Advanced Tools for Astronomical Time Series and Image Analysis

J.D. Scargle

"The unconscious goal of the scientific philosopher is the automation of science."
Irving John Good, *The Estimation of Probabilities*, 1965

"Automate or die."
Silicon Valley Billboard, June, 2001

ABSTRACT The algorithms described here, which I have developed for applications in X-ray and γ-ray astronomy, will hopefully be of use in other ways, perhaps aiding in the exploration of modern astronomy's data cornucopia. The goal is to describe principled approaches to some ubiquitous problems, such as detection and characterization of periodic and aperiodic signals, estimation of time delays between multiple time series, and source detection in noisy images with noisy backgrounds. The latter problem is related to detection of clusters in data spaces of various dimensions. A goal of this work is to achieve a unifying view of several related topics: signal detection and characterization, cluster identification, classification, density estimation, and multivariate regression. In addition to being useful for analysis of data from space-based and ground-based missions, these algorithms may be a basis for a future automatic science discovery facility, and in turn provide analysis tools for the Virtual Observatory. This chapter has ties to those by Larry Bretthorst, Tom Loredo, Alanna Connors, Fionn Murtagh, Jim Berger, David van Dyk, Adrian Raftery, Vincent Martinez, and Enn Saar.

1 Statistical Challenges in Modern Astronomy

One of the most important statistical challenges in science today is the effective analysis of data from NASA's observational astronomy programs. The work discussed here is meant to provide algorithms of general applicability in the framework of automated science analysis. It is hoped that they will be useful in addressing various challenges in astronomy – such as mining information from the Sloan Digital Sky Survey (see presentation by Michael Strauss) and other cosmological datasets (presentations by Vincent Martinez and Enn Saar, and A. H. Jaffe).

Automated processing already plays a large role in astronomical data analysis, and will be increasingly important as astronomy progresses into the Century of Data. How far along the path to the final scientific output can automatic processing be taken? I feel artificially intelligent data analysis will soon become surprisingly practical. See (Glymour *et al.* 1997, Glymour and Cooper 1999, Heckerman 1997, and Shalizi and Crutchfield 1999) for modern approaches to automatic analysis of data.

2 Periodic Signals

Definitive presentations of the modern approach to detection of a sinusoidal signal in the presence of noise appear in (Bretthorst, 1988, 2001). A key result that we will need for future reference is that the posterior probability density for the frequency ω of a single component is

$$P(\omega) \propto e^{\frac{C(\omega)}{\sigma^2}} \ , \tag{1.1}$$

[Bretthorst, 1988, Eq. (2.7)] where $C(\omega)$ is the ordinary *Schuster periodogram*, and σ is the variance of the noise, here assumed known. This equation shows that the periodogram is a sufficient statistic for this problem, and contains all information needed to compute frequency estimates and their uncertainties. (Bretthorst 2001) shows that the Lomb-Scargle periodogram serves the same role for unevenly spaced data.

The situation just described is an instructive case study in the relation between the frequentist approach employing a *statistic* and the Bayesian computation of a *posterior distribution*:

- As initially introduced, the periodogram is an *ad hoc* frequentist statistic. Since it is the inner product of a sinusoid and the data, it is reasonable that the periodogram will be large at frequencies at which a harmonic signal is present, small otherwise. But otherwise it is "pulled out of a hat" – an interesting quantity offered with minimal motivation, no justification[1] for preferring it over other possibilities, and only an indirect connection to detection probabilities.
- The Bayesian approach, so eloquently expounded in (Bretthorst, 1988), computes directly and straightforwardly the probability of sinusoidal

[1] Of course the periodogram's statistical behavior more or less validates its choice, after the fact. Indeed, the reason for constructing a modified periodogram for unevenly spaced data was to make its statistical behavior the same simple behavior shown by the Schuster periodogram for even spacing (Scargle 1982, 1989).

signal being present. It devolves that the resulting expression contains the periodogram, nicely clarifying its meaning – but this was by no means guaranteed.

Which of these two approaches is more satisfying is a matter of some debate.

3 Time Delays and Scaling

One often wants to determine the *lag* between two time series. That is, we picture the process generating the second time series as a delayed and possibly scaled version of that generating the first, and we wish to estimate the value of the delay. The approach here follows closely Bretthorst's, mentioned in §2. Only results are given here; see (Scargle 2001b) for details.

Assume that the underlying process is a signal, S, superimposed on a background, B. Take as given the two background rates, B_X and B_Y. We seek to characterize the signals that rise above these backgrounds. In some applications the backgrounds should be treated as unknown nuisance parameters, assigned a prior probability distribution, and then marginalized. In one case of special interest (gamma-ray bursts), the background levels are well determined by other data, and can properly be fixed at known constant values. Even here the ideal procedure is to represent this extrinsic data with a prior distribution for the background and marginalize it.

The complete model, expressing delay and scaling between the two signals, is:

$$X_m^{Model} = \quad S_m \quad + B_X \tag{1.2}$$

$$Y_m^{Model} = \quad aS_{m-\tau} \quad + B_Y \ , \tag{1.3}$$

where the lag is τ, and the Y-signal is an overall factor a times the X-signal.

For TTE data, m is measured is quantized units – here called time *ticks*, as defined by the electronics of the data acquisition system – and the above equations give the probability of a photon being detecting during tick m. The observed values, X_m, Y_m have values 1 or 0, depending on whether or not an event was recorded at tick m. After the usual procedure of writing down the likelihoods and marginalizing[2] the signal amplitudes, we find the posterior probability density for τ and a is

$$G_{total}(\tau, a) = G_0 \ e^{\frac{\gamma_{X,Y}(\tau)}{\Sigma^2}} \tag{1.4}$$

where

$$\gamma_{X,Y}(\tau) = \sum_{m=1}^{M} X_{m+\tau} Y_m \tag{1.5}$$

[2]That is, integrating out.

is the cross-correlation function of X and Y, and M is the length of the observation interval in ticks. This function arises naturally in the development, and is not introduced in an *ad hoc* manner. It can be readily and rapidly computed using the fast Fourier transform, representing X and Y as arrays of zeros punctuated by unit amplitude δ-functions at the values of m at which photons were detected. The coefficients g_0 and Σ (given in Scargle 2001b) depend on a and the backgrounds, but not on τ. The posterior for evenly binned count data (Scargle 2001b), at least for the case where the variances are independent of time, has exactly the same form.

Note that Eq. (1.4) has a clear similarity to the probability density for ω quoted above, Eq. (1.1) in §2. **The cross correlation function, γ is a sufficient statistic for lags, just as the periodogram is for frequencies.** The maximum likelihood value of the lag is just the value of τ that maximizes the cross-correlation function, so the main added feature is the ability to compute the full distributions of τ and a.

4 Signal Structure: Segmentation Yields Structure

Now turn to the problem of detecting and characterizing signal structure, from time series data. This section described a very practical representation of time-domain structures corrupted by observational noise[3], namely *partitioning of the data space into subsets in which the signal is assumed constant.*

4.1 Data

We consider data consisting of signal measurements, corrupted by noise, blurring, or other instrumental effects. These measurements may be in spaces of one dimension (*e.g.*, time series, energy spectra, *etc.*), two dimensions (images), or more (galaxy redshift/position catalogs).

I distinguish three types of measurement. The first is *event data*[4]. One measures positions of discrete points in the data space under consideration. Examples from the Compton Gamma Ray Observatory are time-tagged photon data from BATSE and sky-image data from EGRET, consisting of lists of photon positions, energies and times. While the usual coordinate representation of such points uses real numbers, in practice the corresponding infinite accuracy or resolution is not achievable. The coordinate is

[3] An important point, often leading to confusion, is that *noise* in astronomy has two quite distinct meanings: random observational errors, and random variability intrinsic to the source. The latter, part of the signal, is often just what one is studying, whereas observational noise is a corruption, to be eliminated as much as possible.

[4] This term is appropriate to the context of 1D time series; *point data* is used in the context of 2D images.

quantized in some small unit. In time series from high energy astrophyiscs, *e.g.*, the points are the times of detection of individual photons, and the corresponding quantum is the res olution of the spacecraft clock, typically somewhere in the range of microseconds to milliseconds.

In the second type of measurement, the entire observation interval (or area, or volume) is partitioned into pre-specified bins (or pixels, or cells), and one records the number of events in each. Event data can be converted to this mode, by adopting a set of bins and counting the points that fall in each bin. This process discards information, diminishes the resolution to that of the bins, and makes the results dependent on the sizes and locations of the bins.

The third type of sequential measurement does not involve explicit counting of events, but some other measurement of a quantity at a set of times or points in space. Here the statistical distribution of the observational errors is not tied to the Poisson distribution, as for the other two types, but can in principle be anything – most commonly normal (Gaussian). The values of the independent variable can be points, intervals, or defined by a spread-out sampling function. For example, spatial power spectra of cosmic microwave background measurements are typically reported in terms of window functions with various shapes; Bharat Ratra and Tarun Souradeep maintain a WWW site (`http://www.phys.ksu.edu/~tarun/CMBwindows/wincomb/wincomb_tf.html`) that gives details for many CMB experiments.

4.2 The Model

A key step in any likelihood analysis is definition of a model representing the underlying process (*i.e.*, the true signal) and the corruption process obscuring the true signal. We must compute the probability that the observed data would be obtained, given the model and its parameters. This function, called the *likelihood*, depends on the data mode, the sampling process, and the nature of the signal, the noise and other corruption processes.

A big advantage of point data is that they are efficiently described by a single, very simple model. The *Poisson process* is appropriate whenever the events are independent of each other. By this is meant that the occurrence of one event does not change the probability of any others. A common example of dependence is *dead time* in time series data: each photon is followed by an interval in which the detection of a second photon is inhibited. See (Stoyan, Kendall and Mecke 1995) for an excellent discussion of point processes in general, Poisson point processes in particular, and a number of ways that real world data can depart from being Poisson.

Independence implies that the probability an event will occur in any element of data space is proportional to the volume of that element. The proportionality coefficient is the local event rate, often called the *Poisson parameter* λ. It need not be constant, but can vary in an arbitrary way over the data space.

If λ varies randomly, the process is said to be a *Cox process*, or more descriptively a *doubly-stochastic process.* In such cases it is important to distinguish the two random processes at play. (The usual assumption is that these two are independent of each other.) Keeping this distinction clearly in mind, one can show that events occurring at two different locations are independent[5], even if the event rates at the two points are strongly correlated.

It is remarkable that the seemingly highly special Poisson model is in reality quite general – and surprisingly appropriate for most astronomical processes. All that is required is that the events are independent of each other, and their rate is described by an unrestricted function of position in the data space. Even dependences can be accounted for by incorporating them into the likelihood.

This function representing λ's dependence on location can be either parametric or nonparametric[6]. Since we do not want to impose an explicit signal shape, we use a nonparametric model, namely *piecewise constant* functions. This very convenient model class has the following properties:

- nonparametric
- general: capable of representing any reasonable signal
- simple, easy to compute: rate constant on finite intervals
- useful, *i.e.* easy computation of physically significant properties:
  - pulse peak times, widths, rise times, and decay times
  - pulse amplitudes
  - background level
- extendible to 2D and higher data spaces
- data adaptive, *i.e.* can respond to local features

This representation is also useful in domains such as classification, cluster detection, regression, and density estimation. One can think of it as implementing density estimation with blocks taking on the role of bins. Importantly, bin locations and sizes are determined by the data, through the condition that the blocks represent the statistically significant variations in the signal.

---

[5] *I.e.*, their joint probability is the product of the individual probabilities.

[6] Somewhat paradoxically, the number of parameters of a nonparametric model depends on the number of data points (Rissanen 1989). Examples are polynomial fitting, Fourier analysis, and wavelets. The basic idea is that one is really representing the structure in terms of elementary basis functions, whose number depends on how much information is present – rather than fitting a predefined shape to the signal.

Note that we don't really assert that the underlying physical process has a rate that changes in this blocky, discontinuous way. The true signal is no doubt relatively smooth. We represent it as piecewise constant in the same spirit as step-function approximations of a smooth curve. The idea is not that this representation is exact in some limit (often the justification for blocky models; *cf.* wavelet theory, especially the innovative ideas of Donoho 1994a,b), but simply that the blockiness reflects the statistical uncertainty of the data.

One could consider more accurate, *e.g.* piecewise *linear*, representations. But if continuity is imposed, the number of free parameters is almost the same as for piecewise constant models. For the most part the added accuracy is illusory and merely serves to complicate model interpretation.

Another issue has to do with what use the model will be put to. Often we are not really interested in the true shape itself, but in more generic information. For example, in the study of impulsive phenomena, such as Gamma ray bursts, one is interested in rise times, decay times, and other pulse properties. Since there are convenient ways to estimate these parameters directly from the blocks, our seemingly crude representation may adequately encode all usable shape information.

4.3 Algorithms

Three algorithms for implementing this Bayesian approach to modeling time series have been described elsewhere (Scargle 1998, 2001a), so only a brief sketch will be given here. The basic component of the model, called a *block* and denoted $\mathcal{B}_i$, comprises a time interval of length T_i and ascribes the N_i data points within this interval to a Poisson process with event rate λ_i. The posterior for this model is

$$P(\mathcal{B}_i) = \Phi(N_i, T_i) = \frac{\Gamma(N_i+1)\Gamma(T_i - N_i + 1)}{\Gamma(T_i+2)} = \frac{N_i!(T_i - N_i)!}{(T_i+1)!} \ . \tag{1.6}$$

Note that λ_i does not appear, since it has been marginalized. $P(\mathcal{B}_i)$ depends on only the size of the block and the number of data points in it. The posterior for the whole model is just $\prod_{i=1} P(\mathcal{B}_i)$, where i ranges over all elements of the partition.

Broadly, the three approaches are:

- **Divide and Conquer**: model comparison specifies the optimum *changepoint* at which to subdivide the interval; apply iteratively to all sub-intervals
- **Markov Chain Monte Carlo** (MCMC): sum the posterior probability by expeditiously exploring changepoint space
- **Cell Coalescence**: start from an ultra-fine representation assigning

one block to each datum; merge block pairs based on model comparison

The first and last can be thought of as *top-down* and *bottom-up* approaches, respectively. Consider two adjacent intervals, described by N_1, V_1 and N_2, V_2. The corresponding *Bayes merge factor* is computed using Eq. (1.6) to give the ratio of posteriors for the two regions merged and not merged, respectively:

$$\frac{P(\text{Merged})}{P(\text{Not Merged})} = \frac{\Phi(N_1 + N_2, V_1 + V_2)}{\Phi(N_1, V_1)\Phi(N_2, V_2)} \ . \tag{1.7}$$

In both cases one iterates until subdivision or merge operations no longer improve the posterior probability of the model. They are *greedy* algorithms, meaning that they choose the greatest gain possible at each step of the numerical optimization. This is sometimes called *myopic optimization* – a "take what you can now, with no regard for the future" strategy. On termination, the result may be a local optimum, perhaps a good approximate solution – but not guaranteed to be the global optimum. Cell Coalescence is easily generalizable to higher dimensions, as we will soon see.

MCMC (*e.g.*, Gilks, Richardson and Spiegelhalter 1996) is the most rigorous approach, as it solves for all changepoints simultaneously. Convergence of MCMC algorithms is a subtle issue.

5 High Dimensional Structure: Cluster Analysis, Classification

Cluster analysis in data spaces of higher dimension faces many vexing problems (Backer 1995, Gordon 1999), including determination of the number of clusters, a bewildering variety of proposed methods, loss of information due to restricted data modes, incorporation of prior information, nuisance parameters, and *post facto* validation of clusters. The Bayesian approach deals effectively with all of these issues. This section sketches an extension of the cell coalescence version of Bayesian Blocks to higher dimensions. The posterior in Eq. (1.6) applies unchanged in a space of any dimension, and the principles of the algorithm are identical to those in 1D.

Happily use of the *Voronoi tessellation (Okabe, Boots, Sugihara, and Chiu 2000)*[7] unravels the only real complication that arises in higher dimension – namely the geometry. The Voronoi tessellation partitions the data space into cells: cell i is that region of the data space closer to datum i than to any other datum.

[7] Due to their importance in computer graphics, fast algorithms yielding the unique Voronoi cell partition of a space of arbitrary dimension are readily available. MatLab (© The MathWorks, Inc.), *e.g.*, has one that represents the resulting data structures in a form very convenient for present purposes.

The Voronoi tessellation is an excellent representation of the data. It contains all relevant information in the raw data. It reduces the search space from the hugely infinite space of all possible partitions to the quite finite space of all possible Voronoi cell subsets which form a partition. It provides a simple generalization of the notion of adjacent intervals: blocks containing cells that touch at one or more points. And it even provides a crude but effective density estimation right off the bat, through the relation that the local point density is the reciprocal of the volume of the Voronoi cell.

The greedy cell coalescence algorithm collects Voronoi cells into larger and larger blocks by iteratively merging the pair of blocks with the largest *merge factor* from Eq. (1.7). In many applications it is both required and efficient to permit only blocks touching each other to merge. The iteration halts if the maximum merge factor falls below 1, at which point the data space has typically been partitioned into blocks much fewer in number than the original data points. Each block has a density equal to the number of data points in it divided by its volume. Then, if desired, high-density blocks adjacent to each other can be collected into *clusters*.

A slightly more detailed discussion of this work in progress is in (Scargle 2001c).

I am greatly indebted to Larry Bretthorst, Alanna Connors, Ayman Farahat, Karl Young, Tom Loredo, Jay Norris, Peter Cheeseman, and Peter Sturrock for comments and suggestions.

References

Backer, E. 1995, *Computer-assisted Reasoning in Cluster Analysis*, Prentice Hall, New York

Bretthorst, G. L. (1988), *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture Notes in Statistics, Springer-Verlag. Available (legally) by downloading from `http://bayes.wustl.edu/`.

Bretthorst, G. L. (2001), "Frequency Estimation And Generalized Lomb-Scargle Periodograms," in this volume, and other papers on his www site `http://bayes.wustl.edu/glb/bib.html`

Donoho, D.L., (1994a), "Smooth Wavelet Decompositions with Blocky Coefficient Kernels," in *Recent Advances in Wavelet Analysis*, L Schumaker and G. Webb, eds., Academic Press, pp. 259-308.

Donoho, D. L. (1994b), "On Minimum Entropy Segmentation," in *Wavelets: Theory, Algorithms, and Applications*, ed. Chui, C.K., Montefusco, L., and Pucciio, L., Academic Press: New York, pp. 233-269.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., eds., (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. 1997, "Statistical Themes and Lessons for Data Mining," in *Data Mining and Knowledge Discovery*, Vol. **1**, p. 11

Glymour, C., and Cooper, G. 1999, *Computation, Causation and Discovery,* MIT/AAAI.

Gordon, A. D.(1999), Classification, 2nd Edition, Monographs on Statistics and Applied Probability 82, Chapman & Hall/CRC, New York

Heckerman, D. 1997, "Bayesian Networks for Data Mining," in *Data Mining and Knowledge Discovery*, Vol. **1**, p. 79

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N., 2000, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd Edition, John Wiley & Sons: New York.

Rissanen, Jorma, 1989, *Stochastic Complexity and Statistical Inquiry*, Singapore: World Scientific.

Scargle, J. 1982, "Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data," 1982, *Astrophysical Journal*, **263**, pp. 835-853.

Scargle, J. 1989, "Studies in Astronomical Time Series Analysis. III. Fourier Transforms, Autocorrelation and Cross-correlation Functions of Unevenly Spaced Data," 1989, *Astrophysical Journal*, **343**, pp. 874-887.

Scargle, J. "Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, A New Method to Analyze Structure in Photon Counting Data," 1998, *Astrophysical Journal*, **504**, p. 405-418, September 1, 1998. `http://xxx.lanl.gov/abs/astro-ph/9711233`

Scargle, J. (2001a), "Bayesian Blocks: Divide and Conquer, MCMC, and Cell Coalescence Approaches," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering,* 19th International Workshop, Boise, Idaho, 2-5 August, 1999. Eds. Josh Rychert, Gary Erickson and Ray Smith, AIP Conference Proceedings, Vol. 567, p. 245-256.

Scargle, J. (2001b), "Bayesian Estimation of Time Series Lags and Structure," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

Scargle, J. D. (2001c), "Bayesian Blocks in Two or More Dimensions: Image Segmentation and Cluster Analysis," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

Shalizi, C. and Crutchfield, J. 1999, "Computational Mechanics: Pattern and Prediction, Structure and Simplicity," `http://www.santafe.edu/sfi/publications/Abstracts/99-07-044abs.html`

Stoyan, D., Kendall, W. S., and Mecke, J. (1995), *Stochastic Geometry and its Applications,* 2nd edition, John Wiley & Sons: New York