Analysis of
Astrophysics and Particle Physics Data
using Optimal Segmentation

Jeffrey.D.Scargle@nasa.gov

Space Science Division
NASA Ames Research Center

Santa Cruz Institute for Particle Physics
May 17, 2005
Outline

- Goal: Detect/Characterize Local Structures
- Data Cells
- Piecewise Constant Models
- Fitness Functions
- Optimization
- Error analysis
- Interpretation
- Extension to Higher Dimensions
The Main Goal is to Detect and Characterize Local Structures
From Data to Astronomical Goals

Data

Intermediate product
(estimate of signal, image, density …)

End goal
Estimate scientifically relevant quantities
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
Data: Measurements Distributed in a *Data Space*

**Independent variable** (data space)
  e.g. time, position, wavelength, ...

**Dependent variable**
  e.g. event locations, counts-in-bins, measurements, ...

**Examples:**
  time series, spectra
  images, photon maps
  redshift surveys
  higher dimensional data
DATA CELLS: Definition

*data space*: set of all allowed values of the independent variable

*data cell*: a data structure representing an individual measurement

For a segmented model, the cells must contain all information needed to compute the model *cost function*.

The data cells typically:

- are in one-to-one correspondence to the measurements
- partition the entire data space (no gaps or overlap)
- contain information on adjacency to other cells

… but any of these conditions may be violated.
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 2 3 4 5 | block 2 7 8 9 10 11 12 13 | block 3 14 15 16 17 | block 4 18 19 20 21 22 23 24 25 26 27 28 29 30 31 | 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!
Bootstrap Method:
Time Series of N Discrete Events

For many iterations:
- Randomly select N of the observed events with replacement
- Analyze this sample just as if it were real data

Compute mean and variance of the bootstrap samples

\[ \text{Bias} = \text{result for real data} - \text{bootstrap mean} \]
\[ \text{RMS error derived from bootstrap variance} \]

Caveat: The real data does not have the repeated events in bootstrap samples. I am not sure what effect this has.
Piecewise Constant Model
(partitions the data space)

Signal modeled as constant over each partition element (block).
Optimum Partitions in Higher Dimensions

- Blocks are collections of Voronoi cells (1D, 2D, ...)
- Relax condition that blocks be connected
- Cell location now irrelevant
- Order cells by volume

Theorem: Optimum partition consists of blocks that are connected in this ordering

- Now can use the 1D algorithm, O(N^2)
- Postprocessing: identify connected block fragments
Blocks

Block: a set of data cells

Two cases:
● Connected (can't break into distinct parts)
● Not constrained to be connected

Model = set of blocks

Fitness function:

F( Model ) = sum over blocks F( Block )
Connected vs. Arbitrary Blocks

Arbitrary Blocks Allowed

2 or fewer persons per square mile

3-8 persons per square mile

12-32 persons per square mile

44-284 persons per square mile

379 or more persons per square mile

An Optimal Partition of California Counties with respect to 1980 Population Density Persons/Square Mile Into Connected Blocks

Roughly 2 Persons Per Square Mile

Roughly 6 persons Per Square Mile

Roughly 12 Persons Per Square Mile

Roughly 21 Persons Per Square Mile

Roughly 61 Persons Per Square Mile

Roughly 771 Persons Per Square Mile

Roughly 1072 Persons Per Square Mile
Local Mean & Variance of Area/Energy (idea due to Bill Atwood)