ON THE STATISTICAL METHODS OF THE SIGNAL SIGNIFICANCE ESTIMATION IN THE DETECTION OF THE SIGNAL FROM THE POINT SOURCES OF COSMIC RAYS

A. CHILINGARIAN, G. GHARAGYOZYAN, G. HOVSEPYAN and G. KARAPETYAN

Cosmic Ray Division, Yerevan Physics Institute, Alikhanian Brothers 2, 375036, Yerevan, Armenia

Received (25 October 2004)
Revised (27 June 2005)

The estimation of the significance of the peaks in 1 and 2-dimensional distributions is one of the most important problems in high-energy physics and astrophysics. The physical inference from low-statistics experiments usually is biased and many discoveries lack further confirmation. We analyze the significance of the experimental evidence in the on-going efforts of detecting the point source of Cosmic Rays. The analysis is carried out in the framework of two models utilizing the extremum statistics: first – using the fixed grid of celestial coordinates, and second – using the tuned grid (introducing more degrees of freedom in the search). The test distributions for the significance estimation are obtained both from simulation models and from the analytical treatment of the problem.

Keywords: Statistical inference, extremely statistics, cosmic rays, point sources

1. Introduction

All sky searches of the Cosmic Ray sources is one of the most promising ways to gain insight in the long-standing problem of the origin of these particles. While many experiments have shown that the distribution of arrival directions are isotropic (see, for example, Ref. 1), existence of the small-angle anisotropies has been climbed by several groups in "knee" energy region $10^{14} - 10^{16}$ eV[3,9] and for the ultrahigh energy range $> 10^{19}$ eV.[11]

Physicists, observing positive fluctuation considerably greater than expected statistically usually ascribe this effect to a "source". However, as mentioned by Michael Hillas,[6] our experience has shown that large excesses up to 6σ are more common than were expected. When consistent and reliable statistical test are applied we can't obtain convincing prove for point sources. In 1973 there was demonstrated,[7] that the evidence of many of claimed γ-ray sources when properly treated is rather weak. Another striking illustration of the importance of assessing accurately the significance of peaks embedded in the low background is the "discovery" of so-called pentaquark: particle which contain four quarks and one antiquark. In 2003 physicists from many laboratories worldwide made headlines announcing that they had found new particle. They were above 10 detections, with very high confidence level of 5 and even 6σ. Unfortunately, new experiments with more statistics don't confirm existence of new particle. The "overwhelming body of negative evidence"[10] indicates that pentaquark might be an artifact.

In present publication we use both analytical approach and Monte-Carlo method to obtain the statistical model adequately describing signal searches with MAKET-ANI array.[2] The sources of possible erroneous physical inference based on biased models are discussed.

2. Gaussian approximation

In the experiment we measure the horizontal coordinates of the incident particle – zenith angle θ and azimuth ϕ, and then transform them to equatorial coordinates – right ascension RA and declination δ, according to transformation equations.[8] In our probabilistic treatment of the problem we convolute the uniform distribution of RA and treat number of events hitting different bins as realizations of the multinomial random process with $\bar{N}_j$, $j = 1, 20$ fixed probabilities. Then, by normalization of the bin contain according to the mean and variance of corresponding declination belt we obtain standard Gaussian distribution $N(0, 1)$ to be used further as the test statistics:

$$\sigma_{i,j} = \frac{N_{i,j} - \bar{N}_j}{\sqrt{\bar{N}_j}}, \quad i = 1, N_\alpha, 3, \quad j = N_{\delta 1, \delta 2, 3} \tag{1}$$

where $N_{i,j}$ is the number of events in the rectangular bins, $\bar{N}_j$ is the RA averaged mean of number of events fallen in the bin of j^{th} belt, $\sqrt{\bar{N}_j}$ is the j^{th} belt mean square deviation, $N_\alpha = 360$ is maximal value of RA; $N_{\delta 1} = 6.6$ is the first declination, $N_{\delta 2} = 66.6$ is the last declination for a total 20 declination belts, each of 3°.

More than 2 million particle showers detected by the MAKET-ANI detector, with sizes starting from $N_e > 10^5$ electrons were distributed among the $M = 2400$ angular bins. The signal was revealed when we examined the sub-sample of $\sim 6 \times 10^4$ events with $N_e > 10^6$.

From the obtained value of 6.04 for the "signal bin" we have calculated corresponding probability of obtaining this value under H_0 hypothesis. We assume that maximal obtained value for the signal bin 6.04 belongs to $N(0, 1)$ distribution. Therefore, the probability density distribution function of obtaining this value as the maximal value among M possibilities is straightaway:[4]

$$P_M(x) = M \cdot g(x) \left(1 - G_{>x}\right)^{M-1} \approx M \cdot g(x) \cdot e^{-M \cdot G_{>x}} \tag{2}$$

Where $g(x)$ is the standard Gaussian probability density for the signal bin; M is the number of bins; $G_{>x} = \int_x^\infty g(t)dt$ is the so-called test statistics p-value: the probability to obtain the value of test statistics in the interval greater than x.

To obtain the chance probability to observe in one from 2400 bins number of events equivalent or more than 6.04 standard deviations (p-value of the distribution $P_M(x)$) we have to integrate $P_M(x)$ within the interval $[6.04\sigma, +\infty)$. For $M = 2400$ we obtain

$$\int_{6.04}^{\infty} P_M(x)dx \sim 2 \times 10^{-6}.$$

Proceeding from this very small value, we rejected the null hypothesis and concluded that the MAKET-ANI has detected signal from the direction of the Monogem Ring.

3. Bin regrouping effect

The physicists usually slightly adjust the grid superimposed on the sky map, to include as much as possible of the "signal" events in selected bin as possible. Each regrouping in order to maximize the effect leads to increase of M in formula (2) thus changing the chance probability. To check and enumerate these changes we develop two simulation models.

Our first model generates the random Gaussian variables in 120 RA bins of each of 20 declination belts according to the belt-specific means and variances as obtained in the experiment. After applying normalizing transformation (1) to the generated random map we obtained $M = 2400$ random variables distributed according standard Gaussian $N(0,1)$. Then the maximal positive deviation from the $N(0,1)$ was stored as value of the test statistics. After more than 1000 independent random experiments with the model we obtained the distribution of the extremum statistics.

Our second model generates a number of events in the same way as the first one. Then the origin of equatorial coordinate system (right ascension and declination) is shifted by 0.1° in each dimension in the range equal to one bin size ($3^\circ \times 3^\circ$). Those instead of one grid 900 different grids are tested. The goal of this procedure is to obtain maximal possible "signal" for given distribution of events. After shift, the current value of extremum is compared with the previous best one, and if the new one is larger, it is saved as the best.

From the Fig. 1, a) we can see that our first model reproduce our results[3] very well. Note the excellent agreement of test statistics distribution obtained from Eq. (2) with $M = 2400$ and distribution obtained with Monte-Carlo method.

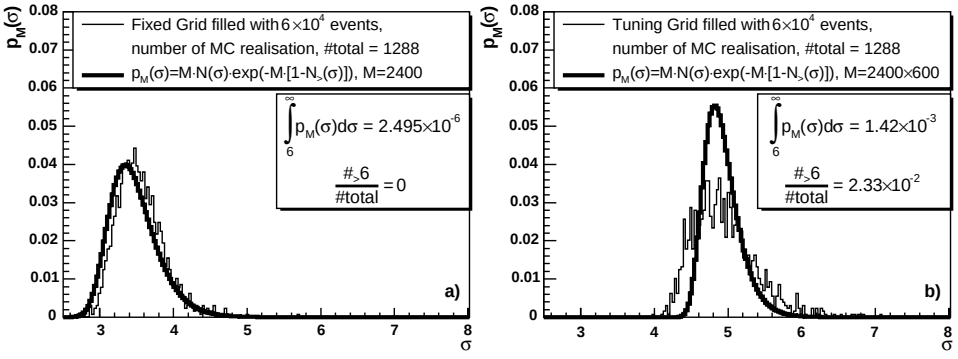

Fig. 1. The comparisons of the 2 ways of point signal searches: left – with one fixed grid, right – with tuned grid to maximize the extremum statistics.

When we are testing numerous grids, we enlarge M in Eq. (2), and the distribution mean is significantly shifted to the higher values, as seen from Fig. 1, b) and probability to obtain large extremum values is dramatically enlarged, especially for the low statistics experiments. For the MAKET-ANI statistics of $\sim 6 \times 10^4$ events with $N_e > 10^6$, the shift of the mean is ~ 1.5 and we can easily obtain significance values exceeding 6 and even 7. The 6σ value for the second model isn't exotic at all; its chance probability is 1 from

thousand (compare with 2 from million for fixed grid model), so we have not enough evidence to reject the H_0 hypothesis. Remember, that H_0 is the statement that the distribution of Cosmic Rays is isotropic.

4. Conclusions

(i) In estimating the significance of the signal detection we are looking for the maximal value of deviation from the normalized belt means, and statistical inference is drawn based on the value of this maximum. Therefore, as the test statistics for estimating the significance of signal, the extreme statistics distribution (2) should be used;
(ii) Both analytical distribution (2) and simulated distribution obtained with Monte Carlo method give very consistent results, proving necessity to account on all choices of data grouping aimed to reveal signal.

Acknowledgement

We thank Suren Chilingarian for writing code and performing simulations with the first and second models of point source detection, Artur Reimers for calculating significance levels and making Figures.

References

1. T. Antoni *et al*. (KASCADE Collaboration), *Astrophys. J.* **604**, 687 (2004).
2. V. V. Avakian, E. B. Bazarov, *et al.*, *Vopr. At. Nauki Tekh. Ser. Tekh. Fiz. Eksp.* **5**, 1 (1986).
3. A. Chilingarian, H. Martirosian and G. Gharagyozian, *Astrophys. J.* **597**, L129 (2003).
4. S. C. Chapman, G. Rowlands, *et al*., in *Nonlinear Processes in Geophysics*, **9**, 409 (2002).
5. D. Hearn *Nucl. Instrum. Meth.* **70**, 200 (1969).
6. M. Hillas, in *Proc. 14th Int. Cosmic Ray Conf., München* **9**, 3439 (1975).
7. E. O'Mongain, Nature**241**, 376 (1973).
8. J. Meeus, *Astronomical Algorithms* (Willmann-Bell Inc., 1991).
9. M. Samorski and W. Stamm, *Astrophys. J.* **268**, L17 (1983).
10. C. Seife, *Science*, **306**, 1281 (2004).
11. M. Takeda *et al*., *Astrophys. J.* **522**, 225 (1999).