DC1 Scientific Data Analysis: Likelihood

Patrick Nolan
Stanford University

Scope

This presentation deals with the justification for basing our analysis on the likelihood tool. Broadly speaking, it is about math.

Jim Chiang will demonstrate the use of the tool.

Claudia Cecchi will speak about the Instrument Response Functions.

Seth Digel presented examples of scientifically interesting analyses of gamma-ray data.

What is Likelihood?

The likelihood, L, of a set of data is the probability of observing that data, given our belief about the physical processes that produced the photons that were detected. When that belief takes the form of a model with adjustable parameters, the likelihood can be expressed as a function of the parameters. The parameter values which produce the maximum value of L are useful estimators of the "true" values. Under fairly mild conditions, this process is unbiased and efficient.

The maximum likelihood value does not provide a test of "goodness of fit." The statistical significance of a point source, for instance, can be determined by the ratio of maximum likelihood values for models with and without the source.

Likelihood vs. χ^2

If the data are governed by Poisson (counting) statistics and the number of counts is large, then the familiar minimum-χ^2 estimator is an approximation to the exact likelihood. Up to an irrelevant additive constant,

$$\chi^2 \approx -2\ln(L)$$

Or when comparing two models,

$$\Delta\chi^2 \approx -2\ln(L_2 / L_1)$$

Wilks's Theorem guarantees that this likelihood ratio will be drawn from a χ^2 distribution if the total number of photons is large, even if they can't be put into bins with large numbers.

Why Likelihood?

- There aren't many photons. In an interesting part of the sky we will collect thousands, but the instrument response has many dimensions: time, angles, energy, and instrument-specific quantities. With sensible binning, most bins won't contain enough photons for χ^2 analysis to be valid.
- With the LAT's broad PSF, many sources will overlap. Care is required to distinguish nearby pairs.
- Direct image deconvolution is dangerous. Poisson noise is amplified to swamp the result. Some sort of regularization is needed, either by making assumptions about the statistical properties of the image (MaxEntropy) or by assuming a simple physical model with adjustable parameters. We choose the latter. Maybe we can use more advanced Bayesian methods some day.
- Pro: Gives quantitative results.
- Con: We won't discover anything we aren't looking for.

TS

A basic tool of the EGRET analysis is the "Test Statistic", or TS.

When two models are compared, $TS \equiv -2\ln(L_2 / L_1)$
Where L_1 and L_2 are the maximum likelihood values for the two models. As we have seen, TS has an asymptotically χ^2 distribution.

TS is applied when the difference between the two models is the presence of an extra point source in model #2. The statistical significance of the new source can be determined by treating TS as a χ^2 value with one degree of freedom, or sqrt(TS) as the gaussian "sigmas" of detection.

A TS map can be made by placing the putative source in many places, calculating TS in each place. This is a means of searching for unknown point sources.

What About XSpec?

Everyone's favorite point-source analysis tool for X-ray astronomy is XSpec. Why not use it?

It has many fine qualities, but it includes some assumptions about the data:

- Point sources are isolated from each other
- The background at each point source can be estimated from nearby areas of the sky. It must have no complex structure on the scale of the PSF.
- The types of IRFs it uses may be less complex than ours.

XSpec has no facility for detecting sources.

We may be able to use Xspec for some strong, isolated, high-latitude sources, but it can't be our workhorse.

The Math of Likelihood

We use Extended Maximum Likelihood (EML). This is the proper form to use when the number of photons is not determined before the observation. The quantity to be maximized is

$$\ln(L) = \sum_i \ln M(x_i) - N_{pred}$$

where x stands for the measured properties of the photons, *i* labels the individual photons, M(x) is the model rate of photon detection, and

$$N_{pred} = \int M(x)dx$$

is the total number of photon detections predicted by the model. We call the two terms the "data sum" and the "model integral."

The Use of Response Functions

For a photon of energy E arriving from direction $\mathbf{p}$ at time t, the probability density of a detection of type k with estimated energy E' and direction $\mathbf{p}'$ is

$$M_k(E', \mathbf{p}', t; E, \mathbf{p}) = h(t)\, H(\mathbf{u}(t) \cdot \mathbf{p}' - \mu_k(E'))$$
$$\times A_k(E, \theta(\mathbf{p}, t), \phi(\mathbf{p}, t))\, P_k(\mathbf{p}'; E, \theta(\mathbf{p}, t), \phi(\mathbf{p}, t))\, D_k(E'; E, \theta(\mathbf{p}, t), \phi(\mathbf{p}, t)).$$

Zenith cutoff

Effective area

Live fraction

Step function

PSF

Zenith direction

Energy dispersion

Angles in spacecraft coordinates

For a point source this must be integrated over the spectrum s(E)dE. If the spectrum is assumed to be steady, it can also be integrated over dt. Otherwise the model variation must be included in the integral.

The N_{pred} term is the integral of this function over all its arguments.

How to Evaluate the Likelihood

The photons are chosen from a Region of Interest (RoI) on the sky. However, some of them are produced by sources outside the RoI. The model must cover a larger area, the Source Region.

The adjustable parameters include the properties of all the point sources in the model (flux, spectrum, and perhaps position) as well as the flux and spectrum of the diffuse sources. In particular the Galactic diffuse flux can never be ignored and it must usually be treated with considerable care; most point sources are located in regions where it dominates the counting rate.

Most of the analysis time is spent evaluating the likelihood function and its derivatives. $M(x_i)$ requires evaluating the source spectrum and all three IRFs. N_{pred} requires integrating these over the whole observation, taking into account zenith cuts and changing angles.

Exposure

The calculation can be simplified by the use of "exposure". If the source spectrum is the only set of parameters that can be adjusted, most of the integral needs to be calculated only once. Exposure has dimensions area × time. It describes how deeply a spot on the sky has been examined. For diffuse sources, exposure must be calculated at many points.

The exposure calculation isn't simple. It looks like this for a particular energy E_m:

$$\mathcal{E}_m(\mathbf{p}) = \int dE' \int d\mathbf{p}' \int dt\, h(t) \sum_k H(\mathbf{u}(t)\cdot\mathbf{p}' - \mu_k(E')) \times A_k(E_m, \theta(\mathbf{p},t), \phi(\mathbf{p},t))\, P_k(\mathbf{p}'; E_m, \theta(\mathbf{p},t), \phi(\mathbf{p},t))\, D_k(E'; E_m, \theta(\mathbf{p},t), \phi(\mathbf{p},t))$$

The d**p**′ integral covers the ROI, but this must be evaluated for directions **p** within the larger source region.

Approximations

To save CPU time, approximations can be made.

- The energy dispersion can be treated as a delta function.
- If a point source is far from the edge of the Source Region, its PSF integrates to 1.
- The effects of the zenith cuts on N_{pred} can be ignored.
- Etc., etc.

The Starting Point

The model must have initial values for its parameters. This is easy for the cosmic and galactic diffuse components – use the EGRET numbers. The EGRET catalog can be used for the bright point sources. However, we expect to discover new point sources. For DC1 this will have to be done by visual inspection of flux maps or by time-consuming production of "TS maps". Later we will need a first-pass source finding tool that's fast and doesn't require subjective judgment. Accurate positions and fluxes are not needed.

There might be some hope of a single tool which will automatically find the sources and decide which ones are statistically significant. This is close to the cutting edge of statistical theory.

How to Fit Data

Finding the parameter values which maximize L is a classic nonlinear optimization problem. We have at our disposal a variety of well-tested, free packages, including Minuit. C++ wrappers have been written, which can be called from python. A Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-newton method gives the best results, so far. For this approach to work, L must be a smooth function and its derivatives must be calculated, which is the case for the models we use.

There is no need to bin the energy values to estimate the spectrum, as was done for EGRET. The spectrum is just a part of the model.

A typical run with a few tens of adjustable parameters and a few thousand photons consumes several minutes of CPU time. The time scales almost linearly with the number of photons. Scaling with the number of parameters seems to be worse than linear. The exposure calculation takes a comparable amount of time, but it doesn't scale with the number of photons.

The EM Algorithm

The EM algorithm is a different method for maximizing the likelihood function. Tests show that it can speed up the fitting process by a factor of ~3 in some cases. EM is used in many branches of science, and its implementation in medical imaging is rather similar to our needs.

EM is based on an underlying distribution which governs the observable data. In our case this is the assignment of photons to particular sources. Two steps alternate until convergence:

- Expectation (E): Based on current parameter estimates, find the expected value of the underlying variables.
- Maximization (M): Using these variables, find new optimum parameter values.

The virtue of EM for us is that the parameters of each source can be optimized separately rather than simultaneously. Lower-dimensional functions are optimized much more easily.