On the Automatic Selection of the Onset of Scaling

Darryl Veitch, Patrice Abry, Murad S. Taqqu

October 14, 2002

Abstract

A method is developed for the automatic detection of the onset of scaling for long-range dependent time series and other asymptotically scale invariant processes. Based on wavelet techniques, it provides the lower cutoff scale for the regression that yields the scaling exponent. The method detects the onset of scaling through the dramatic improvement of a goodness of fit statistic taken as a function of this lower cutoff scale. It relies on qualitative features of the goodness of fit statistic and on features of the wavelet analysis. The method is easy to implement, appropriate for large data sets and highly robust. It is tested against 34 time series models and found to perform very well. Examples involving telecommunications data are presented.

1 Introduction

The practical determination of cutoff scales for processes which are scale invariant in an asymptotic sense is a recurrent problem. In this paper, we deal with Long-Range Dependent (LRD) second-order stationary time series, and consider the problem of selecting the lower cutoff time-scale where the long-range dependence can be deemed to ‘begin’.

We adopt a simple and common definition of LRD, namely that \( f_X(\nu), \nu \in (-1/2, 1/2], \) the spectral density of the discrete time series \( X(t) \), exhibits a power-law ‘infra-red’ divergence:

\[
 f_X(\nu) \sim c_\nu |\nu|^{-\alpha}, \quad |\nu| \to 0, \tag{1}
\]

where \( \alpha \in (0, 1) \) is the scaling exponent and \( c_\nu \) a size parameter. We note immediately that this definition specifies only the asymptotic behaviour of the spectral density. It offers no insight into the frequencies at which the power-law divergence becomes apparent. The same is true of other common definitions, such as replacing \( c_\nu \) with a slowly varying function, or equating LRD to divergence of the spectrum at the origin. Knowledge of the onset frequency is important in several contexts, for example when estimating the scaling exponent using semi-parametric techniques, it is essential to restrict to the frequencies at which the power-law is present to avoid estimation bias.

We see that the question of the automatic selection of a cutoff frequency or time-scale must be preceded by the more fundamental one, namely the definition of such a scale. There is no universally valid definition. Rather, different choices will be appropriate for different purposes. Here we adopt one based on estimation quality, an important practical issue. More specifically, the lower cutoff or onset scale is defined to be the one which gives minimum mean square error (MSE) to a particular semi-parametric wavelet-based estimator of \( \alpha \). This definition was introduced and evaluated in [1]. In the present paper we provide...
a heuristic which allows this ‘optimal’ or ‘true’ onset scale to be robustly estimated from data without requiring visual inspection as in [10] or [9]. Existing theoretical results regarding restricting the estimation to the scaling region, for example [8], [6], [7], are asymptotic in nature or describe some minimax property, and are not immediately applicable. The goal of this article is to develop a practical method which can be implemented.

Although defined via the estimation quality of a particular estimator, the definition provides a selection of onset scale which agrees well with visual inspection of the departure of the process from the asymptotic power-law behaviour, and should prove practically useful in other contexts. Furthermore, although developed in the context of LRD, the estimation method can readily be applied to other scaling processes, for example $1/f^a$ noise, for the estimation of lower cutoff scale, and to a lesser extent for upper cutoff scales also.

The paper is organised as follows. In section 2, background material is presented on time series, wavelets, and estimation using wavelets. In section 3, we define the lower scale and develop a method for finding it. We examine the performance and calibration of the method in section 4, and compare it against a more parametric alternative. Finally we discuss the advantages and drawbacks of the approach in section 5, and give examples using long-range dependent telecommunication data.

## 2 Preliminaries

### 2.1 The FARIMA and fGn processes

The effectiveness of the method will be tested and illustrated using Fractional Autoregressive Integrated Moving Average (FARIMA) time series. For completeness we also include examples of the fractional Gaussian noise (fGn).

The FARIMA family offers a convenient framework to define long-range dependent series in discrete time with flexible and controllable short range dependence. Another key advantage is that their spectra are known in closed form. We now offer a brief introduction to these processes. Further details can be found in [2].

A Gaussian FARIMA$(0,d,0)$ process is defined by $X_i = \Delta^{-d}\epsilon_i$, $i \geq 1$, where the $\epsilon_i$ are independent, identically distributed Gaussian random variables called innovations with zero mean and variance $\sigma^2$, and where $\Delta$ is the differencing operator $\Delta \epsilon_i = \epsilon_i - \epsilon_{i-1}$. For fractional $d$ we interpret $\Delta^{-d}$ via a formal power series expansion: $\Delta^{-d} = \sum_{i=0}^{\infty} b_i(-d) B^i$, where $B$ is the backward operator, $B \epsilon_i = \epsilon_{i-1}$, and $b_i(-d) = \Gamma(i + d)/\Gamma(d) \Gamma(i + 1)$, $i = 1, 2, \ldots$, $\Gamma$ being the gamma function. For $d \in (0,1/2)$ this process is LRD, where $\alpha = 2d$. More generally, a FARIMA$(1,d,1)$ process adds a single autoregressive and moving average term to the fractional differencing, namely:

$$X_i - \phi X_{i-1} = \Delta^{-d}\epsilon_i - \theta \Delta^{-d}\epsilon_{i-1},$$

where $\phi$ and $\theta$ are the autoregressive and moving average coefficient respectively. The Fourier spectrum for this process is

$$f_X (\nu) = \sigma^2 \left| 1 - e^{-2\pi i \nu} \right|^{2d} \left| 1 - \theta e^{-2\pi i \nu} \right|^2 \left| 1 - \phi e^{-2\pi i \nu} \right|^2, \quad -1/2 < \nu < 1/2$$

(we use normalized frequencies and define the Fourier transform of $\psi$ as $\int_{-\infty}^{\infty} e^{-i2\pi \nu t} \psi(t) dt$). The shape parameters $\theta$ and $\phi$ can be used to modulate the degree and nature of the spectrum at high frequency or equivalently, at coarse scales. Higher order autoregressive and moving average components are possible but will not be used here.

The fGn is a one parameter family of stationary discrete time processes with correlation function

$$\gamma_X(0) = 1,$$

$$\gamma_X(k) = \frac{1}{2} \left( (k-1)^{2H} - 2k^{2H} + (k+1)^{2H} \right), \quad k = 1, 2, \ldots,$$
where $H \in (0,1)$ is the Hurst parameter. Members of the fGn family are exactly second-order self-similar, in the sense that for all positive integers $m$, the finite-dimensional distributions of $\sum_{i=1}^{nm} X_i$ are the same as those of $m^H \sum_{i=1}^{n} X_i$. Members of the fGn family are also LRD if $H > 1/2$. The spectrum can not be expressed in closed form but must be calculated through the defining sum for the discrete Fourier transform of $\gamma_X(k)$.

2.2 Wavelet Analysis

As mentioned in the introduction, the onset selection method is built around a wavelet estimation framework. This is justified in general terms by the natural matching of wavelets to scaling phenomena, and by the ensuing low computational cost allowing the analysis of very long series. It is also indicated for a technical reason, the natural weighting of smaller scales, as explained in the next section. We restrict ourselves to a description of the main features of the wavelet analysis as employed for our purposes here. Further details on wavelets can be found for example in [3, 4].

The coefficients $d_k(j,k)$ of the discrete wavelet transform (DWT) result from the comparison, by means of inner products, of the process to be analysed $x$ and a family of functions $\{\psi_{j,k}\}$, called the wavelet basis: $d_k(j,k) = \langle X, \psi_{j,k} \rangle = \int_{\mathbb{R}} X(t) \psi_{j,k}(t) dt$. The wavelet basis $\{\psi_{j,k}\}$ consists of shifted and dilated templates of a single reference pattern $\psi_0$, usually called the mother-wavelet. The mother-wavelet has a time support and frequency support which are strongly concentrated: it therefore acts as an elementary atom of information. From it the time-shift operator and the dilation (or change of scale) operator together generate the full, two parameter set of basis functions:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \psi_0 \left( \frac{t - 2^j k}{2^j} \right),$$

centred on a sparse set of points in the time-scale plane known as the dyadic grid, that is the points $\{(\text{scale } = 2^j, \text{ time } = 2^j k), j, k \in \mathbb{N}\}$. The mother-wavelet is moreover characterised by an integer $N$, called the number of vanishing moments, defined as the largest $N$ such that $\int t^k \psi_0(t) dt \equiv 0$ for $k = 0, 1, 2 \cdots N - 1$.

It can be shown [5] that the second moments of the coefficients $d_k(j,k)$ satisfy

$$\mathbb{E} d_k(j,k)^2 = \int_{\mathbb{R}} \Gamma_X(\nu) 2^j |\Psi_0(2^j \nu)|^2 d\nu,$$

where $\Psi_0(\nu) = \int \psi_0(t) e^{-2\pi i \nu t} dt$ is the Fourier transform of $\psi_0$. These second order quantities take a particularly simple form in the case of Long-Range Dependence (LRD), where by definition the spectral density follows a power-law near the origin. Because of the inherent scaling properties in the wavelet basis which are naturally matched to the scaling properties of LRD processes, one obtains ((11))

$$\mathbb{E} d_k(j,k)^2 \sim 2^{2j} c_j C(\alpha, \psi_0), \quad j \to +\infty,$$

where $C(\alpha, \psi_0)$ is an integral independent of scale $j$.

In practice the DWT coefficients are not calculated from their defining integrals, but through an efficient recursive algorithm. This begins with the initial approximation sequence defined by

$$a(k) = \int_{\mathbb{R}} X(t) \phi_0(t - k) dt,$$

where $\phi_0$ is the scaling function closely associated to the mother wavelet. The wavelet coefficients can be calculated directly from $a(k)$ by a recursive procedure involving simple discrete filtering, with computational complexity of only $O(n)$, where $n$ is the length of $a(k)$. In the case of wavelets with finite temporal support, such as the orthogonal Daubechies wavelets [3] we use in this paper, this filtering can be performed without truncation error.
2.2.1 Estimation

From (5), one can think of extracting the scaling exponent $\alpha$ from a linear fit in a $\log_2(\mathcal{E}d_\chi (j, k)^2$ versus $\log_2(2^j) = j$ plot or Logscale Diagram (LD), namely $\log_2(\mathcal{E}d_\chi (j, k)^2 = \alpha j + c$. Following this idea, a semi-parametric wavelet based estimator reads

\begin{align*}
y_j &= \log_2\left(1/n_j \sum_k d_\chi(j,k)^2 \right) - g_j, \\
\hat{\alpha} &= \sum_j w_j y_j, \\
\hat{\sigma} &= \sum_j v_j y_j,
\end{align*}

where $n_j$ is the number of coefficients available at octave $j$, and the latter sum is over $j \in [j_1, j_2]$, the range of octaves over which the scaling phenomenon is observed and the linear regression performed. The $g_j$ are deterministic quantities that account for the fact $\log_2 \mathcal{E}d_\chi (j, k)^2 \neq \mathcal{E} \log_2 d_\chi(j, k)^2$. The weights $w_j$ follow the standard formulae for weighted linear regression using the variances $\sigma_j^2$ of the $y_j$, namely

\begin{equation}
w_j = \frac{S_j - S_1}{(SS_2 - S_1^2)\sigma_j^2}, \quad v_j = \frac{S_2 - jS_1}{(SS_2 - S_1^2)\sigma_j^2}, \quad S = \sum_{j=j_1}^{j_2} 1/\sigma_j^2, \quad S_1 = \sum_{j=j_1}^{j_2} j/\sigma_j^2, \quad S_2 = \sum_{j=j_1}^{j_2} j^2/\sigma_j^2.
\end{equation}

Because of the ability of wavelets to quasi-decorrelate scaling processes, such an estimate has excellent statistical properties. In addition it can be calculated with a computational time complexity of only $O(n)$, as well as enjoying a variety of robustness advantages. See [11] for full details.

It is important to understand the nature of the variances $\sigma_j^2$ of the (bias corrected) log-estimates $y_j$ at octave $j$. In general these cannot be calculated exactly, but the following expression, obtained by idealizing the weak correlations between the wavelet coefficients to complete decorrelation, is an excellent approximation in the case where the $d_\chi(j, k)$ are Gaussian:

\begin{equation}
\sigma_j^2 = \zeta(2, n_j/2)/\ln^2 2 \sim \frac{2}{n_j \ln 2^2},
\end{equation}

where $\zeta(2, z) = \sum_{n=0}^{\infty} 1/(z + n)^2$ is a generalized Riemann Zeta function. The asymptotic expression for large $n_j$ holds true to a good approximation for data with reasonable departures from Gaussianity (see [11] for further details). For completeness we also quote from [11] the expression for $g_j$

\begin{equation}
g_j = \psi(n_j/2)/\ln 2 - \log_2(n_j/2)
\end{equation}

where $\psi(z)$ is the Psi function.

The left plot in figure 1 illustrates the estimates $y_j$ for five different realisations (dashed lines) of a particular FARIMA($1, d, 1$) process. The vertical lines at each $j$ denote confidence intervals proportional to the $\sigma_j$, which increase geometrically, reflecting the rapid decrease in available data with increasing scale described by $n_j \approx n 2^{-j}$, where $n$ is the length of the sample paths of $X(t)$.

2.2.2 Analysis of Discrete Data

An issue which is often overlooked is the fact that wavelet analysis, including the Discrete Wavelet Transform, is defined for continuous time processes, and not for discrete time series such as FARIMA processes. This is significant as time series models in general, including some of the most popular LRD models, and much real data, fall into this category. Recall from equation (6) that to initialize the DWT the initial approximation sequence $\mathcal{a}(k)$ must be obtained via a sequence of integrals in continuous time, a procedure
that has no meaning for discrete data. Typically what is done if one wants to study a discrete process \( X(k) \) is to set \( a(k) = X(k) \), an ad hoc procedure which makes it unclear what is being studied, and introduces errors. In earlier work [12], we showed how to use the DWT in a well defined way for the study of second order properties of discrete wide sense stationary processes, such as those we consider here. (In fact [12] arose from the need to resolve the discrete data issue as a necessary precursor to the present problem of the detection of the onset scale.)

The approach is conceptually very simple and can be expressed as follows. To a given discrete series \( \{X(k), k \in \mathbb{Z}\} \) we associate a closely related continuous-time process \( \{\tilde{X}(t), t \in \mathbb{R}\} \). In fact the process \( \tilde{X}(t) \) is chosen such that the spectral densities of \( X \) and \( \tilde{X} \) coincide on the principal interval \( \nu \in (-1/2, 1/2] \). The series \( \tilde{X}(t) \), being defined in continuous time, can be studied using the DWT as usual. Since the spectra are equal in \( (-1/2, 1/2] \), any conclusion on the spectral density of \( \tilde{X} \) in this range, for example estimates of \( \alpha \) in the case of LRD, holds automatically for the spectral density of \( X \). The only difficulty is the initialization phase which begins the DWT analysis of \( \tilde{X}(t) \). It can be shown that the two steps of first converting from \( X \) to \( \tilde{X} \), and then initializing for \( \tilde{X} \), can be accomplished in a single linear discrete filtering (convolution) operation:

\[
a_{\tilde{X}}(k) = (X * I)(k),
\]

where \( I \) is a discrete filter given by

\[
I(m) = \int_{-\infty}^{\infty} \frac{\sin(\pi(t + m))}{\pi(t + m)} dt,
\]

which depends only on the wavelet used (\( I \) has infinite support but in practice can be taken to be quite small, say 100 long or less). The prefiltering is easy to perform, and corrects errors which would otherwise have been very significant on the first two octaves, but negligible at coarse scales. Throughout this paper discrete series are analysed, and the prefiltering has been performed.

3 A Robust Selector of Lower Scale

In this section we present the intuition underlying the selection method and then define it. We begin by reviewing the definition of the onset scale presented in [1], supported by the reference example of figure 1. The plots in the figure describe the different aspects of the method for a FARIMA \((1, d, 1)\) process with \( d = 0.2 \) (\( \alpha = 2d = 0.4 \), or \( H = d + 1/2 = 0.7 \) in the labelling convention of fGn), \( \phi = 0.3 \) and \( \theta = 0.7 \).

3.1 Defining the Lower Scale

For a given wavelet, and a given time series, the expectation of equation (4) can in principle be calculated. We do so using numerical integration with Daubechies 3 wavelets \((N = 3)\). The logarithms

\[
s_j \equiv \log_2 E_{\tilde{X}}(j, \cdot)
\]

of these expectations are shown as the circles in the leftmost plot of figure 1. They constitute essentially a wavelet based ‘spectrum’ which we term the Exact Logscale Diagram (exact LD), which serves as a starting point for the wavelet analysis of the model.

Consider relation (5). The logarithm of its left-hand side is \( s_j \), whereas the logarithm of its right-hand side is \( a_j = \alpha j + \log_2 (C(\alpha, \psi_0)) \), corresponding to the asymptotic LRD behaviour. Over the first four octaves \( j \), the exact logscale diagram \( s_j \) deviates strongly from this straight line, shown extending down from the coarsest scales in the figure. This deviation is reflected in the logscale diagrams of five independent realisations of the process overlayed as dashed lines. The exact logscale diagram is essentially the ensemble
average of such estimates. From this clear visual departure we must extract a rigorous definition of an onset scale.

One way to approach this is to examine the effect of the short range dependence on \( \hat{\alpha} \) estimates. To investigate this in an exact setting we adopt the following approach, which formalises the performance of semi-parametric estimation in the face of deviations from exact scaling of the process. Using the exact LD values \( s_j \) to formally replace the estimates \( y_j \), the weighted least squares fit described in section 2.2.1 is applied, resulting in a deterministic fit \( \hat{\alpha} \) over \( [j_1, j_2] \). The quantity \( \hat{\alpha} - \alpha \) can be interpreted as the bias in an idealised sense of the wavelet semi-parametric estimator, and \( \sum_j \omega_j^2 \sigma_j^2 \) as its variance, and both can be computed exactly. They are given as a function of \( j_1 \) (with \( j_2 \) as large as possible) in the middle plot of figure 1. The standard deviations increase monotonically with \( j_1 \), as they must, corresponding to less data being included in the fit. The bias, which is a weighted measure of deviation from the asymptotic behaviour, in general decreases as the smaller scales are progressively excluded. These complementary monotonicities suggest that a mean square error (MSE) will have a distinct minimum value. For a random variable \( \hat{\alpha} \) estimated by \( \hat{\alpha} \), the MSE is defined as \( \text{MSE} = \text{Var}(\hat{\alpha}) \), which in the present context becomes

\[
\text{MSE} = (\hat{\alpha} - \alpha)^2 + \sum_j \omega_j^2 \sigma_j^2.
\]

Inspired by this idea, we define the onset scale \( j_1 = j_1^{\text{MSE}} \) as the one which minimizes the above MSE. With \( n = 10000 \) and \( N = 3 \), this yields \( j_1^{\text{MSE}} = 5 \) for the FARIMA\((1, d, 1)\) model of figure 1 (if multiple \( j_1^{\text{MSE}} \) are found, the smallest is retained thus selecting the largest possible measurement range). This value is in good ‘visual’ agreement with the onset of the asymptotic regime (see figure 1, left plot), and we have found this to be true in all models we have examined.

Although it does not affect our definition, it is important to understand that the bias and variance of the estimator \( \hat{\alpha} \) we actually apply to data, that described in section 2.2.1, differs subtly from the ‘exact’ one \( \alpha \) considered here. This is because the \( g_j \) bias correction terms are not exact, and similarly \( \text{Var}(\hat{\alpha}) \) is not

---

**Figure 1:** Exact LD, \( j_1^{\text{MSE}} \), and Goodness of fit for FARIMA\((1, d, 1)\), \( d = 0.2, \phi = 0.3, \theta = 0.7 \) (model 22 from table 1). Left: the exact LD for \( N = 3 \), and 5 sample paths based on \( n = 10000 \) points. Middle: the quality of ideal estimates of \( \alpha \) based on the exact LD, judged by bias, standard deviation, and root MSE as a function of \( j_1 \). The square marks \( j_1^{\text{MSE}} = 5 \). Right: The goodness of fit measure increases dramatically then levels off. The darker line colour denotes the non-decreasing zone (log-values truncated at \(-10\)). The diamond marks \( j_1^{\text{MSE}} \). The horizontal lines correspond to \( Q = 0.01, 0.05, \) and \( 0.1 \).
exactly equal to $\sum_j \omega_j^2 \sigma_j^2$. As shown in [11] however, for Gaussian data they are very accurate, and so we may effectively regard the estimator as the sample equivalent of the exact LD based ‘estimator’ $\hat{\alpha}$ presented above.

Intuitively it is clear that, at fixed $n$, as we move to lower frequencies with increasing $j_1$, the value of $\hat{\alpha}$ is less and less polluted by departures at small scales, and so the estimation bias decreases, at the expense of an increase in variance due to the effective loss of data. An MSE criterion allows the tradeoff between these two effects to be economically expressed. If we now consider increasing the sample size $n$, the balance shifts toward larger $j_1$ values as the $\sigma_j^2$ all decrease with $n$ (SEE 11), whereas the bias $s_j - a_j$ at each $j$, which depends on the exact LD only, remains constant. Thus $j_1^{\text{MSE}}(n)$, far from being an absolute constant for a process indicating where the LRD ‘truly’ begins, is in fact a non-decreasing function of $n$. Note that $j_2 \approx \log_2(n)$ is also a function of $n$. However, as it increases slowly with $n$, and as the weight in the estimation is very small at the largest scale, the dependence of $\hat{\alpha}$ and hence of $\sigma_j^2$ on $n$ is weak in most situations.

We could also describe this choice of cutoff in another way, as being the crossover point between the SRD and LRD of the process. It is interesting to note that even for a fractional Gaussian noise, which is a second order self-similar process, the scale invariance in this sense does not extend to all scales. This is to be expected since the spectrum of $fGn$ is not a pure power-law. In fact, for $fGn$, when $N = 3$ and $n = 10000$, the MSE defined cutoff is $j_1^{\text{MSE}} = 3$. In the literature however one commonly sees semi-parametric estimators for $fGn$ being applied over all scales.

### 3.2 Finding the Lower Scale

The $j_1^{\text{MSE}}$ above is a theoretical quantity introduced to give a precise meaning to the idea of onset scale. The MSE based definition however is not well suited to a measurement of the onset scale in practice from sample data, as it involves measuring the bias at each octave when the true value of the exponent is unknown.

An alternative measure of quality is a Chi-squared goodness of fit statistic. This can be defined in the exact setting described above as

\[ V(j_1) = \sum_{j=j_1}^{j_2} \frac{(s_j - (\hat{\alpha} j + \hat{c}))^2}{\sigma_j^2}, \]

\[ Q(j_1) = 1 - F_{j-2}(V(j_1)), \]  

where $F_m$ is the cumulative distribution function of a Chi-squared random variable with $m$ degrees of freedom, and $J = j_2 - j_1 + 1$ is the width of the scaling range. The $J - 2$ degrees of freedom arises from having $J$ points with two constraints: the determination of the slope and the intercept. The deterministic quantity $V$ represents a weighted squared distance of the fit from the exact LD, which is then mapped into $Q$, which has a range of $[0, 1]$, and an interpretation as a likelihood. To see this, consider the sample version of relations (15) and (16):

\[ \hat{V}(j_1) = \sum_{j=j_1}^{j_2} \frac{(y_j - (\hat{\alpha} j + \hat{c}))^2}{\sigma_j^2}, \]

\[ \hat{Q}(j_1) = 1 - F_{j-2}(\hat{V}(j_1)), \]

where the $s_j$ in (15) are replaced by the sample values $y_j$ from (7) and the exact fit $\hat{\alpha} j + \hat{c}$ by the sample fit $\hat{\alpha} j + \hat{c}$. Then by discounting the weak correlations as we did in section 2.2.1, and idealising the $y_j$ as Gaussian, the $\hat{V}(j_1)$ are Chi-squared random variables with $J - 2$ degrees of freedom.

The deterministic $Q(j_1)$ from (16) will be called the “exact $Q(j_1)$” whereas the $\hat{Q}(j_1)$ of (18) will be termed the “sample $Q(j_1)$”.
An example of an exact $Q(j_1)$ is given in figure 1 with $j_2 = 10$ (the largest value $j_1$ can take is $j_2 - 2$). We note a remarkable feature, that the increase with $j_1$ is extremely dramatic even on a logarithmic scale until the onset scale is reached, where the graph becomes very level ($Q(1)$ was so small it was not possible to calculate it, its logarithm was therefore truncated at $-10$). This feature has two generic and robust causes.

1. By definition, for $j$ values before the LRD regime there is a systematic departure of the corresponding $s_j$ from the asymptotic line. Consequently, as $j_1$ decreases in that range, the values of $V(j_1)$ become steadily larger, and since a Chi-squared distribution function has a tail with an exponential decay, such increases are seen as wildly improbable, resulting in dramatic (relative) drops in the corresponding $Q(j_1)$. These relative drops translate into absolute drops when considering $\log_2 Q(j_1)$.

2. The first reason is greatly amplified by the fact that, because of the dyadic nature of the wavelet transform, $n_j \approx n 2^{-j}$ and so the weights $w_j$ strongly favor the data at the smaller scales. The regression lines tend therefore to orient themselves towards $s_j$ at low $j$. Thus, as $j_1$ decreases, as soon as a systematic departure begins, the weights underlying $V(j_1)$ make it a large departure.

The first cause applies equally to upper and lower cutoffs, whilst the second involves the scale dependence of the regression weights and so applies only to lower cutoffs. Neither is strongly dependent on distributional assumptions, despite the fact that the Chi-squared statistic is motivated by Gaussian assumptions. The two causes combine to keep the goodness of fit extremely poor until, roughly speaking, each confidence interval in the fitted range cuts the fitted line. Figure 2 shows examples of exact $Q(j_1)$ plots (thick lines in second row) for three other FARIMA models. In each, the transition from radical improvement to ‘constant’ goodness of fit is very sharp, occurring over just one or two octaves.

Figure 2: Exact LD, Exact $Q(j_1)$ and 5 sample $Q(j_1)$ for different FARIMA(1, d, 1) models. Each column corresponds to a model, from left to right models 10, 24, and 7 from table 1: Left: $(\phi, d, \theta) = (0.5, 0, 0)$ Middle: $(\phi, d, \theta) = (0.3, 0.4, 0.7)$ Right: $(\phi, d, \theta) = (0, 0.1, 0)$. The top row gives exact LDs with $j^{MSE}$ marked by black squares, the lower the goodness of fit with $j^*$ indicated by filled circles (thick lines give exact $Q$, broken lines sample $Q$). For the exact $Q(j_1)$, one finds $j^* = j^{MSE}$ for each model (note different horizontal scales). For each model, $N = 3$ and $n = 10000$. 

8
The onset detection method is based on exploiting the robustness and distinctness of the transition in $Q(j_1)$ from the ‘rapid improvement zone’, to the ‘alignment zone’. The idea is that because of this sharpness, a scale $j_1^*$ corresponding to the transition can be found, which should be a good surrogate for $j_1^{MSE}$. The concept is well suited to practical implementation, as a goodness of fit test can be readily evaluated on data, and the dramatic improvement phenomenon is highly robust to statistical variations.

For each model in figure 2, five sample $Q(j_1)$ curves have been superimposed in the lower plots. Two key observations can be made on the difference between the exact and sample goodness of fit functions. First, the strongly increasing nature of the improvement zone is not changed by the addition of fluctuations. Conversely, in the alignment zone the fluctuations dominate and may even generate local sharp increases which rival those from the improvement zone, and could be confused with them. These considerations inform the following algorithm for the detection of the transition scale.

**Transition scale $j^*$ detection algorithm**

1. Determine the $j$ range $[1, j_{ND}]$ in which $Q(j_1)$ is non-decreasing. If $j_{ND} = 1$ then set $j_1^* = 1$, else

2. Calculate the improvement ratios $r_j = Q(j)/Q(j - 1)$ for each $j \in [2, j_{ND}]$. Select a rapid improvement factor $fac$ and find the largest $j$ such that $r_j > fac$. If there is no such $j$, then set $j_1^* = 1$. If there is such a $j$, set $j_1^*$ equal to it. Finally, set $j_1^* = j_1^* + 1$.

The idea behind phase 1 is that a decrease in $Q(j_1)$ is a clear indication that the rapid improvement zone has been passed, so there is no point searching beyond it, indeed it can be counterproductive. The idea behind phase 2 is that by definition the rapid improvement zone continues until there are no more large improvements in $Q$, as measured by a multiplicative increase exceeding a critical magnitude. For example we take $fac = 10$, corresponding to a unit shift in the logarithmic plots of figures 1 and 2, where a symbol is placed over $j_1^*$. The final step of phase 2, incrementing $j_1^*$ by 1, is an empirically justified calibration associated to the value of $fac$, and will be discussed further below. The performance of the method, and the determination of the parameter $fac$, is the subject of the next section.

### 3.3 A User’s Guide

In this section we summarize the steps required to implement the method, as an aid to those wishing to use it. We assume that a discrete Gaussian time series $X(t)$ of length $n$ is being studied, which using a mother wavelet with a number of vanishing moments $N$, yields wavelet coefficients over scales $j = 1$ up to an upper scale $j = j_2$ (the value of $j_2$ is determined automatically by $n$ and the filtering edge effects implied by the choice of $N$).

The steps are:

- **Wavelet analysis:**
  - Perform the prefiltering of $X$ according to equation (6).
  - Recursively calculate the $d_X(j, k)$ using the DWT.
  - Calculate the $y_j$ from equations (7) and (12).

- **Goodness of fit calculation:** For each $j$ in $[1, j_2 - 2]$:
  - Perform the weighted regression according to equations (7), (8), (9), (10) and (11), to obtain the fit: $\hat{est}_j = \hat{\alpha}_j + \hat{\beta}$ over scales $[j_1, j_2]$.
  - Equations (17) and (18) yield the sample goodness of fit $\hat{Q}(j_1)$.

- **Transition Algorithm:**
– operate on the sample \( \hat{Q}(j_i) \) function as described in section 3.2 (transition scale detection algorithm) to obtain the sample \( j_1^* \).

A Matlab based implementation of the above procedure can be found at [13].

4 Performance

4.1 Method comparison

Comparing the upper row of figure 2, where the \( j_1^{\text{MSE}} \) are marked by black squares, and the lower row, where the \( j_1^* \) values are also marked, we see that the algorithm’s performance is promising across a wide range of LRD exponent values as well as a variety of behaviours at small scale. We now investigate its performance more systematically by comparing \( j_1^{\text{MSE}} \) against the exact \( j_1^* \), and the statistics of sample \( j_1^* \), for 34 different Gaussian time series models (the models are the same as those used in [10] and [1]). The models can be split into 7 groups according to the nature of their short range dependence (SRD). Within each group, as the model number advances the LRD exponents steadily increase as detailed in table 1.

<table>
<thead>
<tr>
<th>Model number</th>
<th>Model class</th>
<th>LRD exponent values</th>
<th>SRD parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 5</td>
<td>( \text{fGn}(H) )</td>
<td>( H = {0.5, 0.6, 0.7, 0.8, 0.9} )</td>
<td>( \text{N/A} )</td>
</tr>
<tr>
<td>6 – 9</td>
<td>FARIMA(0, d, 0)</td>
<td>( d = {0, 0.1, 0.2, 0.3} )</td>
<td>( \text{N/A} )</td>
</tr>
<tr>
<td>10 – 14</td>
<td>FARIMA(1, d, 0)</td>
<td>( d = {0, 0.1, 0.2, 0.3, 0.4} )</td>
<td>( \phi = 0.5 )</td>
</tr>
<tr>
<td>15 – 19</td>
<td>FARIMA(0, d, 1)</td>
<td>( d = {0, 0.1, 0.2, 0.3, 0.4} )</td>
<td>( \theta = 0.5 )</td>
</tr>
<tr>
<td>20 – 24</td>
<td>FARIMA(1, d, 1)</td>
<td>( d = {0, 0.1, 0.2, 0.3, 0.4} )</td>
<td>( (\phi, \theta) = (0.3, 0.7) )</td>
</tr>
<tr>
<td>25 – 29</td>
<td>FARIMA(1, d, 1)</td>
<td>( d = {0, 0.1, 0.2, 0.3, 0.4} )</td>
<td>( (\phi, \theta) = (-0.3, -0.7) )</td>
</tr>
<tr>
<td>30 – 34</td>
<td>FARIMA(1, d, 1)</td>
<td>( d = {0, 0.1, 0.2, 0.3, 0.4} )</td>
<td>( (\phi, \theta) = (0.7, 0.3) )</td>
</tr>
</tbody>
</table>

Table 1: Time series models used to test the transition algorithm.

For each model 50 independent realisations were generated using an exact generation method, and for each a \( j_1^* \) sample was calculated using the above algorithm with \( \text{fac} = 10 \) (the realisations are the same as those used in [10] and [1]). The average, rounded to the nearest integer, of these 50 sample \( j_1^* \) was computed and is reported in figure 3. Using the median or the mode instead of this rounded average sample statistic produced virtually identical results, as the sample distributions of \( j_1^* \) have narrow support, typically just two or three integers, with most of the mass on a single value. For example, for model 10 (on the left in figure 2) the support was \( \{4, 5, 6\} \), with respective masses \( \{0.30, 0.68, 0.02\} \). In rare instances the mass was fairly evenly split between adjacent values.

The results are given in figure 3, where the different models are spread out over the horizontal axis. The \( j_1^{\text{MSE}} \) values are marked with a large circle, and we see that they fall into groups following those of table 1, confirming the role of the SRD parameters in fixing the onset scale. Extra tick marks flag the models we have already encountered: model 22 from figure 1, and models 10, 24, and 7 from figure 3.

The exact \( j_1^* \) values are given by the diamonds. They agree in almost all cases with the summary statistic for the corresponding empirical \( j_1^* \) values, marked by the asterisks. Comparing the empirical transition scales with the \( j_1^{\text{MSE}} \) values, we see that in most cases they coincide, and that they differ by at most one, suggesting that the heuristic is effective in practice independently of the details of the SRD or the LRD.

It is significant that the exceptions fall in the fGn and FARIMA(0, d, 0) groups. For these processes, and particularly in the case of FARIMA(0, d, 0), the wavelet spectrum is very close to a straight line, and so the rapid improvement zone is very small (see model 7 in figure 2, right plot) and may even not exist. As the method is based on detecting the end of the improvement zone, it cannot perform perfectly in such cases.
Value-based transition scale $j^*$ detection algorithm

1. Set $j^*_1$ equal to the first $j_1$ such that $Q(j_1) > 0.05$ (corresponding to a 5% significance level under Gaussian assumptions). If there is no such $j$, set $j^*_1$ equal to the first $j_1$ such that $Q(j_1) > 0.1$. If there is no such $j$, set $j^*_1$ equal to the value where $Q(j_1)$ is a maximum.

The essence of this method is that in a Gaussian setting we may interpret $Q$ as a likelihood and select according to its value, viewed as a significance level. Thus the scaling region begins at the first $j_1$ where the graph aligns to a significance level of 5%. This is possible here as the time series, and hence the wavelet coefficients, are Gaussian, and so the $y_j$ values in the Logscale Diagram are approximately Gaussian and we have accurate expressions for their variances.

Figure 4 reports the results for the value-based method. We see that despite the fact that it makes use of stronger assumptions on the data (namely gaussianity), the results are no better, in fact worse, than the heuristic of section 3.2. If we added an ad hoc increment step, $j^*_1 = j^*_1 + 1$, the results would improve, but would still be worse than the method we propose.

4.2 Sensitivity/Evaluation of $fac$

Our transition detection algorithm contains an arbitrary parameter, $fac$, whose value has been set to 10 in the results above. In this section we show that the results are not sensitive to the value of $fac$, provided it is chosen large enough. This can be understood in terms of the idealised form of the graph of exact $Q(j_1)$: a discontinuity followed by a constant. With an infinitely sharp transition, any positive value of $fac$ will select the same transition point. With the addition of variability in the case of sample $Q(j_1)$, the result becomes sensitive to small values of $fac$, but the logic holds true at larger values. With the replacement
Figure 4: **Performance of the value-based method.** For each model (with $N = 3$, and $n = 10000$): $j_1^{MSE}$ (circle), exact $j_1^*$ (diamond), rounded average of 50 sample $j_1^*$ (asterisk). The value-based method performs worse despite using the assumption of Gaussianity explicitly.

of the discontinuity by a zone of very rapid increase, $j_1^*$ becomes a non-increasing function of $fac$, but the sensitivity to its value remains very low.

Figure 5 shows $j_1^*$ values as a function of $fac$ for each model, that is, 34 overlapping non-increasing functions are plotted. Exact $j_1^*$ values are used, since figure 3 confirms that they correspond closely to the rounded mean statistic for the ensemble. In the range $fac \leq 4$ there is considerable variability in the $j_1^*$ values, whereas beyond $fac = 5$ they are almost constant. In particular, in the range $[5, 10]$ the $j_1^*$ are constant for each of the 34 models, justifying our choice of $fac = 10$.

Each $j_1^*$ value is in fact a non-increasing function of $fac$. Beyond the small values where the sample variability plays a role, increasing $fac$ corresponds to detecting the end of the rapid improvement zone earlier, at some distance from the true transition. Thus, as $fac$ increases, it is necessary to correct for this offset by adding a compensating increment. In the case of $fac = 10$, we have empirically chosen this value to be 1, as encoded in the algorithm definition. Increasing $fac$ further is not desirable, as the greater the offset from the true transition point, the less able is the method to detect transition values which are small.
5 Discussion

We have presented a method to detect the onset scale of long-range dependence, or other scaling models characterised by lower cutoff scales. The method has several advantages. Some are derived from the underlying wavelet definition of cutoff scale. These include a uniform approach and a method applicable to scaling processes of different types (LRD, \(1/f^\alpha\) noise, processes with stationary increments and asymptotic scaling, ...), and a non-iterative algorithm with a computational complexity of \(O(n)\). Other advantages stem from the nature of the underlying idea, the rapid drop in goodness of fit when even a small number of ‘wrong’ scales are included in the estimation range. The method is not based on the values of the goodness of fit parameter, which are process dependent, but on robust features of the goodness of fit function.

Figure 6: Sample LD, Sample \(Q(j_1)\) for Internet data. Each column corresponds to a time series (the top row gives the first 1500 samples, the middle row the logscale diagram with the regression line whose slope is \(\hat{\alpha}\), the bottom row the sample goodness of fit): Left: number of TCP connection arrivals in 40ms intervals (\(\hat{\alpha} = 0.59\) over \((j_1, j_2) = (7, 16)\)). Right: number of IP packets carrying TCP in 40ms intervals (\(\hat{\alpha} = 0.72\) over \((j_1, j_2) = (10, 16)\)). The values of \(j^*\) are indicated by the symbols in the bottom plot. Note the scale change in the horizontal axis between the middle and bottom plots.

The method could be readily adapted to detect upper cutoff scale, although its performance would be weaker as the confidence intervals in the wavelet log-spectrum at large scale are much smaller than at small scale, so the departures from linear behaviour would not be greatly amplified as they are for the lower cutoff. However, provided the upper cutoff is well under the length of the data, so that the confidence intervals
remain small, the method should still work adequately. Note that for non-Gaussian data, the variances $\sigma_j^2$ may be larger. In such a case the magnification effect which creates the rapid change in goodness of fit past the transition scale will be lower, and the method more sensitive to sample variability. Of course a rigorous testing of the method for non-Gaussian data would first require that accurate expressions for the $\sigma_j^2$ be found, so that $j_{\text{MSE}}^{\text{MSE}}$ could be calculated.

The method involves two arbitrary choices, the value of the rapid improvement factor $\text{fac}$, and the value of the calibrating increment added to form the final onset scale estimate. However, we have shown that the sensitivity to $\text{fac}$ is very low, and some calibration is unavoidable in a method which involves so few assumptions on the data.

We conclude with two examples using real data. Each of the time series in figure 6 derives from the same raw Internet data, captured from a telecommunications link connecting the University of Auckland, New Zealand, to the external Internet. They describe different aspects of the Transfer Control Protocol (TCP) connection process. TCP traffic currently dominates Internet traffic and is used for example to transport the contents of web pages. Time windows were chosen in which the time series seemed stationary. We see that the data appears long-range dependent, and that the method gives sensible answers for the onset scales. We have found that the method performs well in a wide variety of real world cases as well as for numerous models, including data with infinite variance.

Acknowledgements

We are grateful to the WAND group at the University of Waikato for making the data sets available. Darryl Veitch gratefully acknowledges the support of Ericsson and Murad S. Taqqu that of the US National Science Foundation for the NSF Grants DMS-0102410 and ANI-9805623 at Boston University. Patrice Abry been partially supported by the French MENRT grant “ACI jeune Chercheur 2329”.

References


