Bayesian Blocks
for Particle Physics

Jeff Scargle
NASA Ames Research Center
Fermi Gamma Ray Space Telescope

Special Thanks to Jim Chiang, Jay Norris, Brad Jackson
Banff Discovery Workshop, July 2010
Data Segmentation for Particle Physics

Jeff Scargle
NASA Ames Research Center
Fermi Gamma Ray Space Telescope

Special Thanks to Jim Chiang, Jay Norris, Brad Jackson
Banff Discovery Workshop, July 2010
Two Constructs:

- Segmentation of Data Spaces into Data Cells & Blocks
- Edelson and Krolik algorithm:
  correlation functions of unevenly sampled data

lead to many informative functions:

- Data adaptive histograms and time series representations
- Correlation functions
- Fourier Power spectra (amplitude and phase)
- Structure Functions
- Wavelet Power spectra (scalegrams)
- Time-Scale/Time Frequency Distributions

... in auto- and cross- modes
... for all data modes (events, counts in bins, measurements, etc.)
... for arbitrarily sampled data
Two Constructs:

- Segmentation of Data Spaces into Data Cells & Blocks
- Edelson and Krolik algorithm: correlation functions of unevenly sampled data

lead to many informative functions

- Data adaptive histograms and time series representations
- Correlation functions
- Fourier Power spectra (amplitude and phase)
- Structure Functions
- Wavelet Power spectra (scalegrams)
- Time-Scale/Time Frequency Distributions

... in auto- and cross- modes
... for all data modes (events, counts in bins, measurements, etc.)
... for arbitrarily sampled data
Data Space – Any Dimension
Measurements (any kind) ➔ Data Cells
Collect Data Cells into Blocks
Partition of Data Space
The Bin Myths

I. Point data must be binned in order to make sense out of them.

II. Bins must be equal in size.

III. The bins must be large enough so that each bin has a “statistically significant” sample.
Smoothing and Binning

Old views: the best (only) way to reduce noise is to smooth the data
the best (only) way to deal with point data is to use bins

New philosophy: smoothing and binning should be avoided because they ...
- discard information
- degrade resolution
- introduce dependence on parameters:
  - degree of smoothing
  - bin size and location

Wavelet Denoising (Donoho, Johnstone) multiscale; no explicit smoothing
Adaptive Kernel Smoothing

Optimal Segmentation (e.g. Bayesian Blocks) Omni-scale -- uses neither explicit smoothing nor pre-defined binning
Bayesian Blocks

Piecewise-constant Model of Time Series Data

Optimum Partition of Interval, Maximizing Fitness Of Step Function Model

Segmentation of Interval into Blocks, Representing Data as Constant In the Blocks -- within Statistical Fluctuations

Histogram in Unequal Bins -- not Fixed A Priori but determined by Data


Simple Example of 1D Data Cells and Blocks

(a)  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32

(b)  block 1  block 2  block 3  block 4  block 5
Fitness Functions

- **Block likelihood** = product of likelihoods of its cells

  - **Block Likelihood** depends on:
    - $N =$ The Number of Events in the Block
    - $M =$ The Size of the Block

- **Model likelihood** = product of likelihoods of its blocks

  - Remove the dependence on the block event rates:
    - Marginalize, or
    - Maximize the Likelihood

  - Adopt prior distribution for $N_b$, the number of blocks. (Parameter of this distribution acts like a smoothing parameter.)

  - Take log to yield an additive fitness function.
The Optimiser

best = []; last = []; for R = 1:num_cells
    [ best(R), last(R) ] = max( [0 best] + fitness( cumsum( data_cells(1:R, :) ) ) );

    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

Do not use at home: a few details omitted!
Height = 1 / dt
n / dt
E / dt
Area = $\frac{1}{dt'}$

$n / dt'$

$E / dt'$

d$t' = dt \times \text{exposure}$
Planet Transits in HD 149026
Voronoi Tessellation of data in any dimension
Construct Voronoi cells to represent local photon density

density $\sim 1 / \text{cell area}$
Photons on the Sphere
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 & 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
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

Solar Ca II K Emission Index

Solar Ca II K Emission Index (9 tapers)

t (yr)

freq (c/yr)
Statistical Interlude

• Clinical studies usually small and expensive

• “Meta-analysis” – Increase significance by combining statistical summaries of published studies (not re-analysis of original data)

• Role of publication bias (PB)

• Assess potential for PB with Rosenthal formula
Statistical Interlude

• Publication bias is large!
• Editorial policy: Do not publish a study unless it achieves a 3-sigma positive result
• Rosenthal formula:
  ✓ Completely wrong!
  ✓ Used to justify hundreds of “meta-analytic” results in medicine, and psychology (real and para-)
  ✓ Not a single applied scientist questioned the validity of the formula

• Many medical studies, especially those relevant to decisions about safety of drugs to be released to the market, are based on this statistical blunder.
Statistical Interlude


Pre-election radio interview with the president of a major political polling organization ("Dr. Z").

Caller: “I hang up on polling phone calls – intrusion of my privacy.”

Discussion of this as a potential bias.

Dr. Z: “I don’t worry about such biases. We just get a larger sample.”

JS calls the radio show and tries to verify Dr. Z’s belief that increased sample size can fix a bias.

Dr. Z does not understand; responds by puffing up the reliability of his polling organization.
Voronoi Tessellations on 3+ Scales

Random space-time lattice (T. D. Lee)
- Points: micro-partons?
- Cells: Planck length cells
- Blocks: Elementary Particles

GLAST Source Detection Algorithm
- Points: Photons
- Blocks: Point sources

Cluster detection algorithm:
- Points: Galaxies
- Cells: Galaxy Neighborhoods
- Blocks: Clusters, filaments, ...

Large Scale Structure
- Points: Galaxies
- Cells: Voids

10⁻³⁵ meters
10⁺²² meters

10⁻³⁵ meters
10⁺²² meters