A Practical Guide to Measuring the Hurst Parameter

What you need to know to measure long-range dependence.

Richard G. Clegg (richard@richardclegg.org)

Department of Mathematics,
University of York,
YO10 5DD

UK Performance Engineering Workshop, July 2005

Introduction

- Long-Range Dependence (LRD) is a statistical phenomenon describing persistent correlations.
- Presence and nature of LRD is characterised by H the Hurst Parameter.
- The Hurst Parameter is perfectly well-defined. A large number of theoretically sound estimators exist.

Introduction

- Long-Range Dependence (LRD) is a statistical phenomenon describing persistent correlations.
- Presence and nature of LRD is characterised by H the Hurst Parameter.
- The Hurst Parameter is perfectly well-defined. A large number of theoretically sound estimators exist.
- The existing estimators disagree when applied to the same data.
- Trends and periodicities or other corrupting noise may be mistaken for LRD.
- The literature gives different approaches for pre-processing data before measurement.
- This paper is a very simple overview for someone who wants to measure LRD.

The Autocorrelation Function and Hurst Parameter

Let $\{X_1, X_2, X_3, \ldots\}$ be a weakly stationary time series.

The Autocorrelation Function (ACF)

$$\rho(k) = \frac{(\mathrm{E}\left[X_t - \mu\right])(\mathrm{E}\left[X_{t+k} - \mu\right])}{\sigma^2},$$

where μ is the mean and σ^2 is the variance.

The ACF measures the correlation between X_t and X_{t+k} and is normalised so $\rho(k) \in [-1, 1]$. Note symmetry $\rho(k) = \rho(-k)$.
A process exhibits LRD if $\sum_{k=0}^{\infty} \rho(k)$ diverges (is not finite).

Definition of Hurst Parameter

The following functional form for the ACF is often assumed

$$\rho(k) \sim C_\rho |k|^{-2(1-H)},$$

where $\sim$ means asymptotically equal to, $C_\rho > 0$ and $H \in (1/2, 1)$ is the Hurst Parameter.

More About LRD

The time series $\{X_t : t \in \mathbb{N}\}$ can be converted to the frequency domain using a Fourier transform. Let $f(\lambda)$ be the spectral density of the series at frequeney λ.

LRD in the Spectral domain

In the spectral domain, the previous definition becomes

$$f(\lambda) \sim C_f |\lambda|^{-(2H-1)},$$

as $\lambda \to 0$, where $C_f > 0$ and H is the Hurst paramter.

- Computationally, LRD is difficult to work with.
- LRD is hard to measure — estimates at low frequencies or high lags.
- The sample mean converges at a rate proportional to n^{2H-2} not n^{-1}.
- Standard techniques for confidence intervals fail to work.

Why do we care about LRD?

- In 1993 LRD (and self-similarity) was found in a time series of bytes/unit time measured on an Ethernet LAN [Leland et al '93].
- This finding has been repeated a number of times by a large number of authors (however recent evidence suggests this may not happen in the core).
- A higher Hurst parameter often increases delays in a network. Packet loss also suffers.
- If buffer provisioning is done using the assumption of Poisson traffic then the network will be underspecifed.
- The Hurst parameter is a dominant characteristic for a number of packet traffic engineering problems.
- The origins of LRD are uncertain but the most likely cause seems to be the aggregation of file transfer processes.

LRD, Self-Similarity and Heavy Tails

- Statistically Self-Similar: The distribution of a process $\{Y_t : t \in \mathbb{N}\}$ is the *same* after stretching $Y_t \stackrel{d}{=} c^{-H} Y_{ct}$ for some constant $c > 0$. Examples: coastlines, tree-bark, internet traffic traces.
- If Y_t is stat. self similar with $H \in (1/2, 1)$ with stationary increments $X_t = Y_t - Y_{t-1}$ then X_t has LRD and same Hurst parameter H.
- Heavy Tailed: Distribution where extreme events still have a significant likelihood. $\mathbb{P}[X > x] \sim x^{-\beta}$ for $\beta \in (0, 2)$ Examples: heights of trees, frequencies of words, lengths of file in the internet.
- A process where the lengths of the on and off periods are heavy-tailed will exhibit LRD.

Estimators for the Hurst parameter

There are a large number of estimates for the Hurst parameter. Five are used in this paper.

1. The R/S plot is the oldest and perhaps best known estimator for the Hurst parameter.
2. Aggregated variance looks at how the variance of a series changes as it is aggregated.
3. The Periodogram looks at the behaviour of an estimate for the spectral density.
4. Wavelets are a method which can be considered as a generalisation of Fourier transform.
5. Local Whittle estimator looks at the behaviour of the frequency spectrum near the zero frequency.

The first two estimators are in the time domain and the last three in the frequency domain.

Simulated Data — Procedure

- Simulated data sets with known Hurst parameter are generated using Fractional Gaussian Noise and Fractional ARIMA modelling.
- The data set is then corrupted by the addition of noise of the following types:
  1. An AR(1) process with a high degree of short-range dependence.
  2. A sin wave.
  3. A linear trend.
- All five estimators are then applied to the raw and the corrupted data.

Simulated Data — Results (1)

Added Noise	R/S Plot	Aggreg. Variance	Period. ogram	Wavelet Estimate	Local Whittle
100,000 points FGN — H= 0.9.					
None	0.782	0.864	0.905	0.901 ± 0.009	0.934
AR(1)	0.805	0.784	0.88	0.969 ± 0.042	1.066
Sin	0.772	0.961	0.907	0.901 ± 0.009	0.945
Trend	0.782	0.958	0.928	0.901 ± 0.009	0.939
100,000 points FARIMA (0,d,0) — H = 0.7					
None	0.663	0.692	0.699	0.696 ± 0.004	0.681
AR(1)	0.823	0.673	0.792	0.896 ± 0.033	0.876
Sin	0.665	0.972	0.704	0.696 ± 0.004	0.765
Trend	0.662	0.973	0.786	0.696 ± 0.004	0.746
100,000 points FARIMA (1,d,1) — H= 0.7, $\phi_1 = 0.5, \theta_1 = 0.5$.					
None	0.684	0.693	0.706	0.697 ± 0.006	0.68
AR(1)	0.818	0.656	0.774	0.88 ± 0.041	0.878
Sin	0.689	0.973	0.71	0.697 ± 0.006	0.766
Trend	0.684	0.972	0.786	0.697 ± 0.006	0.743

Simulated Data — Results (2)

Added Noise	R/S Plot	Aggreg. Variance	Period. ogram	Wavelet Estimate	Local Whittle
100,000 points FARIMA (0,d,0) — H $= 0.9$.					
None	0.757	0.882	0.91	0.886 ± 0.004	0.861
AR(1)	0.804	0.789	0.873	0.969 ± 0.036	1.011
Sin	0.764	0.967	0.913	0.886 ± 0.004	0.883
Trend	0.757	0.974	0.933	0.886 ± 0.004	0.875
100,000 points FARIMA (1,d,1) — H$= 0.9$, $\phi_1 = 0.5$, $\theta_1 = 0.5$.					
None	0.856	0.854	0.881	0.887 ± 0.006	0.858
AR(1)	0.888	0.773	0.874	0.959 ± 0.04	1.001
Sin	0.86	0.963	0.885	0.887 ± 0.006	0.879
Trend	0.856	0.968	0.92	0.887 ± 0.006	0.872
100,000 points FARIMA (2,d,1) — H$= 0.7$, $\phi_1 = 0.5$, $\phi_2 = 0.2$, $\theta_1 = 0.1$.					
None	0.807	0.74	0.817	0.966 ± 0.048	1.05
AR(1)	0.814	0.691	0.822	1.007 ± 0.059	1.136
Sin	0.8	0.94	0.821	0.966 ± 0.048	1.052
Trend	0.807	0.939	0.856	0.966 ± 0.048	1.051

Real Data — Procedure

- Two real-life data sets are analysed in this paper:
    1. A data set collected at the University of York in 2001. A 67 minute trace of incoming and outgoing data from the University.
    2. The much studied Bellcore data set — this was collected in 1989 and has been used in many famous papers.
- The literature gives the following suggestions for pre-processing data before estimating the Hurst parameter.
    1. Taking logs of the time series (only appropriate if data is positive).
    2. Removal of mean and a trend from the data.
    3. Removal of a best fit polynomial of high order (ten is chosen here).

Real Data — Results

Filter Type	R/S Plot	Aggreg. Variance	Period. ogram	Wavelet Estimate	Local Whittle
York trace (bytes/tenth) — 40467 points					
None	0.826	0.924	0.928	0.909 ± 0.012	0.881
Trend	0.826	0.923	0.932	0.909 ± 0.012	0.881
Poly	0.827	0.892	0.863	0.909 ± 0.012	0.878
Bellcore data BC-Aug89 (bytes/10ms) — first 1000 secs.					
None	0.707	0.8	0.817	0.786 ± 0.017	0.822
Trend	0.707	0.797	0.815	0.786 ± 0.017	0.822
Poly	0.707	0.789	0.787	0.786 ± 0.017	0.822
Bellcore data BC-Aug89 (bytes/10ms) — second 1000 secs.					
None	0.62	0.802	0.808	0.762 ± 0.012	0.825
Trend	0.62	0.802	0.808	0.762 ± 0.012	0.825
Poly	0.618	0.786	0.777	0.762 ± 0.012	0.824

Conclusions

- Even in artificial data measuring the Hurst parameter can be hit and miss.
- Corrupting noise of various types can harm measurements and all techniques were vulnerable to addition of short-range dependent data.
- Techniques used to pre-process the data seemed to make little difference.
- A researcher relying on a single measure of the Hurst parameter is likely to be drawing false conclusions.

References

1) This talk online at www.richardclegg.org/pubs/.
2) Clegg, R. G. (2004) "The Statistics of Dynamic Networks" PhD Thesis. Dept of Maths. University of York. Online at: www.richardclegg.org/pubs/thesis.pdf
3) Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1993). *On the self-similar nature of Ethernet traffic*. In Proc. ACM SIGCOMM, pages 183193, San Francisco, California.
4) Software used in this paper: http://www.richardclegg.org/lrdsources/software/