Statistica Sinica **14**(2004), 395-412

A PEAKS OVER THRESHOLD MODEL FOR CHANGE-POINT DETECTION BY WAVELETS

Marc Raimondo and Nader Tajvidi

The University of Sydney and Lund University

Abstract: Newly available wavelet bases on multi-resolution analysis have exciting implications for detection of change-points. By checking the absolute value of wavelet coefficients one can detect discontinuities in an otherwise smooth curve even in the presence of additive noise. In this paper, we combine wavelet methods and extreme value theory to test the presence of an arbitrary number of discontinuities in an unknown function observed with noise. Our approach is based on a Peaks Over Threshold modelling of noisy wavelet transforms. Particular features of our method include the estimation of the extreme value index in the tail of the noise distribution. The critical region of our test is derived using a Generalised Pareto Distribution approximation to normalised sums. Asymptotic results show that our method is powerful in a wide range of medium size wavelet frequencies. We compare our test with competing approaches on simulated examples and illustrate the method on Dow-Jones data.

Key words and phrases: Change point, general Pareto distribution, nonparametric regression, peaks over threshold, tail exponent wavelets.

1. Introduction

The emergence of explicit orthonormal bases on multi-resolution analyses, Daubechies (1992), provides an inspiring tool for applied and theoretic statistical problems, see e.g., Härdle, Kerkyacharian, Picard and Tsybakov (1998). Wavelet bases offer a degree of localisation in space as well as in frequency that enables decomposition of a signal into compactly supported oscillating components. The coefficients associated with each of those components are called wavelet coefficients or wavelet transform. A remarkable property of wavelet coefficients is to reflect the local regularity of the original function, being large where the function is irregular and small where the function is smooth. This property is very useful to detect discontinuities or sharp changes in a noisy signal. Wang (1995) has proposed a test statistic based on the optimisation of the absolute value of the wavelet coefficients, Odgen and Parzen (1996) have presented a method based on the cumulative sum of squared wavelet coefficients. These procedures detect "jump" or "cusp" in a differentiable function observed with noise. A key distribution in these procedures is that of the maximum of noisy wavelet transforms.

Generally, critical regions for existing wavelet-based tests are obtained by using asymptotic theory for the maxima of Gaussian processes. In this paper we propose an alternative method to model extreme values of noisy wavelet transforms. Our approach is based on the recent Peaks Over Threshold models, Davison and Smith (1990). The idea, which goes back to Pickands (1975), is that excesses over a high threshold, asymptotically, follow a Generalised Pareto Distribution (GPD). At the finest wavelet frequency we use a GPD-approximation to noisy wavelet transforms to estimate the extreme value index in the tail of the noise distribution, and we present an asymptotic critical region for wavelet change-point detection taking into account such information. Our wavelet change-point detection method can be applied to any resolution level (or wavelet frequency) although asymptotic theory shows that for optimal results one should use medium size frequencies. Our results also suggest that lower frequencies should be used to detect change-points in case of heavy-tailed perturbations. We compare our test with existing methods on simulated examples and illustrate the practical interest of our procedure using the daily closing Dow-Jones index in the period 1967-2000. Alternative approaches, related works and additional references may be found in Müller (1999), Lio and Vannucci (2000), Huh and Carriere (2002) and Raimondo (2002).

The paper is organised as follows: Section 2 presents our model and assumptions, as well as a review on wavelets. Section 3 details the GPD-approximation to noisy wavelet transforms. In Section 4, we derive a peaks over threshold model for wavelet change-point detection. Numerical properties and finite sample behaviour of our method are studied in Section 5. Proofs are summarised in Section 6.

2. Preliminaries

2.1. Model and assumptions

Suppose we observe

$$Y_i = f(i/n) + \mathcal{E}_i,\ i = 1, \ldots, n, \tag{1}$$

where $(\mathcal{E}_i)_{i=1,\ldots,n}$ are centred i.i.d. random variables and f is an unknown mean contribution.

Hypotheses:

- $\mathcal{H}_0$: f is smooth (at least continuously differentiable on $[0, 1]$).
- $\mathcal{H}_1(m)$: f has "at-least 1 and at-most m" jump points and is otherwise smooth.

From observations (1) we wish to test $\mathcal{H}_0$ against $\mathcal{H}_1(m)$. We assume that the number, the locations, and the sizes of jump points in the function f are unknown. However, we suppose that a realistic upper bound to the number of change-points to be tested is known. In our assumption $\mathcal{H}_1(m)$, m denotes such an upper bound, it is supposed to be known. A typical example of $\mathcal{H}_0$, respectively $\mathcal{H}_1$, is given on the left, respectively right, panels of Figure 1.

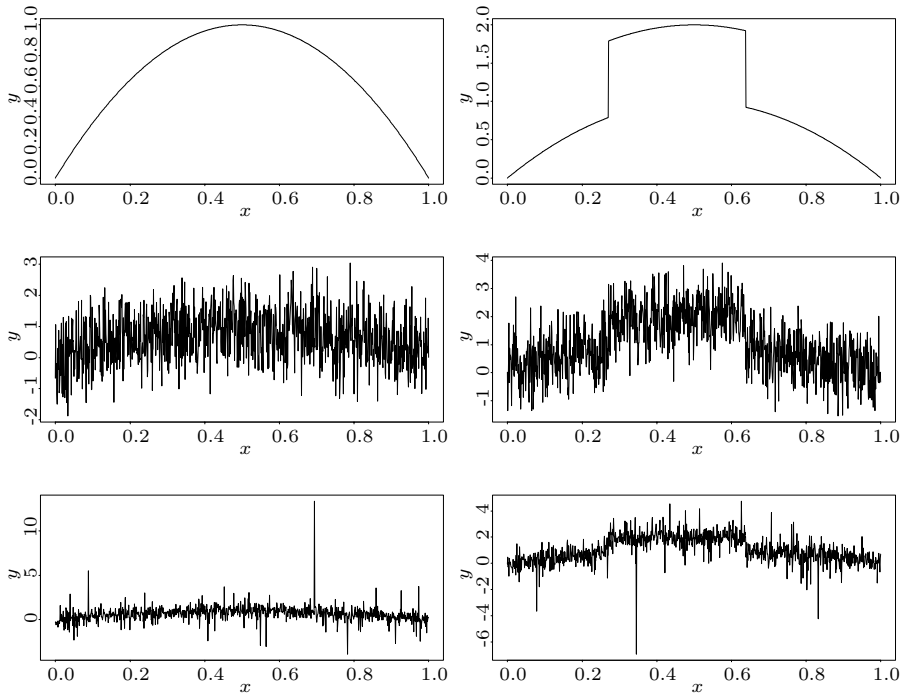

Figure 1. Illustrations of the model (1) and assumption (C_γ) with $n = 1,024$. Top plots: noise-free functions under $\mathcal{H}_0$: $f(x) = 4x(1-x)$ (left) and under $\mathcal{H}_1 : g(x) = f(x) + \mathbb{1}(x > 0.25) - \mathbb{1}(x > 0.5)$ (right). Middle plots: f and g observed with Gaussian noise $\mathcal{N}(0, 0.75^2)$, here $\gamma = 0$ (light-tail). Bottom plots: f and g observed with Student-t_3 noise (standard deviation = 0.75), here $\gamma = -1/3$ (heavy-tail).

Assumption (C_γ). We assume that the tail of the noise depends on γ as follows: for $\gamma > 0$, $|\mathcal{E}_i| \le B$ and $P(|\mathcal{E}_i| > B - x^{-1}) \sim x^{-\frac{1}{\gamma}} L(x)$; for $\gamma \le 0$, we assume that

$$P(|\mathcal{E}_i| > x) \sim \begin{cases} b \exp(-cx), & \text{if } \gamma = 0, \\ \\ x^{1/\gamma} L(x), & \text{if } -\frac{1}{2} < \gamma < 0, \end{cases}$$

where b, c and B are positive constants and $L(x)$ is a slowly varying function.

Remarks. The assumption (C_γ) gives a typical description of distributions which belong to the domain of attraction of an extreme value law. The three cases $\gamma > 0$, $\gamma = 0$ and $\gamma < 0$ corresponding, respectively to the Weibull, Gümbel and Fréchet extreme value distributions, see e.g., Embrechts, Klüppelberg and Mikosch (1997). A distribution with $\gamma < 0$ (respectively, $\gamma \geq 0$) is referred to as heavy-tailed (respectively, light-tailed) distribution. Note that we restrict ourselves to $\gamma > -1/2$ so that the $\mathrm{Var}\,(\mathcal{E}_1) < \infty$. To ease notation, we often assume that $\mathrm{Var}\,(\mathcal{E}_1) = 1$.

To simplify the presentation, we suppose that $\mathcal{E}_1 =^d -\mathcal{E}_1$ although such a symmetry assumption can be removed with minor technical modifications. Indeed, our results apply to any distribution with γ being the index of the heavier of the two tails.

We consider asymptotics in the sample size n, and the notation: $d_n \sim e_n$ means $\lim_{n\to\infty}(d_n/e_n) = 1$.

2.2. Wavelet transforms and local regularity

Wavelet coefficients are discrete transformations of a so-called "mother" wavelet Ψ. First, a doubly indexed family of wavelets is generated by dilating and translating Ψ, $\Psi_{j,k}(u) = 2^{j/2}\Psi(2^j u - k)$, $j \in \mathbb{N}$, $k \in \mathbb{Z}$. Wavelet coefficients are defined by

$$\int f(u)\Psi_{j,k}(u)du, \quad j \in \mathbb{N},\, k \in \mathbb{Z}.$$

The operator which associates wavelet coefficients with a given function f is called the Discrete Wavelet Transform (DWT). Mallat's pyramid algorithm, Mallat (1989), is implemented in the wavelet package *wavetresh* of Nason and Silverman (1994) and can be used to compute the wavelet coefficients of any n-sampled signal $(w_{j,k}) = W(Y)$, where $Y = (Y_1, \ldots, Y_n)$ as in (1) and $(w_{j,k})$ are n-empirical wavelet coefficients. W is an orthogonal transformation which depends on the choice of the wavelet family. The index j, $0 \leq j \leq J$ with $2^{J+1} = n$, is called the resolution level and corresponds to a frequency of 2^{-j}. The index k, $k = 0, 1, \ldots, 2^j - 1$, is called the time (or space) parameter and corresponds to the dyadic position $k/2^j$. To simplify the exposition we denote by w_k the wavelet coefficient computed from the data (1) at resolution level j and time position $k/2^j$. Hence, for any resolution level j and index $k = 0, 1, \ldots, 2^j - 1$, in the wavelet domain (1) becomes

$$w_k = w_k(f) + w_k(\mathcal{E}). \tag{2}$$

Following Härdle, Kerkyacharian, Picard and Tsybakov (1998, p.183), and approximating integrals by sums (taking into account the L^2-normalisation of the

empirical wavelet transform),

$$w_k(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{j,k}(i/n) f(i/n), \quad w_k(\mathcal{E}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{j,k}(i/n) \mathcal{E}_i. \tag{3}$$

We recall some basic properties which can be found in Raimondo (1998).

(i) Under $\mathcal{H}_0$, the function f is differentiable so that for all resolution levels $j \geq j_0 = 3$ and all $k = 0, 1, \ldots, 2^j - 1$, $|w_k(f)| \leq c_1 (n 2^{-3j})^{\frac{1}{2}}$.

(ii) Under $\mathcal{H}_1$, there exists at least a point $x \in [k/2^j, (k+1)/2^j)$ where f has a jump so that $|w_k(f)| \geq c_2 (n 2^{-j})^{\frac{1}{2}}$. Note that a translation invariant wavelet family should be used to ensure that (ii) holds at any resolution level j, see Section 5.2 for details.

The constants c_1, c_2 depend only on f. To simplify the exposition we take $c_1 = c_2 = 1$.

2.3. The CLT approach to noisy wavelet transforms

A classical approach to model noisy wavelet transforms invokes the Central Limit Theorem (CLT). We illustrate the argument using the Haar basis. We recall that the Haar wavelet is the step function $\Psi(x) = \mathbb{1}_{[0,1/2)}(x) - \mathbb{1}_{[1/2,1)}(x)$, so that the support of the $\Psi_{j,k}$-wavelet is exactly the dyadic interval $[k/2^j, (k+1)/2^j)$. This implies the following two important consequences.

- At any level j, the coefficients $w_k(\mathcal{E}), k = 0, 1, \ldots, 2^j$ are independent random variables. (Since the supports of $\Psi_{j,k}, k = 0, 1, \ldots, 2^j$ provide a partition of $[0, 1]$, we see from (3) that the $\mathcal{E}_i$'s involved in $w_k(\mathcal{E})$ are independent of the $\mathcal{E}_i$'s involved in $w_{k'}(\mathcal{E}), k \neq k'$.)

- The number of points i/n, $i = 1, 2, \ldots$ in the dyadic interval $[k/2^j, (k+1)/2^j)$ is $l_n = n/2^j$. By symmetry, we have that

$$w_k(\mathcal{E}) =^d \frac{1}{\sqrt{n}} \sum_{i=1}^{n/2^j} 2^{j/2} (\pm 1) \mathcal{E}_i =^d \frac{1}{\sqrt{l_n}} \sum_{i=1}^{l_n} \mathcal{E}_i = \frac{1}{\sqrt{l_n}} S_{l_n}. \tag{4}$$

  Hence, if $\mathrm{Var}\,(\mathcal{E}_1) = 1$ and $l_n = n/2^j \to \infty$

$$w_k(\mathcal{E}) \longrightarrow^d \mathcal{N}(0, 1), \quad \text{as } n \to \infty. \tag{5}$$

Since the pioneer work of Donoho and Johnstone (1994) on Universal thresholding (see Section 4.3), the CLT approach has been extensively used in the statistical wavelet literature, see e.g., Härdle, Kerkyacharian, Picard and Tsybakov (1998, Chap. 10 and 11). While there is no doubt that, in the limit model, wavelet

coefficients are normally distributed, Wang (1995), for finite sample size, the quality of approximation (5) depends on the parent distribution as well as on the number of terms in the sum (4). Thus, for heavy tailed noise, one may wish to find an alternative to (5).

3. The GPD Approach to Noisy Wavelet Transforms

In this section, we propose an approximation to noisy wavelet transforms based on the Generalised Pareto Distribution (GPD). Our goal is to study the effect of the tail exponent γ of the parent distribution on the distribution of noisy wavelet coefficients. Of particular interest to us, is the effect of γ on the distribution of the maximum (in absolute value) of the wavelet coefficients at any resolution level j.

3.1. Generalised Pareto Distributions

A rather recent approach for modelling extreme events is based on so-called peaks over threshold methods, Davison and Smith (1990). The basic model uses the Generalised Pareto Distribution (GPD) for modelling exceedances of a random variable over a high threshold. The GPD is defined as

$$H(x;\gamma,\sigma) = 1 - (1 - \gamma\frac{x}{\sigma})^{\frac{1}{\gamma}}, \qquad 1 - \gamma\frac{x}{\sigma} > 0.$$

Here $\sigma > 0$ and γ are real parameters and the support of the distribution is $x > 0$ for $\gamma < 0$ and $0 < x < \sigma/\gamma$ for $\gamma > 0$. For $\gamma = 0$ we interpret H to be the exponential distribution $H(x) = 1 - e^{-x/\sigma}$, $x > 0$. Pickands (1975) has shown that the GPD arises as a limiting distribution for the excess over large thresholds. Under (C_γ), for all $x > 0$,

$$P_u(|\mathcal{E}_1| - u > x) \sim_u \bar{H}(x;\gamma,\sigma_u) = 1 - H(x;\gamma,\sigma_u), \tag{6}$$

where P_u denotes the conditional probability given that $\{|\mathcal{E}_1| > u\}$, and $u \to B_\gamma$ or ∞ according to (C_γ). In (6), σ_u is a scale parameter which depends on u as well as on the variance of the parent distribution. More importantly, the shape parameter γ in the GPD-fit (6) is the same as the tail exponent in our assumption (C_γ).

3.2. The GPD-paradigm

Under (C_γ) the approximation (6) holds for the parent distribution, the question remains as to whether such a result can be used for the normalised sums (2) and (3). Borrowing results from large deviation theory (Nagaev (1965) and Petrov (1975)) we show that one can use the GPD-approximation to normalised sums. As it turns out, the GPD-approximation really differs from the

CLT-approximation for heavy-tailed distributions ($\gamma < 0$). For light-tailed distributions ($\gamma \geq 0$) the GPD-approximation to normalised sums reduces to an exponential distribution ($\gamma = 0$) as if the distribution was normal.

Proposition 1. *For any δ, $1 < \delta < 2$, let $e_n = (n2^{-3j})^{\frac{1}{2}}$, $l_n = n/2^j$, $v_n = e_n + d_n$ where*

$$d_n = \begin{cases} \sqrt{\delta \log n} & \text{if } \gamma \geq 0 \\ l_n^{\frac{\delta-1}{2}} & \text{if } -\frac{1}{2} < \gamma < 0. \end{cases}$$

For $\gamma < 0$, define the sequence σ_n by (6) with $u = d_n$. For $\gamma \geq 0$, define the sequence σ_n by (6) with $u = d_n$ and $\mathcal{E}_1 =^d \mathcal{N}(0,1)$. Under assumptions (C_γ) and $\mathcal{H}_0$, for n large enough, and all $x > 0$,

$$P\Big(|w_k| - v_n \geq x\Big) \leq \begin{cases} \bar{H}(x; 0, \sigma_n) & \text{if } \gamma \geq 0 \\ \bar{H}(x; \gamma, \sigma_n) & \text{if } -\frac{1}{2} < \gamma < 0. \end{cases} \tag{7}$$

Of course (7) is relevant only when $|w_k|$ exceeds the threshold v_n. In the next proposition we give a condition on the resolution level j which ensures that there are at least a finite number of exceedances over the threshold v_n. Our condition depends on the sample size, and $a_n \asymp b_n$ means that there exists some positive constants c_1, c_2 such that, for large n, $c_1 b_n \leq a_n \leq c_2 b_n$.

Proposition 2. *Let $|w_{(k)}|$ be the ordered (absolute value of) wavelet coefficients at resolution level j so $|w_{(1)}| \geq |w_{(2)}| \geq \cdots \geq |w_{(2^j)}|$. Let v_n be the threshold defined in Proposition 1. Under assumptions (C_γ) and $\mathcal{H}_0$, if the resolution level $j = j(n)$ satisfies*

$$\begin{cases} 2^j \asymp \frac{n}{(\log n)^\delta}\ , \ 1 < \delta < 2, \ \text{if } \gamma \geq 0, \\ 2^j \asymp n^{\frac{\gamma+1}{\delta}} \quad , \ 1 < \delta < 1 - \gamma, \ \text{if } -\frac{1}{2} < \gamma < 0, \end{cases} \tag{8}$$

then for any arbitrary constant $m \geq 1$,

$$P\Big(|w_{(m)}| > v_n\Big) \longrightarrow 1, \quad \text{as } n \to \infty. \tag{9}$$

3.3. Fitting the GPD to noisy wavelet transforms

Tajvidi (2004) compares performances of different methods for GPD-parameter estimation. It is known that if the sample size is large (e.g., $n > 500$) Maximum Likelihood Estimation (MLE) is preferred because of its efficiency properties c.f. Hosking and Wallis (1987) and Smith (1985). Below and later, we compute the MLE of the GPD parameters, namely $\hat{\gamma}$, $\hat{\sigma}$, from observations $(|w_k| - u_n)_+ = \max(|w_k| - u_n, 0)$, at the highest resolution level J, $2^{J+1} = n$

(where we have the lowest signal to noise ratio). An illustration of a GPD-fit to noisy wavelet transforms is given in Figure 2.

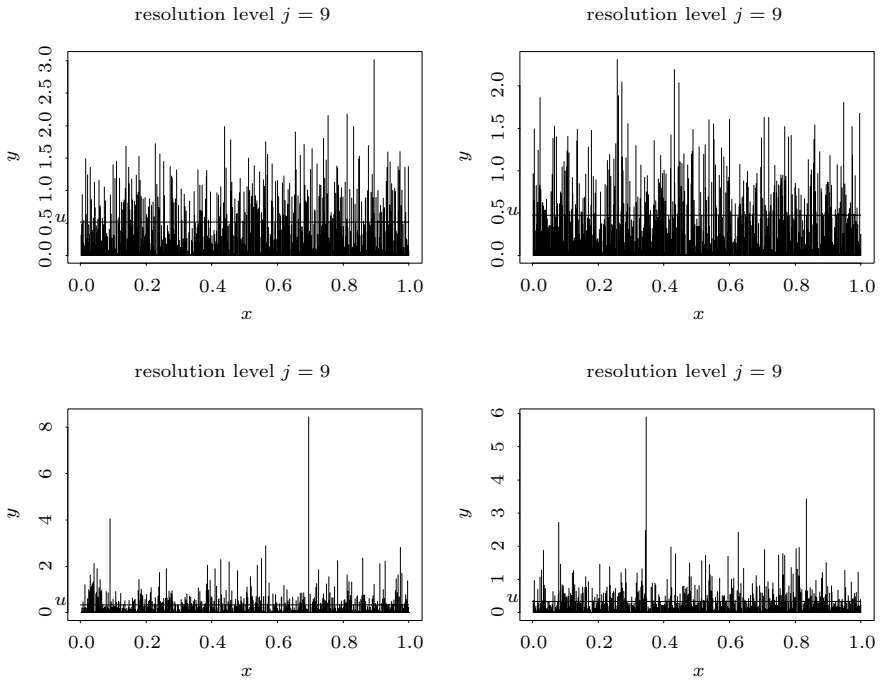

Figure 2. Illustration of the GPD-fit to noisy wavelet transforms. Depicted are the absolute values of wavelet coefficients ($|w_k|$) at the highest accessible resolution level ($J = 9$). Top plots: the wavelet coefficients (w_k) are computed from observations of Figure 1 in light tailed noise. Bottom plots: the wavelet coefficients (w_k) are computed from observations of Figure 1 in heavy tailed noise. GPD-fit: (clockwise from top left) $u_n = 0.51$, $\hat{\gamma} = 0.22$, $\hat{\sigma} = 0.73$; $u_n = 0.47$, $\hat{\gamma} = 0.28$, $\hat{\sigma} = 0.72$; $u_n = 0.34$, $\hat{\gamma} = -0.11$, $\hat{\sigma} = 0.43$; $u_n = 0.33$, $\hat{\gamma} = -0.07$, $\hat{\sigma} = 0.46$.

A common technique to find an appropriate threshold u_n is to use a mean residual life plot, Embrechts, Klüppelberg and Mikosch (1997). Since the expected value of exceedances over the threshold u is a linear function of the threshold, a plot of mean residual life against u should be approximately linear. Here we chose the smallest value of u_n in the region where the plot is approximately linear. Note that for large sample size, one can use a data-driven method for choosing the threshold, see Section 5.3.

4. Peak-Over-Threshold Model for Noisy Wavelet Transforms

Let j be any resolution level, the basic idea is to use the GPD-paradigm

without having to estimate the theoretical threshold v_n. Instead, we consider the statistics

$$T_i = |w_{(i)}| - |w_{(m+1)}| \quad , \ i = 1, \ldots, m. \tag{10}$$

Under $\mathcal{H}_0$, Proposition 2 shows that, with probability tending to one, $T_i < |w_{(i)}| - v_n$. Hence, the GPD-paradigm (Proposition 1) can be used to derive asymptotic confidence bounds for exceedances T_i, c.f. Theorem 1. This is illustrated in the left panels of Figure 3.

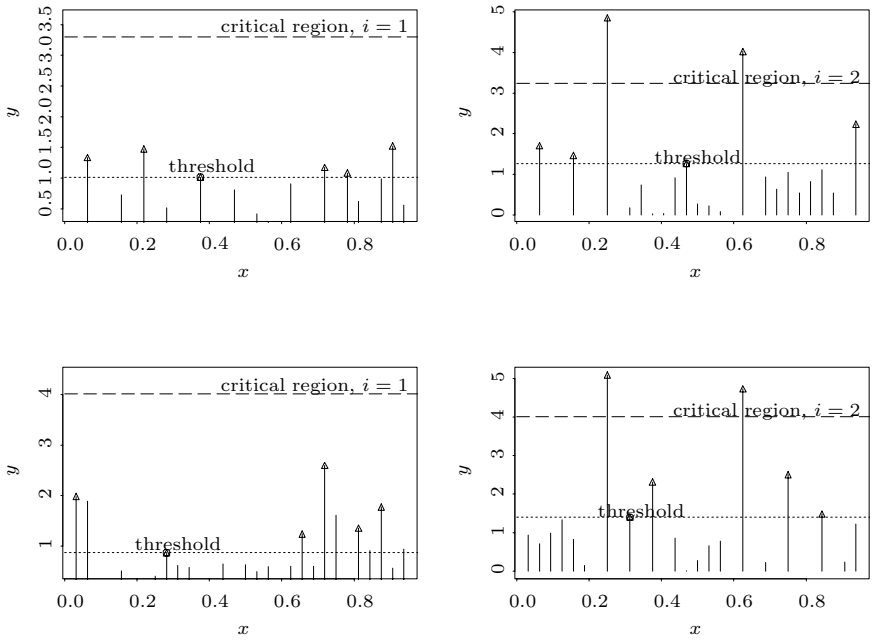

Figure 3. Peak Over Threshold model for noisy wavelet transforms (10), $m = 5$. Depicted are the absolute values of wavelet coefficients ($|w_k|$) at resolution level $j = 5$. Top plots: the wavelet coefficients (w_k) are computed from observations of Figure 1 in light tailed noise. Bottom plots: the wavelet coefficients (w_k) are computed from observations of Figure 1 in heavy tailed noise. The threshold line is drawn at $|w_{(6)}|$ after maximum selection, see Section 5.2. The critical region, at significance level $\beta = 0.05$, is drawn according to (11) using estimates $\hat{\gamma}$ and $\hat{\sigma}$ derived in Figure 2.

4.1. Critical region and confidence bounds

In the theorem below we combine results of propositions 1 and 2 to derive $100(1-\beta)\%$ confidence bounds for the combined m-exceedances (10).

Theorem 1. *Let β be an arbitrary number, $0 < \beta < 1$, and put*

$$c_i = \begin{cases} -\sigma \log\left(\frac{\beta}{m(m-i+1)}\right) & \text{if } \gamma \geq 0, \\ \frac{\sigma}{\gamma}\left(1 - \left(\frac{\beta}{m(m-i+1)}\right)^{\gamma}\right) & \text{if } -\frac{1}{2} < \gamma < 0, \end{cases} \tag{11}$$

$$\mathcal{R}_n(\beta) = \bigcup_{i=1,\ldots,m} \Big\{ T_i > c_i \Big\}. \tag{12}$$

Under assumptions (C_γ) and $\mathcal{H}_0$, if the level j satisfies (8)*, then* $\lim_n P(\mathcal{R}_n(\beta)) \leq \beta$.

Of course when applying these bounds in practice we use estimated values of γ and σ.

4.2. Testing and counting discontinuities

The next theorem states that the previous confidence bounds can be used to detect discontinuities. Indeed condition (8) on the resolution level ensures that the method is powerful. This is illustrated in the right panels of Figure 3.

Theorem 2. *Let $\mathcal{R}_n(\beta)$ be the critical region defined in* (12)*. Under assumptions (C_γ) and $\mathcal{H}_1(m)$, if the resolution level j satisfies* (8)*, then* $\lim_n P(\mathcal{R}_n(\beta)) = 1$.

Remark 1. In the case of light-tailed noise, the GPD-test (12) is quite close to the method initiated by Wang (1995). An examination of the critical value (11) shows that it is not too far from the Universal threshold of Donoho and Johnstone (1994). In contrast, we see that the GPD-test takes into account the tail exponent γ for heavy-tailed distributions.

Remark 2. A close examination of the condition (8) suggests that lower resolution levels should be used to detect discontinuities in the presence of heavy-tailed noise.

Another attractive aspect of wavelet methods is that, with no additional effort, we can estimate the number and the locations of jump points under $\mathcal{H}_1(m)$,

Corollary 1. *Under $\mathcal{H}_1(m)$, denote i_1 and $0 < \theta_1 < \cdots < \theta_{i_1}$ be the number and locations of the jumps of the function f, $1 \leq i_1 \leq m$. Take $\hat{i} = \sup\Big(i : 1 \leq i \leq m,\ T_i > c_i\Big)$ and let $k_1 \leq k_2 \leq \cdots \leq k_{\hat{i}}$ index the ordered DWT, i.e., $|w_{k_1}| \geq |w_{k_2}| \geq \cdots \geq |w_{k_{\hat{i}}}|$. Under the same assumptions as in Theorem 2, $(\hat{i}, k_1/2^j, k_2/2^j, \ldots, k_{\hat{i}}/2^j) \to (i_1, \theta_1, \ldots, \theta_{i_1})$, in probability, as $n \to \infty$.*

An illustration of this result can be seen on the left panels of Figure 3.

4.3. Competing approaches

Wavelet competitors to the GPD-method like Wang (1995) or Odgen and Parzen (1996) derive from arguments in asymptotic theory for Gaussian processes. We refer to these methods as the "Universal" and the "Goodness-Of-Fit" methods, respectively. Let $\hat{s}_1$, $\hat{s}_2$ be the estimated standard deviation of rvs $\mathcal{E}_1$, $(\mathcal{E}_1)^2$ respectively.

- The "Universal"-method (UNI) rejects $\mathcal{H}_0$ when $\max_{k=1,\ldots,2^j} |w_k| > \hat{s}_1\sqrt{2\log n}$.

- The "Goodness-Of-Fit" method (GOF) rejects $\mathcal{H}_0$ when

$$\max_k \frac{1}{\hat{s}_2\sqrt{2^j}}\Big(\sum_{i=1}^{k} w_i^2 - \frac{k}{2^j}\sum_{i=1}^{2^j} w_i^2\Big) > 2.984. \tag{13}$$

A recent non-wavelet approach based on the sums of squared differences of the data has been proposed by Müller (1999). These are formed with various span size L, and are used to estimate the amount of discontinuity in the data $d_1 = \sum_{i=1}^{l} c_i^2$, where c_i's are jump-sizes. We refer to this method as the "U-statistics"-method. Let $\tilde{\mu}_4$, $\tilde{\sigma}^2$ be the estimated 4th moment and variance, respectively.

- The "U-statistics"-method (UST) rejects $\mathcal{H}_0$ when

$$\frac{\sqrt{L}|\hat{d}_1|}{\sqrt{(12/5)(\tilde{\mu}_4 - \tilde{\sigma}^4)}} > 1.961. \tag{14}$$

Critical regions (13) and (14) are given at significance level $\beta = 0.05$.

5. Numerical Properties

5.1. Simulation study

In light of (7), we expect GPD to differ from UNI and GOF for heavy-tailed distributions. This is confirmed by our simulation results where we compared performances of GPD to competing approaches presented in Section 4.3. For each method we calculated Monte-Carlo approximations to the probability of rejecting $\mathcal{H}_0$ under $\mathcal{H}_0$ and under $\mathcal{H}_1$. The results are based on 1000 independent simulations with different noise levels (increasing the standard deviation *sd* from 0.5 to 1) and two different noise types, in all cases $n = 1,024$. We used normal noise to illustrate performances of the methods for light tailed perturbations and Student-t_3 noise to illustrate performances of the methods for heavy-tailed perturbations. We used several test functions f (constant, linear and quadratic)

but we report a detailed summary (Table 1) for only one function f, as a similar pattern was observed for other cases. Our findings are the following.

Table 1. Monte-Carlo approximations to the probability of rejecting $\mathcal{H}_0$.

		Under $\mathcal{H}_0$		Under $\mathcal{H}_1$	
sd	Method	Gaussian	Student-t_3	Gaussian	Student-t_3
0.5	GPD	0.002	0.001	1.000	1.000
0.5	GOF	0.003	0.003	1.000	1.000
0.5	UNI	0.013	**0.164**	1.000	1.000
0.5	UST	0.002	**0.108**	1.000	0.989
0.75	GPD	0.002	0.001	1.000	0.919
0.75	GOF	0.029	0.021	1.000	1.000
0.75	UNI	0.017	**0.185**	1.000	1.000
0.75	UST	0.029	**0.260**	0.991	0.957
1	GPD	0.002	0.001	0.988	0.979
1	GOF	**0.090**	**0.069**	1.000	1.000
1	UNI	0.020	**0.431**	1.000	1.000
1	UST	**0.156**	**0.405**	0.915	0.875

The results are based on 1,000 independent simulations of the model (1) with $f = 0$ under $\mathcal{H}_0$ and $f(x) = \mathbb{1}_{(x>0.25)}$ under $\mathcal{H}_1$.

- The UNI-method works very well under Gaussian noise. It can tolerate low signal-to-noise ratio as seen Table 1, $sd = 1$. On the other hand, it is quite sensitive to heavy-tailed noise perturbations with a large type-I error for Student noise perturbations, even for $sd = 0.5$.
- The GOF-method has good results for both Gaussian and Student noise. It does not tolerate low-signal-to-noise ratio ($sd = 1$) as well as UNI but it has much better performances for Student noise.
- The GPP-method works very well for both Gaussian and Student noise. It can handle low-signal-to-noise ratio ($sd = 1$).
- The UST-method works well for Gaussian noise and constant signal, as seen in Table 1. Also, in Table 1, we see that UST is quite sensitive to heavy-tailed noise perturbations with a large type-I error for Student noise perturbations. Further, we noticed that this method does not work as well as the wavelets method for non-constant signal − we observed a high type-I error when tested with the quadratic function of Figure 1.

Conclusion. The results show that both GOF and GPD can detect discontinuities under heavy-tailed noise as well as under Gaussian noise. We note that

GPD can tolerate a lower signal-to-noise-ratio than GOF. For Gaussian noise, UNI and GPD have better results than GOF.

5.2. Choosing the wavelet family

Existing wavelet methods for change-points detection (UNI or GOF) were originally developed using the Daubechies wavelet families, Daubechies (1992). Such wavelet families are non-translation invariant. However, a simple modification of the wavelet transform can be done to obtain a translation-invariant wavelet family, Raimondo (1998). For change-point estimation and detection we recommend the use of such a translation invariant wavelet family as this enhances the power of detection for discontinuities. Indeed, this is the case here and throughout the simulation study where the translation-invariant Haar-wavelet was used. Note that for counting discontinuities one needs to order the invariant wavelet transforms in the following fashion: if $|w_{(1)}| = |w_k|$ then we disregard $|w_{k\pm1}|$ in the selection of $|w_{(2)}|$ and so on. This avoids counting discontinuities twice, see Raimondo (1998, Section 3).

5.3. Choosing the threshold

The GPD-fit to noisy wavelet transforms, as described in Section 3.3, is based on an appropriate choice of the threshold u_n. Choice of threshold is a matter of trade-off between bias and variance. A too-high threshold results in fewer observations and hence high variance of the estimates, and a too-low threshold makes the estimates severely biased. A good choice of threshold is rather important in small samples, and in practice it is common to use a mean residual plot to find a suitable one. In our simulation study, we used the 10% upper quantiles as the thresholds. This is not expected to affect the results mainly because the sample sizes are fairly large, e.g., 512 or 1,024.

5.4. Application to Dow Jones index

We applied our method to the logarithm of daily closing Dow Jones index starting from October 16th, 1967 until May 1st, 2000 ($2^{13} = 8,192$ data points). We refer to Huang and Litzenberger (1988) for a discussion on the validity of (1) for stock market data. The Dow Jones data, and detected change points, are shown in Figure 4. Interestingly, our analysis not only confirms some well known historical dates of crashes in the Dow Jones index, but also reveals some other important dates of high variation in the Dow Jones index. The first crash is in September 1974 and corresponds to the oil crisis at that time. The next two change points are in August and September of 1982 which might be explained by conflicts in the Middle East. Probably the disturbances in supply of oil from

main Middle East producers have been followed by a boom in the Dow-Jones. October 1987 is the notorious "Black Monday" and the last two crashes in August 1998 and March 2000 are related to the so-called "adjustments" in the market.

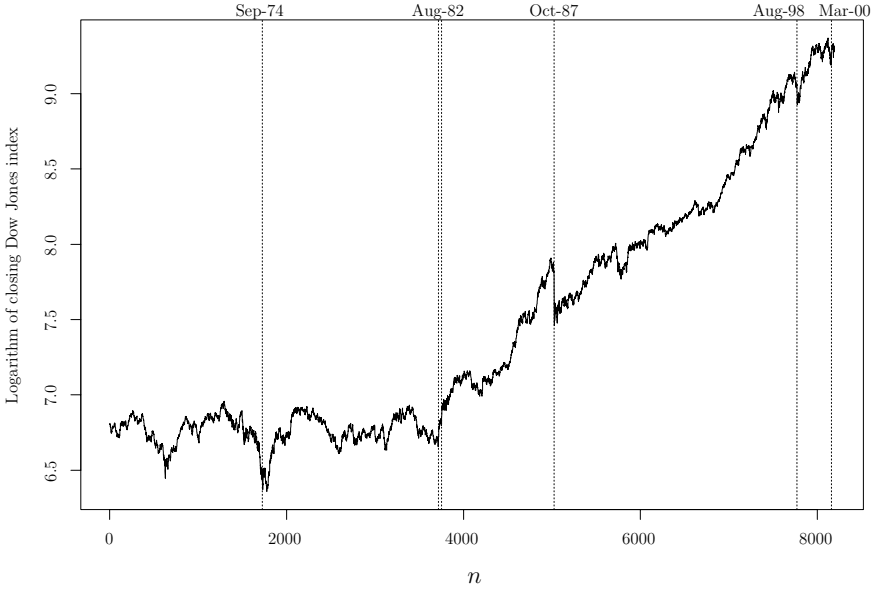

Figure 4. Logarithm of daily Dow Jones index from October 16th, 1967 to May 1st, 2,000 ($2^{13} = 8192$ data points). The dashed lines show the change points detected by GPD-method with $m = 15$, $j = 11$ at significance level $\beta = 0.05$.

6. Proofs

Proof of Proposition 1. Under $\mathcal{H}_0$, $|w_k(f)| \leq e_n$. It follows from the triangle inequality that

$$P\Big(|w_k| - v_n \geq x\Big) \leq P\Big(|w_k(\mathcal{E})| \geq d_n + x\Big) = P\Big(|S_{l_n}/\sqrt{l_n}| \geq d_n + x\Big). \qquad (15)$$

Heavy-tailed case: $d_n = l_n^\nu$ with $0 < \nu = (\delta - 1)/2 < 1/2$. By Petrov (1975, p.251), $P\Big(S_{l_n}/\sqrt{l_n} > l_n^\nu + x\Big) \sim l_n P\Big(\mathcal{E}_1 > \sqrt{l_n}(l_n^\nu + x)\Big)$. Using (C_γ), $P\Big(S_{l_n}/\sqrt{l_n} > l_n^\nu + x\Big) \sim l_n^{1+(1/(2\gamma))}(l_n^\nu + x)^{1/\gamma} L(\sqrt{l_n}(l_n^\nu + x))$. Since $-1/2 < \gamma < 0$ we have $1+(1/(2\gamma)) < 0$. Hence $l_n^{1+(1/(2\gamma))} L(\sqrt{l_n}(l_n^\nu + x)) = o(L(l_n^\nu + x))$ so that, for n large enough ($n \geq n_0$), $P\Big(S_{l_n}/\sqrt{l_n} > l_n^\nu + x\Big) \leq (l_n^\nu + x)^{1/\gamma} L(l_n^\nu + x) \sim P(\mathcal{E}_1 > l_n^\nu + x)$. By symmetry $(\mathcal{E}_1, \ldots, \mathcal{E}_n) =^d (-\mathcal{E}_1, \ldots, -\mathcal{E}_n)$. For $n \geq n_0, P\Big(|S_{l_n}/\sqrt{l_n}| \geq l_n^\nu + x\Big) \leq P(|\mathcal{E}_1| > l_n^\nu + x)$. For all $x > 0$, $P(|\mathcal{E}_1| > l_n^\nu + x) \leq P\Big(|\mathcal{E}_1| > l_n^\nu + x \mid \{|\mathcal{E}_1| > l_n^\nu\}\Big)$.

Combining this with (15) and using (6) with $u_n = l_n^\nu$, shows (7) for $\gamma < 0$.

Light-tailed case: $d_n = \sqrt{\delta \log n}$. By Petrov (1975, p.218), $P\Big(S_{l_n}/\sqrt{l_n} > d_n + x\Big) \sim P\Big(\mathcal{N}(0,1) > d_n + x\Big)$. Combining (15) and (6) with $u_n = \sqrt{\delta \log n}$, as previously, shows (7) for $\gamma \geq 0$.

Proof of Proposition 2. We prove (9) with $m = 1$ (extension to any $m < \infty$ is straightforward). We show that under some condition on the resolution level j,

$$P(|w_{(1)}| \leq v_n) \longrightarrow 0, \quad \text{as} \quad n \to \infty. \tag{16}$$

For the Haar basis we recall that $(w_k)_{k=0,1,\ldots,2^j}$ are independent random variables, hence $P(|w_{(1)}| \leq v_n) = P(\max_{k=0,\ldots,2^j} |w_k| \leq v_n) = \prod_{k=0}^{2^j} P(|w_k| \leq v_n)$. Using $1 - x \leq \exp(-x)$, $0 \leq x \leq 1$, $P(|w_{(1)}| \leq v_n) \leq \exp\Big(-\sum_{k=0}^{2^j} P(|w_k| > v_n)\Big)$. Then (16) will follow if we prove that

$$A_n = \sum_{k=0}^{2^j} P(|w_k| > v_n) \longrightarrow \infty, \quad \text{as} \quad n \to \infty. \tag{17}$$

Let d_n be any positive sequence. Under $\mathcal{H}_0$, $|w_k(f)| \leq (n2^{-3j})^{1/2} = e_n$. Hence with $v_n = e_n + d_n$, by the triangle inequality, $|w_k| \geq |w_k(\mathcal{E})| - e_n$. It follows

$$P(|w_k| > v_n) \geq P(|w_k(\mathcal{E})| > v_n + e_n) = P(|w_k(\mathcal{E})| > d_n + 2e_n). \tag{18}$$

Recall that $l_n = n/2^j$ and let $S_{l_n} = \sum_{i=1}^{l_n} \mathcal{E}_i$. From (4) we see that

$$P(|w_k(\mathcal{E})| > d_n + 2e_n) = P(|S_{l_n}/\sqrt{l_n}| > d_n + 2e_n) \geq P(S_{l_n}/\sqrt{l_n} > d_n + 2e_n). \tag{19}$$

Heavy-tailed case: $d_n = l_n^\nu$ where $0 < \nu < 1/2$. By definition of sequences d_n, e_n we have $d_n + 2e_n = d_n(1 + o(1))$. Thus, there exists a constant $c > 0$ such that, for $n \geq n_0$, $d_n + 2e_n \leq c\, d_n$. From (18), (19) we have that, for $n \geq n_0$,

$$P(|w_k| > v_n) \geq P\Big(S_{l_n}/\sqrt{l_n} > c\, d_n\Big) = P\Big(S_{l_n} > c\, l_n^{\frac{1}{2}+\nu}\Big). \tag{20}$$

If $-1/2 < \gamma < 0$, Nagaev's condition is satisfied, hence by Petrov (1975, p.251), $P\Big(S_{l_n} > c\, l_n^{\frac{1}{2}+\nu}\Big) \sim l_n P\Big(\mathcal{E}_1 > c\, l_n^{\frac{1}{2}+\nu}\Big)$. Using (C_γ), $P\Big(S_{l_n} > c\, l_n^{\frac{1}{2}+\nu}\Big) \sim l_n^{1+\frac{1}{2\gamma}+\frac{\nu}{\gamma}}\, L_2(n)$ for some slowly varying function L_2. Summing terms in (20), for $n \geq n_0$,

$$\sum_{k=0}^{2^j} P(|w_k| > v_n) \geq 2^j P\Big(S_{l_n} > c\, l_n^{\frac{1}{2}+\nu}\Big) \sim 2^j\; n^{1+\frac{1}{2\gamma}+\frac{\nu}{\gamma}}\; 2^{-j(1+\frac{1}{2\gamma}+\frac{\nu}{\gamma})}\; L_2(n).$$

Letting $\delta = 1+2\nu$, it follows that $A_n \geq (n^{(2\gamma+\delta)}2^{-j\delta})^{\frac{1}{2\gamma}}\ L_2(n)$ which, for $-1/2 < \gamma < 0$, proves (17) for all resolution levels j satisfying (8).

Light-tailed case: $d_n = \sqrt{\delta \log n},\ 1 < \delta < 2$. From (8), $d_n + 2e_n = d_n + o(1)$. Using (18) and (19), $P(|w_k| > v_n) \geq P(S_{l_n}/\sqrt{l_n} > d_n + o(1)) \sim P(S_{l_n}/\sqrt{l_n} > \sqrt{\delta \log n})$. For $\gamma \geq 0$, Cramer's condition is satisfied, c.f. Petrov (1975, p.54). Since $d_n = o(\sqrt{l_n})$, we use Petrov (1975, p.218) to let $P(S_{l_n}/\sqrt{l_n} > \sqrt{\delta \log n}) \sim P(\mathcal{N}(0,1) > \sqrt{\delta \log n})$. It follows that $P(S_{l_n}/\sqrt{l_n} > \sqrt{\delta \log n}) \sim n^{-\delta/2}\ (2\pi\delta \log n)^{-1/2}$ and

$$\sum_{k=0}^{2^j} P(|w_k| > v_n) \geq \sum_{k=0}^{2^j} P\Big(S_{l_n}/\sqrt{l_n} > \sqrt{\delta \log n}\Big) \sim 2^j\ n^{-\frac{\delta}{2}} \frac{1}{\sqrt{2\pi\delta \log n}}.$$

Thus, for $\gamma \geq 0$, (17) holds for all resolution levels j satisfying (8).

Proof of Theorem 1. Let $\mathcal{A}_n$ the event that $|w_{(m+1)}| > v_n$; we have $P(T_i > x|\mathcal{A}_n) \leq P(|w_{(i)}| - v_n > x|\mathcal{A}_n)$. With a slight abuse of notation, writing $|w_i|, |w_{i+1}|, \ldots, |w_m|$ for the unordered set of coefficients whose ordered sequence is $|w_{(i)}| \geq |w_{(i+1)}| \geq \cdots \geq |w_{(m)}|$, we obtain

$$P(|w_{(i)}| - v_n > x|\mathcal{A}_n) = P(\max_{k=i,\ldots,m} |w_k| - v_n > x|\mathcal{A}_n) \leq \sum_{k=i}^{m} P(|w_k| - v_n > x|\mathcal{A}_n).$$

By propositions 1 and 2, $P(|w_k| - v_n > x|\mathcal{A}_n) \sim P(|w_k| - v_n > x) \leq \bar{H}(x, \gamma, \sigma_n)$. Writing $\sigma = \sigma_n$ for $n \geq n_0$,

$$P(T_i > x|\mathcal{A}_n) \leq (m - i + 1)\bar{H}(x; \gamma, \sigma), \quad i = 1, \ldots, m. \tag{21}$$

Since there are at most m exceedances, we choose $x = c_i = c(i, m, \beta, \gamma, \sigma)$ such that $\bar{H}(c_i; \gamma, \sigma) = \beta/(m(m - i + 1))$, from which (11) is derived. This, together with (21) gives for $n \geq n_0$,

$$P(\mathcal{R}_n(\beta)|\mathcal{A}_n) \leq \sum_{i=1}^{m} P(T_i > x|\mathcal{A}_n) \leq \sum_{i=1}^{m} (m - i + 1)\bar{H}(c_i; \gamma, \sigma) \leq \beta.$$

Taking the limit on the left hand-side and applying Proposition 2 proves the Theorem.

Proof of Theorem 2. To simplify the exposition we suppose that $m = 1$ and $\mathrm{Var}\,(\mathcal{E}_1) = 1$ (extension to other cases is straightforward). The theorem will follow if we prove that

$$T_1 \longrightarrow^P \infty, \quad \text{as } n \to \infty. \tag{22}$$

By (2), along with the triangle inequaly $|w_k| \leq |w_k(f)| + |w_k(\mathcal{E})|$ and $|w_k| \geq |w_k(f)| - |w_k(\mathcal{E})|$. Let $\mathcal{B}_n = \{\max_{k=0,\ldots,2^j} |w_k(\mathcal{E})| \leq x_n\}$, where

$$x_n = \begin{cases} \sqrt{2 \log n}, & \text{if } \gamma \geq 0, \\ n^{-\frac{\gamma}{2}}, & \text{if } -\frac{1}{2} < \gamma < 0. \end{cases} \tag{23}$$

Under $\mathcal{H}_1(1)$, and for the Haar basis, there exists a unique index k_1 such that $|w_{k_1}(f)| \geq (l_n)^{\frac{1}{2}}$ whereas $|w_k(f)| \leq (n2^{-3j})^{\frac{1}{2}} = e_n$ for any $k \neq k_1$, see Raimondo (1998). Working conditionally on $\mathcal{B}_n$ and using definitions (8), (23) of x_n and j_n in terms of n, it is not hard to check that $e_n = o(x_n)$, $x_n = o(\sqrt{l_n})$. It follows that $|w_k| \leq e_n + x_n = O(x_n)$, $k \neq k_1$ and $|w_{k_1}| \geq \sqrt{l_n} - x_n = \sqrt{l_n}(1+o(1))$. This shows that for $n \geq n_0$, $|w_{(1)}| = \max_k |w_k| = |w_{k_1}| \geq \sqrt{l_n}(1+o(1))$. Since $|w_{(2)}| = |w_k|$ for some $k \neq k_1$, we have $T_1 = |w_{(1)}| - |w_{(2)}| \geq \sqrt{l_n} - O(x_n) = \sqrt{l_n}(1+o(1))$. This proves (22) if we show that $P(\mathcal{B}_n) \longrightarrow 1$, as $n \to \infty$. For light-tailed distributions, this follows from Cramer Large Deviation Theory, see Petrov (1975, p.218). For heavy-tailed distributions we note that

$$P(\mathcal{B}_n^c) = P(\max_{k=0,\ldots,2^j} |w_k(\mathcal{E})| > x_n) \leq \sum_{k=0}^{2^j} P(|w_k(\mathcal{E})| > x_n) = 2\sum_{k=0}^{2^j} P(w_k(\mathcal{E}) > x_n). \tag{24}$$

Recalling that $P(w_k(\mathcal{E}) > x_n) = P(S_{l_n}/\sqrt{l_n} > x_n) = P(S_{l_n} > \sqrt{l_n}\, x_n)$, and applying Petrov (1975, p.251) with $x = \sqrt{l_n}\, x_n = \sqrt{l_n}\, n^{-\frac{\gamma}{2}}$,

$$P(S_{l_n} > \sqrt{l_n} n^{-\frac{\gamma}{2}}) \sim l_n P(\mathcal{E}_1 > \sqrt{l_n} n^{-\frac{\gamma}{2}}) \sim l_n^{1+\frac{1}{2\gamma}} n^{-\frac{1}{2}}\, L_1(n) \tag{25}$$

for some slowly varying function $L_1(n)$. Combining (24) and (25), $P(\mathcal{B}_n^c) = O\Big(2^j\ l_n^{1+\frac{1}{2\gamma}}\ n^{-\frac{1}{2}}\ L_1(n)\Big) = O\Big((n^{\gamma+1}2^{-j})^{\frac{1}{2\gamma}}\ L_1(n)\Big)$ and this gives $P(B_n) \longrightarrow 1$, under condition (8).

Acknowledgements

We are grateful to the Editor, an associate editor and a referee for their helpful comments which led to the revised version of this paper. Part of this paper was written while Nader Tajvidi visited the University of Sydney supported by the Swedish Foundation for International Cooperation in Research and by the University of Sydney.

References

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Vol. 61. SIAM, Philadelphia.

Davison, A. and Smith, R. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52**, 393-442.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Embrechts, P., Klüppelberg, C. and Mikosch, H. (1997). *Modelling of Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, Heidelberg.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*. Vol. 129 of *Lecture Notes in Statistics*, Springer.

Hosking, J. and Wallis, J. (1987). Parameter and quantile estimation for the generalised Pareto distribution. *Technometrics* **29**, 339-349.

Huang, C. and Litzenberger, R. (1988). *Foundations for Financial Economics*. North-Holland, New York.

Huh, J. and Carriere, K. (2002). Estimation of regression functions with a discontinuity in a derivative with local polynomial fits. *Statist. Probab. Lett.* **56**, 329-343.

Lio, P. and Vannucci, M. (2000). Wavelet change-point prediction of transmembrane proteins. *Bioinformatics* **16**, 1-7.

Mallat, S. (1989). A theory for multi-resolution signal decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674-693.

Müller, H. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27**, 299-337.

Nagaev, S. (1965). Some limit theorems for large deviations. *Theory Probab. Appl.* **10**, 215-235.

Nason, G. and Silverman, B. (1994). The discrete wavelet transform in S. *J. Comput. Graph. Statist.* **3**, 163-191.

Odgen, T. and Parzen, O. (1996). Change-point approach to data analytic thresholding. *Statist. Comput.* **6**, 93-99.

Petrov, V. (1975). *Sums of Independent Random Variables*. Springer-Verlag, New York.

Pickands, J. I. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Raimondo, M. (1998). Minimax estimation of sharp change points. *Ann. Statist.* **26**, 1379-1397.

Raimondo, M. (2002). Wavelet shrinkage via peaks over threshold. *InterStat* (May), 1-19. http://interstat.stat.vt.edu/InterStat/intro.html-ssi.

Smith, R. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67-90.

Tajvidi, N. (2004). Confidence intervals and accuracy estimation for heavytailed generalised Pareto distribution. Extremes, to appear.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82**, 385-397.

School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia.

E-mail: marcr@maths.usyd.edu.au

Department of Mathematical Statistics, Lund University, Box 118, SE 22100 Lund, Sweden.

E-mail: Nader.Tajvidi@maths.lth.se

(Received April 2001; accepted August 2003)