SELFIS: A Short Tutorial

Thomas Karagiannis (tkarag@cs.ucr.edu)

November 8, 2002

This document is a short tutorial of the SELF-similarity analysIS software tool. Section 1 presents briefly useful definitions. Section 2 describes the bucket shuffling methodology. Finally, section 3 is a quick manual of how to use SELFIS.

1 Definitions

- **Autocorrelation Function (ACF):** Autocorrelation is a statistical measure of the relationship, if any, between a random variable and itself, at different time lags. The autocorrelation coefficient can range between +1 (very high positive correlation) and -1 (very high negative correlation). The Autocorrelation function (ACF) shows the value of the autocorrelation coefficient for different time lags k:

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

- **Long-Range Dependence:**
  Long-range dependence measures the memory of a process. Intuitively, distant events in time are correlated. This correlation is captured by the autocorrelation function (ACF), which measures how similar is a series with shifted version of itself. For LRD data the ACF decays very slowly to zero. On the contrary, short-range dependence is characterized by quickly decaying correlations (e.g. ARMA processes). The strength of the long-range dependence is quantified by the Hurst exponent (H). A series exhibits LRD when $\frac{1}{2} < H < 1$. Furthermore, the closer H is to 1, the stronger the dependence of the process is.

More rigorously, a stationary process X_t has long-memory or is long-range dependent, if there exists a real number $\alpha \in (0, 1)$ and a constant $c_p > 0$ such that

$$\lim_{k\to\infty} \rho(k)/[c_p k^{-\alpha}] = 1$$

where $\rho(k)$ the **sample Autocorrelation function (ACF)**. The definition states that the autocorrelation function of long-memory processes, decays to zero with rate approximately $k^{-\alpha}$, where $H = 1 - \frac{\alpha}{2}$ is the Hurst exponent.

- **Hurst Estimators:**
  There are many estimators that are used to estimate the value of the Hurst parameter. Below is a list of the estimators implemented in SELFIS.

  1. *Absolute Value method*, where an aggregated series $X^{(m)}$ is defined, using different block sizes m. The log-log plot of the aggregation level versus the absolute first moment of the aggregated series $X^{(m)}$ should be a straight line with slope of H-1, if the data are long-range dependent.
  2. *Variance method*, where the log-log plot of the sample variance versus the aggregation level must be a straight line with slope β greater than -1. In this case $H = 1 + \frac{\beta}{2}$.
  3. *R/S method.*A log-log plot of the R/S statistic versus the number of points of the aggregated series should be a straight line with the slope being an estimation of the Hurst exponent.
  4. *Periodogram method.* This method plots the logarithm of the spectral density of a time series versus the logarithm of the frequencies. The slope provides an estimate of H. The Periodogram is given by

     $$I(\nu) = \frac{1}{2\pi N} |\sum_{j=1}^{N} X(j) e^{ij\nu}|^2$$

     where ν is the frequency, N is the length of the time series and X is the actual time series.
  5. *Whittle* estimator. The method is based on the minimization of a likelihood function, which is applied to the Periodogram of the time series. It gives an estimation of H and produces the confidence interval. It does not produce a graphical output.

6. *Variance of Residuals.* A log-log plot of the aggregation level versus the average of the variance of the residuals of the series should be a straight line with slope of H/2.
7. *Abry-Veitch.* Wavelets are used in order to estimate the Hurst exponent. The energy of the series in various scales is studied to provide an estimate.

2 Bucket Shuffling

Bucket shuffling decouples the short-range from long-range correlations in a series to study the effects of long-range dependence. This is achieved through partitioning the time series into a set of "buckets" of length b. Thus, we define the contents of the uth bucket to be items $X_{u \cdot b}, \ldots, X_{(u+1) \cdot b-1}$ from the series, and the **home** of item X_i to be bucket $H(i) \equiv \lfloor i/b \rfloor$. Also, we say that two items (X_i, X_j) form an **inbucket** pair if $H(i) = H(j)$; otherwise, they form an **outbucket** pair with an **offset** of $|H(i) - H(j)|$ buckets. Note that this classification depends on the (fixed) locations of the bucket boundaries, and not just the separation between two items in the time series. Once the series has been partitioned in this way, we can then apply one of the following bucket shuffling algorithms to reorder its items:

- **External Shuffling (EX):** The order of buckets is shuffled, whereas the content of each bucket remains intact. This can be achieved by labelling each bucket with a bucket-id between 0 and $\lfloor$Time-SeriesLength$/b \rfloor$, and shuffling the bucket-ids. External shuffling preserves all correlations among the inbucket pairs, while equalizing all correlations among the outbucket pairs with different offsets. Thus, if the series is sufficiently long, the ACF should not exhibit significant correlations beyond the bucket size.
- **Internal Shuffling (IN):** The order of the buckets remains unchanged while the contents of each bucket are shuffled. As a result, correlations among the inbucket pairs are equalized, while correlations among the outbucket pairs are preserved, but rounded to a common value for each offset. Thus, if the original signal has long-memory, then the ACF of the internally-shuffled series will still show power-law behavior.
- **Two-Level Shuffling (2L):** Each bucket is further subdivided into a series of "atoms" of size a. Thereafter, we apply external shuffling to the block of $\lfloor b/a \rfloor$ atoms within each bucket. As a result,

both short-range correlations (within each atom) and long-range correlations (across multiple buckets) are preserved, while medium-range correlations (across multiple atoms within the same bucket) are equalized.

3 SELFIS Manual

- **Input:** The input file must be a timeseries in the following two formats:

```
x1 y1
x2 y2
x3 y3
 .
 .
```

OR

```
 y1
 y2
 y3
 y4
 .
 .
```

  The input file must have at least 64 values. When the file is plotted the X axis always denotes the number of observations in the input file and not the true value of the x1,x2,x3... if any. A sample input file is fgn08, which was generated with a fractional Gaussian noise generator.

- **Hurst Exponent Estimation:** In the "Hurst Exponent Estimation" menu item the preferred estimator can be selected after a file was opened. Note that depending on the size of the input file, the Whittle and Abry-Veitch estimators may take up to a few minutes to compute the Hurst exponent. The "Run All Estimators" option opens a new form that displays the results for each of the seven estimators.

- **Bucket Shuffling:** In the "Bucket Shuffling" menu item, three types of bucket shuffling can be performed. The "Restore Original Data" option cancels any shuffling that was performed and re-plots the original

input file. After performing any shuffling all the estimations (Hurst estimation, Autocorrelation function, etc) refer to the new shuffled time series. You can also save the shuffled time series with the save option in the file menu item.

- **Tools:** The "Tools" menu item includes the calculation of basic statistics (mean, variance, standard deviation, skewness, kurtosis), the estimation of the autocorrelation function and the power spectrum and the zoom in and out in the plot. You can also zoom in by defining a rectangle by clicking in the plot and dragging the mouse. The rectangle defines the area that will be shown. Zooming out can be performed by right mouse click inside the plot.

ACKNOWLEDGEMENTS

I am especially grateful to Dr. M. Faloutsos, Dr. M. Molle and Dr. R. H. Riedi for their help and contributions throughout this project. Thanks are also due to Theodoros Folias for the splash screen of SELFIS.