Maximizing Signal Search Sensitivity Using the Likelihood Ratio as Event Weight

Hongbo Hu
Institute of High Energy Physics,P.O.Box 918,Beijing 100039,P.R.of China

Abstract

A statistical analysis using the Likelihood Ratio (LR) as event weight for high statistic experiment is described. Making use of one or more discriminating variables to maximize the signal search sensitivity, LR test-statistic effectively extends the normal event counting to weighted event counting, in case of a high statistic experiment.

1. Introduction

The purpose of this paper is to describe an optimal and intuitive signal search method for high statistic experimets. In recent years, many Higgs search experiments successfully constructed the LR test statistic using one or more discriminating variables and significantly improved the search sensitivity [1,2,3] , both for the exclusion limit and the discovery significance. Like TeV GRBs search,typical Higgs search analyses deal with low-statistic candidate events sample, their LR distributions are mastered in the charactaristic discriminanting variable distribution, no general form can be derived. As for the high statistic experiments, owing to the central limit theorem, the LR test statistic obeys a Gaussian distribution, of which the two parameters, mean value and variance can be calculated in a way, effectively as to perform a weighted event counting.

2. Event counting method and LR estimator

It can be seen from this section that the normal event counting method is the simplest application of the LR test statistic. As an example[4], in searching for a γ ray source with an EAS array, a round shape "on" source window is defined with the center at the point of the source and a radius comparable to the detector's angular resolution, the number of events in this window is counted as N_{on}, while a couple of "off" source windows with the same shape are chosen to estimate the background event number N_{off}. Denoting α the ratio of "on" source and "off" source exposures, significance then can be calculated as:

$$(N_{on} - \alpha N_{off})/\sqrt{\sigma_{on}^2 + \alpha^2\sigma_{off}^2} = (N_{on} - \alpha N_{off})/\sqrt{N_{on} + \alpha^2 N_{off}} \qquad (1)$$

pp. 3027–3030

here, $\sigma_{on,off} = \sqrt{N_{on,off}}$ are the statistic unceratinty on $N_{on,off}$ respectively. It should be mentioned that all the events selected in the window are counted equally, with a unit weight, while all the rejected events with a zero weight. Firstly it is unnature for the events near the boundary to have their weights jumping from zero to one; Secondarily, the window size has to be adjusted in order to have an optimal sensitivity;Thirdly, as the signal events out of window are rejected, the sensitivity in principle is not maximized. The best solution to those arguments is to use LR test statistic, which is defined as:

$$R = C\frac{\mathcal{L}_{s+b}}{\mathcal{L}_b}, \tag{2}$$

here $\mathcal{L}_{s+b,b}$ is the probability density function for signal+background and background only experiments. Constant factor C can be any none zero value and in this work is chosen to be e^s where s is the expected number of signal events, one experiment cannot become more signal-like by choosing a different C. For the same reason, any monotonous transform of R is equivlent to the orignal R. When no discriminanting variable is used, for n events observed experiment:

$$\mathcal{L}_{s+b} = e^{-(s+b)}\frac{(s+b)^n}{n!}, \mathcal{L}_b = e^{-b}\frac{b^n}{n!}, R = e^s\frac{\mathcal{L}_{s+b}}{\mathcal{L}_b} = (1+s/b)^n \tag{3}$$

here, b is the background number. As we have disscussed above, a logrithm of R is an equivlent estimator, denoting W as

$$W = \ln(R) = \ln(1+s/b)^n = n\ln(1+s/b), \tag{4}$$

therefore, the *pdf* of W is a poisson distribution. This simply means that the LR test-statistic leads to the number counting method when no discrminanting variable is used. In general, signal and background events can be distinguished by some observable variables, denoting those variables by $\{\vec{v}_i\}$ for i_th observed event:

$$W = \ln(e^s\frac{\frac{e^{-(s+b)}(s+b)^n}{n!}\prod_{i=1}^{n}\frac{s\rho_s(\{\vec{v}_i\})+b\rho_b(\{\vec{v}_i\})}{s+b}}{\frac{e^{-b}b^n}{n!}\prod_{i=1}^{n}\rho_b(\{\vec{v}_i\})}) = \sum_{i=1}^{n}\ln(1+\frac{s\rho_s(\{\vec{v}_i\})}{b\rho_b(\{\vec{v}_i\})}) \tag{5}$$

In comparison with the number counting method, the weight [2] for i_th event $w(i)$ is thus $\ln(1+\frac{s\rho_s(\{\vec{v}_i\})}{b\rho_b(\{\vec{v}_i\})})$.

3. Weighted event counting method in high statistic experiment

Assuming the *pdf* of one(n) event(s) experiment's weight $w_1(w_n)$ is $\rho_1(w)$ ($\rho_n(w_n)$), then the average and the variance of w_1 can be calculated as

$$\overline{w_1} = \int w_1\rho_1(w_1)dw_1, \sigma_1^2 = \int(w_1-\overline{w_1})^2\rho_1(w_1)dw_1 \tag{6}$$

When an experiment observing large number of events N, mean and variance of w can be calculated directly from N observed events, and approch to the exact values when N approches infinity.

$$\overline{w_1} = \frac{1}{N}\sum_{i=1}^{N} w(i), \sigma_1^2 = \frac{1}{N}\sum_{i=1}^{N}[w(i) - \overline{w_1}]^2 = \frac{1}{N}\sum_{i=1}^{N} w(i)^2 - \overline{w_1}^2 \tag{7}$$

with a simple derivation, the average and variance of w_n for the experiment observing n events can be calculated:

$$\overline{w_n} = \overline{\sum_{i=1}^{n} w(i)} = n\overline{w_1}, \sigma_n^2 = \int (w_n - \overline{w_n})^2 \rho_n(w_n) dw_n = n\sigma_1^2 \tag{8}$$

For overall experiments, composing all possible n observed events while n follows a poisson distribution $P(n, \mu)$, here μ is the expected number of events,

$$\overline{W} = \sum_{n=0}^{\infty} P(n, \mu)\overline{w_n} = \mu\overline{w_1}, \sigma^2 = \sum_{n=0}^{\infty} P(n, \mu) \int (W - \overline{W})^2 \rho_n(w_n) dw_n = \mu\sigma_1^2 + \mu\overline{w_1}^2, \tag{9}$$

which are another form of equation(7) if the observed number of events N is replaced by the expected number of events μ, in another word, the two characteristic variables of the W distribution, average and variance can be calculated from the summation of the observed individual estimator $w(i)$ and their square $w(i)^2$. This in general leads to a weighted event counting. Similar to the normal event counting method, significance can be calculated as:

$$(W_{on} - \alpha W_{off}) / \sqrt{\sigma_{on}^2 + \alpha^2 \sigma_{off}^2} \tag{10}$$

4. Example

As an illustration, considering a search of γ ray point source along the galactic plane with a detector having an angular resolution as 1^o. The signal would be distributed as $\rho_s(\delta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta^2}{2}}$, where δ is the signed angular distance from the source in degree. Further assuming that s, b are 200 and 10000, background is in a flat distribution in the range from -10^o to 10^o, namely, $\rho_b(\delta) = \frac{1}{20}$.

20000 toy MC experiments were generated to check the validity of eq(9). Fig.1a is the $\sum w_i$ distribution for background only experiments, well discribed by a Gaussian, of which the width agrees with the expectation of $\sum w_i^2$ shown in Fig.1b. Fig.2 is the significance comparison between normal event counting method and weighted event counting method. The latter one is about 6% better than the optimal value in the former method.

5. Conclusion

LR test statistic is a well established and widely practiced method. When discriminating variables are used, it naturaly extends the normal(unweighted)

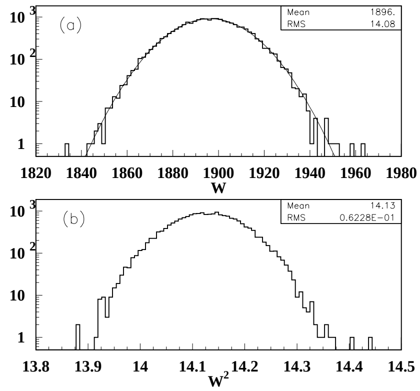

Fig. 1. (a) shows the distribution of the summed weight, fitted well by a Gaussian as expected; (b) shows the distribution of the summed weight square, of which the mean value well represents the width of plot(a).

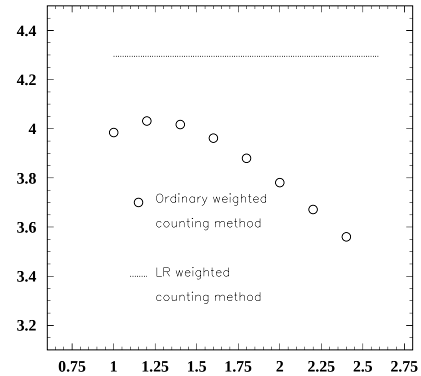

Fig. 2. The averaged significance from 20000 MC experiments as a function of the window size are shown as the open circle, the dashed line is the significance from the event weighting method.

number counting method to weighted event counting, and gives the maximum sensitivity.

References

[1] A.G,Frodesen,O.Skjeggestad,and H.Tofte,*Probability and Statistic in Particle Physics*,Universitetsforlaget,1979,**ISBN** 82-00-01906-3;
S.L.Meyer,*Data Analysis for Scientists and Engineers*,John Wiley and Sons, 1975, **ISBN** 0-471-59995-6;
A.Stuart,A.K.Ord,and Arnold,*Kendall's Advanced Theory of Statistic*, Vol.2 *Classical Inference and Relationship* 6th Ed., (Oxford Univ. Press,1998).

[2] A. L. Read, DELPHI note 97-158 PHYS 737 (1997);

[3] P. Janot and F. Le Diberder, Nucl. Instrum. Methods **A411** (1998) 449;
T. Junk, `hep-ex/9902006` (1999);
H.Hu,J.Nielsen,High Energy Phys. and Nucl. Phys.Vol.24,No.5(2000)450.

[4] Amenomori,M. et al. 2001, Proc. 27th ICRC(Hamburg),6,2395;
Amenomori,M. et al. astro-ph/0304241(2003), submitted to Astrophysical Journal.