Multiple-change-point detection for high dimensional time series via sparsified binary segmentation

Author(s): Haeran Cho and Piotr Fryzlewicz

Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, MARCH 2015, Vol. 77, No. 2 (MARCH 2015), pp. 475-507

Published by: Wiley for the Royal Statistical Society

Stable URL: https://www.jstor.org/stable/24774746

J. R. Statist. Soc. B (2015)
77, *Part* 2, *pp.* 475–507

Multiple-change-point detection for high dimensional time series via sparsified binary segmentation

Haeran Cho

University of Bristol, UK

and Piotr Fryzlewicz

London School of Economics and Political Science, UK

[Received January 2013. Final revision May 2014]

Summary. Time series segmentation, which is also known as multiple-change-point detection, is a well-established problem. However, few solutions have been designed specifically for high dimensional situations. Our interest is in segmenting the second-order structure of a high dimensional time series. In a generic step of a binary segmentation algorithm for multivariate time series, one natural solution is to combine cumulative sum statistics obtained from local periodograms and cross-periodograms of the components of the input time series. However, the standard 'maximum' and 'average' methods for doing so often fail in high dimensions when, for example, the change points are sparse across the panel or the cumulative sum statistics are spuriously large. We propose the sparsified binary segmentation algorithm which aggregates the cumulative sum statistics by adding only those that pass a certain threshold. This 'sparsifying' step reduces the influence of irrelevant noisy contributions, which is particularly beneficial in high dimensions. To show the consistency of sparsified binary segmentation, we introduce the multivariate locally stationary wavelet model for time series, which is a separate contribution of this work.

Keywords: Binary segmentation; Cumulative sum statistic; High dimensional time series; Locally stationary wavelet model; Multiple-change-point detection; Thresholding

1. Introduction

Detecting multiple change points in univariate time series has been widely discussed in various contexts; see Inclán and Tiao (1994), Chen and Gupta (1997), Lavielle and Moulines (2000), Ombao *et al.* (2001) and Davis *et al.* (2006, 2008) for some recent approaches. In this paper, we use the term 'multiple-change-point detection' interchangeably with 'segmentation'. By contrast, segmentation of the second-order structure of multivariate time series, especially those of high dimensionality, is yet to receive much attention although multivariate time series observed in practical problems often appear second-order non-stationary. For example, in financial time series, large panels of asset returns routinely display such non-stationarities (see for example Fan *et al.* (2011) for a comprehensive review of challenges of high dimensionality in finance and economics). Another example can be found in neuroscience, where electroencephalograms

Address for correspondence: Haeran Cho, School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK.
E-mail: haeran.cho@bristol.ac.uk

 1369–7412/15/77475

that are recorded at multiple channels exhibit non-stationarity and high correlations as well as being massive in volume (Ombao *et al.*, 2005). Vert and Bleakley (2010) have described other interesting examples of multivariate, non-stationary time series in many other fields, such as signal processing, biology and medicine.

As arguably one of the simplest forms of departure from stationarity, we consider a class of piecewise stationary, multivariate (possibly high dimensional) time series with a time varying second-order structure, where the autocovariance and cross-covariance functions are asymptotically piecewise constant and hence the time series is approximately stationary between change points in these functions.

We first list some existing approaches to the problem of multiple-change-point detection in multivariate (not necessarily high dimensional) time series. Ombao *et al.* (2005) employed the 'smooth localized complex exponentials' basis for time series segmentation that was originally proposed by Ombao *et al.* (2002). The choice of this basis leads to the segmentation of the time series, achieved via complexity-penalized optimization. Lavielle and Teyssière (2006) introduced a procedure based on the penalized Gaussian log-likelihood as a cost function, where the estimator was computed via dynamic programming. The performance of the method was tested on bivariate examples. Vert and Bleakley (2010) proposed a method for approximating multiple signals, with independent noise, via piecewise constant functions, where the change point detection problem was reformulated as a penalized regression problem and solved by the group lasso (Yuan and Lin, 2006). Note that, in Cho and Fryzlewicz (2011), we argued that the l_1-penalty was suboptimal for change point detection.

Cumulative sum (CUSUM) type statistics have been widely used in time series segmentation. In the context of multivariate time series segmentation, Groen *et al.* (2013) studied the average and the maximum of d CUSUM statistics, each obtained from one component of a d-dimensional time series, and compared their theoretical properties as well as finite sample performance. The average test statistic was also adopted in Horváth and Hušková (2012) for detecting a single change in the mean of a panel data model, and both allowed the dimensionality to increase under the constraint $d^2/T \to 0$, where T denoted the sample size. In Aue *et al.* (2009), a CUSUM statistic was proposed for detecting and locating a single change point in the covariance structure of multivariate time series, where its extension to the detection of multiple change points via binary segmentation was discussed heuristically.

In this paper, we propose a CUSUM-based binary segmentation algorithm, termed 'sparsified binary segmentation' (SBS), for identifying multiple change points in the second-order structure of a multivariate (possibly high dimensional) time series. The input to the SBS algorithm is $\{Y_{t,T}^{(k)}, k = 1, \ldots, d\}$, which is a d-dimensional sequence of localized periodograms and cross-periodograms computed on the original multivariate time series, where the dimensionality d is allowed to diverge with the number of observations T at a certain rate.

A key ingredient of the SBS algorithm is a 'sparsifying' step, where, instead of blindly aggregating all the information about the change points from the d sequences $Y_{t,T}^{(k)}$, we apply a threshold to the individual CUSUM statistics computed on each $Y_{t,T}^{(k)}$, and only those temporal fragments of the CUSUMs that survive after the thresholding are aggregated to have any contribution in detecting and locating the change points. In this manner, we reduce the influence of those sequences that do not contain any change points so that the procedure is less affected by them, which can be particularly beneficial in a high dimensional context. Therefore, we can expect improved performance in comparison with methods without a similar dimension reduction step, and this point is explained in more detail in Section 2.1. Further, owing to the aggregation of the CUSUM statistics, the algorithm automatically identifies common change points, rather than estimating single change points at different locations in different compo-

nents of the time series, which removes the need for post-processing across the d-dimensional sequence. This characteristic is particularly attractive in a high dimensional situation.

As well as formulating the complete SBS algorithm, we show its consistency for the number and the locations of the change points. One theoretical contribution of this work is that our rates of convergence of the location estimators improve on those previously obtained for binary segmentation for univariate time series (Cho and Fryzlewicz, 2012) and are near optimal in the case of the change points being separated by time intervals of length $\asymp T$, where $a_T \asymp b_T$ if $a_T^{-1} b_T \to C$ as $T \to \infty$ for some constant C. This was achieved by adapting, to the high dimensional time series context, the proof techniques from Fryzlewicz (2014) for the univariate signal plus independent, identically distributed (IID) Gaussian noise model. As a theoretical setting for deriving the consistency results, we introduce the multivariate locally stationary wavelet (LSW) model for time series. This, we believe, is a separate contribution of the current work and provides a multivariate extension of the univariate LSW model of Nason *et al.* (2000) and of the bivariate LSW model of Sanderson *et al.* (2010).

The rest of the paper is organized as follows. In Section 2, we introduce the SBS algorithm for segmenting a possibly large number of multiplicative sequences. In Section 3, we introduce a class of piecewise stationary, multivariate time series and discuss the specifics of applying the SBS from Section 2 to detect change points in its second-order structure (the version of the SBS algorithm that is specifically applicable to multivariate time series is labelled SBS-MVTS in the paper). Section 4 illustrates the performance of the proposed methodology on a set of simulated examples, and Section 5 applies it to the multivariate series of Standard and Poors 500 components, observed daily between 2007 and 2011. The proofs are in Appendix A.

2. The sparsified binary segmentation algorithm in a generic setting

In this section, we outline the SBS algorithm for change point detection in a panel of multiplicative sequences, which may share common change points in their expectations. We later consider a piecewise stationary, multivariate time series model and use it to derive a set of statistics, which contain information about the change points in its second-order structure. Those statistics are shown to follow the multiplicative model considered so that SBS can be applied to them. This will enable us to segment the original time series by using the SBS methodology.

The multiplicative model in question is

$$Y_{t,T}^{(k)} = \sigma^{(k)}(t/T) Z_{t,T}^{(k)2}, \qquad t = 0, \ldots, T-1, \quad k = 1, \ldots, d, \tag{1}$$

where $Z_{t,T}^{(k)}$ is a sequence of (possibly) auto-correlated and non-stationary standard normal variables such that $\mathbb{E}[Y_{t,T}^{(k)}] = \sigma^{(k)}(t/T)$, which implies that each $Y_{t,T}^{(k)}$ is a scaled χ_1^2-variable. Extensions to some other distributions are possible but technically involved and we do not pursue them here. Each $\sigma^{(k)}(t/T)$ is a piecewise constant function, and we aim to detect any change points in $\sigma^{(k)}(t/T)$ for $k = 1, \ldots, d$. It is assumed that there are N change points $0 < \eta_1 < \eta_2 < \ldots < \eta_N < T-1$ possibly shared by the d functions $\sigma^{(k)}(t/T)$, in the sense that, for each η_q, there is one or more $\sigma^{(k)}(t/T)$ satisfying $\sigma^{(k)}(\eta_q/T) \neq \sigma^{(k)}\{(\eta_q+1)/T\}$. We impose the following conditions on η_q, $q = 1, \ldots, N$.

Assumption 1.

(a) The distance between any two adjacent change points is bounded from below by $\delta_T \asymp T^{\Theta}$ for $\Theta \in (\frac{3}{4}, 1]$.

(b) The spacings between any three consecutive change points are not too 'unbalanced' in the sense that they satisfy

$$\max\left(\frac{\eta_q-\eta_{q-1}+1}{\eta_{q+1}-\eta_{q-1}+1},\frac{\eta_{q+1}-\eta_q}{\eta_{q+1}-\eta_{q-1}+1}\right)\leqslant c_*, \tag{2}$$

where c_* is a constant satisfying $c_*\in[\frac{1}{2},1)$.

Condition 1(a) determines the upper bound on the total number of change points, which is allowed to diverge with T as long as $\Theta<1$, and is unknown by the user. Cho and Fryzlewicz (2012) proposed a change point detection method for a *single* sequence $Y_{t,T}$ following model (1). The main ingredient of the method proposed in that work was a binary segmentation algorithm which simultaneously located and tested for change points in a recursive manner. Below we provide a sketch of that algorithm, which is referred to as univariate binary segmentation (UBS) throughout the present paper.

Firstly, the likely position of a change point in the interval $[0,T-1]$ is located as the point where the following CUSUM-type statistic is maximized over t:

$$\mathcal{Y}_{0,t,T-1}=\mathcal{Y}_{0,t,T-1}(Y_{u,T})=\left(\frac{1}{T}\sum_{u=0}^{T-1}Y_{u,T}\right)^{-1}\left|\sqrt{\left(\frac{T-t}{Tt}\right)}\sum_{u=0}^{t-1}Y_{u,T}-\sqrt{\left\{\frac{t}{T(T-t)}\right\}}\sum_{u=t}^{T-1}Y_{u,T}\right|. \tag{3}$$

A discussion of the properties of $\mathcal{Y}_{0,t,T-1}$ can be found in Cho and Fryzlewicz (2012); we only remark here that the first term of the product in equation (3) is a normalizing term that is essential in multiplicative settings, which makes our results independent of the level of $\sigma^{(k)}(t/T)$ in model (1). Next, for $b=\arg\max_t \mathcal{Y}_{0,t,T-1}$, if $\mathcal{Y}_{0,b,T-1}<\pi_T$ with a suitably chosen threshold π_T, then we stop; otherwise we add b to the set of estimated change points and continue recursively in the same manner to the left and to the right of b. Details of the UBS algorithm and the theoretical result on its consistency for the number and the locations of the change points can be found in Cho and Fryzlewicz (2012).

2.1. Binary segmentation for high dimensional data

In this section, we extend the UBS algorithm to an algorithm which is applicable to a panel of multiplicative sequences (1) even if its dimensionality d diverges as $T\to\infty$. The resulting SBS algorithm contains a crucial sparsifying step as detailed below.

We firstly note that, in the multivariate case $d>1$, we could proceed by applying the UBS algorithm to each sequence $Y_{t,T}^{(k)}$ separately, and then by pruning the estimated change points by identifying those corresponding to each true change point. However, it is conceivable that such pruning may not be straightforward, particularly in high dimensions. We propose to circumvent this difficulty by segmenting the d sequences $Y_{t,T}^{(k)}$ at the same time by examining the CUSUM statistics $\mathcal{Y}_{0,t,T-1}(Y_{u,T}^{(k)})\equiv\mathcal{Y}_{0,t,T-1}^{(k)}$ in equation (3) simultaneously over k, rather than separately for each k.

Various ways of aggregating information from multiple CUSUM statistics have been proposed in the literature. Groen *et al.* (2013) discussed two popular methods: the pointwise average and the pointwise maximum. Specifically, using our notation, they are respectively defined as

$$\begin{aligned}\tilde{y}_t^{\mathrm{avg}}&=\frac{1}{d}\sum_{k=1}^{d}\mathcal{Y}_{0,t,T-1}^{(k)},\\ \tilde{y}_t^{\max}&=\max_{1\leqslant k\leqslant d}\mathcal{Y}_{0,t,T-1}^{(k)}.\end{aligned} \tag{4}$$

To determine whether $b=\arg\max\tilde{y}_t^{\mathrm{avg}}$ or $b=\arg\max\tilde{y}_t^{\max}$ is regarded as an estimated change

point, $\tilde{y}_b^{\text{avg}}$ or $\tilde{y}_b^{\max}$ respectively needs to be compared against a threshold which takes into account the aggregation step.

In the SBS algorithm, we propose another way of simultaneously considering multiple CUSUM statistics, which integrates a thresholding step that enables us to bypass some difficulties in dealing with high dimensional data which we describe later. For each k, the CUSUM statistic $\mathcal{Y}^{(k)}_{0,t,T-1}$ is compared with a threshold, say π_T (to be specified later in Section 3), and only the contributions from the time intervals where $\mathcal{Y}^{(k)}_{0,t,T-1} > \pi_T$ are taken into account in detecting and locating a change point. Thus $\tilde{y}_t^{\text{thr}}$, the main statistic of interest in the SBS algorithm, is defined as

$$\tilde{y}_t^{\text{thr}} = \sum_{k=1}^{d} \mathcal{Y}^{(k)}_{0,t,T-1} \mathbb{I}(\mathcal{Y}^{(k)}_{0,t,T-1} > \pi_T), \tag{5}$$

where $\mathbb{I}(\cdot)$ is an indicator function returning $\mathbb{I}(\mathcal{A}) = 1$ if the event $\mathcal{A}$ is true and $\mathbb{I}(\mathcal{A}) = 0$ otherwise. In this manner, $\tilde{y}_t^{\text{thr}}$ is non-zero only when at least one of $\mathcal{Y}^{(k)}_{0,t,T-1}$ is greater than the threshold, i.e. a change point is detected in $Y^{(k)}_{t,T}$ for such k. Therefore we can conclude that a change point is detected in the d-dimensional multiplicative sequences and, without applying any pruning, its location is estimated as $b = \arg\max_t \tilde{y}_t^{\text{thr}}$.

Although the empirical study that was conducted in Groen *et al.* (2013) shows the effectiveness of both $\tilde{y}_t^{\text{avg}}$ and $\tilde{y}_t^{\max}$ in detecting a single change point, there are high dimensional scenarios where these two estimators fail. Below we provide examples of high dimensional situations where $\tilde{y}_t^{\text{thr}}$ exhibits better performance than the other two.

2.1.1. *Sparse change points*

We first independently generate two time series $X_t^{(k)}$, $k = 1, 2$, as

$$X^{(1)}_{t,T} = aX^{(1)}_{t-1,T} + \epsilon^{(1)}_{t,T},$$

$$X^{(2)}_{t,T} = \begin{cases} 0.95X^{(2)}_{t-1,T} + \epsilon^{(2)}_{t,T} & \text{for } 1 \leqslant t \leqslant \lfloor T/2 \rfloor, \\ 0.3X^{(2)}_{t-1,T} + \epsilon^{(2)}_{t,T} & \text{for } \lfloor T/2 \rfloor + 1 \leqslant t \leqslant T, \end{cases}$$

with $T = 1024$. The parameter a is randomly generated from a uniform distribution $\mathcal{U}(0.5, 0.99)$ and $\epsilon^{(k)}_{t,T}$ are IID standard normal variables for $k = 1, 2$. We further produce the sequences $Y^{(1)}_{t,T}$ and $Y^{(2)}_{t,T}$ as $Y^{(k)}_{t,T} = 2^{-1}(X^{(k)}_{t,T} - X^{(k)}_{t-1,T})^2, k = 1, 2$, such that $Y^{(1)}_{t,T}$ does not have any change in $\mathbb{E}[Y^{(1)}_{t,T}]$, whereas $\mathbb{E}[Y^{(2)}_{t,T}]$ has one change point at $t = \lfloor T/2 \rfloor$. The rationale behind the choice of $Y^{(k)}_{t,T}$ as well as its relationship to the multiplicative model (1) are discussed in detail in Section 3. As can be seen from Figs 1 (a)–1(c), all three of the corresponding statistics $\tilde{y}_t^{\text{avg}}$, $\tilde{y}_t^{\max}$ and $\tilde{y}_t^{\text{thr}}$ can correctly identify the location of the true change point.

Now, consider the case with $d = 100$ time series where the additional time series $X^{(k)}_{t,T}$, $k = 3, \ldots, d$, are independently generated as $X^{(1)}_{t,T}$ such that, overall, there is only one change point coming from $X^{(2)}_{t,T}$ in the entire panel. Then, in obtaining the pointwise average of the d CUSUM statistics in $\tilde{y}_t^{\text{avg}}$, the $\mathcal{Y}^{(k)}_{0,t,T-1}$ for $k \neq 2$ corrupt the peak that is achieved around $t = \lfloor T/2 \rfloor$ for $\mathcal{Y}^{(2)}_{0,t,T-1}$, and hence the maximum of $\tilde{y}_t^{\text{avg}}$ is attained far from the true change point. In constrast, both $\tilde{y}_t^{\text{thr}}$ and $\tilde{y}_t^{\max}$ are successful in maintaining the peak achieved by $\mathcal{Y}^{(2)}_{0,t,T-1}$ by disregarding most or all of the $\mathcal{Y}^{(k)}_{0,t,T-1}$, $k \neq 2$.

2.1.2. *Spuriously large cumulative sum statistics*

Again, we first independently generate $d = 2$ time series $X^{(k)}_{t,T}$, $k = 1, 2$, with $T = 1024$, where $X^{(1)}_{t,T}$ is identical to that in Section 2.1.1 and

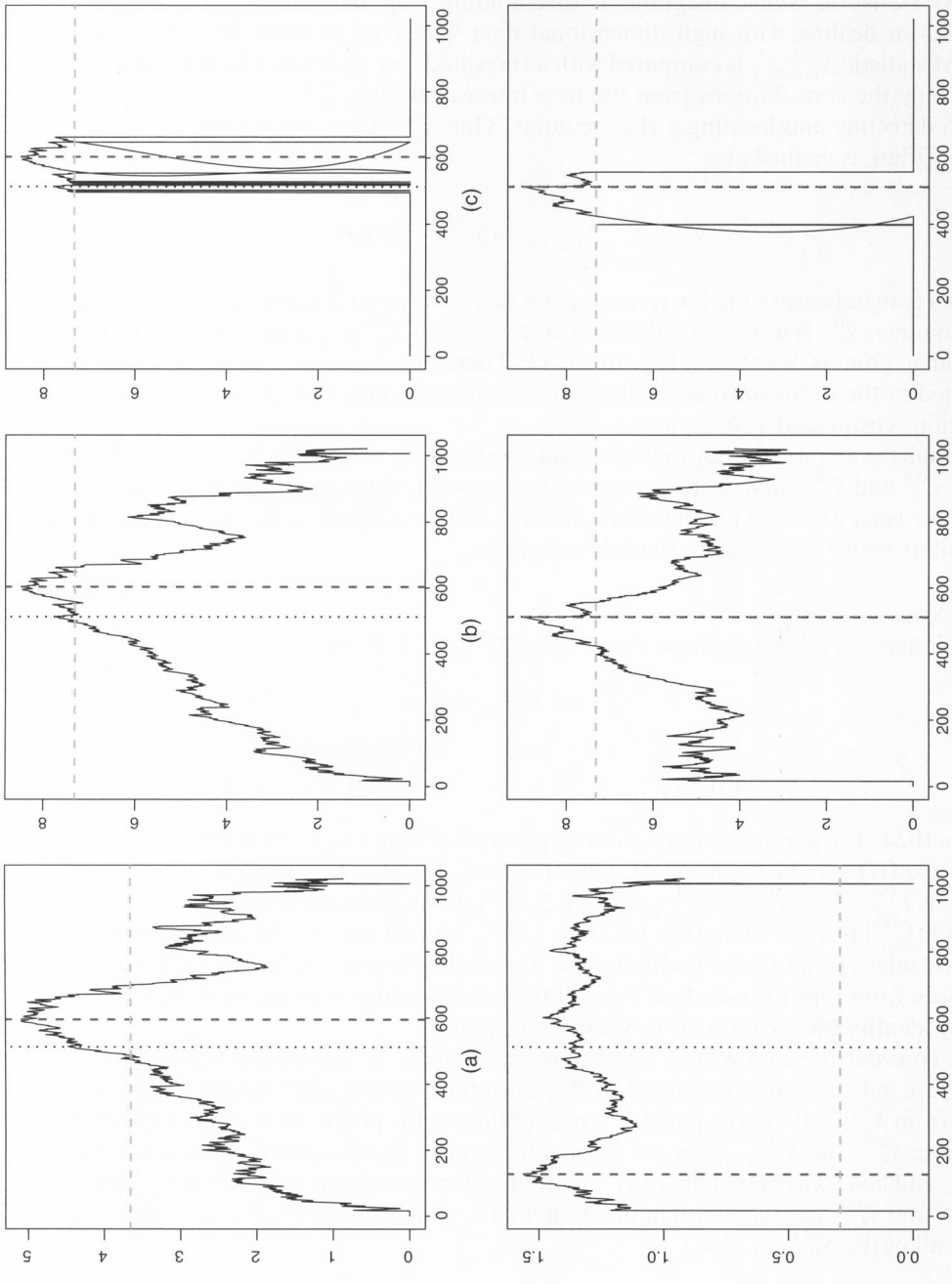

Fig. 1. (a) $\tilde{Y}_t^{\mathrm{avg}}$, (b) $\tilde{Y}_t^{\mathrm{max}}$ and (c) $\tilde{Y}_t^{\mathrm{thr}}$ from the model in Section 2.1.1 when $d = 2$, and (d) $\tilde{Y}_t^{\mathrm{avg}}$, (e) $\tilde{Y}_t^{\mathrm{max}}$ and (f) $\tilde{Y}_t^{\mathrm{thr}}$ when $d = 100$: ⋮, location of the maximum of each of these sequences; ⋮, location of the true change point; – – –, threshold π_T

$$X_{t,T}^{(2)} = \begin{cases} 0.3X_{t-1,T}^{(2)} + \epsilon_{t,T}^{(2)} & \text{for } 1 \leqslant t \leqslant 100, \\ -0.75X_{t-1,T}^{(2)} + \epsilon_{t,T}^{(2)} & \text{for } 101 \leqslant t \leqslant T. \end{cases}$$

$X_{t,T}^{(2)}$ is composed of two stationary segments, where the first segment is relatively short and (weekly) positively auto-correlated, and the second is long and negatively auto-correlated. The negative auto-correlation in $X_{t,T}^{(2)}$ for $t \geqslant 101$ leads to $Y_{t,T}^{(2)}$ being highly auto-correlated, which in turn results in spuriously large values of $\mathcal{Y}_{0,t,T-1}^{(2)}$ for $t \geqslant 101$ even when t is far from the true change point. However, when $d = 2$, all three statistics $\tilde{y}_t^{\text{thr}}$, $\tilde{y}_t^{\max}$ and $\tilde{y}_t^{\text{avg}}$ still manage to locate the true change point around $t = 100$, which is illustrated in Figs 2(a)–2(c).

Now, let $d = 100$ and independently generate 50 time series distributed as $X_{t,T}^{(1)}$ and 50 as $X_{t,T}^{(2)}$ such that the change point is not sparse across the panel. Since there are $d/2 = 50$ sequences $Y_{t,T}^{(k)}$ for which the CUSUM statistics $\mathcal{Y}_{0,t,T-1}^{(k)}$ can take spuriously large values anywhere over $t \in [101, T]$, the statistic $\tilde{y}_t^{\max}$ becomes corrupted and can no longer identify the true change point.

In constrast, $\tilde{y}_t^{\text{thr}}$ not only disregards the contribution from the segments containing no change points but also aggregates the contribution from those containing the change point, and therefore can identify the change point very clearly. In this example, the aggregation effect also causes $\tilde{y}_t^{\text{avg}}$ to work well.

To summarize, $\tilde{y}_t^{\text{thr}}$ is shown to be better at dealing with some difficulties arising from the high dimensionality of the data than either $\tilde{y}_t^{\text{avg}}$ or $\tilde{y}_t^{\max}$ in these two examples. In addition, the superior performance of $\tilde{y}_t^{\text{thr}}$ is attributed to different features of the sparsifying step in the two cases.

Motivated by the above discussion, we now introduce our SBS algorithm for segmenting d-dimensional series. We use j to denote the level index (indicating the progression of the segmentation procedure) and l to denote the location index of the node at each level.

2.1.3. *Sparsified binary segmentation algorithm*

Start with $(j, l) = (1, 1)$, setting $s_{1,1} = 0$, $e_{1,1} = T - 1$ and $n_{1,1} = e_{1,1} - s_{1,1} + 1$.

Step 1: compute the CUSUM statistics $\mathcal{Y}_{s_{j,l},t,e_{j,l}}^{(k)}$ as in expression (3) for all $k = 1, \ldots, d$ over $t \in (s_{j,l}, e_{j,l})$, and obtain $\tilde{y}_t^{\text{thr}}$ as

$$\tilde{y}_t^{\text{thr}} = \sum_{k=1}^{d} \mathcal{Y}_{s_{j,l},t,e_{j,l}}^{(k)} \mathbb{I}(\mathcal{Y}_{s_{j,l},t,e_{j,l}}^{(k)} > \pi_T),$$

with a threshold π_T.

Step 2:

(a) if $\tilde{y}_t^{\text{thr}} = 0$ for all $t \in (s_{j,l}, e_{j,l})$, stop the algorithm for the interval $[s_{j,l}, e_{j,l}]$;

(b) if $\tilde{y}_t^{\text{thr}} \neq 0$, find t that maximizes the corresponding $\tilde{y}_t^{\text{thr}}$ while satisfying

$$\max\left(\frac{t - s_{j,l} + 1}{n_{j,l}}, \frac{e_{j,l} - t}{n_{j,l}}\right) \leqslant c_*, \tag{6}$$

where c_* is identical to the c_* in assumption 1.

(c) If there is any $u \in [t - \Delta_T, t + \Delta_T]$ for which $\tilde{y}_u^{\text{thr}} = 0$, go back to step (b) and find t attaining the next largest $\tilde{y}_t^{\text{thr}}$ while satisfying condition (6). Repeat until a t is found that satisfies $\tilde{y}_u^{\text{thr}} > 0$ for all $u \in [t - \Delta_T, t + \Delta_T]$, set such a t as $b_{j,l}$ and proceed to step 3. If such a t does not exist, stop the algorithm for the interval $[s_{j,l}, e_{j,l}]$.

Step 3: set $b_{j,l}$ as an estimated change point and divide the interval $[s_{j,l}, e_{j,l}]$ into two subintervals $(s_{j+1,2l-1}, e_{j+1,2l-1}) \leftarrow (s_{j,l}, b_{j,l})$ and $(s_{j+1,2l}, e_{j+1,2l}) \leftarrow (b_{j,l} + 1, e_{j,l})$. Update the level j as $j \leftarrow j + 1$ and go to step 1.

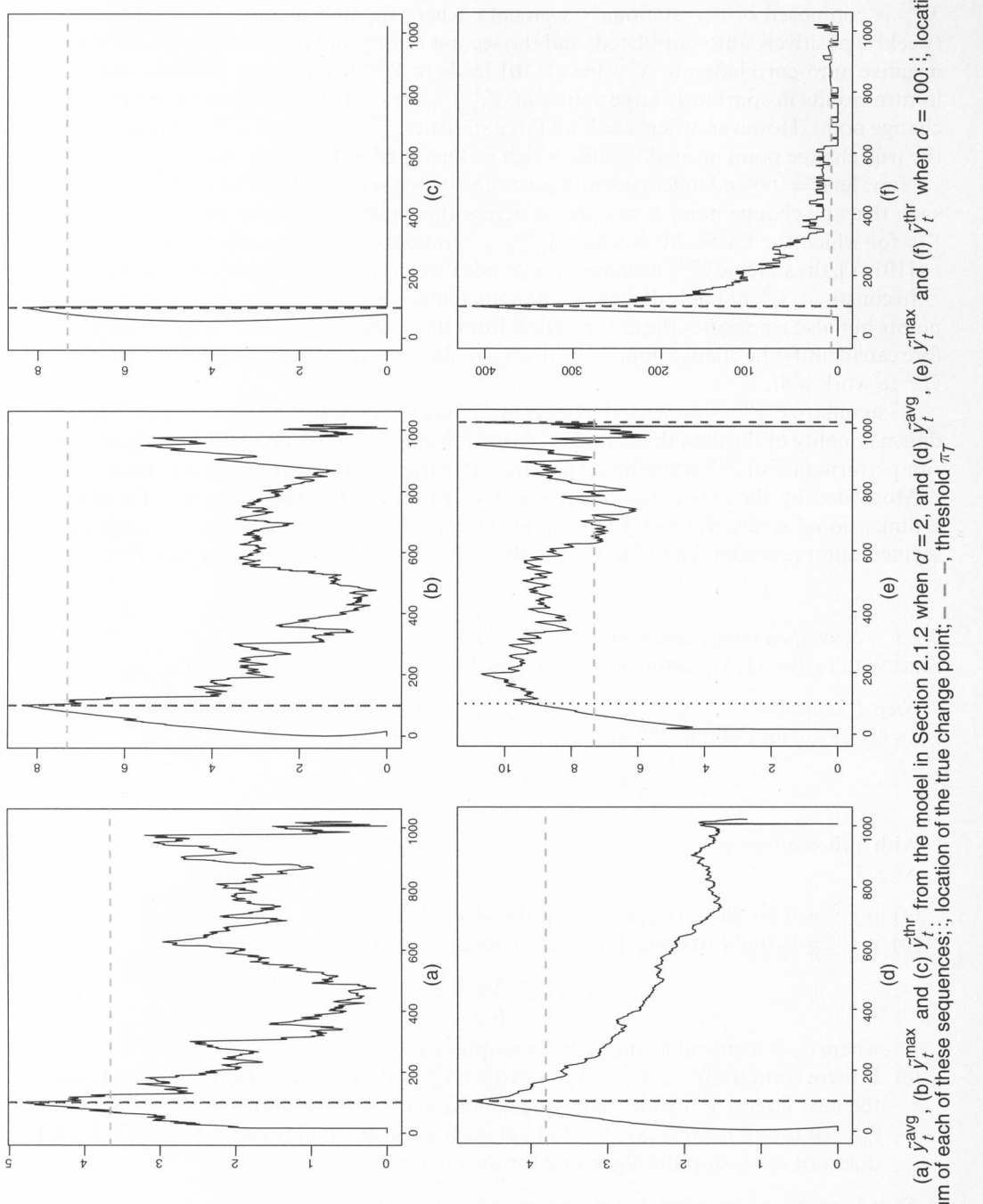

Fig. 2. (a) $\tilde{y}_t^{\mathrm{avg}}$, (b) $\tilde{y}_t^{\mathrm{max}}$ and (c) $\tilde{y}_t^{\mathrm{thr}}$ from the model in Section 2.1.2 when $d = 2$, and (d) $\tilde{y}_t^{\mathrm{avg}}$, (e) $\tilde{y}_t^{\mathrm{max}}$ and (f) $\tilde{y}_t^{\mathrm{thr}}$ when $d = 100$: ⋮, location of the maximum of each of these sequences; :, location of the true change point; – – –, threshold π_T

The algorithm is quitted when step 2(a) is met by all the interval $s_{j,l}, e_{j,l}$ that partition $[1,T]$.

Condition (6) is imposed to prevent the algorithm from detecting a change point that is too close to previously detected ones; note that, in assumption 1, a similar condition is imposed on the locations of the true change points.

As seen in Sections 2.1.1 and 2.1.2 with two motivating examples, the performance of a change point detection method for high dimensional time series depends on many factors besides the underlying dimension, and we cannot set π_T to increase or decrease uniformly with d. Instead, to handle the false alarms in multiple-testing procedures, the threshold π_T is derived such that, on any segment $[s,e]$ containing previously undetected true change points for at least one $k=1,\ldots,d$, the test statistic $\max_{t\in(s,e)} \mathcal{Y}^{(k)}_{s,t,e}$ exceeds π_T with probability converging to 1 for all such k, whereas $\mathcal{Y}^{(k)}_{s,t,e} < \pi_T$ for the remaining ks, as long as d satisfies assumption 4 in Section 2.2.

Also, as the CUSUM statistic $\mathcal{Y}^{(k)}_{s_{j,l},t,e_{j,l}}$ is expected to increase and then to decrease smoothly around true change points without discontinuities, step 2(c) ensures that the algorithm disregards any spurious spikes in $\mathcal{Y}^{(k)}_{s_{j,l},t,e_{j,l}}$. Section 3.3 provides a detailed discussion on the practical selection of the parameters of SBS, including π_T and Δ_T. Steps 2(a) and 2(c) provide a stopping rule for the algorithm on those intervals $[s_{j,l}, e_{j,l}]$ where either no CUSUM statistic $\mathcal{Y}^{(k)}_{s_{j,l},t,e_{j,l}}$ exceeds π_T (step 2(a)), or the exceedance is judged to be spurious (step 2(c)).

As an aside, we note that the mechanics of the SBS algorithm can be applicable in more general situations also, beyond the particular model (1).

2.2. *Consistency of the sparsified binary segmentation algorithm*

To show the consistency of the change points detected by the SBS algorithm in terms of their total number and locations, we impose the following assumptions in addition to assumption 1.

Assumption 2. $\{Z^{(k)}_{t,T}\}_{t=0}^{T-1}$ is a sequence of standard normal variables and $\max_k \phi^{(k)1}_{\infty} < \infty$, where

$$\phi^{(k)}(\tau) = \sup_{t,T} |\mathrm{corr}(Z^{(k)}_{t,T}, Z^{(k)}_{t+\tau,T})|,$$

$$\phi^{(k)r}_{\infty} = \sum_{\tau} |\phi^{(k)}(\tau)|^r.$$

Assumption 3. There are constants $\sigma^*, \sigma_* > 0$ such that $\max_{k,t,T} \sigma^{(k)}(t/T) \leqslant \sigma^*$ and, given any change point η_q in $\sigma^{(k)}(t/T)$,

$$\left|\sigma^{(k)}\left(\frac{\eta_q+1}{T}\right) - \sigma^{(k)}\left(\frac{\eta_q}{T}\right)\right| > \sigma_*$$

uniformly for all $k=1,\ldots,d$.

Assumption 4. d and T satisfy $dT^{-\log(T)} \to 0$.

In particular, condition 4 specifies the maximum rate at which the dimensionality d of model (1) is permitted to increase with the sample size T. Denoting the estimated change points (sorted in increasing order) by $\hat{\eta}_q$, $q=1,\ldots,\hat{N}$, we have the following result.

Theorem 1. Let $\Delta_T \asymp \varepsilon_T$ in the SBS algorithm. Under assumptions 1–4, there exists $C_1 > 0$ such that $\hat{\eta}_q$, $q=1,\ldots,\hat{N}$, satisfy

$$\mathbb{P}\{\hat{N} = N; |\hat{\eta}_q - \eta_q| < C_1\varepsilon_T \text{ for } q=1,\ldots,N\} \to 1$$

as $T \to \infty$, where

(a) if $\delta_T \asymp T$, there is some positive constant κ such that we have $\varepsilon_T = \log^{2+\vartheta}(T)$ with $\pi_T = \kappa \log^{1+\omega}(T)$ for any positive constants ϑ and $\omega > \vartheta/2$ and,
(b) if $\delta_T \asymp T^{\Theta}$ for $\Theta \in (\frac{3}{4}, 1)$, we have $\varepsilon_T = T^{\theta}$ for $\theta = 2 - 2\Theta$ with $\pi_T = \kappa T^{\gamma}$ for some $\kappa > 0$ and any $\gamma \in (1 - \Theta, \Theta - \frac{1}{2})$.

We may define the optimality in change point detection as when each of the true change points and the corresponding estimated change point are within the distance of $O_p(1)$; see for example Korostelev (1987). In this sense, when $\delta_T \asymp T$, the rate of ε_T is near optimal up to a logarithmic factor.

2.3. Post-processing of the change points

We further equip the SBS algorithm with an extra step aimed at reducing the risk of overestimating the number of change points. The step is completely analogous to the corresponding step in the UBS algorithm (see Cho and Fryzlewicz (2012), section 3.2.1), except that it now involves checks of the form

$$\exists k \quad \mathcal{Y}^{(k)}_{\hat{\eta}_{q-1}+1, \hat{\eta}_q, \hat{\eta}_{q+1}} > \pi_T, \tag{7}$$

with the convention $\hat{\eta}_0 = 0$ and $\hat{\eta}_{\hat{N}+1} = T - 1$. In other words, we compute the CUSUM statistic $\mathcal{Y}^{(k)}_{\cdot,\cdot,\cdot}$ on each triple of neighbouring change point estimates for each k and retain only those $\hat{\eta}_q$s for which that statistic exceeds the threshold π_T for at least one k. The reader is referred to Cho and Fryzlewicz (2012) for details. As in the UBS algorithm, the consistency result of theorem 1 is preserved even after performing this extra post-processing.

3. The sparsified binary segmentation algorithm in the multivariate locally stationary wavelet model

In this section, we demonstrate how the SBS algorithm can be used for detecting multiple change points in the second-order (i.e. autocovariance and cross-covariance) structure of multivariate, possibly high dimensional time series.

For this purpose, we first define the multivariate LSW model, in which wavelets act as building blocks analogous to the Fourier exponentials in the classical Cramér representation for stationary processes. Our choice of the LSW model as the theoretical setting is motivated by the attractive features of the univariate LSW model, listed in Cho and Fryzlewicz (2012).

As the simplest example of a wavelet system, we consider Haar wavelets defined as

$$\psi^{\mathrm{H}}_{i,k} = 2^{i/2}\, \mathbb{I}(0 \leqslant k \leqslant 2^{-i-1} - 1) - 2^{i/2}\, \mathbb{I}(2^{-i-1} \leqslant k \leqslant 2^{-i} - 1),$$

where $i \in \{-1, -2, \ldots\}$ and $k \in \mathbb{Z}$ denote scale and location parameters respectively. Small negative values of the scale parameter i denote fine scales where the wavelet vectors are the most localized and oscillatory, whereas large negative values denote coarser scales with longer, less oscillatory wavelet vectors. For a more detailed introduction to wavelets, see for example Nason and Silverman (1995) and Vidakovic (1999). With such wavelets as building blocks, we define the p-variate, piecewise stationary LSW model as follows.

Definition 1. The p-variate LSW process $\{\mathbf{X}_{t,T} = (X^{(1)}_{t,T}, \ldots, X^{(p)}_{t,T})'\}_{t=0}^{T-1}$ for $T = 1, 2, \ldots$, is a triangular stochastic array with the representation

$$X^{(j)}_{t,T} = \sum_{i=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W^{(j)}_i(k/T) \psi_{i,t-k} \xi^{(j)}_{i,k} \qquad \text{for each } j = 1, \ldots, p, \tag{8}$$

where $\boldsymbol{\xi}_{i,k} = (\xi_{i,k}^{(1)}, \xi_{i,k}^{(2)}, \ldots, \xi_{i,k}^{(p)})'$ are independently generated from multivariate normal distributions $\mathcal{N}_p\{\mathbf{0}, \boldsymbol{\Sigma}_i(k/T)\}$, with $\Sigma_i^{(j,j)}(k/T) \equiv 1$ and

$$\operatorname{cov}(\xi_{i,k}^{(j)}, \xi_{i',k'}^{(l)}) = \begin{cases} \delta_{i,i'}\delta_{k,k'}\Sigma_i^{(j,j)}(k/T) = \delta_{i,i'}\delta_{k,k'} & \text{when } j = l, \\ \delta_{i,i'}\delta_{k,k'}\Sigma_i^{(j,l)}(k/T) & \text{when } j \neq l. \end{cases}$$

The parameters $i \in \{-1, -2, \ldots\}$ and $k \in \mathbb{Z}$ denote scale and location respectively, and the Kronecker delta function $\delta_{i,i'}$ returns 1 when $i = i'$ and 0 otherwise. For each i and $j, l = 1, \ldots, p$, the functions $W_i^{(j)}(k/T) : [0,1] \to \mathbb{R}$ and $\Sigma_i^{(j,l)}(k/T) : [0,1] \to \mathbb{R}$ are piecewise constant with an unknown number of change points, and we denote the sets of change points as

$$\mathbb{B}_i^{(j)} = \{z \in (0,1) : \lim_{u \to z-} W_i^{(j)}(u) \neq \lim_{u \to z+} W_i^{(j)}(u)\},$$
$$\mathbb{B}_i^{(j,l)} = \{z \in (0,1) : \lim_{u \to z-} \Sigma_i^{(j,l)}(u) \neq \lim_{u \to z+} \Sigma_i^{(j,l)}(u)\}.$$

In comparison with the Cramér representation for stationary processes, the functions $W_i^{(j)}(k/T)$ can be thought of as scale- and location-dependent transfer functions, whereas the wavelet vectors ψ_i can be thought of as building blocks analogous to the Fourier exponentials.

The autocovariance and the cross-covariance functions of $X_{t,T}^{(j)}$, $j = 1, \ldots, p$, which are defined in Section 3.1.1 below, inherit the piecewise constancy of $W_i^{(j)}(\cdot)$ and $\Sigma_i^{(j,l)}(\cdot)$, with identical change point locations. We denote the set of those change points by

$$\mathbb{B} = \{\cup_{j=1}^p \mathbb{B}^{(j)}\} \cup \{\cup_{j,l=1}^p \mathbb{B}^{(j,l)}\} \equiv \{\nu_r, r = 1, \ldots, N\}. \tag{9}$$

3.1. Wavelet periodograms and cross-periodograms

In this section, we construct particular wavelet-based local periodogram sequences from the LSW time series $\mathbf{X}_{t,T}$ in expression (8), to which the SBS algorithm of Section 2.1 will be applied to detect the change points in the second-order structure of $\mathbf{X}_{t,T}$.

Recall that, in the examples in Section 2.1.1 and Section 2.1.2, the multiplicative sequences were constructed as $Y_{t,T}^{(k)} = 2^{-1}(X_{t+1,T}^{(k)} - X_{t,T}^{(k)})^2$. Note that each element of $Y_{t,T}^{(k)}$ is simply the squared wavelet coefficient of $X_{t,T}^{(k)}$ with respect to Haar wavelets at scale -1, i.e.

$$Y_{t,T}^{(k)} = 2^{-1}(X_{t,T}^{(k)} - X_{t-1,T}^{(k)})^2 = \left(\sum_u X_{u,T}^{(k)} \psi_{-1,t-u}^H\right)^2,$$

or the (Haar) *wavelet periodogram* of $X_{t,T}^{(k)}$ at scale -1. In the two examples, it was shown that the change points in the auto-regressive (AR) coefficients of $X_{t,T}^{(k)}$ (and hence in its second-order structure) were detectable from the wavelet periodograms. In this section, we study the properties of the wavelet periodogram and cross-periodogram sequences, and we discuss the applicability of the SBS algorithm to the segmentation of $\mathbf{X}_{t,T}$ defined as expression (8), with the wavelet periodograms and cross-periodograms of $\mathbf{X}_{t,T}$ as an input.

3.1.1. Definitions and properties

Given a p-variate LSW time series $\mathbf{X}_{t,T} = (X_{t,T}^{(1)}, \ldots, X_{t,T}^{(p)})'$, its empirical wavelet coefficients at scale i are denoted by $w_{i,t,T}^{(j)} = \Sigma_u X_{u,T}^{(j)} \psi_{i,t-u}$ for each $X_{t,T}^{(j)}$, $j = 1, \ldots, p$. Then, the *wavelet periodogram* of $X_{t,T}^{(j)}$ and the *wavelet cross-periodogram* between $X_{t,T}^{(j)}$ and $X_{t,T}^{(l)}$ at scale i are defined as

$$I_{i,t,T}^{(j,j)} \equiv I_{i,t,T}^{(j)} = |w_{i,t,T}^{(j)}|^2,$$
$$I_{i,t,T}^{(j,l)} = w_{i,t,T}^{(j)} w_{i,t,T}^{(l)}$$

respectively. The Gaussianity of $X_{t,T}^{(j)}$ implies the Gaussianity of $w_{i,t,T}^{(j)}$, and hence $I_{i,t,T}^{(j)}$ and $I_{i,t,T}^{(j,l)}$ admit the decompositions

$$I_{i,t,T}^{(j)} = \mathbb{E}[I_{i,t,T}^{(j)}] Z_{t,T}^{(j)2}, \qquad t = 0, \ldots, T-1, \tag{10}$$

$$I_{i,t,T}^{(j,l)} = \mathbb{E}[I_{i,t,T}^{(j,l)}] Z_{t,T}^{(j)} Z_{t,T}^{(l)}, \qquad t = 0, \ldots, T-1, \tag{11}$$

where $\{Z_{t,T}^{(j)}\}_{t=0}^{T-1}$ is a sequence of (correlated and non-stationary) standard normal variables for each $j = 1, \ldots, p$. Therefore each $I_{i,t,T}^{(j)}$ follows a scaled χ_1^2-distribution.

It has been shown in the literature that, for a univariate LSW process $X_{t,T}$, there is an asymptotic one-to-one correspondence between its time varying autocovariance functions $c_T(z,\tau) = \mathrm{cov}(X_{\lfloor zT \rfloor, T}, X_{\lfloor zT \rfloor + \tau, T})$, $\tau = 0, 1, \ldots$, transfer functions $W_i^2(z)$ and the expectations of wavelet periodograms $\mathbb{E}[I_{i,t,T}]$ at multiple scales (see for example Cho and Fryzlewicz (2012)), i.e. any change points in the set of piecewise constant functions $\{W_i^2(z)\}_i$ correspond to change points in the (asymptotic limits of the) autocovariance functions $\{c_T(z,\tau)\}_\tau$, which in turn correspond to the change points in the (asymptotic limits of the) functions $\{\mathbb{E}[I_{i,t,T}]\}_i$, and thus are asymptotically detectable by examining $I_{i,t,T}$, $i = -1, -2, \ldots$. For a multivariate LSW process $\mathbf{X}_{t,T}$, its autocovariance and cross-covariance functions are defined as

$$\begin{aligned} c_T^{(j,j)}(z,\tau) &= c_T^{(j)}(z,\tau) = \mathrm{cov}(X_{\lfloor zT \rfloor, T}^{(j)}, X_{\lfloor zT \rfloor + \tau, T}^{(j)}), \\ c_T^{(j,l)}(z,\tau) &= \mathrm{cov}(X_{\lfloor zT \rfloor, T}^{(j)}, X_{\lfloor zT \rfloor + \tau, T}^{(l)}). \end{aligned} \tag{12}$$

In the multivariate LSW model, analogous one-to-one correspondence can be shown for any pair of $X_{t,T}^{(j)}$ and $X_{t,T}^{(l)}$ between the following quantities: the autocovariance and cross-covariance functions $c_T^{(j)}(z,\tau)$, $c_T^{(l)}(z,\tau)$ and $c_T^{(j,l)}(z,\tau)$ at lags $\tau = 0, 1, \ldots$, piecewise constant functions $W_i^{(j)2}(z)$, $W_i^{(l)2}(z)$ and $\Sigma_i^{(j,l)}(z)$, and the expectations of wavelet periodograms and cross-periodograms $\mathbb{E}[I_{i,t,T}^{(j)}]$, $\mathbb{E}[I_{i,t,T}^{(l)}]$ and $\mathbb{E}[I_{i,t,T}^{(j,l)}]$ at scales $i = -1, -2, \ldots$. Therefore, any change points in the second-order structure of the multivariate time series $\mathbf{X}_{t,T}$ are detectable from the wavelet periodograms and cross-periodograms at multiple scales. Formal derivation of this one-to-one correspondence is provided in Appendix B.

Thus we now focus on wavelet periodogram $I_{i,t,T}^{(j)}$ and cross-periodogram $I_{i,t,T}^{(j,l)}$ as the input to the SBS algorithm. We firstly note that $\mathbb{E}[I_{i,t,T}^{(j)}]$ are piecewise constant except for negligible biases around the change points (which are accounted for in our results; see Appendix B.1), and thus $I_{i,t,T}^{(j)}$ almost follow the multiplicative model (1). However, $I_{i,t,T}^{(j,l)}$ is not of the form specified in model (1) and the next section introduces an alternative to $I_{i,t,T}^{(j,l)}$ which does follow model (1) (again, up to the negligible biases) and contains the same information about the change points as does $I_{i,t,T}^{(j,l)}$.

3.1.2. *Non-negative multiplicative alternative to the cross-periodogram*

To gain an insight into obtaining a possible alternative to $I_{i,t,T}^{(j,l)}$, we first present a toy example. Consider two sequences of zero-mean, serially independent normal variables $\{a_t\}_{t=1}^T$ and $\{b_t\}_{t=1}^T$ where the correlation between a_t and b_t satisfies $\mathrm{corr}(a_t, b_t) = 0$ for $t \leqslant \lfloor T/2 \rfloor$ and $\mathrm{corr}(a_t, b_t) = 0.9$ for $t \geqslant \lfloor T/2 \rfloor + 1$, whereas $\mathrm{var}(a_t)$ and $\mathrm{var}(b_t)$ are constant over time. The change in the second-order structure of $(a_t, b_t)'$ originates solely from the change in the correlation between the two sequences and thus cannot be detected from $\{a_t^2\}_{t=1}^T$ and $\{b_t^2\}_{t=1}^T$ alone. Fig. 3 confirms this, and

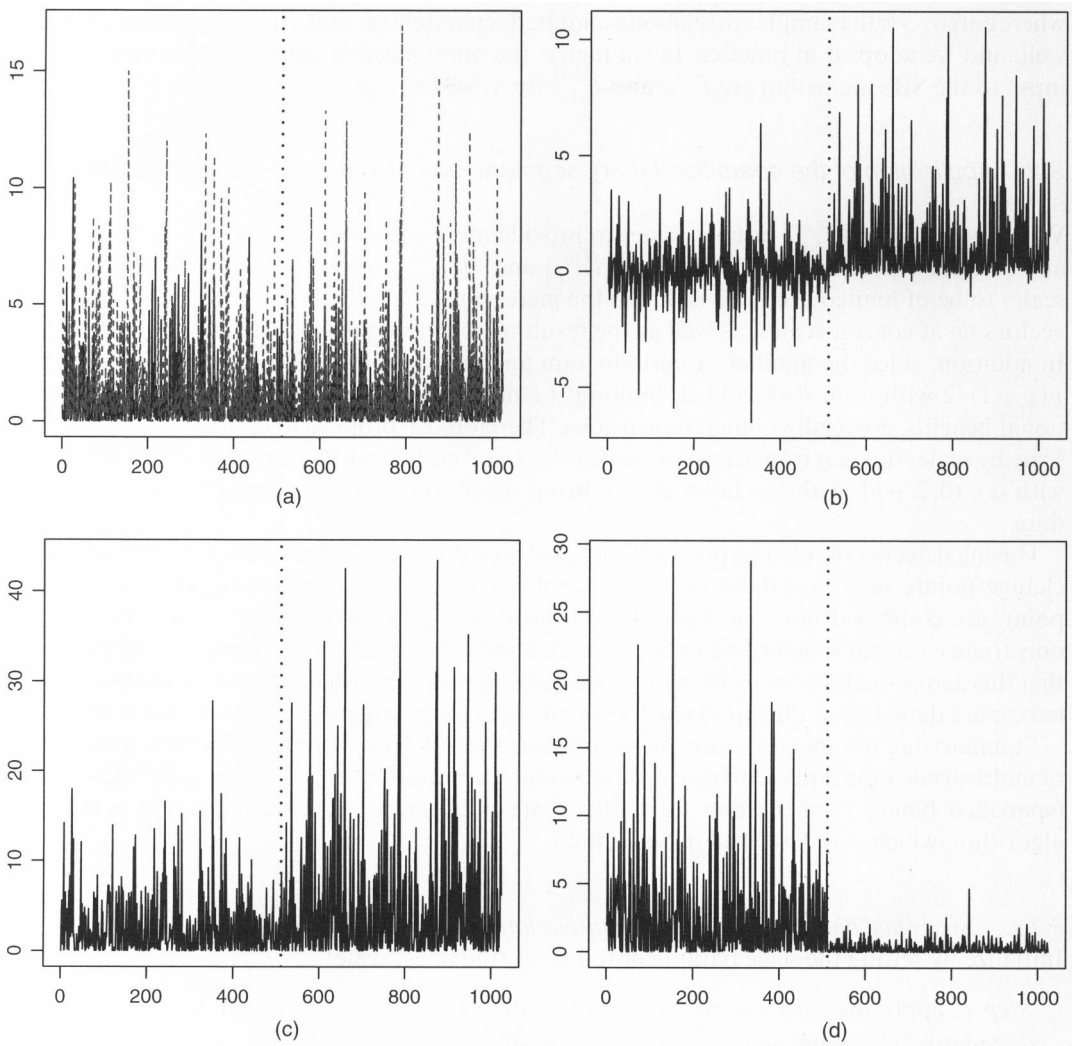

Fig. 3. (a) a_t^2 (———) and b_t^2 (------) from the example of Section 3.1.2, (b) $a_t b_t$, (c) $(a_t + b_t)^2$ and (d) $(a_t - b_t)^2$:, position where $(a_t, b_t)'$ have a change point

it is the sequence $\{(a_t - b_t)^2\}_{t=1}^T$ that exhibits the change point more prominently than $\{a_t b_t\}_{t=1}^T$ or $\{(a_t + b_t)^2\}_{t=1}^T$.

Identifying a_t with $w_{i,t,T}^{(j)}$ and b_t with $w_{i,t,T}^{(l)}$, it becomes apparent that we may detect any change in the covariance structure between $w_{i,t,T}^{(j)}$ and $w_{i,t,T}^{(l)}$ by examining $I_{i,t,T}^{(j)}$, $I_{i,t,T}^{(l)}$ and either $(w_{i,t,T}^{(j)} + w_{i,t,T}^{(l)})^2$ or $(w_{i,t,T}^{(j)} - w_{i,t,T}^{(l)})^2$ instead of $I_{i,t,T}^{(j,l)} = w_{i,t,T}^{(j)} w_{i,t,T}^{(l)}$. Since each variable $w_{i,t,T}^{(j)}$ is zero mean normal, both $(w_{i,t,T}^{(j)} + w_{i,t,T}^{(l)})^2$ and $(w_{i,t,T}^{(j)} - w_{i,t,T}^{(l)})^2$ are scaled χ_1^2-variables, and so either of these sequences can serve as an input to the SBS algorithm. Although both lead to identical results theoretically, there remains the choice between the signs '+' and '–' to optimize finite sample performance. Our empirical observation is that the choice

$$\tilde{I}_{i,t,T}^{(j,l)} = [w_{i,t,T}^{(j)} - \operatorname{sgn}\{\widehat{\operatorname{corr}}(w_{i,t,T}^{(j)}, w_{i,t,T}^{(l)})\} w_{i,t,T}^{(l)}]^2, \tag{13}$$

where $\widehat{\mathrm{corr}}(\cdot,\cdot)$ is the sample correlation computed separately on each current segment, performs well, and we adopt it in practice. In summary, the multiplicative sequences that comprise the input to the SBS algorithm are $I^{(j)}_{i,t,T}$ and $\tilde{I}^{(j,l)}_{i,t,T}$ for $j,l=1,\ldots,p$.

3.2. Application of the sparsified binary segmentation algorithm to multivariate time series

We expect $I^{(j)}_{i,t,T}$ and $\tilde{I}^{(j,l)}_{i,t,T}$ at finer scales to provide more accurate information on the presence and locations of the change points in $\mathbb{E}[I^{(j)}_{i,t,T}]$ and $\mathbb{E}[\tilde{I}^{(j,l)}_{i,t,T}]$, respectively, and those at coarser scales to be of limited use. This is due to the increasing length $\mathcal{L}_i$ of the support of the wavelet vectors $\boldsymbol{\psi}_i$ at coarser scales, as well as the resulting increasing auto-correlation in $\{w^{(j)}_{i,t,T}\}_{t=0}^{T-1}$. In addition, since the number of periodogram and cross-periodogram sequences increases by $p(p+1)/2$ with each scale added, limiting the number of scales also carries clear computational benefits, especially in high dimensions. Therefore we propose to consider $I^{(j)}_{i,t,T}$ and $\tilde{I}^{(j,l)}_{i,t,T}$ scale by scale, starting from the finest scale $i=-1$ and ending with scale $I_T^*=-\lfloor\alpha\log\{\log(T)\}\rfloor$ with $\alpha\in(0,2+\vartheta]$, with the latter choice being made to guarantee consistency of our procedure.

Having detected the change points at each scale separately, we then reduce the set of estimated change points such that those estimated on different scales, yet indicating the same change point, are combined into one with high probability. This is done in the same way as in the univariate case and is described in detail in Cho and Fryzlewicz (2012). Here, we mention only that this across-scales post-processing procedure involves a parameter Λ_T which determines the maximum diameter of the initial clusters of change points originating from different scales.

Summarizing the above arguments, we propose the following algorithm for the segmentation of multivariate time series with piecewise constant second-order structure. We call it SBS-MVTS (sparsified binary segmentation for multivariate time series). Its core ingredient is the SBS algorithm, which was described in Section 2.1.

3.2.1. Sparsified binary segmentation multivariate time series algorithm

Initialize by setting the scale parameter to $i=-1$ (the finest scale).

Step 1: apply the SBS algorithm as well as the post-processing step of Section 2.3 to the $d\equiv p(p+1)/2$ sequences $I^{(j)}_{i,t,T}$, $j=1,\ldots,p$, and $\tilde{I}^{(j,l)}_{i,t,T}$, $j\neq l$, $j,l=1,\ldots,p$, and denote the detected change points by $\hat{\nu}_{i,r}$, $r=1,\ldots,\hat{N}_i$.

Step 2: update $i\leftarrow i-1$ and repeat step 1 until i reaches I_T^*. Apply the across-scales post-processing (which was described earlier in this section) to the change points $\hat{\nu}_{i,r}, r=1,\ldots,\hat{N}_i$, detected from the scales $i=-1,\ldots,I_T^*$, and obtain the final set of estimated change points $\hat{\nu}_r$, $r=1,\ldots,\hat{N}$.

The following theorem demonstrates that the consistency of the SBS algorithm for the multiplicative sequences in model (1) carries over to that of the SBS-MVTS algorithm, provided that the p-variate LSW time series $\mathbf{X}_{t,T}$ on input satisfies conditions 5–9 (in Appendix B), which are analogues of conditions 1–4 but phrased in the specific context of LSW processes. In particular, condition 9 states that the dimensionality p of the input time series $\mathbf{X}_{t,T}$ is permitted to increase with T as long as $p^2T^{-\log(T)}\to 0$.

Theorem 2. Let $\Delta_T\asymp\varepsilon_T$ in the SBS algorithm and $\Lambda_T\asymp\varepsilon_T$ in the across-scales post-processing. Under assumptions 5–9 in Appendix B there exists $C_2>0$ such that $\hat{\nu}_r$, $r=1,\ldots,\hat{N}$, estimated with $I_T^*=-\lfloor\alpha\log\{\log(T)\}\rfloor$ for $\alpha\in(0,2+\vartheta]$, satisfy

$$\mathbb{P}\{\hat{N} = N; |\hat{\nu}_r - \nu_r| < C_2\varepsilon_T \text{ for } r = 1, \ldots, N\} \to 1$$

as $T \to \infty$, where

(a) if $\delta_T \asymp T$, there is some positive constant κ such that we have $\varepsilon_T = \log^{2+\vartheta}(T)$ with $\pi_T = \kappa \log^{1+\omega}(T)$ for any positive constants ϑ and $\omega > \vartheta/2$, and
(b) if $\delta_T \asymp T^{\Theta}$ for $\Theta \in (\frac{3}{4}, 1)$, we have $\varepsilon_T = T^{\theta}$ for $\theta = 2 - 2\Theta$ with $\pi_T = \kappa T^{\gamma}$ for some $\kappa > 0$ and any $\gamma \in (1 - \Theta, \Theta - \frac{1}{2})$.

3.3. *Practical choice of threshold and other quantities*

The aim of this section is to provide some practical guidance on the choice of various parameters of the SBS-MVTS algorithm. We provide heuristic justification for the chosen values below. They have been found to work well in our extensive simulation studies across a range of models; however, we do not claim that other values would not work equally well or better in practice.

Importantly, we also note that the necessity of calibrating these parameters is not specific to the SBS-MVTS algorithm in the sense that they would also need to be set if, for example, $\tilde{y}_t^{\text{avg}}$ or $\tilde{y}_t^{\max}$ were used instead of $\tilde{y}_t^{\text{thr}}$ in a binary segmentation framework.

From the conditions of theorem 1, we have $\gamma \in (1 - \Theta, \Theta - \frac{1}{2})$ in the threshold $\pi_T = \kappa T^{\gamma}$ when $\Theta \in (\frac{3}{4}, 1)$, and ω is any positive constant greater than $\vartheta/2$ in $\pi_T = \kappa \log^{1+\omega}(T)$ when $\Theta = 1$. We propose to set γ as conservatively as $\gamma = 0.499$ and we focus on the choice the constant κ for each $X_{t,T}^{(j)}$, by simulating wavelet periodograms under the null hypothesis of no change points as below. With this approach to the selection of κ, finite sample performance is little affected by whether T^{γ} or $\log^{1+\omega}(T)$ is used as the rate of π_T, and thus we do not expand on the choice of ω here.

For each univariate process $X_{t,T}^{(j)}$, we estimate $\hat{a}_j$, its lag 1 auto-correlation. Then, generating auto-regressive AR(1) time series of length T with the AR parameter $\hat{a}_j$ repeatedly R times, we compute the following statistic for each realization m:

$$\mathbb{J}_i^{(j,m)} = \max_t \left(\frac{1}{T} \sum_{u=1}^{T} I_{i,u}^{(j,m)} \right)^{-1} \left| \sqrt{\left(\frac{T-t}{Tt} \right)} \sum_{u=1}^{t} I_{i,u}^{(j,m)} - \sqrt{\left\{ \frac{t}{T(T-t)} \right\}} \sum_{u=t+1}^{T} I_{i,u}^{(j,m)} \right|,$$

where $I_{i,t}^{(j,m)}$ denotes the scale i wavelet periodogram of the mth AR(1) process generated with the AR parameter $\hat{a}_j$. Note that $\mathbb{J}_i^{(j,m)}$ is of the same form as the test statistic that is used in the SBS algorithm. Since the AR processes have been generated under the null hypothesis of no change points in their second-order structure, $T^{-\gamma}\mathbb{J}_i^{(j,m)}$ may serve as a proxy for κ for the wavelet periodograms generated from $X_{t,T}^{(j)}$. We have observed that the values of $\mathbb{J}_i^{(j,m)}$ tend to increase at coarser scales because of the increasing support of the wavelet vector ψ_i. Therefore, we select κ to be scale dependent as $\kappa_i^{(j)}$ for each $i = -1, -2, \ldots$ and $j = 1, \ldots, p$. In the SBS algorithm, we choose it to be the 99% quantile of $T^{-\gamma}\mathbb{J}_i^{(j,m)}$ over all $m = 1, \ldots, R$. In the case of wavelet cross-periodograms, we use the first-lag sample auto-correlation of $X_{t,T}^{(j)} - \text{sgn}\{\hat{\text{corr}}(w_{i,t,T}^{(j)}, w_{i,t,T}^{(l)})\} X_{t,T}^{(l)}$ in place of $\hat{a}_j$.

As for the choice of Δ_T in step 2(c) of the SBS algorithm, since $\Delta_T \asymp \varepsilon_T$, we choose $\Delta_T = \lfloor \sqrt{T}/2 \rfloor$ to be on conservative side and use it in our implementation for the simulations study that is reported in the next section. Also we use $\alpha = 2$ and $\Lambda_T = \lfloor \sqrt{T}/2 \rfloor$. Finally, rather than choosing a fixed constant as c_*, we make sure that a newly detected change point is distanced from the previously detected change points by at least Δ_T.

4. Simulation study

In this section, we study the performance of the SBS-MVTS algorithm on simulated multivariate

time series with time varying second-order structure. All simulated data sets are generated with $T = 1024$, and the sparsity of the change points across the p-dimensional time series is controlled such that $\lfloor \varrho p \rfloor$ processes out of the p have at least one change point, from a sparse case ($\varrho = 0.05$) through moderate cases ($\varrho = 0.25, 0.5$) to a dense case ($\varrho = 1$).

(a) *Model 1 (AR time series)*: we simulate the p time series as AR(1) processes

$$X_t^{(j)} = \alpha^{(j)} X_{t-1}^{(j)} + \sigma^{(j)} \epsilon_t^{(j)}, \qquad j = 1, \ldots, p. \tag{14}$$

The AR coefficients are independently generated from the uniform distribution $\mathcal{U}(-0.5, 0.999)$, and $\sigma^{(j)}$ from $\mathcal{U}(\frac{1}{2}, 2)$. The error terms $\boldsymbol{\epsilon}_t = (\epsilon_t^{(1)}, \ldots, \epsilon_t^{(p)})'$ are generated from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ with $\boldsymbol{\Sigma}_\epsilon$ specified below. There are three change points at $t = 341, 614, 838$ which occur in the following ways.

(i) At each change point, both $\alpha^{(j)}$ and $\sigma^{(j)}$ are regenerated for randomly chosen $\lfloor \varrho p \rfloor$ time series $X_t^{(j)}$, whereas $\boldsymbol{\Sigma}_\epsilon = 4\mathbf{I}_p$ and remains unchanged throughout.

(ii) Originally, $\boldsymbol{\epsilon}_t$ is generated with a block diagonal variance–covariance matrix $\boldsymbol{\Sigma}_\epsilon = (\Sigma_{j,l})_{j,l=1}^p$, where $\Sigma_{j,j} = 4$ for $j = 1, \ldots, p$, and $\Sigma_{j,l} = 4(-0.95)^{|j-l|}$ for $j, l = 1, \ldots, p/2$ and $\Sigma_{j,l} = 0$ elsewhere. The cross-correlation structure of $\boldsymbol{\epsilon}_t$ changes at each change point as the locations of randomly chosen $\lfloor \varrho p/2 \rfloor$ elements of $\boldsymbol{\epsilon}_t$ are swapped with those of other $\lfloor \varrho p/2 \rfloor$ randomly chosen elements on each stationary segment.

This model has been chosen for the simplicity of the AR(1) dependence structure and for the fact that it permits easy manipulation of the cross-dependence between the component series.

(b) *Model 2 (factor)*: the p time series are generated from a factor model

$$\mathbf{X}_t = \mathbf{A}\boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t,$$

where $\mathbf{A}$ is a $p \times 5$ factor loading matrix with each element $A_{j,l}$ generated from a uniform distribution $\mathcal{U}(0.5, 1.5)$. The vector $\boldsymbol{\eta}_t$ contains five factors, each of which is an independent AR(1) time series generated as $X_t^{(j)}$ in expression (14) with $\boldsymbol{\Sigma}_\epsilon = 4\mathbf{I}_p$. The error terms $\boldsymbol{\epsilon}_t$ follow $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ with the same covariance matrix as that in model 1(ii). There are three change points at $t = 341, 614, 838$ which occur in the following ways.

(i) At each change point, $\lfloor \varrho p \rfloor$ randomly chosen rows of the factor loading matrix $\mathbf{A}$ are regenerated, each from $\mathcal{N}(0, 1)$.

(ii) The cross-correlation structure of $\boldsymbol{\epsilon}_t$ changes as in model 1(ii).

The aim of this model is to investigate the performance of our algorithm when the dependence structure is governed by a factor model, which is a popular dimensionality reduction tool for high dimensional time series.

(c) *Model 3 (AR(1)+MA(2))*: in this example, the p-variate time series $\mathbf{X}_t$ is generated such that

$$X_t^{(j)} = \begin{cases} \epsilon_t^{(j)} + \beta_1^{(j)} \epsilon_{t-1}^{(j)} + \beta_2^{(j)} \epsilon_{t-2}^{(j)} & \text{for } 1 \leqslant t \leqslant 512, \\ \alpha^{(j)} X_t^{(j)} + \sigma^{(j)} \epsilon_t^{(j)} & \text{for } 513 \leqslant t \leqslant 1024, \end{cases}$$

for $j = 1, \ldots, \lfloor \varrho p \rfloor$, and $X_t^{(j)}$, $j = \lfloor \varrho p \rfloor + 1, \ldots, p$, are stationary AR(1) processes with the AR parameters generated from $\mathcal{U}(-0.5, 0.999)$ and $\mathrm{var}(\epsilon_t^{(j)}) = 1$. The coefficients $\beta_1^{(j)}$, $\beta_2^{(j)}$, $\alpha^{(j)}$ and $\sigma^{(j)}$ are generated such that for $X_t^{(j)}$, $j = 1, \ldots, \lfloor \varrho p \rfloor$, the variance and the first-lag auto-correlation remain constant before and after the change point at $t = 512$, whereas auto-correlations at other lags have a change point at $t = 512$. The purpose of this model is to investigate whether the SBS-MVTS algorithm can perform well when the change points are not detectable at the finest scale $i = -1$.

(d) *Model 4 (short segment)*: inspired by the example in Section 2.1.2, the p-variate time series $\mathbf{X}_t$ is generated such that the first $\lfloor \varrho p \rfloor$ processes follow

$$X_t^{(j)} = \begin{cases} \alpha^{(j)} X_t^{(j)} + \epsilon_t^{(j)} & \text{for } 1 \leqslant t \leqslant 100, \\ \beta^{(j)} X_t^{(j)} + \epsilon_t^{(j)} & \text{for } 101 \leqslant t \leqslant 1024, \end{cases}$$

with $\alpha^{(j)}$ drawn from $\mathcal{U}(0.5, 0.59)$ and $\beta^{(j)}$ from $\mathcal{U}(-0.79, -0.5)$. The remaining $p - \lfloor \varrho p \rfloor$ time series are generated as stationary AR(1) processes with the AR parameters drawn from the same distribution as $\beta^{(j)}$. The purpose of this model is to investigate whether the SBS-MVTS algorithm performs well when the finest scale wavelet periodograms suffer from high auto-correlation while the two stationary segments that are defined by the change point are of substantially different lengths.

Most methods for multivariate time series segmentation that have been proposed in the literature, such as those cited in Section 1, have not been designed for data of the dimensionality or size that is considered in this paper, which are $p = 50, 100$ and $T = 1024$ respectively (recall that d is quadratic in p).

In what follows, we compare the performance of the SBS-MVTS algorithm with that of identical binary segmentation algorithms but constructed using $\tilde{y}_t^{\text{avg}}$ and $\tilde{y}_t^{\max}$ in expression (4) instead of $\tilde{y}_t^{\text{thr}}$. For clarity, in the remainder of this section, we refer to the three algorithms as THR (which is equivalent to SBS-MVTS), AVG and MAX. Identical thresholds π_T are applied in THR and MAX. As for AVG, we test $\tilde{y}_t^{\text{avg}}$ by using a scaled threshold $d^{-1} \sum_{k=1}^{d} \mathbb{I}(\max_{t \in (s,e)} \mathcal{Y}_{s,t,e}^{(k)} > \pi_T) \pi_T$ to ensure fairer comparison. As an aside, we note that the threshold selection via simulation is easier for the THR and MAX algorithms than for the AVG algorithm, the reason being that in the first two cases it can be reduced to the problem of threshold selection for univariate time series, which is not so for AVG.

Tables 1–4 report the results of applying the three segmentation algorithms to the simulated data sets from models 1–4. Each table reports the mean and standard deviation of the total number of detected change points over 100 simulated time series, and the percentage of correctly identifying each change point in the time series (in the sense that it lies within the distance of $\lfloor \sqrt{T}/2 \rfloor$ from the true change points).

Overall, it is evident that the THR algorithm outperforms the other two. In particular, the performance of AVG does not match that of THR or MAX especially when the change points are sparse: in some of the models, there is a tendency for AVG to overestimate the number of change points. Besides, the standard deviation of the number of change points detected by AVG tends to be larger than those for the other two algorithms.

In terms of the *number* of detected change points, THR and MAX perform similarly well. However, the accuracy of the detected change point *locations* is significantly better for THR than for MAX, especially in models 3 and 4. This is unsurprising as, effectively, the MAX algorithm locates change points on the basis of one individual component of the input time series, whereas THR typically averages information across many components. We also note that the performance of the THR algorithm does not differ greatly between the cases when $p = 50$ and when $p = 100$.

As noted earlier, the input sequences to the segmentation algorithms, $I_{i,t,T}^{(j)}$ and $\tilde{I}_{i,t,T}^{(j,l)}$, have expectations which are almost piecewise constant but not completely so, owing to negligible biases around the change points (see Appendix B.1). In deriving theorem 2, these biases have fully been taken into account, which implies that the consistency of SBS-MVTS is extended to the case where changes occur in the second-order structure of $\mathbf{X}_{t,T}$ within a short period of time (to be precise, of length $C \log^{\alpha}(T)$ for some $C > 0$ and α from I_T^*), but not entirely

Table 1. Summary of the change points detected from model 1: mean and standard deviation of the total number of change points detected, and the percentage of correctly identifying each change point at $t = 341, 614, 838$ over 100 simulated time series

ϱ		*Results for* $p = 50$						*Results for* $p = 100$					
		Model 1(i)			*Model 1(ii)*			*Model 1(i)*			*Model 1(ii)*		
		THR	*AVG*	*MAX*	*THR*	*AVG*	*MAX*	*THR*	*AVG*	*MAX*	*THR*	*AVG*	*MAX*
0.05	Mean	3.03	2.61	3.01	2.81	3.78	2.8	3.06	3	3.02	3.33	4.97	3.34
	Standard deviation	0.17	0.71	0.1	0.44	1.34	0.45	0.24	0.83	0.14	0.55	1.23	0.54
	$t = 341$	98	71	95	91	65	88	97	55	96	97	55	96
	$t = 614$	89	75	92	91	67	92	99	55	91	99	55	91
	$t = 838$	92	76	91	93	60	91	94	50	87	94	50	87
0.25	Mean	3.03	3.23	3.07	3.01	4.8	3.03	3.08	3.27	3.14	3.02	4.92	3.01
	Standard deviation	0.17	0.58	0.26	0.1	1.13	0.17	0.27	0.57	0.4	0.14	1.24	0.1
	$t = 341$	100	100	86	100	73	89	98	100	84	100	65	87
	$t = 614$	89	100	91	100	57	88	89	99	88	100	66	93
	$t = 838$	99	99	95	100	55	92	99	100	92	100	66	88
0.5	Mean	3.05	3.21	3.05	3.01	4.66	3	3.15	3.48	3.24	3.04	4.9	3.06
	Standard deviation	0.22	0.52	0.22	0.1	1.02	0	0.36	0.64	0.51	0.2	1.14	0.24
	$t = 341$	100	100	85	99	70	90	100	100	80	100	67	88
	$t = 614$	91	100	83	100	69	88	100	100	82	100	68	91
	$t = 838$	98	100	80	100	58	86	100	100	84	100	65	87
1	Mean	3.07	3.25	3.13	3.01	4.76	3.04	3.11	3.59	3.24	3.04	5.03	3.09
	Standard deviation	0.26	0.52	0.37	0.1	1.1	0.2	0.31	0.81	0.45	0.24	1.27	0.29
	$t = 341$	98	100	72	100	61	88	100	100	65	100	75	84
	$t = 614$	99	100	82	100	65	83	100	100	79	100	73	88
	$t = 838$	100	100	89	100	63	85	100	100	88	100	67	84

Table 2. Summary of the change points detected from model 2

ϱ		*Results for* $p=50$						*Results for* $p=100$					
		Model 2(i)			*Model 2(ii)*			*Model 2(i)*			*Model 2(ii)*		
		THR	*AVG*	*MAX*	*THR*	*AVG*	*MAX*	*THR*	*AVG*	*MAX*	*THR*	*AVG*	*MAX*
0.05	Mean	3.04	1.86	3.11	2.81	3.07	2.79	2.98	3.01	2.95	3.27	3.94	3.29
	Standard deviation	0.2	0.96	0.35	0.44	1.28	0.46	0.28	1	0.3	0.57	1.04	0.59
	$t=341$	91	44	88	91	61	92	90	67	81	99	50	86
	$t=614$	89	35	90	92	55	87	89	52	77	96	44	84
	$t=838$	95	41	85	89	59	83	86	52	79	96	43	88
0.25	Mean	3.02	3.07	3.05	3	4.11	3.01	3.01	3.43	3.05	3.01	4.42	3.01
	Standard deviation	0.14	0.76	0.22	0	0.96	0.1	0.1	0.78	0.22	0.1	1.16	0.1
	$t=341$	95	66	93	98	70	95	93	77	80	100	71	88
	$t=614$	95	66	91	100	74	86	90	69	84	100	79	92
	$t=838$	93	58	85	100	70	93	94	74	84	100	70	91
0.5	Mean	3.01	3.07	3.02	3	4.43	3.02	3.03	3.11	3.07	3	4.32	3.04
	Standard deviation	0.1	0.48	0.14	0	1.08	0.14	0.17	0.31	0.26	0	1.07	0.2
	$t=341$	97	82	86	97	75	85	96	88	83	100	79	93
	$t=614$	93	80	87	100	76	94	92	98	77	100	73	89
	$t=838$	93	90	80	98	67	87	95	95	82	99	70	93
1	Mean	3	3.1	3.01	3.01	4.05	3.02	3	3.2	3.03	3	4.27	3.05
	Standard deviation	0	0.36	0.1	0.1	1.08	0.14	0	0.57	0.17	0	1.12	0.22
	$t=341$	99	71	94	99	71	94	94	100	84	98	70	86
	$t=614$	99	72	89	99	72	89	93	100	89	100	78	86
	$t=838$	100	69	91	100	69	91	94	100	81	100	66	80

Table 3. Summary of the change points detected from model 3

ϱ		*Results for* $p=50$			*Results for* $p=100$		
		THR	*AVG*	*MAX*	*THR*	*AVG*	*MAX*
0.05	Mean	0.63	0.03	0.69	1.08	0.09	0.98
	Standard deviation	0.53	0.17	0.49	0.46	0.32	0.62
	$t=512$	51	1	49	72	3	64
0.25	Mean	1.01	0.11	1.04	1.02	0.32	0.99
	Standard deviation	0.22	0.31	0.28	0.2	0.51	0.33
	$t=512$	92	9	73	92	25	75
0.5	Mean	1.01	0.31	1	1.05	0.47	1.07
	Standard deviation	0.17	0.51	0.28	0.22	0.56	0.29
	$t=512$	92	29	77	90	39	68
1	Mean	1.02	0.36	1.03	1.13	0.68	1.22
	Standard deviation	0.14	0.5	0.22	0.34	0.65	0.46
	$t=512$	95	34	67	95	53	66

Table 4. Summary of the change points detected from model 4

ϱ		*Results for* $p=50$			*Results for* $p=100$		
		THR	*AVG*	*MAX*	*THR*	*AVG*	*MAX*
0.05	Mean	0.99	1.03	0.98	0.88	4.12	0.89
	Standard deviation	0.52	1.49	0.51	0.38	2.06	0.37
	$t=100$	80	35	73	80	87	78
0.25	Mean	1.04	1.72	0.98	1.06	4.74	1.12
	Standard deviation	0.24	1.58	0.51	0.34	1.98	0.67
	$t=100$	91	74	73	93	97	73
0.5	Mean	1.14	2.1	1.03	1.09	5.56	1.09
	Standard deviation	0.47	1.64	0.41	0.32	2.26	0.38
	$t=100$	92	92	74	97	100	62
1	Mean	1.09	2.94	1.01	1.28	0.02	1.05
	Standard deviation	0.32	1.9	0.41	0.73	0.14	0.39
	$t=100$	94	99	50	97	2	49

synchronized. To confirm this, we performed a further simulation study where the p-variate time series was generated from model M3, except that the change points were allowed to be anywhere within an interval of length $\lfloor 2\log(T) \rfloor$ around $t=512$. Although not reported here, we obtained the change point detection results with $T=1024$ and varying ϱ and p, which were comparable with those reported in Table 3. More specifically, although the number of detected change points had greater variance, the accuracy in their locations was preserved even when the change points were not aligned. Also, overall, the THR algorithm still outperformed the two other competitors in terms of both the total number of the detected change points and their locations.

(We now abandon the THR notation and revert to the SBS-MVTS notation in the remainder of the paper.)

5. Detecting change points in the component processes of the Standard and Poors 500 index

We further study the performance of the SBS-MVTS algorithm by applying it to the multivariate time series of *daily closing prices* of the constituents of the Standard and Poors 500 stock market index. The period considered is between January 1st, 2007, and December 31st, 2011, overlapping with the period of the recent financial crisis. We have chosen only those 456 constituents that remained in the index over the entire period; the resulting time series is of dimensionality $p = 456$ and length $T = 1260$ (we recall that d is quadratic in p and therefore much larger than T in this example).

Before presenting the change point detection results, we briefly mention the rationale behind our approach to this data set. As noted in Section 3.1, the wavelet periodograms computed with Haar wavelets at scale $i = -1$ take the form $I_{i,t,T} = 2^{-1}(X_{t+1,T} - X_{t,T})^2$ and thus reflect the behaviour of return series, and these periodograms comprise the input multiplicative sequences to SBS-MVTS. Mikosch and Stărică (2004) discussed that the 'stylized facts' observable in financial time series, such as long-range dependence of the absolute returns, might be artefacts that are induced by change points in the second-order structure of the series. It was further discussed in Fryzlewicz (2005) where a class of Gaussian LSW time series was shown to embed these stylized facts.

When first applied to the first 100 component processes, the algorithm returns $t = 67$, 129, 198, 276, 427, 554, 718, 864, 1044, 1147 as change points. We then apply the algorithm to the first 200 processes to obtain $t = 67$, 126, 198, 270, 333, 427, 554, 652, 718, 867, 1022, 1086, 1148 as change points. Comparing the two sets of change points detected, it is reassuring to see that those from the former set also appear to have their counterparts in the latter, as expected, since the latter data set contains the former. When applied to the entire p-variate time series, the SBS-MVTS algorithm returns the change points that are summarized in Table 5, which also lists some historical events that occurred close to some of the change points detected.

The 'T-bill and ED' (TED) spread is the difference between the interest rate at which the US

Table 5. Summary of the change points detected from the component processes of the Standard and Poors 500 index

t	*Date*	*Historical event*
66	April 9th, 2007	
126	July 3rd, 2007	TED spread†
199	October 16th, 2007	US stock market peaked in October 2007
274	February 4th, 2008	
426	September 10th, 2008	TED spread†
550	March 10th, 2009	Dow Jones average index reached a trough of around 6600 by March 2009; identified by the *New York Times* as the 'nadir of the crisis'
711	October 27th, 2009	
864	June 8th, 2010	TED spread†
1017	January 13th, 2011	
1088	April 27th, 2011	
1148	July 22nd, 2011	Global stock markets fell owing to fears of contagion of the European sovereign debt crisis

†Refer to the TED spread in Fig. 4.

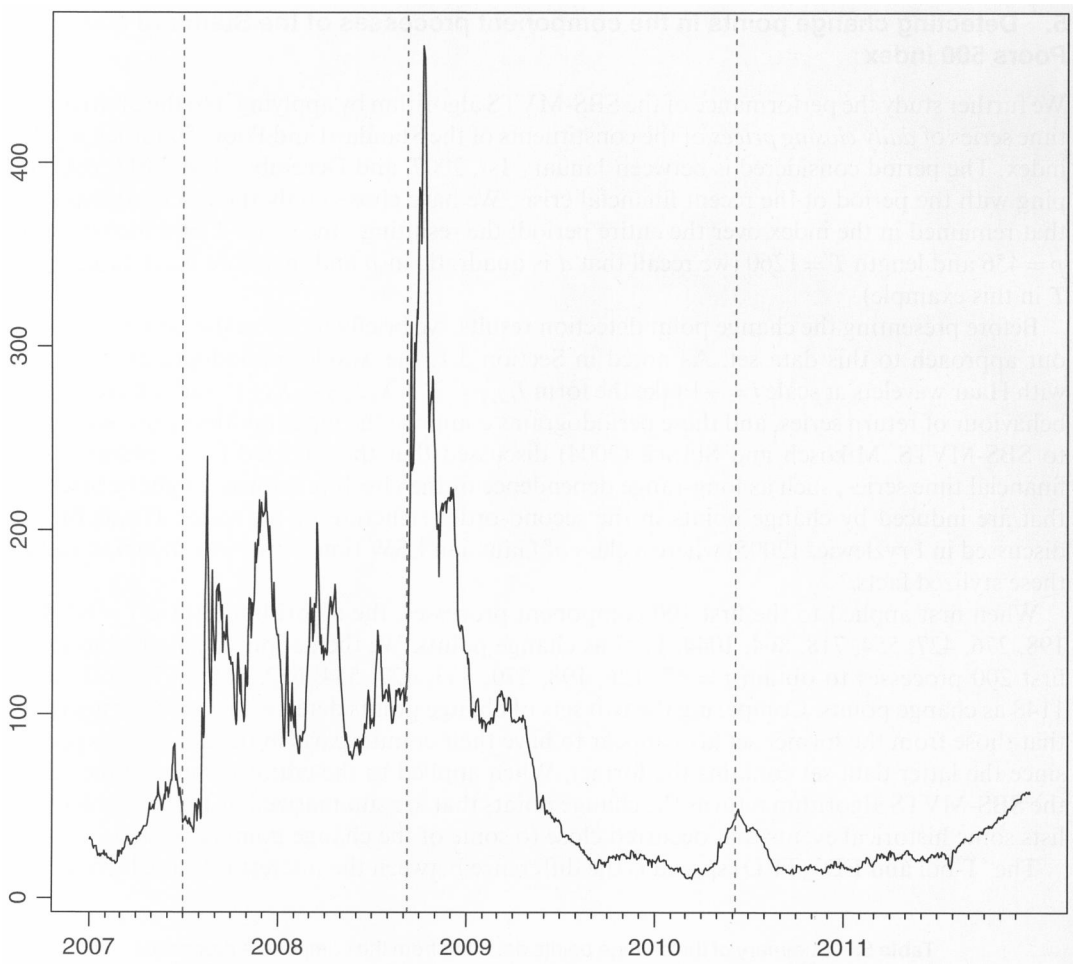

Fig. 4. TED spread between 2007 and 2011: ⋮, estimated change points indicated in Table 5

Government can borrow over a 3-month period (T-bill) and the rate at which banks lend to each other over the same period (measured by the Libor index) and therefore can serve as an indicator of perceived credit risk in the general economy. During 2007, the TED spread rapidly increased to around 150–200 basis points, which coincided with the 'subprime' mortgage crisis, and, in mid-September 2008, it exceeded 300 basis points. In 2010, it returned to its long-term average of 30 basis points. However, it started to rise again with the beginning of the European debt crisis and reached above 45 basis points by mid-June. The volatile behaviour of the TED spread during 2007–2011 is reflected in some of the change points that are detected by the SBS-MVTS algorithm as shown in Fig. 4.

To check the validity of the detected change points further, we tested the stationarity of the series within the segments that are examined at each iteration of the SBS-MVTS algorithm. The problem of testing stationarity for multivariate time series has not been widely studied; Jentsch and Subba Rao (2013) noted that only few procedures exist for such a purpose and those are not easily applicable to the current data set with dimensionality as large as $p = 456$.

Instead, we chose to examine the stationarity of the first few principal component series obtained over each segment. Various methods have been proposed for testing second-order stationarity of univariate time series and, among them, the multiple-testing procedure that was proposed in Nason (2013) is available in the format of an R package. However, since its test statistics are close to ours except that they are computed at the locations which are a power of 2, we concluded that performing this procedure would not be suitable for our purpose.

Alternatively, we adopted the stationarity test that was proposed in Dwivedi and Subba Rao (2011) (R code is available from `http://www.stat.tamu.edu/~suhasini/Rcode.html`), which tests whether the correlations between the discrete Fourier transforms of the series are close to 0. We applied the testing procedure to each segment examined as the SBS-MVTS algorithm proceeded, i.e., since change points were detected in the order 550, {426, 1148}, {199, 1017}, {126, 274, 711, 1088} and {66, 864} (those detected at the same 'level' were grouped together), we investigated the segments [1, 1260], [1, 549], [550, 1260], [1, 425], [426, 549] and so forth. Within each segment $[s, e]$, principal component analysis was performed on $\mathbf{X}_t$, producing two factor series as the first two principal components. As these factors often exhibited high auto-correlations (which might falsely lead to rejecting the null hypothesis), we fitted an AR(1) process to each factor and tested the stationarity of these residual series.

Furthermore, we checked whether the resulting residuals behaved like Gaussian white noise. It may be expected that, if $\mathbf{X}_t$ is stationary within $t \in [s, e]$, the residuals behave like Gaussian white noise under our LSW model, whereas, if its second-order structure undergoes a change, departure from Gaussianity is observable from the distribution of the residuals. To do so, we adopted the normality tests which were implemented in R (packages `tseries` and `nortest`), namely Lilliefors, Anderson–Darling, Pearson, Shapiro–Francia and Jarque–Bera tests. Although failing to reject the null hypothesis via these tests does not guarantee that the residual series follows a normal distribution, they can serve as an indicator that certain moments and quantiles of the residuals behave like those of Gaussian random variables.

Adopting the Bonferroni correction as in Nason (2013), we rejected the null hypothesis of stationarity or normality when the corresponding p-value was smaller than $\alpha^* = 0.05/23 = 0.00212$ (dependence in the test statistics was not taken into account). For most of the segments containing any change points, the p-values were smaller than α^* for at least one of the factors, except for [119, 425] (for normality tests) and [1017, 1147] (for both tests). In contrast, p-values were generally greater than α^* over the segments which did not contain any change point, indicating that the residuals over these segments behaved similarly to Gaussian white noise. For some segments, such as [1, 65], both of the null hypotheses were rejected, which implies that further change points could have been detected but the restriction imposed on change point dispersion in the SBS algorithm prevented them from being detected.

Overall, the findings support the use of the SBS-MVTS methodology in this case-study.

Acknowledgements

We thank the Joint Editor, Associate Editor and two referees for very helpful comments which led to a substantial improvement of this manuscript.

Appendix A: Proof of theorem 1

We first prove a set of lemmas that are essential in proving theorem 1 for a single multiplicative sequence following model (1). Note that, when $d = 1$, the algorithm returns identical change points no matter whether $\tilde{y}_t^{\mathrm{thr}}$ or the raw CUSUM statistic $\mathcal{Y}_{s,t,e}^{(1)}$ is used. In this section, the superscripts are suppressed where there is no confusion. Define $\mathbb{Y}_{s,b,e}$ as

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

for $n = e-s+1$, and $\mathbb{S}_{s,b,e}$ is defined similarly with $\sigma(t/T)$ replacing $Y_{t,T}$. Further, let $\eta_1 < \eta_2 < \ldots < \eta_N$ be the change points in $\sigma(t/T)$ (with the convention of $\eta_0 = 0$ and $\eta_{N+1} = T-1$). In what follows, $c_i, i = 1, 2, \ldots$, are used to denote specific positive constants and C and C' to denote generic constants.

Let s and e denote the 'start' and the 'end' of a segment to be examined at some stage of the algorithm. Further, we assume that s and e satisfy

$$\eta_{q_1} \leqslant s < \eta_{q_1+1} < \ldots < \eta_{q_2} < e \leqslant \eta_{q_2+1}$$

for $0 \leqslant q_1 < q_2 \leqslant N$. In lemmas 1–5, we impose at least one of following conditions:

$$s < \eta_{q_1+q} - c_1\delta_T < \eta_{q_1+q} + c_1\delta_T < e \qquad \text{for some } 1 \leqslant q \leqslant q_2 - q_1, \tag{15}$$

$$\{(\eta_{q_1+1} - s) \wedge (s - \eta_{q_1})\} \vee \{(\eta_{q_2+1} - e) \wedge (e - \eta_{q_2})\} \leqslant c_2\varepsilon_T, \tag{16}$$

where '$\wedge$' and '$\vee$' are the minimum and maximum operators. We later show that, under assumptions 1–4 in Section 2, both conditions (15) and (16) hold throughout the algorithm for all those segments which contain change points still to be detected. Finally, throughout the following proofs, δ_T and ε_T are as assumed in theorem 1 along with the threshold π_T and other quantities that are involved in their definitions, i.e. $\theta, \vartheta, \kappa, \gamma$ and ω.

Lemma 1. Let s and e satisfy condition (15). Then there exists $1 \leqslant q^* \leqslant q_2 - q_1$ such that

$$|\mathbb{S}_{s,\eta_{q_1+q^*},e}| \geqslant c_3 \frac{\delta_T}{\sqrt{T}}. \tag{17}$$

Proof. When there is a single change point in $\sigma(z)$ over (s, e), we have $q^* = 1$ and thus use the constancy of $\sigma(z)$ to the left and the right of $\eta_{q_1+q^*}$ to show that

$$|\mathbb{S}_{s,\eta_{q_1+q^*},e}| = \sqrt{\left\{ \frac{(\eta_{q_1+q^*} - s + 1)(e - \eta_{q_1+q^*})}{n} \right\}} \left| \sigma\left(\frac{\eta_{q_1+q^*}+1}{T} \right) - \sigma\left(\frac{\eta_{q_1+q^*}}{T} \right) \right|,$$

which is bounded from below by $\sigma_* c_1 \delta_T / \sqrt{T}$ from assumptions in Section 2. In the case of multiple change points, we remark that, for any q satisfying condition (15), there is at least one q^* for which

$$\left| \frac{1}{\eta_{q_1+q^*} - s + 1} \sum_{t=s}^{\eta_{q_1+q^*}} \sigma\left(\frac{t}{T}\right) - \frac{1}{e - \eta_{q_1+q^*}} \sum_{t=\eta_{q_1+q^*}+1}^{e} \sigma\left(\frac{t}{T}\right) \right| \tag{18}$$

is bounded away from zero under condition 3. Therefore, the same arguments apply as in the case of a single change point and condition (17) follows.

Lemma 2. Suppose that condition (15) holds. Then there exists $c_0 \in (0, \infty)$ such that, for b satisfying $|\eta_{q_1+q} - b| \geqslant c_0\varepsilon_T$ and $\mathbb{S}_{s,b,e} < \mathbb{S}_{s,\eta_{q_1+q},e}$ for some q, we have $\mathbb{S}_{s,\eta_{q_1+q},e} \geqslant \mathbb{S}_{s,b,e} + C\varepsilon_T/\sqrt{T}$.

Proof. Without loss of generality, let $\eta \equiv \eta_{q_1+q} < b$. Then we have

$$\mathbb{S}_{s,b,e} = \frac{\sqrt{(\eta - s + 1)}\sqrt{(e-b)}}{\sqrt{(e-\eta)}\sqrt{(b-s+1)}} \mathbb{S}_{s,\eta,e},$$

and therefore, using the Taylor expansion and lemma 1,

$$\begin{aligned}
\mathbb{S}_{s,\eta,e} - \mathbb{S}_{s,b,e} &= \left\{ 1 - \frac{\sqrt{(\eta - s + 1)}\sqrt{(e-b)}}{\sqrt{(e-\eta)}\sqrt{(b-s+1)}} \right\} \mathbb{S}_{s,\eta,e} \\
&= \frac{\sqrt{\{1 + (b-\eta)/(\eta - s + 1)\}} - \sqrt{\{1 - (b-\eta)/(e-\eta)\}}}{\sqrt{\{1 + (b-\eta)/(\eta - s + 1)\}}} \mathbb{S}_{s,\eta,e} \\
&\geqslant \frac{1 + c_0\varepsilon_T/(2c_1\delta_T) - \{1 - c_0\varepsilon_T/(2c_1\delta_T)\} + o(c_0\varepsilon_T/n)}{\sqrt{2}} \mathbb{S}_{s,\eta,e} \geqslant \frac{c_0\varepsilon_T}{c_1\sqrt{2}\delta_T} c_3 \frac{\delta_T}{\sqrt{T}} = C\frac{\varepsilon_T}{\sqrt{T}}.
\end{aligned}$$

Lemma 3. Define

$$\mathcal{D}=\left\{1\leqslant s<b<e\leqslant T; n\equiv e-s+1\geqslant \delta_T \text{ and } \max\left(\frac{b-s+1}{n}, \frac{e-b}{n}\right)\leqslant c_*\right\}$$

for the same c_* as that used in expression (2). Then, as $T\to\infty$,

$$\mathbb{P}\{\max_{(s,b,e)\in\mathcal{D}} |\mathbb{Y}_{s,b,e}-\mathbb{S}_{s,b,e}|>\log(T)\}\to 0. \tag{19}$$

Proof. We first study the probability of the event

$$\frac{1}{\sqrt{n}}\left|\sum_{t=s}^{e} c_t\,\sigma\left(\frac{t}{T}\right)(Z_{t,T}^2-1)\right|>\log(T), \tag{20}$$

where $c_t=\sqrt{\{(e-b)/(b-s+1)\}}$ for $s\leqslant t\leqslant b$ and $c_t=-\sqrt{\{(b-s+1)/(e-b)\}}$ for $b+1\leqslant t\leqslant e$. From the definition of $\mathcal{D}$, we have $|c_t|\leqslant c^*\equiv\sqrt{\{c_*/(1-c_*)\}}<\infty$. Let $\{U_i\}_{i=1}^n$ denote IID standard normal variables, $\mathbf{V}=(v_{i,j})_{i,j=1}^n$ with $v_{i,j}=\mathrm{corr}\left(Z_{i,T}, Z_{j,T}\right)$ and $\mathbf{W}=(w_{i,j})_{i,j=1}^n$ be a diagonal matrix with $w_{i,i}=c_t\,\sigma(t/T)$ where $t=i+s-1$. By standard results (see for example Johnson and Kotz (1970), page 151), the probability of event (20) equals $\mathbb{P}\{n^{-1/2}|\Sigma_{i=1}^n\lambda_i(U_i^2-1)|>\log(T)\}$, where λ_i are eigenvalues of the matrix $\mathbf{VW}$. Because of the Gaussianity of U_i, it follows that $\lambda_i(U_i^2-1)$ satisfy Cramér's condition, i.e. there is a constant $C>0$ such that

$$\mathbb{E}[|\lambda_i(U_i^2-1)|^k]\leqslant C^{k-2}k!\,\mathbb{E}[|\lambda_i(U_i^2-1)|^2], \qquad k=3,4,\ldots.$$

Therefore we can apply the Bernstein inequality (Bosq, 1998) and obtain

$$\mathbb{P}\left\{\left|\sum_{t=s}^{e} c_t\,\sigma\left(\frac{t}{T}\right)(Z_{t,T}^2-1)\right|>\sqrt{n}\log(T)\right\}\leqslant 2\exp\left\{-\frac{n\log^2(T)}{4\sum_{i=1}^{n}\lambda_i^2+2\max_i|\lambda_i|C\sqrt{n}\log(T)}\right\}.$$

It holds that $\Sigma_{i=1}^n\lambda_i^2=\mathrm{tr}(\mathbf{VW})^2\leqslant c^{*2}\max_z\{\sigma^2(z)\}n\phi_\infty^2$. We also note that $\max_i|\lambda_i|\leqslant c^*\max_z\sigma(z)\|\mathbf{V}\|_2$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix, and that $\|\mathbf{V}\|_2\leqslant\phi_\infty^1$. Then, the probability in expression (19) is bounded from above by

$$\sum_{(s,b,e)\in\mathcal{D}} 2\exp\left\{-\frac{n\log^2(T)}{4c^{*2}\max_z\{\sigma^2(z)\}n\phi_\infty^2+2c^*\max_z\{\sigma(z)\}\sqrt{n\log(T)}\phi_\infty^1}\right\}\leqslant 2T^3\exp\{-C'\log^2(T)\}$$

which converges to 0, since $\phi_\infty^1<\infty$ from assumption 2 in Section 2, $n\geqslant\delta_T>\log(T)$ and $c^*<\infty$.

Lemma 4. Under conditions (15) and (16), define an interval

$$\mathcal{D}_{s,e}=\{t\in(s,e); \max\{(t-s+1)/n, (e-t)/n\}\leqslant c_*\}\subset[s,e].$$

Then there exists $1\leqslant q^*\leqslant q_2-q_1$ such that $\eta_{q_1+q^*}\in\mathcal{D}_{s,e}$ and $|\hat{\eta}-\eta_{q_1+q^*}|<c_0\varepsilon_T$ for $\hat{\eta}=\arg\max_{t\in\mathcal{D}_{s,e}}|\mathbb{Y}_{s,t,e}|$.

Proof. The following proof is an adaptation of the proof of theorem 3.1 in Fryzlewicz (2014) to non-Gaussian and non-IID noise.

We note that model (1) can be rewritten as

$$Y_{t,T}=\sigma(t/T)+\sigma(t/T)(Z_{t,T}^2-1), \qquad t=0,\ldots,T-1,$$

which in turn can be regarded as a generic additive model $y_t=f_t+\epsilon_t$ with a piecewise constant signal f_t by setting $y_t=Y_{t,T}$, $f_t=\sigma(t/T)$ and $\epsilon_t=\sigma(t/T)(Z_{t,T}^2-1)$.

On a given segment $[s,e]$, detecting a change point is equivalent to fitting the best step function (i.e. a piecewise constant function with one change point) $\hat{f}_t$ which minimizes $\Sigma_{t=s}^e(y_t-g_t)^2$ among all step functions g_t defined on $[s,e]$. Let f_t^0 denote the best step function approximation to f_t with its change point within $\mathcal{D}_{s,e}$, i.e. any g_t which has its change point in $\mathcal{D}_{s,e}$ and minimizes $\Sigma_{t=s}^e(f_t-g_t)^2$ (f_t^0 may or may not be unique). Under conditions 1 and (15)–(16), lemmas 2.2–2.3 in Venkatraman (1992) imply that the single change point in f_t^0 coincides with one of any undetected change points of f_t in $\mathcal{D}_{s,e}$, and we denote such a change point by η.

Let us assume that $\hat{f}_t$ has a change point at $t=\hat{\eta}$ and it satisfies $|\hat{\eta}-\eta|=c_0\varepsilon_T$. Then, if we show that

$$\sum_{t=s}^{e}(y_t-f_t^0)^2-\sum_{t=s}^{e}(y_t-\hat{f}_t)^2<0, \tag{21}$$

it would prove that $\hat{\eta}$ must be within the distance less than $c_0\varepsilon_T$ from η. Expanding the left-hand side of inequality (21), we obtain

$$\sum_{t=s}^{e}(\epsilon_t+f_t-f_t^0)^2-\sum_{t=s}^{e}(\epsilon_t+f_t-\hat{f}_t)^2=2\sum_{t=s}^{e}\epsilon_t(\hat{f}_t-f_t^0)+\sum_{t=s}^{e}\{(f_t-f_t^0)^2-(f_t-\hat{f}_t)^2\}\equiv \mathrm{I}+\mathrm{II}.$$

From the definition of f_t^0, it is clear that $\mathrm{II}<0$. Let $\mathcal{F}$ be the set of vectors whose elements are initially positive and constant, and then, after a change point, are negative and constant; moreover, they sum to 0 and to 1 when squared. Let $\bar{f}$ be the mean of f_t on $t\in[s,e]$, and let the vector $\psi^0\in\mathcal{F}$ satisfy $f_t^0=\bar{f}+\langle f,\psi^0\rangle\psi^0$. Then we have

$$\begin{aligned}\sum_{t=s}^{e}(f_t-f_t^0)^2&=\sum_{t=s}^{e}(f_t-\bar{f})^2-2\langle f,\psi^0\rangle\sum_{t=s}^{e}(f_t-\bar{f})\psi^0+\langle f,\psi^0\rangle^2\sum_{t=s}^{e}(\psi^0)^2\\&=\sum_{t=s}^{e}(f_t-\bar{f})^2-\langle f,\psi^0\rangle^2.\end{aligned} \tag{22}$$

Let a step function $\tilde{f}_t$ be chosen to minimize $\Sigma_{t=s}^{e}(f_t-g_t)^2$ under the constraint that g_t shares the same change point as $\hat{f}_t$. Then we have

$$\sum_{t=s}^{e}(f_t-\tilde{f}_t)^2\leqslant\sum_{t=s}^{e}(f_t-\hat{f}_t)^2 \tag{23}$$

Representing $\tilde{f}_t=\bar{f}+\langle f,\tilde{\psi}\rangle\tilde{\psi}$ for another vector $\tilde{\psi}\in\mathcal{F}$ and using expressions (22) and (23),

$$\begin{aligned}\sum_{t=s}^{e}\{(f_t-f_t^0)^2-(f_t-\hat{f}_t)^2\}&\leqslant\sum_{t=s}^{e}\{(f_t-f_t^0)^2-(f_t-\tilde{f}_t)^2\}=\langle f,\tilde{\psi}\rangle^2-\langle f,\psi^0\rangle^2\\&=(|\langle f,\tilde{\psi}\rangle|-|\langle f,\psi^0\rangle|)(|\langle f,\tilde{\psi}\rangle|+|\langle f,\psi^0\rangle|)\leqslant(|\langle f,\tilde{\psi}\rangle|-|\langle f,\psi^0\rangle|)|\langle f,\psi^0\rangle|.\end{aligned}$$

Since $|\langle f,\psi^0\rangle|=\mathbb{S}_{s,\eta,e}$ and $|\langle f,\tilde{\psi}\rangle|=\mathbb{S}_{s,\hat{\eta},e}$ with the distance between η and $\hat{\eta}$ being at least $c_0\varepsilon_T$, the above expression is bounded from above by $-(C\delta_T/\sqrt{T})(\varepsilon_T/\sqrt{T})=-C\varepsilon_T\delta_T/T$ from lemmas 1 and 2. Turning to term I, we can decompose it as

$$\sum_{t=s}^{e}\epsilon_t(\hat{f}_t-f_t^0)=\sum_{t=s}^{e}\epsilon_t(\hat{f}_t-\tilde{f}_t)+\sum_{t=s}^{e}\epsilon_t(\tilde{f}_t-f_t^0),$$

and each of the two sums is split into subsums computed over the intervals of constancy of $\hat{f}_t-\tilde{f}_t$ and $\tilde{f}_t-f_t^0$ respectively. Assume that $\hat{\eta}\geqslant\eta$; without loss of generality, we have

$$\sum_{t=s}^{e}\epsilon_t(\tilde{f}_t-f_t^0)=\left(\sum_{t=s}^{\eta}+\sum_{t=\eta+1}^{\hat{\eta}}+\sum_{t=\hat{\eta}+1}^{e}\right)\epsilon_t(\tilde{f}_t-f_t^0)\equiv\mathrm{III}+\mathrm{IV}+\mathrm{V}.$$

As $T\to\infty$, we have with probability tending to 1 (lemma 3)

$$\begin{aligned}|\mathrm{III}|&=\left|\frac{1}{\sqrt{(\eta-s+1)}}\sum_{t=s}^{\eta}\epsilon_t\right|\sqrt{(\eta-s+1)}\left|\frac{1}{\hat{\eta}-s+1}\sum_{t=s}^{\hat{\eta}}f_i-\frac{1}{\eta-s+1}\sum_{t=s}^{\eta}f_i\right|\\&\leqslant C\log(T)\sqrt{(\eta-s+1)}\frac{c_0\varepsilon_T}{\hat{\eta}-s+1}\leqslant C'\varepsilon_T\delta_T^{-1/2}\log(T).\end{aligned}$$

Term |V| is of the same order as |III| and similarly |IV| is bounded by $C\varepsilon_T^{1/2}\log(T)$.

As for $\Sigma_{t=s}^{e}\epsilon_t(\hat{f}_t-\tilde{f}_t)$, we have

$$\sum_{t=s}^{e}\epsilon_t(\hat{f}_t-\tilde{f}_t)=\left(\sum_{t=s}^{\hat{\eta}}+\sum_{t=\hat{\eta}+1}^{e}\right)\epsilon_t(\hat{f}_t-\tilde{f}_t)\equiv\mathrm{VI}+\mathrm{VII}.$$

Note that terms VI and VII are of the same order and, with probability converging to 1 as $T\to\infty$,

$$|\mathrm{VI}|=\frac{1}{\hat{\eta}-s+1}\left(\sum_{t=s}^{\hat{\eta}}\epsilon_t\right)^2=\log^2(T).$$

Putting together all these requirements, as long as

$$\frac{\varepsilon_T \delta_T}{T} > \{\varepsilon_T \delta_T^{-1/2} \log(T)\} \vee \{\varepsilon_T^{1/2} \log(T)\} \vee \log^2(T), \tag{24}$$

the dominance of term II over I holds and thus we prove the lemma.

From expression (24), it is derived that $\Theta > \frac{2}{3}$ and $\varepsilon_T > \delta_T^{-2} T^2 \log^2(T)$, i.e. letting $\varepsilon_T = \max\{T^\theta, \log^{2+\vartheta}(T)\}$, it is sufficient to have $\theta \geqslant 2 - 2\Theta$ and $\vartheta > 0$. Also the proofs of lemmas 5 and 6 require $\delta_T^{-1} T \log(T) \ll \sqrt{\varepsilon_T} \ll \pi_T \ll \delta_T \{T \log(T)\}^{-1/2}$, which is satisfied by $\theta = 2 - 2\Theta$ and $\pi_T = \kappa T^\gamma$ with any $\gamma \in (1 - \Theta, \Theta - \frac{1}{2})$ when $\Theta \in (\frac{3}{4}, 1)$, and by $\pi_T = \kappa \log^{1+\omega}(T)$ with any $\omega > \vartheta/2$ when $\Theta = 1$.

Lemma 5. Under conditions (15) and (16), we have

$$\mathbb{P}\left(|\mathbb{Y}_{s,b,e}| < \pi_T n^{-1} \sum_{t=s}^{e} Y_{t,T} \right) \to 0 \tag{25}$$

for $b = \arg\max_{t \in \mathcal{D}_{s,e}} |\mathbb{Y}_{s,t,e}|$, as $T \to \infty$.

Proof. Define the two events $\mathcal{A}$ and $\mathcal{B}$ as

$$\mathcal{A} = \left\{ |\mathbb{Y}_{s,b,e}| < \pi_T \frac{1}{n} \sum_{t=s}^{e} Y_{t,T} \right\},$$

$$\mathcal{B} = \left\{ \frac{1}{n} \left| \sum_{t=s}^{e} Y_{t,T} - \sum_{t=s}^{e} \sigma\left(\frac{t}{T}\right) \right| < \bar{\sigma} \equiv \frac{1}{2n} \sum_{t=s}^{e} \sigma\left(\frac{t}{T}\right) \right\}.$$

We can show that $\mathbb{P}(\mathcal{B}) \to 1$ as $T \to \infty$ by using the Bernstein inequality as in the proof of lemma 3 and that the rate of convergence is faster than that of expression (19). Hence $\mathbb{P}\{n^{-1} \Sigma_{t=s}^{e} Y_{t,T} \in (\bar{\sigma}/2, 3\bar{\sigma}/2)\} \to 1$. Since the probability in expression (25) is bounded from above by $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) + \mathbb{P}(\mathcal{B}^c)$, we need to show only that $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) \to 0$. From lemma 4, we have some $\eta \equiv \eta_{q_1+q}$ satisfying $|b - \eta| < c_0 \varepsilon_T$. Without loss of generality, let $\eta < b$ and define

$$\sigma_1 \equiv \sigma\left(\frac{\eta}{T}\right) \neq \sigma\left(\frac{\eta+1}{T}\right) \equiv \sigma_2.$$

From lemma 3, and conditions (15), (16) and 1, the following inequality holds with probability tending to 1 as $\gamma < \Theta - \frac{1}{2}$:

$$\begin{aligned}
|\mathbb{Y}_{s,b,e}| &\geqslant |\mathbb{S}_{s,b,e}| - \log(T) \\
&= \sqrt{\left\{ \frac{(b-s+1)(e-b)}{n} \right\}} \left| \frac{\sigma_1(\eta - s + 1) + \sigma_2(b - \eta)}{b - s + 1} - \sigma_2 \right| - \log(T) \\
&\geqslant \sqrt{\left\{ \frac{e-b}{n(b-s+1)} \right\}} \sigma_*(\eta - s + 1) - \log(T) \geqslant \sqrt{\left(\frac{1 - c_*}{n c_*} \right)} \sigma_*(\eta - s + 1) - \log(T) \\
&\geqslant \frac{C \delta_T}{c^* \sqrt{T}} - \log(T) > \pi_T \frac{3\bar{\sigma}}{2}.
\end{aligned}$$

Lemma 6. For some positive constants C and C', let s and e satisfy either

(a) $\exists 1 \leqslant q \leqslant N$ such that $s \leqslant \eta_q \leqslant e$ and $(\eta_q - s + 1) \wedge (e - \eta_q) \leqslant C \varepsilon_T$ or
(b) $\exists 1 \leqslant q \leqslant N$ such that $s \leqslant \eta_q < \eta_{q+1} \leqslant e$ and $(\eta_q - s + 1) \vee (e - \eta_{q+1}) \leqslant C' \varepsilon_T$.

Then, as $T \to \infty$,

$$\mathbb{P}\left(|\mathbb{Y}_{s,b,e}| > \pi_T n^{-1} \sum_{t=s}^{e} Y_{t,T} \right) \to 0 \tag{26}$$

for $b = \arg\max_{t \in \mathcal{D}_{s,e}} |\mathbb{Y}_{s,t,e}|$.

Proof. First we assume condition (a). We define the event $\mathcal{A}'$ as $\mathcal{A}' = \{|\mathbb{Y}_{s,b,e}| > \pi_T n^{-1} \Sigma_{t=s}^{e} Y_{t,T}\}$ and adopt the event $\mathcal{B}$ from the proof of lemma 5. Since $\mathbb{P}(\mathcal{B}) \to 1$, the probability in expression (26) is bounded from above by $\mathbb{P}(\mathcal{A}' \cap \mathcal{B}) + \mathbb{P}(\mathcal{B}^c)$ and it remains to show only that $\mathbb{P}(\mathcal{A}' \cap \mathcal{B}) \to 0$. Assuming that $\eta_q - s + 1 \leqslant C \varepsilon_T$ leads to $b > \eta_q \equiv \eta$, and, using the same notation as in lemma 5, we have

$$
\begin{aligned}
|\mathbb{Y}_{s,b,e}| &\leqslant |\mathbb{S}_{s,b,e}| + \log(T) \\
&\leqslant \sqrt{\left\{\frac{(b-s+1)(e-b)}{n}\right\}} \left| \frac{\sigma_1(\eta-s+1)+\sigma_2(b-\eta)}{b-s+1} - \sigma_2 \right| + \log(T) \\
&\leqslant \sqrt{\left\{\frac{e-b}{n(b-s+1)}\right\}} \times 2\sigma^*(\eta-s+1) + \log(T) \leqslant \sqrt{\left\{\frac{e-\eta}{n(\eta-s+1)}\right\}} \times 2\sigma^*(\eta-s+1) + \log(T) \\
&\leqslant 2\sigma^* \sqrt{(C\varepsilon_T)} + \log(T) < \pi_T \frac{\bar{\sigma}}{2}.
\end{aligned}
$$

The proof in the case of condition (b) takes similar arguments and thus lemma 6 follows. □

When applying the algorithm to a single sequence with N change points, lemmas 1–6 show the consistency of the algorithm as follows. At the start of the binary segmentation algorithm, we have $s=0$ and $e=T-1$, and thus all the conditions that are required by lemma 5 are met. Then the algorithm detects and locates a change point which is within the distance of $c_0\varepsilon_T$ from a true change point (lemma 4) such that any segments that are defined by the change points detected also satisfy the conditions in lemma 5, from the assumptions on the spread of η_q, $q=1,\ldots,N$, in condition 1. The algorithm iteratively proceeds in this manner until all the N change points have been detected and, since thus-determined segments meet either of the two conditions in lemma 6, change point detection is completed.

Now we turn our attention to the case of $d>1$ sequences and prove theorem 1. When necessary to highlight the dependence of $Y_{t,T}^{(k)}$ on k in deriving $\mathbb{S}_{s,t,e}$ and $\mathbb{Y}_{s,t,e}$, we use the notation $\mathbb{S}_{s,t,e}^{(k)}$ and $\mathbb{Y}_{s,t,e}^{(k)}$. The index set $\{1,\ldots,d\}$ is denoted by $\mathcal{K}$. From lemma 3, we have $\max_k \max_{(s,t,e)\in\mathcal{D}} |\mathbb{Y}_{s,t,e}^{(k)} - \mathbb{S}_{s,t,e}^{(k)}| \leqslant \log(T)$ with the probability bounded from below by $1 - CdT^3 \exp\{-C' \log^2(T)\} \to 1$ under condition 4 in Section 2. Therefore, the following arguments are made conditional on this event.

Let $\mathcal{K}_{s,e} \subset \mathcal{K}$ denote the index set corresponding to those $Y_{t,T}^{(k)}$ with at least one change point in $\sigma^{(k)}(t/T)$ on $t \in (s,e)$. Lemma 6 shows that $\mathcal{Y}_{s,t,e}^{(k)}, k \in \mathcal{K} \setminus \mathcal{K}_{s,e}$ do not pass the thresholding at any $t \in (s,e)$, i.e. $\mathcal{I}_{s,t,e}^{(k)} = \mathbb{I}(\mathcal{Y}_{s,t,e}^{(k)} > \pi_T) = 0$ for all $t \in (s,e)$. In contrast lemma 5 indicates that all $\mathcal{Y}_{s,t,e}^{(k)}, k \in \mathcal{K}_{s,e}$, survive after thresholding in the sense that $\mathcal{I}_{s,t,e}^{(k)} = 1$ over the intervals around the true change points. Besides, in Venkatraman (1992), each $\mathbb{S}_{s,t,e}^{(k)}$ is shown to be of the functional form $g^{(k)}(x) = \{x(1-x)\}^{-1/2}(\alpha_x^{(k)} x + \beta_x^{(k)})$ for $x = (t-s+1)/n \in (0,1)$, where $\alpha_x^{(k)}$ and $\beta_x^{(k)}$ are determined by the magnitude of the jumps at the change points of $\sigma^{(k)}(t/T)$ as well as their locations, and constant between any two adjacent change points. Note that scaling of $\mathbb{S}_{s,t,e}^{(k)}$ by $n^{-1}\sum_{t=s}^{e} Y_{t,T}^{(k)}$ scales the values of α_x and β_x only, and does not change the shape of $g^{(k)}(x)$. Each function $g^{(k)}(x)$

(a) is either monotonic or decreasing and then increasing on any interval that is defined by two adjacent change points of $\sigma^{(k)}(t/T)$ and
(b) achieves the maximum at one of the change points of $\sigma^{(k)}(t/T)$ in (s,e);

see lemma 2.2 of Venkatraman (1992). Since the pointwise summation of $g^{(k)}(\cdot)$ over $k \in \mathcal{K}_{s,e}$ takes the functional form $g(x) = \{x(1-x)\}^{-1/2}(\alpha_x x + \beta_x)$ which is identical to that of each individual $g^{(k)}(\cdot)$, it satisfies the above (a) and (b) as well.

Denoting $\bar{Y}_{s,e} = (1/n)\sum_{u=s}^{e} Y_{u,T}^{(k)}$, we decompose $\tilde{y}_t^{\mathrm{thr}}$ as

$$
\begin{aligned}
\frac{\tilde{y}_t^{\mathrm{thr}}}{\sum_{k\in\mathcal{K}} \mathcal{I}_{s,t,e}^{(k)}} &= \frac{\sum_{k\in\mathcal{K}} \bar{Y}_{s,e}^{-1} \mathbb{Y}_{s,t,e}^{(k)} \mathcal{I}_{s,t,e}^{(k)}}{\sum_{k\in\mathcal{K}} \mathcal{I}_{s,t,e}^{(k)}} \\
&= \frac{\sum_{k\in\mathcal{K}_{s,e}} \bar{Y}_{s,e}^{-1} \mathbb{S}_{s,t,e}^{(k)} \mathcal{I}_{s,t,e}^{(k)}}{|\mathcal{K}_{s,e}|} + \frac{\sum_{k\in\mathcal{K}_{s,e}} \bar{Y}_{s,e}^{-1} (\mathbb{Y}_{s,t,e}^{(k)} - \mathbb{S}_{s,t,e}^{(k)}) \mathcal{I}_{s,t,e}^{(k)}}{|\mathcal{K}_{s,e}|} = \mathrm{I} + \mathrm{II}
\end{aligned}
$$

where $\mathrm{II} \leqslant C\log(T)$ (lemma 3). Note that we can construct an additive model $y_t = f_t + \epsilon_t$ over $t \in [s,e]$ like that introduced in lemma 4, such that the CUSUM statistic of the piecewise constant signal f_t (i.e. $\mathbb{S}_{s,t,e}$ with f_t replacing $\sigma(t/T)$) is equal to $|\mathcal{K}_{s,e}|^{-1}\sum_{k\in\mathcal{K}_{s,e}} \bar{Y}_{s,e}^{-1} \mathbb{S}_{s,t,e}^{(k)}$. Since thresholding does not have any effect on the peak that is formed around the change points within the distance of $C\varepsilon_T$, I is of the same functional form as the CUSUM statistic of f_t in that region around the change points. Therefore, from lemma 4, $b = \arg\max_{t\in(s,e)} \tilde{y}_t^{\mathrm{thr}}$ satisfies $|b - \eta_q| < c_0\varepsilon_T$ for some $q = 1,\ldots,N$.

The SBS algorithm continues the change point detection procedure on the segments that are defined by

previously detected change points, which satisfy both condition (15) and condition (16) for at least one of $k \in \mathcal{K}$ until every change point is detected (as in the case of $d = 1$). Once all $\eta_1, \ldots, \eta_N$ have been identified, each of the resulting segments satisfies either condition (a) or condition (b) in lemma 6 for all $k \in \mathcal{K}$ such that the termination condition of the SBS algorithm (step 2(a)) is met.

For any $k \in \mathcal{K}_{s,e}$, a simple modification of the proof of lemma 2 leads to the existence of a positive constant C satisfying $\mathbb{S}^{(k)}_{s,t,e} > \pi_T$ for $|t - \eta| \leqslant C\varepsilon_T$, where η is any of the change points of $Y^{(k)}_{t,T}$ within (s, e) at which expression (18) is not equal to 0. Then, the corresponding $\mathbb{Y}^{(k)}_{s,t,e}$ is also greater than π_T within the distance of $\Delta_T \asymp \varepsilon_T$ from $b = \arg\max_{t \in [s,e]} \mathcal{Y}^{(k)}_{s,t,e}$, and hence the condition on the change point estimates in step 2(c) is justified with the choice of $\Delta_T = \lfloor \sqrt{T}/2 \rfloor$.

Appendix B: Multivariate locally stationary wavelet time series

The LSW model enables a timescale decomposition of a multivariate, possibly high dimensional process and thus permits a rigorous estimation of its second-order structure as shown in this section. The following conditions are imposed on the piecewise constant functions $W^{(j)}_i(z)$ and $\Sigma^{(j,l)}_i(k/T)$, as well as on the change points in the second-order structure for the p-variate LSW time series defined in definition 1.

Assumption 5. The following condition holds for each of the piecewise constant functions $W^{(j)}_i(z)$ and $\Sigma^{(j,l)}_i(z)$ for $j, l = 1, \ldots, p$, $i = -1, -2, \ldots$.

(a) Denoting by $L^{(j)}_i$ the total magnitude of jumps in $W^{(j)}_i(z)^2$, the variability of the functions $W^{(j)}_i(z)$, $i = -1, -2, \ldots$, is controlled such that $\Sigma^{-1}_{i=-I_T} 2^{-i} L^{(j)}_i = O\{\log(T)\}$ uniformly in j where $I_T = \lfloor \log(T) \rfloor$. Also, there is a positive constant $C > 0$ such that $|W^{(j)}_i(z)| \leqslant 2^{i/2} C$ uniformly over all $i \leqslant -1$ and $j = 1, \ldots, p$.

(b) Denoting the total magnitude of jumps in $\Sigma^{(j,l)}_i(z)$ by $R^{(j,l)}_i$, the variability of the functions $\Sigma^{(j,l)}_i(z)$, $i = -1, -2, \ldots$, is controlled such that $\Sigma^{-1}_{i=-I_T} 2^{-i} R^{(j,l)}_i = O\{\log(T)\}$ uniformly in $j \neq l$.

Assumption 6. Recall $\mathbb{B}$, the set of all change points in the second-order structure of $\mathbf{X}_{t,T}$ defined in expression (9). Then $\nu_r \in \mathbb{B}$, $r = 1, \ldots, N$, satisfy the conditions in assumption 1 in Section 2 in place of η_q, $q = 1, \ldots, N$.

The quantities that are of interest in modelling a multivariate LSW time series are the evolutionary wavelet spectrum and the evolutionary wavelet cross-spectrum, which are defined respectively as

$$S^{(j)}_i(z) = S^{(j,j)}_i(z) = W^{(j)}_i(z)^2 \qquad \text{for } j = 1, \ldots, p,$$
$$S^{(j,l)}_i(z) = W^{(j)}_i(z)\, W^{(l)}_i(z)\, \Sigma^{(j,l)}_i(z) \qquad \text{for } j \neq l, \quad j, l = 1, \ldots, p.$$

To study the connection between the evolutionary wavelet spectrum and the second-order structure of $\mathbf{X}_{t,T}$, we adopt the following quantities from Nason *et al.* (2000): with the same wavelet system as that used in the definition of $\mathbf{X}_{t,T}$, we define the auto-correlation wavelets as $\Psi_i(\tau) = \Sigma_k \psi_{i,k} \psi_{i,k+\tau}$, the cross-scale auto-correlation wavelets as $\Psi_{i,i'}(\tau) = \Sigma_k \psi_{i,k} \psi_{i',k+\tau}$ and the auto-correlation wavelet inner product matrix as $\mathbf{A} = (A_{i,i'})_{i,i'<0}$ with $A_{i,i'} = \Sigma_\tau \Psi_i(\tau) \Psi_{i'}(\tau) = \Sigma_\tau \Psi^2_{i,i'}(\tau) > 0$. Then, the *local* autocovariance and cross-covariance functions of $\mathbf{X}_{t,T}$ are defined as

$$c^{(j)}(z, \tau) = c^{(j,j)}(z, \tau) = \sum_{i=-\infty}^{-1} S^{(j)}_i(z)\, \Psi_i(\tau)$$

(from Nason *et al.* (2000)) and

$$c^{(j,l)}(z, \tau) = \sum_{i=-\infty}^{-1} S^{(j,l)}_i(z)\, \Psi_i(\tau)$$

(from Sanderson *et al.* (2010)).

Recalling the definition of time varying autocovariance and cross-covariance functions, the functions $c^{(j,l)}(z, \tau)$ and $c^{(j,l)}_T(z, \tau)$ are close to each other in the following sense.

Proposition 1. Under assumptions 5 and 6, $c^{(j,l)}_T(z, \tau)$ converges to $c^{(j,l)}(z, \tau)$ as

$$\frac{1}{T}\sum_{t=0}^{T-1}\left|c_T^{(j,l)}\left(\frac{t}{T},\tau\right)-c^{(j,l)}\left(\frac{t}{T},\tau\right)\right|=o(1) \tag{27}$$

for all $j,l=1,\ldots,p$.

The proof is provided in Appendix B.2. From equation (27), we can see that there is an asymptotic one-to-one relationship respectively between the evolutionary wavelet spectrum and the evolutionary wavelet cross-spectrum $S_i^{(j,l)}(z)$ and the autocovariance and cross-covariance functions $c_T^{(j,l)}(z,\tau),\tau\in\mathbb{Z}$, for all $j,l=1,\ldots,p$ such that, if there is a change point in $c_T^{(j,l)}(z,\tau)$ at some lag τ, at least one of the corresponding $S_i^{(j,l)}(z), i=-1,-2,\ldots$, has a change point at the same location z, and vice versa.

Furthermore, we can also show a one-to-one correspondence between the evolutionary wavelet spectrum and the evolutionary wavelet cross-spectrum and the wavelet periodograms and cross-periodograms respectively. Let $\beta_i^{(j,l)}(z)$ be a linear transformation of the evolutionary wavelet spectrum and the evolutionary wavelet cross-spectrum defined as $\beta_i^{(j,l)}(z)=\Sigma_{i'=-\infty}^{-1}S_{i'}^{(j,l)}(z)A_{i,i'}$. Then, the function $\beta_i^{(j,l)}(z)$ is piecewise constant with its change points corresponding to those of $\{S_{i'}^{(j,l)}(z)\}_{i'}$ owing to the invertibility of $\mathbf{A}$.

Proposition 2. Under assumptions 5 and 6, $\mathbb{E}[I_{i,t,T}^{(j,l)}]$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\left|\mathbb{E}[I_{i,t,T}^{(j,l)}]-\beta_i^{(j,l)}\left(\frac{t}{T}\right)\right|^2=2^{-i}\,O(T^{-1})+b_{i,T}^{(j,l)}, \tag{28}$$

where $b_{i,T}^{(j,l)}$ depends on the corresponding sequence $\{L_i^{(j)}\}_i$ or $\{R_i^{(j,l)}\}_i$.

The proof of equation (28) is a direct modification of that of proposition 2.1 of Fryzlewicz and Nason (2006) and thus has been omitted. In summary, from propositions 1 and 2 and the invertibility of $\mathbf{A}$, there is an asymptotic one-to-one correspondence between the autocovariance and cross-covariance functions and the expectations of wavelet periodograms and cross-periodograms respectively as noted in Section 3. Therefore, any change points in the autocovariance or cross-covariance functions are detectable by examining the corresponding wavelet periodogram or cross-periodogram sequences respectively.

B.1. Proof of theorem 2

From its construction, $\mathbb{E}[I_{i,t,T}^{(j,l)}]$ is piecewise constant and 'almost' satisfies conditions 1 and 3 in Section 2 in the sense that, for any change point ν in $\beta_i^{(j,l)}(t/T)$,

(a) $\mathbb{E}[I_{i,t,T}^{(j,l)}]$ is piecewise constant apart from the intervals $[\nu-2^{-i}K,\nu+2^{-i}K]$ for some $K>0$, where it shows smoother transitions, and
(b) $\mathbb{E}[I_{i,t,T}^{(j,l)}]$ has at least one change point within the intervals $[\nu-2^{-i}K,\nu+2^{-i}K]$, such that $|\mathbb{E}[I_{i,t_1,T}^{(j,l)}]-\mathbb{E}[I_{i,t_2,T}^{(j,l)}]|$ is bounded away from zero for $t_1=\nu-2^{-i}K-1$ and $t_2=\nu+2^{-i}K+1$.

Note that conditions (a) and (b) also hold for $\mathbb{E}[\tilde{I}_{i,t,T}^{(j,l)}]$ for $j\neq l$ defined as in expression (13). To accommodate these features of $I_{i,t,T}^{(j)}$ and $\tilde{I}_{i,t,T}^{(j,l)}$, we propose a modification of the multiplicative model (1):

$$\tilde{Y}_{t,T}^{(k)}=\sigma_{t,T}^{(k)}\tilde{Z}_{t,T}^{(k)2},\qquad t=0,\ldots,T-1,\quad k=1,\ldots,d. \tag{29}$$

The difference between the two models (1) and (29) comes from the function $\mathbb{E}[\tilde{Y}_{t,T}^{(k)}]=\sigma_{t,T}^{(k)}$ which is close to a piecewise constant function $\sigma^{(k)}(t/T)$ as $\mathbb{E}[I_{i,t,T}^{(j,l)}]$ is close to $\beta_i^{(j,l)}(z)$ (see equation (28)).

We also adapt assumptions 2–4 in Section 2, to the multivariate time series set-up and denote their analogues in this setting by assumptions 7–9. The latter assumptions are imposed on $I_{i,t,T}^{(j)}, j=1,\ldots,p$, and $\tilde{I}_{i,t,T}^{(j,l)}, j\neq l, j,l=1,\ldots,p$, at scales $i=-1,\ldots,I_T^*$, using the representation of these quantities as in expression (29) (the notation below refers to that representation).

Assumption 7. $\{\tilde{Z}_{t,T}^{(k)}\}_{t=0}^{T-1}$ is a sequence of standard normal variables and $\max_k\phi_\infty^{(k)1}<\infty$, where $\phi^{(k)}(\tau)=\sup_{t,T}|\mathrm{corr}(\tilde{Z}_{t,T}^{(k)},\tilde{Z}_{t+\tau,T}^{(k)})|$ and $\phi_\infty^{(k)r}=\Sigma_\tau|\phi^{(k)}(\tau)|^r$.

Assumption 8. There are positive constants $\sigma^*,\sigma_*>0$ such that $\{\max_{k,t,T}\sigma^{(k)}(t/T)\vee\max_{k,t,T}\sigma_{t,T}^{(k)}\}\leqslant\sigma^*$, and, given any change point η_q in $\sigma^{(k)}(t/T)$, we have $|\sigma^{(k)}\{(\eta_q+1)/T\}-\sigma^{(k)}(\eta_q/T)|>\sigma_*$ uniformly for all k.

Assumption 9. p and T satisfy $p^2T^{-\log(T)}\to 0$.

The following proposition shows that applying the SBS algorithm to $\tilde{Y}_{t,T}^{(k)}$ instead of $Y_{t,T}^{(k)}$ also leads to consistent change point estimates $\tilde{\eta}_q$, $q=1,\ldots,N$.

Proposition 3. Under assumptions 1, 4, 5 and 6, letting $\Delta_T \asymp \varepsilon_T$, we have that $\tilde{\eta}_q$, $q=1,\ldots,\tilde{N}$, satisfy

$$\mathbb{P}\{\tilde{N}=N; |\tilde{\eta}_q-\eta_q|<C_3\varepsilon_T \text{ for } q=1,\ldots,N\}\to 1$$

as $T\to\infty$ for some $C_3>0$, where ε_T and π_T are identical to those in theorem 1.

For a proof, see Appendix B.3.

We are now ready to prove theorem 2. Proposition 3 implies that the SBS algorithm is consistent in detecting change points from wavelet periodograms and cross-periodograms at a single scale i, i.e. all the change points that are detectable from scale i are identified by applying the SBS algorithm to the wavelet periodograms and cross-periodograms at the same scale. Besides, coupled with assumption 8, the condition on the magnitude of $|W_i^{(j)}(z)|$ in assumption 5 implies that, for each change point, the finest scale i at which it is detected satisfies $i \geqslant I_T^* = -\lfloor \alpha \log\{\log(T)\}\rfloor$. Suppose that $t=\nu$ is a change point in $S_i^{(j)}(t/T)$ which can be detected only at the scales that are coarser than I_T^*. Then, the corresponding jump in $\beta_i^{(j)}(t/T)$ is of the magnitude bounded from above by

$$\left|\sum_{i'=-\infty}^{I_T^*-1}\left\{S_i^{(j)}\left(\frac{\nu+1}{T}\right)-S_i^{(j)}\left(\frac{\nu}{T}\right)\right\}A_{i,i'}\right| \leqslant C\sum_{i'=-\infty}^{I_T^*-1}2^{i'}A_{i,i'}\to 0,$$

since $A_{i,i'}>0$ and $\Sigma_{i'=-\infty}^{-1}2^{i'}A_{i,i'}=1$ from Fryzlewicz *et al.* (2003), and thus assumption 8 is violated. We conclude that there exists I_T^* such that all the change points ν_r, $r=1,\ldots,N$, are detectable from examining the scales $i=-1,\ldots,I_T^*$. Since the SBS-MVTS algorithm repeatedly applies the SBS algorithm to the finest $|I_T^*|$-scales, its consistency with the required rates is a simple consequence of the argument about across-scales post-processing from Cho and Fryzlewicz (2012).

B.2. Proof of proposition 1

Let $t=\lfloor zT\rfloor$. Then we have

$$\begin{aligned}c_T^{(j)}(z,\tau) &= \mathbb{E}\left\{\sum_{i=-\infty}^{-1}\sum_{k=-\infty}^{\infty}W_i^{(j)}\left(\frac{k}{T}\right)\psi_{i,t-k}\xi_{i,k}^{(j)}\sum_{i'=-\infty}^{-1}\sum_{k'=-\infty}^{\infty}W_{i'}^{(j)}\left(\frac{k'}{T}\right)\psi_{i',t+\tau-k'}\xi_{i',k'}^{(j)}\right\}\\ &= \sum_{i=-\infty}^{-1}\sum_{k=-\infty}^{\infty}S_i^{(j)}\left(\frac{k}{T}\right)\psi_{i,t-k}\psi_{i,t+\tau-k}.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left|c_T^{(j)}(z,\tau)-c^{(j)}(z,\tau)\right| &\leqslant \frac{1}{T}\sum_{t=0}^{T-1}\left|\sum_{i=-\infty}^{-1}\sum_{k=-\infty}^{\infty}\left\{S_i^{(j)}\left(\frac{k}{T}\right)-S_i^{(j)}\left(\frac{t}{T}\right)\right\}\psi_{i,t-k}\psi_{i,t+\tau-k}\right|\\ &= \frac{1}{T}\sum_{t=0}^{T-1}\left|\sum_{i=-J_T}^{-1}+\sum_{i=-\infty}^{-J_T-1}\left[\sum_{k=-\infty}^{\infty}\left\{S_i^{(j)}\left(\frac{k}{T}\right)-S_i^{(j)}\left(\frac{t}{T}\right)\right\}\psi_{i,t-k}\psi_{i,t+\tau-k}\right]\right|\\ &\equiv \mathrm{I}+\mathrm{II},\end{aligned}$$

where the cut-off index is set as $J_T=\varrho\log_2(T)$ for $\varrho\in(0,1)$. For all $i=-1,\ldots,-J_T$, the length of support of $\psi_{i,t-k}\psi_{i,t+\tau-k}$ is bounded from above by $2^{J_T}K$ uniformly for some $K>0$. Therefore, the summands of term I are equal to 0 except for those t which are within the distance of $2^{J_T}K$ from any change point of $S_i^{(j)}(z), i=-1,\ldots,-J_T$. Then, from assumptions 5 and 6, term I is bounded by

$$\begin{aligned}\frac{2^{J_T}NK}{T}\left|\sum_{i=-J_T}^{-1}L_i^{(j)}\sum_{k=-\infty}^{\infty}\psi_{i,t-k}\psi_{i,t+\tau-k}\right| &= \frac{2^{J_T}NK}{T}\left|\sum_{i=-J_T}^{-1}L_i^{(j)}\Psi_i(\tau)\right|\\ &\leqslant \frac{2^{J_T}NK}{T}\left|\sum_{i=-J_T}^{-1}L_i^{(j)}\right| = O\left\{\frac{2^{J_T}N\log(T)}{T}\right\},\end{aligned} \tag{30}$$

where the first inequality comes from the fact that $\Psi_i(\tau)=O(1)$ uniformly in τ. Term II is bounded by

$$T^{-1}\sum_{t=0}^{T-1}\left|\sum_{i=-\infty}^{-J_T-1}L_i^{(j)}\Psi_i(\tau)\right|\leqslant\left|\sum_{i=-\infty}^{-J_T-1}L_i^{(j)}\right|.$$

Owing to the bound imposed on $W_i^{(j)}(z)$, assumption 5 implies that $|L_i^{(j)}|\leqslant 2^iC$ and therefore $\Sigma_{i=-\infty}^{-J_T}L_i^{(j)}\leqslant 2^{-J_T}C\to 0$, and combined with expression (30) we have $\mathrm{I}+\mathrm{II}=o(1)$.

As for the relationship between $c_T^{(j,l)}(z,\tau)$ and $c^{(j,l)}(z,\tau)$, we note that assumptions 5 and 6 imply that the total magnitude of jumps in $S_i^{(j,l)}(z,\tau)=W_i^{(j)}(z,\tau)\,W_i^{(l)}(z,\tau)\,\Sigma_i^{(j,l)}(z,\tau)$ is bounded from above by $K'\max(L_i^{(j)},L_i^{(l)},R_i^{(j,l)})$ for some positive constant K'. Then the proof follows exactly the same arguments as above and thus has been omitted.

B.3. Proof of proposition 3

Proposition 3 can be proved by showing that any change point in $\sigma^{(k)}(t/T)$ is detectable from $\tilde{Y}_{t,T}^{(k)}$ within the distance of $O(\varepsilon_T)$. Let $\sigma^{(k)}(t/T)$ have a change point at $t=\eta_{q_1+q}\equiv\eta$ within a segment (s,e), where s,η_{q_1+q} and e satisfy conditions (15) and (16). From the compactness of wavelet vector ψ_i, the sequence $\sigma_{t,T}^{(k)}$ also has a change point within the interval containing η, such that there exists $\bar{\eta}\in[\eta-2^{-I_T^*}K,\eta+2^{-I_T^*}K]$ where $|\sigma_{\bar{\eta},T}^{(k)}-\sigma_{\bar{\eta}+1,T}^{(k)}|>0$ although it may not be unique. Let $\bar{\eta}<\eta$. Since such $\bar{\eta}$ still satisfies condition (15) in place of η, proposition 2 implies that

$$\begin{aligned}
&\left\{\left|\frac{1}{\bar{\eta}-s+1}\sum_{t=s}^{\bar{\eta}}\sigma_{t,T}^{(k)}-\frac{1}{e-\bar{\eta}}\sum_{t=\bar{\eta}+1}^{e}\sigma_{t,T}^{(k)}\right|-\left|\frac{1}{\eta-s+1}\sum_{t=s}^{\eta}\sigma^{(k)}\left(\frac{t}{T}\right)-\frac{1}{e-\eta}\sum_{t=\eta+1}^{e}\sigma^{(k)}\left(\frac{t}{T}\right)\right|\right\}^2\\
&\qquad\leqslant C\delta_T^{-2}\left|\sum_{t=s}^{\bar{\eta}}\left\{\sigma_{t,T}^{(k)}-\sigma^{(k)}\left(\frac{t}{T}\right)\right\}-\sum_{t=\eta+1}^{e}\left\{\sigma_{t,T}^{(k)}-\sigma^{(k)}\left(\frac{t}{T}\right)\right\}-\sum_{t=\bar{\eta}+1}^{\eta}\left\{\sigma_{t,T}^{(k)}+\sigma^{(k)}\left(\frac{t}{T}\right)\right\}\right|^2\\
&\qquad\leqslant C\delta_T^{-1}\sum_{t=s}^{e}\left|\sigma_{t,T}^{(k)}-\sigma^{(k)}\left(\frac{t}{T}\right)\right|^2+2^{-2I_T^*}C'\delta_T^{-2}\sigma^{*2}\to 0,
\end{aligned}$$

i.e. the CUSUM statistics that are computed from $\tilde{Y}_{t,T}^{(k)}$ are of the same order as those from $Y_{t,T}^{(k)}$ around $t=\eta$. Therefore, the arguments that were used in lemmas 1–5 also apply to $\tilde{Y}_{t,T}^{(k)}$, and $\hat{\eta}=\arg\max_{t\in(s,e)}\tilde{\mathbb{Y}}_{s,t,e}$ satisfies $|\hat{\eta}-\bar{\eta}|\leqslant c_0\varepsilon_T$. Then, with $I_T^*=-\lfloor\alpha\log\{\log(T)\}\rfloor$, we have $|\hat{\eta}-\eta|\leqslant|\hat{\eta}-\bar{\eta}|+|\bar{\eta}-\eta|\leqslant c_0\varepsilon_T+C\log^{2+\vartheta}(T)$. Besides, once a change point has been detected within such an interval, condition (6) in step 2(b) does not allow any more change points to be detected too close to previously detected change points, and therefore any $t\in[\eta-2^{-I_T^*}K,\eta+2^{-I_T^*}K]$ is disregarded from future change point detection. Hence, despite the bias between $\sigma_{t,T}^{(k)}$ and $\sigma^{(k)}(t/T)$, the consistency of the SBS algorithm still holds for $\tilde{Y}_{t,T}^{(k)}$ in place of $Y_{t,T}^{(k)}$.

References

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009) Break detection in the covariance structure of multivariate time series models. *Ann. Statist.*, **37**, 4046–4087.

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. New York: Springer.

Chen, J. and Gupta, A. K. (1997) Testing and locating variance change-points with application to stock prices. *J. Am. Statist. Ass.*, **92**, 739–747.

Cho, H. and Fryzlewicz, P. (2011) Multiscale interpretation of taut string estimation and its connection to Unbalanced Haar wavelets. *Statist. Comput.*, **21**, 671–681.

Cho, H. and Fryzlewicz, P. (2012) Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statist. Sin.*, **22**, 207–229.

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

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2008) Break detection for a class of nonlinear time series models. *J. Time Ser. Anal.*, **29**, 834–867.

Dwivedi, Y. and Subba Rao, S. (2011) A test for second-order stationarity of a time series based on the discrete Fourier transform. *J. Time Ser. Anal.*, **32**, 68–91.

Fan, J., Lv, J. and Qi, L. (2011) Sparse high-dimensional models in economics. *Ann. Rev. Econ.*, **3**, 291–317.

Fryzlewicz, P. (2005) Modelling and forecasting financial log-returns as locally stationary wavelet processes. *J. App. Statist.*, **32**, 503–528.

Fryzlewicz, P. (2014) Wild Binary Segmentation for multiple change-point detection. *Ann. Statist.*, to be published.
Fryzlewicz, P. and Nason, G. (2006) Haar–Fisz estimation of evolutionary wavelet spectra. *J. R. Statist. Soc.* B, **68**, 611–634.
Fryzlewicz, P., Van Bellegem, S. and von Sachs, R. (2003) Forecasting non-stationary time series by wavelet process modelling. *Ann. Inst. Statist. Math.*, **55**, 737–764.
Groen, J., Kapetanios, G. and Price, S. (2013) Multivariate methods for monitoring structural change. *J. Multiv. Time Ser.*, **28**, 250–274.
Horváth, L. and Hušková, M. (2012) Change-point detection in panel data. *J. Time Ser. Anal.*, **33**, 631–648.
Inclán, C. and Tiao, G. C. (1994) Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Am. Statist. Ass.*, **89**, 913–923.
Jentsch, C. and Subba Rao, S. (2013) A test for second order stationarity of a multivariate time series. *Preprint.* (Available from `http://www.stat.tamu.edu/~suhasini/papers/multivariate_test_stationarity_revisionR1.pdf`.)
Johnson, N. and Kotz, S. (1970) *Distributions in Statistics: Continuous Univariate Distributions*, vol. 1. Boston: Houghton Mifflin.
Korostelev, A. (1987) On minimax estimation of a discontinuous signal. *Theor. Probab. Appl.*, **32**, 727–730.
Lavielle, M. and Moulines, E. (2000) Least-squares estimation of an unknown number of shifts in a time series. *J. Time Ser. Anal.*, **21**, 33–59.
Lavielle, M. and Teyssière, G. (2006) Detection of multiple change-points in multivariate time series. *Lith. Math. J.*, **46**, 287–306.
Mikosch, T. and Stărică, C. (2004) Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Rev. Econ. Statist.*, **86**, 378–390.
Nason, G. P. (2013) A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *J. R. Statist. Soc.* B, **75**, 879–904.
Nason, G. P., von Sachs, R. and Kroisandt, G. (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Statist. Soc.* B, **62**, 271–292.
Nason, G. P. and Silverman, B. W. (1995) *The Stationary Wavelet Transform and Some Statistical Applications*, pp. 281–300. New York: Springer.
Ombao, H. C., Raz, J. A., von Sachs, R. and Guo, W. (2002) The SLEX model of a non-stationary random process. *Ann. Inst. Statist. Math.*, **54**, 171–200.
Ombao, H. C., Raz, J. A., von Sachs, R. and Malow, B. A. (2001) Automatic statistical analysis of bivariate nonstationary time series. *J. Am. Statist. Ass.*, **96**, 543–560.
Ombao, H., von Sachs, R. and Guo, W. (2005) SLEX analysis of multivariate nonstationary time series. *J. Am. Statist. Ass.*, **100**, 519–531.
Sanderson, J., Fryzlewicz, P. and Jones, M. (2010) Estimating linear dependence between non-stationary time series using the locally stationary wavelet model. *Biometrika*, **97**, 435–446.
Venkatraman, E. S. (1992) Consistency results in multiple change-point problems. *Technical Report 24.* Department of Statistics, Stanford University, Stanford.
Vert, J. and Bleakley, K. (2010) Fast detection of multiple change-points shared by many signals using group LARS. *Adv. Neur. Inform. Process. Syst.*, **23**, 2343–2351.
Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. New York: Wiley.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.