Statistical Methods for Investigating the Cosmic Ray Energy Spectrum

J. D. Hague1, B. R. Becker1, M. S. Gold1, J. A. J. Matthews1, J. Urbán1.
1University of New Mexico, Albuquerque, New Mexico USA
jhague@unm.edu

Abstract: Two separate statistical tests are described and developed in order to test unbinned data sets for adherence to the power-law form. The first test employs the TP-statistic, a function defined to deviate from zero when the sample deviates from the power-law form, regardless of the value of the power index. The second test employs a likelihood ratio test to reject a power-law background in favor of a model signal distribution with a cut-off. (ICRC07 Abs.I.D.=1217)

Introduction and Formalism

The question of whether the cosmic ray energy spectrum exhibits a cut-off at the very highest energies is of central interest to the cosmic ray (CR) physics[4, 10]. The flux of CR's at these energies is very small - about 3/km² steradian century - and, therefore, statistical analysis techniques which clearly quantify one's knowledge of flux suppression are useful. In this note we apply the statistics first developed for binned CR data sets in [6] to an un-binned analysis. We also introduce a new test based on a likelihood ratio test and show that both statistics can quantify our knowledge of a flux suppression.

We first establish the mathematical foundations of the analysis. The CR flux follows a power-law for over 10 orders of magnitude. The fundamental probability distribution function (p.d.f.) governing the power-law assumption (normalized such that \( \int_{x_{\text{min}}}^{\infty} f(x; x_{\text{min}}, \gamma) dx = 1 \)) is

\[
f(x; x_{\text{min}}, \gamma) = A x^{-\gamma},
\]

where \( A = (\gamma - 1)x_{\text{min}}^{\gamma-1} \) and the parameter \( \gamma \) is referred to as the spectral index.

The \( n^{th} \) raw moment of this distribution diverges[8] for \( n \geq 2 \) with \( \gamma \leq 3 \). Alternatively, the expected value of \( \ln(x/x_{\text{min}}) \) is better behaved and offers a crucial result of this analysis. Analytically we find,

\[
\nu_n \equiv \left\langle \ln^n \left( \frac{x}{x_{\text{min}}} \right) \right\rangle_X = \frac{n!}{(\gamma - 1)^n}.
\]

For a given sample we use,

\[
\hat{\nu}_n(X_{(j)}) \equiv \frac{1}{N - (j - 1)} \sum_{i=j}^{N} \ln^n \left( \frac{X_{(i)}}{X_{(j)}} \right).
\]

In eq.3 we denote the sorted (from least to greatest) data set as \( \{X_{(1)}, X_{(2)}, \ldots, X_{(N)}\} \).

To apply these statistics to an un-binned data set we calculate \( \hat{\nu}_n(X_{(j)}) \) for each minimum \( X_{(j)} \).

We also study a toy p.d.f. which is designed to mimic a power-law up to a certain energy but then exhibit a sharp "Fermi-Dirac like" cut-off above that energy[6]. We follow the parameterization used in [2],

\[
f_{FD}(x; x_c, w_c, \gamma) = \frac{B x^{-\gamma}}{1 + \exp \left( \frac{\log x - \log x_c}{w_c} \right)},
\]

where \( B \) is chosen such that \( f_{FD} \) is normalized over the interval \([x_{\text{min}}, \infty)\), i.e. \( \langle f_{FD} \rangle = 1 \).
Binned vs Un-binned Spectral-Index Estimators

Under the power-law assumption, we can take the log of both sides of eq.1 to yield \( \log f_x = \log ((\gamma - 1)/x_{\text{min}}) - \gamma \log (x/x_{\text{min}}) \). The slope, \( \hat{\gamma}_{\text{lb}} \), of the line which results in the minimum \( \chi^2 \) fit to the log-binned ("LB") histogram of a particular data. The un-binned spectral index can be found analytically[8]:

\[
\hat{\gamma}_{\text{lb}}(x_{(j)}) = 1 + 1/\hat{\nu}_1(X_{(j)}). \quad (5)
\]

This estimator is within 1% of the true \( \gamma \) for \( N \gtrsim 100 \) and it is asymptotically unbiased. The variance of this estimator is within 1% of the Cramer-Rao lower bound, given by \( \sigma_{\hat{\gamma}} \gtrsim (\gamma - 1)/\sqrt{N} \), for \( N \gtrsim 100 \). As derived in [5], we write the asymptotic p.d.f. of \( \hat{\gamma}_{\text{lb}} \) as \( f_{\nu_{1,2}}(\nu_1, \nu_2) \).

To illustrate the benefits of using un-binned estimators \( 10^5 \) Monte-Carlo trials were conducted. For each trail we draw \( N = 3500 \) events from a power-law with \( \gamma = 2.5 \) and calculate \( \hat{\gamma}_{\text{lb}} \) and \( \hat{\gamma}_{\text{ub}} \). These numbers are chosen to be approximately consistent with the flux reported[1] by the Auger Collaboration at ICRC 2005, as studied in [6]. In Figure 1 we plot histograms of these estimators and we note that the analytic prediction (\( f_{\nu} \)) is not a “fit” but represents a good approximation for the distribution of \( \hat{\gamma}_{\text{lb}} \). The mean (over the trials) of \( \hat{\gamma}_{\text{lb}} \) is 2.47 with deviation 0.034 while the corresponding values for \( \hat{\gamma}_{\text{ub}} \) are 2.50 and 0.025, verifying that \( \hat{\gamma}_{\text{ub}} \) has smaller error and less bias than \( \hat{\gamma}_{\text{lb}} \). This fact was first pointed out by Goldstein[3] and can be traced to the “heavy-tailed” nature of power-laws.

TP-statistic

We define the TP-statistic to be,

\[
\tau = \nu_1^2 - \nu_2^2/2 = 0 \quad (6)
\]

\[
\hat{\tau}(X_{(j)}) = \hat{\nu}_1^2(X_{(j)}) - \frac{1}{2} \hat{\nu}_2(X_{(j)}). \quad (7)
\]

The utility of using this statistic comes from the fact[9] that eq.6 is zero and thus, eq.7 will tend to zero as \( N \to \infty \), regardless of the value of \( \gamma \).

We may approximate the asymptotic joint distribution of \( \hat{\nu}_1 \) and \( \hat{\nu}_2 \) as a bivariate Gaussian \( f_{\nu_1,\nu_2}(\nu_1, \nu_2) \) with known means, variances and correlation coefficient[5]. Thus, for a given \( N \) and \( \gamma \), we calculate the p.d.f. of \( \tau \) to be,

\[
f_{\tau}(\tau; N, \gamma) = \int_{-\infty}^{\infty} f_{\nu_1,\nu_2}(t, 2(t^2 - \tau))dt. \quad (8)
\]

The analytic “location” \( \langle \tau \rangle_{\tau^p} \) and “shape” \( \langle \sigma_{\tau} \rangle_{\tau^p} = \sqrt{\langle \tau^2 \rangle_{\tau^p} - \langle \tau \rangle_{\tau^p}^2} \) parameters of this distribution are consistent with simulation generated values. Since the numeric integration required to calculate these quantities can be carried out faster than the requisite simulations we use the former to estimate the expected mean and variance of the power-law sample TP-statistic.

We estimate the significance of the TP-statistic for a given sample as

\[
(\hat{\tau} - \langle \tau \rangle_{\tau^p})/\langle \sigma_{\tau} \rangle_{\tau^p}. \quad (9)
\]

A spectrum with flux suppression in the tail (like that in eq.4) will result in a positive significance[6]. We note from [5] that \( \langle \sigma_{\tau} \rangle_{\tau^p} \sim N^{-1/2}(\gamma - 1)^{-2} \).

In Figure 2 we illustrate the behavior of this statistic when applied to a distribution with suppression in the tail. Using eq. 4 we analytically calculate \( \hat{\gamma}_{\text{ub}} = 1 + (\ln(x/x_{\text{min}}))_{\tau^p} \) (lower left) and \( \tau = (\ln(x/x_{\text{min}}))_{\tau^p}^2 -
The TP-statistic is sensitive to flux suppression for these toy distributions, see text for explanation.

0.5(ln²(x/x₁))ₓ tà (upper right) with γ = 2.5, log x₀ = 1.0, for three choices of w_c and as a function of x₂. We also calculate the expected value (and deviation) of these quantities when applied to a data set containing 3500 events, drawn from a pure power-law with values greater than 1.0. For each x₂ we estimate the number of events N with value greater than x₂ as 3500x₁⁻γ. The upper left panel shows the p.d.f.’s (on a log-log scale) normalized to unity on [1.0, ∞). The lower right contains the significance of the TP-statistic: for the lowest x₂ (i.e. N = 3500) the model cut-off distributions can reject the power-law assumption at the ∼4σ confidence level.

A Likelihood Ratio Test

Here we introduce a likelihood ratio test designed to discriminate a model signal (power-law with a cut) from a background (pure power-law) hypothesis and to be weakly dependent on γ. We may write the natural log of the ratio of the signal likelihood Lₓ tà = ∏ fₓ tà(xᵢ) to that of the background Lₓ = ∏ fₓ(xᵢ) as,

R(γ, log x_c, w_c) = N ln {C(γ, log x_c, w_c)}

- N ln N \sum_{i=1}^{N} \ln \left(1 + \exp \left(\frac{\log x - \log x_c}{w_c}\right)\right)

(10)

We note that C = B/A (see eqs. 4 and 1) contains the only dependence on γ and is independent of the data points under study, i.e. R contains no term involving log xᵢ⁻γ. Indeed, for any given log x_c and w_c, the quantity ln C is linearly dependent on γ with slope ∼ 0.125. In this sense the ratio test is weakly dependent on γ. However, in order to evaluate the efficiency of this test to reject a particular power-law background in favor of the cut-off signal we must choose γ a priori.

To illustrate how this test could be applied to a CR data set we generate 3500 “toy” events from fₓ tà with input parameters γ = 2.5, log x_c = 1 and w_c = 0.1 (see Figure 3). With the a priori choice of γ = 2.5, we then calculate R(2.5, log x_c, w_c) by scanning over the ranges 0.03 ≤ w_c ≤ 0.17 and 0.93 ≤ log x_c ≤ 1.07. The maximum ln Rₘₐₓ = 81.83 gives us the fit parameter estimates log x_c = 0.97 ± 0.04 and w_c = 0.10 ± 0.03, where the 68% confi-
Figure 3: The ratio test is sensitive to flux suppression for this MC set, see text for explanation.

By simulating $N_{bg} = 10^4$ sets of 3500 background events drawn from a pure power law (with $\gamma = 2.5$) and performing the same parameter scan over $\log x_c$ and $w_c$, we can estimate the efficiency $\beta$ of this test to reject the power-law in favor of the toy cut-off model, i.e. $\beta \sim N_{ln R \geq ln R_{max}}/N_{bg}$. From the right panel of Figure 3 we note that none of the $10^4$ background sets have $\ln R \geq \ln R_{max}$; we can reject the power-law in favor of the model cut-off at the $\sim 4\sigma$ confidence level.

When applying this test to a real CR data set $\gamma$ is not known a priori and one would want to estimate it. Studies of the ratio test with this extra degree of freedom are currently underway.

Conclusions

We began this note by verifying that the log-binned spectral index estimator has more bias and a larger error than the un-binned (maximum likelihood) estimator. We then detailed two un-binned statistical tests sensitive to flux suppression. We show that both tests show high sensitivity for rejecting the power-law hypothesis in favor of a toy flux suppression model and depend only weakly on the true spectral index. Applying these tests to 3500 events drawn from a toy cut-off distribution (see eq. 4) we can reject the power-law model in favor of the cut-off model at a confidence level $\sim 4$ standard deviations.

References

[1] Auger Collaboration. 29th ICRC. Pune, India, 10:115, 2005.
[2] Auger Collaboration. The UHECR spectrum measured at Auger and its astrophysical implications. ICRC-07 Abs.Num.318, 00:000–001, 2007.
[3] M.L. Goldstein, Morris S.A., and Yen G.G. Problems with fitting to the power-law distribution. Eur. Phys. J. B., 41:255–258, 2004.
[4] K. Greisen. End to the cosmic ray spectrum? Phys. Rev. Lett., 16:748–750, 1966.
[5] J.D. Hague et. al. Power Laws and the Cosmic Ray Energy Spectrum: Un-binned. In prep. for Astroparticle Physics.
[6] J.D. Hague et. al. Power Laws and the Cosmic Ray Energy Spectrum. Astroparticle Physics, 27:455–464, 2007.
[7] L. W. Howell. Statistical Properties of Maximum Likelihood Estimators of Power Law Spectra Information. NASA/TP-2002-212020/REV1, Marshall Space Flight Center, 2002.
[8] M. E. J. Newman. Power Laws, Pareto distributions and Zipf’s Law. Contemporary Physics, 46:323–351, 2005.
[9] V. Pisarenko, Sornette D., and Rodkin M. Deviations of the Distributions of Seismic Energies from the Gutenberg-Richter Law. Computational Seismology, 35:138–159, 2004.
[10] G.T. Zatsepin and Kuzmin V.A. Upper limit of the spectrum of cosmic rays. JETP Lett., 4:78–80, 1966.