Sequential Analysis Techniques for Correlation Studies in Particle Astronomy

Segev Y. BenZvi*, Brian M. Connolly†, and Stefan Westerhoff*

*University of Wisconsin-Madison, Department of Physics, Madison, WI 53706, USA
†University of Pennsylvania, Department of Physics and Astronomy, Philadelphia, PA 19104, USA

***Abstract*. A common analysis in ultrahigh energy cosmic ray physics is the search for statistically significant correlations between cosmic ray arrival directions and classes of astrophysical objects. We present a method to test a potential correlation signal sequentially, i.e., after each incoming new event in a running experiment. The sequential method can be applied to data taken after the test has concluded, adheres to the likelihood principle and rigorously accounts for our ignorance of the signal strength.**

***Keywords*: cosmic rays, correlation with catalogs, sequential analysis**

I. Motivation

One of the major goals in astroparticle physics is the identification and the study of sources of ultra-high energy cosmic rays, defined as cosmic rays with energies larger than 10^{18} eV. The discovery of discrete sources would answer longstanding questions about how and where particles are accelerated to such energies. So far, no discrete sources have been positively identified. A major obstacle for the identification of potential sources is the small number of detected events. Until a few years ago, the published world data set of cosmic rays with energies above 4×10^{19} eV consisted of little more than 100 events, mainly recorded with the Akeno Giant Air Shower Array (AGASA) in Japan between 1984 and 2003 [1], and the High Resolution Fly's Eye (HiRes) Experiment in Utah between 1997 and 2006 [2]. With a new generation of large-aperture astroparticle physics detectors like the Pierre Auger Observatory in Malargüe, Argentina and the Telescope Array detector in Utah, the amount of ultra-high energy data is now growing at an unprecedented pace.

The fact that previous experiments have failed to find statistically significant deviations from isotropy in skymaps of ultra-high energy cosmic rays can be seen as an indication that the sources are weak. In this case, the most promising correlation searches are not those which aim at finding sources individually, but rather those conducted on a statistical basis; i.e., searches for significant correlations of cosmic ray arrival directions with catalogs of astrophysical objects.

When studying correlations with objects from a source catalog, one tests whether the probability p of a given event to arrive from the direction of an object in the catalog is significantly larger than the probability p_0 of the correlation occurring by chance. These analyses typically have a number of parameters that are not known a priori, for example the angular scale of the correlation, the energy above which a correlation signal is expected, or the catalog of astrophysical objects itself.

Typically, potential signals are identified after intensive searches using different angular scales, different energy thresholds, different source catalogs, and other parameters that are found to maximize the signal strength. Therefore, an unbiased chance probability for the observed signal can only be established by discarding the data set used to find the signal and testing the signal with statistically independent data. For the test, the source catalog and all analysis parameters are fixed a priori to obtain an unbiased chance probability for the signal.

Once the a priori analysis parameters are identified, the problem is easily formulated in terms of a classical hypothesis test, in which new data are checked for compatibility with a null hypothesis $\mathcal{H}_0$ ("the data exhibit no significant correlation") or an alternative "signal" hypothesis $\mathcal{H}_1$. There are several ways to perform such a test. For example, one can run the test after the new data set has reached a certain size n, or after the experiment has run for a certain fixed amount of time.

However, it is often desirable to evaluate and test the signal sequentially, i.e., after each new event, rather than at the end of the test. This approach allows for the possibility of claiming a statistically significant result earlier than with methods that check the signal only once, a distinct advantage when event rates are quite low. It also avoids another practical disadvantage of hypothesis tests that arises when the experiment, for one reason or another, has to discontinue data taking before the predefined number of events is taken. In that case, the "one-shot" analysis does not lead to a conclusion.

The size of the data set and the acceptance or rejection of the null hypothesis are determined by two probabilities, α and β, which are usually chosen before the start of the test: α is the probability of wrongly rejecting the null hypothesis when $\mathcal{H}_0$ is true (type-1 error); and β is the probability of wrongly accepting the null hypothesis when $\mathcal{H}_0$ is false (type-2 error). A p-value P is used to estimate the agreement of the data with the null hypothesis.

A sequential analysis can be performed in several ways. If P is evaluated after every incoming event and not just once after all n events are collected, a "penalty" factor has to be inserted to account for the fact that

there are now more opportunities to satisfy the test by chance [3], [4]. This penalty factor can be evaluated with simulations and will depend on n. The dependence of P on n is an undesirable feature of the method; rather than depending on the data that were actually recorded, P now depends on the number of events that an observer would have recorded had he decided to perform a "one-shot" test. The interpretation of the data therefore depends on data not actually taken. This feature of the test violates the likelihood principle [5].

In addition, the inclusion of the penalty factor means that data arriving after the test has ended cannot be used to calculate P for the entire data set. It is therefore not possible to include new data in the calculation of the probability. In many practical situations, data taking continues after the test has ended, and it is highly desirable to monitor the signal probability with new data.

The classical sequential likelihood ratio test developed by Wald [6], [7] avoids the limitations that arise when using the p-value P. Wald defines the likelihood ratio as

$$\mathcal{R} = \frac{P(\mathcal{D}|\mathcal{H}_1)}{P(\mathcal{D}|\mathcal{H}_0)} \quad , \tag{1}$$

where the denominator and numerator represent the probability of observing a data set $\mathcal{D}$ given a null hypothesis (no correlation) and an alternative (correlation). The ratio $\mathcal{R}$ can be evaluated after each incoming event without statistical penalty, and the test stops with the acceptance or rejection of the null hypothesis when $\mathcal{R}$ falls below or exceeds a predefined value. In addition, the evaluation of $\mathcal{R}$ can continue after the decision to see whether new data continue to favor or disfavor the selected hypothesis.

The probabilities $P(\mathcal{D}|\mathcal{H}_0)$ and $P(\mathcal{D}|\mathcal{H}_1)$ in eq. 1 depend on the expected correlations in case of random coincidences and true signals, respectively. In correlation studies, the strength of the signal is typically not known before the test is complete; so in the analysis proposed by Wald [6], [7], one simply takes a "best guess" at the lower bound of the signal strength. Here, we extend Wald's technique to marginalize the signal strength, which more rigorously accounts for our ignorance of the true signal.

II. The Method

We consider the case of an analysis searching for correlations between cosmic ray arrival directions and objects from a catalog. The background probability p_0 is the probability that a given event correlates by chance. We want to test the signal probability p_1 against p_0. If two point hypotheses are tested against each other, p_0 and p_1 are single numbers; but in general, p_1 can also have a range of values. If, for example, the "signal" corresponds to a stronger correlation than can be expected by chance, then $p_1 > p_0$.

Since an event can either be correlated with an object from the catalog or not, the probability of observing a data set $\mathcal{D}$ in which k out of n events correlate with sources is given by the binomial distribution

$$P(\mathcal{D}|p) = P(n,k|p) = \binom{n}{k}\ p^k\ (1-p)^{n-k} \tag{2}$$

where p is the probability of a given event to correlate. If the data show no significant correlations in addition to those occurring by chance, then $p = p_0$.

In a sequential analysis that tests hypothesis $\mathcal{H}_1$ against $\mathcal{H}_0$ with data $\mathcal{D}$, the probability ratio $\mathcal{R}_n$ of eq. 1 is calculated after each incoming event, and is then compared to two positive constants A and B (where $B < A$). During each step in the sequence, the experimenter is presented with the following possible outcomes:

1) $\mathcal{R}_n > A$: the test terminates with the rejection of $\mathcal{H}_0$.
2) $\mathcal{R}_n < B$: the test terminates with the acceptance of $\mathcal{H}_0$.
3) $B < \mathcal{R}_n < A$: the test continues to record data.

Wald [6], [7] showed that the constants A and B are closely related to the probabilities α and β of type-1 and type-2 errors:

$$A \leq \frac{1-\beta}{\alpha} \quad \text{and} \quad B \geq \frac{\beta}{1-\alpha} \quad . \tag{3}$$

While it is difficult in most practical situations to estimate exact values for A and B, Wald showed that simply choosing

$$A = \frac{1-\beta}{\alpha} \quad \text{and} \quad B = \frac{\beta}{1-\alpha} \quad , \tag{4}$$

as the test boundaries leads to adequate results if α and β are small (typically, they are not larger than 0.05). By adequate, we mean that the true type-1 and type-2 rates will never exceed α and β. In fact, the true error rates will often be smaller than the nominal α and β specified before the start of the experiment.

For a data set that contains n events and k correlations, the likelihood ratio is given by

$$\mathcal{R}'_n = \frac{P(\mathcal{D}|p_1)}{P(\mathcal{D}|p_0)} = \frac{p_1^k(1-p_1)^{n-k}}{p_0^k(1-p_0)^{n-k}} \quad . \tag{5}$$

In practice, the signal strength p_1 is often not known. We consider here the common case of a one-sided test where $p_0 < p_1 \leq 1$. The confidence in rejecting $\mathcal{H}_0$ typically increases with increasing p. To evaluate $\mathcal{R}_n$ in this case, we can expand the numerator and denominator of eq. 1 in terms of p:

$$\mathcal{R}_n = \frac{\int_0^1 P(D|p)\ P(p|\mathcal{H}_1)\ dp}{\int_0^1 P(D|p)\ P(p|\mathcal{H}_0)\ dp} \ . \tag{6}$$

The quantities $P(p|\mathcal{H}_1)$ and $P(p|\mathcal{H}_0)$ represent our prior assumptions about p in the cases of true signal vs. chance correlations. In cosmic ray studies, the probability p_0 of a chance correlation with a catalog object is estimated from the a priori parameters of the test: e.g., the detector exposure to the catalog, the angular

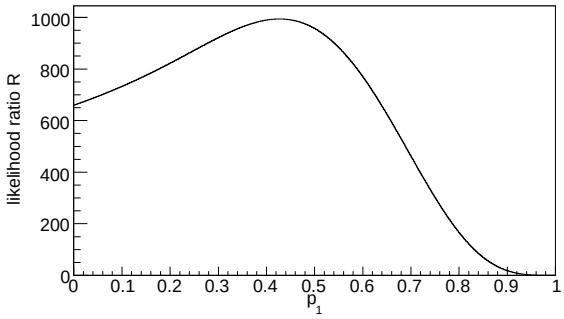

Fig. 1: Likelihood ratio as a function of p_1 for $n = 10$, $k = 6$, and $p_0 = 0.1$.

bin size θ, etc. In contrast, it is fairly uncommon to have a reliable estimate of the signal probability p_1 beyond the fact that $p_1 > p_0$. Absent further knowledge of the signal, we can therefore treat the probability as uniformly distributed on the interval $[p_1, 1]$. Hence, we summarize our prior knowledge of the two cases by

$$P(p|\mathcal{H}_1) = \frac{\Theta(p - p_1)}{1 - p_1} \quad , \tag{7}$$

$$P(p|\mathcal{H}_0) = \delta(p - p_0) \quad . \tag{8}$$

Note that p is not time-dependent, although we do not see anything inherently problematic in inserting a time-dependence.

Solving for the likelihood ratio $\mathcal{R}_n$, we have

$$\mathcal{R}_n = \frac{\int_{p_1}^{1} p^k \ (1-p)^{n-k} \ dp}{p_0^k \ (1-p_0)^{n-k} \ (1-p_1)} \tag{9}$$

$$= \frac{\mathrm{B_c} - \mathrm{B_{ic}}}{p_0^k \ (1-p_0)^{n-k} \ (1-p_1)} \tag{10}$$

where

$$\mathrm{B_c} = \mathrm{B}(k+1, n-k+1) \tag{11}$$

and

$$\mathrm{B_{ic}} = \mathrm{B}(p_1; k+1, n-k+1) \tag{12}$$

where $\mathrm{B}(a, b)$ and $\mathrm{B}(x; a, b)$ are the complete and incomplete beta functions.

When nothing is known *a priori* about the strength of the signal, p_1 will be chosen close to p_0 to test as large a signal space p as possible. In practice, one would choose $p_1 = p_0 + \delta$, where δ is a positive constant. The particular choice of δ is somewhat ad hoc, since it mainly reflects the experimenter's degree of belief about the strength of the signal. If more information on p were available — for example, if it were known that p is larger than some value p_{min} — then the range of integration could be made smaller. To illustrate the merits of improved knowledge, fig. 1 shows $\mathcal{R}_n$ as a function of p_1 for $n = 10$, $k = 6$, and $p_0 = 0.1$. Since the "true" probability for an event to correlate is $p = 6/10 = 0.6$, choosing p_1 close to p increases $\mathcal{R}_n$ and therefore minimizes the time necessary to confirm the signal. As p_1 continues to increase beyond the true signal probability, $\mathcal{R}_n$ decreases, as expected.

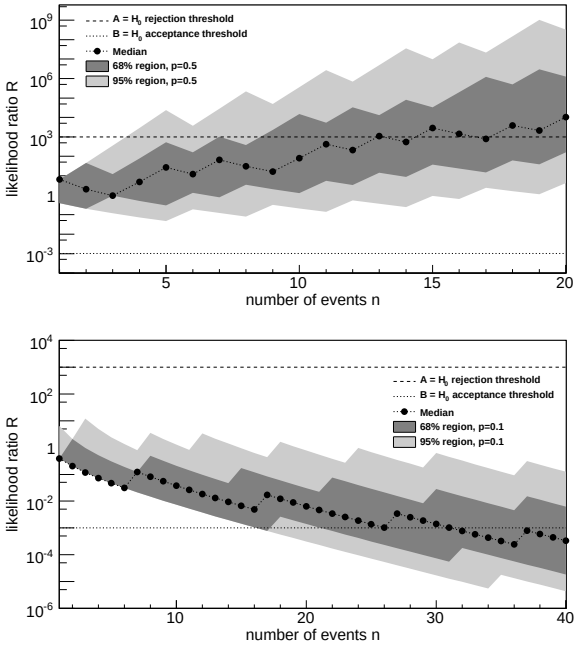

Fig. 2: Likelihood ratio as a function of the number of events for a background probability $p_0 = 0.1$, $p_1 = 0.3$, and a signal probability $p = 0.5$ (top) and $p = 0.1$ (bottom). The ratio is calculated for 10^5 random data sets. The plots show the median (dark dots) together with the range that includes 68 % and 95 % of the data sets (dark and light areas). The values for the test boundaries A and B for $\alpha = \beta = 0.001$ are indicated as dashed and dotted lines.

Fig. 2 shows the results of the sequential analysis described above when applied to simulated data sets. The background probability is $p_0 = 0.1$; $p_1 = 0.3$ is the minimum signal we choose to distinguish from the background; and $\alpha = \beta = 0.001$. The upper plot shows the result of the test for data sets with a correlation probability of $p = 0.5$ ($\mathcal{H}_0$ is false), whereas for the bottom plot, $p = 0.1$ ($\mathcal{H}_0$ is true). For both plots, the analysis is performed for 10^5 Monte Carlo data sets, and the dark and light grey areas indicate the range that includes 68% and 95% of the data sets.

III. Discussion

Our approach of accounting for our ignorance of the true correlation probability p of the given data set by marginalizing p in the likelihoods has an important consequence if one were to interpret the results of the hypothesis test in terms of the probabilities α and β, for example by using $(1 - \alpha)$ as a confidence level for the rejection of the null hypothesis. Since the numerator now allows for $p_1 < p < 1$, α and β have, strictly speaking, only meaning for a data set that has similar properties, i.e. has a correlation probability that is not a single value,

but spread over the interval $[p_1, 1]$. Consequently, the parameters α and β have lost their intuitive meaning if the method is applied to data sets where p is fixed, as is typically the case for real data. However, it can be shown [8] that for most values of p and δ that occur in correlation searches, the type-1 and type-2 error rates of the sequential analysis are consistent with the classical interpretations of the probabilities α and β.

Note that we have run a test with one of two outcomes (i.e., an acceptance or rejection of $\mathcal{H}_0$), defining α and β, rather than one outcome (say, only a rejection of $\mathcal{H}_0$) such as in [9]. The latter case supposes that we are only concerned about reporting a signal. However, we would assert that it is important to state a null result at some point in the interest of reducing reporting bias. That is, it is important to ensure that 1% of the results that claim an excess of events are indeed a 1% effect.

In summary, the sequential analysis technique proposed here is efficient, allows the signal significance to be evaluated after the test has been fulfilled, adheres to the likelihood principle, and rigorously accounts for our ignorance of the signal strength.

The method is applied in the search for correlations between ultra-high energy cosmic rays recorded with the Pierre Auger Observatory and locations of nearly active galactic nuclei [10]. An update on this search appears in these proceedings [11].

This work is supported by the National Science Foundation under contract number NSF-PHY-0636875.

REFERENCES

[1] M. Takeda et al. 1999, Astrophys. J., 522, 225.
[2] R. U. Abbasi et al. 2004, Astrophys. J., 610, L73.
[3] F. J. Anscombe 1954, Biometrics, 10, 89.
[4] P. Armitage, C. K. McPherson, and B. C. Rowe 1969, J. Roy. Stat. Soc. A, 132, 235.
[5] D. A. Berry 1987, Amer. Stat., 41, 117.
[6] A. Wald 1945, Ann. Math. Stat., 16, 117.
[7] —. 1947, Sequential Analysis (New York, NY: John Wiley and Sons).
[8] S. Y. BenZvi, B. M. Connolly, and S. Westerhoff 2008, Astrophys. J. 687, 1035.
[9] D. A. Darling, and H. Robbins 1968, Proc. Nat. Acad. Sci. USA, 61, 804.
[10] J. Abraham et al. (The Pierre Auger Collaboration) 2008, Astroparticle Phys., 29, 188.
[11] J. D. Hague for the Pierre Auger Collaboration 2009, Proc. 31^{st} ICRC, Łódź.