielle (1999) for a discussion and some extensions of this result and Lavielle (2005) for some practical proposals regarding the adaptive choice of the penalty parameter. For a fixed N, Pan and Chen (2006) propose a likelihood criterion with a penalty depending not only on the number, but also on the locations of change-points, favouring more uniformly-spread estimated change-points. For an unknown N, Lebarbier (2005) propose least-squares estimation with a penalty originating from the model selection approach of Birgé and Massart (2001) and show the least-squares consistency of the resulting estimator of f_t (not of the estimated change-points themselves). Boysen et al. (2009) use the least-squares criterion with a linear penalty on the number of change-points and, under the assumption of a finite but arbitrary N, show various theoretical results including analogues of those of Yao and Au (1989). More general forms of Schwarz-like penalties are studied, for example, in Wu (2008) and Ciuperca (2011, 2014).

Often, a major drawback of change-point estimators formulated as multivariate optimisation problems, such as those based on penalised least-squares or log-likelihood fits, is their computational complexity, which is typically of order $O(T^2)$ [see, e.g., Auger and Lawrence (1989) and Jackson et al. (2005)], a prohibitively slow speed for large datasets. Killick, Fearnhead and Eckley (2012) propose an algorithm, called PELT, that reduces the complexity to $O(T)$ but under the assumption of change-points being separated by time intervals drawn independently from a probability distribution, a set-up under which considerations of statistical consistency are impossible due to these spacings being too short. Rigaill (2010) proposes an alternative 'pruned dynamic programming' algorithm with the aim of reducing the computational effort, which, however, remains of order $O(T^2)$ in the worst case. Both algorithms are revisited in the simulations section of this paper. An interesting approach to change-point detection, in the context of piecewise-stationary AR time series models rather than in model (1), appears in Davis, Lee and Rodriguez-Yam (2006): the minimum description length is used as the criterion for segmentation, and it is minimised using a genetic algorithm to reduce computational complexity.

A different route to reducing the computational complexity of the multiple change-point detection problem is taken by Harchaoui and Lévy-Leduc (2010) who consider the least-squares criterion with a total variation penalty, which enables them to use the LARS algorithm of Efron et al. (2004) to compute the solution in $O(NT\log(T))$ time. For a known N (only), they prove consistency of the resulting estimated change-point locations with near-optimal rates. We note, however, that the total variation penalty is not an optimal one for change-point detection; see Cho and Fryzlewicz (2011), who reiterate an argument made earlier in Brodsky and Darkhovsky (1993). The total variation penalty is also considered in the context of peak/trough detection by Davies and Kovac (2001), who propose the 'taut string' approach for fast computation, and in the context of multiple change-point detection by Rinaldo (2009) [as part of the fused lasso penalty, proposed by Tibshirani et al. (2005) and equivalent to taut string in model (1)] and Rojas and Wahlberg (2014), who also point out that the main result in Rinaldo (2009) is erroneous. On the other hand, Wang (1995) uses the traditional fast discrete wavelet transform to detect change-points.

An informative review of some multiple change-point detection methods (in the context of DNA segmentation, but applicable more widely) appears in Braun and Mueller (1998). Killick et al. (2012) is an online repository of publications and software related to change-point detection.

Binary segmentation (BS) is a generic technique for multiple change-point detection in which, initially, the entire dataset is searched for one change-point, typically via a CUSUM-like procedure. If and once a change-point is detected, the data are then split into two (hence the name 'binary') subsegments, defined by the detected change-point. A similar search is then performed on either subsegment, possibly resulting in further splits. The recursion on a given segment continues

until a certain criterion is satisfied on it. Unlike estimators resulting from multi-dimensional optimisation of a certain global criterion, such as the least-squares estimators reviewed above, BS is a 'greedy' procedure in the sense that it is performed sequentially, with each stage depending on the previous ones, which are never re-visited. On the other hand, each stage is particularly simple and involves one-, rather than multi-dimensional optimisation. To the best of our knowledge, the first work to propose BS in a stochastic process setting was Vostrikova (1981), who showed consistency of BS for the number and locations of change-points for a fixed N, with rates of convergence of the estimators of locations, under certain technical conditions on the norm of the cumulative sum of the process X_t, which in that work was assumed to be multivariate. Testing for change-points at each stage of the BS procedure was performed via a simple CUSUM test; however, the stopping criterion was not easy to compute in practice due to randomness in the previously detected change-points. Venkatraman (1992) outlines an interesting proof of the consistency of BS for N and for the change-point locations, even for N increasing with T, albeit with sub-optimal rates for the locations.

Interestingly, BS in a setting similar to Vostrikova (1981) (for a fixed N and with ε_t following a linear process), reappears in Bai (1997), but without references to the earlier works cited above. Chen, Cohen and Sackrowitz (2011) provide a proof of consistency of BS for the number of change-points in the case of fixed N and i.i.d. normal ε_t; however, links between their result and the analogous consistency results obtained in the above papers are not established.

We also note that BS has an interpretation in terms of 'unbalanced Haar' wavelets; see Fryzlewicz (2007). BS is used for univariate time series segmentation in Fryzlewicz and Subba Rao (2014) and Cho and Fryzlewicz (2012), and for multivariate, possibly high-dimensional time series segmentation in Cho and Fryzlewicz (2014).

The benefits of BS include low computational complexity [typically of order $O(T \log T)$], conceptual simplicity, and the fact that it is usually easy to code, even in more complex models than (1). Killick, Fearnhead and Eckley (2012) describe it as 'arguably the most widely used change-point search method'. On the other hand, the fact that each stage of BS involves search for a *single* change-point means that BS may be unsuitable for some functions containing multiple change-points in certain configurations. Indeed, in one of our side results of the paper, we show that BS is only consistent when the minimum spacing between any two adjacent change-points is of order greater than $T^{3/4}$ (even in the 'easiest' case of jump magnitudes being bounded away from zero), so relatively large.

In this work, we attempt to capitalise on the popularity and other benefits of BS and propose a multiple change-point detection procedure, termed wild binary segmentation (WBS), which inherits the main strengths of BS but attempts to eliminate its weaknesses. The main idea is simple. In the first stage, rather than using a global CUSUM statistic that uses the entire data sample $(X_1, X_2, \ldots, X_T)$, we randomly draw (hence the term 'wild') a number of subsamples, that is, vectors

$(X_s, X_{s+1}, \ldots, X_e)$, where s and e are integers such that $1 \le s < e \le T$, and compute the CUSUM statistic on each subsample. We then maximise each CUSUM, choose the *largest* maximiser over the entire collection of CUSUMs, and take it to be the first change-point candidate to be tested against a certain threshold. If it is considered to be significant, the same procedure is then repeated recursively to the left and to the right of it. The hope is that even a relatively small number of random draws will contain a particularly 'favourable' draw in which, for example, the randomly drawn interval (s, e) contains only one change-point, sufficiently separated from both s and e: a set-up in which our CUSUM estimator of the change-point location works particularly well as it coincides with the maximum likelihood estimator (in the case of ε_t being i.i.d. Gaussian). We provide a lower bound for the number of draws that guarantees such favourable draws with a high probability. Apart from the threshold-based stopping criterion for WBS, we also introduce another, based on what we call the *strengthened Schwarz information criterion*.

By 'localising' our CUSUM statistic in this randomised manner, we overcome the issue of the 'global' CUSUM being unsuitable for certain configurations of multiple change-points. We also dramatically reduce the permitted spacing between neighbouring change-points in comparison to standard BS, as well as the permitted jump magnitudes. Moreover, by drawing intervals of different lengths, we avoid the problem of span or window selection, present in some existing approaches to localising the CUSUM statistic, for example in the 'moving sum' (MOSUM) technique of Hušková and Slabý (2001) and Kirch and Muhsal (2014), and the (windowed) 'circular' binary segmentation of Olshen et al. (2004). We note that Matteson and James (2014) provide theoretical consistency results for a method related to the latter, but not windowed and hence computationally intensive, in the case of a bounded number of change-points.

The WBS procedure is computationally fast, consistent, as well as being provably better than BS and near-optimal in terms of the rates of convergence of the estimated locations of change-points even for very short spacings between neighbouring change-points and for N increasing with T. It also performs very well in practice and is easy to code. Its R implementation is provided in the R package `wbs` [Baranowski and Fryzlewicz (2014)], available from CRAN.

The paper is organised as follows. In Section 2, we motivate the WBS procedure. In Section 3, we recall standard binary segmentation (with some new consistency results) and outline the WBS technique in more detail, also with corresponding results. In Section 4, we give recommendations on default parameter values and illustrate the performance of WBS in a comparative simulation study. In Section 5, we exhibit its performance in the problem of segmenting a time series arising in finance.

2. Motivation. In this work, we consider the model

$$X_t = f_t + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{2}$$

where f_t is a deterministic, one-dimensional, piecewise-constant signal with change-points whose number N and locations $\eta_1, \ldots, \eta_N$ are unknown. Further technical assumptions on f_t and ε_t will be specified later.

The basic ingredient of both the standard BS algorithm and WBS is the CUSUM statistic defined by the inner product between the vector $(X_s, \ldots, X_e)$ and a particular vector of 'contrast' weights given below:

$$\tilde{X}^b_{s,e} = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} X_t - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} X_t, \tag{3}$$

where $s \le b < e$, with $n = e - s + 1$. It is used in different ways in both algorithms. In its first step, the BS algorithm computes $\tilde{X}^b_{1,T}$ and then takes $b_{1,1} = \arg\max_{b:1\le b<T} |\tilde{X}^b_{1,T}|$ to be the first change-point candidate, whose significance is to be judged against a certain criterion. If it is considered significant, the domain $[1, T]$ is split into two sub-intervals to the left and to the right of $b_{1,1}$ (hence the name 'binary segmentation'), and the recursion continues by computing $\tilde{X}^b_{1,b_{1,1}}$ and $\tilde{X}^b_{b_{1,1}+1,T}$, possibly resulting in further splits. The complete BS algorithm is outlined in Section 3.2.

We note that the maximisation of $|\tilde{X}^b_{s,e}|$ is equivalent to the least squares fit of a piecewise-constant function with one change-point to $X^e_s = (X_s, \ldots, X_e)'$, in the following sense. Define $\mathcal{F}^b_{s,e}$ to be the set of vectors supported on $[s, e]$ with a single change-point at b. We have

$$\arg\max_{b:s\le b<e} |\tilde{X}^b_{s,e}| = \arg\min_{b:s\le b<e} \min_{\bar{f}^b_{s,e}\in\mathcal{F}^b_{s,e}} \|X^e_s - \bar{f}^b_{s,e}\|_2^2.$$

Therefore, if the true function f_t contains only one change-point b_0 on $[s, e]$, then $\hat{b}_0 = \arg\max_{b:s\le b<e} |\tilde{X}^b_{s,e}|$ is the least-squares estimator of b_0, coinciding with the MLE in the case of ε_t being i.i.d. Gaussian. Speaking heuristically, this means that if f_t contains only one change-point on its entire domain $[1, T]$, then $b_{1,1}$, the estimator of its location from the first step of the BS algorithm, is likely to perform well.

However, in the case of more than one change-point, the first step of the BS algorithm amounts to fitting $\bar{f}^b_{1,T}$, a function with a single change-point, to data with underlying multiple change-points, that is, to fitting the wrong model. This may have disastrous consequences, as the following example demonstrates.

The function $\{f_t\}_{t=1}^{300}$ in Figure 1 has three change-points (at $t = 130, 150, 170$) which are concentrated in the middle of f_t, and which 'work against each other' in the sense that the jump at $t = 150$ is offset by the two jumps at $t = 130, 170$. In the first step of BS, $|\tilde{X}^b_{1,300}|$ is computed. However, because of this unfavourable configuration of the change-points, its maximum, occuring around $b = 100$, completely misses all of them.

On the other hand, $|\tilde{X}^b_{101,200}|$ is successful in locating the middle change-point. Heuristically speaking, this is because the localised feature (defined by the three

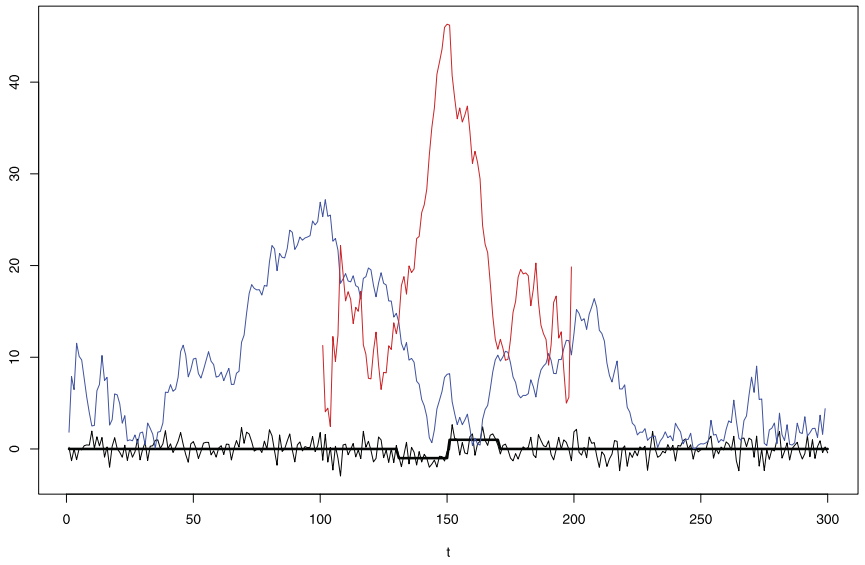

FIG. 1. *True function* f_t, $t = 1, \ldots, T = 300$ (*thick black*), *observed* X_t (*thin black*), $|\tilde{X}^b_{1,300}|$ *plotted for* $b = 1, \ldots, 299$ (*blue*), *and* $|\tilde{X}^b_{101,200}|$ *plotted for* $b = 101, \ldots, 199$ (*red*).

change-points) is more 'obvious' when considered as part of the interval $[101, 200]$ than $[1, 300]$, in the sense of the absolute inner product $|\tilde{f}^{150}_{101,200}|$ being much higher than $|\tilde{f}^{150}_{1,300}|$ [where $\tilde{f}^b_{s,e}$ is defined as in (3) but with X replaced by f]. This effect would be even more pronounced if we 'moved' the starting point of the inner product from $s = 101$ towards the first change-point $t = 130$, and analogously the end point $e = 200$ towards $t = 170$. In this example, the inner product $|\tilde{f}^{150}_{s,e}|$ is maximised exactly when $s = 131$, $e = 170$ (i.e., when s, e coincide with the two outside change-points), as this creates the 'maximal' interval $[s, e]$ containing only the one change-point at $t = 150$. This is further illustrated in Figure 2.

Obviously, in practice, we cannot use the knowledge of the change-point locations to choose favourable locations for the start-point s and the endpoint e of the inner product $|\tilde{X}^b_{s,e}|$. We also cannot test all possible locations s, e as this would be computationally prohibitive. Our main proposal in this work is to randomly draw a number of pairs (s, e) and find $\arg\max_{b:s\le b<e} |\tilde{X}^b_{s,e}|$ for each draw. If the number of draws is suitably large, we will be able to guarantee, with high probability, a particularly favourable draw for which $[s, e]$ is long enough and only contains one change-point at a sufficient distance from its endpoints (or is sufficiently 'close' to that situation, as in the example above). The hope is that $\arg\max_{b:s\le b<e} |\tilde{X}^b_{s,e}|$ corresponding to that particular draw will be a clear indicator of a true change-point in f_t. One perhaps surprising aspect of this procedure is that the number of draws

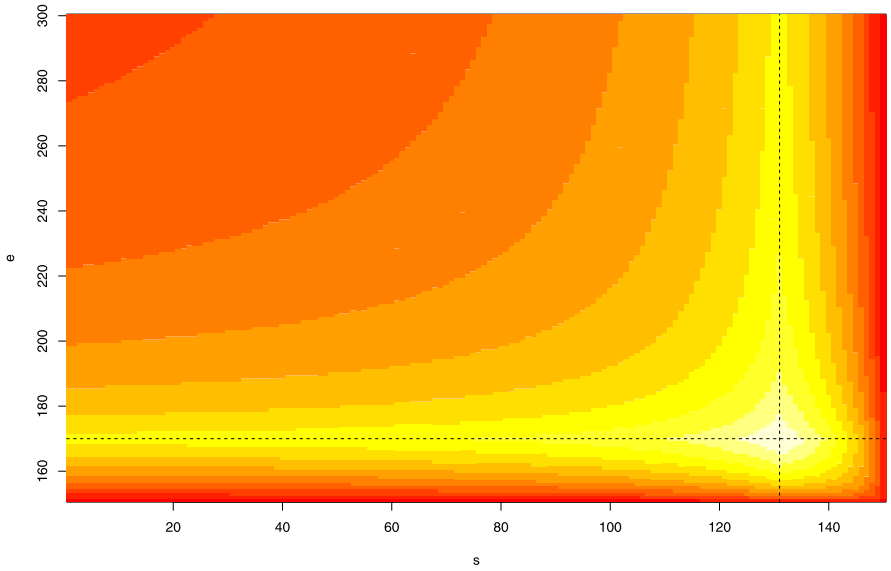

FIG. 2. *Heat map of the values of* $|\tilde{f}^{150}_{s,e}|$ *as a function of* s *and* e*: the lighter the colour, the higher the value. The two dashed lines indicate the location of the maximum,* $(s, e) = (131, 170)$.

guaranteed to achieve this (for all change-points at once) is not large, as will be shown later.

This motivating discussion leads us to propose, in the next section, the wild binary segmentation algorithm for multiple change-point detection.

3. Methodology and theory of wild binary segmentation.

3.1. *Model and technical assumptions.* We make the following assumption.

ASSUMPTION 3.1. (i) The random sequence $\{\varepsilon_t\}_{t=1}^T$ is i.i.d. Gaussian with mean zero and variance 1.

(ii) The sequence $\{f_t\}_{t=1}^T$ is bounded, that is, $|f_t| < \bar{f} < \infty$ for $t = 1, \ldots, T$.

Assumption 3.1(i) is made both for technical convenience and for clarity of exposition; it is reasonable to expect that it could in principle be extended to dependent, heterogeneous and/or non-Gaussian noise. We assume that $\mathrm{Var}(\varepsilon_t)$ is known, the reason being that in practice it can usually be estimated accurately using, for example, median absolute deviation [Hampel (1974)]. Such an assumption is standard in the literature on function estimation in Gaussian noise.

Different assumptions on the spacing between change-points and on the jump magnitudes will be needed by standard binary segmentation and by WBS. In what follows, denote $\eta_0 = 0$, $\eta_{N+1} = T$.

ASSUMPTION 3.2 (for standard binary segmentation). The minimum spacing between change-points satisfies $\min_{i=1,\ldots,N+1} |\eta_i - \eta_{i-1}| \geq \delta_T$, where $\delta_T \geq CT^{\Theta}$ for $C > 0$, with $\Theta \leq 1$. In addition, the magnitudes $f_i' = |f_{\eta_i} - f_{\eta_i - 1}|$ of the jumps satisfy $\min_{i=1,\ldots,N} f_i' \geq \underline{f}_T$, where $\underline{f}_T \geq CT^{-\varpi}$, with $\varpi \geq 0$. The parameters Θ and ϖ satisfy $\Theta - \frac{\varpi}{2} > \frac{3}{4}$.

ASSUMPTION 3.3 (for WBS). The minimum spacing between change-points satisfies $\min_{i=1,\ldots,N+1} |\eta_i - \eta_{i-1}| \geq \delta_T$, and the magnitudes $f_i' = |f_{\eta_i} - f_{\eta_i - 1}|$ of the jumps satisfy $\min_{i=1,\ldots,N} f_i' \geq \underline{f}_T$, where δ_T and $\underline{f}_T$ are linked by the requirement $\delta_T^{1/2} \underline{f}_T \geq C \log^{1/2} T$ for a large enough C.

It is worth noting that we do not assume any further upper bounds on the number N of change-points, other than those implied by the minimum spacing δ_T. In other words, N can be as large as allowed by δ_T, and in particular can increase to infinity with T. Therefore, formally, we have $N = N(T)$ and $\eta_i = \eta_i(T)$ for $i = 1, \ldots, N + 1$. However, for economy of notation and keeping in line with many other papers on change-point detection, in the remainder of the paper we use the shorthand notation N, η_i rather than the longer notation $N(T), \eta_i(T)$.

The quantity $\delta_T^{1/2} \underline{f}_T$ appearing in Assumption 3.3 is well known in the 'statistical signal detection' literature. For example, Chan and Walther (2013) summarise results which show that detection of hat-shaped signals observed in Gaussian noise is impossible if (the equivalent of) this quantity is below a certain threshold. See also Dümbgen and Spokoiny (2001) and Frick, Munk and Sieling (2014) for related discussions. We will argue in Section 3.2 that our Assumption 3.3 is rate-near-optimal from this point of view.

3.2. *Standard binary segmentation.* To gain a better understanding of the improvement offered by WBS over standard BS, we first provide a theoretical consistency result for the latter. The BS algorithm is best defined recursively and hence described by pseudocode. The main function is defined as follows.

function BINSEG(s, e, ζ_T)
  if $e - s < 1$ **then**
    STOP
  else
    $b_0 := \arg\max_{b \in \{s,\ldots,e-1\}} |\tilde{X}^b_{s,e}|$
    if $|\tilde{X}^{b_0}_{s,e}| > \zeta_T$ **then**
      add b_0 to the set of estimated change-points
      BINSEG(s, b_0, ζ_T)
      BINSEG($b_0 + 1$, e, ζ_T)
    else
      STOP

end if
end if
end function

Given the above definition, the standard BS procedure is launched by the call BINSEG$(1, T, \zeta_T)$, where ζ_T is a threshold parameter. Let $\hat{N}$ denote the number of change-points estimated by the BS algorithm, and $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$ their locations, sorted in increasing order. The following consistency theorem holds.

THEOREM 3.1. *Let X_t follow model (2), and suppose Assumptions 3.1 and 3.2 hold. Let N and $\eta_1, \ldots, \eta_N$ denote, respectively, the number and locations of change-points. Let $\hat{N}$ denote the number, and $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$ the locations, sorted in increasing order, of the change-point estimates obtained by the standard binary segmentation algorithm. Let the threshold parameter satisfy $\zeta_T = c_1 T^\theta$ where $\theta \in (1-\Theta, \Theta - 1/2 - \varpi)$ if $\Theta \in (\frac{3}{4}, 1)$, or $\zeta_T \geq c_2 \log^p T$ $(p > 1/2)$ and $\zeta_T \leq c_3 T^\theta$ $(\theta < 1/2 - \varpi)$ if $\Theta = 1$, for any positive constants c_1, c_2, c_3. Then there exist positive constants C, C_1 such that $P(\mathcal{A}_T) \geq 1 - C_1 T^{-1}$, where*

$$\mathcal{A}_T = \Big\{ \hat{N} = N; \max_{i=1,\ldots,N} |\hat{\eta}_i - \eta_i| \leq C\epsilon_T \Big\}$$

with $\epsilon_T = T^2 \delta_T^{-2} (\underline{f}_T)^{-2} \log T$.

We note that the rates of convergence of $\hat{\eta}_i$ are better than those obtained by Venkatraman (1992) and Fryzlewicz and Subba Rao (2014), both of which consider consistency of the BS procedure for the number of change-points N possibly increasing with T; they are also better than those in Cho and Fryzlewicz (2012) (where N is assumed to be bounded). The latter three papers use the assumption that $\underline{f}_T$ is bounded away from zero. The improvement is due to the crucial and new Lemma A.3. Rates are particularly important here, as they inform the stopping criterion (i.e., the admissible magnitude of the threshold ζ_T), rather than merely quantifying the performance of the procedure.

As an aside, we mention that in the case $\delta_T = o(T)$, it is possible to further improve our rates via a simple trick, whereby change-point locations are re-estimated by maximising the CUSUM statistic $|\tilde{X}^b_{s,e}|$ on each interval $[s, e]$ where s, e are respective mid-points of two adjacent intervals $[\hat{\eta}_{i-1} + 1, \hat{\eta}_i]$, $[\hat{\eta}_i + 1, \hat{\eta}_{i+1}]$ (with the convention $\hat{\eta}_0 = 0$, $\hat{\eta}_{\hat{N}+1} = T$). This refinement can be applied to any multiple change-point detection procedure, not just BS. However, even with this refinement, the BS procedure as defined above is only guaranteed to produce valid results under Assumption 3.2, which is rather restrictive in terms of the permitted distance between change-points and the magnitudes of the jumps.

3.3. *Wild binary segmentation.* Denote by F_T^M a set of M random intervals $[s_m, e_m]$, $m = 1, \ldots, M$, whose start- and end-points have been drawn (independently with replacement) uniformly from the set $\{1, \ldots, T\}$. Guidance as to a suitable choice of M will be given later. Again using pseudocode, the main function of the WBS algorithm is defined as follows.

function WILDBINSEG(s, e, ζ_T)
  if $e - s < 1$ **then**
    STOP
  else
    $\mathcal{M}_{s,e} :=$ set of those indices m for which $[s_m, e_m] \in F_T^M$ is such that $[s_m, e_m] \subseteq [s, e]$
    (Optional: augment $\mathcal{M}_{s,e} := \mathcal{M}_{s,e} \cup \{0\}$, where $[s_0, e_0] = [s, e]$)
    $(m_0, b_0) := \arg\max_{m \in \mathcal{M}_{s,e}, b \in \{s_m, \ldots, e_m - 1\}} |\tilde{X}^b_{s_m, e_m}|$
    if $|\tilde{X}^{b_0}_{s_{m_0}, e_{m_0}}| > \zeta_T$ **then**
      add b_0 to the set of estimated change-points
      WILDBINSEG(s, b_0, ζ_T)
      WILDBINSEG($b_0 + 1$, e, ζ_T)
    else
      STOP
    end if
  end if
end function

The WBS procedure is launched by the call WILDBINSEG(1, T, ζ_T). We believe that the WBS procedure is not difficult to code even for the nonexpert, unlike some change-point detection algorithms based on dynamic programming. Let $\hat{N}$ denote the number of change-points estimated by the WBS procedure, and $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$ their locations, sorted in increasing order.

The optional augmentation of $\mathcal{M}_{s,e}$ by $\{0\}$ is done to ensure that the algorithm also examines the entire current interval $[s, e]$, and not only its randomly drawn subintervals, in case $[s, e]$ only contains one change-point and hence it is optimal to examine $[s, e]$ in its entirety. We note that unlike the BS procedure, the WBS algorithm (in the case without the optional augmentation) returns estimated change-points in the order corresponding to decreasing maxima of $|\tilde{X}^b_{s_m, e_m}|$, which is due to the maximisation over m. There is no corresponding maximisation in the BS procedure, which means that the maxima of the CUSUM statistics corresponding to estimated change-points in the latter procedure are not necessarily arranged in decreasing order.

Finally, we motivate the use of random, rather than fixed, intervals. As demonstrated in Section 2, some change-points require narrow intervals $[s, e]$ around them in order to be detectable. For such change-points, the use of random intervals, as in the WBS algorithm, means that there is always a positive probability, sometimes high, of there being a suitably narrow interval around them in the set

F_T^M. On the other hand, consider a fixed design, where the start-points s_m and end-points e_m take all possible values from a fixed subset of $\{1, \ldots, T\}$, of such cardinality that the number of resulting intervals is the same as in the random design. For such a fixed design (however it is chosen), at least some of the intervals will inevitably be significantly longer than the corresponding random ones, so that they may not permit detection of such change-points if those happen to lie within them. Another reason is that through the use of randomness, we avoid having to make the subjective choice of a particular fixed design. Finally, if the number of intervals drawn turns out to be insufficient, it is particularly easy to add further intervals if the design is random; this is achieved simply by drawing further intervals from the same distribution. In the case of a fixed design, the entire collection may need to be re-drawn if the distribution of interval lengths is to be preserved. However, for a very large number M of intervals, the difference in performance between the random and deterministic designs is likely to be minimal.

The following theoretical result holds for the WBS algorithm.

THEOREM 3.2. *Let X_t follow model (2), and suppose Assumptions 3.1 and 3.3 hold. Let N and $\eta_1, \ldots, \eta_N$ denote, respectively, the number and locations of change-points. Let $\hat{N}$ denote the number, and $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$ the locations, sorted in increasing order, of the change-point estimates obtained by the wild binary segmentation algorithm. There exist two constants $\underline{C}$, $\overline{C}$ such that if $\underline{C} \log^{1/2} T \le \zeta_T \le \overline{C} \delta_T^{1/2} \underline{f}_T$, then $P(\mathcal{A}_T) \ge 1 - C_1 T^{-1} - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/9)^M$, where*

$$\mathcal{A}_T = \Big\{ \hat{N} = N; \max_{i=1,\ldots,N} |\hat{\eta}_i - \eta_i| \le C \log T (\underline{f}_T)^{-2} \Big\}$$

for certain positive C, C_1.

Some remarks are in order. Firstly, we note that Assumption 3.3 is much milder than Assumption 3.2. As an illustration, consider the case when $\underline{f}_T$ is bounded away from zero (although we emphasise that both algorithms permit $\underline{f}_T \to 0$, albeit at different rates). In this case, the WBS method produces consistent results even if the minimum spacing δ_T between the true change-points is logarithmic in T, whereas δ_T must be larger than $O(T^{3/4})$ in standard BS. Furthermore, for a given separation δ_T and minimum jump height $\underline{f}_T$, the admissible range of threshold rates for the WBS method is always larger than that for BS. In this sense, the WBS method may be viewed as more robust than BS to the possible misspecification of the value of the threshold.

Secondly, unlike the BS algorithm, the lower bound for the threshold ζ_T in the WBS method is always square-root logarithmic in T, irrespective of the spacing δ_T. This is also the only threshold rate that yields consistency for any admissible separation δ_T and minimum jump size $\underline{f}_T$. For this reason, we use the rate $\log^{1/2} T$ as the default rate for the magnitude of the threshold, and hence, in the

remainder of the article, we consider thresholds of the form $\zeta_T = C\sqrt{2}\log^{1/2} T$ (we introduce the factor of $\sqrt{2}$ in order to facilitate the comparison of ζ_T to the 'universal' threshold in the wavelet thresholding literature, which is of the form $\sqrt{2}\log^{1/2} T$). Practical choice of the constant C will be discussed in Section 4. In BS, the only threshold rate that leads to consistency for any admissible δ_T is $\zeta_T \sim T^{1/4-\varpi/2}$ (where $\sim$ means 'of the order of' throughout the paper).

Thirdly, again unlike the BS algorithm, the rate of convergence of the estimated change-point locations in the WBS method does not depend on the spacing δ_T (as long as $\delta_T^{1/2} \underline{f}_T$ is large enough in the sense of Assumption 3.3) but only on the minimum jump height $\underline{f}_T$. We now consider the special case of $\underline{f}_T$ being bounded away from zero, and discuss the optimality, up to at most a logarithmic factor, of wild binary segmentation in estimating the change-point locations in this setting. In the case $\delta_T \sim T$, the optimal rate in detecting change-point locations is $O_P(1)$ in the sense that for any estimator $\hat{\eta}_i$ of η_i, we have $|\hat{\eta}_i - \eta_i| = O_P(1)$ at best; see, for example, Korostelëv (1987). This can be reformulated as $P(|\hat{\eta}_i - \eta_i| \geq a_T) \to 0$ for any sequence $a_T \to \infty$. In the case $\underline{f}_T > \underline{f} > 0$, the result of Theorem 3.2 implies $P(\exists_i |\hat{\eta}_i - \eta_i| \geq C \log T) \to 0$, thus matching the above minimax result up to a logarithmic term. However, we emphasise that this is in the (more challenging) context where (i) the number N of change-points is possibly unbounded with T, and (ii) the spacing δ_T between change-points can be much shorter than of order T.

We now discuss the issue of the minimum number M of random draws needed to ensure that the bound on the speed of convergence of $P(\mathcal{A}_T)$ to 1 in Theorem 3.2 is suitably small. Suppose that we wish to ensure

$$T\delta_T^{-1}\big(1 - \delta_T^2 T^{-2}/9\big)^M \leq T^{-1}$$

in order to match the rate of the term $C_1 T^{-1}$ in the upper bound for $1 - P(\mathcal{A}_T)$ in Theorem 3.2. Bearing in mind that $\log(1-y) \approx -y$ around $y = 0$, this is, after simple algebra, (practically) equivalent to

$$M \geq \frac{9T^2}{\delta_T^2} \log\big(T^2 \delta_T^{-1}\big).$$

In the 'easiest' case $\delta_T \sim T$, this results in a logarithmic number of draws, which leads to particularly low computational complexity. Naturally, the required M progressively increases as δ_T decreases. Our practical recommendations for the choice of M are discussed in Section 4.

Furthermore, we explain why the binary recursion is needed in the WBS algorithm at all: the careful reader may wonder why change-points are not estimated simply by taking *all* those points that attain the maxima of $|\tilde{X}^b_{s_m,e_m}|$ exceeding the threshold ζ_T, for all intervals $[s_m, e_m] \in F_T^M$. This is because such a procedure would very likely lead to some true change-points being estimated more than once at different locations. By proceeding sequentially as in the WBS algorithm, and by restricting ourselves to those intervals $[s_m, e_m]$ that fully fall within the current

interval of interest $[s, e]$, we ensure that this problem does not arise. Another reason for proceeding sequentially is the optional augmentation of $\mathcal{M}_{s,e}$ by $\{0\}$ in the WBS algorithm, which depends on the previously detected change-points and hence is not feasible in a nonsequential setting.

Regarding the optimality of the lowest permitted rate for $\delta_T^{1/2}\underline{f}_T$ in Assumption 3.3, recall that, by Theorem 3.2, δ_T must be at least as large as $\max_{i=1,\dots,N}|\hat{\eta}_i - \eta_i|$, or it would not be possible to match the estimated change-point locations with the true ones. Therefore, δ_T cannot be of a smaller order than $\log T$. By the minimax arguments summarised in Chan and Walther (2013) (but using our notation), the rate of the smallest possible $\delta_T^{1/2}\underline{f}_T$ that permits change-point detection (by any method) for this range of δ_T is $(\log T - \log\log T)^{1/2}$. Our Assumption 3.2 achieves this rate up to the negligible double-logarithmic factor and therefore is optimal under the circumstances.

Randomised methods are not commonly used in nonparametric statistics (indeed, we are not aware of any other commonly used such method); however, randomised techniques are beginning to make headway in statistics in the context of 'big data'; see, for example, the review articles Mahoney (2010) and Halko, Martinsson and Tropp (2011). The proof technique in Theorem 3.2 relies on some subtle arguments regarding the guarantees of quality of the randomly drawn intervals.

3.4. *Strengthened Schwarz information criterion for WBS.* Naturally, the estimated number $\hat{N}$ and locations $\hat{\eta}_1, \dots, \hat{\eta}_{\hat{N}}$ of change-points depend on the selected threshold ζ_T. For the purpose of this paragraph, denote $\hat{N}(\zeta_T) = \hat{N}$ and $\mathcal{C}(\zeta_T) = \{\hat{\eta}_1, \dots, \hat{\eta}_{\hat{N}(\zeta_T)}\}$. It is a property of the WBS method that $\hat{N}(\zeta_T)$ is a nondecreasing function of ζ_T, each increase has size 1 almost-surely, and the collection $\mathcal{C}(\zeta_T)$ is nested in the sense that if $\zeta_T' < \zeta_T''$ then $\mathcal{C}(\zeta_T'') \subseteq \mathcal{C}(\zeta_T')$. Consider any decreasing sequence $\{\zeta_T^k\}_{k=0}^K$ of thresholds such that $|\mathcal{C}(\zeta_T^k)| = k$ for a certain fixed constant K, and assume $N \le K$. One may perform model selection either by choosing a suitable threshold ζ_T and hence selecting the associated model $\mathcal{C}(\zeta_T)$, or alternatively by considering the sequence of model candidates $\{\mathcal{C}(\zeta_T^k)\}_{k=0}^K$ and choosing one that optimises a certain criterion, thereby by-passing the question of threshold choice entirely. Thus it is a viable alternative to view the 'solution path' $\mathcal{C}(\zeta_T^k)$ not as a function of threshold ζ_T^k, but as a function of the number k of change-point candidates. We define $\mathcal{C}_k = \mathcal{C}(\zeta_T^k)$. In this section, we propose to select a model out of the collection $\{\mathcal{C}_k\}_{k=0}^K$ by minimising what we term the 'strengthened Schwarz information criterion' (sSIC), defined as follows.

For any candidate model $\mathcal{C}_k$, denote by $\hat{f}_t^k$ the estimate of f_t defined by $\hat{f}_t^k = (\hat{\eta}_{i+1} - \hat{\eta}_i)^{-1}\sum_{j=\hat{\eta}_i+1}^{\hat{\eta}_{i+1}} X_j$ for $\hat{\eta}_i + 1 \le t \le \hat{\eta}_{i+1}$. Let $\hat{\sigma}_k^2 = T^{-1}\sum_{t=1}^T (X_t - \hat{f}_t^k)^2$ be the corresponding maximum likelihood estimator of the residual variance. We

define

$$\text{sSIC}(k) = \frac{T}{2}\log\hat{\sigma}_k^2 + k\log^\alpha T. \tag{4}$$

We remark that the choice $\alpha = 1$ corresponds to the standard SIC penalty, considered, for example, by Yao (1988) in the context of multiple change-point detection in a model similar to ours performed via a full penalised least-squares minimisation. The following result holds.

THEOREM 3.3. *Let X_t follow model (2), and let the assumptions of Theorem 3.2 hold. Let N and $\eta_1, \ldots, \eta_N$ denote, respectively, the number and locations of change-points. Let $N \le K$, where K is a certain constant independent of T. Let the constant $\alpha > 1$ be such that $\log^\alpha T = o(\delta_T \underline{f}_T^2)$. Let the candidate models $\{\mathcal{C}_k\}_{k=0}^K$ be produced by the WBS algorithm, and let $\hat{N} = \arg\min_{k=0,\ldots,K} \text{sSIC}(k)$. Then $P(\mathcal{A}_T) \ge 1 - C_1 T^{-1} - T\delta_T^{-1}(1 - \delta_T^2 T^{-2}/9)^M$, where*

$$\mathcal{A}_T = \Big\{\hat{N} = N; \max_{i=1,\ldots,N} |\hat{\eta}_i - \eta_i| \le C\log T(\underline{f}_T)^{-2}\Big\}$$

for certain positive C, C_1.

The only parameter of the above procedure is the constant α, and we require that $\alpha > 1$, which results in a stronger penalty than in the standard SIC, hence the term 'strengthened' SIC. Noting the requirement that $\log^\alpha T = o(\delta_T \underline{f}_T^2)$, we focus attention on values of α close to 1, to ensure the admissibility of the sSIC criterion for as large a class of signals as possible; from this point of view, it is tempting to regard this region of the parameter space for α as a natural default choice. With this in mind, in the remainder of the paper, we report the performance of sSIC with $\alpha = 1.01$, which also ensures that the results remain close to those obtained by SIC.

We further note that unlike in thresholding, where the magnitude of the threshold is sensitive to $\text{Var}(\varepsilon_t)$, the minimisation of the sSIC penalty in (4) is independent of $\text{Var}(\varepsilon_t)$ due to the use of the logarithmic transformation in $\log\hat{\sigma}_k^2$. This logarithmic transformation causes $\text{Var}(\varepsilon_t)$ to have an additive contribution to the sSIC criterion in (4), and therefore this term has no impact on the minimisation.

In summary, the attraction of the sSIC approach lies in the fact that the default choice of the parameter of the procedure is perhaps easier than in the thresholding approach. On the other hand, the theoretical validity of sSIC in the version of Theorem 3.3 requires that $N \le K$ for a finite K and that the lowest admissible $\delta_T \underline{f}_T^2$ is (marginally) larger than in the thresholding approach. The requirement of a finite K is common to penalised approaches to multiple change-point detection; see, for example, Yao (1988) and Ciuperca (2014).

4. Parameter choice and simulation study.

4.1. *Parameter choice*. We now elaborate on the choice of the number M of the random draws, and the threshold constant C.

Choice of M. The parameter M should be chosen to be 'as large as possible' subject to computational constraints. We note that with the optional augmentation of $\mathcal{M}_{s,e}$ by $\{0\}$, the WBS reduces to standard BS for $M = 0$, so even a relatively small value of M is likely to bring benefits in terms of performance. Our recommendation is to set $M = 5000$ for datasets of length T not exceeding a few thousand. As an example, with this value of M, we achieved the average computation time of 1.20 seconds for a dataset of length $T = 2000$. The code was written in a combination of R and C, and executed on a 3.40 GHz quad-core with 8 GB of RAM, running Windows 7. The implementation of WBS in the R package `wbs` is faster still.

Moreover, the larger the value of M, the more negligible the dependence of the solution on the particular random draw. For $M = 5000$, this dependence has been observed to be very minimal.

Choice of the threshold constant C. In Section 3.3, we motivate the use of thresholds of the form $\zeta_T = C\sqrt{2}\log^{1/2} T$. There remains the question of how to choose the threshold constant C. We firstly remark that from the theoretical point of view, it is challenging to propose a particular choice of C without having a specific cost function in mind, which the thresholding approach inherently avoids. Therefore, one possibility is to use a large-scale simulation study to select a default value of C that works well across a range of signals.

With this in mind, we conducted the following simulation study. For a given average number $N_{\mathrm{avg}} \in \{4, 8\}$ of change-points, we simulated a Poisson number of change-points $N = \mathrm{Pois}(N_{\mathrm{avg}})$ and distributed them uniformly on $[0, 1]$. At each change-point, we introduced a jump whose height had been drawn from the normal distribution with mean zero and variance $\sigma^2_{\mathrm{jmp}} \in \{1, 3, 10\}$. We sampled the thus-constructed function at $T \in \{100, 200, 500, 1000, 2000\}$ equispaced points, and contaminated it with Gaussian noise with mean zero and variance one. Based on a large number of replicates, we considered the quantity $|\hat{N} - N|$, where $\hat{N}$ was produced by the WBS algorithm with threshold $\zeta_T = C\sqrt{2}\log^{1/2} T$, and found the value of C that minimised it. The minimiser was sufficiently close to $C = 1$ for us to use this value as the default one.

We add that our theoretical results do not permit a data-dependent choice of the threshold constant C, so having a reliable default choice is essential. The hope is that choosing such a default constant via extensive simulation should lead to good calibration of our method for a wide range of signals.

When the variance of ε_t is unknown, we use $\zeta_T = \hat{\sigma}\, C(2\log T)^{1/2}$, where $\hat{\sigma}$ is the median absolute deviation estimator of $\mathrm{Var}^{1/2}(\varepsilon_t)$.

Finally, we remark that in our comparative simulation study reported below, we apply two threshold constants: the default value of $C = 1$ and a higher value of

$C = 1.3$. The latter is used for comparative purposes as it was also used in the example considered in Fryzlewicz (2014).

Users with a preference for a method whose default parameters are not chosen by simulation are encouraged to use the WBS method with the sSIC stopping criterion described in Section 3.4, rather than with thresholding. This method is also part of the simulation study below.

4.2. *Simulation study.* In this section, we compare the performance of WBS (and BS) against the best available competitors implemented in R packages, most of which are publicly available on CRAN. The competing packages are: `strucchange`, which implements the multiple change-point detection method of Bai and Perron (2003), `Segmentor3IsBack`, which implements the method of Rigaill (2010) with the model selection methodology from Lebarbier (2005), `changepoint`, which implements the PELT methodology of Killick, Fearnhead and Eckley (2012), `cumSeg`, which implements the method from Muggeo and Adelfio (2011), and `stepR`, which implements the SMUCE method of Frick, Munk and Sieling (2014). In the remainder of this section, we refer to these methods as, respectively, B&P, S3IB, PELT, cumSeg, and SMUCE. Appendix B provides an extra discussion of how these methods were used in our simulation study. With the exception of `stepR`, which is available from http://www.stochastik.math.uni-goettingen.de/smuce at the time of writing, the remaining packages are available on CRAN.

In this section, the WBS algorithm uses the default value of $M = 5000$ random draws. In the thresholding stopping rule, we use the threshold $\zeta_T = C\hat{\sigma}\sqrt{2\log T}$, where $\hat{\sigma}$ is the median absolute deviation estimator of σ suitable for i.i.d. Gaussian noise, T is the sample size, and the constant C is set to 1 and 1.3 as motivated earlier. The WBS method combined with the sSIC stopping criterion is referred to as 'WBS sSIC' and uses $\alpha = 1.01$, again as justified earlier, and $K = 20$. The BS method uses the same thresholds as WBS, for comparability.

Our test signals, fully specified in Appendix B along with the sample sizes and noise standard deviations used, are (1) `blocks`, (2) `fms`, (3) `mix`, (4) `teeth10`, and (5) `stairs10`. Tables 1 and 2 show the results. We describe the performance of each method below.

B&P. The B&P method performs poorly, which may be partly due to the default minimum segment size set to 15% of the sample size, an assumption violated by several of our test signals. However, resetting this parameter to 1 or even 1% of the sample size resulted in exceptionally slow computation times, which prevented us from reporting the results in our comparisons.

S3IB. This method offers excellent performance for the `blocks` signal, and very good performance for the `fms` signal. The `mix` signal is more challenging, and the S3IB method does not perform well here, with a tendency to underestimate the number of change-points, sometimes by as many as 12. Performance is rather

TABLE 1

Distribution of $\hat{N} - N$ for the various competing methods and models, over 100 simulated sample paths. Also the average mean-square error of the resulting estimate of f_t. Bold: methods with the highest empirical frequency of $\hat{N} - N = 0$, and those with frequencies within 10% off the highest

		$\hat{N} - N$							
Method	**Model**	**≤ -3**	**-2**	**-1**	**0**	**1**	**2**	**≥ 3**	**MSE**
PELT	(1)	0	0	0	8	9	9	74	4.3
B&P		100	0	0	0	0	0	0	14.3
cumSeg		53	21	24	2	0	0	0	7.26
S3IB		0	5	42	**51**	1	1	0	2.55
SMUCE		54	42	4	0	0	0	0	6.66
WBS $C = 1.0$		0	0	24	38	15	18	5	2.77
WBS $C = 1.3$		1	13	78	8	0	0	0	3.02
WBS sSIC		0	1	51	**46**	2	0	0	2.65
BS $C = 1.0$		0	1	40	39	16	2	2	3.12
BS $C = 1.3$		9	27	56	6	2	0	0	4.27
PELT	(2)	0	0	0	15	11	20	54	79×10^{-4}
B&P		99	1	0	0	0	0	0	399×10^{-4}
cumSeg		0	73	1	24	1	1	0	127×10^{-4}
S3IB		0	0	0	**89**	9	2	0	37×10^{-4}
SMUCE		0	8	46	46	0	0	0	157×10^{-4}
WBS $C = 1.0$		0	0	0	32	25	16	27	54×10^{-4}
WBS $C = 1.3$		0	0	6	**92**	2	0	0	43×10^{-4}
WBS sSIC		0	0	0	**95**	5	0	0	40×10^{-4}
BS $C = 1.0$		0	0	30	49	16	4	1	75×10^{-4}
BS $C = 1.3$		1	10	65	23	1	0	0	109×10^{-4}
PELT	(3)	0	0	3	11	16	17	53	2.08
B&P		100	0	0	0	0	0	0	11.82
cumSeg		99	1	0	0	0	0	0	8.59
S3IB		34	34	18	14	0	0	0	1.96
SMUCE		63	28	8	1	0	0	0	4.35
WBS $C = 1.0$		0	9	22	**32**	21	13	3	1.67
WBS $C = 1.3$		15	41	32	12	0	0	0	1.91
WBS sSIC		7	28	23	**33**	6	1	2	1.62
BS $C = 1.0$		10	30	26	19	13	2	0	2.34
BS $C = 1.3$		80	18	2	0	0	0	0	3.99
PELT	(4)	0	0	0	38	28	18	16	55×10^{-3}
B&P		100	0	0	0	0	0	0	251×10^{-3}
cumSeg		100	0	0	0	0	0	0	251×10^{-3}
S3IB		36	16	1	47	0	0	0	116×10^{-3}
SMUCE		98	1	0	1	0	0	0	215×10^{-3}
WBS $C = 1.0$		0	1	7	**77**	11	2	2	51×10^{-3}
WBS $C = 1.3$		22	11	28	38	1	0	0	80×10^{-3}
WBS sSIC		4	1	4	**80**	7	4	0	55×10^{-3}
BS $C = 1.0$		49	9	19	14	9	0	0	129×10^{-3}
BS $C = 1.3$		94	3	3	0	0	0	0	210×10^{-3}

TABLE 1
(*Continued*)

Method	Model	$\hat{N} - N$							MSE
		≤ -3	-2	-1	**0**	**1**	**2**	≥ 3	
PELT	(5)	0	0	0	34	24	19	23	26×10^{-3}
B&P		100	0	0	0	0	0	0	554×10^{-3}
cumSeg		3	1	11	77	8	0	0	63×10^{-3}
S3IB		97	1	2	0	0	0	0	210×10^{-3}
SMUCE		64	17	11	8	0	0	0	185×10^{-3}
WBS $C = 1.0$		0	0	0	63	31	4	2	24×10^{-3}
WBS $C = 1.3$		0	0	4	**87**	9	0	0	27×10^{-3}
WBS sSIC		0	0	0	61	35	4	0	23×10^{-3}
BS $C = 1.0$		0	0	0	79	20	1	0	24×10^{-3}
BS $C = 1.3$		0	0	5	**88**	7	0	0	27×10^{-3}

average for the `teeth10` signal, and systematically poor for the `stairs10` signal.

PELT. The PELT method has a tendency to overestimate the number of change-points, which is apparent in all of the examples studied.

cumSeg. Apart from the `stairs10` signal for which it offers acceptable performance, the cumSeg method tends to heavily underestimate the number of change-points.

SMUCE. The SMUCE method tends to underestimate the true number of change-points. However, its performance for the `fms` signal is acceptable.

BS. For $C = 1$, the method performs acceptably for the `blocks` and `stairs10` signals, has rather average performance for the `fms` and `mix` signals, and performs poorly for `teeth10`. For $C = 1.3$, performance is excellent for the `stairs10` signal; otherwise poor. Overall, our test signals clearly demonstrate the shortcomings of classical binary segmentation.

WBS. With the threshold constant $C = 1$, WBS works well for the `blocks` and `stairs10` signals, although in both cases it is behind the best performers. For the `fms` signal, it tends to overestimate the number of change-points, although not by many. It offers (relatively) excellent performance for `mix` and `teeth10`.

For $C = 1.3$, WBS performs excellently for `fms` and `stairs10`, while it underestimates the number of change-points for the other signals, although again, not by many.

WBS sSIC performs the best or very close to the best for all signals bar `stairs10`; however, for the latter, if it overestimates the number of change-points, then it does so mostly by one change-point only. If one overall 'winner' were to be chosen out of the methods studied, it would clearly have to be WBS sSIC.

TABLE 2
Summary statistics for the empirical distribution of $\hat{N} - N$ for the various competing methods and models, over 100 *simulated sample paths*

		Summary of $\hat{N} - N$					
Method	**Model**	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
PELT	(1)	0	2	4	4.78	7	18
B&P		−7	−7	−7	−7	−7	−7
cumSeg		−7	−4	−3	−2.84	−1	0
S3IB		−2	−1	0	−0.49	0	2
SMUCE		−4	−3	−3	−2.63	−2	−1
WBS $C = 1.0$		−1	0	0	0.45	1	5
WBS $C = 1.3$		−3	−1	−1	−1.07	−1	0
WBS sSIC		−2	−1	−1	−0.51	0	1
BS $C = 1.0$		−2	−1	0	−0.16	0	3
BS $C = 1.3$		−3	−2	−1	−1.35	−1	1
PELT	(2)	0	1	3	3.39	5	10
B&P		−3	−3	−3	−2.99	−3	−2
cumSeg		−2	−2	−2	−1.44	0	2
S3IB		0	0	0	0.13	0	2
SMUCE		−2	−1	−1	−0.62	0	0
WBS $C = 1.0$		0	0	1	1.56	3	5
WBS $C = 1.3$		−1	0	0	−0.04	0	1
WBS sSIC		0	0	0	0.05	0	1
BS $C = 1.0$		−1	−1	0	−0.03	0	3
BS $C = 1.3$		−3	−1	−1	−0.87	−1	1
PELT	(3)	−1	1	3	3.17	4.25	12
B&P		−13	−13	−12.5	−12.44	−12	−10
cumSeg		−13	−13	−9	−9	−5	−2
S3IB		−12	−3	−2	−2.15	−1	0
SMUCE		−6	−4	−3	−2.95	−2	0
WBS $C = 1.0$		−2	−1	0	0.16	1	3
WBS $C = 1.3$		−5	−2	−2	−1.64	−1	0
WBS sSIC		−5	−2	−1	−0.88	0	4
BS $C = 1.0$		−5	−2	−1	−1.03	0	2
BS $C = 1.3$		−8	−4	−3	−3.56	−3	−1
PELT	(4)	0	0	1	1.26	2	7
B&P		−13	−13	−13	−12.98	−13	−12
cumSeg		−13	−13	−13	−12.94	−13	−11
S3IB		−13	−11.25	−2	−4.17	0	0
SMUCE		−12	−8	−6	−6.35	−5	0
WBS $C = 1.0$		−2	0	0	0.12	0	3
WBS $C = 1.3$		−8	−2	−1	−1.46	0	1
WBS sSIC		−12	0	0	−0.21	0	2
BS $C = 1.0$		−11	−4.25	−2	−2.70	−1	1
BS $C = 1.3$		−13	−11	−9	−8.42	−7	−1

TABLE 2
(*Continued*)

Method	Model	Summary of $\hat{N} - N$					
		Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
PELT	(5)	0	0	1	1.55	2	8
B&P		−9	−9	−9	−9	−9	−9
cumSeg		−5	0	0	−0.17	0	1
S3IB		−7	−7	−6	−5.71	−5	−1
SMUCE		−6	−4	−3	−2.85	−2	0
WBS $C = 1.0$		0	0	0	0.46	1	4
WBS $C = 1.3$		−1	0	0	0.05	0	1
WBS sSIC		0	0	0	0.43	1	2
BS $C = 1.0$		0	0	0	0.22	0	2
BS $C = 1.3$		−1	0	0	0.02	0	1

Our overall recommendation is to use WBS sSIC first. If the visual inspection of the residuals from the fit reveals any obvious patterns neglected by WBS sSIC, then WBS with $C = 1.3$ should be used next. Since the latter has a tendency to underestimate the number of change-points, the hope is that it does not detect any spurious ones. If patterns in residuals remain, WBS with $C = 1$ should be used next.

Furthermore, Appendix C contains a small-scale simulation study and brief discussion regarding the performance of WBS in the presence of linear trends.

5. Real data example. In this section, we apply the WBS method to the detection of trends in the S&P 500 index. We consider the time series of log-returns on the daily closing values of S&P 500, of length $T = 2000$ (i.e., approximately 8 trading years) ending 26 October 2012. We then remove the volatility of this series by fitting the GARCH(1, 1) model with Gaussian innovations, and apply the WBS procedure to the residuals X_t from the fit, both with the thresholding and the sSIC stopping criteria. To obtain a more complete picture of the estimated change-point structure, it is instructive to carry out the WBS procedure for a range of thresholds ζ_T.

The results, for ζ_T changing from 0 to 5, are presented in the 'time-threshold map' [see Fryzlewicz (2012) for more details of this generic concept] in Figure 3. The map should be read as follows. The x-coordinates of the vertical lines indicate the estimated change-point locations, detected for the range of thresholds equal to the range of the given line on the y-axis. For example, for $\zeta_T = \hat{\sigma}(2 \log T)^{1/2} \approx 3.83$, we have 5 estimated change-points, since the horizontal blue line (corresponding to $\zeta_T = 3.83$) in Figure 3 crosses 5 vertical lines. The 5 estimated change-points are concentrated in or around 3 separate locations.

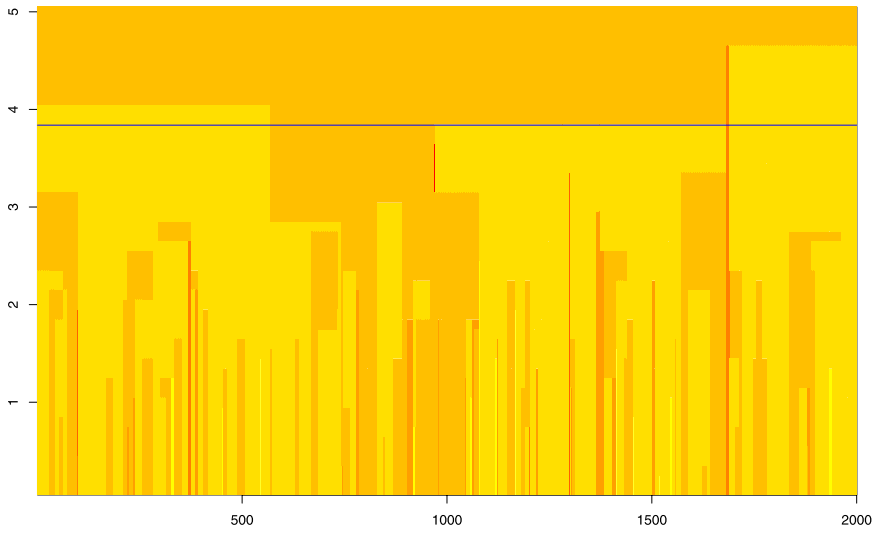

FIG. 3. *Time-threshold map of the WBS acting on the series X_t from Section 5. The horizontal blue line is the threshold $\zeta_T \approx 3.83$.*

Figure 4 shows the corresponding cumulative sum of the residuals from the GARCH fit (which can be viewed as the logged S&P 500 index with its volatility removed), with the estimated change-point locations corresponding to the thresh-

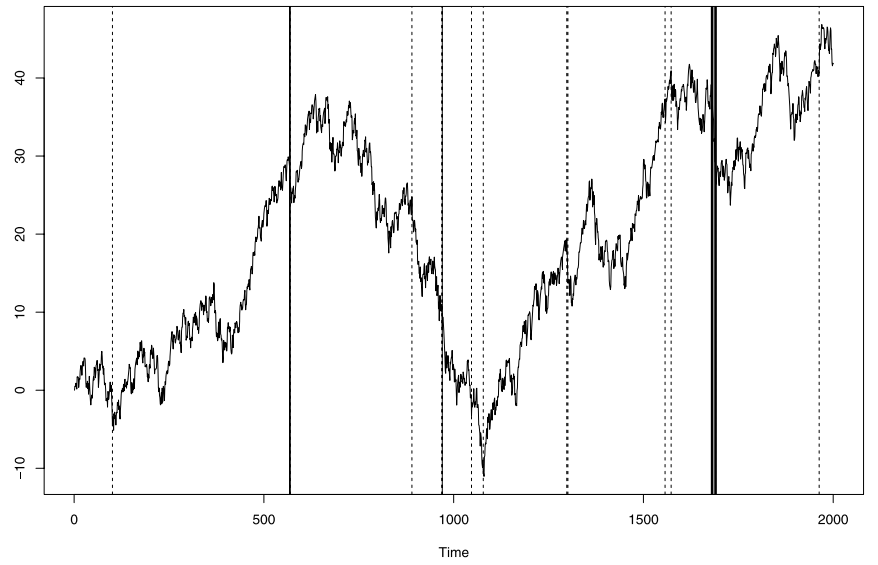

FIG. 4. *Cumulative sum of X_t, change-points corresponding to sSIC (thick solid vertical lines), $\zeta_T = 3.83$ (thin and thick solid vertical lines), $\zeta_T = 3.1$ (all vertical lines).*

olds $\zeta_T = 3.83$ and $\zeta_T = 3.1$, as well as the sSIC criterion. Interestingly, the sSIC criterion estimates only 2 change-points, both concentrated around time $t = 1700$.

As with any other financial data, it is difficult to speak of the number of estimated change-points being 'right' or 'wrong' here: for example, some more frequent traders may naturally be more interested in trend changes on the scale of weeks or months, rather than years, in which case a lower threshold might be more suitable. However, it is interesting to observe that both the sSIC criterion, the most accurate estimator of $\hat{N}$ from our simulation study, and the thresholding criterion with $\zeta_T = 3.83$, which corresponds to the threshold constant $C = 1$ and tended to slightly overestimate the number of change-points in the simulation study, point to a rather low number of estimated change-points in this example.

APPENDIX A: PROOFS

PROOF OF THEOREM 3.1. We first introduce some notation. Denoting $n = e - s + 1$, we define

$$\tilde{X}^b_{s,e} = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} X_t - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} X_t, \tag{5}$$

$$\tilde{f}^b_{s,e} = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} f_t - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} f_t. \tag{6}$$

Let s, e satisfy

$$\eta_{p_0} \le s < \eta_{p_0+1} < \cdots < \eta_{p_0+q} < e \le \eta_{p_0+q+1} \tag{7}$$

for $0 \le p_0 \le N - q$, which will be the case at all stages of the algorithm while there are still undetected change-points remaining. In Lemmas A.2–A.4, we impose the following conditions:

$$s < \eta_{p_0+r} - C\delta_T < \eta_{p_0+r} + C\delta_T < e \qquad \text{for some } 1 \le r \le q, \tag{8}$$

$$\max\bigl(\min(\eta_{p_0+1} - s, s - \eta_{p_0}), \min(\eta_{p_0+q+1} - e, e - \eta_{p_0+q})\bigr) \le C\epsilon_T. \tag{9}$$

Both (8) and (9) hold throughout the algorithm for all those segments starting at s and ending at e which contain previously undetected change-points. As Lemma A.5 concerns the case where all change-points have been detected, it does not use either of these conditions.

We also introduce a set A_T defined by

$$A_T = \left\{ \left| (e-b+1)^{-1/2} \sum_{i=b}^{e} \varepsilon_i \right| < \lambda_2 \ \forall 1 \le b \le e \le T \right\}. \tag{10}$$

Note that by Bonferroni's inequality, $P(A_T) \ge 1 - CT^{-1}$ for $\lambda_2 \ge (6 \log T)^{1/2}$, where C is a positive constant.

Before presenting the formal proof, we informally discuss some of its aspects to facilitate understanding.

Informal discussion of some aspects of the proof. The performance of the binary segmentation algorithm analysed in Theorem 3.1 can be seen as 'deterministic on a random set whose probability approaches one', in the sense that for a T large enough and in a certain subset of the probability space whose probability approaches one, the algorithm is guaranteed to detect all true change-points before being stopped at the right time by the application of threshold ζ_T. We further clarify this observation below.

Heuristically speaking, on the set $A_T \cap B_T$, where A_T is defined in (10) and B_T in Lemma A.1, the innovations ε_t are well behaved in the sense that the empirical CUSUM statistics $\tilde{X}^b_{s,e}$ are uniformly close to the corresponding unobserved true quantities $\tilde{f}^b_{s,e}$ in the particular sense described in Lemmas A.1 and A.3. It is this closeness that causes the following behaviour: if there are still previously undetected change-points within the current interval $[s, e]$ (by which we mean that there are change-points for which there is no estimated change-point within the distance of $C\epsilon_T$), and $[s, e]$ satisfies (8) and (9), then (i) by Lemma A.3, $b_0 = \arg\max_{t:s\le t<e} |\tilde{X}^t_{s,e}|$ falls within the distance of $C\epsilon_T$ of one of the previously undetected change-points in $[s, e]$ (denote that change-point here by η_{p_0+r}), and (ii) by Lemma A.4, we have $|\tilde{X}^{b_0}_{s,e}| > \zeta_T$.

The consequence of (i) and (ii) is that b_0 passes the thresholding test for the significance of a change-point and is from now on considered to be an estimate of η_{p_0+r}. Note that the assignment of b_0 to η_{p_0+r} is unambiguous: b_0 cannot be an estimate of any of the other change-points as they are too far; the nearest left- or right-neighbour of η_{p_0+r} is at a distance of no less than δ_T of it, which means not nearer than $\delta_T - C\epsilon_T$ from b_0, which is orders of magnitude larger than $C\epsilon_T$ as specified in Theorem 3.1 and its assumptions.

As a consequence, the procedure then moves on to operate on the intervals $[s, b_0]$ and $[b_0, e]$. Without loss of generality, suppose there are previously undetected change-points on $[s, b_0]$. We now demonstrate that (8) and (9) hold for that interval. Since b_0 is close to η_{p_0+r} (which is 'previously detected'), it must be far from all other true change-points in the sense described in the previous paragraph. In particular, for any previously undetected change-point $\eta_{p_0+r'} \in [s, b_0]$, we must have $b_0 - \eta_{p_0+r'} \ge \delta_T - C\epsilon_T$ by the argument from the previous paragraph, which is larger than $C\delta_T$ for some $C > 0$, by the assumptions of Theorem 3.1. Hence $[s, b_0]$ satisfies (8).

Similarly, $[s, b_0]$ satisfies (9) as b_0 is within the distance of $C\epsilon_T$ of one of its neighbouring change-points, namely η_{p_0+r}.

Thus (8) and (9) are both valid as the algorithm progresses for any interval $[s, e]$ on which there are still previously undetected change-points. Therefore, for a large enough T and on $A_T \cap B_T$, all change-points will be detected one by one. At that point, by Lemma A.5, the statistics $|\tilde{X}^b_{s,e}|$ will become uniformly smaller than the threshold ζ_T and the algorithm will stop.

We are now in a position to turn to the formal proof, which is split into a number of lemmas.

LEMMA A.1. *Let X_t follow model (2), and let the assumptions of Theorem 3.1 hold. Let $\tilde{X}^b_{s,e}$ and $\tilde{f}^b_{s,e}$ be defined by (5) and (6), respectively. We then have $P(B_T) \geq 1 - CT^{-1}$, where*

$$B_T = \Big\{ \max_{s,b,e:\, 1 \leq s \leq b < e \leq T} |\tilde{X}^b_{s,e} - \tilde{f}^b_{s,e}| \leq \lambda_1 \Big\},$$

$\lambda_1 \geq \sqrt{8 \log T}$, and C is a positive constant.

PROOF. The proof proceeds via a simple Bonferroni inequality,

$$1 - P(B_T) \leq \sum_{s,b,e} P\big(|Z| > \lambda_1\big) \leq T^3 \frac{\phi_Z(\lambda_1)}{\lambda_1} \leq \frac{C}{T},$$

where Z is a standard normal and $\phi_Z(\cdot)$ is its p.d.f. □

We conjecture that more accurate bounds for λ_1 of Lemma A.1 and λ_2 in formula (10) can be obtained, for example, using techniques as in Taylor, Worsley and Gosselin (2007), Antoch and Jarušková (2013), or especially Lemma 1 of Yao (1988). However, we note that even with the use of the suboptimal Bonferroni inequality, λ_1 and λ_2 are already rate-optimal, which is what matters for the rates of convergence in Theorems 3.1–3.3. Improving the multiplicative constants in λ_1 and λ_2 would bring no further practical benefits in terms of choosing the stopping criterion for BS or WBS, the main reason for this being that the result of Lemma A.5 (and its equivalent in the proof of Theorem 3.2) is dependent on a different constant C anyway, which is not straightforward to evaluate in theory.

LEMMA A.2. *Let X_t follow model (2) and let the assumptions of Theorem 3.1 hold. Let $\tilde{X}^b_{s,e}$ and $\tilde{f}^b_{s,e}$ be defined by (5) and (6), respectively. Assume (7), (8), and (9). On set B_T of Lemma A.1, the following holds. For $b = \arg\max_{t:s \leq t < e} |\tilde{X}^t_{s,e}|$, there exists $1 \leq r \leq q$ such that for large T, $|b - \eta_{p_0+r}| \leq C_1 \gamma_T$ with $\gamma_T = T^{1/2}\lambda_1 / f'_{p_0+r}$ (λ_1 as in Lemma A.1). In addition, $|\tilde{f}^t_{s,e}|$ must then have a local maximum at $t = \eta_{p_0+r}$, and we must have*

$$\frac{|\tilde{f}^{\eta_{p_0+r}}_{s,e}|}{\max_{t:\, s \leq t < e} |\tilde{f}^t_{s,e}|} \geq C_2,$$

where C_1, C_2 are positive constants.

PROOF. We first note that $\gamma_T = o(\delta_T)$ since $1/2 + \varpi < 1/2 + 2\Theta - 3/2 \leq \Theta$. Note also that $\delta_T T^{-1/2} \underline{f}_T \geq CT^{\varphi}$ for C, φ positive. Let $b_1 = \arg\max_{t:s \leq t < e} |\tilde{f}^t_{s,e}|$. From Lemma A.1, we have

$$|\tilde{f}^{b_1}_{s,e}| \leq |\tilde{X}^{b_1}_{s,e}| + \lambda_1 \leq |\tilde{X}^b_{s,e}| + \lambda_1 \leq |\tilde{f}^b_{s,e}| + 2\lambda_1. \tag{11}$$

Assume $b \in (\eta_{p_0+r} + C\gamma_T, \eta_{p_0+r+1} - C\gamma_T)$ for a large enough constant C for some r and w.l.o.g. $\tilde{f}^b_{s,e} > 0$. From Lemma 2.2 in Venkatraman (1992), $\tilde{f}^t_{s,e}$ is either monotonic or decreasing and then increasing on $[\eta_{p_0+r}, \eta_{p_0+r+1}]$ and $\max(\tilde{f}^{\eta_{p_0+r}}_{s,e}, \tilde{f}^{\eta_{p_0+r+1}}_{s,e}) > \tilde{f}^b_{s,e}$. If $\tilde{f}^b_{s,e}$ locally decreases at b, then $\tilde{f}^{\eta_{p_0+r}}_{s,e} > \tilde{f}^b_{s,e}$, and arguing exactly as in Lemma 2 of Cho and Fryzlewicz (2012), there exists $b' \in (\eta_{p_0+r}, \eta_{p_0+r} + C\gamma_T]$ such that $\tilde{f}^{\eta_{p_0+r}}_{s,e} \geq \tilde{f}^{b'}_{s,e} + 2\lambda_1$. This would in turn lead to $|\tilde{f}^{b_1}_{s,e}| > |\tilde{f}^b_{s,e}| + 2\lambda_1$, a contradiction of (11). Similar arguments apply if $\tilde{f}^b_{s,e}$ locally increases at b.

Let r be as in the statement of this lemma. Then $|\tilde{f}^{\eta_{p_0+r}}_{s,e}|$ must be a local maximum, as if it were not, we would have $\max(|\tilde{f}^{\eta_{p_0+r-1}}_{s,e}|, |\tilde{f}^{\eta_{p_0+r+1}}_{s,e}|) > |\tilde{f}^{\eta_{p_0+r}}_{s,e}|$, and arguing exactly as above, this maximum would have to be sufficiently larger than $|\tilde{f}^{\eta_{p_0+r}}_{s,e}|$ for b to fall near the change-point achieving this maximum, rather than near η_{p_0+r}, which is a contradiction.

Finally, using the same argumentation again, $|\tilde{f}^{\eta_{p_0+r}}_{s,e}| / \max_{t:s\leq t<e} |\tilde{f}^t_{s,e}|$ must be bounded from below, as if were not, then recalling that $\max_{t:s\leq t<e} |\tilde{f}^t_{s,e}| \geq C\delta_T T^{-1/2} \underline{f}_T$ by Lemma 1 of Cho and Fryzlewicz (2012), b would have to fall near the change-point achieving this maximum, rather than near η_{p_0+r}, which is again a contradiction. This completes the proof of the lemma. □

LEMMA A.3. *Let the conditions of Lemma A.2 hold, and let the notation be as in that lemma. On set $B_T \cap A_T$, where B_T is defined in Lemma A.1 and A_T in (10), we have for large T, $|b - \eta_{p_0+r}| \leq C\epsilon_T$, where $\epsilon_T = \lambda_2^2 T^2 \delta_T^{-2} (f'_{p_0+r})^{-2}$ and C is a positive constant.*

PROOF. Let $\langle \cdot, \cdot \rangle$ denote the inner product between two vectors. Let $\psi^d_{s,e}$ be a vector whose elements $\psi^d_{s,e,t}$ are constant and positive for $t = s, \ldots, d$, constant and negative for $t = d+1, \ldots, e$, sum to zero and such that their squares sum to one. Then it is easy to see that $\tilde{X}^d_{s,e} = \sum_{t=s}^e \psi^d_{s,e,t} X_t = \langle \psi^d_{s,e}, X \rangle$ and similarly $\tilde{f}^d_{s,e} = \langle \psi^d_{s,e}, f \rangle$. For any vector v supported on $[s, e]$, we have $\arg\max_{d:s\leq d<e} |\langle \psi^d_{s,e}, v \rangle| = \arg\min_{d:s\leq d<e} \sum_{t=s}^e (v_t - \bar{v}^d_{s,e,t})^2$, where $\bar{v}^d_{s,e}$ is an orthogonal projection of v on the space of step functions constant on $s, \ldots, d$ and constant on $d+1, \ldots, e$; this is immediate by noting that $\bar{v}^d_{s,e} = \bar{v} + \langle v, \psi^d_{s,e} \rangle \psi^d_{s,e}$, where $\bar{v}$ is the mean of v. From Lemma A.2,

$$\sum_{t=s}^e (X_t - \bar{X}^b_{s,e,t})^2 \leq \sum_{t=s}^e (X_t - \bar{f}^{\eta_{p_0+r}}_{s,e,t})^2.$$

Therefore, if it can be shown that for a certain $\epsilon_T < C_1\gamma_T$, we have

$$\sum_{t=s}^e (X_t - \bar{X}^d_{s,e,t})^2 > \sum_{t=s}^e (X_t - \bar{f}^{\eta_{p_0+r}}_{s,e,t})^2 \tag{12}$$

as long as

$$\epsilon_T < |d - \eta_{p_0+r}| \le C_1 \gamma_T, \tag{13}$$

then this would prove that necessarily, $|b - \eta_{p_0+r}| \le \epsilon_T$. Recalling that $X_t = f_t + \varepsilon_t$, (12) is equivalent to

$$2\sum_{t=s}^{e} \varepsilon_t (\bar{X}^d_{s,e,t} - \bar{f}^{\eta_{p_0+r}}_{s,e,t}) < \sum_{t=s}^{e} (f_t - \bar{X}^d_{s,e,t})^2 - \sum_{t=s}^{e} (f_t - \bar{f}^{\eta_{p_0+r}}_{s,e,t})^2,$$

and implied by

$$2\sum_{t=s}^{e} \varepsilon_t (\bar{X}^d_{s,e,t} - \bar{f}^{\eta_{p_0+r}}_{s,e,t}) < \sum_{t=s}^{e} (f_t - \bar{f}^d_{s,e,t})^2 - \sum_{t=s}^{e} (f_t - \bar{f}^{\eta_{p_0+r}}_{s,e,t})^2, \tag{14}$$

since obviously $\sum_{t=s}^{e} (f_t - \bar{f}^d_{s,e,t})^2 \le \sum_{t=s}^{e} (f_t - \bar{X}^d_{s,e,t})^2$. For any d, we have an ANOVA-type decomposition

$$\sum_{t=s}^{e} (f_t - \bar{f}^d_{s,e,t})^2 = \sum_{t=s}^{e} (f_t - \bar{f} - \langle f, \psi^d_{s,e}\rangle \psi^d_{s,e,t})^2 = \sum_{t=s}^{e} (f_t - \bar{f})^2 - \langle f, \psi^d_{s,e}\rangle^2.$$

Therefore, the right-hand side of (14) reduces to

$$\begin{aligned} \langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle^2 - \langle f, \psi^d_{s,e}\rangle^2 &= (|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle| - |\langle f, \psi^d_{s,e}\rangle|)(|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle| + |\langle f, \psi^d_{s,e}\rangle|) \\ &\ge (|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle| - |\langle f, \psi^d_{s,e}\rangle|)|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle|. \end{aligned}$$

Since by Lemma A.2, $|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle|$ is a local maximum, we can invoke Lemma 2 of Cho and Fryzlewicz (2012), by which we obtain

$$|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle| - |\langle f, \psi^d_{s,e}\rangle| \ge C|d - \eta_{p_0+r}| T^{-1/2} f'_{p_0+r}.$$

Combining Lemma 1 of Cho and Fryzlewicz (2012) with the last assertion of Lemma A.2, we obtain $|\langle f, \psi^{\eta_{p_0+r}}_{s,e}\rangle| \ge C\delta_T T^{-1/2} f'_{p_0+r}$. This finally yields

$$\sum_{t=s}^{e} (f_t - \bar{f}^d_{s,e,t})^2 - \sum_{t=s}^{e} (f_t - \bar{f}^{\eta_{p_0+r}}_{s,e,t})^2 \ge C|d - \eta_{p_0+r}| \delta_T (f'_{p_0+r})^2 / T.$$

We decompose the left-hand side of (14) as

$$\begin{aligned} &2\sum_{t=s}^{e} \varepsilon_t (\bar{X}^d_{s,e,t} - \bar{f}^{\eta_{p_0+r}}_{s,e,t}) \\ &\quad = 2\sum_{t=s}^{e} \varepsilon_t (\bar{X}^d_{s,e,t} - \bar{f}^d_{s,e,t}) + 2\sum_{t=s}^{e} \varepsilon_t (\bar{f}^d_{s,e,t} - \bar{f}^{\eta_{p_0+r}}_{s,e,t}). \end{aligned} \tag{15}$$

Without loss of generality, assume $d \geq \eta_{p_0+r}$. The second term on the right-hand side of (15) decomposes as

$$\sum_{t=s}^{e} \varepsilon_t (\bar{f}^d_{s,e,t} - \bar{f}^{\eta_{p_0+r}}_{s,e,t}) = \left(\sum_{t=s}^{\eta_{p_0+r}} + \sum_{t=\eta_{p_0+r}+1}^{d} + \sum_{t=d+1}^{e} \right) \varepsilon_t (\bar{f}^d_{s,e,t} - \bar{f}^{\eta_{p_0+r}}_{s,e,t})$$
$$= I + II + III.$$

We bound

$$|I| \leq \sqrt{\eta_{p_0+r} - s + 1} \left| \frac{1}{\sqrt{\eta_{p_0+r} - s + 1}} \sum_{t=s}^{\eta_{p_0+r}} \varepsilon_t \right|$$
$$\times \left| \frac{1}{d - s + 1} \sum_{t=s}^{d} f_t - \frac{1}{\eta_{p_0+r} - s + 1} \sum_{t=s}^{\eta_{p_0+r}} f_t \right|$$
$$\leq \sqrt{\eta_{p_0+r} - s + 1} \lambda_2 \frac{C|d - \eta_{p_0+r}| f'_{p_0+r}}{\eta_{p_0+r} - s + 1} \leq C\lambda_2 |d - \eta_{p_0+r}| f'_{p_0+r} \delta_T^{-1/2},$$

and we note that the bound for III is of the same order. Similarly, the bound for II is $C\lambda_2 |d - \eta_{p_0+r}|^{1/2} f'_{p_0+r}$. The first term on the right-hand side of (15) decomposes as

$$\sum_{t=s}^{e} \varepsilon_t (\bar{X}^d_{s,e,t} - \bar{f}^d_{s,e,t}) = \left(\sum_{t=s}^{d} + \sum_{t=d+1}^{e} \right) \varepsilon_t (\bar{X}^d_{s,e,t} - \bar{f}^d_{s,e,t}) = IV + V.$$

Note that IV and V are of the same order. We have

$$IV = \frac{1}{d - s + 1} \left(\sum_{t=s}^{d} \varepsilon_t \right)^2 \leq \lambda_2^2.$$

Combining all of the above bounds, there exists a constant C such that (14) holds if

$$|d - \eta_{p_0+r}| \delta_T T^{-1} (f'_{p_0+r})^2 \geq C \max\left(\lambda_2 |d - \eta_{p_0+r}| \delta_T^{-1/2} f'_{p_0+r}, \lambda_2 |d - \eta_{p_0+r}|^{1/2} f'_{p_0+r}, \lambda_2^2\right). \tag{16}$$

These three inequalities yield, respectively, $\delta_T \geq (C\lambda_2 T / f'_{p_0+r})^{2/3}$, $|d - \eta_{p_0+r}| \geq C^2 \lambda_2^2 T^2 (\delta_T f'_{p_0+r})^{-2}$, $|d - \eta_{p_0+r}| \geq C\lambda_2^2 T \delta_T^{-1} (f'_{p_0+r})^{-2}$. The last inequality can be ignored as it is implied by the second if $C \geq 1$. The first inequality can also be ignored as the second inequality and (13) together imply

$$C^2 \lambda_2^2 T^2 \delta_T^{-2} (f'_{p_0+r})^{-2} \leq C_1 \lambda_1 T^{1/2} (f'_{p_0+r})^{-1},$$

which leads to $\delta_T \geq C C_1^{-1/2} \lambda_2 \lambda_1^{-1/2} T^{3/4} (f'_{p_0+r})^{-1/2}$, a stronger requirement that in the first inequality since $\varpi < 1/2$, but automatically satisfied since

$3/4 + \varpi/2 < \Theta$. Therefore by the second inequality, ϵ_T can be taken to be $\max(1, C^2)\lambda_2^2 T^2 \delta_T^{-2} (f'_{p_0+r})^{-2}$. It remains for us to note that (13) is automatically satisfied, as required. This completes the proof of the lemma. □

LEMMA A.4. *Let X_t follow model (2), and let the assumptions of Theorem 3.1 hold. Let $\tilde{X}^b_{s,e}$ be defined by (5). Assume (7), (8), and (9). On the event B_T from Lemma A.1, we have $|\tilde{X}^b_{s,e}| > CT^{\Theta-1/2-\varpi}$, where $b = \arg\max_{t:s\le t<e} |\tilde{X}^t_{s,e}|$.*

PROOF. Let r be as in Lemma A.2. We have

$$|\tilde{X}^b_{s,e}| \ge |\tilde{X}^{\eta_{p_0+r}}_{s,e}| \ge |\tilde{f}^{\eta_{p_0+r}}_{s,e}| - \lambda_1 \ge C\delta_T T^{-1/2} \underline{f}_T - \lambda_1 > C_1 T^{\Theta-1/2-\varpi},$$

which completes the proof. □

LEMMA A.5. *Let X_t follow model (2), and let the assumptions of Theorem 3.1 hold. Let $\tilde{X}^b_{s,e}$ be defined by (5). For some positive constants C, C', let s, e satisfy one of three conditions*:

(i) $\exists! 1 \le p \le N$ *such that* $s \le \eta_p \le e$ *and* $(\eta_p - s + 1) \wedge (e - \eta_p) \le C\epsilon_T$, *or*
(ii) $\exists 1 \le p \le N$ *such that* $s \le \eta_p \le \eta_{p+1} \le e$ *and* $(\eta_p - s + 1) \vee (e - \eta_{p+1}) \le C'\epsilon_T$, *or*
(iii) $\exists 1 \le p \le N$ *such that* $\eta_p < s < e \le \eta_{p+1}$.

On the event $B_T \cap A_T$ from Lemma A.3, we have $|\tilde{X}^b_{s,e}| < C\lambda_2 T^{1-\Theta} + \lambda_1$, where $b = \arg\max_{t:s\le t<e} |\tilde{X}^t_{s,e}|$.

PROOF. We show case (ii); the remaining two cases are similar and simpler.

$$|\tilde{X}^b_{s,e}| \le |\tilde{f}^b_{s,e}| + \lambda_1 \le \max\bigl(|\tilde{f}^{\eta_p}_{s,e}|, |\tilde{f}^{\eta_{p+1}}_{s,e}|\bigr) + \lambda_1 = |\tilde{f}^{\eta_{p_0}}_{s,e}| + \lambda_1 \le C\epsilon_T^{1/2} f'_{p_0} + \lambda_1,$$

where the last inequality uses the definition of $\tilde{f}^t_{s,e}$. Continuing, for large T,

$$C\epsilon_T^{1/2} f'_{p_0} + \lambda_1 \le C\lambda_2 T \delta_T^{-1} + \lambda_1 \le C\lambda_2 T^{1-\Theta} + \lambda_1,$$

which completes the proof. □

With the use of Lemmas A.1 to A.5, the proof of the theorem is simple; the following occurs on the event $B_T \cap A_T$, which has probability $\ge 1 - C_1 T^{-1}$. At the start of the algorithm, as $s = 0$ and $e = T - 1$, all conditions for Lemma A.3 are met and it finds a change-point within the distance of $C\epsilon_T$ from the true change-point, by Lemma A.4. Under the assumption of the theorem, both (8) and (9) are satisfied within each segment until every change-point in f_t has been identified. Then one of the three conditions, (i), (ii), or (iii) of Lemma A.5, are met, and no further change-points are detected. □

PROOF OF THEOREM 3.2. We start by defining intervals $\mathcal{I}_i$ between change-points in such a way that their lengths are at least of order δ_T, and they are separated from the change-points also by distances at least of order δ_T. To fix ideas, define $\mathcal{I}_i = [\eta_{i-1} + \frac{1}{3}(\eta_i - \eta_{i-1}), \eta_{i-1} + \frac{2}{3}(\eta_i - \eta_{i-1})]$, $i = 1, \ldots, N+1$.

Each stage of the algorithm uses CUSUM statistics computed over M intervals (s_m, e_m), $m = 1, \ldots, M$, drawn uniformly (independently with replacement) from the set $\{(s, e) : s < e, 1 \le s \le T-1, 2 \le e \le T\}$. Define the event D_T^M as follows:

$$D_T^M = \big\{\forall i = 1, \ldots, N \ \exists m = 1, \ldots, M \ (s_m, e_m) \in \mathcal{I}_i \times \mathcal{I}_{i+1}\big\}.$$

Note that

$$P\big((D_T^M)^c\big) \le \sum_{i=1}^{N} \prod_{m=1}^{M} \big(1 - P\big((s_m, e_m) \in \mathcal{I}_i \times \mathcal{I}_{i+1}\big)\big) \le T\delta_T^{-1}\big(1 - \delta_T^2 T^{-2}/9\big)^M.$$

The remaining arguments will be valid on the set D_T^M. If an interval (s_m, e_m) is such that $(s_m, e_m) \in \mathcal{I}_i \times \mathcal{I}_{i+1}$, and thus (s_m, e_m) contains one change-point only, η_i, then arguing as in Cho and Fryzlewicz (2012), Lemma 1, we have

$$|\tilde{f}^{\eta_i}_{s_m, e_m}| = \max_{t : s_m \le t < e_m} |\tilde{f}^{t}_{s_m, e_m}| \ge C\delta_T^{1/2} f_i'. \tag{17}$$

Let (s, e) now be a generic interval satisfying (8) and (9), with ϵ_T and δ_T as in the statement of this theorem. The remaining arguments are valid on the set B_T. Consider

$$(m_0, b) = \arg \max_{(m,t) \,:\, m \in \mathcal{M}_{s,e}, s_m \le t < e_m} |\tilde{X}^{t}_{s_m, e_m}|, \tag{18}$$

where $\mathcal{M}_{s,e} = \{m : (s_m, e_m) \subseteq (s, e), 1 \le m \le M\}$. Imposing the condition

$$\delta_T \ge 3\epsilon_T, \tag{19}$$

we guarantee that both s and e are sufficiently bounded away from all the previously undetected change-points $\eta_i \in (s, e)$ in the sense that $\mathcal{I}_i \cup \mathcal{I}_{i+1} \subset (s, e)$ for all such i. Denote the set of these i's by $\mathcal{J}_{s,e}$. For each $i \in \mathcal{J}_{s,e}$, there exists an $m_i \in \mathcal{M}_{s,e}$ such that $(s_{m_i}, e_{m_i}) \in \mathcal{I}_i \times \mathcal{I}_{i+1}$, and thus

$$\begin{aligned} |\tilde{X}^{b}_{s_{m_0}, e_{m_0}}| &\ge \max_{t : s_{m_i} \le t < e_{m_i}} |\tilde{X}^{t}_{s_{m_i}, e_{m_i}}| \\ &\ge |\tilde{X}^{\eta_i}_{s_{m_i}, e_{m_i}}| \ge |\tilde{f}^{\eta_i}_{s_{m_i}, e_{m_i}}| - \lambda_1 \ge C_1 \delta_T^{1/2} f_i', \end{aligned} \tag{20}$$

provided that

$$\delta_T \ge C_8 \lambda_1^2 (\underline{f}_T)^{-2}. \tag{21}$$

Therefore,

$$|\tilde{f}^{b}_{s_{m_0}, e_{m_0}}| \ge |\tilde{X}^{b}_{s_{m_0}, e_{m_0}}| - \lambda_1 \ge C_2 \delta_T^{1/2} \max_{i \in \mathcal{J}_{s,e}} f_i'. \tag{22}$$

By Lemma 2.2 in Venkatraman (1992), there exists a change-point η_{p_0+r}, immediately to the left or to the right of b such that

$$|\tilde{f}^{\eta_{p_0+r}}_{s_{m_0},e_{m_0}}| > |\tilde{f}^{b}_{s_{m_0},e_{m_0}}| \geq C_2\delta_T^{1/2} \max_{i\in\mathcal{J}_{s,e}} f'_i. \tag{23}$$

Now, the following two situations are impossible:

(1) (s_{m_0}, e_{m_0}) contains one change-point only, η_{p_0+r}, and either $\eta_{p_0+r} - s_{m_0}$ or $e_{m_0} - \eta_{p_0+r}$ is not bounded from below by $C_3\delta_T$;
(2) (s_{m_0}, e_{m_0}) contains two change-points only, say η_{p_0+r} and η_{p_0+r+1}, and both $\eta_{p_0+r} - s_{m_0}$ and $e_{m_0} - \eta_{p_0+r+1}$ are not bounded from below by $C_3\delta_T$.

Indeed, if either situation were true, then using arguments as in Lemma A.5, we would obtain that $\max_{t:s_{m_0}\leq t<e_{m_0}} |\tilde{f}^{t}_{s_{m_0},e_{m_0}}|$ were not bounded from below by $C_2\delta_T^{1/2}\max_{i\in\mathcal{J}_{s,e}} f'_i$, a contradiction to (22). This proves that the interval (s_{m_0}, e_{m_0}) satisfies condition (8) (with δ_T as in the statement of this theorem), and thus we can follow the argument from the proof of Lemma 2 in Cho and Fryzlewicz (2012) to establish that if $|b' - \eta_{p_0+r}| = C\gamma_T$ for a certain C, with $\gamma_T = \delta_T^{1/2}\lambda_1/f'_{p_0+r}$, and if $\tilde{f}^{\eta_{p_0+r}}_{s_{m_0},e_{m_0}} > \tilde{f}^{b'}_{s_{m_0},e_{m_0}}$ (assuming w.l.o.g. $\tilde{f}^{\eta_{p_0+r}}_{s_{m_0},e_{m_0}} > 0$), then $\tilde{f}^{\eta_{p_0+r}}_{s_{m_0},e_{m_0}} \geq \tilde{f}^{b'}_{s_{m_0},e_{m_0}} + 2\lambda_1$.

With this result, it is then straightforward to proceed like in the proof of Lemma A.2 to show that $|b - \eta_{p_0+r}| \leq C_4\gamma_T$, and that $|\tilde{f}^{t}_{s_{m_0},e_{m_0}}|$ must have a local maximum at $t = \eta_{p_0+r}$.

To establish that $|b - \eta_{p_0+r}| \leq C_7\epsilon_T$, we need to use the above results to obtain an improved version of Lemma A.3. The arguments in the remainder of the proof are valid on the set A_T. Following the proof of Lemma A.3 for the interval (s_{m_0}, e_{m_0}) with $\gamma_T = \delta_T^{1/2}\lambda_1/f'_{p_0+r}$, in the notation of that lemma and using an argument like in Lemma 2 of Cho and Fryzlewicz (2012), we obtain

$$|\langle f, \psi^{\eta_{p_0+r}}_{s_{m_0},e_{m_0}}\rangle| - |\langle f, \psi^{d}_{s_{m_0},e_{m_0}}\rangle| \geq C|d - \eta_{p_0+r}|\delta_T^{-1/2} f'_{p_0+r}.$$

Additionally, by (23), $|\langle f, \psi^{\eta_{p_0+r}}_{s_{m_0},e_{m_0}}\rangle| \geq C_5\delta_T^{1/2} f'_{p_0+r}$, which combined yields

$$\sum_{t=s_{m_0}}^{e_{m_0}} (f_t - \bar{f}^{d}_{s_{m_0},e_{m_0},t})^2 - \sum_{t=s_{m_0}}^{e_{m_0}} (f_t - \bar{f}^{\eta_{p_0+r}}_{s_{m_0},e_{m_0},t})^2 \geq C|d - \eta_{p_0+r}|(f'_{p_0+r})^2.$$

This in turn leads to the following replacement for the triple inequality (16):

$$|d - \eta_{p_0+r}|(f'_{p_0+r})^2 \geq C\max(\lambda_2|d - \eta_{p_0+r}|\delta_T^{-1/2} f'_{p_0+r}, \lambda_2|d - \eta_{p_0+r}|^{1/2} f'_{p_0+r}, \lambda_2^2).$$

These three inequalities yield, respectively, $\delta_T \geq C^2\lambda_2^2/(f'_{p_0+r})^2$, $|d-\eta_{p_0+r}| \geq C^2\lambda_2^2/(f'_{p_0+r})^2$, $|d-\eta_{p_0+r}| \geq C\lambda_2^2/(f'_{p_0+r})^2$. The second inequality and the requirement that $|d-\eta_{p_0+r}| \leq C_6\gamma_T = C_6\lambda_1\delta_T^{1/2}/f'_{p_0+r}$ (see the proof of Lemma A.3) together imply $\delta_T \geq C^4C_6^{-2}\lambda_1^{-2}\lambda_2^4(f'_{p_0+r})^{-2}$. Combining this with the first inequality, we obtain

$$\delta_T \geq C^2\lambda_2^2\bigl(f'_{p_0+r}\bigr)^{-2}\max\bigl(C^2C_6^{-2}\lambda_1^{-2}\lambda_2^2,1\bigr). \tag{24}$$

By the second and third inequalities, ϵ_T can be taken to be $\max(1,C^2)\lambda_2^2/(f'_{p_0+r})^2$. At this point, we recall the constraints (19) and (21). Taking λ_1 and λ_2 to be of the lowest permissible order of magnitude, that is $\sim \log^{1/2}T$, these constraints together with (24) stipulate that we must have $\delta_T \geq C_9\log T/(\underline{f}_T)^2$ for a large enough C_9.

With the use of the above results, the proof of the theorem proceeds as follows; the following occurs on the event $B_T \cap A_T \cap D_T^M$, which has probability $\geq 1-C_1T^{-1}-T\delta_T^{-1}(1-\delta_T^2T^{-2}/9)^M$. At the start of the algorithm, as $s=0$ and $e=T-1$, (8) and (9) (with δ_T and ϵ_T as in the statement of this theorem) are satisfied, and therefore, by formula (20), the algorithm detects a change-point b on that interval, defined by formula (18). By the above discussion, b is within the distance of $C\epsilon_T$ from the change-point. Then (8) and (9) (with δ_T and ϵ_T as in the statement of this theorem) are satisfied within each segment until every change-point in f_t has been identified. Once this has happened, we note that every subsequent interval (s_m,e_m) satisfies the assumptions on (s,e) from Lemma A.5 and therefore $|\tilde{X}^b_{s_{m_0},e_{m_0}}| < C\lambda_2+\lambda_1 \leq \zeta_T$, which means that no further change-points are detected. □

PROOF OF THEOREM 3.3. The following considerations are valid on the set $A_T \cap B_T \cap D_T^M$ (from Theorem 3.2) which has probability $\geq 1-C_1T^{-1}-T\delta_T^{-1}(1-\delta_T^2T^{-2}/9)^M$. First consider the case $k>N$. Let $\bar{X}_{s,e}$ be the sample mean of X_t on the interval $[s,e]$ and recall the definition of $\psi^d_{s,e}$ from Lemma A.3. The difference $\hat{\sigma}^2_{k-1}-\hat{\sigma}^2_k$ must necessarily be of the form

$$\begin{aligned}\hat{\sigma}^2_{k-1}-\hat{\sigma}^2_k &= \frac{1}{T}\biggl\{\sum_{i=s}^{e}(X_i-\bar{X}_{s,e})^2-\sum_{i=s}^{e}\bigl(X_i-\bar{X}_{s,e}-\langle X,\psi^d_{s,e}\rangle\psi^d_{s,e,i}\bigr)^2\biggr\}\\ &= \frac{1}{T}\biggl\{2\sum_{i=s}^{e}(X_i-\bar{X}_{s,e})\langle X,\psi^d_{s,e}\rangle\psi^d_{s,e,i}-\sum_{i=s}^{e}\langle X,\psi^d_{s,e}\rangle^2\bigl(\psi^d_{s,e,i}\bigr)^2\biggr\} \qquad (25)\\ &= \frac{\langle X,\psi^d_{s,e}\rangle^2}{T}.\end{aligned}$$

From the proof of Theorem 3.2, in the case $k>N$, that is, once all the change-points have been detected, we have $\langle X,\psi^d_{s,e}\rangle^2 \leq C(\lambda_1^2+\lambda_2^2) \leq C\log T$. Therefore, for a constant $\upsilon>0$, and using the fact that on the set A_T, we have

$|\hat{\sigma}_N^2 - \mathrm{Var}(\varepsilon_t)| \leq CT^{-1}\log T$, we obtain

$$\begin{aligned}
\mathrm{sSIC}(k) - \mathrm{sSIC}(N) &= \frac{T}{2}\log\frac{\hat{\sigma}_k^2}{\hat{\sigma}_N^2} + (k-N)\log^\alpha T \\
&= \frac{T}{2}\log\biggl(1 - \frac{\hat{\sigma}_N^2 - \hat{\sigma}_k^2}{\hat{\sigma}_N^2}\biggr) + (k-N)\log^\alpha T \\
&\geq -\frac{T}{2}(1+\upsilon)\frac{\hat{\sigma}_N^2 - \hat{\sigma}_k^2}{\hat{\sigma}_N^2} + (k-N)\log^\alpha T \\
&\geq -C_1 \log T + (k-N)\log^\alpha T,
\end{aligned}$$

which is guaranteed to be positive for T large enough. Conversely, if $k < N$, then by formulae (25) and (20), we have $\hat{\sigma}_k^2 - \hat{\sigma}_{k+1}^2 \geq C\delta_T \underline{f}_T^2/T$ and hence

$$\begin{aligned}
\mathrm{sSIC}(k) - \mathrm{sSIC}(N) &= \frac{T}{2}\log\frac{\hat{\sigma}_k^2}{\hat{\sigma}_N^2} + (k-N)\log^\alpha T \\
&= \frac{T}{2}\log\biggl(1 + \frac{\hat{\sigma}_k^2 - \hat{\sigma}_N^2}{\hat{\sigma}_N^2}\biggr) + (k-N)\log^\alpha T \\
&\geq \frac{T}{2}(1-\upsilon)\frac{\hat{\sigma}_k^2 - \hat{\sigma}_N^2}{\hat{\sigma}_N^2} - N\log^\alpha T \\
&\geq C\delta_T \underline{f}_T^2 - N\log^\alpha T,
\end{aligned}$$

which is again guaranteed to be positive for T large enough. Hence for T large enough and on the set $A_T \cap B_T \cap D_T^M$, $\mathrm{sSIC}(k)$ is necessarily minimised at N and therefore $\hat{N} = N$, as required. □

APPENDIX B: TEST MODELS AND METHODS USED IN THE SIMULATION STUDY

In the list below, we provide specifications of the test signals f_t and standard deviations σ of the noise ε_t used in the simulation study of Section 4.2, as well as reasons why these particular signals were used.

(1) `blocks`: length 2048, change-points at 205, 267, 308, 472, 512, 820, 902, 1332, 1557, 1598, 1659, values between change-points $0, 14.64, -3.66, 7.32, -7.32, 10.98, -4.39, 3.29, 19.03, 7.68, 15.37, 0$. Standard deviation of the noise $\sigma = 10$. Reason for choice: a standard piecewise-constant test signal widely analysed in the literature.

(2) `fms`: length 497, change-points at 139, 226, 243, 300, 309, 333, values between change-points $-0.18, 0.08, 1.07, -0.53, 0.16, -0.69, -0.16$. Standard deviation of the noise $\sigma = 0.3$. Reason for choice: a test signal proposed in Frick, Munk and Sieling (2014).

(3) `mix`: length 560, change-points at $11, 21, 41, 61, 91, 121, 161, 201, 251, 301, 361, 421, 491$, values between change-points $7, -7, 6, -6, 5, -5, 4, -4, 3, -3, 2, -2, 1, -1$. Standard deviation of the noise $\sigma = 4$. Reason for choice: a mix of prominent change-points between short intervals of constancy and less prominent change-points between longer intervals.

(4) `teeth10`: length 140, change-points at $11, 21, 31, 41, 51, 61, 71, 81, 91, 101, 111, 121, 131$, values between change-points $0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1$. Standard deviation of the noise $\sigma = 0.4$. Reason for choice: frequent change-points, occurring every 10th observation, in the shape of 'teeth'.

(5) `stairs10`: length 150, change-points at $11, 21, 31, 41, 51, 61, 71, 81, 91, 101, 111, 121, 131, 141$, values between change-points $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$. Standard deviation of the noise $\sigma = 0.3$. Reason for choice: frequent change-points, occurring every 10th observation, in the shape of 'stairs'.

The list below provides extra details of the competing methods used in the simulation study.

`strucchange`: the main routine for estimating the number and locations of change-points is `breakpoints`. It implements the procedure by Bai and Perron (2003). It is suitable for use in general regression problems, but also in the signal plus noise set-up. Given an input vector `x`, the command we use is `breakpoints(x ~ 1)`. The `breakpoints` routine requires a minimum segment size, which makes it not fully automatic. The results reported in the paper are with the default minimum segment size, which may not be the optimal choice for our test signals. We tried changing the minimum segment size to 1, but this resulted in execution times that were too long to permit inclusion of the method in our simulation study. We refer to the method as 'B&P' throughout the paper.

`Segmentor3IsBack`: the main routine is `Segmentor`. It implements a fast algorithm for minimising the least-squares cost function for change-point detection, as described in Rigaill (2010). The function `SelectModel` then selects the best model according to (by default) the 'oracle' penalisation as described in Lebarbier (2005). Our execution is

```
z <- Segmentor(x, model=2)
SelectModel(z)
```

The routine `Segmentor` requires specification of the maximum number of segments, which is set to 15 by default. We do not change this default setting. None of our test signals exceed this maximum number of segments. We refer to this method as 'S3IB'.

`changepoint`: the main routine is `cpt.mean`. It implements a (different) fast algorithm for minimising the least-squares cost function for change-point detection, as described in Killick, Fearnhead and Eckley (2012). The best model is then selected, by default, via the SIC penalty. Our execution is

```
cpt.mean(x/mad(diff(x)/sqrt(2)), method="PELT")@cpts,
```

where the `mad` function implements the median absolute deviation estimator. We refer to this method as 'PELT'.

`cumSeg`: the main routine is `jumpoints`, implementing an algorithm described in Muggeo and Adelfio (2011). We do not change the default setting which requires 'the starting number of changepoints', which 'should be quite larger than the supposed number of (true) changepoints' (quotes from the package manual) and is set to `min(30, round(length(x)/10))` by default. None of our test signals violates this. Our execution is `jumpoints(x)`. We refer to this method as 'cumSeg'.

`stepR`: the main routine is `smuceR`, implementing a multiscale algorithm described in Frick, Munk and Sieling (2014). We leave the default settings unchanged. Our execution is

```
smuceR(x, 1:length(x), family="gauss")
```

We refer to this method as 'SMUCE'.

APPENDIX C: PERFORMANCE OF WBS IN THE PRESENCE OF LINEAR TRENDS

Figure 5 shows the results of a small-scale simulation study aimed at obtaining some insight into the performance of WBS under model misspecification, namely in cases where the true function f_t exhibits linear trends.

In the example from the top row of that figure, the linear trends are so flat that they are almost completely ignored by WBS. However, in the example from the second row, the linear trends are more pronounced, and spurious detection of change-points within the trend sections tends to occur towards their middle parts. This can be interpreted in at least two ways: (i) WBS considers the middle part of a section with a linear trend as the most likely location of a change-point in the piecewise-constant approximation of that linear trend, which is natural, and (ii) the change-points (spuriously) detected within the trend sections tend to be separated from the main (correctly detected) change-point in the middle of the time domain, which is beneficial for the interpretability of the main change-point.

In the bottom two examples, spurious detection of change-points within the trend sections tends to occur towards their middle parts and towards their edges. This can be interpreted as the algorithm producing piecewise-constant approximations to the linear trends in which the change-points are spaced out rather than being clustered together, which hopefully leads to those approximations being visually attractive.

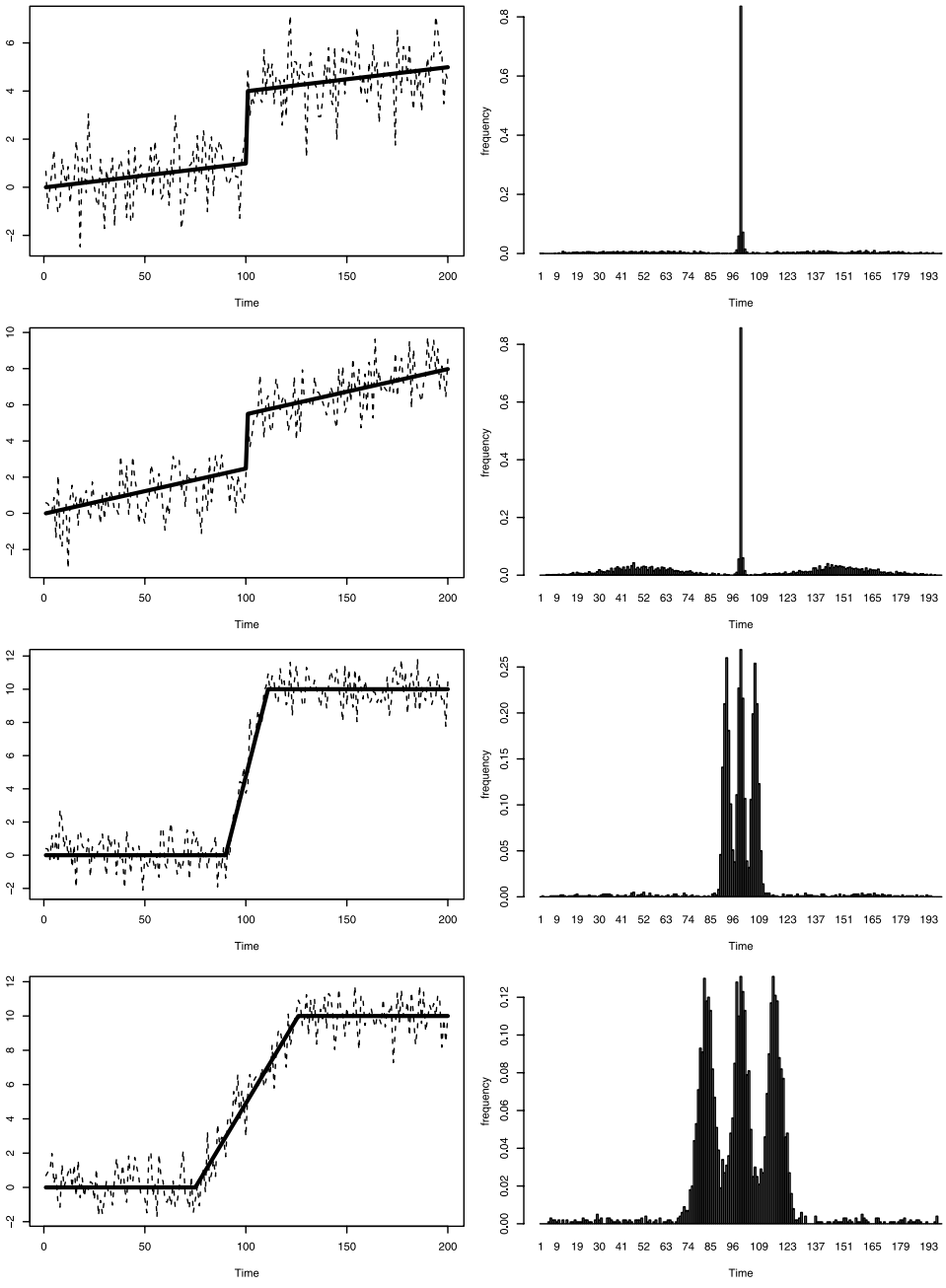

FIG. 5. *Left column*: *functions* f_t *containing linear trends* (*thick solid lines*) *and typical realisations of model* (1) *with* ε_t *i.i.d. standard normal* (*thin dashed lines*). *Right column*: *the corresponding bar plots of the frequencies with which change-points were detected at each time* t, *using the WBS method with threshold constant* $C = 1.3$, *over* 1000 *realisations of each model.*

REFERENCES

ANTOCH, J. and JARUŠKOVÁ, D. (2013). Testing for multiple change points. *Comput. Statist.* **28** 2161–2183. MR3107296

AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. MR0978902

BAI, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory* **13** 315–352. MR1455175

BAI, J. and PERRON, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18** 1–22.

BARANOWSKI, R. and FRYZLEWICZ, P. (2014). wbs: Wild Binary Segmentation for multiple change-point detection, 2014. R package version 1.1.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946

BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348

BRAUN, J. and MUELLER, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statist. Sci.* **13** 142–162.

BRODSKY, B. E. and DARKHOVSKY, B. S. (1993). *Nonparametric Methods in Change-Point Problems*. *Mathematics and Its Applications* **243**. Kluwer Academic, Dordrecht. MR1228205

CHAN, H. P. and WALTHER, G. (2013). Detection with the scan and the average likelihood ratio. *Statist. Sinica* **23** 409–428. MR3076173

CHEN, K.-M., COHEN, A. and SACKROWITZ, H. (2011). Consistent multiple testing for change points. *J. Multivariate Anal.* **102** 1339–1343. MR2819952

CHO, H. and FRYZLEWICZ, P. (2014). Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* To appear.

CHO, H. and FRYZLEWICZ, P. (2011). Multiscale interpretation of taut string estimation and its connection to unbalanced Haar wavelets. *Stat. Comput.* **21** 671–681. MR2826700

CHO, H. and FRYZLEWICZ, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statist. Sinica* **22** 207–229. MR2933173

CIUPERCA, G. (2011). A general criterion to determine the number of change-points. *Statist. Probab. Lett.* **81** 1267–1275. MR2803773

CIUPERCA, G. (2014). Model selection by LASSO methods in a change-point model. *Statist. Papers* **55** 349–374. MR3188408

DAVIES, P. L. and KOVAC, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29** 1–65. MR1833958

DAVIS, R. A., LEE, T. C. M. and RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101** 223–239. MR2268041

DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change-point inference (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76** 495–580.

FRYZLEWICZ, P. and SUBBA RAO, S. (2014). Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76** 903–924.

FRYZLEWICZ, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Statist. Assoc.* **102** 1318–1327. MR2412552

FRYZLEWICZ, P. (2012). Time-threshold maps: Using information from wavelet reconstructions with all threshold values simultaneously (with discussion). *J. Korean Statist. Soc.* **41** 145–159.

FRYZLEWICZ, P. (2014). Discussion contribution to 'Multiscale change-point inference' by Frick, Munk and Sieling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76** 547–548.

HALKO, N., MARTINSSON, P. G. and TROPP, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53** 217–288. MR2806637

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

HUŠKOVÁ, M. and SLABÝ, A. (2001). Permutation tests for multiple changes. *Kybernetika* (*Prague*) **37** 605–622. MR1877077

JACKSON, B., SARGLE, J., BARNES, D., ARABHI, S., ALT, A., GIOUMOUSIS, P., GWIN, E., SANGTRAKULCHAROEN, P., TAN, L. and TSAI, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters* **12** 105–108.

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

KILLICK, R., NAM, C., ASTON, J. and ECKLEY, I. (2012). changepoint.info: The changepoint repository. Available at http://changepoint.info.

KIRCH, C. and MUHSAL, B. (2014). A MOSUM procedure for the estimation of multiple random change points. Preprint.

KOROSTELËV, A. P. (1987). Minimax estimation of a discontinuous signal. *Theory Probab. Appl.* **32** 727–730.

LAVIELLE, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Process. Appl.* **83** 79–102. MR1705601

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

LAVIELLE, M. and MOULINES, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *J. Time Series Anal.* **21** 33–59. MR1766173

LEBARBIER, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85** 717–736.

LEE, C.-B. (1995). Estimating the number of change points in a sequence of independent normal random variables. *Statist. Probab. Lett.* **25** 241–248. MR1369518

MAHONEY, M. (2010). Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning* **3** 123–224.

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. MR3180567

MUGGEO, V. and ADELFIO, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* **27** 161–166.

OLSHEN, A., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

PAN, J. and CHEN, J. (2006). Application of modified information criterion to multiple change point problems. *J. Multivariate Anal.* **97** 2221–2241. MR2301636

RIGAILL, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. Preprint.

RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37** 2922–2952. MR2541451

ROJAS, C. and WAHLBERG, B. (2014). On change point detection using the fused lasso method. Preprint.

TAYLOR, J. E., WORSLEY, K. J. and GOSSELIN, F. (2007). Maxima of discretely sampled random fields, with an application to 'bubbles'. *Biometrika* **94** 1–18. MR2307898

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 91–108. MR2136641

VENKATRAMAN, E. S. (1992). Consistency results in multiple change-point problems. Ph.D. thesis, Stanford Univ., ProQuest LLC, Ann Arbor, MI. MR2687536

VOSTRIKOVA, L. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Math. Dokl.* **24** 55–59.

WANG, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385–397. MR1354236

WU, Y. (2008). Simultaneous change point analysis and variable selection in a regression problem. *J. Multivariate Anal.* **99** 2154–2171. MR2466557

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. MR1175613

DEPARTMENT OF STATISTICS
LONDON SCHOOL OF ECONOMICS
HOUGHTON STREET
LONDON WC2A 2AE
UNITED KINGDOM
E-MAIL: p.fryzlewicz@lse.ac.uk