MULTIPLE CHANGE-POINT DETECTION FOR NON-STATIONARY TIME SERIES USING WILD BINARY SEGMENTATION

Karolos K. Korkas and Piotr Fryzlewicz

London School of Economics

Abstract: We propose a new technique for consistent estimation of the number and locations of the change-points in the second-order structure of a time series. The core of the segmentation procedure is the Wild Binary Segmentation method (WBS), a technique which involves a certain randomised mechanism. The advantage of WBS over the standard Binary Segmentation lies in its localisation feature, thanks to which it works in cases where the spacings between change-points are short. In addition, we do not restrict the total number of change-points a time series can have. We also ameliorate the performance of our method by combining the CUSUM statistics obtained at different scales of the wavelet periodogram, our main change-point detection statistic, which allows a rigorous estimation of the local autocovariance of a piecewise-stationary process. We provide a simulation study to examine the performance of our method for different types of scenarios. A proof of consistency is also provided. Our methodology is implemented in the R package `wbsts`, available from CRAN.

Key words and phrases: non-stationarity, binary segmentation, change-points, locally stationary wavelet processes

1 Introduction

The assumption of stationarity has been the dominant framework for the analysis of many real data. However, in practice, time series entail changes in their dependence structure and therefore modelling non-stationary processes using stationary methods to capture their time-evolving dependence aspects will most likely result in a crude approximation. As pointed out by Mercurio and Spokoiny (2004) the risk of fitting a stationary model to non-stationary data can be high in terms of prediction and forecasting. Many examples of non-stationary data exist; for example, in biomedical signal processing of electroencephalograms (EEG) see Ombao et al. (2001); in audio signal processing see Davies and Bland (2010); in finance see Stărică and Granger (2005); in oceanography see Killick et al.

(2013). In this paper we deal with piecewise stationarity, arguably the simplest type of deviation from stationarity. This implies a time-varying process where its parameters evolve through time but remain constant for a specific period of time.

The problem of change-point estimation has attracted significant attention. A branch of the literature deals with the estimation of a single change-point (for a change in mean see e.g. Sen and Srivastava (1975); for time series see Davis et al. (1995), Gombay (2008), Gombay and Serban (2009) and references therein) while another extends it to multiple change-points with many changing parameters such as Ombao et al. (2001) who divide a time series into dyadic segments and choose the one with the minimum cost. The latter branch can be further categorised. On the one hand, the multiple change-point estimation can be formulated through an optimisation task i.e. minimising a multivariate cost function (or criterion). When the number of change-points N is unknown then a penalty is typically added e.g. the Schwarz criterion (see Yao (1988)). In addition, the user can adopt certain cost functions to deal with the estimation of specific models: the least-squares for change in the mean of a series (Yao and Au (1989) or Lavielle and Moulines (2000)), the Minimum Description Length criterion (MDL) for non-stationary time series (Davis et al. (2006)), the Gaussian log-likelihood function for changes in the volatility (Lavielle and Teyssiere (2007)) or the covariance structure of a multivariate time series (Lavielle and Teyssiere (2006)).

Several algorithms for minimising a cost function are based on dynamic programming (Bellman and Dreyfus (1966) and Kay (1998)) and they are often used in solving change-point problems, see e.g. Perron (2006) and references therein. Auger and Lawrence (1989) propose the Segment Neighbourhood method with complexity $\mathcal{O}(QT^2)$ where Q is the maximum number of change-points. An alternative method is the exact method of Optimal Partitioning by Jackson et al. (2005), but its complexity of $\mathcal{O}(T^2)$ still makes it suitable mostly for smaller samples.

Change-point estimators that adopt a multivariate cost function often come with a high computational cost. An attempt to reduce the computational burden is found in Killick et al. (2012) who extend the Optimal Partitioning method of

Jackson et al. (2005) (termed PELT) and show that the computational cost is $\mathcal{O}(T)$ when the number of change-points increases linearly with T. Another attempt is found in Davis et al. (2006) and Davis et al. (2008) who suggest a genetic algorithm to detect change-points in a piecewise-constant AR model or non-linear processes, respectively, where the MDL criterion is used.

On the other hand, the estimation of change-points can be formulated as a problem of minimising a series of univariate cost functions i.e. detecting a single change-point and then progressively moving to identify more. The Binary Segmentation method (BS) belongs to this category and uses a certain test statistic (such as the CUSUM) to reject the null hypothesis of no change-point. The BS has been widely used and the main reasons are its low computational complexity and the fact that it is conceptually easy to implement: after identifying a change-point the detection of further change-points continues to the left and to the right of the initial change-point until no further changes are found.

The BS method has been adopted to solve different types of problems. Inclan and Tiao (1994) detect breaks in the variance of a sequence of independent observations; Berkes et al. (2009) use a weighted CUSUM to reveal changes in the mean or the covariance structure of a linear process; Lee et al. (2003) apply the test in the residuals obtained from a least squares estimator; and Kim et al. (2000) and Lee and Park (2001) extend Inclan and Tiao (1994) method to a GARCH(1,1) model and linear processes, respectively. A common factor of most of these methods is the estimation of the long-term variance or autocovariance; a rather difficult task when the observations are dependent. Cho and Fryzlewicz (2012) apply the binary segmentation method on the wavelet periodograms with the purpose of detecting change-points in the second-order structure of a non-stationary process. Using the wavelet periodogram, Killick et al. (2013) propose a likelihood ratio test under the null and alternative hypotheses. The authors apply the binary segmentation algorithm but assume an upper bound for the number of change-points. Fryzlewicz and Subba Rao (2014) adopt the binary segmentation search to test for multiple change-points in a piecewise constant ARCH model. BS is also used for multivariate (possibly high-dimensional) time series segmentation in Cho and Fryzlewicz (2015) and in Schröder and Fryzlewicz (2013) in the context of trend detection for financial time series.

In this paper we develop a detection method to estimate the number and locations of change-points in the second-order structure of a piecewise stationary time series model using the non-parametric Locally Stationary Wavelet (LSW) process of Nason et al. (2000). The LSW model provides a complete description of the second-order structure of a stochastic process and, hence, it permits a fast estimation of the local autocovariance through the evolutionary wavelet spectrum. This choice, however, should not be seen as a restriction and potentially other models can form the basis for our algorithm.

In order to implement the change-point detection we adopt the Wild Binary Segmentation (WBS) method, proposed in the signal+iid Gaussian noise set-up by Fryzlewicz (2014), which attempts to overcome the limitations of the BS method. Our motivation for doing so is the good practical performance of the WBS method in this setting. Under specific models in which many change-points are present the BS search may be inefficient in detecting them. This stems from the fact that the BS starts its search assuming a single change-point. To correct this limitation, Fryzlewicz (2014) proposes the WBS algorithm that involves a "certain random localisation mechanism". His method can be summarised as follows. At the beginning of the algorithm the CUSUM statistic is not calculated over the entire set $\{0, ..., T-1\}$ where T is the sample size but over M local segments $[s, e]$. The starting s and ending e points are randomly drawn from a uniform distribution $U(0, T-1)$ and the hope is that for a large enough M, at least some of the intervals drawn will only contain single change-points, and therefore be particularly suitable for CUSUM-based detection. The location where the largest maximum CUSUM over all intervals drawn is achieved serves as the first change-point candidate. The method then proceeds similarly to BS: if the obtained CUSUM statistic exceeds a threshold then it is deemed to be a change-point and the same procedure continues to its left and right.

In order to adapt the WBS technique to our aim of detecting change-points in the second-order structure of a time series, we firstly adapt WBS for use in the multiplicative model setting, where the input sequence is a typically autocorrelated random scaled χ^2_1-distributed sequence with a piecewise constant variance. This is more challenging to achieve than in the standard BS setting (Cho and Fryzlewicz (2012)) due to the fact that many of the intervals consid-

ered are short, which typically causes spurious behaviour of the corresponding CUSUM statistics. We note here that this phenomenon does not arise in the signal+iid Gaussian noise setting (Fryzlewicz (2014)) and is entirely due to the distributional features of the above multiplicative setting. This challenge requires a number of new solutions proposed in this work, which include introducing the smallest possible interval length, and limiting the permitted "unbalancedness" of the CUSUM statistics, which is achieved without a detrimental effect on their operability thanks to the suitably large number of intervals of differing lengths considered at each stage by WBS.

Change-point detection for the second-order structure of a time series is achieved by combining information from local wavelet periodograms (each of which can be viewed as coming from the multiplicative model described above) across the resolution scales at which they were computed. This paper introduces a new way of combining this information across the scales, with the aim of further improving the practical performance of the proposed methodology.

One "high-level" message that this work attempts to convey is the introduction of a new *modus operandi* in time series analysis, whereby, in order to solve a specific problem, a large number of simple problems are solved on sub-samples of the data of differing lengths, and then the results combined to create an overall answer. In this work, this is done via the WBS technique with the aim of detecting change-points, but related techniques could be envisaged e.g. for trend and seasonality detection, stationarity testing or forecasting. We hope that our work will stimulate further work in this direction.

The paper is structured as follows: in Section 2 we present and review the WBS algorithm in the context of time series. The reasons for selecting the LSW model as the core of our detection algorithm are given in Section 3. The main algorithm is presented in Section 4 along with its theoretical consistency in estimating the number and locations of change-points. In addition, we conduct a simulation study to examine the performance of the algorithm; the results are given in Section 5. Finally, in Section 6 we apply our method to two real data sets. Proofs of our results are in the Appendix. Our methodology is implemented in the R package `wbsts`, available from CRAN.

2 The Wild Binary Segmentation Algorithm

The BS algorithm for a stochastic process was first introduced by Vostrikova (1981) who showed its consistency for the number and locations of change-points for a fixed N. A proof of its consistency is also given by Venkatraman (1992) for the Gaussian function+noise model, though the rates for the locations of the change-points are suboptimal. Improved rates of convergence of the locations of the change-points for the BS method are given by Fryzlewicz (2014).

As a preparatory exercise before considering segmentation in the full time series model (1.6) we first examine the following multiplicative model

$$Y_{t,T}^2 = \sigma_{t,T}^2 Z_{t,T}^2,\ t = 0, ..., T-1 \tag{1.1}$$

where $\sigma_{t,T}^2$ is a piecewise constant function and the series $Z_{t,T}$ are possibly autocorrelated standard normal variables. This generic set-up is of interest to us because the wavelet periodogram, used later in the segmentation of (1.6), follows model (1.1), up to a small amount of bias which we show can provably be neglected.

A potential change-point b_0 on a segment $[s, e]$ is given by

$$b_0 = \arg\max_b \left| \tilde{Y}_{s,e}^b / q_{s,e} \right|$$

where $\tilde{Y}_{s,e}^b$ is the CUSUM statistic

$$\tilde{Y}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} Y_t^2 - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} Y_t^2, \tag{1.2}$$

$q_{s,e} = \sum_{t=s}^{e} Y_t^2/n$ and $n = e - s + 1$. It can be shown that b_0 is the least-squares estimator of the change-point location in the case of $[s, e]$ containing exactly one change-point.

The value $|\tilde{Y}_{s,e}^{b_0}/q_{s,e}| = \max_b |\tilde{Y}_{s,e}^b/q_{s,e}|$ will be tested against a threshold ω_T in order to decide whether the null hypothesis of no change-point is rejected or not. The BS proceeds by recursively applying the above CUSUM on the two, newly-created segments defined by the already detected b_0, i.e $[s, b_0]$ and $[b_0+1, e]$. The algorithm stops in each current interval when no further change-points are detected, that is, the obtained CUSUM values fall below threshold ω_T.

The BS method has the disadvantage of possibly fitting the wrong model when multiple change-points are present as it searches the whole series. The

application of the CUSUM statistic (1.2) can result in spurious change-point detection when e.g. the true change-points occur close to each other. This is due to the fact that the BS method begins by assuming a single change-point exists in the series and, hence, the CUSUM statistic may be flatter (as a function of b) in the presence of multiple change-points. Especially, the BS method can fail to detect a small change in the middle of a large segment (Olshen et al. (2004)) which is illustrated in Fryzlewicz (2014).

Fryzlewicz (2014) proposes a randomised binary segmentation (termed Wild Binary Segmentation – WBS) where the search for change-points proceeds by calculating the CUSUM statistic in smaller segments whose length is random. By doing so, the user is guaranteed, with probability tending to one with the sample size, to draw favourable intervals containing at most a single change-point, which means the CUSUM statistic will be an appropriate one to use over those intervals from the point of view of model choice. The maximum of the CUSUM statistics in absolute value, taken over a large collection of random intervals (see Figure 1.1 for an illustration), is considered to be the first change-point candidate, and is tested for significance. The binary segmentation procedure is not altered, meaning that after identifying a change-point the problem is divided into two sub-problems where for each segment we again test for further change-points in exactly the same way. The computational complexity of the method can be reduced by noticing that the randomly drawn intervals and their corresponding CUSUM statistics can be calculated once at the start of the algorithm. Then, as the algorithm proceeds at a generic segment $[s, e]$, the obtained statistics can be reused making sure the random starting and end points fall within $[s, e]$.

The main steps of the WBS algorithm modified for the model (1.1) are outlined below.

- Calculate the CUSUM statistics over a collection of random intervals $[s_m, e_m]$. The starting and ending points are not fixed but are sampled from a uniform distribution with replacement making sure that

$$e_m \geq s_m + \Delta_T \tag{1.3}$$

  where $\Delta_T > 0$ defines the minimum size of the interval drawn.

  Denote by $\mathcal{M}_{s,e}$ the set of indices m of all random intervals $[s_m, e_m]$ where

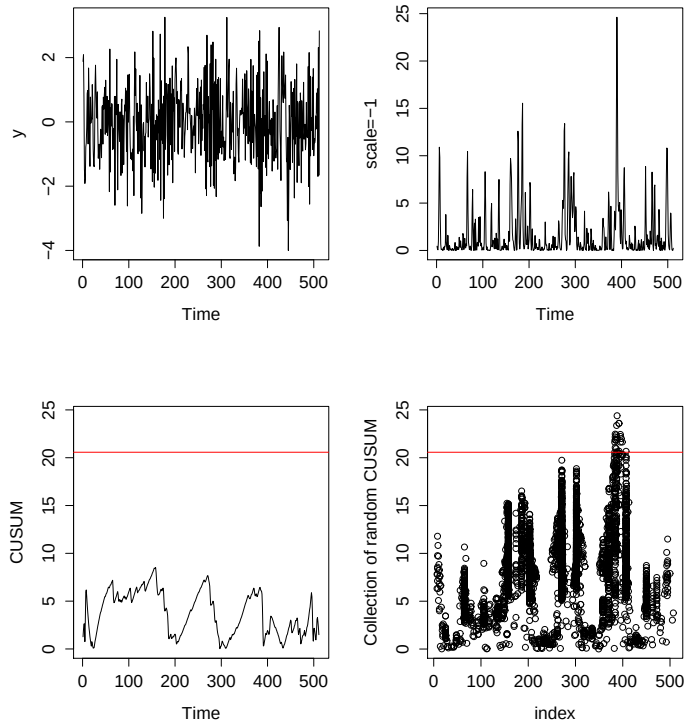

Figure 1.1: A simulated series (top-left) of an AR(1) model $y_t = \phi_t y_{t-1} + \varepsilon_t$ with $\phi_t = (0.5, 0.0)$ and change-points at $\{50, 100, ..., 450\}$. The Wavelet Periodogram at scale -1 (top-right). The CUSUM statistic of scale -1 (bottom-left) as in the BS method; the red line is threshold $C\log(T)$ with C chosen arbitrarily (for comparative illustration only). The rescaled $\mathbb{Y}^b_{s_m,e_m}$ for $m \in \mathcal{M}_{s,e}$ and $b \in s_m, ..., e_m - 1$ (bottom-right) as in the WBS method; the red line is the same threshold.

$m = 1, ..., M$ such that $[s_m, e_m] \subseteq [s, e]$; then the likely location of a change-point is

$$(m_0, b_0) = \underset{(m\in\mathcal{M}_{s,e}, b\in s_m,...,e_m-1)}{\arg\max} \left|\tilde{Y}^b_{s_m,e_m}/q_{s_m,e_m}\right| \tag{1.4}$$

such that

$$\max\left(\frac{e_{m_0} - b_0}{e_{m_0} - s_{m_0} + 1}, \frac{b_0 - s_{m_0} + 1}{e_{m_0} - s_{m_0} + 1}\right) \leq c_\star \tag{1.5}$$

where $c_\star$ is a constant satisfying $c_\star \in [2/3, 1)$. The conditions (1.3) and (1.5) do not appear in the original work by Fryzlewicz (2014), but they are necessary from the point of view of both theory and practical performance in the multiplicative model (1.1).

- The obtained CUSUM values are rescaled and tested against a threshold ω_T. This will ensure that with probability tending to one with the sample size, only the significant change-points will survive. The choice of the threshold ω_T is discussed in Section 4. If the obtained CUSUM statistic is significant then the search is continued to the left and to the right of b_0; otherwise the algorithm stops. This step differs from the original WBS method of Fryzlewicz (2014) in that the CUSUM statistics are rescaled using q_{s_m,e_m} so that ω_T does not depend on $\sigma^2_{t,T}$.

Although from the theoretical point of view, sampling distributions other than the uniform are also possible and lead to practically the same theoretical results (as long as their probability mass functions, supported on $[0, 1, \ldots, T-1]$ are uniformly of order T^{-1}, as is apparent by examining formula (6.3)), the uniform distribution plays a special role here as it provides a natural and fair subsampling of the set of *all* possible sub-intervals of $[s, e]$, which would be the optimal set to consider were it not for the prohibitive computational time of such an operation.

3 Locally Stationary Wavelets and the Multiplicative Model

In this section we introduce the reader to the LSW modelling paradigm of Nason et al. (2000). The LSW process enables a time-scale decomposition of a process and thus permits a rigorous estimation of the evolutionary wavelet spectrum and the local autocovariance and can be seen as an alternative to the Fourier based approach for modelling time series.

We now provide the definition of the LSW from Fryzlewicz and Nason (2006): a triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$ for $T = 1, 2, ...$, is in a class of Locally Stationary Wavelet (LSW) processes if there exists a mean-square representation

$$X_{t,T} = \sum_{i=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W_i(k/T)\psi_{i,t-k}\xi_{i,k} \tag{1.6}$$

with $i \in -1, -2, ...$ and $k \in \mathbb{Z}$ are, respectively, scale and location parameters, $(\psi_{i,0}, ..., \psi_{i,\mathcal{L}-1})$ are discrete, real-valued, compactly supported, non-decimated wavelet vectors with support length $\mathcal{L} = O(2^{-i})$, and the $\xi_{i,k}$ are zero-mean, orthonormal, identically distributed random variables. In this set-up we replace the Lipschitz-continuity constraint on $W_i(z)$ by the piecewise constant constraint,

which allows us to model a process whose second-order structure evolves in a piecewise constant manner over time with a finite but unknown number of change-points. Let L_i be the total magnitude of change-points in $W_i^2(z)$, then the functions $W_i(z)$ satisfy

- $\sum_{i=-\infty}^{-1} W_i^2 < \infty$ uniformly in z
- $\sum_{i=-I}^{-1} 2^{-i} L_i = \mathcal{O}(\log T)$ where $I = \log_2 T$.

The simplest type of a wavelet system that can be used in formula (1.6) are the Haar wavelets. Specifically,

$$\psi_{i,k} = 2^{i/2}\mathbb{I}_{0,\dots,2^{-j-1}-1}(k) - 2^{i/2}\mathbb{I}_{2^{-j-1},\dots,2^{-i}-1}(k)$$

for $i = -1, -2, ..., k$ $k \in \mathbb{Z}$ where $\mathbb{I}_A(k)$ is 1 if $k \in A$ and 0 otherwise. Further, small absolute values of the scale parameter i denote "fine" scales, while large absolute values denote "coarser" scales. In fine scales the wavelet vectors are most oscillatory and localised. By contrast, coarser scales have longer, less oscillatory wavelet vectors. Throughout the paper, we only use Haar wavelets, noting that the theoretical analysis using any other compactly supported wavelets would not be straightforward due to the unavailability of a closed formula for their coefficient values.

Throughout this paper, we assume that $\xi_{i,k}$ are distributed as $N(0,1)$ even though extensions to other cases are possible but technically challenging as they would entail consideration of quadratic forms of correlated non-Gaussian variables.

Of main interest in the LSW set-up is the Evolutionary Wavelet Spectrum (EWS) $S_i(z) = W_i^2(z)$, $i = -1, -2, ...$, defined on the rescaled-time interval $z \in [0, 1]$. The estimation of the EWS is done through the wavelet periodogram (Nason et al. (2000)) and its definition is given below:

Definition: Let $X_{t,T}$ be an LSW process constructed using the wavelet system ψ. The triangular stochastic array

$$I_{t,T}^{(i)} = \left| \sum_s X_{s,T} \psi_{i,s-t} \right|^2 \tag{1.7}$$

is called the wavelet periodogram of $X_{t,T}$ at scale i.

The wavelet periodogram is a convenient statistic for us to use, for the following reasons: (a) wavelet periodograms are fast to compute, (b) for Gaussian processes X_t, they arise as χ^2_1-type sequences which are easier to segment than, for example, empirical autocovariance sequences of the type $\{X_t X_{t+\tau}\}_t$, (c) because wavelets "decorrelate" a wide range of time series dependence structures and (d) because the expectations of wavelet periodograms encode, in a one-to-one way, the entire autocovariance structure of a time series, so it suffices to estimate change-points in those expectations in order to obtain segmentation of the autocovariance structure of X_t, which is our ultimate goal.

We also recall two further definitions from Nason et al. (2000): the autocorrelation wavelets $\Psi_i(\tau) = \sum_k \psi_{i,k}\psi_{i,k-\tau}$ and the autocorrelation wavelet inner product matrix $A_{i,k} = \sum_\tau \Psi_i(\tau)\Psi_k(\tau)$. Fryzlewicz and Nason (2006) show that $\mathbb{E}I^{(i)}_{t,T}$ is "close" (in the sense that the integrated squared bias converges to zero) to the function $\beta_i(z) = \sum_{j=-\infty}^{-1} S_j(z)A_{i,j}$, a piecewise constant function with at most N change-points, whose set is denoted by $\mathcal{N}$. Every change-point in the autocovariance structure of the time series results in a change-point in at least one of the $\beta_i(z)$; therefore, detecting a change-point in the wavelet periodogram implies a change-point in the autocovariance structure of the process.

In addition, note that each wavelet periodogram ordinate is a squared wavelet coefficient of a standard Gaussian time series and it satisfies

$$I^{(i)}_{t,T} = \mathbb{E}I^{(i)}_{t,T} Z^2_{t,T} \tag{1.8}$$

where $\{Z_{t,T}\}_{t=0}^{T-1}$ are autocorrelated standard normal variables (or equivalently the distribution of the squared wavelet coefficient $I^{(i)}_{t,T}$ is that of a scaled χ^2_1 variable). Then, the quantities $I^{(i)}_{t,T}$ and $\mathbb{E}I^{(i)}_{t,T}$ can be seen as special cases of $Y^2_{t,T}$ and $\sigma^2_{t,T}$ respectively of the multiplicative model (1.1). To enable the application of the model (1.8) in this context, we assume the following condition:

(A0): $\sigma^2_{t,T}$ is deterministic and "close" to a piecewise constant function $\sigma^2(t/T)$ (apart from intervals around the discontinuities in $\sigma^2(t/T)$ which have length at most $K2^{-i}$) in the sense that $T^{-1}\sum_{t=0}^{T-1} |\sigma^2_{t,T} - \sigma^2(t/T)|^2 = o(\log^{-1} T)$ where the rate of convergence comes from the integrated squared bias between $\beta_i(t/T)$ and $\mathbb{E}I^{(i)}_{t,T}$ (see Fryzlewicz and Nason (2006)).

4 The Algorithm

In this section we present the WBS algorithm within the framework of the LSW model. First, we form the following CUSUM-type statistic

$$\mathbb{Y}^{b(i)}_{s_m,e_m} = \sqrt{\frac{e_m - b}{n(b - s_m + 1)}} \sum_{t=s_m}^{b} I^{(i)}_{t,T} - \sqrt{\frac{b - s_m + 1}{n(e_m - b)}} \sum_{t=b+1}^{e_m} I^{(i)}_{t,T} \tag{3.1}$$

where the subscript $(.)_m$ denotes an element chosen randomly from the set $\{0, ..., T-1\}$ as in (1.3), $n = e_m - s_m + 1$ and $I^{(i)}_{t,T}$ are the wavelet periodogram ordinates at scale i that form the multiplicative model $I^{(i)}_{t,T} = \mathbb{E}I^{(i)}_{t,T} Z^2_{t,T}$ discussed in Section 3. The likely location of a change-point b_0 is then given by (1.4).

The following stages summarise the recursive procedure:

Stage I: Start with $s = 1$ and $e = T$.

Stage II: Examine whether $h_{m_0} = |\mathbb{Y}^{b_0}_{s_{m_0},e_{m_0}}|/q_{s_{m_0},e_{m_0}} > \omega_T = C\log(T)$ where $q_{s_{m_0},e_{m_0}} = \sum_{t=s_{m_0}}^{e_{m_0}} I^{(i)}_{t,T}/n_{m_0}$, $n_{m_0} = e_{m_0} - s_{m_0} + 1$ and m_0, b_0 as in (1.4); C is a parameter that remains constant and only varies between scales. Define $h'_{m_0} = h_{m_0}\mathbb{I}(h_{m_0} > \omega_T)$ where $\mathbb{I}(.)$ is 1 if the inequality is satisfied and 0 otherwise.

Stage III: If $h'_{m_0} > 0$, then add b_0 to the set of estimated change-points; otherwise if $h'_{m_0} = 0$ stop the algorithm.

Stage IV: Repeat stages II-III to each of the two segments $(s, e) = (1, b_0)$ and $(s, e) = (b_0 + 1, T)$ if their length is more than Δ_T.

The choice of parameters C and Δ_T is described in Section 4.4. We note that in addition to the random intervals $[s_m, e_m]$ we also include into $\mathcal{M}_{s,e}$ the index (labelled 0) corresponding to the interval $[s, e]$. This does *not* mean that the WBS procedure "includes" the classical BS, as at the first stage the WBS and BS are not guaranteed to locate the same change-point (even if WBS also examines the full interval $[s, e]$), so the two procedures can "go their separate ways" after examining the first full interval. The reason for manually including the full interval $[s, e]$ is that if there is at most one change-point in $[s, e]$, considering the entire interval $[s, e]$ is an optimal thing to do.

Further, we expect that finer scales will be more useful in detecting the number and locations of the change-points in $\mathbb{E}I^{(i)}_{t,T}$. This is because as we move to coarser scales the autocorrelation within $I^{(i)}_{t,T}$ becomes stronger and the in-

tervals on which a wavelet periodogram sequence is not piecewise constant become longer. Hence, we select the scale $i < -I^\star$ where $I^\star = \lfloor \alpha \log\log T \rfloor$ and $\alpha \in (0, 3\lambda]$ for $\lambda > 0$ such that the consistency of our method is retained.

In stage II, we rescale the statistic h_{m_0} before we test it against the threshold. This division plays the role of stabilising the variance, which is exact in the multiplicative model in which the observations are independent, over intervals where the variance is constant. In all other cases, the variance stabilisation cannot be guaranteed to be exact, but in the case where the process under consideration is stationary over the given interval (i.e. there are no change-points), the cancellation of the variance parameter still takes place and therefore the distribution of the rescaled CUSUM is a function of the *autocorrelation* of the process, rather than its entire *autocovariance*. This reduces the difficulty in choosing the threshold parameter ω_T and one can still hope to obtain "universal" thresholds which work well over a wide range of dependence structures. Exact variance stabilisation in the non-independent case would require estimating what is referred to as the "long-run variance" parameter (the variance of the sample mean of a time series), which is a known difficult problem in time series analysis and if we were to pursue it, the estimation error would likely not make it worthwhile. In other words, we choose this rescaling method as a mid-way compromise between doing nothing and having to estimate the long-run variance. In Section 4 of the Supplementary material, we illustrate the essence of this issue, and provide a simple but informative example showing that the variance stabilisation, performed as described above, is still desirable, despite this "non-exactness" problem.

Finally, we notice that Horváth et al. (2008) propose a similar type of CUSUM statistic which does not require an estimate of the variance of a stochastic process by using the ratio of the maximum of two local means. The authors apply the method to detect a single change-point in the mean of a stochastic process under independent, correlated or heteroscedastic error settings.

4.1 Technical assumptions and consistency

In this section we present the consistency theorem for the WBS algorithm for the total number N and locations of the change-points $0 < \eta_1 < ... < \eta_N < T - 1$ with $\eta_0 = 0$ and $\eta_{N+1} = T$. To achieve consistency, we impose the following assumptions:

(A1): $\sigma^2(t/T)$ is bounded from above and away from zero, i.e. $0 < \sigma^2(t/T) < \sigma^\star < \infty$ where $\sigma^\star \leq \max_{t,T} \sigma^2(t/T)$. Further, the number of change-points N in (1.1) is unknown and allowed to increase with T i.e. only the minimum distance between the change-points can restrict the maximum number of N.

(A2): $\{Z_{t,T}\}_{t=0}^{T-1}$ is a sequence of standard Gaussian variables and the autocorrelation function $\rho(\tau) = \sup_{t,T} |\text{cor}(Z_{t,T}, Z_{t+\tau,T})|$ is absolutely summable, that is it satisfies $\rho_\infty^1 < \infty$ where $\rho_\infty^p = \sum_\tau |\rho(\tau)|^p$.

(A3): The distance between any two adjacent change-points satisfies $\min_{r=1,\dots,N+1} |\eta_r - \eta_{r-1}| \geq \delta_T$, where $\delta_T \geq C \log^2 T$ for a large enough C.

(A4): The magnitude of the change-points satisfy $\inf_{1\leq r\leq N} |\sigma((\eta_r+1)/T) - \sigma(\eta_r/T)| \geq \sigma_\star$ where $\sigma_\star > 0$.

(A5): $\Delta_T \asymp \delta_T$ where Δ_T as defined in (1.3).

Theorem 1 *Let $Y_{t,T}^2$ follow model (1.1), and suppose that Assumptions (A0)-(A5) hold. Denote the number of change-points in $\sigma^2(t/T)$ by N and the locations of those change-points as $\eta_1, ..., \eta_N$. Let $\hat{N}$ and $\hat{\eta}_1, ..., \hat{\eta}_N$ be the number and locations of the change-points (in ascending order), respectively, estimated by the Wild Binary Segmentation algorithm. There exist two constants C_1 and C_2 such that if $C_1 \log T \leq \omega_T \leq C_2\sqrt{\delta_T}$, then $P(\mathcal{Z}_T) \to 1$, where*

$$\mathcal{Z}_T = \{\hat{N} = N; \quad \max_{r=1,\dots,N} |\hat{\eta}_r - \eta_r| \leq C \log^2 T\}$$

for a certain $C > 0$, where the guaranteed speed of convergence of $P(\mathcal{Z}_T)$ to 1 is no faster than $T\delta_T^{-1}(1 - \delta_T^2(1-\bar{c})^2 T^{-2}/9)^M$ where M is the number of random draws and $\bar{c} = 3 - 2/c_\star$.

For the purpose of comparison we note that the rate of convergence for the estimated change-points obtained for the BS method by Cho and Fryzlewicz (2015) is of order $\mathcal{O}(\sqrt{T}\log^{(2+\vartheta)} T)$ and $\mathcal{O}(\log^{(2+\vartheta)} T)$ for $\vartheta > 0$ when δ_T is $T^{3/4}$ and T respectively. In the WBS setting, the rate is square logarithmic when δ_T is of order $\log^2 T$, which represents an improvement. In addition, the lower threshold is always of order $\log T$ regardless of the minimum space between the change-points.

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 $P(\mathcal{Z}_T)$ to 1 in Theorem

1 is suitably small. Suppose that we wish to ensure

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

Bearing in mind that $\log(1-y) \approx -y$ around $y=0$, this is, after simple algebra, (practically) equivalent to

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

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 4.4.

4.2 Simultaneous across-scale post-processing

Theorem 1 covers the case of the multiplicative model (1.1). We now consider change-point detection in the full model (1.6). Recall from Section 3 that a change-point in $\beta_i(z)$ for $i=-1,-2,...,-I^\star$ would signal a change-point in the second-order covariance structure of $X_{t,T}$. To accomplish this we propose two methods.

Method 1: The search for further change-points in each interval (s_m, e_m) proceeds to the next scale $i-1$ only if no change-points are detected at scale i on that interval. It therefore ensures that the finest scales are preferred (since change-points detected at the finest scales are likely to be more accurate) and only moves to coarser if necessary. Cho and Fryzlewicz (2012) use a similar technique to combine across scales change-points, but involving an extra parameter. The role of this parameter is to create groups of estimated change-points which are close to each other. Then, only one change-point (detected at the finest scale) from each of these groups will survive the post-processing. Hence, their method will be used as a benchmark for our first type of across-scale post-processing.

Method 2: Alternatively, we suggest a method that simultaneously joins the estimated change-points across all the scales such that all the information from every scale is combined. Namely, motivated by Cho and Fryzlewicz (2015) who propose an alternative aggregation method to these of Groen et al. (2013) in order to detect change-points in the second order structure of a high-dimensional

time series we define the following statistic

$$\mathbb{Y}_t^{thr} = \sum_{i=-I^\star}^{-1} \mathcal{Y}_t^{(i)} \mathbb{I}(\mathcal{Y}_t^{(i)} > \omega_T^{(i)}) \text{ for } i = -1, ..., -I^\star \tag{3.2}$$

where $\mathcal{Y}_t^{(i)} = |\mathbb{Y}_{s_m,e_m}^{b(i)}|/q_{s_m,e_m}^{(i)}|$. This statistic differs from that of Cho and Fryzlewicz (2015) in that it applies across the scales $i = -1, -2, ..., -I^\star$ of a univariate time series, whereas Cho and Fryzlewicz (2015) calculate their statistic across multiple time series.

The algorithm is identical to the algorithm in Section 4 except for replacing (3.1) with (3.2). In addition, if the obtained $\mathbb{Y}_t^{thr} > 0$ there is no need to test further for the significance of b_0.

Below, we present the consistency theorem for the across-scale post-processing algorithm:

Theorem 2 *Let X_t follow model (1.6), and suppose that Assumptions (A0)-(A5) for $\sigma^2(t/T)$ hold for each $\beta_i(z)$. Denote the number of change-points in $\beta_i(z)$ as N and the locations of those change-points as $\theta_1, ..., \theta_N$. Let $\hat{N}$ and $\hat{\theta}_1, ..., \hat{\theta}_N$ be the number and locations of the change-points (in ascending order), respectively, estimated by the across-scale post-processing method 1 or 2. There exist two constants C_3 and C_4 such that if $C_3 \log T \leq \omega_T \leq C_4 \delta_T$, then $P(\mathcal{U}_T) \to 1$, where*

$$\mathcal{U}_T = \{\hat{N} = N; \quad \max_{r=1,...,N} |\hat{\theta}_r - \theta_r| \leq C' \log^2 T\}$$

for a certain $C' > 0$, where the guaranteed speed of convergence is the same as in Theorem 1.

Even though the two Methods 1 and 2 achieve the same rate of convergence for the estimated change-points, their relative performance is empirically examined in Section 5.

4.3 Post-processing

In order to control the number of change-points estimated from the WBS algorithm and to reduce the risk of over-segmentation we propose a post-processing method similar to Cho and Fryzlewicz (2012) and Inclan and Tiao (1994). More specifically, we compare every change-point against the adjacent ones using the CUSUM statistic making sure that (1.5) is satisfied. That is, for a set $\hat{\mathcal{N}} =$

$\{\hat{\theta}_0, ..., \hat{\theta}_{N+1}\}$ where $\hat{\theta}_0 = 0$ and $\hat{\theta}_{N+1} = T$ we test whether $\hat{\theta}_r$ satisfies

$$\mathbb{Y}_t^{thr} = \sum_{i=-I^\star}^{-1} \mathcal{Y}_t^{(i)} \mathbb{I}(\mathcal{Y}_t^{(i)} > \omega_T^{(i)}) > 0 \text{ for } i = -1, ..., -I^\star$$

where $\mathcal{Y}_t^{(i)} = |\mathbb{Y}^{\hat{\theta}_r(i)}_{\hat{\theta}_{r-1},\hat{\theta}_{r+1}}| / q^{(i)}_{\hat{\theta}_{r-1},\hat{\theta}_{r+1}}|$ and

$$\max\left(\frac{\hat{\theta}_{r+1} - \hat{\theta}_r}{\hat{\theta}_{r+1} - \hat{\theta}_{r-1} + 1}, \frac{\hat{\theta}_r - \hat{\theta}_{r-1} + 1}{\hat{\theta}_{r+1} - \hat{\theta}_{r-1} + 1}\right) \leq c_\star. \tag{3.3}$$

If $\mathbb{Y}_t^{thr} = 0$ then change-point $\hat{\theta}_r$ is temporarily eliminated from set $\hat{\mathcal{N}}$. In the next run, when considering change-point $\hat{\theta}_{r+1}$, the adjacent change-points are $\hat{\theta}_{r-1}$ and $\hat{\theta}_{r+2}$. When the post-processing finishes its cycle all temporarily eliminated change-points are reconsidered using as adjacent change-points those that have survived the first cycle. It is necessary for $\hat{\theta}_r$ to satisfy (3.3) with its adjacent estimated change-points $\hat{\theta}_{r-1}$ and $\hat{\theta}_{r+1}$, otherwise it is never eliminated. The algorithm is terminated when the set of change-points does not change.

The post-processing step does not involve any extra parameters since it only uses those already mentioned in Section 4. In the next section we discuss the choice of the parameters.

4.4 Choice of threshold and parameters

In this section we present the choices of the parameters involved in the algorithms. From Theorems 1 and 2 we have that the threshold ω_T includes the constant $C^{(i)}$ which varies between the scales. The values of $C^{(i)}$ will be the same for all the methods presented, either BS/WBS or the Methods 1 and 2 in Section 4.2. Therefore, we can use the thresholds by Cho and Fryzlewicz (2012) who conduct experiments to establish the value of the threshold parameter under the null hypothesis of no change-points such that when the obtained statistic exceeds the threshold the null hypothesis is rejected. However, in that work the threshold is of the form $\tau_0 T^{\vartheta_0}\sqrt{\log T}$ where $\vartheta_0 \in (1/4, 1/2)$ and $\tau_0 > 0$ is the parameter that changes across scales. For that reason, we repeat the experiments which are described below.

We generate a vector $\mathbf{X} \sim N(0, \Sigma)$ where the covariance matrix $\Sigma = (\sigma_{\kappa,\kappa'})_{\kappa,\kappa'=1}^T$

and $\sigma_{\kappa,\kappa'} = \rho^{|\kappa-\kappa'|}$. Then we find v that maximises (3.1). The following ratio

$$C_T^{(i)} = \mathbb{Y}_v^{(i)} (\log T)^{-1} \left(\sum_{t=1}^{T} I_{t,T}^{(i)} \right)^{-1} T$$

gives us an insight into the magnitude of parameter $C^{(i)}$. We repeat the experiment for different values of ρ and for every scale i we select $C^{(i)}$ as the 95% quantile. The same values are used for the post-processing method explained in Section 4.3. Our results indicate that $C^{(i)}$ tends to increase as we move to coarser scales due to the increasing dependence in the wavelet periodogram sequences. Since our method applies to non-dyadic structures it is reasonable to propose a general rule that will apply in most cases. To accomplish this we repeated the simulation study described above for $T = 50, 100, ..., 6000$. Then, for each scale i we fitted the following regression

$$C^{(i)} = c_0^{(i)} + c_1^{(i)} T + c_2^{(i)} \frac{1}{T} + c_3^{(i)} T^2 + \varepsilon.$$

The adjusted R^2 was above 90% for all the scales. Having estimated the values for $\hat{c}_0^{(i)}, \hat{c}_1^{(i)}, \hat{c}_2^{(i)}, \hat{c}_3^{(i)}$ we were able to use fitted values for any sample size T. For samples larger than $T = 6000$ we used the same $C^{(i)}$ values as for $T = 6000$.

Further, based on empirical evidence (see the online supplementary material) we select the scale $I^\star$ by setting $\lambda = 0.7$. In stage III of the algorithm, the procedure is terminated when either the CUSUM statistic does not exceed a certain threshold or the length of the respective segment is Δ_T. This also defines the minimum length of a favourable draw from (1.3). We choose Δ_T to be of the same order as δ_T since this is the lowest permissible order of magnitude according to (A5). Practically, we find that the choice $\Delta_T = \lfloor \log^2 T/3 \rfloor$ works well. In addition, a simulation study found in the online supplementary material provides empirical arguments for the choice $c_\star = 0.75$. The main idea of this parameter is to ensure that long enough stretches of data are included in the computation of our CUSUM statistics, or otherwise the computed CUSUM statistics will be too variable to be reliable. This is particularly important in the autocorrelated multiplicative setting where there tends to be a large amount of noise so the use of such a parameter is needed to suppress the variance of the CUSUM statistics. Finally, our recommendation for the parameter M is 3500 when T does not exceed 10000. These values are used in the remainder of the paper.

Table 4.1: Stationary processes results. For all the models the sample size is 1024 and there are no change-points. Figures show the number of occasions the methods detected change-points with the universal thresholds $C^{(i)}$ obtained as described in Section 4.4. Figures in brackets are the number of occasions the methods detected change-points with the thresholds $C^{(i)}$ obtained as described in Section 5.1.

Model	BS1	WBS1	BS2	WBS2	CF
S1: iid standard normal	1 [0]	3 [2]	0 [0]	1 [0]	4
S2: AR(1) with parameter 0.9	3 [1]	5 [1]	1 [1]	5 [1]	9
S3: AR(1) with parameter -0.9	58 [0]	93 [0]	46 [0]	48 [5]	79
S4: MA(1) with parameter 0.8	2 [3]	7 [4]	3 [3]	1 [0]	7
S5: MA(1) with parameter -0.8	2 [0]	4 [2]	4 [0]	0 [0]	7
S6: ARMA(1,0,2) with AR= $\{-0.4\}$ and MA= $\{-0.8, 0.4\}$	8 [0]	27 [0]	8 [0]	8 [0]	25
S7: AR(2) with parameters 1.39 and -0.96	88 [3]	99 [4]	88 [3]	88 [5]	96

5 Simulation study

We present a set of simulation studies to assess the performance of our methods. In all the simulations we assume sample sizes to be 1024 over 100 iterations. In the online supplementary material smaller and larger sample sizes are also considered. For comparison we also report the performance of the method by Cho and Fryzlewicz (2012) – henceforth CF – using the default values specified in their paper. BS1 and BS2 refer to the Method 1 and Method 2 of aggregation (as described in Section 4.2) using the BS technique, respectively. WBS1 and WBS2 refer to the Method 1 and Method 2 of aggregation (as in Section 4.2) using the Wild Binary Segmentation technique, respectively. To briefly illustrate computation times, our code, executed on a standard PC, runs in approximately 25 seconds for a time series of length 10000 with 10 change-points.

5.1 Models with no change-points

We simulate stationary time series with innovations $\varepsilon_t \sim N(0, 1)$ and we report the number of occasions (out of 100) the methods incorrectly rejected the null hypothesis of no change-points. The models S1-S7 (Table 4.1) we consider here are taken from Nason (2013).

The results of Table 4.1 indicate our methods' good performance over that of Cho and Fryzlewicz (2012) apart from models S3 and S7 where all methods incorrectly reject the null hypothesis on many occasions. A visual inspection of an AR(1) process with $\phi = -0.9$ could confirm that this type of process exhibits

a "clustering behaviour" which mimics changing variance. Hence, the process is interpreted as non-stationary by the wavelet periodogram resulting in erroneous outcomes. A similar argument is valid for S7 model. To correct that limitation, parameter $C^{(i)}$ should be chosen with care. Higher values will ensure that the null hypothesis is not rejected frequently. This is achieved by not using universal thresholds (as shown in Section 4.4) but calculating them for every instance. Specifically, given a time series y_t we fit an AR(p) model. Then we generate 100 instances of the same length and with the same AR(p) coefficients. Similarly with Section 4.4 we select $C^{(i)}$ as the 95% quantile. This procedure is more computationally intensive but improves the method significantly; see the figures in brackets (Table 4.1). An alternative approach in obtaining thresholds, by taking time-averages of spectrum values for each $i = -1, -2, ..., -I^\star$ and then simulating stationary models, described in the online supplementary material, does generally well but not as well as our suggestion above.

5.2 Non-stationary models

We now examine the performance of our method for a set of non-stationary models by using and extending the examples from Cho and Fryzlewicz (2012). Since the WBS method has improved rates of convergence new simulation results are presented which assess how close to the real change-points the estimated ones are. For this reason we report the total number of change-points identified within $\lfloor 5\% \cdot T \rfloor$ from the real ones. Results for $\lfloor 2.5\% \cdot T \rfloor$ distances are reported in the online supplementary material.

The accuracy of a method should be also judged in parallel with the total number of change-points identified. We propose a test that tries to accomplish this. Assuming that we define the maximum distance from a real change-point η as $d_{\max}$, an estimated change-point $\hat{\eta}$ is correctly identified if $|\eta - \hat{\eta}| \leq d_{\max}$ (here within 5% of the sample size). If two (or more) estimated change-points are within this distance then only one change-point which is the closest to the real change-point is classified as correct. The rest are deemed to be false, except if any of these are close to another change-point. An estimator performs well when the following hit ratio HR

$$HR = \frac{\#\text{correct change-points identified}}{\max(N, \hat{N})}$$

is close to 1. By using the term $\max(N, \hat{N})$ we aim to penalise cases where, for example, the estimator correctly identifies a certain number of change-points all within the distance $d_{\max}$ but $\hat{N} < N$. It also penalises the estimator when $\hat{N} > N$ and all $\hat{N}$ estimated change-points are within the distance $d_{\max}$ of the true ones.

Tables 4.2 and 4.3 summarise the results, and histograms of the estimated change-point locations for every model can be found in the supplementary material.

Model A: *A non-stationary process that includes one AR(1) and two AR(2) processes with two clearly observable change-points*

$$y_t = \begin{cases} 0.9y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 1 \le t \le 512 \\ 1.68y_{t-1} - 0.81y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 513 \le t \le 768 \\ 1.32y_{t-1} - 0.81y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 769 \le t \le 1024. \end{cases}$$

BS2 is the best option, marginally ahead of WBS1 and WBS2. The fact that BS performs well here is unsurprising given the fact that the change-points are far apart and prominent. However, it is reassuring to see the WBS methods also performing well.

Model B: *A non-stationary process with two less clearly observable change-points*

$$y_t = \begin{cases} 0.4y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 1 \le t \le 400 \\ -0.6y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 401 \le t \le 612 \\ 0.5y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 613 \le t \le 1024. \end{cases}$$

The WBS methods move into the lead, marginally ahead of the BS methods. This is again not unexpected given the fact that the change-points here are less prominent than in Model A.

Model C: *A non-stationary process with a short segment at the start*

$$y_t = \begin{cases} 0.75y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 1 \le t \le 50 \\ -0.5y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 51 \le t \le 1024. \end{cases}$$

In this type of model both BS2 and CF perform well compared to the BS1, WBS1 and WBS2 methods. It is expected that binary segmentation methods

will perform better due to the fact that they start their search assuming a single change-point.

Model D: *A non-stationary process similar to model B but with the two change-points at a short distance from each other.* In this model, the two change-points occur very close to each other i.e. $(400, 470)$ instead of $(400, 612)$. The CF method, BS1 and BS2 do not perform well as the two change-points were detected in less than half of the cases. By contrast, the WBS1 and WBS2 methods achieved high hit ratio (almost double that of the BS methods).

Model E: *A highly persistent non-stationary process with time-varying variance*

$$y_t = \begin{cases} 1.399y_{t-1} - 0.4y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim N(0, 0.8) & \text{for } 1 \leq t \leq 400 \\ 0.999y_{t-1} + \epsilon_t,\ \varepsilon_t \sim N(0, 1.2^2) & \text{for } 401 \leq t \leq 750 \\ 0.699y_{t-1} + 0.3y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, 1) & \text{for } 751 \leq t \leq 1024. \end{cases}$$

The CF and BS1 methods perform well since they detect most of the change-points within 5% distance from the real ones. From our simulations we noticed that in most cases the two change-points where found in the finest scale ($i = -1$). The aggregation Method 2 does not improve the estimation since its purpose is to simultaneously combine the information from different scales not just from a single one. On the other hand, the CF method and Method 1 favour change-points detected in the finest scales and this is the reason for their good performance.

Model F: *A piecewise constant ARMA(1,1) process*

$$y_t = \begin{cases} 0.7y_{t-1} + \epsilon_t + 0.6\varepsilon_{t-1}, & \text{for } 1 \leq t \leq 125 \\ 0.3y_{t-1} + \epsilon_t + 0.3\varepsilon_{t-1}, & \text{for } 126 \leq t \leq 532 \\ 0.9y_{t-1} + \epsilon_t, & \text{for } 533 \leq t \leq 704 \\ 0.1y_{t-1} + \epsilon_t - 0.5\varepsilon_{t-1}, & \text{for } 704 \leq t \leq 1024. \end{cases}$$

The first change-point is the least apparent and is left undetected in most cases when applying the CF method. Our methods are capable of capturing this point more frequently and within 5% from its real position.

Model G: *A near-unit-root non-stationary process with time-varying variance*

$$y_t = \begin{cases} 0.999y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1) & \text{for } 1 \le t \le 200,\ 401 \le t \le 600 \text{ and } 801 \le t \le 1024 \\ 0.999y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0,1.5^2) & \text{for } 201 \le t \le 400 \text{ and } 601 \le t \le 800. \end{cases}$$

In this near-unit-root process there are 4 change-points in its variance. All binary segmentation methods do not perform well as they often miss the middle change-points. Both WBS1 and WBS2 manage to detect most of the change-points achieving a hit ratio almost three times higher than BS2. In almost 70% of the occasions WBS2 detects at least 4 change-points.

Model H: *A non-stationary process similar to model F but with the three change-points at a short distance from each other.* In this model the three change-points occur close to each other, i.e. $\mathcal{N} = (125, 325, 550)$. The first two change-points fail to be detected by the CF in many instances. By contrast, BS1 and BS2 do well while WBS1 and WBS2 perform slightly better in this case by identifying them more often. This results in a higher hit ratio.

Model I: *A non-stationary AR process with many changes within close distances.* We simulate instances with 5 change-points occurring at uniformly distributed positions. We allow the distances to be as small as 30 and not larger than 100.

In this scenario, CF correctly identifies more than 4 change-points in 15% instances while BS1 and BS2 in 24% and 23% respectively. Again, the WBS methods do well in revealing the majority of the change-points and in many cases close to the real ones.

In summary, the WBS methods offer a reliable default choice. In terms of the hit ratio, they perform the best or nearly the best in 7 of the 9 models studied, and do not perform particularly poorly in the other 2 models, especially if the total number of detected change-points is also taken into account. All of: BS1, BS1 and CF perform poorly in at least 3 of the models. In terms of the hit ratio, both BS methods are in or close to the lead only in 2 models. Overall, the WBS methods seem to be clear winners here. Our recommendation to the user is to try the WBS2 method first since overall it appears to be the most reliable one.

Table 4.2: Non-stationary processes results for $T = 1024$ (Models A - I). Table shows the number of occasions a method detected the given number of change-points within a distance of 5% from the real ones. Bold: the method with the highest hit ratio or within 10% from the highest.

Number of Change-points															
Model	A					B					C				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	2	0	1	0	3	0	0	0	0	0	39	12	35	21	6
1	29	15	16	21	29	11	8	4	9	7	61	88	65	79	94
2	69	85	83	79	68	89	92	96	91	93	-	-	-	-	-
Hit ratio	**0.768**	**0.850**	**0.817**	**0.808**	0.712	**0.928**	**0.921**	**0.966**	**0.928**	0.865	0.580	**0.860**	0.600	0.746	**0.853**
Model	D					E					F				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	36	52	12	11	48	6	12	8	11	1	2	0	0	0	1
1	58	14	9	11	12	40	42	59	53	40	18	6	5	3	7
2	6	34	79	78	40	54	46	33	36	59	32	32	22	24	45
3	-	-	-	-	-	-	-	-	-	-	48	62	73	73	47
Hit ratio	0.428	0.403	**0.835**	**0.835**	0.436	**0.712**	0.649	0.610	0.611	**0.743**	0.744	**0.847**	**0.890**	**0.894**	0.765
Model	G					H					I				
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	58	60	9	11	39	0	0	2	2	0	0	2	1	0	0
1	11	11	13	6	20	40	33	23	16	29	39	33	8	8	39
2	20	21	20	20	30	38	37	38	40	57	16	15	8	7	27
3	6	5	15	22	5	22	30	37	42	14	23	27	20	18	25
4	5	3	43	41	6	-	-	-	-	-	14	11	22	18	3
5	-	-	-	-	-	-	-	-	-	-	8	12	41	49	6
Hit ratio	0.222	0.200	**0.671**	**0.686**	0.297	0.605	0.654	**0.693**	**0.732**	0.603	0.472	0.496	**0.745**	**0.779**	0.419

Table 4.3: Non-stationary processes results for $T = 1024$ (Models A - I). Table shows the percentage of occasions a method detected the given number of change-points. True number of change-points is in bold.

Number of Change-points															
Model			A					B					C		
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	0	0	0	0	0	0	0	0	0	0	34	9	6	4	2
1	5	0	0	1	0	0	0	0	0	1	**59**	**86**	**78**	**84**	**81**
2	**59**	**77**	**65**	**70**	**65**	**78**	**81**	**79**	**80**	**70**	7	5	11	8	16
$\geq$ **3**	36	23	35	30	35	22	19	21	20	29	0	0	5	4	1
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Model			D					E					F		
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	49	42	8	3	38	0	0	1	1	0	0	0	0	0	0
1	5	9	0	1	17	22	21	22	24	19	14	0	0	0	1
2	**45**	**45**	**87**	**88**	**38**	**63**	**65**	**65**	**61**	**65**	13	9	12	8	19
3	1	4	5	8	7	14	11	10	12	15	**63**	**82**	**78**	**81**	**65**
$\geq$ **4**	0	0	0	0	0	1	3	2	2	1	10	9	10	11	15
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Model			G					H					I		
	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF	BS1	BS2	WBS1	WBS2	CF
0	59	59	7	4	38	0	0	0	0	0	0	0	0	0	0
1	7	8	3	2	16	24	20	13	16	12	33	30	8	1	22
2	23	21	17	22	32	32	24	30	22	51	9	6	2	2	28
3	1	2	4	2	3	**41**	**50**	**48**	**55**	**30**	22	23	10	11	24
4	**9**	**10**	**62**	**66**	**11**	3	6	7	6	7	12	18	14	13	11
$\geq$ **5**	1	0	7	4	0	0	0	2	1	0	**24**	**23**	**66**	**73**	**15**
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

6 Applications

6.1 US Gross National Product series (GNP)

We obtain the GNP time series from the Federal Reserve Bank of St. Louis web page (http://research.stlouisfed.org/fred2/series/GNP). The seasonally adjusted and quarterly data is expressed in billions of dollars and spans from 1947:1 until 2013:1 but we only use the last 256 observations. In the left panel of Figure 5.2 one can see the logarithm of the GNP series. As in Shumway and Stoffer (2011), we only examine the first difference of the logarithm of the GNP (also called the growth rate) since there is an obvious linear trend. In the right panel of the same figure, which illustrates the growth rate, it is visually clear that the series exhibits less variability in the latter part. We are interested in finding whether

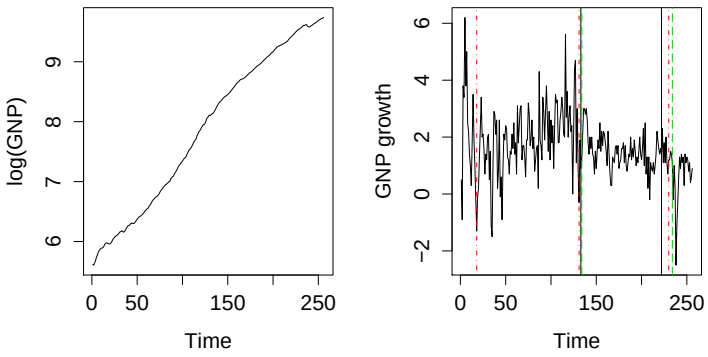

Figure 5.2: Natural logarithm of the GNP series (left) and its first difference (right). The black, green and red vertical lines are the change-points as estimated by BS2, CF and WBS2 respectively.

our method is capable of spotting this change and/or possibly others.

Applying our method i.e. BS2 and WBS2 (BS1 and WBS1 produced identical results) we find that BS2 detects two change-points $\hat{\eta} = \{133, 222\}$ while the WBS2 detects three at positions $\{18, 131, 230\}$. For the sake of comparison, CF detects two possible change-points i.e. $\hat{\eta} = \{134, 234\}$. The acf graphs (not shown here) confirm that there may be changes in the autocovariance structure occurring at all of these estimated change-points.

Change-point 18 i.e. 1953(3) almost exactly coincides with a peak of the GNP growth as decided by the Business Cycle Dating Committee of the National Bureau of Economic Research where the official date is July 1953 (note that cycles do not necessarily overlap with the quarterly publications of the GNP). In addition, change-points 131, 133 and 134 lie within a cycle that peaks in January 1981 and has a trough in November 1982. This cycle corresponds to the start of the Great Moderation (around 1980s), a period that experienced more efficient monetary policy and shocks of small magnitude, see Clark (2009) and references therein. Finally, we note that all three methods detected a change-point towards the end of the series - $222, 230, 234$ which are dated 2004(3), 2006(3) and 2007(3) respectively. According to e.g. Clark (2009) the Great Moderation had reversed and the decline was offset by negative growth rates due to the recent economic recession.

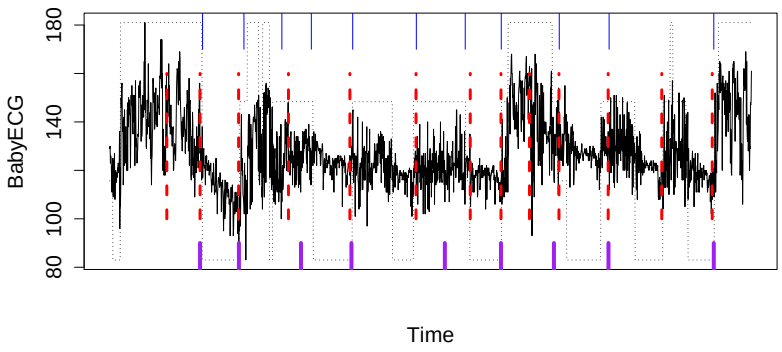

Figure 5.3: Plot of BabyECG data. The top blue, middle red and bottom purple vertical lines are the change-points as estimated by CF, WBS2 and BS2 respectively. The horizontal dotted line represents the sleep states i.e. 1 = quiet sleep, 2 = quiet-to-active sleep, 3 = active sleep, 4 =awake.

6.2 Infant Electrocardiogram Data (ECG)

We apply the three methods (CF, BS2, WBS2) to the ECG data of an infant found at the *R* package *wavethresh*. This is a popular example of a non-stationary time series and it has been analysed in e.g. Nason et al. (2000). The local segments of possible stationarity indicate the sleep state of the infant and it is classified on a scale from 1 to 4, see the caption of Figure 5.3. The same figure plots the time series with the respective estimated change-points (the methods were applied on the first difference so that its mean is approximately zero). All methods identify most of the sleep states and, notably, WBS2 detects an abrupt change of short duration (quiet sleep-awake-quiet sleep) towards the end of the series.

7 Conclusion

The work in this paper has addressed the problem of detecting the change-points in the autocovariance structure of a univariate time series. As discussed in the Introduction, there are many types of non-stationary time series which require segmentation methods. Using the WBS framework we are able to detect multiple change-points that are small in magnitude and/or close to each other. The simulation study in Section 5 indicates that the WBS mechanism leads to a well-performing methodology for this task.

8 Supplementary Materials

The online supplementary material contains additional simulation studies supporting the choice of the default parameters of our procedure, empirical performance evaluation for small and large samples and using other error measures, as well as additional material on the variance stabilization.

9 Acknowledgements

Piotr Fryzlewicz's work was supported by the Engineering and Physical Sciences Research Council grant no. EP/L014246/1.

Appendix

Proof of Theorem 1

The proof of consistency is based on the following multiplicative model

$$\tilde{Y}_{t,T} = \sigma(t/T)^2 Z_{t,T}^2, \;\; t = 0, ..., T-1.$$

We define the following two CUSUM statistics

$$\mathbb{Y}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \tilde{Y}_{t,T} - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \tilde{Y}_{t,T}$$

and

$$\mathbb{S}_{s,e}^b = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^{b} \sigma^2(t/T) - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^{e} \sigma^2(t/T)$$

where $n = e - s + 1$, the size of the segment defined by (s, e).

$\mathbb{Y}_{s,e}^b$ can be seen as the inner product between sequence $\{\tilde{Y}_{t,T}\}_{t=s,...,e}$ and a vector $\psi_{s,e}^b$ whose elements $\psi_{s,e,t}^b$ are constant and positive for $t \leq b$ and constant and negative for $t > b$ such that they sum to zero and sum to one when squared. Similarly for $\mathbb{S}_{s,e}^b$.

Let s, e satisfy $\eta_{p_0} \leq s < \eta_{p_0+1} < ... < \eta_{p_0+q} < e \leq \eta_{p_0+q+1}$ for $0 \leq p_0 \leq N - q$. The inequality will hold at all stages of the algorithm until no undetected change-points are remained. We impose at least one of the following conditions

$$s < \eta_{p_0+r'} - C\delta_T < \eta_{p_0+r'} + C\delta_T < e, \quad \text{for some } 1 \leq r' \leq q \tag{6.1}$$

$$\{(\eta_{p_0+1} - s) \wedge (s - \eta_{p_0})\} \vee \{(\eta_{p_0+q+1} - e) \wedge (e - \eta_{p_0+q})\} \leq C\epsilon_T \tag{6.2}$$

where $\wedge$ and $\vee$ denote the minimum and maximum operators, respectively. These inequalities will hold throughout the algorithm until no further change-points are detected.

We define symmetric intervals $\mathcal{I}_r^L$ and $\mathcal{I}_r^R$ around change-points such that for every triplet $\{\eta_{r-1}, \eta_r, \eta_{r+1}\}$

$$\mathcal{I}_r^L = \left[\eta_r - \frac{2}{3}\delta_{\min}^r, \eta_r - \frac{1}{3}\delta_{\min}^r (1+\bar{c})\right]$$

and

$$\mathcal{I}_r^R = \left[\eta_r + \frac{1}{3}\delta_{\min}^r (1+\bar{c}), \eta_r + \frac{2}{3}\delta_{\min}^r\right] \quad \text{for } r = 1, ..., N+1$$

where $\delta_{\min}^r = \min\{\eta_r - \eta_{r-1}, \eta_{r+1} - \eta_r\}$ and $\bar{c} = 3 - \frac{2}{c_\star}$ for $c_\star$ as in (1.5). We recall that at every stage of the WBS algorithm M intervals (s_m, e_m), $m = 1, ..., M$ are drawn from a discrete uniform distribution over the set $\{(s,e) : s < e, 0 \le s \le T-2, 1 \le e \le T-1\}$.

We define the event D_T^M as

$$D_T^M = \{\forall r = 1, ..., N \;\; \exists\, m = 1, ..., M \;\; (s_m, e_m) \in \mathcal{I}_r^L \times \mathcal{I}_r^R\}.$$

Also, note that

$$P((D_T^M)^c) \le \sum_{r=1}^{N} \prod_{m-1}^{M} (1 - P((s_m, e_m) \in \mathcal{I}_r^L \times \mathcal{I}_r^R)) \le \frac{T}{\delta_T}(1 - \delta_T^2(1-\bar{c})^2 T^{-2}/9)^M. \tag{6.3}$$

On a generic interval satisfying (6.1) and (6.2) we consider

$$(m_0, b) = \arg \max_{(m,t): m \in \mathcal{M}_{s,e}, s_m \le t \le e_m} |\tilde{Y}_{s_m, e_m}^t| \tag{6.4}$$

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

Lemma 1

$$P\left(\max_{(s,b,e)} \left|\mathbb{Y}_{s,e}^b - \mathbb{S}_{s,e}^b\right| > \lambda_1\right) \to 0 \tag{6.5}$$

for

$$\lambda_1 \ge \log T.$$

Proof: We start by studying the following event

$$\left|\sum_{t=s}^{e} c_t \sigma(t/T)^2 (Z_{t,T}^2 - 1)\right| > \sqrt{n}\lambda_1$$

where $c_t = \sqrt{(e-b)/(b-s+1)}$ and $c_t = \sqrt{(b-s+1)/(e-b)}$ for $t \leq b$ and $b+1 \leq t$ respectively. From (1.5), we have that $c_t \leq c_\star \equiv \sqrt{\frac{c_\star}{1-c_\star}} < \infty$. The proof proceeds as in Cho and Fryzlewicz (2015) and we have that (6.5) is bounded by

$$\sum_{(s,b,e)} 2\exp\left(-\frac{n\lambda_1^2}{4c_\star^2 \max_z \sigma^2(z) n\rho_\infty^2 + 2c_\star \max_z \sigma(z)\sqrt{n}\lambda_1\rho_\infty^1}\right) \leq 2T^3 \exp\left(-C_1'(c_{\star^{-2}})\log^2 T\right)$$

which converges to 0 since $n \geq \delta_T = \mathcal{O}(\log^2 T)$ and $\rho_\infty^1 < \infty$ from (A2).

Lemma 2 *Assuming that (6.1) holds, then there exists $C_2 > 0$ such that for b satisfying $|b - \eta_{p_0+r'}| = C_2\gamma_T$ for some r', we have $|\mathbb{S}_{s_{m_0},e_{m_0}}^{\eta_{p_0}+r'}| \geq |\mathbb{S}_{s_{m_0},e_{m_0}}^{b}| + C\gamma_T\delta_T^{-1/2} \geq |\mathbb{S}_{s_{m_0},e_{m_0}}^{b}| + 2\lambda_1$, where $\gamma_T = \sqrt{\delta_T}\lambda_1$.*

Proof: From the proof of Theorem 3.2 in Fryzlewicz (2014) and Lemma 1 in Cho and Fryzlewicz (2012) we have the following result

$$|\mathbb{S}_{s_{m_0},e_{m_0}}^{b}| \geq |\mathbb{Y}_{s_{m_0},e_{m_0}}^{b}| - \lambda_1 \geq C_3\sqrt{\delta_T} \tag{6.6}$$

provided that $\delta_T \geq C_4\lambda_1^2$.

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

$$|\mathbb{S}_{s_{m_0},e_{m_0}}^{\eta_{p_0}+r'}| > |\mathbb{S}_{s_{m_0},e_{m_0}}^{b}| \geq C_3\sqrt{\delta_T}.$$

Now, the following three cases are not possible:

1. (s_{m_0}, e_{m_0}) contains a single change-point, $\eta_{p_0+r'}$, and both $\eta_{p_0+r'} - s_{m_0}$ and $e_{m_0} - \eta_{p_0+r'}$ are not bounded from below by $c_1\delta_T$.
2. (s_{m_0}, e_{m_0}) contains a single change-point, $\eta_{p_0+r'}$, and either $\eta_{p_0+r'} - s_{m_0}$ or $e_{m_0} - \eta_{p_0+r'}$ are not bounded from below by $c_1\delta_T$.
3. (s_{m_0}, e_{m_0}) contains two change-points, $\eta_{p_0+r'}$ and $\eta_{p_0+r'+1}$, and both $\eta_{p_0+r'} - s_{m_0}$ and $e_{m_0} - \eta_{p_0+r'+1}$ are not bounded from below by $c_1\delta_T$.

The first case is not permitted by (A5). For the last two, if either case were true, then following the arguments as in Lemma A.5 of Fryzlewicz (2014), we would obtain that $\max_{t:s_{m_0}\leq t\leq e_{m_0}} |\mathbb{S}_{s_{m_0},e_{m_0}}^{t}|$ was not bounded from below by

$C_3\sqrt{\delta_T}$ which contradicted (6.6). Hence, interval (s_{m_0}, e_{m_0}) satisfies condition (6.1) and following a similar argument to the proof of Lemma 2 in Cho and Fryzlewicz (2012) we can show that for any b satisfying $|b - \eta_{p_0+r'}| = C_2\gamma_T$, we have $|\mathbb{S}^{\eta_{p_0}+r'}_{s_{m_0},e_{m_0}}| \geq |\mathbb{S}^{b}_{s_{m_0},e_{m_0}}| + C\gamma_T\delta_T^{-1/2}$.

Lemma 3 *Under conditions (6.1) and (6.2) there exists* $1 \leq r' \leq q$ *such that* $|b - \eta_{p_0+r'}| \leq \epsilon_T$, *where* b *is given in (6.4) and* $\epsilon_T = C\log^2 T$ *for a positive constant* C.

Proof: First, we mention that the model (1.1) can be written as $\tilde{Y}_{t,T} = \sigma(t/T)^2 + \sigma(t/T)^2(Z^2_{t,T} - 1)$ which has the form of a signal+noise model i.e. $Y_t = f_t + \varepsilon_t$. Now, let $\bar{f}^d_{s_{m_0},e_{m_0}}$ define the best function approximation to f_t such that $\arg\max_d |\langle \psi^d_{s_{m_0},e_{m_0}}, f\rangle| = \arg\min_d \sum_{t=s_{m_0}}^{e_{m_0}} (f_t - \bar{f}^d_{s_{m_0},e_{m_0}})$ where $\bar{f}^d_{s_{m_0},e_{m_0}} = \bar{f} + \langle f, \psi^d_{s_{m_0},e_{m_0}}\rangle \psi^d_{s_{m_0},e_{m_0}}$, $\bar{f}$ is the mean of f and $\psi^d_{s_{m_0},e_{m_0}}$ is a set of vectors that are constant and positive until d and then constant and negative from $d+1$ until e_{m_0}.

If it can be shown that for a certain $\epsilon_T < C_2\gamma_T$, we have

$$\sum_{t=s_{m_0}}^{e_{m_0}} (Y_t - \bar{Y}^d_{s_{m_0},e_{m_0},t})^2 > \sum_{t=s_{m_0}}^{e_{m_0}} (Y_t - \bar{f}^{\eta_{p_0+r'}}_{s_{m_0},e_{m_0},t})^2 \tag{6.7}$$

as long as

$$\epsilon_T \leq |d - \eta_{p_0+r'}|$$

then this would prove necessarily that $|b - \eta_{p_0+r'}| \leq \epsilon_T$.

By Lemma 2 and Lemma A.3 in Fryzlewicz (2014), we have the same triplet of inequalities as in the argument in the proof of Theorem 3.2 in Fryzlewicz (2014) i.e.

$$|d - \eta_{p_0+r'}| \geq C(\lambda_2|d - \eta_{p_0+r'}|\delta_T^{-1/2}) \vee (\lambda_2|d - \eta_{p_0+r'}|^{-1/2}) \vee (\lambda_2^2). \tag{6.8}$$

Hence, with the requirement that $|d - \eta_{p_0+r'}| \leq C_2\gamma_T = C_2\lambda_1\sqrt{\delta_T}$ we obtain

$$\delta_T > C^2\lambda_2^2 \max(C^2C_2^{-2}\lambda_1^{-2}\lambda_2^2, 1)$$

and $\epsilon_T = \max(1, C^2)\lambda_2^2$. From Lemma 1 λ_1 is of order $\mathcal{O}(\log T)$. For λ_2, which appears in the following two terms of the decomposition of (6.7)

$$I = \frac{1}{d - s_{m_0} + 1}\left(\sum_{t=s_{m_0}}^{d} \varepsilon_t\right)^2 \quad \text{and} \quad II = \frac{1}{e_{m_0} - d + 1}\left(\sum_{t=d+1}^{e_{m_0}} \varepsilon_t\right)^2$$

we show below that with probability tending to 1, $I \leq \lambda_2^2 = \log^2 T$. From Lemma 1 we have that $c_t = 1$ for $t = s_{m_0}, ..., d$ and thus

$$P\left(\frac{1}{\sqrt{d - s_{m_0} + 1}}\left|\sum_{t=s_{m_0}}^{d} \varepsilon_t\right| > \lambda_2\right) \to 0$$

since by the Bernstein inequality the probability is bounded by

$$2T^2 \exp\left(-\frac{(d - s_{m_0} + 1)\lambda_2^2}{4\max_z \sigma^2(z)(d - s_{m_0} + 1)\rho_\infty^2 + 2c' \max_z \sigma(z)\sqrt{d - s_{m_0} + 1}\lambda_2 \rho_\infty^1}\right) \leq 2T^2 \exp\left(-C_3'\lambda_2^2\right)$$

which converges to 0 due to $(d - s_{m_0} + 1) = \mathcal{O}(\delta_T)$ from (1.5). Note that II has similar order and we omit the details. This concludes the lemma.

Lemma 4 *Under conditions (6.1) and (6.2)*

$$P\left(|\mathbb{Y}^b_{s_{m_0}, e_{m_0}}| > \omega_T \frac{\sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_t}{n_{m_0}}\right) \to 1$$

where b is given in (6.4).

Proof: We define the following two events $\mathcal{A} = \left\{|\mathbb{Y}^b_{s_{m_0}, e_{m_0}}| < \omega_T \frac{1}{n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T}\right\}$ and $\mathcal{B} = \left\{\frac{1}{n_{m_0}}\left|\sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} - \sum_{t=s_{m_0}}^{e_{m_0}} \sigma(t/T)^2\right| < \bar{\sigma} = \frac{1}{2n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \sigma^2(t/T)\right\}$.

Since $P(\mathcal{A}) \leq P(\mathcal{A} \cap \mathcal{B}) + P(\mathcal{B}^c)$ we need to show that $P(\mathcal{B}) \to 1$ and $P(\mathcal{A} \cap \mathcal{B}) \to 0$. To show that $P(\mathcal{B}) = P\left(\frac{1}{n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} \in (\bar{\sigma}/2, 3\bar{\sigma}/2)\right) \to 1$ we apply the Bernstein inequality as in Lemma 1 and we have that

$$P(\mathcal{B}') = P\left(\frac{1}{n_{m_0}}\left|\sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} - \sum_{t=s_{m_0}}^{e_{m_0}} \sigma(t/T)^2\right| > \bar{\sigma}\right) = P\left(\left|\sum_{t=s_{m_0}}^{e_{m_0}} \sigma(t/T)^2 (Z_{t,T}^2 - 1)\right| > n_{m_0}\bar{\sigma}\right).$$

Hence,

$$P(\mathcal{B}') \leq 2\exp\left(-\frac{n_{m_0}^2 \bar{\sigma}^2}{4\max_z \sigma^2(z) n_{m_0} \rho_\infty^2 + 2c' \max_z \sigma(z) n_{m_0} \bar{\sigma} \rho_\infty^1}\right) \leq 2T^2 \exp\left(-C_4' \log^2 T\right)$$

which converges to 0 since $n_{m_0} \geq \delta_T = \mathcal{O}(\log^2 T)$ and $\rho_\infty^1 < \infty$ from (A2).

Now, from Lemma (3), we have some $\eta \equiv \eta_{p_0+r'}$ satisfying $|b - \eta| \leq C\epsilon_T$.

Turning to $P(\mathcal{A} \cap \mathcal{B})$ we have from conditions (6.1) and (6.2)

$$\begin{aligned}
|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| \geq |\mathbb{Y}^\eta_{s_{m_0},e_{m_0}}| &\geq |\mathbb{S}^\eta_{s_{m_0},e_{m_0}}| - \log T \\
&= \left| \sqrt{\frac{(\eta - s_{m_0} + 1)(e_{m_0} - \eta)}{n_{m_0}}} \left(\sigma \left(\frac{\eta}{T} \right)^2 - \sigma \left(\frac{\eta + 1}{T} \right)^2 \right) \right| - \log T \\
&= \sqrt{\frac{e_{m_0} - \eta}{n_{m_0}(\eta - s_{m_0} + 1)}} (\eta - s_{m_0} + 1)\sigma_\star - \log T \geq C\sqrt{\delta_T} - \log T > \omega_T 3\bar{\sigma}/2,
\end{aligned}$$

which concludes the Lemma.

Lemma 5 *For some positive constants C, C', let s, e satisfy either*

- *$\exists 1 \leq p \leq N$ such that $s \leq \eta_p \leq e$ and $(\eta_p - s + 1) \wedge (e - \eta_p) \leq C\epsilon_T$ or*
- *$\exists 1 \leq p \leq N$ such that $s \leq \eta_{p+1} \leq e$ and $(\eta_p - s + 1) \vee (e - \eta_{p+1}) \leq C'\epsilon_T$.*

Then,

$$P\left(|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| < \omega_T \frac{\sum_{t=s_{m_0}}^{e_{m_0}} Y_t}{n_{m_0}} \right) \to 1$$

where b is given in (6.4).

Proof: A similar argument to the proof of Lemma 5 is applied here. We only need to show that $P(\mathcal{A} \cap \mathcal{B}) \to 0$ where now event $\mathcal{A} = \left\{ |\mathbb{Y}_{s_{m_0},b,e_{m_0}}| > \omega_T \frac{1}{n_{m_0}} \sum_{t=s_{m_0}}^{e_{m_0}} \tilde{Y}_{t,T} \right\}$. Using condition (i) or (ii) we have that

$$\begin{aligned}
|\mathbb{Y}^b_{s_{m_0},e_{m_0}}| &\leq |\mathbb{S}^b_{s_{m_0},e_{m_0}}| + \log T \\
&= \left| \frac{\sqrt{b - s_{m_0} + 1}\sqrt{e_{m_0} - b}}{\sqrt{n_{m_0}}} \left(\sigma^2(b/T) - \sigma^2((b+1)/T) \right) \right| + \log T \\
&\leq \sigma^* C\sqrt{\epsilon_T} + \log T < \omega_T \bar{\sigma}/2.
\end{aligned}$$

The proof of Theorem 1 proceeds as follows: at the start of the algorithm when $s = 0$ and $e = T - 1$ all the conditions of (6.1) & (6.2) required by Lemma 4 are met and thus it detects a change-point on that interval defined by formula (6.4) within the distance of $C\epsilon_T$ (by Lemma 3). The conditions of Lemma 4 are satisfied until all change-points have been identified. Then, every random interval (s_m, e_m) does not contain a change-point or the conditions of Lemma 5 are met; hence no more change-points are detected and the algorithm stops.

Finally, we examine whether the bias present in $\mathbb{E}I^{(i)}_{t,T}$ (see condition (A0)) will affect the above result. We define $\tilde{\mathbb{S}}^t_{s,e}$ similarly to $\mathbb{S}^t_{s,e}$ by replacing $\sigma(t/T)^2$ with $\sigma^2_{t,T}$. Assume that η_r is a change-point within the interval $[s_{m_0}, e_{m_0}]$ and $b = \arg\max_{t\in(s_{m_0},e_{m_0})} |\mathbb{S}^b_{s_{m_0},e_{m_0}}|$ and $\hat{b} = \arg\max_{t\in(s_{m_0},e_{m_0})} |\tilde{\mathbb{S}}^b_{s_{m_0},e_{m_0}}|$. Recall that $\mathbb{E}I^{(i)}_{t,T}$ is constant within each segment apart from short intervals around true change-point η_r i.e. $[\eta_r - K2^{-i}, \eta_r + K2^{-i}]$. In addition, from Theorem 2 in Cho and Fryzlewicz (2015) the finest scale should satisfy $i \geq I^\star = -\lfloor \alpha \log\log T \rfloor$ in order for (A4) to hold. Then, $|\hat{b}-b| \leq K2^{I^\star} < \epsilon_T$ holds since $I^\star = \mathcal{O}(\log\log T)$. Therefore, bias does not affect the above result and the consistency is preserved.

Proof of Theorem 2

We start by the first method of aggregation. From the invertibility of the autocorrelation wavelet inner product matrix A, there exists at least one ordinate of wavelet periodogram in which a change-point θ_r is detected. From Theorem 1 it holds that $|\theta_r - \hat{\theta}_r| \leq C\epsilon_T$ with probability converging to 1 regardless of the scale i. Since the algorithm begins its search from the finest scale and only proceeds to the next one if no change-point is detected (until scale $I^\star$) then consistency is preserved.

We now turn to the second method of aggregation. We note that $\mathbb{Y}^{thr}_t$ has the same functional form with each of $\mathcal{Y}^{(i)}_t$ i.e. $h^{(i)}(x) = (x(1-x))^{-1/2}(c^{(i)}_x x + d^{(i)}_x x)$ for $x = (t - s_m + 1)/n \in (0,1)$, where $c^{(i)}_x$, $d^{(i)}_x$ are determined by the location and the magnitude of the change-points of $I^{(i)}_{t,T}$. Let $b = \arg\max_{s_{m_0}<t<e_{m_0}} \mathbb{Y}^{thr}_t$; then following a similar argument to Lemma 2 of Fryzlewicz (2014) we can show that $\mathbb{Y}^{thr}_t$ must have a local maximum at $t = \theta_{p_0+r'}$ and that $|b - \theta_{p_0+r'}| \leq C_5\gamma_T$. With this result, we can show that $|b - \theta_{p_0+r}| \leq C'\epsilon_T$ for some $1 \leq r' \leq q$ as in Lemma 3 above by constructing a signal+noise model $y_t = f_t + \varepsilon_t$ and substituting f_t with $\sum_{i=-I^\star}^{-1} \mathbb{E}I^{(i)}_{t,T}\mathbb{I}(\mathcal{Y}^{(i)}_t > \omega^{(i)}_T)/q^{(i)}_{s_m,e_m}$. Then, conditions (6.1) and (6.2) are satisfied within each segment for at least one scale $i \in \{-1, ..., -I^\star\}$. When all change-points have been detected every subsequent random interval (s_m, e_m) will satisfy the conditions of Lemma 5 for every $i \in \{-1, ..., -I^\star\}$ and the algorithm stops.

References

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

R. Bellman and S. Dreyfus. *Applied Dynamic Programming*, volume 2. Princeton University Press, 1966.

I. Berkes, E. Gombay, and L. Horváth. Testing for changes in the covariance structure of linear processes. *Journal of Statistical Planning and Inference*, 139:2044–2063, 2009.

H. Cho and P. Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22:207–229, 2012.

H. Cho and P. Fryzlewicz. Multiple change-point detection for high-dimensional time series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society Series B*, 77: 475–507, 2015.

T. Clark. Is the great moderation over? An empirical analysis. *Federal Reserve Bank of Kansas City Economic Review*, 94:5–42, 2009.

S. Davies and D. Bland. Interestingness detection in sports audio broadcasts. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 643–648. IEEE, 2010.

R. Davis, D. Huang, and Y. Yao. Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, 23:282–304, 1995.

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

R. Davis, T. Lee, and G. Rodriguez-Yam. Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29:834–867, 2008.

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

P. Fryzlewicz and G. Nason. Haar–Fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society: Series B*, 68:611–634, 2006.

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

E. Gombay. Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99:451–464, 2008.

E. Gombay and D. Serban. Monitoring parameter change in time series models. *Journal of Multivariate Analysis*, 100:715–725, 2009.

J. Groen, G. Kapetanios, and S. Price. Multivariate methods for monitoring structural change. *Journal of Applied Econometrics*, 28:250–274, 2013.

L. Horváth, Z. Horváth, and M. Husková. Ratio tests for change point detection. *Inst. Math. Stat. Collections*, 1:293–304, 2008.

C. Inclan and G. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89:913–923, 1994.

B. Jackson, J. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. Tsai. An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12:105–108, 2005.

S. Kay. *Fundamentals of Statistical Signal Srocessing, Volume 2: Detection Theory*. Prentice Hall PTR, 1998.

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

R. Killick, I. Eckley, and P. Jonathan. A wavelet-based approach for detecting changes in second order structure within nonstationary time series. *Electronic Journal of Statistics*, 7: 1167–1183, 2013.

S. Kim, S. Cho, and S. Lee. On the cusum test for parameter changes in GARCH$(1, 1)$ models. *Communications in Statistics-Theory and Methods*, 29:445–462, 2000.

M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21:33–59, 2000.

M. Lavielle and G. Teyssiere. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.

M. Lavielle and G. Teyssiere. Adaptive detection of multiple change-points in asset price volatility. *Long Memory in Economics*, pages 129–156, 2007.

S. Lee and S. Park. The cusum of squares test for scale changes in infinite order moving average processes. *Scandinavian Journal of Statistics*, 28:625–644, 2001.

S. Lee, O. Na, and S. Na. On the cusum of squares test for variance change in nonstationary and nonparametric time series models. *Annals of the Institute of Statistical Mathematics*, 55: 467–485, 2003.

D. Mercurio and V. Spokoiny. Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics*, 32:577–602, 2004.

G. Nason. A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal of the Royal Statistical Society: Series B*, 75:879–904, 2013.

G. Nason, R. Von Sachs, and G. Kroisandt. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B*, 62:271–292, 2000.

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

H. Ombao, J. Raz, R. von Sachs, and B. Malow. Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96:543–560, 2001.

P. Perron. Dealing with structural breaks. *Palgrave Handbook of Econometrics*, 1:278–352, 2006.

A. Schröder and P. Fryzlewicz. Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and Its Interface*, 6:449–461, 2013.

A. Sen and M. Srivastava. On tests for detecting change in mean. *The Annals of Statistics*, 3: 98–108, 1975.

R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications: with R Examples*. Springer, 2011.

C. Stărică and C. Granger. Nonstationarities in stock returns. *Review of economics and statistics*, 87:503–522, 2005.

E. Venkatraman. *Consistency results in multiple change-point problems*. PhD thesis, Department of Statistics. Stanford University, 1992.

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

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

Y. Yao and S. Au. Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, 51:370–381, 1989.

Department of Statistics, London School of Economics, Columbia House, LSE, Houghton Street, London, WC2A 2AE, UK

E-mail: k.k.korkas@lse.ac.uk

Department of Statistics, London School of Economics, Columbia House, LSE, Houghton Street, London, WC2A 2AE, UK

E-mail: p.fryzlewicz@lse.ac.uk