

## Fitting Cosmic Ray Spectrum Data

R. J. Wilkes

2/1/97

Summary: The comprehensive JACEE database on proton and helium events covering flights JACEE-1-12 is used as an example to describe application of the maximum likelihood method for fitting spectra. The method can be implemented as a simple EXCEL spreadsheet for databases of this size (<1000 events).

### 1. Primary cosmic ray spectra.

In the energy range well above the region of significant solar modulation ( $10^9$  eV) and below  $10^{15}$  eV, we expect the Galactic primary cosmic ray spectra for protons and nuclei to be of the general form

$$dN/dE d\Omega dt \sim E^{-\gamma},$$

with the differential spectral coefficient  $\gamma \sim 2.75$  for protons, and slightly smaller in magnitude for nuclei. Here, the differential spectrum is usually given as the number of particles observed per GeV, per  $\text{cm}^2$ -sr, per sec. The latter factors, representing the geometrical collecting efficiency of the detector and its exposure time, are assumed to be known (at least, on an event-by-event basis) and treated as constants. The database to be analyzed consists of a list of events, each characterized by an energy parameter, and part of a known exposure factor  $\Omega t$ .

As is well known, the spectral coefficient slowly increases (ie becomes more negative; the spectrum steepens) around  $10^{15}$  eV, the “knee” of the spectrum. This behavior is consistent with both observation and expected behavior from source and propagation models. The detailed behavior of the spectra around the knee is of special significance to theorists working on these models, hence wide interest in JACEE results.

Note that the power law form of the spectrum means that the spectrum is expected to follow a straight line of negative slope on a log-log plot of intensity  $dN/dE$  vs energy.

Also, if one integrates from some energy  $E_{\min}$  to infinity, one expects

$N(>E_{\min}) \sim E_{\min}^{\Gamma}$ , where  $\Gamma = \gamma - 1$ . This is the integral spectrum, the number of events above energy  $E_{\min}$  GeV, per  $\Omega t$ .

### 2. Methods Used to Fit Spectral Data

There are three ways commonly used to fit a slope to a set of data.

1. Least squares fit to binned data
2. Least squares fit to integral spectrum
3. Maximum likelihood fit to unbinned data

Method 1 is to bin the data by energy, represent each bin as a data point on a plot of intensity vs energy, and use least squares fitting techniques to find the best-fit slope. The error bars on each bin are assumed to be proportional to the square root of the number of

events in the bin; ie Poisson statistics are assumed applicable. The least squares method itself implicitly assumes that fluctuations in the data about the fitted line are Gaussian in character, which follows from the central limit theorem only for high-statistics experiments. Different results are obtained for different binnings of the data, and a variety of methods are used (eg, equal-population bins, equal  $\Delta(\log E)$  bins, etc). Errors on the fitted slope are usually reported straight from the fitting program's mouth, without proper evaluation such estimates' validity.

Method 2 recognizes the impropriety of making the fits dependent upon choices of energy bins, and is binning free. One simply plots the integral spectrum, expected to be a straight line of slope  $-\Gamma$  on a log-log scale, and fits to the slope. On the other hand, it is absolutely incorrect to apply any kind of weighted least squares method to the data in the integral spectrum, since each point is correlated strongly with its predecessor. Usually what is done is to treat each point as unweighted for purposes of the fit. But it is a rare experiment in which each individual event can be treated with equal weight, so this usually amounts to sweeping valuable information under the rug. The result is an incorrect slope and a statistically unsound estimate of its errors.

The maximum likelihood method treats each event independently, with its own weighting factor. Any subset of events can be used for fitting. Confidence limits are directly calculable and are not based on any assumptions regarding the character of the population distribution. There exist straightforward and accepted methods for weighting data and computing the effect of weighting on error estimates. For a good textbook discussion, see Ref. [1].

### 3. Maximum Likelihood Method.

Specifically applying the ML method to the problem of fitting cosmic ray power-law spectra, we hypothesize:

$$dN/dE = A E^{-\gamma},$$

and so the integral spectrum is

$$N(>E_{\min}) = \int_{E_{\min}} A E^{-\gamma} = A E_{\min}^{-(\gamma-1)/(\gamma-1)}.$$

If we observe  $N$  events over energy  $E_{\min}$  in an experiment, the constant

$$A = N (\gamma-1) E_{\min}^{(\gamma-1)}.$$

Therefore

$$dN/dE = N (\gamma-1) E_{\min}^{(\gamma-1)} E^{-\gamma}$$

and so, rearranging,

$$P(E|\gamma) = \{N (\gamma-1)/E_{\min}\} (E/E_{\min})^{-\gamma}$$

gives the properly normalized probability density function for observing an event with energy  $E$ , in an experiment in which  $N$  total events are seen over energy  $E_{\min}$ , and the spectrum has negative power law form with parameter  $\gamma$ . Therefore the joint probability of observing the ensemble of events actually recorded is

$$\mathcal{L}(\{E_i\}|\gamma) = \prod_i P(E_i|\gamma) = \prod_i \{N (\gamma-1)/E_{\min}\} (E_i/E_{\min})^{-\gamma}$$

assuming the spectral slope is  $\gamma$ ; ie,  $\mathcal{L}$  is a function of  $\gamma$ . Since products are hard to handle, we take the logarithm of both sides:

$$\ln \mathcal{L} = \sum \ln(P(E_i|\gamma)) = \sum_i \{ \ln (N / E_{\min}) + \ln (\gamma-1) - \gamma \ln (E_i/ E_{\min}) \}.$$

The best estimate for  $\gamma$  is the value which maximizes  $\ln \mathcal{L}$ , ie which makes  $d(\ln \mathcal{L})/d\gamma = 0$ .

Neglecting the first term in the sum, which is not a function of  $\gamma$ , we have

$$\ln \mathcal{L} \sim N \ln(\gamma-1) - \gamma \sum \ln (E_i/ E_{\min}).$$

One may of course use iterative methods to find the value of  $\gamma$  which minimizes the value of  $d(\ln \mathcal{L})/d\gamma$ . However, since in practice one knows the general range of the  $\gamma$  value required, and precision to two decimals is sufficient, it is quite simple to just tabulate  $\ln \mathcal{L}$  values as a function of  $\gamma$  and find the value which maximizes  $\ln \mathcal{L}$  directly.

Statistical theory tells us that the 68% confidence limits on  $\gamma$  (conventional 1- $\sigma$  error bars) are those values of  $\gamma$  for which the log-likelihood is 0.5 less than the maximum on either side. Confidence limits are obtained by intersecting the plot of  $\ln \mathcal{L}$  vs  $\gamma$  with horizontal lines  $\ln \mathcal{L} = \mathcal{L}_{\max} \exp(-a)$ , where  $a=n^2/2$  for  $n\sigma$  limits. Thus for 68% confidence limits, one finds the  $\gamma$  values at which  $\ln \mathcal{L}$  falls 0.5 units below its maximum, for 95% CLs, the points 2.0 units below maximum, and for 99.7% CLs, 4.5 units below maximum. If the underlying probability density distribution is Gaussian, the  $\ln \mathcal{L}$  plot will be an inverted parabola and the error bars will be symmetrical. If not, the  $\ln \mathcal{L}$  plot will in general be asymmetrical and the high and low error bars will differ in size. These limits are not strictly “confidence limits” but “likelihood limits” since they are upper bounds on the CLs, exact only for the Gaussian case (which is theoretically the minimum variance case). Further discussion takes us into the realm of “Bayesian prior knowledge” arguments which are unresolvable and thus more appropriate to philosophy than physics.

Note that this procedure for finding error bars assumes an absolute normalized scale for  $\ln \mathcal{L}$ , so it is essential that the likelihood function be a properly normalized probability density function (ie, integral over all possible values =1) in order to properly estimate error bars.

One can introduce a weighted fit by simply multiplying each event’s contribution to the  $\ln \mathcal{L}$  sum by a weight factor. It is of course necessary to keep track of the sum of all weights applied, so that the resulting  $\ln \mathcal{L}$  value can be rescaled to give a properly normalized value for error determination.

#### **4. Simple Implementation of ML method as an EXCEL Spreadsheet**

For the JACEE database, we need to extract the following information for each event:

1. Energy parameter: this should be  $\sum E_\gamma$ , but one can use  $E_0$  estimates if desired.
2. Geometrical and detection efficiency factors: these represent collection efficiency as a function of detector geometry, location of vertex, atmospheric overburden, etc, calculated by Watts, Roberts and others over the years for each specific detector design. The individual efficiency factors are simply multiplied together to give an overall efficiency factor, which functions as an inverse weight.

3. Exposure factor for the sample from which the event was derived: since we collect the highest energy events from all flights, but do not scan recent flights for low energy data, the lower energy events in the database represent a smaller exposure factor sample and must be appropriately weighted.

These data are put into a spreadsheet in which each line contains the relevant items for one event. The events are sorted by energy, highest first.

The efficiency factors and exposure factor for each event determine its weight:

$$1/W_i = \epsilon_1 \epsilon_2 \epsilon_3 \epsilon_4 \Omega_i / \Omega_{\text{tot}}$$

Here  $\Omega_{\text{tot}}$  represents the total exposure factor for the database being fitted. If all flights were scanned for all events, it would be the total exposure represented by the series of flights. However, in the present database, lower energy events come from scans covering smaller subsets of the exposure. Note that each event is thus assigned a weight in effect corresponding to the number of events of that energy that would be in the database if all efficiency and exposures were equal. The sum of the weights gives the effective number of events in the database, taking into account the weighting. This must be considered in ML fitting, unlike LSQ fitting, where only relative weights matter: confidence limits on the slope depend upon absolute, not relative, values of the calculated likelihood.

In the spreadsheet we can then put an appropriate range of  $\gamma$  values (eg, 1.50, 1.51...3.00) in a column and calculate the corresponding  $\ln \mathcal{L}$  value for each:

$$\ln \mathcal{L}(\gamma) = N \ln(\gamma-1) - \gamma \sum_i \ln(E_i / E_{\text{min}}).$$

If we apply a weight factor  $W_i$  to each event, we must simply scale the second term by the total of applied weights:

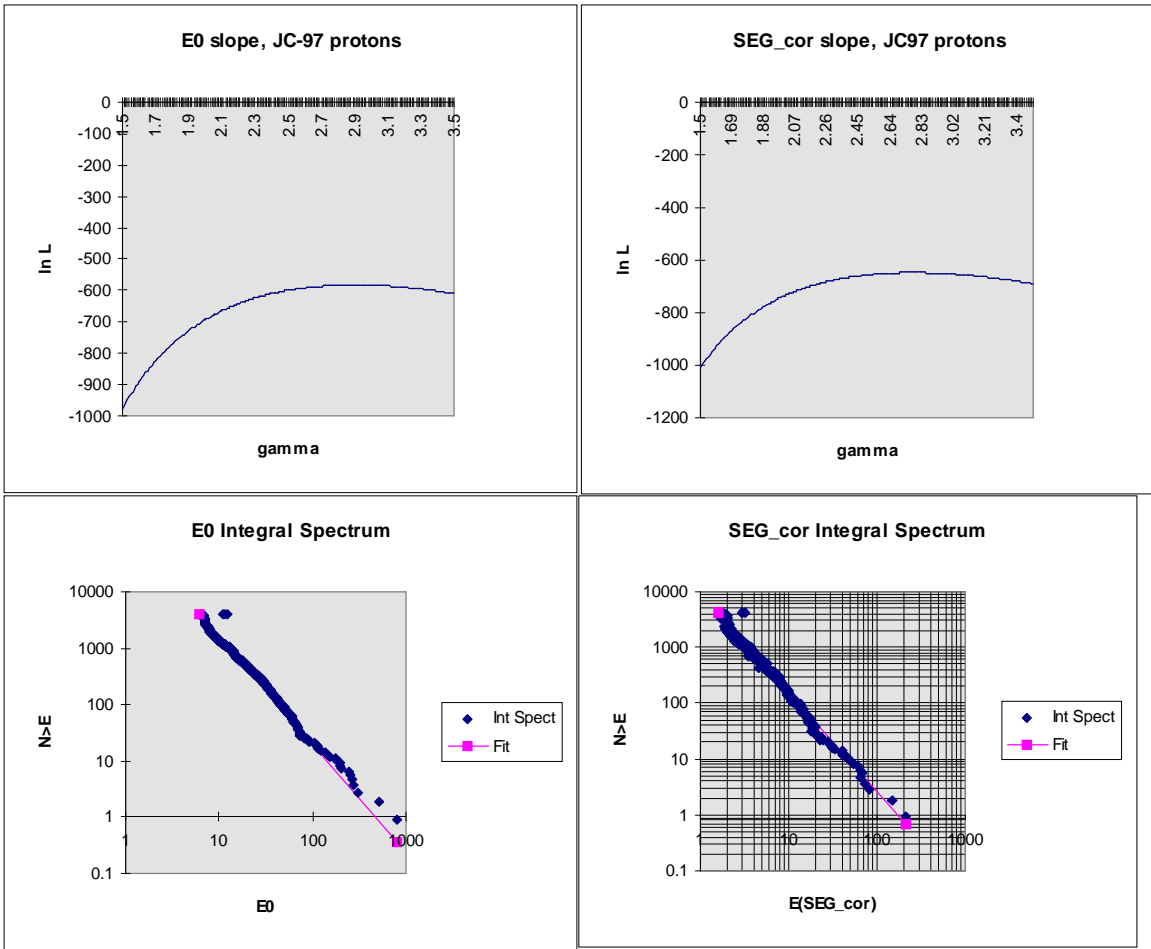
$$\ln \mathcal{L}(\gamma) = N \ln(\gamma-1) - \gamma (1/\sum W_i) \sum_i W_i \ln(E_i / E_{\text{min}}).$$

Attached are graphs showing the  $\ln \mathcal{L}$  plots and corresponding integral spectra for JACEE 1-12 proton and He databases as of January 1997. Also shown are portions of the relevant spreadsheets. A summary of the fit results is as follows:

| Spectrum | Energy used                         | Slope | Error |
|----------|-------------------------------------|-------|-------|
| proton   | $E_0$                               | 2.91  | 0.08  |
|          | $\Sigma E_\gamma(\text{corrected})$ | 2.79  | 0.07  |
| helium   | $E_0$                               | 2.81  | 0.09  |
|          | $\Sigma E_\gamma(\text{corrected})$ | 2.64  | 0.08  |

## 5. References

1. A. Frodesen, O. Skjeggstad, and H. Tofte, *Probability and Statistics in Particle Physics*, Universitetsforlaget, Bergen, 1979 (Columbia Univ. Press, New York, 1979).
2. W. Eadie, D. Drijard, F. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics*, North-Holland, Amsterdam, 1971.



Max Likelihood Spectral Slope

JC97 proton E0 spectrum

|        |             |      |                        |                      |          |            |
|--------|-------------|------|------------------------|----------------------|----------|------------|
| NBINS= | 200         |      |                        |                      | <b>E</b> | <b>Fit</b> |
| Emin=  | 6 N=        | 658  | $\ln E\_sum=$ 346.4046 | $1/Wt\ sum$ 4090.779 | 6        | 4090.779   |
| gamma= | 2.91 1-sig= | 0.08 |                        |                      | 792.2    | 0.364167   |

Max Likelihood Spectral Slope

JC97 proton SEG\_cor spectrum

|        |                 |      |                        |                      |          |            |
|--------|-----------------|------|------------------------|----------------------|----------|------------|
| NBINS= | 200             |      |                        |                      | <b>E</b> | <b>Fit</b> |
| Emin=  | 1.6 N=          | 658  | $\ln E\_sum=$ 369.5753 | $1/Wt\ sum$ 4090.779 | 1.6      | 4090.779   |
| gamma= | 2.79 +/- 1-sig= | 0.07 |                        |                      | 209.93   | 0.66171    |

