

Detecting linear trend changes and point anomalies in data sequences

Hyeyoung Maeng and Piotr Fryzlewicz
Department of Statistics, London School of Economics

May 30, 2019

Abstract

We propose TrendSegment, a methodology for detecting multiple change-points corresponding to linear trend changes or point anomalies in one dimensional data. A core ingredient of TrendSegment is a new Tail-Greedy Unbalanced Wavelet transform: a conditionally orthonormal, bottom-up transformation of the data through an adaptively constructed unbalanced wavelet basis, which results in a sparse representation of the data. The bottom-up nature of this multiscale decomposition enables the detection of point anomalies and linear trend changes at once as the decomposition focuses on local features in its early stages and on global features next. To reduce the computational complexity, the proposed method merges multiple regions in a single pass over the data. We show the consistency of the estimated number and locations of change-points. The practicality of our approach is demonstrated through simulations and two real data examples, involving Iceland temperature data and sea ice extent of the Arctic and the Antarctic. Our methodology is implemented in the R package `trendsegmentR`, available from CRAN.

Keywords: change-point detection; bottom-up algorithms; piecewise-linear signal; anomaly detection; wavelets

1 Introduction

Multiple change-point detection is a problem of importance in many applications; recent examples include automatic detection of change-points in cloud data to maintain the performance and availability of an app or a website (James et al., 2016), climate change detection in tropical cyclone records (Robbins et al., 2011), detecting exoplanets from light curve data (Fisch et al., 2018), detecting changes in the DNA copy number (Olshen et al., 2004; Jeng et al., 2012; Bardwell et al., 2017), estimation of stationary intervals in potentially cointegrated stock prices (Matteson et al., 2013), estimation of change-points in multi-subject fMRI data (Robinson et al., 2010) and detecting changes in vegetation trends (Jamali et al., 2015).

This paper considers the change-point model

$$X_t = f_t + \varepsilon_t, \quad t = 1, \dots, T, \quad (1)$$

where f_t is a deterministic and piecewise-linear signal containing N change-points, i.e. time indices at which the slope and/or the intercept in f_t undergoes changes. These changes occur at unknown locations $\eta_1, \eta_2, \dots, \eta_N$. The ε_t 's are iid random errors following the normal distribution with mean zero and variance σ^2 . Both continuous and discontinuous changes in the linear trend are permitted. A point anomaly can be viewed as a separate data segment containing only one data point. Therefore, if f_{η_i} is a point anomaly, then the two consecutive change-points that define it, η_{i-1} and η_i , are linked via $\eta_{i-1} = \eta_i - 1$ under the definition of a change-point specified later in (15). Our main interest is in the estimation of N and $\eta_1, \eta_2, \dots, \eta_N$ under some assumptions that quantify the difficulty of detecting each η_i ; therefore, our aim is to segment the data into sections of linearity and/or point anomalies in f_t . In particular, a point anomaly can only be detected when it has a large enough jump size with respect to the signal levels to its right and left, while a change-point capturing a small size of linear trend change requires a longer distance from its adjacent change-points to be detected. Detecting both linear trend changes and point anomalies

is an important applied problem in a variety of fields, including climate change, as illustrated in Section 5.

The change-point detection procedure proposed in this paper is referred to as TrendSegment; it is designed to work well in detecting not only long trend segments and point anomalies, but also short trend segments that are not necessarily classified as point anomalies. The engine underlying TrendSegment is a new Tail-Greedy Unbalanced Wavelet (TGUW) transform: a conditionally orthonormal, bottom-up transformation for univariate data sequences through an adaptively constructed unbalanced wavelet basis, which results in a sparse representation of the data. In this article, we show that TrendSegment offers good performance in estimating the number and locations of change-points across a wide range of signals containing constant and/or linear segments and/or point anomalies. TrendSegment is also shown to be statistically consistent and computationally efficient.

In earlier related work regarding linear trend changes, Bai and Perron (1998) consider the estimation of linear models with multiple structural changes by least-squares and present Wald-type tests for the null hypothesis of no change. Kim et al. (2009) and Tibshirani et al. (2014) consider ‘trend filtering’ with the L_1 penalty and Maidstone et al. (2017) detect changes in the slope with an L_0 regularisation via a dynamic programming algorithm. Spiriti et al. (2013) study two algorithms for optimising the knot locations in least-squares and penalised splines. Baranowski et al. (2016) propose a multiple change-point detection device termed Narrowest-Over-Threshold (NOT), which focuses on the narrowest segment among those whose contrast exceeds a pre-specified threshold. Anastasiou and Fryzlewicz (2018) propose the Isolate-Detect (ID) approach which continuously searches expanding data segments for changes.

Keogh et al. (2004) mention that sliding windows, top-down and bottom-up approaches are three principal categories which most time series segmentation algorithms can be grouped into. Keogh et al. (2004) apply those three approaches to the detection of changes in linear trends (but not point anomalies) in 10 different signals and discover that the performance of bottom-

up methods is better than that of top-down methods and sliding windows, notably when the underlying signal has jumps, sharp cusps or large fluctuations. Bottom-up procedures have rarely been used in change-point detection. Matteson and James (2014) use an agglomerative algorithm for hierarchical clustering in the context of change-point analysis. Keogh et al. (2004) merge adjacent segments of the data according to a criterion involving the minimum residual sum of squares (RSS) from a linear fit, until the RSS falls under a certain threshold; but the lack of precise recipes for the choice of this threshold parameter causes the performance of this method to be somewhat unstable, as we report in Section 4.

As illustrated later in this paper, many existing change-point detection methods for the piecewise-linear model fail in signals that include frequent change-points or abrupt local features. The TGUW transform, which underlies TrendSegment, is able to handle scenarios involving possibly frequent change-points. It constructs, in a bottom-up way, an adaptive wavelet basis by consecutively merging neighbouring segments of the data starting from the finest level (throughout the paper, we refer to a wavelet basis as adaptive if it is constructed in a data-driven way). This enables it to identify local features at an early stage, before it proceeds to focus on more global features corresponding to longer data segments.

Fryzlewicz (2018) introduces the Tail-Greedy Unbalanced Haar (TGUH) transform, a bottom-up, agglomerative, data-adaptive transformation of univariate sequences that facilitates change-point detection in the piecewise-constant sequence model. The current paper extends this idea to adaptive wavelets other than adaptive Haar, which enables change-point detection in the piecewise-linear model (and, in principle, to higher-order piecewise polynomials, but we do not pursue this in the current work). We emphasise that this extension from TGUH to TGUW is both conceptually and technically non-trivial, due to the fact that it is not a priori clear how to construct a suitable wavelet basis in TGUW for wavelets other than adaptive Haar; this is due to the non-uniqueness of the local orthonormal matrix transformation for performing each merge in TGUW, which does not occur in TGUH. We solve this issue by imposing certain guiding principles in

the way the merges are performed, which enables the detection of changes in the linear trend and point anomalies. The TGUW transform is fast and its computational cost is the same as that of TGUH. Important properties of the TGUW transform include orthonormality conditional on the merging order, nonlinearity and “tail-greediness”, and will be investigated in Section 2. The TGUW transform is the first step of our proposed TrendSegment procedure, which involves four steps.

The detection of point anomalies has been widely studied in both time series and machine learning literature and the reader is referred to Chandola et al. (2009) for an extensive review. Our framework is different from a model recently studied by Fisch et al. (2018) in that our focus is on linear trend changes and point anomalies, while they do not focus on trends but only on point and collective anomalies with respect to a constant baseline distribution.

The remainder of the article is organised as follows. Section 2 gives a full description of the TrendSegment procedure and the relevant theoretical results are presented in Section 3. The supporting simulation studies are described in Section 4 and our methodology is illustrated in Section 5 through climate datasets. The proofs of our main theoretical results are in Appendix A and Section A of the supplementary materials, and the geometric interpretation of the TGUW transformation can be found in Section D of the supplementary materials. The TrendSegment procedure is implemented in the R package `trendsegmentR`.

2 Methodology

2.1 Summary of TrendSegment

The TrendSegment procedure for estimating the number and the locations of change-points includes four steps. We give a broad picture first and outline details in later sections.

1. *TGUW transformation.* Perform the TGUW transform; a bottom-up unbalanced adaptive wavelet transformation of the input data X_1, \dots, X_T by recursively applying local conditionally orthonormal transformations. This produces a data-adaptive multiscale decomposition of the data with $T - 2$ detail-type coefficients and 2 smooth coefficients. The resulting conditionally orthonormal transform of the data hopes to encode most of the energy of the signal in only a few detail-type coefficients arising at coarse levels. This sparse representation of the data justifies thresholding in the next step.
2. *Thresholding.* Set to zero those detail coefficients whose magnitude is smaller than a pre-specified threshold as long as all the non-zero detail coefficients are connected to each other in the tree structure. This step performs “pruning” as a way of deciding the significance of the sparse representation obtained in step 1.
3. *Inverse TGUW transformation.* Obtain an initial estimate of f_t by carrying out the inverse TGUW transformation of the thresholded coefficient tree. The resulting estimator can be shown to be l_2 -consistent, but not yet consistent for N or η_1, \dots, η_N .
4. *Post-processing.* Post-process the estimate from step 3 by removing some change-points perceived to be spurious, which enables us to achieve estimation consistency for N and η_1, \dots, η_N .

We devote the following four sections to describing each step above in order.

2.2 TGUW transformation

2.2.1 Key principles of the TGUW transform

In the initial stage, the data are considered smooth coefficients and the TGUW transform iteratively updates the sequence of smooth coefficients by merging the adjacent sections of the data which are the most likely to belong to the same segment. The merging is done by performing an adaptively constructed orthonormal transformation to the chosen triplet of the smooth coefficients

and in doing so, a data-adaptive unbalanced wavelet basis is established. The TGUW transform is completed after $T - 2$ such orthonormal transformations and each merge is performed under the following principles.

1. In each merge, three adjacent smooth coefficients are selected and the orthonormal transformation converts those three values into one detail and two (updated) smooth coefficients. The size of the detail coefficient gives information about the strength of the local linearity and the two updated smooth coefficients are associated with the estimated parameters (intercept and slope) of the local linear regression performed on the raw observations corresponding to the initially chosen three smooth coefficients.
2. “*Two together*” rule. The two smooth coefficients returned by the orthonormal transformation are paired in the sense that both contain information about one local linear regression fit. Thus, we require that any such pair of smooth coefficients cannot be separated in any subsequent merges. We refer to this recipe as the “two together” rule.
3. To decide which triplet of smooth coefficients should be merged next, we compare the corresponding detail coefficients as their magnitude represents the strength of the corresponding local linear trend; the smaller the (absolute) size of the detail, the smaller the local deviation from linearity. Smooth coefficients corresponding to the smallest detail coefficients have priority in merging.

As merging continues under the “two together” rule, all mergings can be classified into one of three forms, Type 1: merging three initial smooth coefficients, Type 2: merging one initial and a paired smooth coefficient and Type 3: merging two sets of (paired) smooth coefficients (this is composed of two merges of triplets; more details are given later).

2.2.2 Example

We now provide a simple example of the TGUW transformation; the accompanying illustration is in Figure 1. The notation for this example and for the general algorithm introduced later is in

Table 1: Notation. See Section 2.2.3 for formulae for the terms listed.

X_p	p^{th} element of the observation vector $\mathbf{X} = \{X_1, X_2, \dots, X_T\}^T$.
$s_{p,p}^0$	p^{th} initial smooth coefficient of the vector \mathbf{s}^0 where $\mathbf{X} = \mathbf{s}^0$.
$d_{p,q,r}$	detail coefficient obtained from $\{X_p, \dots, X_r\}$ (merges of Types 1 or 2).
$s_{p,r}^1, s_{p,r}^2$	smooth coefficients obtained from $\{X_p, \dots, X_r\}$, paired under the “two together” rule.
$d_{p,q,r}^1, d_{p,q,r}^2$	paired detail coefficients obtained by merging two adjacent subintervals, $\{X_p, \dots, X_q\}$ and $\{X_{q+1}, \dots, X_r\}$, where $r > q + 2$ and $q > p + 1$ (merge of Type 3).
\mathbf{s}	data sequence vector containing the (recursively updated) smooth and detail coefficients from the initial input \mathbf{s}^0 .

Table 1. This example shows single merges at each pass through the data. We will later generalise it to multiple passes through the data, which will speed up computation (this device is referred to as “tail-greediness”). We refer to j^{th} pass through the data as scale j . Assume that we have the initial input $\mathbf{s}^0 = (X_1, X_2, \dots, X_8)$, so that the complete TGUW transform consists of 6 merges. We show 6 example merges one by one under the rules introduced in Section 2.2.1. This example demonstrates all three possible types of merges.

Scale $j = 1$. From the initial input $\mathbf{s}^0 = (X_1, \dots, X_8)$, we consider 6 triplets (X_1, X_2, X_3) , (X_2, X_3, X_4) , (X_3, X_4, X_5) , (X_4, X_5, X_6) , (X_5, X_6, X_7) , (X_6, X_7, X_8) and compute the size of the detail for each triplet, where the formula can be found in (2). Suppose that (X_2, X_3, X_4) gives the smallest size of detail, $|d_{2,3,4}|$, then merge (X_2, X_3, X_4) through the orthogonal transformation formulated in (4) and update the data sequence into $\mathbf{s} = (X_1, s_{2,4}^1, s_{2,4}^2, d_{2,3,4}, X_5, X_6, X_7, X_8)$. We categorise this transformation into Type 1 (merging three initial smooth coefficients).

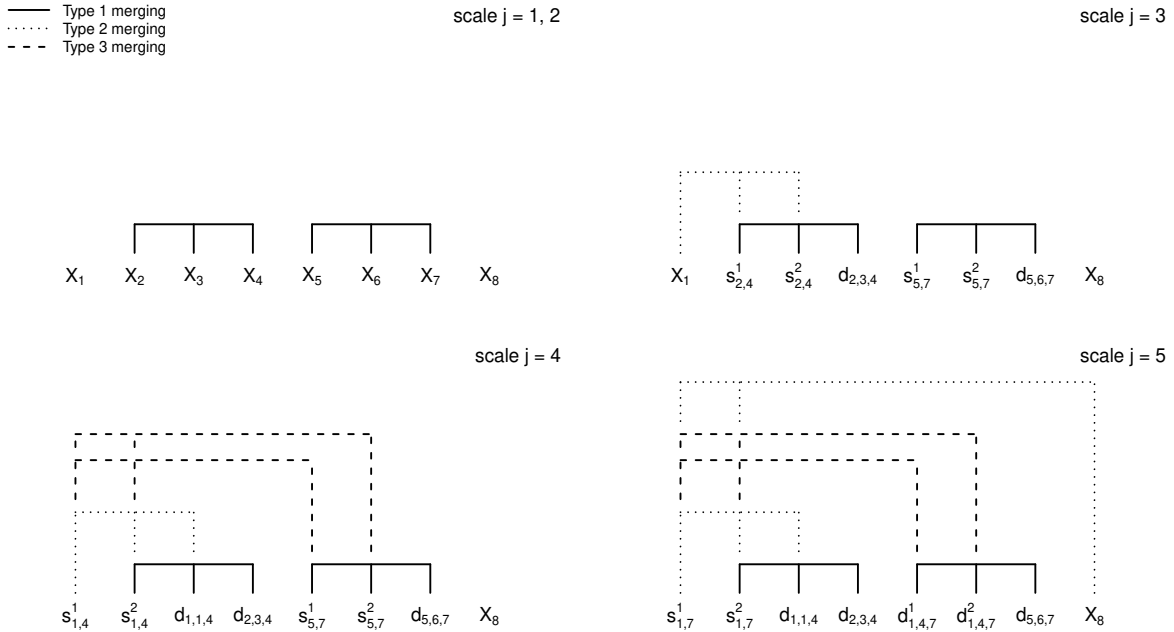


Figure 1: Construction of tree for the example in Section 2.2.2; each diagram shows all merges performed up to the given scale.

Scale $j = 2$. From now on, the “two together” rule is applied. Ignoring any detail coefficients in s , the possible triplets for next merging are $(X_1, s_{2,4}^1, s_{2,4}^2)$, $(s_{2,4}^1, s_{2,4}^2, X_5)$, (X_5, X_6, X_7) , (X_6, X_7, X_8) . We note that $(s_{2,4}^2, X_5, X_6)$ cannot be considered as a candidate for next merging under the two “together rule” as this triplet contains only one (not both) of the paired smooth coefficients returned by the previous merging. Assume that (X_5, X_6, X_7) gives the smallest size of detail coefficient $|d_{5,6,7}|$ among the four candidates, then we merge them through the orthogonal transformation formulated in (4) and now update the sequence into $s = (X_1, s_{2,4}^1, s_{2,4}^2, d_{2,3,4}, s_{5,7}^1, s_{5,7}^2, d_{5,6,7}, X_8)$. This transformation is also Type 1.

Scale $j = 3$. We now compare four candidates for merging, $(X_1, s_{2,4}^1, s_{2,4}^2)$, $(s_{2,4}^1, s_{2,4}^2, s_{5,7}^1)$, $(s_{2,4}^2, s_{5,7}^1, s_{5,7}^2)$ and $(s_{5,7}^1, s_{5,7}^2, X_8)$. The two triplets in middle, $(s_{2,4}^1, s_{2,4}^2, s_{5,7}^1)$ and $(s_{2,4}^2, s_{5,7}^1, s_{5,7}^2)$, are

paired together as they contain two sets of paired smooth coefficients, $(s_{2,4}^1, s_{2,4}^2)$ and $(s_{5,7}^1, s_{5,7}^2)$, and if we were to treat these two triplets separately, we would be violating the “two together” rule. The summary detail coefficient for this pair of triplets is obtained as $d_{2,4,7} = \max(|d_{2,4,7}^1|, |d_{2,4,7}^2|)$, which is compared with those of the other triplets. Now suppose that $(X_1, s_{2,4}^1, s_{2,4}^2)$ has the smallest size of detail; we merge this triplet and update the data sequence into $s = (s_{1,4}^1, s_{1,4}^2, d_{1,1,4}, d_{2,3,4}, s_{5,7}^1, s_{5,7}^2, d_{5,6,7}, X_8)$. This transformation is of Type 2.

Scale $j = 4$. We now have two pairs of paired coefficients: $(s_{1,4}^1, s_{1,4}^2)$ and $(s_{5,7}^1, s_{5,7}^2)$. Therefore, with the “two together” rule in mind, the only possible options for merging are: to merge the two pairs into $(s_{1,4}^1, s_{1,4}^2, s_{5,7}^1, s_{5,7}^2)$, or to merge $(s_{5,7}^1, s_{5,7}^2)$ with X_8 . Suppose that the first merging is preferred. The merge of $(s_{1,4}^1, s_{1,4}^2)$ and $(s_{5,7}^1, s_{5,7}^2)$ into $(s_{1,4}^1, s_{1,4}^2, s_{5,7}^1, s_{5,7}^2)$ is of Type 3 and is performed in two stages as follows. In the first stage, we merge $(s_{1,4}^1, s_{1,4}^2, s_{5,7}^1)$ and then update the sequence temporarily as $s = (s_{1,7}^1, s_{1,7}^2, d_{1,1,4}, d_{2,3,4}, d_{1,4,7}^1, s_{5,7}^2, d_{5,6,7}, X_8)$. In the second stage, we merge $(s_{1,7}^1, s_{1,7}^2, s_{5,7}^2)$, which gives the updated sequence $s = (s_{1,7}^1, s_{1,7}^2, d_{1,1,4}, d_{2,3,4}, d_{1,4,7}^1, d_{1,4,7}^2, d_{5,6,7}, X_8)$. As a summary detail coefficients for this merge, we use $d_{1,4,7} = \max(|d_{1,4,7}^1|, |d_{1,4,7}^2|)$.

Scale $j = 5$. The only available triplet is now $(s_{1,7}^1, s_{1,7}^2, X_8)$, thus we perform this Type 2 merge and update the data sequence into $s = (s_{1,8}^1, s_{1,8}^2, d_{1,1,4}, d_{2,3,4}, d_{1,4,7}^1, d_{1,4,7}^2, d_{5,6,7}, d_{1,7,8})$. The transformation is completed with the updated data sequence which contains $T - 2 = 6$ detail and 2 smooth coefficients.

2.2.3 TGUW transformation: general algorithm

In this Section, we formulate in generality the TGUW transformation illustrated in Section 2.2.2. One of the important principles is “tail-greediness” (Fryzlewicz, 2018) which enables us to reduce the computational complexity by performing multiple merges over non-overlapping regions in a single pass over the data. More specifically, it allows us to perform up to $\max\{2, \lceil \rho \alpha_j \rceil\}$ merges at each scale j , where α_j is the number of smooth coefficients in the data sequence s and $\rho \in (0, 1)$ (the lower bound of 2 is essential to permit a Type 3 transformation, which consists of two

merges).

Sometimes, we will be referring to a detail coefficient $d_{p,q,r}^j$ as $d_{p,q,r}^{(j,k)}$ or $d^{(j,k)}$, where $j = 1, \dots, J$ is the scale of the transform (i.e. the consecutive pass through the data) at which $d_{p,q,r}^j$ was computed, $k = 1, \dots, K(j)$ is the location index of $d_{p,q,r}^j$ within all scale j coefficients, and $d_{p,q,r}^j$ is $d_{p,q,r}^1$ or $d_{p,q,r}^2$ or $d_{p,q,r}$, depending on the type of merge. We now describe the TGUW algorithm.

1. At each scale j , find the set of triplets that are candidates for merging under the “two together” rule and compute the corresponding detail coefficients. Regardless of the type of merge, a detail coefficient $d_{p,q,r}^j$ is, in general, obtained as

$$d_{p,q,r}^j = as_{p:r}^1 + bs_{p:r}^2 + cs_{p:r}^3, \quad (2)$$

where $p \leq q < r$, $s_{p:r}^k$ is the k^{th} smooth coefficient of the subvector $s_{p:r}$ with a length of $r - p + 1$ and the constants a, b, c are the elements of the detail filter $\mathbf{h} = (a, b, c)^T$. We note that (a, b, c) also depends on (p, q, r) , but this is not reflected in the notation, for simplicity. The detail filter is a weight vector used in computing the weighted sum of a triplet of smooth coefficients which should satisfy the condition that the detail coefficient is zero if and only if the corresponding raw observations over the merged regions have a perfect linear trend. If (X_p, \dots, X_r) are the raw observations associated with the triplet of the smooth coefficients $(s_{p:r}^1, s_{p:r}^2, s_{p:r}^3)$ under consideration, then the detail filter \mathbf{h} is obtained in such a way as to produce zero detail coefficient only when (X_p, \dots, X_r) has a perfect linear trend, as the detail coefficient itself represents the extent of non-linearity in the corresponding region of data. This implies that the smaller the size of the detail coefficient, the closer the alignment of the corresponding data section with linearity. Specifically, the detail filter $\mathbf{h} = (a, b, c)^T$ is

established by solving the following equations,

$$\begin{aligned}
a\mathbf{w}_{p:r}^{c,1} + b\mathbf{w}_{p:r}^{c,2} + c\mathbf{w}_{p:r}^{c,3} &= 0, \\
a\mathbf{w}_{p:r}^{l,1} + b\mathbf{w}_{p:r}^{l,2} + c\mathbf{w}_{p:r}^{l,3} &= 0, \\
a^2 + b^2 + c^2 &= 1,
\end{aligned} \tag{3}$$

where $\mathbf{w}_{p:r}^{:,k}$ is k^{th} nonzero element of the subvector $\mathbf{w}_{p:r}$ with a length of $r - p + 1$, and \mathbf{w}^c and \mathbf{w}^l are weight vectors of constancy and linearity, respectively, in which the initial inputs have a form of $\mathbf{w}_0^c = (1, 1, \dots, 1)^\top$, $\mathbf{w}_0^l = (1, 2, \dots, T)^\top$. The last condition in (3) is to preserve the orthonormality of the transform. The solution to (3) is unique up to multiplication by -1 .

2. Summarise all $d_{p,q,r}^i$ constructed in step 1 to a (equal length or shorter) sequence of $d_{p,q,r}$ by finding a summary detail coefficient $d_{p,q,r} = \max(|d_{p,q,r}^1|, |d_{p,q,r}^2|)$ for any pair of detail coefficients constructed by type 3 merges.
3. Sort the size of the summarised detail coefficients $|d_{p,q,r}|$ obtained in step 2 in non-decreasing order.
4. Extract the (non-summarised) detail coefficient(s) $|d_{p,q,r}^i|$ corresponding to the smallest (summarised) detail coefficient $|d_{p,q,r}|$ where both $|d_{p,q,r}^1|$ and $|d_{p,q,r}^2|$ should be extracted only if $d_{p,q,r} = \max(|d_{p,q,r}^1|, |d_{p,q,r}^2|)$. Repeat the extraction until $\max\{2, \lceil \rho \alpha_j \rceil\}$ (or all possible, whichever is the smaller number) detail coefficients have been obtained, as long as the region of the data corresponding to each detail coefficient extracted does not overlap with the regions corresponding to the detail coefficients already drawn.
5. For each $|d_{p,q,r}^i|$ extracted in step 4, merge the corresponding smooth coefficients by updating the corresponding triplet in \mathbf{s} , \mathbf{w}^c and \mathbf{w}^l through the orthonormal transform as follows,

$$\begin{pmatrix} s_{p,r}^1 \\ s_{p,r}^2 \\ d_{p,q,r}^i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\ell}_1^\top \\ \boldsymbol{\ell}_2^\top \\ \mathbf{h}^\top \end{pmatrix} \begin{pmatrix} s_{p:r}^1 \\ s_{p:r}^2 \\ s_{p:r}^3 \end{pmatrix} = \Lambda \begin{pmatrix} s_{p:r}^1 \\ s_{p:r}^2 \\ s_{p:r}^3 \end{pmatrix}, \tag{4}$$

$$\begin{pmatrix} w_{p,r}^{c,1} \\ w_{p,r}^{c,2} \\ 0 \end{pmatrix} = \Lambda \begin{pmatrix} \mathbf{w}_{p:r}^{c,1} \\ \mathbf{w}_{p:r}^{c,2} \\ \mathbf{w}_{p:r}^{c,3} \end{pmatrix}, \quad \begin{pmatrix} w_{p,r}^{l,1} \\ w_{p,r}^{l,2} \\ 0 \end{pmatrix} = \Lambda \begin{pmatrix} \mathbf{w}_{p:r}^{l,1} \\ \mathbf{w}_{p:r}^{l,2} \\ \mathbf{w}_{p:r}^{l,3} \end{pmatrix}. \quad (5)$$

The key step is finding the 3×3 orthonormal matrix, Λ , which is composed of one detail and two low-pass filter vectors in its rows. Firstly the detail filter \mathbf{h}^T is determined to satisfy the conditions in (3), and then the two low-pass filters (ℓ_1^T, ℓ_2^T) are obtained by satisfying the orthonormality of Λ . There is no uniqueness in the choice of (ℓ_1^T, ℓ_2^T), but this has no effect on the transformation itself. The details of this mechanism can be found in Section D of the supplementary materials.

6. Go to step 1 and repeat at new scale $j = j + 1$ as long as we have at least three smooth coefficients in the updated data sequence s .

More specifically, the detail coefficient in (2) is formulated for each type of merging introduced in Section 2.2.2 as follows.

Type 1: merging three initial smooth coefficients ($s_{p,p}^0, s_{p+1,p+1}^0, s_{p+2,p+2}^0$),

$$d_{p,p+1,p+2} = a_{p,p+1,p+2} s_{p,p}^0 + b_{p,p+1,p+2} s_{p+1,p+1}^0 + c_{p,p+1,p+2} s_{p+2,p+2}^0. \quad (6)$$

Type 2: merging one initial and a paired smooth coefficient ($s_{p,p}^0, s_{p+1,r}^1, s_{p+1,r}^2$),

$$d_{p,p,r} = a_{p,p,r} s_{p,p}^0 + b_{p,p,r} s_{p+1,r}^1 + c_{p,p,r} s_{p+1,r}^2, \quad \text{where } p+2 < r, \quad (7)$$

similarly, when merging a paired smooth coefficient and one initial, ($s_{p,r-1}^1, s_{p,r-1}^2, s_{r,r}^0$),

$$d_{p,r-1,r} = a_{p,r-1,r} s_{p,r-1}^1 + b_{p,r-1,r} s_{p,r-1}^2 + c_{p,r-1,r} s_{r,r}^0, \quad \text{where } p+2 < r. \quad (8)$$

Type 3: merging two sets of (paired) smooth coefficients, ($s_{p,q}^1, s_{p,q}^2$) and ($s_{q+1,r}^1, s_{q+1,r}^2$),

$$\begin{aligned} d_{p,q,r}^1 &= a_{p,q,r}^1 s_{p,q}^1 + b_{p,q,r}^1 s_{p,q}^2 + c_{p,q,r}^1 s_{q+1,r}^1 \\ d_{p,q,r}^2 &= a_{p,q,r}^2 s_{p,q}^{01} + b_{p,q,r}^2 s_{p,q}^{02} + c_{p,q,r}^2 s_{q+1,r}^2 \end{aligned} \quad \Rightarrow \quad d_{p,q,r} = \max(|d_{p,q,r}^1|, |d_{p,q,r}^2|), \quad (9)$$

where $q > p+1$ and $r > q+2$. Importantly, the two consecutive merges in (9) are achieved by visiting the same two adjacent data regions twice. In this case, after the first detail coefficient, $d_{p,q,r}^1$, has been obtained, we instantly update the corresponding triplets s , w^c and w^l via an orthonormal transform as defined in (4) and (5). Therefore, the second detail filter, $(a_{p,q,r}^2, b_{p,q,r}^2, c_{p,q,r}^2)$, is constructed with the updated w^c and w^l in a way that satisfies the conditions (3).

The TGUW transform eventually converts the input data sequence X of length T into the sequence containing 2 smooth and $T - 2$ detail coefficients through $T - 2$ orthonormal transforms. The detail coefficients $d^{(j,k)}$ can be regarded as scalar products between X and a particular unbalanced wavelet basis $\psi^{(j,k)}$, where the formal representation is given as $\{d^{(j,k)} = \langle X, \psi^{(j,k)} \rangle, j=1, \dots, J, k=1, \dots, K(j)\}$ for detail coefficients and $s_{1,T}^1 = \langle X, \psi^{(0,1)} \rangle$, $s_{1,T}^2 = \langle X, \psi^{(0,2)} \rangle$ for the two smooth coefficients. The set $\{\psi^{(j,k)}\}$ is an orthonormal unbalanced wavelet basis for \mathbb{R}^T . Some additional properties of the TGUW transform such as sparse representation and computational complexity are discussed in Section 2.6.

2.3 Thresholding

Because at each stage, the TGUW transform constructs the smallest possible detail coefficients, but it is at the same time orthonormal and so preserves the l_2 energy of the input data, the variability (= deviation from linearity) of the signal tends to be mainly encoded in only a few detail coefficients computed at the later stages of the transform. The resulting sparsity of representation of the input data in the domain of TGUW coefficients justifies thresholding as a way of deciding the significance of each detail coefficient (which measures the local deviation from linearity).

We propose to threshold the TGUW detail coefficients under two important rules, which should simultaneously be satisfied; we refer to these as the “connected” rule and the “two together” rule. The “connected” rule prunes the branches of the TGUW detail coefficients if and only if the detail coefficient itself and all of its children coefficients fall below a certain threshold

in absolute value. For instance, referring to the example of Section 2.2.2, if both $d_{1,1,4}$ and $d_{1,7,8}$ were to survive the initial thresholding, the “connected” rule would mean we also had to keep $d_{1,4,7}^1$ and $d_{1,4,7}^2$, which are the children of $d_{1,7,8}$ and the parents of $d_{1,1,4}$ in the TGUW coefficient tree.

The “two together” rule in thresholding is similar to the one in the TGUW transformation except it targets pairs of detail rather than smooth coefficients, and only applies to pairs of detail coefficients arising from Type 3 merges. One such pair in the example of Section 2.2.2 is $(d_{1,4,7}^1, d_{1,4,7}^2)$. The “two together” rule means that both such detail coefficients should be kept if at least one survives the initial thresholding. This is a natural requirement as a pair of Type 3 detail coefficients effectively corresponds to a single merge of two adjacent regions.

Through the thresholding, we wish to estimate the underlying signal f in (1) by estimating $\mu^{(j,k)} = \langle f, \psi^{(j,k)} \rangle$ where $\psi^{(j,k)}$ is an orthonormal unbalanced wavelet basis constructed in the TGUW transform from the data. Throughout the thresholding procedure, the “connected” and “two together” rules are applied in this order. We firstly threshold and apply the “connected” rule, which gives us $\hat{\mu}_0^{(j,k)}$, the initial estimator of $\mu^{(j,k)}$, as

$$\hat{\mu}_0^{(j,k)} = d_{p,q,r}^{(j,k)} \cdot \mathbb{I} \left\{ \exists (j', k') \in C_{j,k} \quad |d_{p',q',r'}^{(j',k')}| > \lambda \right\}, \quad (10)$$

where \mathbb{I} is an indicator function and

$$C_{j,k} = \{(j', k'), j' = 1, \dots, j, k' = 1, \dots, K(j') : d_{p',q',r'}^{(j',k')} \text{ is such that } [p', r'] \subseteq [p, r]\}. \quad (11)$$

Now the “two together” rule is applied to the initial estimators $\hat{\mu}_0^{(j,k)}$ to obtain the final estimators $\hat{\mu}^{(j,k)}$. We firstly note that two detail coefficients, $d_{p,q,r}^{(j,k)}$ and $d_{p',q',r'}^{(j',k+1)}$ are called “paired” when they are formed by Type 3 mergings and when $(j, p, q, r) = (j', p', q', r')$. The “two together” rule is formulated as below,

$$\hat{\mu}^{(j,k)} = \begin{cases} \hat{\mu}_0^{(j,k)}, & \text{if } d_{p,q,r}^{(j,k)} \text{ is not paired,} \\ \hat{\mu}_0^{(j,k)}, & \text{if } d_{p,q,r}^{(j,k)} \text{ is paired with } d_{p,q,r}^{(j,k')} \text{ and both } \hat{\mu}_0^{(j,k)} \text{ and } \hat{\mu}_0^{(j,k')} \text{ are zero or nonzero,} \\ d_{p,q,r}^{(j,k)}, & \text{if } d_{p,q,r}^{(j,k)} \text{ is paired with } d_{p,q,r}^{(j,k')} \text{ and } \hat{\mu}_0^{(j,k')} \neq 0 \text{ and } \hat{\mu}_0^{(j,k)} = 0. \end{cases} \quad (12)$$

It is important to note that the application of the two rules ensures that \tilde{f} is a piecewise-linear function composed of best linear fits (in the least-squares sense) for each estimated interval of linearity. As an aside, we note that the number of survived detail coefficients does not necessarily equal the number of change-points in \tilde{f} as a pair of detail coefficients arising from a Type 3 merge are associated with a single change-point.

2.4 Inverse TGUW transformation

The estimator \tilde{f} of the true signal f in (1) is obtained by inverting (= transposing) the orthonormal transformations in (4) in reverse order to that in which they were originally performed. This inverse TGUW transformation is referred to as TGUW^{-1} , and thus

$$\tilde{f} = \text{TGUW}^{-1} \{ \hat{\mu}^{(j,k)}, j = 1, \dots, J, k = 1, \dots, K(j) \parallel s_{1,T}^1, s_{1,T}^2 \}, \quad (13)$$

where \parallel denotes vector concatenation.

2.5 Post processing for consistency of change-point detection

As will be specified in Theorem 1 of Section 3, the piecewise-linear estimator \tilde{f} in (13) possibly overestimates the number of change-points. To remove the spurious estimated change-points and to achieve the consistency of the number and the locations of the estimated change-points, we borrow the post-processing framework of Fryzlewicz (2018). Lin et al. (2017) show that we can usually post-process l_2 -consistent estimators in this way as a fast enough l_2 error rate implies that each true change-point has an estimator nearby. The post-processing methodology includes two stages, i) execution of three steps, TGUW transform, thresholding and inverse TGUW transform, again to the estimator \tilde{f} in (13) and ii) examination of regions containing only one estimated change-point to check for its significance.

Stage 1. We transform the estimated function \tilde{f} in (13) with change-points $(\tilde{\eta}_1, \tilde{\eta}_2, \dots, \tilde{\eta}_{\tilde{N}})$ into a new estimator $\tilde{\tilde{f}}$ with corresponding change-points $(\tilde{\tilde{\eta}}_1, \tilde{\tilde{\eta}}_2, \dots, \tilde{\tilde{\eta}}_{\tilde{N}})$. Using \tilde{f} in (13) as an input data sequence s , we perform the TGUW transform as presented in Section 2.2.3, but in a greedy rather than tail-greedy way such that only one detail coefficient $d^{(j,1)}$ is produced at each scale j , and thus $K(j) = 1$ for all j . We repeat to produce detail coefficients until the first detail coefficient such that $|d^{(j,1)}| > \lambda$ is obtained where λ is the parameter used in the thresholding procedure described in Section 2.3. Once the condition, $|d^{(j,1)}| > \lambda$, is satisfied, stop merging and relabel the surviving change-points as $(\tilde{\tilde{\eta}}_1, \tilde{\tilde{\eta}}_2, \dots, \tilde{\tilde{\eta}}_{\tilde{N}})$ and construct the new estimator $\tilde{\tilde{f}}$ as

$$\tilde{\tilde{f}}_t = \hat{\theta}_{i,1} + \hat{\theta}_{i,2} t \quad \text{for } t \in [\tilde{\tilde{\eta}}_{i-1} + 1, \tilde{\tilde{\eta}}_i], \quad i = 1, \dots, \tilde{N}, \quad (14)$$

where $\tilde{\tilde{\eta}}_0 = 1$, $\tilde{\tilde{\eta}}_{\tilde{N}+1} = T+1$ and $(\hat{\theta}_{i,1}, \hat{\theta}_{i,2})$ are the OLS intercept and slope coefficients, respectively, for the corresponding pairs $\{(t, X_t), t \in [\tilde{\tilde{\eta}}_{i-1} + 1, \tilde{\tilde{\eta}}_i]\}$. The exception is when the region under consideration only contains a single data point X_{t_0} (a situation we refer to as a point anomaly throughout the paper), in which case fitting a linear regression is impossible, so we then set $\tilde{\tilde{f}}_{t_0} = X_{t_0}$.

Stage 2. From the estimator $\tilde{\tilde{f}}_t$ in Stage 1, we obtain the final estimator \hat{f} by pruning the change-points $(\tilde{\tilde{\eta}}_1, \tilde{\tilde{\eta}}_2, \dots, \tilde{\tilde{\eta}}_{\tilde{N}})$ in $\tilde{\tilde{f}}_t$. For each $i = 1, \dots, \tilde{N}$, compute the corresponding detail coefficient d_{p_i, q_i, r_i} as described in (7)-(9), where $p_i = \lfloor \frac{\tilde{\tilde{\eta}}_{i-1} + \tilde{\tilde{\eta}}_i}{2} \rfloor + 1$, $q_i = \tilde{\tilde{\eta}}_i$ and $r_i = \lceil \frac{\tilde{\tilde{\eta}}_i + \tilde{\tilde{\eta}}_{i+1}}{2} \rceil$. Now prune by finding the minimiser $i_0 = \arg \min_i |d_{p_i, q_i, r_i}|$ and removing $\tilde{\tilde{\eta}}_{i_0}$ and setting $\tilde{N} := \tilde{N} - 1$ if $|d_{p_{i_0}, q_{i_0}, r_{i_0}}| \leq \lambda$ where λ is same as in Section 2.3. Then relabel the change-points with the subscripts $i = 1, \dots, \tilde{N}$ under the convention $\tilde{\tilde{\eta}}_0 = 0$, $\tilde{\tilde{\eta}}_{\tilde{N}+1} = T$. Repeat the pruning while we can find i_0 which satisfies the condition $|d_{p_{i_0}, q_{i_0}, r_{i_0}}| < \lambda$. Otherwise, stop, set \hat{N} as the number of detected change-points and reconstruct the change-points $\hat{\eta}_i$ in increasing order for $i = 0, \dots, \hat{N} + 1$ where $\hat{\eta}_0 = 1$ and $\hat{\eta}_{\hat{N}+1} = T + 1$. The estimated function \hat{f} is obtained by simple linear regression for each region determined by the final change-points $\hat{\eta}_1, \dots, \hat{\eta}_{\hat{N}}$ as in (14), with the exception for point anomalies as described in Stage 1 above.

Through these two stages of post processing, the estimation of the number and the locations of change-points becomes consistent, and further details can be found in Section 3.

2.6 Extra discussion of TGUW transformation

Sparse representation. The TGUW transform is nonlinear, but it is also linear and orthonormal conditional on the order in which the merges are performed. The orthonormality of the unbalanced wavelet basis, $\{\psi^{(j,k)}\}$, implies Parseval's identity, $\sum_{t=1}^T X_t^2 = \sum_{j=1}^J \sum_{k=1}^{K(j)} (d^{(j,k)})^2 + (s_{1,T}^1)^2 + (s_{1,T}^2)^2$ where $d^{(j,k)} = \langle X, \psi^{(j,k)} \rangle$, $s_{1,T}^1 = \langle X, \psi^{(0,1)} \rangle$ and $s_{1,T}^2 = \langle X, \psi^{(0,2)} \rangle$. Furthermore, the filters $(\psi^{(0,1)}, \psi^{(0,2)})$ corresponding to the two smooth coefficients $s_{1,T}^1$ and $s_{1,T}^2$ form an orthonormal basis of the subspace $\{(x_1, x_2, \dots, x_T) \mid x_1 - x_2 = x_2 - x_3 = \dots = x_{T-1} - x_T\}$ of \mathbb{R}^T ; see Section D of the supplementary materials for further details. This implies that $\sum_{t=1}^T X_t^2 - (s_{1,T}^1)^2 - (s_{1,T}^2)^2 = \sum_{t=1}^T (X_t - \hat{X}_t)^2$, where $\hat{X} = s_{1,T}^1 \psi^{(0,1)} + s_{1,T}^2 \psi^{(0,2)}$ is the best linear regression fit to X achieved by minimising the sum of squared errors. This, combined with the Parseval's identity above, implies $\sum_{t=1}^T (X_t - \hat{X}_t)^2 = \sum_{j=1}^J \sum_{k=1}^{K(j)} (d^{(j,k)})^2$.

By construction, the detail coefficients $|d^{(j,k)}|$ obtained in the initial stages of the TGUW transform tend to be small in magnitude. Therefore, the above Parseval's identity implies that a large portion of $\sum_{t=1}^T (X_t - \hat{X}_t)^2$ is explained by only a few large $|d^{(j,k)}|$'s arising in the later stages of the transform; in this sense, the TGUW transform provides sparsity of signal representation.

Computational complexity. Assume that α_j smooth coefficients are available in the data sequence s at scale j . We allow the algorithm to merge up to $\lceil \rho \alpha_j \rceil$ many triplets (unless their corresponding data regions overlap) where $\rho \in (0, 1)$ is a constant. This gives us at most $(1 - \rho)^j T$ smooth coefficients remaining in s after j scales. Solving for $(1 - \rho)^j T \leq 2$ gives the largest number of scales J as $\lceil \log(T) / \log((1 - \rho)^{-1}) + \log(2) / \log(1 - \rho) \rceil$, at which point the TGUW transform terminates with two smooth coefficients remaining. Considering that the most expensive step at each scale is sorting which takes $O(T \log(T))$ operations, the computational complexity of the

TGUW transformation is $O(T \log^2(T))$.

3 Theoretical results

We study the l_2 consistency of \tilde{f} and $\tilde{\tilde{f}}$, and the change-point detection consistency of \hat{f} , where the estimators are defined in Section 2. The l_2 risk of an estimator \tilde{f} is defined as $\|\tilde{f} - f\|_T^2 = T^{-1} \sum_{i=1}^T (\tilde{f}_i - f_i)^2$, where f is the underlying signal as in (1). We note the true change-points $\{\eta_i, i = 1, \dots, N\}$ are such that,

$$\begin{aligned} f_t &= \theta_{\ell,1} + \theta_{\ell,2} t \text{ for } t \in [\eta_{\ell-1} + 1, \eta_\ell], \ell = 1, \dots, N + 1 \\ &\text{where } f_{\eta_\ell} + \theta_{\ell,2} \neq f_{\eta_{\ell+1}} \text{ for } \ell = 1, \dots, N. \end{aligned} \quad (15)$$

This definition permits both continuous and discontinuous changes and if f_{η_i} is a point anomaly, there exist two consecutive change-points at η_{i-1} and η_i where $\eta_{i-1} = \eta_i - 1$. We firstly investigate the l_2 behaviour of \tilde{f} . The proofs of Theorems 1-3 can be found in Appendix A.

Theorem 1 *X_t follows model (1) with $\sigma = 1$ and \tilde{f} is the estimator in (13). If the threshold $\lambda = C_1 \{2 \log(T)\}^{1/2}$ with a constant C_1 large enough, then we have*

$$\mathbb{P} \left(\|\tilde{f} - f\|_T^2 \leq C_1^2 \frac{1}{T} \log(T) \left\{ 4 + 8N \lceil \log(T) / \log(1 - \rho)^{-1} \rceil \right\} \right) \rightarrow 1, \quad (16)$$

as $T \rightarrow \infty$ and the piecewise-linear estimator \tilde{f} contains $\tilde{N} \leq CN \log(T)$ change-points where C is a constant.

Thus, \tilde{f} is l_2 consistent under the strong sparsity assumption i.e. if N is finite. The crucial mechanism of l_2 consistency is the ‘‘tail-greediness’’ which allows up to $K(j) \geq 1$ smooth coefficients to be removed at each scale j . In other words, consistency is generally unachievable if we proceed in a greedy (as opposed to tail-greedy) way, i.e. if we only merge one triplet at each scale of the TGUW transformation.

We now move onto the estimator $\tilde{\tilde{f}}$ obtained in the first stage of post-processing.

Theorem 2 X_t follows model (1) with $\sigma = 1$ and \tilde{f} is the estimator in (14). Let the threshold λ be as in Theorem 1 and let the number of true change-points, N , be finite. Then we have $\|\tilde{f} - f\|_T^2 = O(NT^{-1} \log^2(T))$ with probability approaching 1 as $T \rightarrow \infty$ and there exist at most two estimated change-points between each pair of true change-points (η_i, η_{i+1}) for $i = 0, \dots, N$, where $\eta_0 = 0$ and $\eta_{N+1} = T$. Therefore $\tilde{N} \leq 2(N + 1)$.

We see that \tilde{f} is l_2 consistent, but inconsistent for the number of change-points. Now we investigate the final estimators, \hat{f} and \hat{N} .

Theorem 3 X_t follows model (1) with $\sigma = 1$ and (\hat{f}, \hat{N}) are the estimators obtained in Section 2.5. Let the threshold λ be as in Theorem 1 and suppose that the number of true change-points, N , be finite. Let $\Delta_T = \min_{i=1, \dots, N} \left\{ \left(\underline{f}_T^i \right)^{2/3} \cdot \delta_T^i \right\}$ where $\underline{f}_T^i = \min(|f_{\eta_{i+1}} - 2f_{\eta_i} + f_{\eta_{i-1}}|, |f_{\eta_{i+2}} - 2f_{\eta_{i+1}} + f_{\eta_i}|)$ and $\delta_T^i = \min(|\eta_i - \eta_{i-1}|, |\eta_{i+1} - \eta_i|)$. Assume that $T^{1/3}R_T^{1/3} = o(\Delta_T)$ where $\|\tilde{f} - f\|_T^2 = O_p(R_T)$ is as in Theorem 2. Then we have

$$\mathbb{P} \left(\hat{N} = N, \max_{i=1, \dots, N} \left\{ |\hat{\eta}_i - \eta_i| \cdot \left(\underline{f}_T^i \right)^{2/3} \right\} \leq CT^{1/3}R_T^{1/3} \right) \rightarrow 1, \quad (17)$$

as $T \rightarrow \infty$ where C is a constant.

Our theory indicates that in the case in which $\min_i \underline{f}_T^i$ is bounded away from zero, the consistent estimation of the number and locations of change-point is achieved by assuming $T^{1/3}R_T^{1/3} = o(\delta_T)$ where $\delta_T = \min_{i=1, \dots, N+1} |\eta_i - \eta_{i-1}|$. In addition, when point anomalies exist in the set of true change-points, a point anomaly η_k and its neighbouring change-point $\eta_{k-1} = \eta_k - 1$ can be detected exactly at their true locations only if the corresponding \underline{f}_T^i s satisfy the condition $\min(\underline{f}_T^k, \underline{f}_T^{k-1}) \gtrsim \log(T)$.

4 Simulation study

4.1 Parameter choice and setting

Post-processing. In what follows, we disable Stages 1 and 2 of post-processing by default: our empirical experience is that Stage 1 rarely makes a difference in practice but comes with an additional computational cost, and Stage 2 occasionally over-prunes change-point estimates.

Choice of threshold λ . Motivated by Theorem 1, we use the threshold of the form $\lambda = C\sigma(2 \log T)^{1/2}$ and estimate σ using the Median Absolute Deviation (MAD) estimator (Hampel, 1974) defined as $\hat{\sigma} = \text{Median}(|X_1 - 2X_2 + X_3|, \dots, |X_{T-2} - 2X_{T-1} + X_T|) / (\Phi^{-1}(3/4) \sqrt{6})$ where Φ^{-1} is the quantile function of the Gaussian distribution. We use $C = 1.3$ as a default as it empirically led to the best performance over the range $C \in [1, 1.4]$.

Choice of the “tail-greediness” parameter. $\rho \in (0, 1)$ is a constant which controls the greediness level of the TGUW transformation in the sense that it decides how many merges are performed in a single pass over the data. A large ρ can reduce the computational cost but it makes the procedure less adaptive, whereas a small ρ gives the opposite effect. Based on our empirical experience, the best performance is achieved in the range $\rho \in (0, 0.05]$ and we use $\rho = 0.04$ as a default in the simulation study and data analyses.

We consider i.i.d. Gaussian noise and simulate data from model (1) using 8 signals, (M1) wave1, (M2) wave2, (M3) mix1, (M4) mix2, (M5) mix3, (M6) lin.sgmts, (M7) teeth and (M8) lin, shown in Figure 2. (M1) is continuous at change-points, while (M2) has discontinuities. (M3) has a mix of continuous and discontinuous change-points and contains both constant and linear segments, whereas (M4) is of the same type but also contains two point anomalies. In addition, (M5) has two particularly short segments. (M6) contains isolated spike-type short segments. (M7) is piecewise-constant, and (M8) is a linear signal without change-points. The signals and R code for all simulations can be downloaded from our GitHub repository (Maeng and Fryzlewicz,

2019) and the simulation results under dependent or heavy-tailed errors can be found in Section B of the supplementary materials.

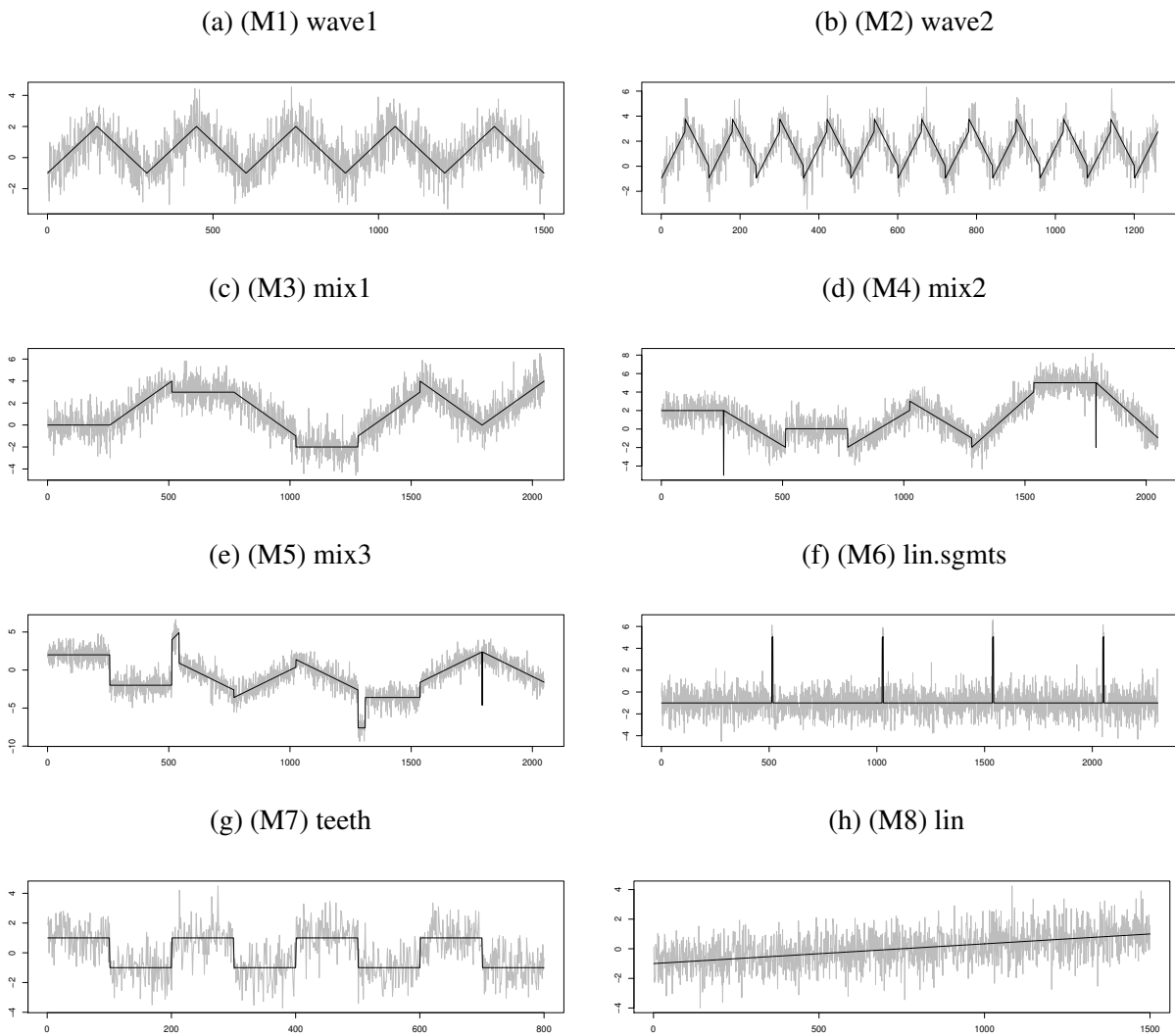


Figure 2: Examples of data with their underlying signals studied in Section 4. (a)-(g) data series X_t (light grey) and true signal f_t (black).

4.2 Competing methods and estimators

We perform the TrendSegment procedure based on the parameter choice in Section 4.1 and compare the performance with that of the following competitors: Narrowest-Over-Threshold detection (**NOT**, Baranowski et al. (2016)) implemented in the R package `not` from CRAN, Isolate-Detect (**ID**, Anastasiou and Fryzlewicz (2018)) available in the R package `IDetect`, trend filtering (**TF**, Kim et al. (2009)) available from <https://github.com/glmgen/genlasso>, Continuous-piecewise-linear Pruned Optimal Partitioning (**CPOP**, Maidstone et al. (2017)) available from <https://www.maths.lancs.ac.uk/~fearnhea/Publications.html> and a bottom-up algorithm based on the residual sum of squares (RSS) from a linear fit (**BUP**, Keogh et al. (2004)). The TrendSegment methodology is implemented in the R package `trendsegmentR`.

As BUP requires a pre-specified number of change-points (or a well-chosen stopping criterion which can vary depending on the data), we include it in the simulation study (with the stopping criterion optimised for the best performance using the knowledge of the truth) but not in data applications. We do not include the methods of Spiriti et al. (2013) and Bai and Perron (2003) implemented in the R packages `freeknotsplines` and `strucchange` as we have found them to be particularly slow. For instance, the minimum segment size in `strucchange` can be adjusted to be small as long as it is greater than or equal to 3 for detecting linear trend changes. This cannot capture point anomalies but is suitable for detecting very short segments (e.g in (M6) `lin.sgmts`). However, this setting is accompanied by extremely heavy computation: with this minimum segment size constraint in place, a single signal simulated from (M6) took us over three hours to process on a standard PC.

Out of the competing methods tested, ID, TF and CPOP are in principle able to classify two consecutive time point as change-points, and therefore they are able to detect point anomalies. NOT and BUP are not designed to detect point anomalies as their minimum distance between two consecutive change-points is restricted to be at least two. For NOT, we use the contrast

function for not necessarily continuous piecewise-linear signals. Regarding the tuning parameters for the competing methods, we follow the recommendation of each respective paper or the corresponding R package.

4.3 Results

The summary of the results for all models and methods can be found in Tables 2 and 3. We run 100 simulations and as a measure of accuracy of estimators, we use Monte-Carlo estimates of the Mean Squared Error of the estimated signal defined as $\text{MSE} = \mathbb{E}\{(1/T) \sum_{t=1}^T (f_t - \hat{f}_t)^2\}$. The empirical distribution of $\hat{N} - N$ is also reported where \hat{N} is the estimated number of change-points and N is the true one. In addition to this, for comparing the accuracy of the locations of the estimated change-points $\hat{\eta}_i$, we show estimates of the scaled Hausdorff distance given by

$$d_H = \frac{1}{T} \mathbb{E} \max \left\{ \max_i \min_j |\eta_i - \hat{\eta}_j|, \max_j \min_i |\hat{\eta}_j - \eta_i| \right\}, \quad (18)$$

where $i = 0, \dots, N + 1$ and $j = 0, \dots, \hat{N} + 1$ with the convention $\eta_0 = \hat{\eta}_0 = 0, \eta_{N+1} = \hat{\eta}_{N+1} = T$ and $\hat{\eta}$ and η denote estimated and true locations of the change-points. The smaller the Hausdorff distance, the better the estimation of the change-point locations. For each method, the average computation time in seconds is shown.

The results for (M1) and (M2) are similar. TrendSegment shows comparable performance to NOT, ID and CPOP in terms of the estimation of the number of change-points, while it is slightly less attractive in terms of the estimated locations of change-points. TF tends to overestimate the number of change-points throughout all models. When the signal is a mix of constant and linear trends as in (M3), TrendSegment, NOT and ID still perform well in terms of the estimation of the number of change-points, while CPOP tends to overestimate. We see that TrendSegment has a particular advantage over the other methods especially in (M4) and (M5), when point anomalies exist or in the case of frequent change-points. TrendSegment shows its relative robustness in

Table 2: Distribution of $\hat{N} - N$ for models (M1)-(M4) and all methods listed in Section 4.1 and 4.2 over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal \hat{f}_i defined in Section 4.3, the average Hausdorff distance d_H given by (18) and the average computational time in seconds using an Intel Core i5 2.9 GHz CPU with 8 GB of RAM, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$.

Model	Method	$\hat{N} - N$							MSE	$d_H(\times 10^2)$	time
		≤ -3	-2	-1	0	1	2	≥ 3			
(M1)	TS	0	0	0	99	1	0	0	0.044	2.79	1.12
	NOT	0	0	0	99	1	0	0	0.034	2.09	0.29
	ID	0	0	0	99	1	0	0	0.029	1.45	0.22
	TF	0	0	0	0	0	0	100	0.016	4.29	36.30
	CPOP	0	0	0	99	1	0	0	0.014	0.78	8.55
	BUP	0	1	18	81	0	0	0	0.069	3.88	2.62
(M2)	TS	0	0	2	98	0	0	0	0.109	1.90	1.06
	NOT	0	0	2	98	0	0	0	0.092	1.56	0.35
	ID	0	0	0	94	6	0	0	0.089	1.44	0.23
	TF	0	0	0	0	0	0	100	0.065	2.31	31.34
	CPOP	0	0	0	93	7	0	0	0.065	1.15	2.09
	BUP	100	0	0	0	0	0	0	0.752	4.69	2.21
(M3)	TS	0	0	1	97	2	0	0	0.032	3.23	1.47
	NOT	0	0	0	100	0	0	0	0.020	2.35	0.36
	ID	0	0	1	94	5	0	0	0.047	2.37	0.33
	TF	0	0	0	0	0	0	100	0.023	5.87	45.31
	CPOP	0	0	0	61	32	6	1	0.024	2.34	21.11
	BUP	0	0	0	3	18	47	32	0.041	5.41	3.50
(M4)	TS	0	0	5	76	18	1	0	0.030	1.81	1.48
	NOT	0	100	0	0	0	0	0	0.066	2.10	0.33
	ID	0	11	52	35	2	0	0	0.163	1.83	0.30
	TF	0	0	0	0	0	0	100	0.080	6.10	44.78
	CPOP	0	0	2	22	45	27	4	0.025	1.60	7.79
	BUP	0	0	8	31	45	13	3	0.092	5.30	3.62

Table 3: Distribution of $\hat{N} - N$ for models (M5)-(M8) and all methods listed in Section 4.1 and 4.2 over 100 simulation runs. Also the average MSE (Mean Squared Error) of the estimated signal \hat{f}_i defined in Section 4.3, the average Hausdorff distance d_H given by (18) and the average computational time in seconds using an Intel Core i5 2.9 GHz CPU with 8 GB of RAM, all over 100 simulations. Bold: methods within 10% of the highest empirical frequency of $\hat{N} - N = 0$ or within 10% of the lowest empirical average $d_H(\times 10^2)$.

Model	Method	$\hat{N} - N$							MSE	$d_H(\times 10^2)$	time
		≤ -3	-2	-1	0	1	2	≥ 3			
(M5)	TS	0	0	1	71	24	4	0	0.031	1.42	1.49
	NOT	0	0	99	1	0	0	0	0.040	1.20	0.29
	ID	0	0	1	2	14	32	51	0.277	8.28	0.30
	TF	0	0	0	0	0	0	100	0.116	6.17	43.13
	CPOP	0	0	0	11	22	39	28	0.023	1.41	5.12
	BUP	0	0	10	45	37	7	1	0.090	4.78	3.64
(M6)	TS	0	0	0	96	4	0	0	0.013	0.05	1.65
	NOT	63	22	4	2	3	0	6	0.240	15.51	0.28
	ID	3	16	0	9	44	1	27	0.151	16.37	0.37
	TF	0	0	0	0	0	0	100	0.134	10.98	48.19
	CPOP	0	0	0	20	41	24	15	0.034	0.13	5.11
	BUP	0	0	0	0	0	0	100	0.135	10.17	4.00
(M7)	TS	0	5	21	40	28	6	0	0.119	7.02	0.65
	NOT	1	1	8	56	31	3	0	0.065	2.62	0.25
	ID	3	0	16	14	26	13	28	0.320	10.87	0.12
	TF	0	0	0	0	0	0	100	0.097	6.11	23.19
	CPOP	0	0	1	1	3	17	78	0.055	3.37	1.19
	BUP	70	25	5	0	0	0	0	0.277	11.89	1.58
(M8)	TS	0	0	0	100	0	0	0	0.001	0.00	1.01
	NOT	0	0	0	100	0	0	0	0.001	0.00	0.17
	ID	0	0	0	100	0	0	0	0.001	0.00	0.59
	TF	0	0	0	78	5	2	15	0.002	9.08	35.79
	CPOP	0	0	0	100	0	0	0	0.001	0.00	12.96
	BUP	0	0	0	0	0	0	100	0.011	46.34	2.63

estimating the number and the location of change-points while ID and CPOP significantly underperform and NOT ignores the point anomalies, as expected. (M6) is another example where only TrendSegment exhibits good performance. For the estimation of the piecewise-constant signal (M7), no method performs well and NOT, ID and TrendSegment tend to underestimate the number of change-points while CPOP and TF overestimate. In the case of the no-change-point signal (M8), all methods except TF perform well.

In summary, TrendSegment is always among the best methods, and is particularly attractive for signals with point anomalies or short segments. With respect to computation time, NOT and ID are very fast in all cases, TrendSegment is slower than these two but is faster than TF, CPOP and BUP, especially when the length of the time series is larger than 2000.

5 Data applications

5.1 Average January temperatures in Iceland

We analyse a land temperature dataset available from <http://berkeleyearth.org>, consisting of average temperatures in January recorded in Reykjavik recorded from 1763 to 2013. Figure 3a shows the data; the point corresponding to 1918 appears to be a point anomaly, and we comment on this aspect later on in this section.

The TrendSegment estimate of the piecewise-linear trend is shown in Figure 3b. It identifies 2 change-points, 1917 and 1918, where the temperature in 1918 is fitted as a single point as it is much lower than in other years. Figures 3c and 3d show that NOT and CPOP detect the change of slope in 1974, ID returns an increasing function with no change-points and TF reports 6 points with the most recent one in 1981, but none of them detect the point anomaly. To assess the goodness of fit of the TrendSegment estimate, we computed the sample autocorrelation and partial autocorrelation functions of the empirical residuals from the TrendSegment fit and both

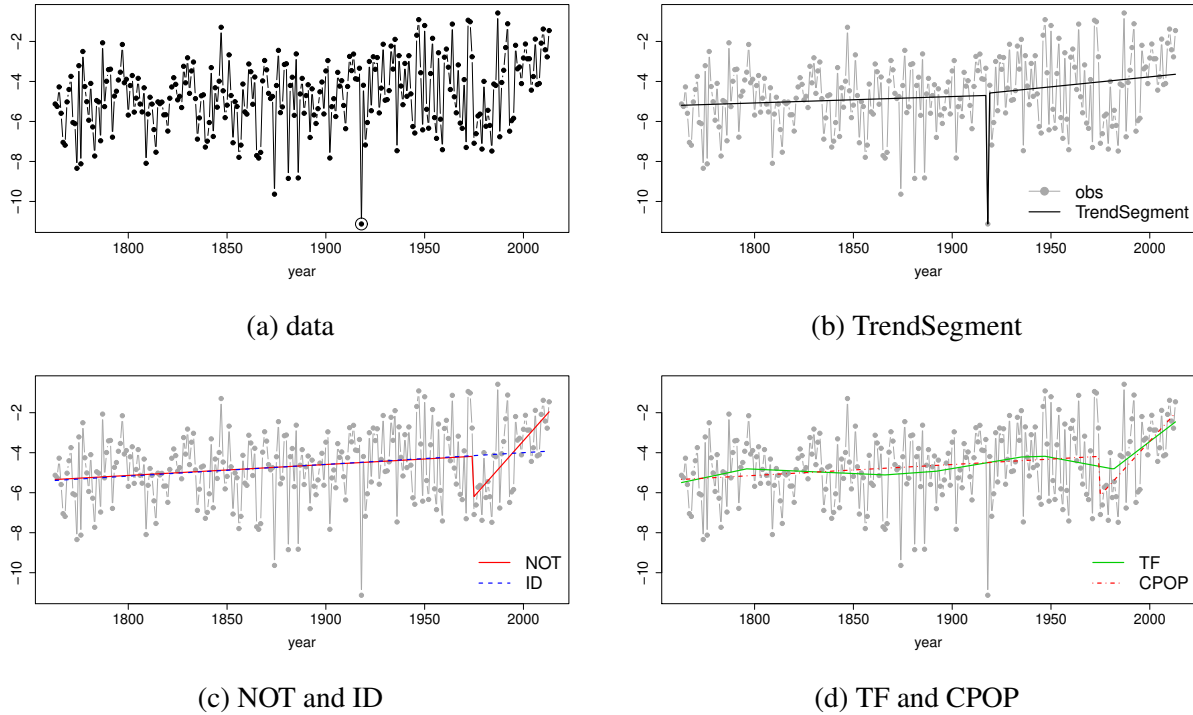


Figure 3: Change-point analysis for January average temperature in Reykjavik from 1763 to 2013 in Section 5.1. (a) the data series, (b) the data series (grey dots) and estimated signal with change-points returned by TrendSegment(—), (c) estimated signal with change-points returned by NOT (—) and ID (- - -), (d) estimated signal with change-points returned by TF (—) and CPOP (- - -).

were consistent with white noise.

Regarding the 1918 observation, Moore and Babij (2017) report that “[t]he winter of 1917/1918 is referred to as the Great Frost Winter in Iceland. It was the coldest winter in the region during the twentieth century. It was remarkable for the presence of sea ice in Reykjavik Harbour as well as for the unusually large number of polar bear sightings in northern Iceland.” This example illustrates the flexibility of the TrendSegment as it detects not only change-points in linear trend but it can identify a point anomaly at the same time, which the competing methods do not achieve.

5.2 Monthly average sea ice extent of Arctic and Antarctic

We analyse the average sea ice extent of the Arctic and the Antarctic available from <https://nsidc.org> to estimate the change-points in its trend. As mentioned in Serreze and Meier (2018), sea ice extent is the most common measure for assessing the condition of high-latitude oceans and it is defined as the area covered with an ice concentration of at least 15%. Here we use the average ice extent in February and September as it is known that the Arctic has the maximum ice extent typically in February while the minimum occurs in September and the Antarctic experiences the opposite.

Serreze and Meier (2018) indicate that the clear decreasing trend of sea ice extent of the Arctic in September is one of the most important indicator of climate change. In contrast to the Arctic, the sea ice extent of the Antarctic has been known to be stable in the sense that it shows a weak increasing trend in the decades preceding 2016 (Comiso et al., 2017; Serreze and Meier, 2018). However, Rintoul et al. (2018) warn of a possible collapse of the past stability by citing a significant decline of the sea ice extent in 2016. We now use the most up-to-date records (to 2018) and re-examine the concerns expressed in Rintoul et al. (2018) with the help of our change-point detection methodology.

Figures 4a and 4c show the well-known decreasing trend of the average sea ice extent in the Arctic both in its winter (February) and summer (September). In Figure 4a, the TrendSegment estimate identifies change-points in 2004 and 2007 and detects a sudden drop during 2005-2007. One change-point in 2006 is identified in Figure 4c, which differentiates the decreasing speed of winter ice extent in the Arctic before and after 2006. As observed in the above-mentioned literature, the sea ice extent of the Antarctic shows a modest increasing trend up until recently (Figures 4b and 4d); however, we observe a strong decreasing trend from the detected change-point in 2016 for the Antarctic summer (February) and from 2015 for the Antarctic winter (September), which is in line with the message of Rintoul et al. (2018). The results for the other competing

methods can be found in Section C of the supplementary materials.

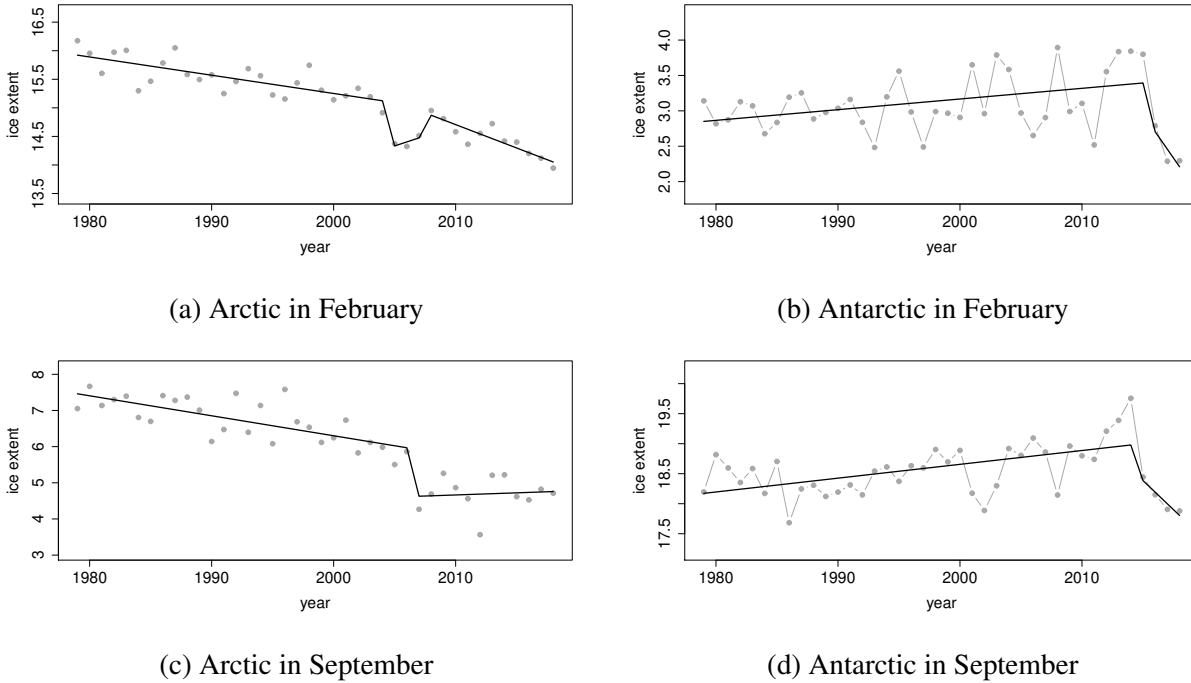


Figure 4: The TrendSegment estimate of piecewise-linear trend for the monthly average sea ice extent from 1979 to 2018 in Section 5.2. (a) the data series (grey dots); the TrendSegment estimate (—) for average sea ice extent of the Arctic in February, (b) Antarctic in February, (c) Arctic in September, (d) Antarctic in September.

A Technical proofs

The proof of Theorem 1-3 are below and Lemmas 1 and 2 can be found in Section A of the supplementary materials.

Proof of Theorem 1. Let S_j^1 and S_j^0 be as in Lemma 2. From the conditional orthonormality of

the unbalanced wavelet transform, on the set A_T defined in Lemma 1, we have

$$\begin{aligned}
\|\tilde{f} - f\|_T^2 &= \frac{1}{T} \sum_{j=1}^J \sum_{k=1}^{K(j)} \left(d^{(j,k)} \cdot \mathbb{I}\{\exists(j',k') \in C_{j,k} \quad |d^{(j',k')}| > \lambda\} - \mu^{(j,k)} \right)^2 + T^{-1}(s_{1,T}^1 - \mu^{(0,1)})^2 + T^{-1}(s_{1,T}^2 - \mu^{(0,2)})^2 \\
&\leq \frac{1}{T} \sum_{j=1}^J \left(\sum_{k \in \mathcal{S}_j^0} + \sum_{k \in \mathcal{S}_j^1} \right) \left(d^{(j,k)} \cdot \mathbb{I}\{\exists(j',k') \in C_{j,k} \quad |d^{(j',k')}| > \lambda\} - \mu^{(j,k)} \right)^2 + 4C_1^2 T^{-1} \log T \\
&=: I + II + 4C_1^2 T^{-1} \log T.
\end{aligned} \tag{19}$$

where $\mu^{(0,1)} = \langle f, \psi^{(0,1)} \rangle$ and $\mu^{(0,2)} = \langle f, \psi^{(0,2)} \rangle$. For $j = 1, \dots, J, k \in \mathcal{S}_j^0$, we have $|d^{(j,k)}| \leq \lambda$, where λ is as in Theorem 1. By Lemma 2, $\mathbb{I}\{\exists(j',k') \in C_{j,k} \quad |d^{(j',k')}| > \lambda\} = 0$ for $k \in \mathcal{S}_j^0$ and also by the fact that $\mu^{(j,k)} = 0$ for $j = 1, \dots, J, k \in \mathcal{S}_j^0$, we have $I = 0$. For II , we denote $\mathcal{B} = \{\exists(j',k') \in C_{j,k} \quad |d^{(j',k')}| > \lambda\}$ and have

$$\begin{aligned}
(d^{(j,k)} \cdot \mathbb{I}\{\mathcal{B}\} - \mu^{(j,k)})^2 &= (d^{(j,k)} \cdot \mathbb{I}\{\mathcal{B}\} - d^{(j,k)} + d^{(j,k)} - \mu^{(j,k)})^2 \\
&\leq (d^{(j,k)})^2 \mathbb{I}(|d^{(j',k')}| \leq \lambda) + 2|d^{(j,k)}| \mathbb{I}(|d^{(j',k')}| \leq \lambda) |d^{(j,k)} - \mu^{(j,k)}| + (d^{(j,k)} - \mu^{(j,k)})^2 \\
&\leq \lambda^2 + 2\lambda C_1 \{2 \log T\}^{1/2} + 2C_1^2 \log T.
\end{aligned} \tag{20}$$

Combining with the upper bound of J , $\lceil \log(T)/\log(1 - \rho)^{-1} \rceil$, and the fact that $|\mathcal{S}_j^1| \leq N$, we have $II \leq 8C_1^2 N T^{-1} \lceil \log(T)/\log(1 - \rho)^{-1} \rceil \log T$, and therefore $\|\tilde{f} - f\|_T^2 \leq C_1^2 \frac{1}{T} \log(T) \{4 + 8N \lceil \log(T)/\log(1 - \rho)^{-1} \rceil\}$. Also, at each scale, the estimated change-points are obtained up to size N , combining it with the largest scale J , the number of change-points in \tilde{f} returned from the inverse TGUW transformation is up to $CN \log T$ where C is a constant.

Proof of Theorem 2. Let \tilde{B} and $\tilde{\tilde{B}}$ the unbalanced wavelet bases corresponding to \tilde{f} and $\tilde{\tilde{f}}$, respectively. As the change-points in $\tilde{\tilde{f}}$ are a subset of those in \tilde{f} , establishing $\tilde{\tilde{f}}$ can be regarded as applying the TGUW transform again to \tilde{f} , which is just a repetition of the estimation procedure \tilde{f} but performed in a greedy way. Thus $\tilde{\tilde{B}}$ is classified into two categories, 1) all basis vectors $\psi^{(j,k)} \in \tilde{\tilde{B}}$ such that $\psi^{(j,k)}$ is not associated with the change-points in \tilde{f} and $|\langle X, \psi^{(j,k)} \rangle| = |d^{(j,k)}| < \lambda$ and 2) all vectors $\psi^{(j,1)}$ produced in Stage 1 of post-processing.

We now investigate how many scales are used for this particular transform. Firstly, the detail coefficients $d^{(j,k)}$ corresponding to the basis vectors $\psi^{(j,k)} \in \tilde{\tilde{B}}$ live on no more than $J = O(\log T)$

scales and we have $|\mathcal{S}_j^1| \leq N$ by the argument used in the proof of Theorem 1. In addition, the vectors $\psi^{(j,1)}$ in the second category above correspond to different change-points in \tilde{f} and there exist at most $\tilde{N} = O(N \log T)$ change-points in \tilde{f} which we examine one at once (i.e. $|\mathcal{S}_j^1| \leq 1$), thus at most \tilde{N} scales are required for $d^{(j,1)}$. Combining the results of the two categories, the equivalent of quantity II in the proof of Theorem 1 for \tilde{f} is bounded by $II \leq C_3 N T^{-1} \log^2 T$ and this completes the proof of the L_2 result, $\|\tilde{f} - f\|_T^2 = O(N T^{-1} \log^2(T))$ where C_3 is a large enough positive constant.

Finally, we show that there exist at most two change-points in \tilde{f} between true change points (η_i, η_{i+1}) for $i = 0, \dots, N$ where $\eta_0 = 0$ and $\eta_{N+1} = T$. Consider the case where three change-point for instance $(\tilde{\eta}_l, \tilde{\eta}_{l+1}, \tilde{\eta}_{l+2})$ lie between a pair of true change-points, (η_i, η_{i+1}) . In this case, by Lemma 2, the maximum magnitude of two detail coefficients computed from the adjacent intervals, $[\tilde{\eta}_l + 1, \tilde{\eta}_{l+1}]$ and $[\tilde{\eta}_{l+1} + 1, \tilde{\eta}_{l+2}]$, is less than λ and $\tilde{\eta}_{l+1}$ would get removed from the set of estimated change-points. This leads to $\tilde{N} \leq 2(N + 1)$.

Proof of Theorem 3. From the assumptions of Theorem 3, 1) given any $\epsilon > 0$ and $C > 0$, for some T_1 and all $T > T_1$, it holds that $\mathbb{P}(\|\tilde{f} - f\|_T^2 > \frac{C^3}{4} R_T) \leq \epsilon$ where \tilde{f} is the estimated signal specified in Theorem 2 and 2) For some T_2 , and all $T > T_2$, it holds that $C^{1/3} T^{1/3} R_T^{1/3} (\Delta_i^f)^{-2/3} < \Delta_i^\eta$ for all $i = 1, \dots, N$. Similar to the argument of Theorem 19 in Lin et al. (2016), we take $T \geq T^*$ where $T^* = \max\{T_1, T_2\}$ and let $r_{i,T} = \lfloor C^{1/3} T^{1/3} R_T^{1/3} (\Delta_i^f)^{-2/3} \rfloor$ for $i = 1, \dots, N$. Suppose that there exist at least one η_i whose closest estimated change-point is not within the distance of $r_{i,T}$. Then there are no estimated change-points in \tilde{f} within $r_{i,T}$ of η_i which means that \tilde{f}_j displays a linear trend over the entire segment $j \in \{\eta_i - r_{i,T}, \dots, \eta_i + r_{i,T}\}$. Hence, $\frac{1}{T} \sum_{j=\eta_i - r_{i,T}}^{\eta_i + r_{i,T}} (\tilde{f}_j - f_j)^2 \geq \frac{13r_{i,T}^3}{24T} (\Delta_i^f)^2 > \frac{C^3}{4} R_T$. We see that assuming that at least one η_i does not have any estimated change-point within the distance of $r_{i,T}$ implies the estimation error exceeds $\frac{C^3}{4} R_T$ which is a contradiction as it is an event that we know occurs with probability at most ϵ . Therefore, there must exist at least one estimated change-point within the distance of $r_{i,T}$ from each true change point η_i .

Throughout Stage 2 of post processing, $\tilde{\eta}_{i_0}$ is either the closest estimated change-point of any η_i or not. If $\tilde{\eta}_{i_0}$ is not the closest estimated change-point to the nearest true change-point on either its left or its right, by the construction of detail coefficients in Stage 2 of post processing, Lemma 2 guarantees that the corresponding detail coefficient has the magnitude less than λ and $\tilde{\eta}_{i_0}$ gets removed. Suppose $\tilde{\eta}_{i_0}$ is the closest estimated change-point of a true change-point η_i and it is within the distance of $CT^{1/3}R_T^{1/3}(\Delta_i^f)^{-2/3}$ from η_i . If the corresponding detail coefficient has the magnitude less than λ and $\tilde{\eta}_{i_0}$ is removed, there must exist another $\tilde{\eta}_i$ within the distance of $CT^{1/3}R_T^{1/3}(\Delta_i^f)^{-2/3}$ from η_i . If there are no such $\tilde{\eta}_i$, then by the construction of the detail coefficient, the order of magnitude of $|d_{p_{i_0}, q_{i_0}, r_{i_0}}|$ would be such that $|d_{p_{i_0}, q_{i_0}, r_{i_0}}| > \lambda$ thus $\tilde{\eta}_{i_0}$ would not get removed. Therefore, after Stage 2 of post processing is finished, each true change-point η_i has its unique estimator within the distance of $CT^{1/3}R_T^{1/3}(\Delta_i^f)^{-2/3}$.

SUPPLEMENTARY MATERIAL

Title: Supplementary materials for “Detecting linear trend changes and point anomalies in data sequences” (.pdf file)

R-package for TrendSegment: R package `trendsegmentR` available from CRAN. The package contains code to perform the TrendSegment method described in the article.

References

- Anastasiou, A. and Fryzlewicz, P. (2018). Detecting multiple generalized change-points by isolating single ones. *Preprint*.
- Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78.
- Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18:1–22.

- Baranowski, R., Chen, Y., and Fryzlewicz, P. (2016). Narrowest-over-threshold detection of multiple change-points and change-point-like features. *arXiv preprint arXiv:1609.00293*.
- Bardwell, L., Fearnhead, P., et al. (2017). Bayesian detection of abnormal segments in multiple time series. *Bayesian Analysis*, 12:193–218.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41:Article 15.
- Comiso, J. C., Gersten, R. A., Stock, L. V., Turner, J., Perez, G. J., and Cho, K. (2017). Positive trend in the antarctic sea ice cover and associated changes in surface temperature. *Journal of Climate*, 30:2251–2267.
- Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2018). A linear time method for the detection of point and collective anomalies. *arXiv preprint arXiv:1806.01947*.
- Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46:3390–3421.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69:383–393.
- Jamali, S., Jönsson, P., Eklundh, L., Ardö, J., and Seaquist, J. (2015). Detecting changes in vegetation trends using time series segmentation. *Remote Sensing of Environment*, 156:182–195.
- James, N. A., Kejariwal, A., and Matteson, D. S. (2016). Leveraging cloud data to mitigate user experience from breaking bad. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 3499–3508. IEEE.
- Jeng, X. J., Cai, T. T., and Li, H. (2012). Simultaneous discovery of rare and common segment variants. *Biometrika*, 100:157–172.
- Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2004). Segmenting time series: A survey and novel approach. In *Data mining in time series databases*, pages 1–21. World Scientific.
- Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM review*, 51:339–360.
- Lin, K., Sharpnack, J., Rinaldo, A., and Tibshirani, R. J. (2016). Approximate recovery in changepoint problems, from ℓ_2 estimation error rates. *arXiv preprint arXiv:1606.06746*.

- Lin, K., Sharpnack, J. L., Rinaldo, A., and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6884–6893.
- Maeng, H. and Fryzlewicz, P. (2019). Detecting linear trend changes and point anomalies in data sequences: Simulation code. URL <https://github.com/hmaeng/trendsegment>.
- Maidstone, R., Fearnhead, P., and Letchford, A. (2017). Detecting changes in slope with an l_0 penalty. *arXiv preprint arXiv:1701.01672*.
- Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109:334–345.
- Matteson, D. S., James, N. A., Nicholson, W. B., and Segalini, L. C. (2013). Locally stationary vector processes and adaptive multivariate modeling. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8722–8726. IEEE.
- Moore, G. and Babij, M. (2017). Iceland’s great frost winter of 1917/1918 and its representation in reanalyses of the twentieth century. *Quarterly Journal of the Royal Meteorological Society*, 143:508–520.
- Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5:557–572.
- Rintoul, S., Chown, S., DeConto, R., England, M., Fricker, H., Masson-Delmotte, V., Naish, T., Siebert, M., and Xavier, J. (2018). Choosing the future of antarctica. *Nature*, 558:233–241.
- Robbins, M. W., Lund, R. B., Gallagher, C. M., and Lu, Q. (2011). Changepoints in the north atlantic tropical cyclone record. *Journal of the American Statistical Association*, 106:89–99.
- Robinson, L. F., Wager, T. D., and Lindquist, M. A. (2010). Change point estimation in multi-subject fmri studies. *Neuroimage*, 49:1581–1592.
- Serreze, M. C. and Meier, W. N. (2018). The arctic’s sea ice cover: trends, variability, predictability, and comparisons to the antarctic. *Annals of the New York Academy of Sciences*.
- Spiriti, S., Eubank, R., Smith, P. W., and Young, D. (2013). Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation*, 83:1020–1036.

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