ID 1217

30TH INTERNATIONAL COSMIC RAY CONFERENCE

Statistical Methods for Investigating the Cosmic Ray Energy Spectrum

J. D. HAGUE[1], B. R. BECKER[1], M. S. GOLD[1], J.A.J. MATTHEWS[1], J. URBÁŘ[1].
[1]*University of New Mexico, Albuquerque, New Mexico USA*
jhague@unm.edu

Abstract: Two separate statistical tests are described and developed in order to test un-binned 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.

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[1, 2]. The flux of CR's at these energies is very small - about $3/\text{km}^2$ steradian century - and, therefore, statistical analysis techniques which clearly quantify ones knowledge of flux suppression are useful. In this note we apply the statistics first developed for binned CR data sets in [3] 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 $\langle\rangle_X \equiv \int_{x_{min}}^{\infty} f_X(x; x_{min}, \gamma)dx = 1$) is

$$f_X(x; x_{min}, \gamma) = A\,x^{-\gamma}, \quad (1)$$

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

The nth raw moment of this distribution diverges[4] for $n \geq 2$ with $\gamma \leq 3$. Alternatively, the expected value of $\ln(x/x_{min})$ is better behaved and offers a crucial result of this analysis.

Analytically we find,

$$\nu_n \equiv \left\langle \ln^n\left(\frac{x}{x_{min}}\right)\right\rangle_X = \frac{n!}{(\gamma - 1)^n}. \quad (2)$$

For a given sample we use,

$$\hat{\nu}_n(X_{(j)}) \equiv \frac{1}{N - (j-1)} \sum_{i=j}^{N} \ln^n\left(\frac{X_{(i)}}{X_{(j)}}\right). \quad (3)$$

In eq.3 we denote the sorted (from least to greatest) data set as $\left\{X_{(1)}, X_{(2)}, \ldots, X_{(N)}\right\}$. 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[3]. We follow the parameterization used in [5],

$$f_{FD}(x; x_c, w_c, \gamma) = \frac{B\,x^{-\gamma}}{1 + \exp\left(\frac{\log x - \log x_c}{w_c}\right)}, \quad (4)$$

where B is chosen such that f_{FD} is normalized over the interval $[x_{min}, \infty)$, i.e. $\langle\rangle_{FD} = 1$.

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 $\log f_X = \log((\gamma -$
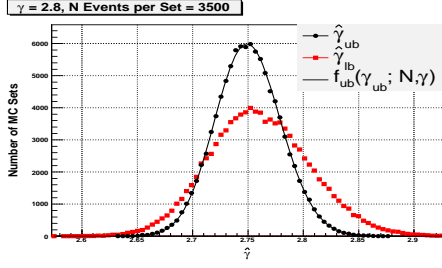
Figure 1: Estimates of the redlog-binned ($\hat{\gamma}_{lb}$) and the blackun-binned ($\hat{\gamma}_{ub}$). for 10^5 Monte-Carlo trials. For each trail we draw $N = 3500$ events from a power-law with $\gamma = 2.75$.

$1)/x_{min}) - \gamma \log(x/x_{min})$. The slope, $\hat{\gamma}_{lb}$, of the line which results in the minimum χ^2 fit to the logarithmically binned ("LB") histogram of a particular data. The un-binned maximum likelihood ("ub") estimate of the spectral index can be found analytically[4]:

$$\hat{\gamma}_{ub}(X_{(j)}) = 1 + 1/\hat{\nu}_1(X_{(j)}). \qquad (5)$$

This estimator is within 1% of the true γ for $N100$ and it is asymptotically unbiased. The variance of this estimator is within 1% of the Cramer-Rao lower bound, given by $\sigma_{\hat{\gamma}} \geq (\gamma - 1)/\sqrt{N}$, for[6] $N100$. As derived in [7], we write the asymptotic p.d.f. of $\hat{\gamma}_{ub}$ as $f_{ub}(\gamma_{ub}; N, \gamma)$.

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 $\gamma = 2.75$ ($x_{min} = 1$) and calculate $\hat{\gamma}_{lb}$ and $\hat{\gamma}_{ub}$. These numbers are chosen to be approximately consistent with the flux reported[8] by the Auger Collaboration at ICRC 2005, as studied in [3]. In Figure 1 we plot histograms of these estimators and we note that the analytic prediction (f_{ub} is not a "fit") represents a good approximation for the distribution of $\hat{\gamma}_{lb}$. The mean (over the trials) of $\hat{\gamma}_{lb}$ is 2.76 with deviation 0.045 while the corresponding values for $\hat{\gamma}_{ub}$ are 2.75 and 0.030, verifying that $\hat{\gamma}_{ub}$ has smaller error and less bias[9] than $\hat{\gamma}_{lb}$. Since we use

TP-statistic

We define the TP-statistic to be,

$$\tau = \nu_1^2 - \nu_2/2 = 0 \qquad (6)$$

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

The utility of using this statistic comes from the fact[10] that eq.6 is zero and thus, eq.7 will tend to zero as $N \to \infty$, regardless of the value of γ.

We may approximate the asymptotic joint distribution of $\hat{\nu}_1$ and $\hat{\nu}_2$ as a bivariate Gaussian $f_{V_1 V_2}(\nu_1, \nu_2)$ with known means, variances and correlation coefficient[7]. Thus, for a given N and γ, we calculate the p.d.f. of τ to be,

$$f_{TP}(\tau; N, \gamma) = \int_{-\infty}^{\infty} f_{V_1 V_2}(t, 2(t^2 - \tau))dt. \qquad (8)$$

The analytic "location" $\langle\tau\rangle_{TP}$ and "shape" $\langle\sigma_\tau\rangle_{TP} = \sqrt{\langle\tau^2\rangle_{TP} - \langle\tau\rangle_{TP}^2}$ 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

$$(\hat{\tau} - \langle\tau\rangle_{TP})/\langle\sigma_\tau\rangle_{TP}. \qquad (9)$$

A spectrum with flux suppression in the tail (like that in eq.4) will result in a *positive* significance[3]. We note from [7] that $\langle\sigma_\tau\rangle_{TP} \sim N^{-1/2}(\gamma - 1)^{-2}$. 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 $\hat{\gamma}_{ub} = 1 + \langle\ln(x/x_{min})\rangle_{FD}$ (lower left) and $\tau = \langle\ln(x/x_{min})\rangle_{FD}^2 - 0.5\langle\ln^2(x/x_{min})\rangle_{FD}$ (upper right) with $\gamma = 2.75$, $\log x_c = 1.0$, for three choices of w_c and as a function of x_{min}. 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 x_{min} we estimate the number of events N with value greater than x_{min} as $3500x_{min}^{1-\gamma}$. The upper left panel shows the p.d.f.'s (on a log-log scale) normalized to unity on $[1.0, \infty)$. The lower right contains the
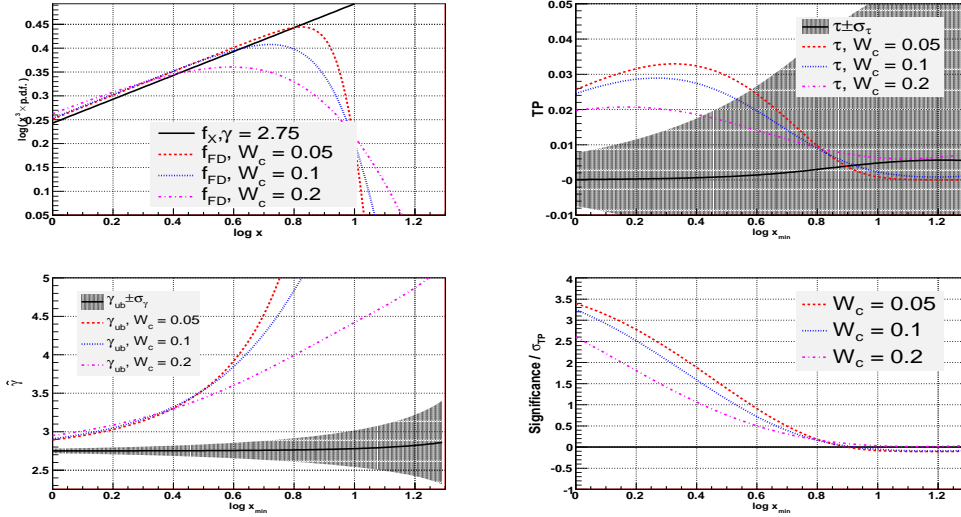
Figure 2: The TP-statistic is sensitive to flux suppression for these toy distributions, see text for explanation.

significance of the TP-statistic; for the lowest x_{min} (i.e. $N = 3500$) the model cut-off distributions can reject the power-law assumption at the $\sim 4\sigma$ 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 $L_{FD} = \prod f_{FD}(x_i)$ to that of the background $L_X = \prod f_X(x_i)$ as,

$$R(\gamma, \log x_c, w_c) = N \ln \{C(\gamma, \log x_c, w_c)\}$$
$$-\sum_{i=1}^{N} \ln \left\{1 + \exp \left(\frac{\log x - \log x_c}{w_c}\right)\right\}. \quad (10)$$

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 $\log x_i^{-\gamma}$. 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 f_{FD} with input parameters $\gamma = 2.75$, $\log x_c = 1$ and $w_c = 0.1$ (see Figure 3). With the a priori choice of $\gamma = 2.75$, we then calculate $R(2.75, \log x_c, w_c)$ by scanning over the ranges $0.03 \leq w_c \leq 0.17$ and $0.93 \leq \log x_c \leq 1.07$. The maximum $\ln R_{max} = 81.83$ gives us the fit parameter estimates $\log \hat{x}_c = 0.97 \pm 0.04$ and $\hat{w}_c = 0.10 \pm 0.03$, where the 68% confidence interval is approximated by the contour $\ln R_{max} - \ln R(2.75, \log x_c, w_c) = 2.30/2$.

By simulating $N_{bg} = 10^4$ sets of 3500 background events drawn from a *pure power law* (with $\gamma = 2.75$) and performing the same parameter scan over $\log x_c$ and w_c, we can estimate the efficiency β of this test to reject the power-law in favor 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 γ 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.
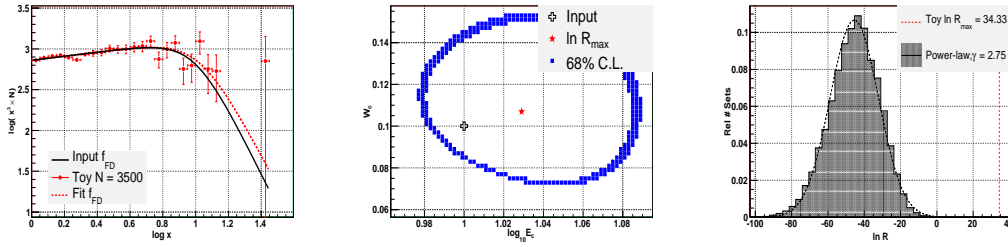
Figure 3: The ratio test is sensitive to flux suppression for this MC set, see text for explanation.

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 ~ 4 standard deviations.

References

[1] K. Greisen, End to the cosmic ray spectrum?, Phys. Rev. Lett. 16 (1966) 748–750.

[2] G. Zatsepin, K. V.A., Upper limit of the spectrum of cosmic rays, JETP Lett. 4 (1966) 78–80.

[3] J. Hague et. al., Power Laws and the Cosmic Ray Energy Spectrum, Astropartical Physics 27 (2007) 455–464.

[4] M. E. J. Newman, Power Laws, Pareto distributions and Zipf's Law, Contemporary Physics 46 (2005) 323–351.

[5] Auger Collaboration, The UHECR spectrum measured at Auger and its astrophysical implications., ICRC-07 Abs.Num.318 00 (2007) 000–001.

[6] L. W. Howell, Statistical Properties of Maximum Likelihood Estimators of Power Law Spectra Information, NASA/TP-2002-212020/REV1, Marshall Space Flight Center.

[7] J. Hague et. al., Power Laws and the Cosmic Ray Energy Spectrum: Un-binned, In prep. for Astropartical Physics.

[8] Auger Collaboration, 29th ICRC, Pune, India 10 (2005) 115.

[9] M. Goldstein, M. S.A., Y. G.G., Problems with fitting to the power-law distribution., Eur. Phys. J. B. 41 (2004) 255–258.

[10] V. Pisarenko, S. D., R. M., Deviations of the Distributions of Seiesmic Energies from the Gutenberg-Richter Law, Computational Seismology 35 (2004) 138–159.