# Time-Threshold Maps: using information from wavelet reconstructions with all threshold values simultaneously

Piotr Fryzlewicz[*]

February 22, 2012

**Abstract**

Wavelets are a commonly used tool in science and technology. Often, their use involves applying a wavelet transform to the data, thresholding the coefficients and applying the inverse transform to obtain an estimate of the desired quantities. In this paper, we argue that it is often possible to gain more insight into the data by producing not just one, but many wavelet reconstructions using a range of threshold values and analysing the resulting object, which we term the Time-Threshold Map (TTM) of the input data. We discuss elementary properties of the TTM, in its "basic" and "derivative" versions, using both Haar and Unbalanced Haar wavelet families. We then show how the TTM can help in solving two statistical problems in the signal + noise model: breakpoint detection, and estimating the longest interval of approximate stationarity. We illustrate both applications with examples involving volatility of financial returns. We also briefly discuss other possible uses of the TTM.

**Keywords:** Time-Threshold Maps; wavelets; thresholding; breakpoint detection; Unbalanced Haar; volatility.

## 1 Introduction

Wavelets can be informally described as oscillatory functions, typically compactly supported in the domain they live on and also localised, to some extent, in the corresponding frequency domain. For the purpose of data analysis, they are often arranged into multiscale orthonormal bases with a dyadic parent-children structure, which lead to decompositions of data that (a) are fast to compute, (b) are stable and fast to invert, (c) provide a scale-location resolution of the data and (d) are often sparse, i.e. only a small proportion of the coefficients of the decomposition tend to explain a large portion of the variability of the data. There are many wavelet families to choose from. Section 2 provides a very brief introduction to wavelets; for a more complete overview the reader is referred to one of the many

[*] Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: `p.fryzlewicz@lse.ac.uk`.

monographs on wavelets that have appeared since the late eighties, e.g. Daubechies (1992), Vidakovic (1999), Härdle et al. (2000), Nason (2008) or the overview papers Antoniadis (2007), Fryzlewicz (2010a).

Thanks to their attractive properties, an incomplete list of which appears above, wavelets have attracted enormous attention in many branches of science and engineering. The Web of Knowledge lists over 70,000 articles on the topic of wavelets published to date. Amongst those published in 2011 alone, one can find applications of wavelets in climatology, astrophysics, harmonic analysis, genome biology, signal, image and video processing, material science, neuroscience, statistical time series analysis, and others.

Central to the popularity of wavelets and the breadth of their applicability is the concept of sparsity of wavelet representations (already mentioned in the first paragraph above) and the associated concept of complexity reduction via wavelet thresholding, which first appeared in the seminal paper by Donoho and Johnstone (1994) in the context of statistical signal denoising. In a typical wavelet decomposition, the main salient features of the data are well described by only a few large wavelet coefficients; the rest carry the "residual" component of the data which in many applications can be safely omitted (e.g. the information discarded in lossy image compression) or is indeed unwanted (e.g. the noise in signal denoising). Thresholding, whereby small wavelet coefficients get set to zero and large ones are preserved, is often used to separate the two groups and reduce the complexity of the data (e.g. denoise the dataset or compress the image) as desired in the particular application.

Due to the importance of the concept of thresholding in wavelet applications, the topic of threshold selection has attracted considerable attention in various contexts. The first work in which a particular threshold selection method was proposed was again Donoho and Johnstone (1994), who introduced the so-called universal threshold in the problem of estimating a function contaminated with iid Gaussian noise. The universal threshold guaranteed noise removal with probability converging to one with the sample size. Due to the simplicity of the function + Gaussian noise model and its importance as a "canonical" model in nonparametric statistics, we will also use it to introduce the main ideas of this paper; therefore, it is appropriate that we mention other statistical techniques used for wavelet threshold selection in this setting; these include Stein's shrinkage estimation (Donoho and Johnstone, 1995), cross-validation (Nason, 1996), false discovery rate (Abramovich and Benjamini, 1996) and empirical Bayes (Johnstone and Silverman, 2005).

We note that all of the above threshold selection procedures in the function estimation problem advocate the choice of one single threshold value per each wavelet coefficient, which leads to inference about data being based on the resulting single wavelet reconstruction. Intuitively, it would appear to us that in some statistical problems (including, but not necessarily limited to problems in which wavelet thresholding was already routinely used, such as function estimation), more insight into the data could be gained by applying *more than one threshold value to each wavelet coefficient* and analysing the resulting *family of wavelet reconstructions*. Motivated by this observation, this work proposes a generic data-analytic tool, the Time-Threshold Map (TTM), where the dataset at hand gets decomposed in a given wavelet basis, the wavelet coefficients get thresholded using a range of threshold values, and the resulting sequence of wavelet reconstructions of the data (one for each threshold value) is used as an output of the procedure and a basis for drawing conclusions about the data. Roughly speaking, for a one-dimensional time-ordered dataset $X_t$ (e.g. a time series or a signal), the TTM of $X_t$ in its most basic version is defined as the matrix whose successive rows are the wavelet reconstructions of $X_t$ obtained for successive values of the wavelet threshold parameters from a specified range.

The TTM is not the first method for data analysis that proceeds by applying the same statistical procedure for a range of parameter values; however, to the best of our knowledge, it is the first one in a wavelet context. To quote some other examples, SiZer (Chaudhuri and Marron, 1999) is a data visualisation technique for displaying features of kernel-smoothed data as a function of location and bandwidth, simultaneously over a range of bandwidths. Thick Pen (Fryzlewicz and Oh, 2011) is a technique for displaying and analysing time series which uses the idea of plotting the time series data with a range of pens with varying thicknesses. Finally, we note that the idea of visualisation of the output of a statistical algorithm for a range of tuning parameter values simultaneously also appears in a number of modern variable selection techniques such as LARS (Efron et al., 2004) where entire solution paths are computed and displayed at once to aid the analysis.

Having introduced the TTM, we provide two examples of its application: one to the problem of breakpoint detection in time series, and the other to the problem of the estimation of the longest interval of parameter constancy, also in a time series context. We also briefly discuss how it can possibly be applied in a selection of other statistical problems.

The paper is organised as follows. Section 2 provides a very brief introduction to wavelets and defines the "basic" and "derivative" TTM, as well as discussing their basic properties. Section 3 uses an example of a well-known contrived dataset to illustrate the typical features of TTMs and discusses how they can potentially aid in the analysis and understanding of some types of data. Section 4 applies and extends these ideas to propose solutions, based on the TTM, to the statistical problems listed in the previous paragraph.

## 2 Time-Threshold Maps: motivation, definition, versions

### 2.1 Haar and Unbalanced Haar wavelets

For a data vector $\mathbf{x} = (x_1, \ldots, x_n)^T$, a wavelet transform of $\mathbf{x}$ is a linear orthonormal transform $W\mathbf{x}$ which provides a certain scale-location decomposition of $\mathbf{x}$, is computable in $O(n)$ operations, and is able to represent $\mathbf{x}$ sparsely in the sense that many elements of $W\mathbf{x}$ will be close to or exactly zero. Rather than making these statements more precise for a general wavelet transform, we provide two examples of wavelet transforms which will be used throughout the paper: those involving Haar and Unbalanced Haar wavelets. For a more complete introduction to wavelets, the reader is referred to the monographs and overview papers listed in the Introduction.

In Haar wavelets, the rows of $W = W_H$ are given by vectors $\psi_{j,k}$ with elements of the form

$$\begin{aligned}\psi_{j,k}(l) &= 2^{-(J-j)/2}\mathbb{I}_{\{1+(k-1)2^{J-j},\ldots,2^{j-1}+(k-1)2^{J-j}\}}(l) \\ &- 2^{-(J-j)/2}\mathbb{I}_{\{2^{j-1}+1+(k-1)2^{J-j},\ldots,2^{j}+(k-1)2^{J-j}\}}(l),\end{aligned}$$

for $l = 1, \ldots, n$, where $j, k$ are (respectively) scale and location parameters with ranges $j = 0, \ldots, J-1$ and $k = 1, \ldots, 2^j$, with $n = 2^J$ (the function $\mathbb{I}_A(\cdot)$ is the indicator function of the set $A$). The exception is the first row, given by $\psi_{-1,1}(l) = n^{-1/2}$. For example, when

$n = 8$, we have the matrix

$$W_H = \begin{pmatrix} \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

The thus-defined $W_H$ is orthonormal (its rows define an orthonormal basis of $\mathbb{R}^n$), it extracts information from $\mathbf{x}$ at certain dyadic scales and locations, and its structure enables the computation of $W_H\mathbf{x}$ in $O(n)$ computational time via the pyramid algorithm of Mallat (1989). It also offers a sparse representation of piecewise-constant vectors in the sense that for a vector $\mathbf{x}$ with $M$ breakpoints, only at most $M \log_2 n + 1$ coefficients of $W_H\mathbf{x}$ are non-zero. Lower values of $j$ correspond to "coarser" and higher – to "finer" scales of the decomposition.

We note that each Haar vector $\psi_{j,k}$ changes signs from positive to negative exactly in the middle of its support, which may be restrictive in some statistical applications, such as in breakpoint detection in the piecewise constant function + noise model (see e.g. Brodsky and Darkhovsky, 1993, for the problem set-up and some early references). Another restriction is that the definition of a Haar basis is only straightforward if $n$ is an integer power of 2. To allow greater flexibility on both counts, it is possible to define Unbalanced Haar (UH) wavelets, in which the sign change occurs not necessarily in the middle of the support of the wavelets.

The construction of UH wavelets proceeds as follows. First, a vector $\psi_{0,1}(l)$ is formed, which is constant and positive for $l = 1, \ldots, b^{0,1}$, and constant and negative for $l = b^{0,1} + 1, \ldots, n$. The breakpoint $b^{0,1} < n$ is to be chosen by the analyst. Then this construction is repeated on the two parts of the domain determined by $\psi_{0,1}$: that is, provided that $b^{0,1} \geq 2$, we construct (in a similar fashion) a vector $\psi_{1,1}$ supported on $l = 1, \ldots, b^{0,1}$, with a breakpoint $b^{1,1}$. Also, provided that $n - b^{0,1} \geq 2$, we construct a vector $\psi_{1,2}$ supported on $l = b^{0,1} + 1, \ldots, n$ with a breakpoint $b^{1,2}$. The recursion then continues in the same manner for as long as feasible, with each vector $\psi_{j,k}$ having at most two "children" vectors $\psi_{j+1,2k-1}$ and $\psi_{j+1,2k}$. For each vector $\psi_{j,k}$, their start, breakpoint and end indices are denoted by $s^{j,k}$, $b^{j,k}$ and $e^{j,k}$, respectively. As in the Haar wavelets, the indices $j, k$ are scale and location parameters, respectively. Small (large) values of $j$ can be thought of as corresponding to "coarse" ("fine") scales.

We consider an example of a set of UH vectors for $n = 6$. The rows of the matrix $W_{UH}$ defined below contain (from top to bottom) vectors $\psi_{-1,1}$, $\psi_{0,1}$, $\psi_{1,2}$, $\psi_{2,3}$, $\psi_{2,4}$ and $\psi_{3,7}$ determined by the following set of breakpoints: $(b^{0,1}, b^{1,2}, b^{2,3}, b^{2,4}, b^{3,7}) = (1, 3, 2, 5, 4)$.

$$W_{UH} = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{30}} & -\frac{1}{\sqrt{30}} & -\frac{1}{\sqrt{30}} & -\frac{1}{\sqrt{30}} & -\frac{1}{\sqrt{30}} \\ 0 & \frac{\sqrt{3}}{\sqrt{10}} & \frac{\sqrt{3}}{\sqrt{10}} & -\frac{\sqrt{2}}{\sqrt{15}} & -\frac{\sqrt{2}}{\sqrt{15}} & -\frac{\sqrt{2}}{\sqrt{15}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{\sqrt{2}}{\sqrt{3}} \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{pmatrix}$$

We note that similarly to $W_H$, the transform $W_{UH}$ is orthonormal, provides a scale-location decomposition of its input and represents it sparsely if it is (close to) piecewise constant. If the set of breakpoints is fixed, the UH decomposition can also be computed in time $O(n)$.

The running example of the use of the TTM involving Unbalanced Haar wavelets in this paper will be in the problem of breakpoint detection in the function + noise model. In this type of application, selecting a suitable UH basis is of utmost importance. One basis selection procedure described in Fryzlewicz (2007) is the following greedy forward stagewise procedure, related to the matching pursuit algorithm of Mallat and Zhang (1993) and to the binary segmentation technique of Sen and Srivastava (1975). We first define the *UH mother vector* $\psi_{s,b,e}$ with elements defined by

$$\psi_{s,b,e}(l) = \left\{\frac{1}{b-s+1} - \frac{1}{e-s+1}\right\}^{1/2} I(s \le l \le b) - \left\{\frac{1}{e-b} - \frac{1}{e-s+1}\right\}^{1/2} I(b+1 \le l \le e).$$

- The breakpoint $b^{0,1}$ is chosen such that the inner product $\langle \mathbf{x}, \psi_{1,b^{0,1},n}\rangle$ is maximised in absolute value.
- Similarly, $b^{j+1,l} := \text{argmax}_b |\langle \mathbf{x}, \psi_{s^{j+1,l},b,e^{j+1,l}}\rangle|$, where $l = 2k-1, 2k$.

Under a mild assumption on the permitted degree of "unbalancedness" of the thus-constructed UH basis, the computational complexity of the above procedure is $O(n \log n)$.

A large variety of other wavelet families have been used in various statistical contexts: these include Shannon's, Meyer's, Franklin's and Daubechies' wavelets. The reader is refereed to Vidakovic (1999), Section 3.4, for a concise description of these wavelets.

## 2.2 Time-Threshold Maps

For any wavelet transform $W$ (with rows $\psi_{j,k}$) applied to a data vector $\mathbf{x}$, we denote $d_{j,k} = \langle \psi_{j,k}, \mathbf{x}\rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product. The sparsity property of wavelets means that for a "typical" input vector $\mathbf{x}$, the sequence $d_{j,k}$ will be sparse, with many elements close to or even exactly zero. The meaning of "typical" depends on what wavelet basis is used: for example, Haar and Unbalanced Haar wavelets produce sparse decompositions of vectors which are exactly or close to piecewise constant.

A canonical example of a statistical application where the sparsity property of wavelets has been used is that of estimating a function from noisy observations (a.k.a. smoothing or denoising). In the most basic setup, we observe

$$x_i = f_i + \varepsilon_i, \tag{1}$$

where $f_i = f(i/n)$ are the function values to be estimated and $\varepsilon_i$ is noise, which in this illustration we assume to be iid Gaussian with mean zero. A wavelet decomposition $W$ of the terms of this regression equation using any fixed orthonormal wavelet basis yields, respectively, $d_{j,k} = \mu_{j,k} + \varepsilon_{j,k}$, where many of the $\mu_{j,k}$'s are hopefully close to zero and only a few are large in magnitude, and $\varepsilon_{j,k}$'s are again iid Gaussian due to the orthonormality of $W$. Because of this separation of the coefficients $d_{j,k}$ into a few large and many small ones, a natural estimator for $\mu_{j,k}$ is the (hard) thresholding estimator

$$\hat{\mu}_{j,k} = d_{j,k}\mathbb{I}(|d_{j,k}| \geq \lambda), \quad (j,k) \neq (-1,1)$$

(and $\hat{\mu}_{-1,1} = d_{-1,1}$), which yields an estimator $\hat{f}$ of $f$ upon applying the inverse wavelet transform $W^{-1} = W^T$. For any $j,k$, we introduce the functional notation $t(d_{j,k}, \lambda) := \hat{\mu}_{j,k}$, which will be useful later. If $\lambda$ is chosen well, $\hat{f}$ can often be shown to possess several attractive properties, e.g. mean-square consistency with a near-optimal rate for a wide range of signals $f$, see e.g. Vidakovic (1999), Section 6.6, for a summary. Arguably the simplest "good" choice of $\lambda$ is the universal threshold of Donoho and Johnstone (1994) of the form $\lambda = \sigma\sqrt{2\log n}$ where $\sigma^2 = \mathrm{Var}(\varepsilon_i)$ (and can be easily estimated from the data), which leads to $\hat{f}$ being noise-free in a suitable sense, with high probability. The principle of function estimation via wavelet thresholding has been extended to a variety of other settings with more complicated noise structure, including Poisson intensity estimation (see Besbeas et al., 2004, for a review), spectral density estimation in time series (Neumann, 1996, amongst others), or time-varying parameter estimation in locally stationary time series models (Nason et al., 2000, amongst others). Although details differ, meaningful application of wavelet thresholding in these settings often requires the use of a different threshold value $\lambda_{j,k}$ for each $j,k$.

For notational convenience later, we define the vectors $\mathbf{d} = (d_{j,k})_{j,k}$ and $\Lambda = (\lambda_{j,k})_{j,k}$ (where the indices $j,k$, here and below, are arranged in the same order as in the rows of $W$). We extend the definition of the function $t(\cdot)$ to vector arguments as follows: $t(\mathbf{d}, \lambda) = (t(d_{j,k}, \lambda))_{j,k}$ and $t(\mathbf{d}, \Lambda) = (t(d_{j,k}, \lambda_{j,k}))_{j,k}$. We also define the term "wavelet reconstruction" or "reconstruction" of a data vector $\mathbf{x}$ with (vector) threshold $\Lambda$ as follows:

$$\mathbf{x}_\Lambda := W^{-1}t(W\mathbf{x}, \Lambda), \tag{2}$$

with $\mathbf{x}_\lambda$ defined analogously. We note that this definition is model-free and that $\mathbf{x}_\Lambda$ should not necessarily be viewed as an estimator of any quantity, even if there is a stochastic model for $\mathbf{x}$. However, if, for example, $\mathbf{x}$ follows model (1) and $\lambda = \sigma\sqrt{2\log n}$, then $\mathbf{x}_\lambda$ reduces to the estimator $\hat{f}$ described above. We also observe that $\mathbf{x}_0 = \mathbf{x}$, and $\mathbf{x}_\infty$ is a vector whose elements are the sample means of $\mathbf{x}$.

The overwhelming majority of existing applications of wavelet thresholding in statistics involve a *single* reconstruction $\mathbf{x}_\Lambda$, corresponding to a single (possibly vector) threshold $\Lambda$. The canonical example is again the denoising problem in which we typically search for a suitable threshold $\Lambda$ such that $\mathbf{x}_\Lambda$ is a good *estimator* of $f$.

Our main proposition is this paper is to argue that for a range of statistical problems, more insight into the data $\mathbf{x}$ could be gained by producing reconstructions $\mathbf{x}_\Lambda$ for an entire range of thresholds $\Lambda$ and analysing them jointly. How precisely to do this and what insight can be gained is of course problem-dependent, and we will have ample examples in the paper. We first construct the necessary toolbox. Starting with scalar thresholds $\lambda$ (rather than vector thresholds $\Lambda$), we define the TTM of $\mathbf{x}$ as

$$T(\mathbf{x}) = \{\mathbf{x}_\lambda\}_{\lambda=0}^{\bar{d}},$$

where $\bar{d} = \max_{(j,k)\neq(-1,1)} |d_{j,k}|$. It is enough to stop at $\bar{d}$ as $\mathbf{x}_\lambda = \mathbf{x}_\infty$ for $\lambda > \bar{d}$. We will occasionally refer to $T(\mathbf{x})$ as the "basic" Time-Threshold map to differentiate it from the Derivative TTM below.

Each component of the vector $\mathbf{x}_\lambda$ is a piecewise-constant (right-continuous) function of $\lambda$. We define

$$\Delta\mathbf{x}_\lambda = \lim_{h\to 0} \mathbf{x}_\lambda - \mathbf{x}_{\lambda+h}.$$

The interpretation of $\Delta\mathbf{x}_\lambda$ is simple: it is the effect of the inverse wavelet transform $W^{-1}$ applied to only those coefficients $d_{j,k}$ ($(j,k) \neq (-1,1)$) of $\mathbf{x}$ whose absolute value equals exactly $\lambda$. Therefore, it is the "detail" present in the reconstruction $\mathbf{x}_\lambda$ but not in any reconstructions $\mathbf{x}_{\lambda+h}$ for $h > 0$.

The introduction of $\Delta\mathbf{x}_\lambda$ invites the definition of the Derivative TTM of $\mathbf{x}$ as

$$\Delta T(\mathbf{x}) = \{\Delta\mathbf{x}_\lambda\}_{\lambda=0}^{\bar{d}} \cup \{\mathbf{x}_\infty\}.$$

We note that $\Delta T(\mathbf{x})$ provides a decomposition of $\mathbf{x}$ which is orthogonal and invertible, in the sense that

$$\begin{aligned} \langle \Delta\mathbf{x}_{\lambda_1}, \Delta\mathbf{x}_{\lambda_2} \rangle &= 0 \quad \text{if} \quad \lambda_1 \neq \lambda_2 & (3)\\ \mathbf{x} &= \mathbf{x}_\infty + \sum_{\lambda\in[0,\bar{d}]} \Delta\mathbf{x}_\lambda. & (4) \end{aligned}$$

(3) is the result of the fact that a coefficient $d_{j,k}$ cannot simultaneously have magnitude $\lambda_1$ and $\lambda_2$, and of the orthonormality of $W$. In (4), all but at most $n$ terms $\Delta\mathbf{x}_\lambda$ in the range $\lambda \in [0, \bar{d}]$ are zero.

$T(\mathbf{x})$ can be interpreted as a visualisation of how quickly the nonlinear approximation $\mathbf{x}_\lambda$ of $\mathbf{x} = \mathbf{x}_0$ reaches the latter as $\lambda$ decreases. Note that this is presented as a function of the threshold $\lambda$, rather than of the number of terms in the nonlinear approximation. For more on nonlinear approximation, the reader is referred to DeVore (1998).

To define the TTM for vector thresholds $\Lambda$, we restrict our attention to separable thresholds $\Lambda(\lambda)$ for which

$$\lambda_{j,k} = \lambda\, r(\mathbf{x}, j, k)$$

and define

$$T(\mathbf{x}) = \{\mathbf{x}_{\Lambda(\lambda)}\}_{\lambda=0}^{\bar{\lambda}},$$

where

$$\bar{\lambda} = \min\{\lambda : \lambda_{j,k} \geq |d_{j,k}| \quad \forall\, (j,k) \neq (-1,1)\}.$$

We end this section by mentioning that the TTM methodology does *not* provide a new wavelet thresholding procedure; instead, it helps visualise existing procedures in a way that displays more information at once and can therefore potentially lead to improved inference.

# 3 Basic properties and features of the TTMs

In this section, we illustrate some generic features of the Time-Threshold Map, both in its basic and derivative version. As we restrict our attention to Haar and Unbalanced Haar wavelets, which are piecewise-constant, we use the well-known piecewise-constant "blocks" signal (first having appeared in Donoho and Johnstone, 1994) as a running example.

## 3.1 Basic Time-Threshold Map

### 3.1.1 Ordering the importance of features

We provide the first illustration of the TTM on the blocks signal, using the Unbalanced Haar wavelets with the basis selection procedure described in Section 2.1. The middle plot in Figure 1 shows the signal, $\mathbf{x}$, sampled at $n = 1000$ equispaced time points. The top-plot shows the values of the TTM $T(\mathbf{x}) = \{\mathbf{x}_\lambda\}_{\lambda=0}^{\bar{d}}$ (lighter colours correspond to higher values) as a function of time (on the $x$-axis) and $\lambda$ (on the $y$-axis). The threshold parameter $\lambda$ has been sampled at 50 equispaced points between 0 and $\bar{d}$ (thus the size of the plotted matrix is $50 \times 1000$). The bottom plot shows $\mathbf{x}_\lambda$ for $\lambda = \frac{k}{5}\bar{d}$, $k = 0, 1, \ldots, 4$.

Note that each vertical line in the TTM corresponds to a breakpoint in $\mathbf{x}_\lambda$ for the values of $\lambda$ within the range of that particular vertical line. One characteristic of the TTM of a piecewise-constant signal (such as blocks) is that the time-locations of the vertical lines in the TTM correspond exactly to the locations of breakpoints in the input signal $\mathbf{x}$. This is guaranteed by Lemma 2.2 in Venkatraman (1993) which, translated into the notation of our paper, states that the breakpoint $b^{j,k}$ in each selected UH basis vector $\psi_{j,k}$ coincides with one of the breakpoints in $\mathbf{x}$ (provided there are any breakpoints in $\mathbf{x}$ contained in the support of $\psi_{j,k}$).

We further note that the length of each vertical line can be interpreted as a measure of "importance" or "prominence" of the given breakpoint in $\mathbf{x}$. For example, in the blocks signal, the "most prominent" feature is the one defined by the two breakpoints at times $t = 650$ and $t = 810$, since the vertical lines corresponding to these two breakpoints are present for $\lambda \in [0, \bar{d})$, i.e. for the entire permitted range. Similarly, the "least important" feature is the breakpoint at time $t = 780$ (as it corresponds to the shortest vertical line) or, interpreting features as peaks or troughs rather than individual breakpoints, the small trough between times $t = 760$ and $t = 780$ as it is defined by two vertical lines the sum of whose lengths is the shortest among all pairs of vertical lines in the TTM. Similarly, quantities such as the ratio or the difference of the lengths of two vertical lines can serve as a measure of the relative importance of two breakpoints in the signal.

We note at this point that the TTM does not necessarily preserve the order in which the UH basis vectors have been chosen, i.e. the vertical lines corresponding to breakpoints $b^{j,k}$ at the coarsest scales (= those for the lowest values of $j$) are not necessarily the longest (since the corresponding UH coefficients are not necessarily of the largest magnitude). This makes the TTM inherently different from dendrogram-type plots in which, using our terminology, coarser-scale splits would be presented as more prominent than finer-scale ones.

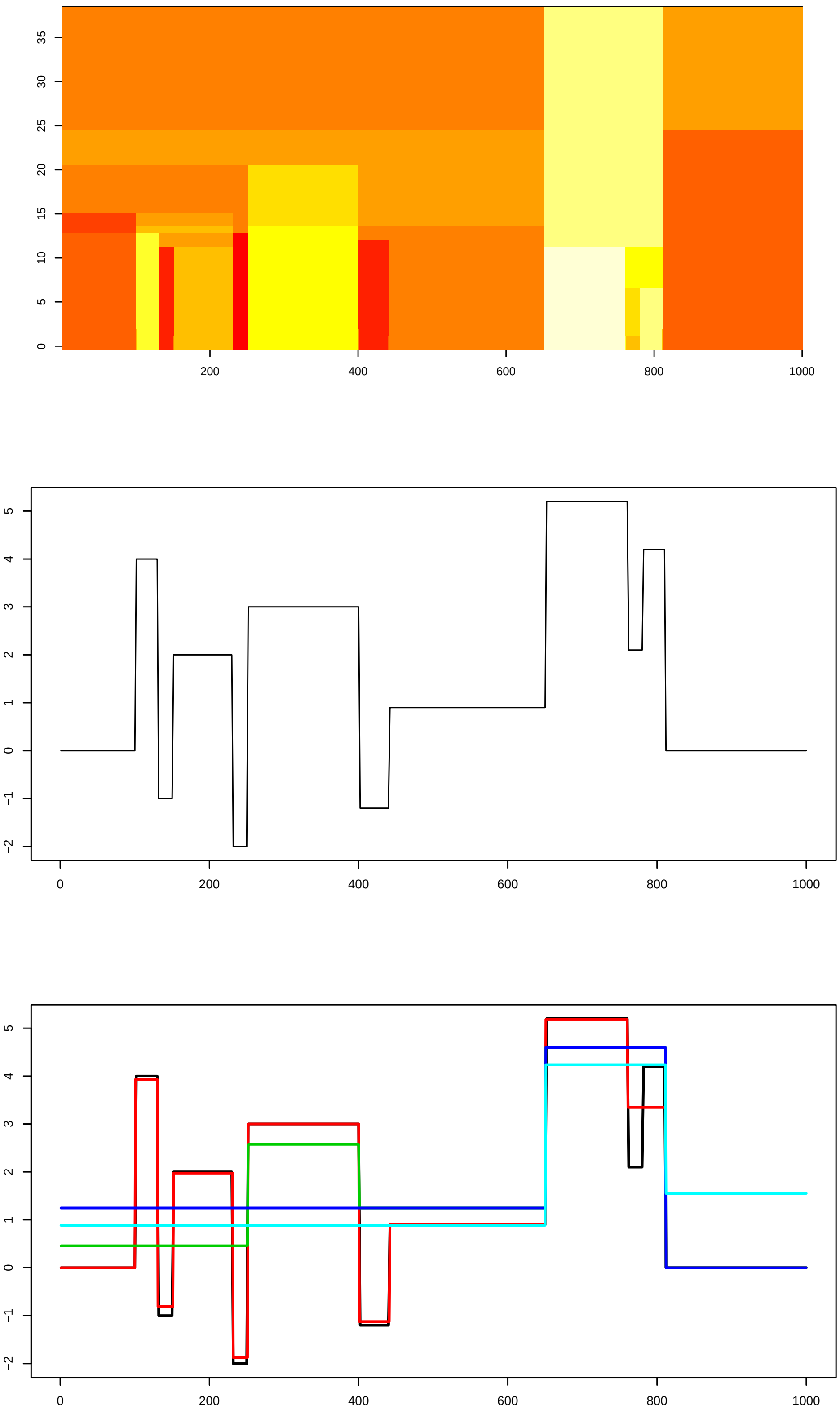

Figure 1: Middle: the blocks signal $\mathbf{x}$. Top: the TTM of $\mathbf{x}$ using Unbalanced Haar wavelets; see Section 3.1.1 for details. Bottom: $\mathbf{x}_\lambda$ for $\lambda = \frac{k}{5}\bar{d}$, $k = 0, 1, \ldots, 4$ (black, red, green, dark blue, light blue).

### 3.1.2 Separation of signal from noise

In Section 3.1.1, we considered the TTM of the noise-free blocks signal. In this section, we illustrate the behaviour of the TTM of the blocks signal with added iid Gaussian noise. We are particularly interested in very noisy set-ups where the human eye cannot be relied on to denoise the signal and more sophisticated techniques are needed.

The middle plot in Figure 2 shows the blocks signal from Section 3.1.1, contaminated with iid Gaussian noise with mean zero and standard deviation 2, resulting in the root-signal-to-noise ratio of 0.952. Although the overall shape of the signal is clear, it would be challenging to an untrained eye to give an accurate estimate of the number of breakpoints in the underlying true signal. The top plot shows the TTM of the noisy blocks, again using Unbalanced Haar wavelets with the basis selection procedure from Section 2.1, with superimposed true locations of the breakpoints in blocks.

It is unsurprising to observe that the bottom part of the TTM shows those reconstructions $\mathbf{x}_\lambda$ which still contain noise. It is more interesting to note that the visibly noisy part of the TTM is confined to a relatively narrow strip of the TTM, reaching only as far as $\lambda = 5$ or $\lambda = 6$ but not above these values. This may give the impression of the TTM being “cleaner” and providing more distinct separation of signal from noise than the plot of the original signal. Indeed, all the true breakpoints in blocks are clearly reflected in the TTM for some values of $\lambda$ significantly above $\lambda = 6$. The spurious features in the TTM occurring before the first true breakpoint are unsurprising given the appearance of this particular simulated data sample in that time region.

The bottom plot in Figure 2, showing $\mathbf{x}_\lambda$ for $\lambda = 7.5$, confirms the good noise separation property of the TTM in this example: the reconstruction is almost perfect except for the spurious break before the first true breakpoint.

We do not formally quantify the above noise-separation property of the TTM in this section; however, we provide some rigorous (albeit asymptotic) results concerning this property in Section 4.1 where we apply it to the problem of breakpoint detection in a particular signal + noise set-up.

We end this section by noting that the noise-separation property can be viewed as an instance of the feature-ordering property from Section 3.1.1: the fact that the vertical lines corresponding to noise tend to be shorter than those corresponding to the signal can be interpreted as noise being “less prominent” than signal in this example.

## 3.2 Derivative Time-Threshold Map

### 3.2.1 Visualising basis vectors on the Time-Threshold plane

In this section, we illustrate some features of the Derivative Time-Threshold Map, using again the blocks example. We now use both Haar and Unbalanced Haar wavelets and therefore consider length $n = 1024$, which is a power of two as required by Haar wavelets (but not by Unbalanced Haar). As in the previous section, we add independent Gaussian noise with mean zero and standard deviation of 2. Algorithmically, it is straightforward to compute the Derivative TTM simply by taking row-wise differences of the matrix representing the basic TTM (whose construction is described in Section 3.1.1).

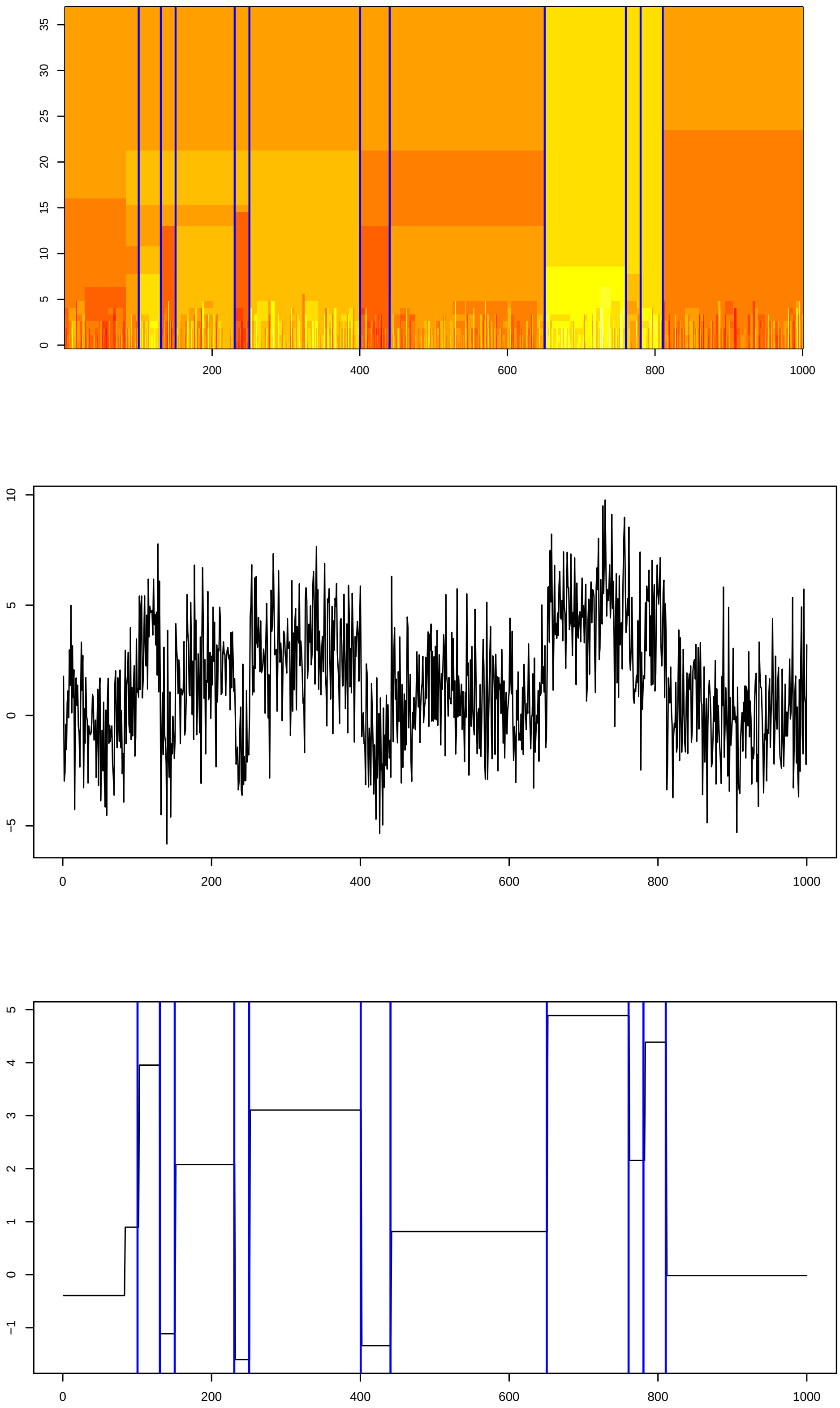

Figure 2: Middle: the noisy blocks signal $\mathbf{x}$. Top: the TTM of $\mathbf{x}$ using Unbalanced Haar wavelets, with true breakpoint locations (blue); see Section 3.1.2 for details. Bottom: $\mathbf{x}_{7.5}$, with true breakpoint locations (blue).

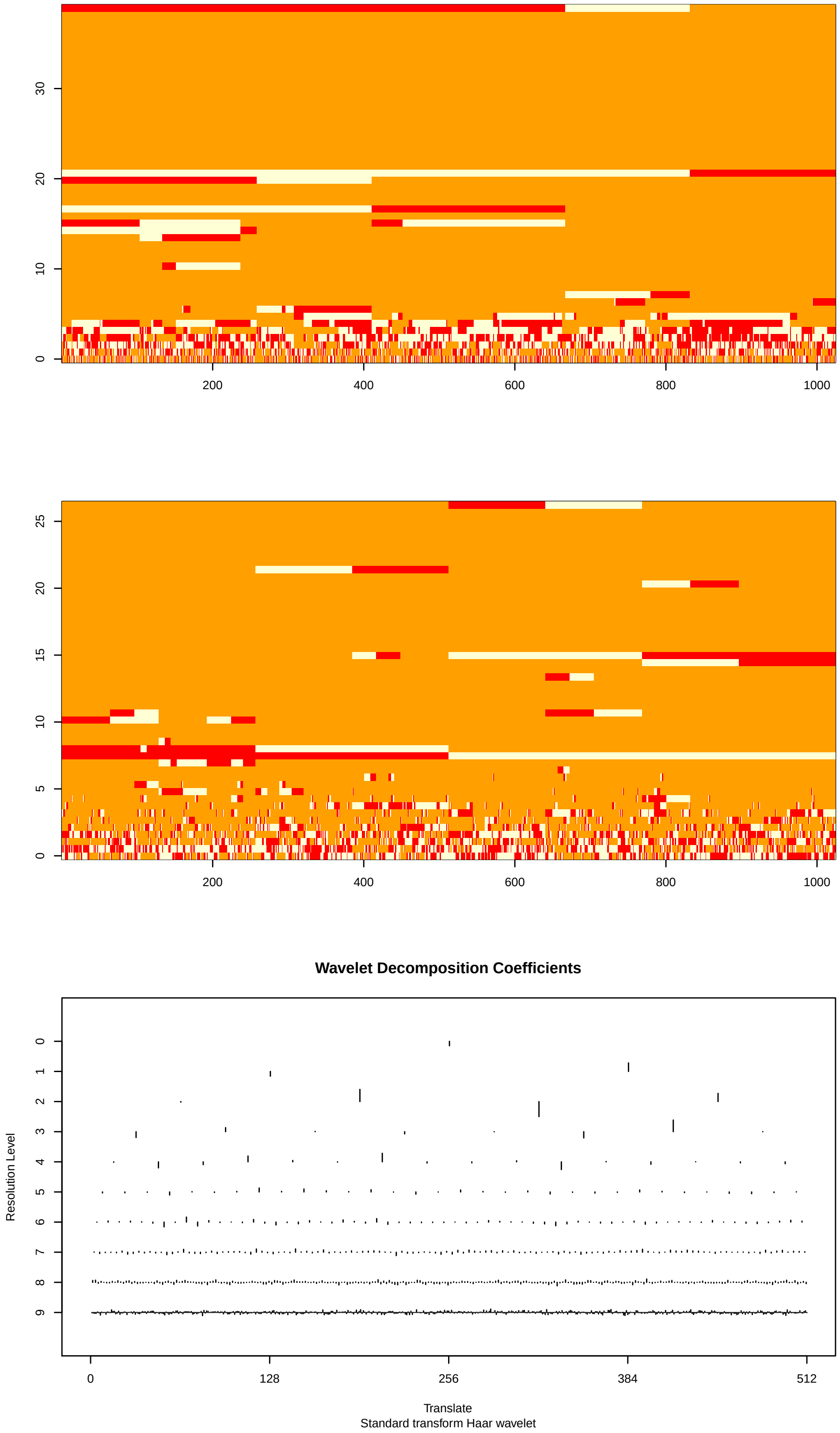

Figure 3: Top: sign of the Derivative TTM of noisy blocks using Unbalanced Haar wavelets; middle: the same using Haar wavelets; bottom: the time-scale plot of noisy blocks. See Section 3.2.1 for details.

The top and middle plots in Figure 3 show the sign of the Derivative TTMs of the noisy blocks. The reason why $\text{sign}\{\Delta T(\mathbf{x})\}$, rather than $\Delta T(\mathbf{x})$ itself, is shown, is that the magnitudes of the non-zero elements in $\Delta T(\mathbf{x})$ are much lower for small values of $\lambda$ than they are for its large values, which makes $\Delta T(\mathbf{x})$ inconvenient to view as a heat map due to insufficient contrast for small values of $\lambda$. The interpretation of the plots is simple: each light-dark colour strip represents a different component (wavelet basis vector $\psi_{j,k}$ times the wavelet coefficient $d_{j,k}$) of $\mathbf{x}$. The $y$-coordinate determines the magnitude of the wavelet coefficient $d_{j,k}$. The light-colour part of the strip coincides with the positive part of the support of $d_{j,k}\psi_{j,k}$, and the dark colour – with the negative part. In other words, the Derivative TTM provides a natural way of visualising the contribution of each wavelet $\psi_{j,k}$ to the input signal.

Traditionally, visualising the wavelet decomposition of a signal has mostly been done via so-called time-scale plots (see e.g. Vidakovic, 1999, Section 3.1), in which, in contrast to the Derivative TTM, the displayed object is the $d_{j,k}$'s, arranged as a function of $k$ (on the $x$-axis) and $j$ (on the $y$-axis). A time-scale plot of the Haar decomposition of the same noisy blocks signal, computed in the R package `wavethresh`, is shown in the bottom plot of Figure 3.

The two approaches, the time-scale plots and the Derivative TTM, aim to visualise the same information in two different ways and should be viewed as 'complementary' rather than 'competing'. However, one scenario in which the Derivative TTM might be a more attractive option than the time-scale plot is the case where the given wavelet system does not have a clear notion of scale. One example of such a wavelet system is the Unbalanced Haar basis. Although notionally, the $j$ parameter in the Unbalanced Haar wavelets $\psi_{j,k}$ is referred to as "scale", parameters such as the shape or frequency characteristics of different Unbalanced Haar wavelets for the same value of $j$ can be dramatically different as they heavily depend on the previously selected Unbalanced Haar basis vectors in the basis selection algorithm as well as on the shape of the input signal over the current sub-interval. Note that this is different from the Haar wavelet case where a wavelet at scale $j$ always has the same shape and length of support. Since the Derivative Time-Threshold Map completely circumvents the notion of "scale" (as it only uses 'time' and 'threshold' as the free variables), it might be a more natural visualisation tool for such types of wavelets.

Another classic example of a wavelet system for which it is not obvious how to define scale is when it arises from a lifting transform, see e.g. Jansen and Oonincx (2005) for an overview of the latter. For lack of a better term, we shall refer to such wavelets as "lifted". Indeed, Knight et al. (2011) in their work on lifted wavelet spectra for time series sampled on a non-equispaced grid provide one particular assignment of "scales" to lifted wavelets, but refer to such scales as "artificial". Many other assignments are possible. Since it avoids the concept of scale altogether, the Derivative Time-Threshold Map might provide a less ambiguous framework in which to visualise and analyse lifted wavelet decompositions.

### 3.2.2 Orthogonal Feature Decomposition via the Derivative TTM

It is a straightforward but interesting consequence of (3) that for any disjoint intervals $[\lambda_1, \lambda_2)$, $[\lambda_3, \lambda_4)$, the vectors $\mathbf{x}_{\lambda_1} - \mathbf{x}_{\lambda_2}$ and $\mathbf{x}_{\lambda_3} - \mathbf{x}_{\lambda_4}$ are exactly orthogonal. To see this, recall the definition of $\mathbf{x}_\lambda$ from (2), the fact that $W^{-1}$ is orthonormal and that the supports of $t(W\mathbf{x}, \lambda_1) - t(W\mathbf{x}, \lambda_2)$ and $t(W\mathbf{x}, \lambda_3) - t(W\mathbf{x}, \lambda_4)$ are disjoint.

The implication is that for any sampling of the threshold parameter $\lambda$, the rows of the Derivative TTM matrix, computed from the basic TTM by row-wise differencing as described in Section 3.2.1, are exactly orthogonal to each other (with some of them possibly being exactly zero).

As an example, consider again the noisy blocks signal from Section 3.2.1 and its Derivative TTM computed using Unbalanced Haar wavelets, where the threshold parameter $\lambda$ has been sampled at six points $\lambda_i$, $i = 1, \dots, 6$, equispaced between 0 and $\bar{d} = 38.91$. Additionally, we denote $\lambda_7 = \infty$.

Figure 4 shows the rows of the Derivative TTM, that is the vectors $\mathbf{x}_{\lambda_i} - \mathbf{x}_{\lambda_{i+1}}$ $(i = 1, \dots, 6)$, which provide the following orthogonal decomposition:

$$\mathbf{x} = \mathbf{x}_{\lambda_7} + \sum_{i=1}^{6} \mathbf{x}_{\lambda_i} - \mathbf{x}_{\lambda_{i+1}}. \tag{5}$$

The fact that the orthogonal components $\mathbf{x}_{\lambda_i} - \mathbf{x}_{\lambda_{i+1}}$ correspond to different features of the input signal $\mathbf{x}$ motivates calling these components the ‘orthogonal features’ of $\mathbf{x}$ and the resolution of identity in (5) – the ‘Orthogonal Feature Decomposition’. Obviously, the number of non-zero features in an Orthogonal Feature Decomposition is bounded from above by the length of the input signal.

At this point, we note the difference between (5) and the wavelet multiresolution decomposition of Mallat (1989): the latter provides a linear decomposition of $\mathbf{x}$ whereas the former is nonlinear. Indeed, the components of a multiresolution decomposition are defined by the scales of the underlying wavelet basis, whereas in the Orthogonal Feature Decomposition they are defined not by scales but by the magnitudes of the wavelet coefficients (hence the nonlinearity). As an aside, note that additionally, in the case of the Unbalanced Haar wavelets (as in the data example considered in this section), the Orthogonal Feature Decomposition is ‘highly nonlinear’ in the sense of DeVore (1998) since the wavelet functions themselves are chosen adaptively.

Finally, we observe that the Orthogonal Feature Decomposition can be viewed as a generalisation of the ‘noise separation’ property of Section 3.1.2 in the sense that it provides a decomposition of the input signal into a larger number of orthogonal components, rather than merely into what can be viewed as ‘signal’ and ‘noise’.

# 4 Possible applications of the TTMs

## 4.1 Aiding breakpoint detection in signals and time series

In this section, we demonstrate how the TTM can be used to improve existing breakpoint detection procedures for signals and time series. We first describe our general model; $C$ below denotes a generic constant. Suppose we observe a realisation of

$$X_t = g_t + \varepsilon_t, \quad t = 1, \dots, n \tag{6}$$

where $g_t$ is close to a piecewise-constant function $\tilde{g}_t$ in the sense that $\sum_{t=1}^{n} |g_t - \tilde{g}_t| \le C$, and $\varepsilon_t$ is a sequence of bounded random variables with mean zero, strongly mixing with a geometric rate $\rho$. The number $N$ of breakpoints in $\tilde{g}$ is bounded by a constant (this can be

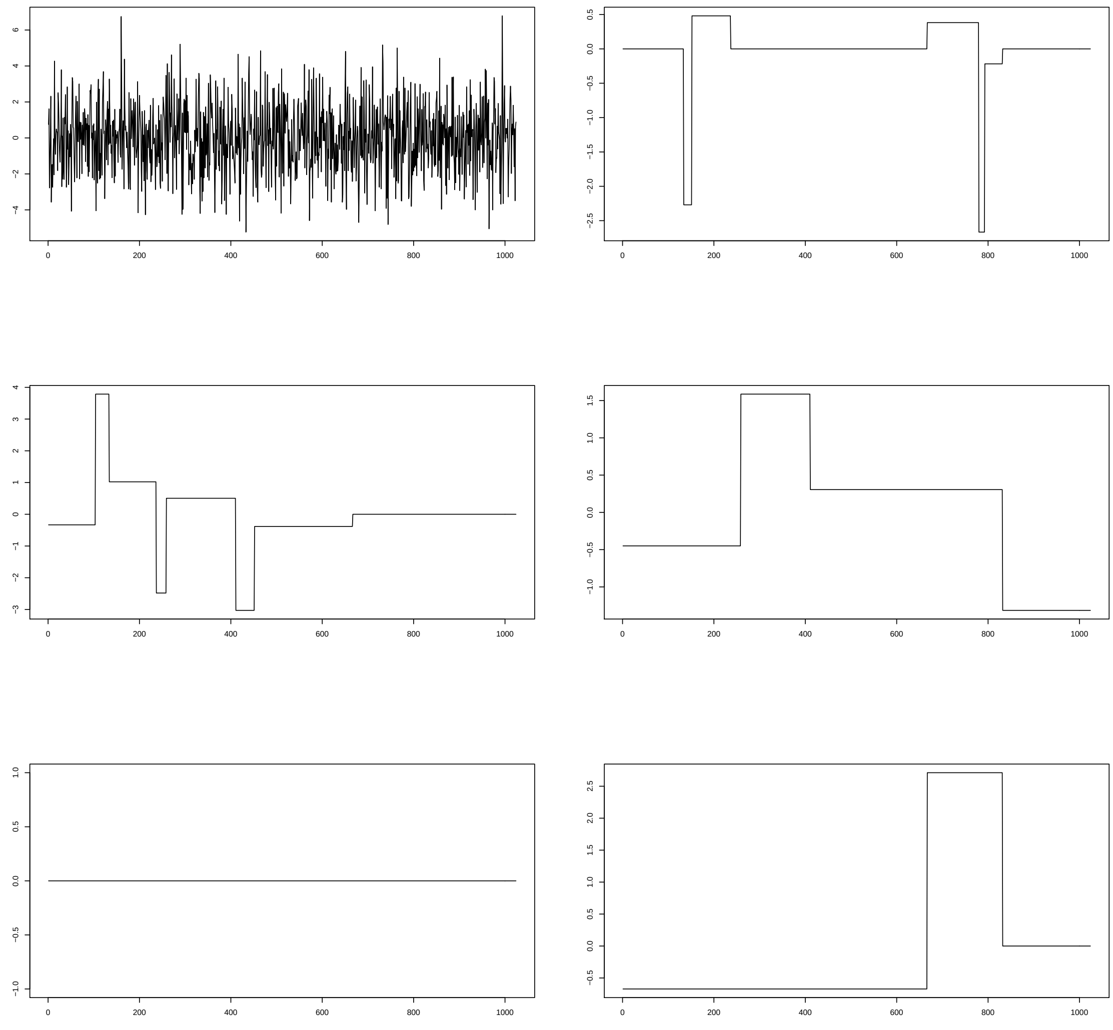

Figure 4: From top to bottom and from left to right: orthogonal features $i = 1, \ldots, 6$ of the noisy blocks signal using Unbalanced Haar wavelets; see Section 3.2.2 for details.

extended to $N$ increasing with $n$ but we do not pursue it in this work) and their locations, sorted in an increasing order, are denoted by $\eta_1, \ldots \eta_N$ with $\eta_0 = 1$ and $\eta_{N+1} = n$. The breakpoints are such that $\inf_{i=0,\ldots,N}\{\eta_{i+1} - \eta_i\} \geq Cn$, and $\inf_{i=1,\ldots,N} |\tilde{g}_{\eta_i} - \tilde{g}_{\eta_i - 1}| \geq C$. The task is to estimate the number $N$ and the locations $\{\eta_i\}_i$ of the breakpoints. The above setting was considered in Fryzlewicz and Subba Rao (2011) in the context of segmenting ARCH processes, which also forms the basis of the examples in this section.

Theorem 4.1 below, which concerns consistent estimation of $N$ and $\{\eta_i\}_i$, is a re-interpretation of Theorem 3.1 from Fryzlewicz and Subba Rao (2011) in the language of the present paper; the proof is identical so we omit it.

**Theorem 4.1** *Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ follow model (6). Let $W$ denote the Unbalanced Haar transform with the basis selection procedure performed on $\mathbf{X}$ as described in Section 2.1. Let the threshold $\lambda_n = cn^\theta$, with $\theta \in (\frac{1}{4}, \frac{1}{2})$ and $c$ being a positive constant. Produce the reconstruction $\mathbf{X}_{\lambda_n} = W^{-1}t(W\mathbf{X}, \lambda_n)$ and denote the number of breakpoints in $\mathbf{X}_{\lambda_n}$ by $\hat{N}$ and their locations, sorted in an increasing order, by $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$. Then there exist positive constants $C$, $\alpha$ such that $P(A_n) \to 1$, where*

$$A_n = \{\hat{N} = N; \quad |\hat{\eta}_i - \eta_i| \leq C\epsilon_n \quad \mathit{for} \quad 1 \leq i \leq N\},$$

*with $\epsilon_n = n^{1/2} \log^\alpha n$.*

We note that the breakpoint detection procedure from Theorem 4.1 is based on a *single* wavelet reconstruction of the input data $\mathbf{X}$. In the spirit of this paper, our proposal in this section is to consider instead a range of reconstructions, i.e. the TTM of $\mathbf{X}$, and argue that it has the potential to lead to improved breakpoint detection. In this section, we always use Unbalanced Haar wavelets to produce the TTM.

We use the following motivating example to introduce our TTM-based methodology. The middle left plot in Figure 5 shows an example of a sequence $X_t$ of length 200, following model (6), arising from a financial time series context which will be described later. The true underlying $\tilde{g}_t$ has one breakpoint at $t = 100$. The method of Theorem 4.1, with the threshold constants $c = 0.5$ and $\theta = 3/8$ (recommended as default in Fryzlewicz and Subba Rao (2011)), leading to the threshold value $\lambda = 3.65$, fails by detecting no breakpoints in this sequence. Consider also the TTM of the sequence $X_t$ (see the top right plot in Figure 5). The TTM appears to indicate the existence of a breakpoint and indeed it shows that the use of any threshold between 2.14 and $\bar{d} = 3.51$ would have lead to the correct detection of the single breakpoint. Since $\bar{d} < \lambda$, we may conclude that the breakpoint detection in this example failed because the breakpoint did not feature prominently enough in the sequence $X_t$ for the default threshold to 'catch' it.

To remedy this issue, we use the TTM of $X_t$ to construct an artificial signal $\tilde{X}_t$ which 'amplifies' the prominent features of $X_t$ (here: the single breakpoint) and suppresses the less prominent ones. We do this by taking the average of the rows of the TTM matrix of $X_t$. Since, as argued in Section 3.1.2, the more prominent breakpoints in $X_t$ are reflected in a larger number of rows of the TTM of $X_t$ than the less prominent ones, the hope is that this construction will lead to a signal $\tilde{X}_t$ in which the prominent features of $X_t$ are exposed in a clearer manner than in $X_t$ itself.

More formally, using the notation $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)^T$, we define

$$\tilde{\mathbf{X}} = \frac{1}{\bar{d}} \int_0^{\bar{d}} \mathbf{X}_\lambda d\lambda. \tag{7}$$

For practical purposes, the integral in (7) is approximated by taking the average of the rows of the TTM matrix of $X_t$.

We now formulate and prove a result which we will interpret to mean that the 'noise separation property' in $\tilde{X}_t$ is stronger than that in $X_t$ itself, i.e. that it is easier to identify the breakpoints in $X_t$ by considering $\tilde{X}_t$ rather than $X_t$.

**Proposition 4.1** *Let $W$ be the Unbalanced Haar transform from Theorem 4.1. Let the threshold $\beta_n = cn^\theta$, with $\theta \in (0, \frac{1}{2})$ and $c$ being a positive constant. Produce the reconstruction $\tilde{\mathbf{X}}_{\beta_n} = W^{-1} t(W\tilde{\mathbf{X}}, \beta_n)$ and denote the number of breakpoints in $\tilde{\mathbf{X}}_{\beta_n}$ by $\hat{N}$ and their locations, sorted in an increasing order, by $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$. Then there exist positive constants $C$, $\alpha$ such that $P(A_n) \to 1$, where*

$$A_n = \{\hat{N} = N; \quad |\hat{\eta}_i - \eta_i| \le C\epsilon_n \quad \textit{for} \quad 1 \le i \le N\},$$

*with $\epsilon_n = n^{1/2} \log^\alpha n$.*

The proof appears in the Appendix. Proposition 4.1 states that reconstructions $\tilde{\mathbf{X}}_{\beta_n}$ consistently estimate the number and locations of the breakpoints in $\tilde{g}_t$ for an asymptotically larger range of thresholds $\beta_n = cn^\theta$ (note the range of the exponent $\theta \in (0, 1/2)$) than the reconstructions $\mathbf{X}_{\lambda_n}$ in which the thresholds are of the form $\lambda_n = cn^\theta$ with $\theta \in (1/4, 1/2)$. This can be interpreted as $\tilde{\mathbf{X}}$ providing better separation of signal from noise than $\mathbf{X}$.

Returning to the example from the start of the section, the middle right plot in Figure 5 shows the sequence $\tilde{X}_t$ corresponding to the sequence $X_t$ from the middle left plot, rescaled such that their sample variances are equal. In confirmation of the above theory, $\tilde{X}_t$ seems to expose the breakpoint in $\tilde{g}_t$ more clearly. Indeed, the breakpoint detection method described in Theorem 4.1, applied to $\tilde{X}_t$ with the default values of $c = 0.5$ and $\theta = 3/8$, leads to one estimated breakpoint, as required.

Motivated by the above theory and the example, we propose the following TTM-based refinement to any breakpoint estimation procedure (not necessarily based on the method from Theorem 4.1):

1. Given an input sequence $X_t$, produce $\tilde{X}_t$ as described above.
2. Rescale $\tilde{X}_t$ so that its sample variance matches that of $X_t$.
3. Use $\tilde{X}_t$, instead of $X_t$, as input to the breakpoint estimation procedure.

We now describe a simulation study which will illustrate the effectiveness of this procedure in the simulation set-up of Fryzlewicz and Subba Rao (2011). They apply the method of Theorem 4.1 to the detection of breakpoints in GARCH processes with piecewise constant parameters. A GARCH(1,1) process $Y_t$, possibly the most widely used model for low-frequency financial return data, has the form

$$\begin{aligned} Y_t &= \sigma_t Z_t \\ \sigma_t^2 &= a_0 + a_1 Y_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{aligned}$$

where $Z_t$ is a sequence of iid innovations with mean zero and variance one (assumed Gaussian in the remainder of this section), and $a_0, a_1$ and $b_1$ are positive constants.

Following Davis et al. (2008), Fryzlewicz and Subba Rao (2011) consider ten GARCH(1,1) models with sample size $n = 1000$, and with at most one breakpoint occuring in the triple $(a_0, a_1, b_1)$ at time $t = 501$ as follows:

(a) $(0.4, 0.1, 0.5) \to (0.4, 0.1, 0.5)$ [note that this model is stationary]

(b) $(0.1, 0.1, 0.8) \to (0.1, 0.1, 0.8)$ [note that this model is stationary]

(c) $(0.4, 0.1, 0.5) \to (0.4, 0.1, 0.6)$

(d) $(0.4, 0.1, 0.5) \to (0.4, 0.1, 0.8)$

(e) $(0.1, 0.1, 0.8) \to (0.1, 0.1, 0.7)$

(f) $(0.1, 0.1, 0.8) \to (0.1, 0.1, 0.4)$

(g) $(0.4, 0.1, 0.5) \to (0.5, 0.1, 0.5)$

(h) $(0.4, 0.1, 0.5) \to (0.8, 0.1, 0.5)$

(i) $(0.1, 0.1, 0.8) \to (0.3, 0.1, 0.8)$

(j) $(0.1, 0.1, 0.8) \to (0.5, 0.1, 0.8)$.

The BaSTA-avg method of Fryzlewicz and Subba Rao (2011) transforms the sample path $Y_t$ into another sequence $X_t$ and searches for breakpoints in the (asymptotic) mean of the latter. For example, a sample path $Y_t$ corresponding to model (c) is shown in the top left plot of Figure 5; the middle left plot shows the corresponding $X_t$.

Table 1 shows the proportion of correctly detected number of breakpoints, averaged over models (a)–(j), for the BaSTA-avg method of Fryzlewicz and Subba Rao (2011) (with the span parameter $s$ set to 5 as recommended in that work) and the TTM-refined BaSTA-avg according to the recipe from this section. The results are shown for a range of values of the threshold constants $c$ and $\theta$. It is encouraging to note that the TTM-refined BaSTA-avg is uniformly better, by a large margin, than the original BaSTA-avg.

The model-by-model breakdown of the results, for the best parameter configurations ($(c, \theta) = (0.5, 0.375)$ for BaSTA-avg and $(c, \theta) = (0.6, 0.375)$ for the TTM-refined BaSTA-avg) is in Table 2.

### 4.2 Visualising intervals of stationarity

In this section, we demonstrate how the Derivative TTM can be of use in the following statistical problem: given observations from the model $X_t = f_t + \varepsilon_t$, $t = \ldots, n-1, n$, where $f_t$ is deterministic and $\varepsilon_t$ are random variables (assumed iid $N(0, \sigma^2)$ for the purpose of this simple illustration), find the longest segment $[n - M, n]$ on which $f_t$ can be approximated by a constant. One instance of this problem arises in financial statistics where it is of interest to estimate the longest "recent" period of constant volatility in financial asset returns to ensure as much stability and accuracy of the estimated current volatility as possible. We expand on this example in our illustration below.

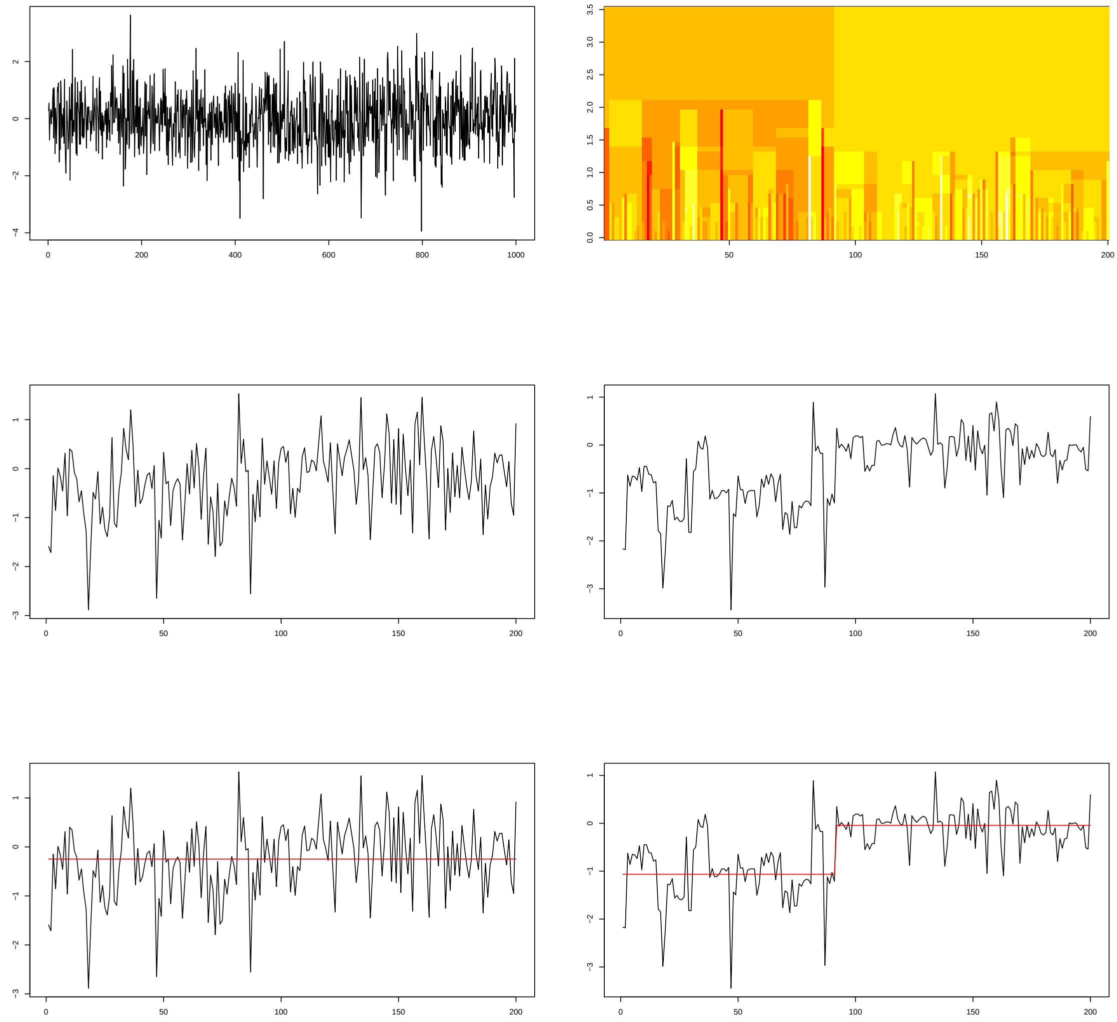

Figure 5: Left column, from top to bottom: sample path $Y_t$ from model (c) of Section 4.1; corresponding $X_t$; result of BaSTA-avg on $X_t$ with $c = 0.5$, $\theta = 3/8$. Right column, from top to bottom: TTM of $X_t$ with Unbalanced Haar wavelets; corresponding $\tilde{X}_t$; result of BaSTA-avg on $\tilde{X}_t$ with $c = 0.5$, $\theta = 3/8$.

| $c \backslash \theta$ | 0.26 | 0.375 | 0.49 |
|---|---|---|---|
| 0.1 | **4**, 0 | **20**, 0 | **51**, 24 |
| 0.2 | **21**, 1 | **54**, 26 | **81**, 76 |
| 0.3 | **40**, 8 | **72**, 64 | **85**, 77 |
| 0.4 | **55**, 30 | **82**, 77 | **81**, 64 |
| 0.5 | **67**, 52 | **85**, 79 | **79**, 51 |
| 0.6 | **73**, 65 | **85**, 76 | **76**, 52 |
| 0.7 | **79**, 75 | **83**, 69 | **69**, 42 |
| 0.8 | **83**, 78 | **81**, 63 | **61**, 30 |
| 0.9 | **85**, 79 | **80**, 60 | **56**, 22 |
| 1 | **85**, 79 | **78**, 56 | **45**, 20 |

Table 1: Number of simulation runs (out of 100) for which the number of breakpoints has been correctly detected, averaged over models (a)–(j) from Section 4.1, for BaSTA-avg with span $s = 5$ (normal font) and TTM-refined BaSTA-avg with span $s = 5$ (bold font).

| Method | (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) | (i) | (j) |
|---|---|---|---|---|---|---|---|---|---|---|
| BaSTA-avg | 1.00 | 0.99 | 0.11 | 0.98 | 0.88 | 0.99 | 0.02 | 0.98 | 0.98 | 0.98 |
| TTM BaSTA-avg | 1.00 | 0.86 | 0.37 | 0.99 | 0.98 | 1.00 | 0.29 | 1.00 | 0.98 | 1.00 |

Table 2: Proportion of simulation runs for which the number of breakpoints has been correctly detected, for models (a)–(j) from Section 4.1, for BaSTA-avg with span $s = 5$ ($(c, \theta) = (0.5, 0.375)$) and TTM-refined BaSTA-avg with span $s = 5$ ($(c, \theta) = (0.6, 0.375)$).

On first glance, a wavelet-based solution to the above problem appears straightforward. Indeed, one can take a sample $\mathbf{X}^{(N)} = \{X_t\}_{t=n-N+1}^{n}$ of dyadic length $N = 2^J$, take its Haar wavelet decomposition, apply the universal threshold $\lambda = \sigma\{2 \log N\}^{1/2}$ and take the inverse Haar transform of the thresholded coefficients to obtain a piecewise-constant reconstruction $\mathbf{X}_{\lambda}^{(N)}$. In the case when $f_t = f(t/N)$ and the function $f(u)$ is piecewise-constant, the last constant segment in $\mathbf{X}_{\lambda}^{(N)}$ is, with probability tending to one with $N$, the longest segment of a dyadic length that is still contained within the last segment of constancy of $f_t$. The proof of this statement is straightforward and proceeds along the lines of Fryzlewicz (2010b), Theorem 3.3, so we omit it.

However, one issue with this simple solution is that its theoretical validity heavily relies on the “rescaled time” (a.k.a. “in-fill asymptotics”) concept which assumes that $f_t$ is a sampled version of a compactly supported function where the sampling grid becomes finer and finer as the sample size increases. This framework may be difficult to justify in some time series contexts, such as in the volatility example from this section, where an increasing sample size simply means progression of time rather than finer sampling in the past.

In the spirit of the TTM approach, the alternative solution proposed in this section is to apply a *family* of thresholds and investigate the lengths of the last segments of constancy yielded by each of them. This would give the analyst a more complete picture of the data and facilitate the final choice of the desired interval. One possibility would be, for example, to consider all possible samples $\mathbf{X}^{(N)} = \{X_t\}_{t=n-N+1}^{n}$ for $N = 2^1, 2^2, 2^3, \ldots$, apply the corresponding universal thresholds and investigate what intervals of constancy they lead to. This is the route that we follow in the example below.

The example we consider in this section is that of the Dow Jones Industrial Average index, one of the most widely reported stock market indices. Let $Y_t$ denote the logged and differenced daily closing values of the DJIA index on 10240 consecutive trading days (roughly 40 trading years) ending on 21 October 2011. Local constancy of the expectation of $Y_t^2$ would correspond to the volatility of the returns being constant over the corresponding time interval.

To stabilise the variance of $Y_t^2$ and bring the data closer to the "function + iid Gaussian noise" set-up, we take local averages of $Y_t^2$ over non-overlapping windows of 5 days, and take the logarithmic transform. To be more precise, we form

$$X_t = \log \left\{ \frac{1}{5} \sum_{s=5(t-1)+1}^{5t} Y_s^2 \right\}, \quad t = 1, \ldots, 2048.$$

$X_t$ is shown in the top left plot of Figure 6. From initial inspection, it appears sensible to model $X_t$ as $X_t = f_t + \varepsilon_t$, $t = 1, \ldots, n = 2048$ with $\sigma = \mathrm{Var}^{1/2}(\varepsilon_t) \approx 0.68$. The sign of the Derivative TTM of $X_t$ is shown in the top right plot of Figure 6.

The bottom left plot of Figure 6 shows the portion of the Derivative TTM map of $X_t$ for $t = (2048-64+1), \ldots, 2048$, i.e. for the last 64 time units (the $y$-axis has also been changed for clarity). The horizontal black lines show the family of universal thresholds corresponding to sample sizes $N = 2048, 1024, \ldots, 2$ (obviously, the larger the sample size, the higher the threshold). For each threshold, the implied longest segment of constancy $[n - M, n]$ is the longest segment for which *there are no sign changes in any wavelet functions that appear above the given threshold* (which means that the reconstruction over that segment is constant). It is remarkable to note that *all* considered threshold values indicate that the longest segment of constancy is the segment $[n - 7, n]$ of length 8. The constancy of the expectation of $X_t$ in the last 8 time units corresponds to the constancy of the expectation of $Y_t^2$ over the last 40 days. $Y_t$ for the last 80 days, with the location of the beginning of the last estimated segment of constancy, is shown in the bottom right plot of Figure 6. The answer is visually plausible. One way of summarising/interpreting this outcome would be to say that "it is significant for sample sizes $N = 2, \ldots, 2048$ with respect to the corresponding universal thresholds", which is a stronger statement than if one were to say "it is significant for one particular sample size with respect to the corresponding universal threshold" (as would the likely conclusion be if one were to apply the usual "rescaled time" asymptotics).

We note that the role of the TTM in the example of this section is mainly that of a "visualiser": it provides a natural and convenient framework in which to visually read and interpret a family of solutions, each yielded by a different threshold, at once.

### 4.3 Discussion: other possible uses of the TTMs

We end by listing two other statistical problems where we envisage that the TTM approach could possibly be of use.

1. *Classification of time series, curves or signals.* In this task, the TTM or Derivative TTM could serve as a classification "signature" of a given dataset in the sense that the membership of a class would be determined by certain properties of the (Derivative) TTM of the dataset.

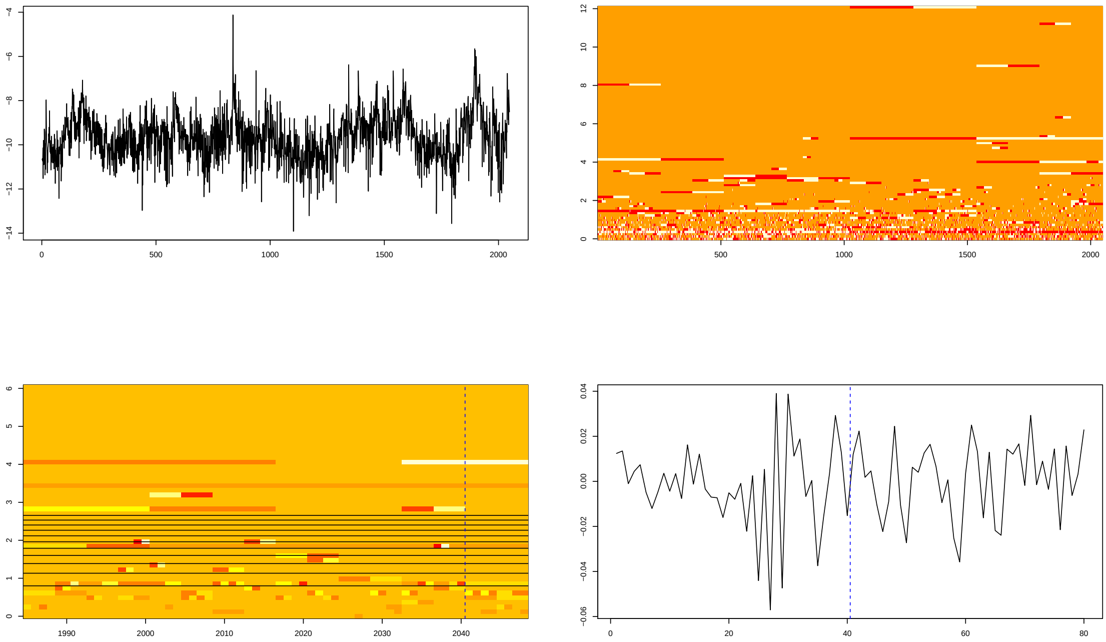

Figure 6: From top to bottom and from left to right: $X_t$ from Section 4.2; the sign of its Derivative TTM using Haar wavelets; a zoom-in of its Derivative TTM with superimposed universal thresholds for $N = 2, 4, 8, \ldots, 2048$ (black lines) and the starting point of the longest estimated interval of constancy (dashed blue line); the last 80 observations of $Y_t$ with the corresponding starting point.

2. *Testing for time series stationarity.* Since the Derivative TTM is invertible, its distribution completely determines the distribution of the input time series. Therefore, tracking changes in the distribution of the Derivative TTM over time could possibly help in assessing the (strict) stationarity, or otherwise, of the input time series.

# A Proof of Proposition 4.1

Let $d_{j,k}$ denote the Unbalanced Haar coefficients of $X_t$, obtained by applying the matrix $W$. Let $\tilde{d}_{j,k}$ denote the Unbalanced Haar coefficients of $\tilde{X}_t$ obtained by applying the same matrix. Below, $C$ denotes a generic positive constant, not necessarily the same in value each time it is used. From (7), we obtain

$$\begin{aligned} \tilde{\mathbf{X}} &= \frac{1}{\bar{d}} \int_0^{\bar{d}} W^{-1} t(d_{j,k}, \lambda) d\lambda \\ \tilde{d}_{j,k} &= \frac{1}{\bar{d}} \int_0^{\bar{d}} t(d_{j,k}, \lambda) d\lambda, \end{aligned}$$

which leads to

$$\tilde{d}_{j,k} = \text{sign}(d_{j,k}) \frac{d_{j,k}^2}{\bar{d}}$$

for $(j,k) \neq (-1,1)$ (and $\tilde{d}_{-1,1} = d_{-1,1}$). We consider separately two cases, Case 1 when the number $N$ of breakpoints satisfies $N \geq 1$ and Case 2 when $N = 0$. The terminology below partly follows Fryzlewicz and Subba Rao (2011).

*Case 1.* We divide the double indices $(j,k) \neq (-1,1)$ into two subsets: those for which the corresponding $b^{j,k}$ estimate previously undetected breakpoints (denote this set by $\mathcal{I}_1$) and the remaining ones ($\mathcal{I}_2$). By Lemma A.6 of Fryzlewicz and Subba Rao (2011), on the set $A_n$, the $d_{j,k}$ from the set $\mathcal{I}_1$ satisfy $|d_{j,k}| \geq C(n^{1/2} - \log^\alpha n)$. By similar arguments, $\bar{d} \leq C(n^{1/2} + \log^\alpha n)$. This leads to the $\tilde{d}_{j,k}$'s from the set $\mathcal{I}_1$ satisfying

$$|\tilde{d}_{j,k}| \geq C(n^{1/2} - \log^\alpha n).$$

On the other hand, by Lemma A.7 of Fryzlewicz and Subba Rao (2011), on the set $A_n$, the $d_{j,k}$ from the set $\mathcal{I}_2$ satisfy $|d_{j,k}| \leq C(n^{1/4} \log^{\alpha/2} n + \log^\alpha n)$, which leads to $\tilde{d}_{j,k}$'s from the set $\mathcal{I}_2$ satisfying

$$|\tilde{d}_{j,k}| \leq C \log^\alpha n.$$

Thus, any threshold of the form $\beta_n = cn^\theta$, $\theta \in (0, 1/2)$, successfully separates, for $n$ large enough, between the two groups of coefficients, leading to $\tilde{\mathbf{X}}_{\beta_n}$ coinciding with $\mathbf{X}_{\alpha_n}$ from Theorem 4.1 and thereby ensuring consistency of the estimated number and locations of the breakpoints with identical rates.

*Case 2.* By Lemma A.4 of Fryzlewicz and Subba Rao (2011), on the set $A_n$, we have $|d_{j,k}| \leq C \log^\alpha n$. Because $|\tilde{d}_{j,k}| \leq |d_{j,k}|$, the same upper bound applies to $|\tilde{d}_{j,k}|$, and thus $\beta_n$ sets all $\tilde{d}_{j,k}$'s to zero, thereby ensuring that $\hat{N} = 0$ as required. □

# References

F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.*, 22:351–361, 1996.

A. Antoniadis. Wavelet methods in statistics: some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.

P. Besbeas, I. De Feis, and T. Sapatinas. A comparative study of wavelet shrinkage estimators for Poisson counts. *Int. Statist. Review*, 72:209–237, 2004.

B. Brodsky and B. Darkhovsky. *Nonparametric Methods in Change-Point Problems.* Kluwer Academic Publishers, 1993.

P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94:807–823, 1999.

I. Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, Pa., 1992.

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

R. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

P. Fryzlewicz. Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Stat. Assoc.*, 102:1318–1327, 2007.

P. Fryzlewicz. Wavelet methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:654–667, 2010a.

P. Fryzlewicz. Haar-Fisz wavelet method for interpretable estimation of large, sparse, time-varying volatility matrices. *Preprint*, 2010b.

P. Fryzlewicz and H.-S. Oh. Thick-pen transformation for time series. *Journal of the Royal Statistical Society Series B*, 73:499–529, 2011.

P. Fryzlewicz and S. Subba Rao. BaSTA: consistent multiscale multiple change-point detection for ARCH processes. *Preprint*, 2011.

W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications (Lecture Notes in Statistics).* Springer, 2000.

M. Jansen and P. Oonincx. *Second Generation Wavelets and Applications.* Springer, 2005.

I. Johnstone and B. Silverman. Empirical Bayes selection of wavelet thresholds. *Ann. Statist*, 33:1700–1752, 2005.

M. Knight, M. Nunes, and G.P. Nason. Spectral estimation for locally stationary time series with missing observations. *Statistics and Computing,* to appear, 2011.

S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, 11:674–693, 1989.

S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Sig. Proc.*, 41:3397–3415, 1993.

G. P. Nason. Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, 58: 463–479, 1996.

G. P. Nason. *Wavelet Methods in Statistics with R.* Springer, New York, 2008.

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

M. H. Neumann. Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *Journal of Time Series Analysis*, 17:601–633, 1996.

A. Sen and M.S. Srivastava. On tests for detecting change in mean. *Ann. Stat.*, 3:98–108, 1975.

E.S. Venkatraman. *Consistency results in multiple change-point problems.* PhD thesis, Stanford University, 1993.

B. Vidakovic. *Statistical Modeling by Wavelets.* Wiley, New York, 1999.