New Developments in Time Series Analysis

Jeff Scargle
NASA Ames Research Center

- Data Cell Representation
- Time Series Models
- Time-Domain Analysis
- Frequency-Domain Analysis
- Time-Scale/Frequency Distributions

Digging Deeper: Algorithms for Computationally-Limited Searches in Astronomy
June 7, 2011  Keck Institute for Space Studies
Data Cell

Contains all information relevant to an analysis task. Index $i = 1, 2, \ldots, N$ denotes individual observations.
N arbitrary time-ordered data cells make a time series.
Height = 1 / dt
n / dt
E / dt
Area = $\frac{1}{dt'}$

\[
\begin{align*}
n / dt' \\
E / dt'
\end{align*}
\]

$dt' = dt \times \text{exposure}$
Models
The Wold Decomposition Theorem

Herman Wold (1908–1992) Econometrician

\[ X(t) = C \ast R(t) + D(t) \]

Any stationary process is the sum of:
- a moving average (white noise R input to a filter C)
- a linearly deterministic process D

Pulse shape C is …
- Convergent
- Causal, by assumption
- Minimum delay
- Constant

Pulse amplitudes R are …
- Uncorrelated (“white”)

An extension of the theorem (Scargle 1981) relaxes some restrictions

**Powerful existence proof of a random flare representation for any stationary time series**
Testing for Stationarity

- Formal definition requires infinite amount of data
- Local stationarity depends on scale
- Construct stationarity measure $S[ x(t) ]$
  - E.g. variance of TF distribution vs. time marginal
  - Any such measure has statistical fluctuations
  - Simulate surrogate data: scramble Fourier phase
- Construct distribution of $S( \text{surrogate data} )$

*Testing Stationarity with Time-Frequency Surrogates*, Jun Xiao, Pierre Borgnat, and Patrick Flandrin
From: Flandrin & Borgnat
“Revisiting and testing stationarity,” 2008

... interpreted as “stationary” or “nonstationary” depending on the observation scale ...

TL: nonstationary

TR: stationary (periodic)

BL: nonstationary

BR: stationary (homogeneous texture)
Processes Involved in Astronomical Time Series

- Variable Source
- Propagation To Observer
- Photon Detection

- Luminosity fluctuations (Random or Deterministic)
  - Photon Emission (Poisson Process)
  - Scintillation
  - Dispersion
  - … (Random)
  - Photon Detection: (Poisson Process)

Correlation here … does not mean … Correlation here!
Time Domain
Bayesian (or Maximum Likelihood) Blocks

Signal represented as *constant in elements of data-space partition* Optimize by maximizing model fitness over all possible partitions.

- Nonparametric

- Assumptions: prior on amplitudes and number of blocks

- No limitation on resolution in the independent variable (no bins!)

- *Real-time mode: first significant signal above background*

- Representation, discontinuous, convenient for further analysis

- Local structure, not global

- $O N^2$ (maybe can be faster)
Blocks

Block: a set of data cells

Fitness function = sum over blocks F( Block )
The Optimizer

best = []; last = []; 
for R = 1:num_cells 
    [ best(R), last(R) ] = max( [0 best] + ...
        reverse( log_post( cumsum( data_cells(R:-1:1, :) ), prior, type ) ) );

    if first > 0 & last(R) > first  % Option: trigger on first significant block 
        changepoints = last(R); return
    end
end

% Now locate all the changepoints 
index = last( num_cells );
changepoints = [];
while index > 1
    changepoints = [ index changepoints ];
    index = last( index - 1 );
end
Most likely of the $10^{468}$ possible segmentations of these 1554 data points!
Cross- and Auto- Correlation Functions for unevenly spaced data

Edelson and Krolik:

The Discrete Correlation Function: a New Method for Analyzing Unevenly Sampled Variability Data
NGC 4151 ASCA (Greg Madejski)
Event Data Cells in any dimension

Measurements: Point coordinates
Data Space: Space of any dimension
Signal: Point density (deterministic or probabilistic)
Data Cell: Voronoi cells for the data points

Suf. Statistics N = number of points in block
V = volume of block

Max Likelihood: \( \left( \frac{N}{V} \right)^N e^{-N} \)
Posterior: \( N! \frac{(V-N)!}{(V+1)!} \)

Example: any problem usually approached with histograms (1D)
positions of objects from a sky survey (2D)
positions of objects in a redshift survey (3D)
Wavelet Kurtosis

New Statistic to Detect and Characterize Intermittency

Daniel Engavatov, Elliott Bloom, JS; SLAC PhD Thesis
Frequency Domain
Data Mode
- Photon events
- Time-to-Spill
- Counts in bins
- Flux measurements
- Any Mode/Sampling!

Universal Time Series Analysis Machine

Auto-
- Correlation Function
- Fourier Power Spectrum
- Fourier Phase Spectrum
- Wavelet scalgram
- Wavelet scaleogram
- Structure Function
- Time-Frequency Distribution
- Time-Scale Distribution
- ...

Extension of Edelson & Krolik Algorithm for Correlation Function of Unevenly Sampled Data

Jeff Scargle
**Data Mode**

- Photon events
- Time-to-Spill
- Counts in bins
- Flux measurements
- Any Mode/Sampling!

**Universal Time Series Analysis Machine**

**Cross-**

- Correlation Function
- Fourier Power Spectrum
- Fourier Phase Spectrum
- Wavelet scalgram
- Wavelet scaleogram
- Structure Function
- Time-Frequency Distribution
- Time-Scale Distribution
- ...

**Extension of Edelson & Krolik Algorithm for Correlation Function of Unevenly Sampled Data**

Jeff Scargle
Time-Scale

Time-Frequency
Time-Frequency/Time-Scale Analysis

*Transform to a new view of the time series information.*

- A Reality in joint time & frequency (or scale) representation
- Atomic decomposition
  - Time-frequency atoms
  - Over-complete representations
  - Optimal Basis Pursuit (Mallat), etc.
- Uncertainty Principle: T-F resolution tradeoff
- Non-stationary processes
  - Flares
  - Trends and Modulations
  - Statistical change-points
- Instantaneous Frequency
- Local vs. Global structure
- Interference (cross-terms in bi-linear representation)

Time-Frequency/Time-Scale Analysis (Temps-Fréquence) Patrick Flandrin
http://perso.ens-lyon.fr/patrick.flandrin/publis.html; A Wavelet tour of Signal Processing (Une Exploration des Signaux en Ondelettes) Stéphane Mallat
SYNCHROSQUEEZEING

A consistent and invertible time-frequency analysis tool that can identify and extract oscillating components (of time-varying frequency and amplitude)


Matlab tools: http://www.math.princeton.edu/~ebrevdo/synsq/
Multi-taper Analysis (Thomson 1982)

- Tapers (windows) reduce sidelobe leakage = bias
- Incomplete use of data ➞ loss of information
- Multitapers recover this information
- Leakage minimization = eigenvalue problem
  - Eigenfunctions: efficient window functions
  - Eigenvalues
    - measure effectiveness
    - determine how many terms to include

Multivariate Hermite Taper Functions

Taper 1

Taper 2

Taper 3

Taper 4
<table>
<thead>
<tr>
<th>Function</th>
<th>Domain</th>
<th>Range</th>
<th>Auto-</th>
<th>Cross-</th>
<th>Physical Interp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian blk. Light Curve</td>
<td>Time</td>
<td>Flux</td>
<td>✔</td>
<td>✔ multivar. BB Flares, events etc.</td>
<td></td>
</tr>
<tr>
<td>Scatter Plot</td>
<td>Flux 1</td>
<td>Flux 2</td>
<td>✔</td>
<td>✔</td>
<td>Dependency (not just cor.)</td>
</tr>
<tr>
<td>Correlation</td>
<td>Lag</td>
<td>&lt;X²&gt; &lt;XY&gt;</td>
<td>✔</td>
<td>✔</td>
<td>Correlated behavior/lags</td>
</tr>
<tr>
<td>Spectrum</td>
<td>Frequency</td>
<td>Power</td>
<td>✔</td>
<td>✔</td>
<td>Periodicity 1/f noise ...</td>
</tr>
<tr>
<td>Scallogram</td>
<td>Scale/Time</td>
<td>Power</td>
<td>✔</td>
<td>✔</td>
<td>Dynamic behavior</td>
</tr>
<tr>
<td>Scalegram</td>
<td>Scale</td>
<td>Power</td>
<td>✔</td>
<td>✔</td>
<td>1/f noise QPOs</td>
</tr>
<tr>
<td>Distribution</td>
<td>Time/scale/frequency</td>
<td>Power</td>
<td>✔</td>
<td>✔</td>
<td>Dynamic behavior</td>
</tr>
</tbody>
</table>
Open Problems

- The Arrow of Time
- Bayesian Blocks $O(N^2) \rightarrow O(N)$ or $O(N \log N)$
- Optimal Partitioning on the Circle
- Negative Power Spectrum Estimates
- Phase Spectrum from Edelson and Krolik-style Correlations
- Exploration of Time-Scale/Time-Frequency Variants
- Understand Wold Representation if $C = C(t)$
- Automatic Classification of Time Series
Handbook of Statistical Analysis of Event Data

MatLab Code; Documentation; Examples; Tutorial
... funded by the NASA AISR Program

Advances in Machine Learning and Data Mining for Astronomy

Editors: Michael Way, Jeff Scargle
Kamal Ali, Ashok Srivastava

Chapman and Hall, an imprint of CRC Press
(a division of Taylor and Francis)
Randi Cohen, Computer Science Acquisitions Editor Data Mining and Knowledge Discovery series
Back-ups
Practical Suggestions

*(somewhat exaggerated)*

- Study distribution of sample intervals $dt_n = t_{n+1} - t_n$
- Never subtract mean of time series
- Edelson and Krolik CF is the source of all other analysis
- Use self terms in E&K CF to assess observational errors
- Don’t confuse: source randomness/observational noise
- $H_0$: AGNs are identical stochastic dynamical systems
- Stationarity is a local property
- Any stationary random process is exactly shot noise (random pulses; the Wold Decomposition Theorem)
- Linearity is a physical property, not one of time series
- Do not bin data