Draft

On the Unified Method With Nuisance Parameters

Bodhisattva Sen, Matthew Walker, and Michael Woodroofe

The University of Michigan

March 10, 2006

1 Introduction

The unified method of Feldman and Cousins [5] has attracted wide interest among high-energy physicists since its appearance. It was originally recommended for problems with a restricted parameter space, for example a normal mean that is known to be positive, and called "unified" because it makes a natural transition from an one-sided confidence bound to a two sided confidence interval. Only problems without nuisance parameters were considered in [5]. Interest here centers on problems in which there is a nuisance parameter in addition to the parameter of primary interest. Consideration of several examples reveal some interesting differences (from the cases without nuisance parameters).

To describe the unified method and understand the issues, suppose that a data vector X has a probability density (or mass function, in the discrete case) $f_{\theta,\eta}$ where θ is the parameter of interest and η is a nuisance parameter. For example, if a mass θ is measured with normally distributed error with an unknown standard deviation, then θ is of primary interest and the standard deviation of the measurement is a nuisance. Let L denote the likelihood function

$$L(\theta, \eta|x) = f_{\theta,\eta}(x);$$

further, let $\hat{\eta}_\theta = \hat{\eta}_\theta(x)$ be the value of η that maximizes $L(\theta, \eta|x)$ for a fixed θ; let $\hat{\theta} = \hat{\theta}(x)$ and $\hat{\eta} = \hat{\eta}(x)$ be the values of θ and η that maximize $L(\theta, \eta|x)$ over all allowable values; and let

$$\Lambda_\theta(x) = \frac{L(\theta, \hat{\eta}_\theta(x)|x)}{L(\hat{\theta}(x), \hat{\eta}(x)|x)}. \tag{1}$$

Then unified confidence intervals consist of θ for which $\Lambda_\theta(x) \geq c_\theta$, where c_θ is a value whose computation is discussed below.

For a desired level of coverage $1-\alpha$, a literal (and correct) interpretation of "confidence" requires that $P_{\theta,\eta}[\Lambda_\theta(X) \geq c_\theta] \geq 1-\alpha$ for all θ and η, where $P_{\theta,\eta}$ denotes probability computed under the assumption that the parameter values are θ and η. Equivalently it requires $\min_\eta P_{\theta,\eta}[\Lambda_\theta(X) \geq c_\theta] \geq 1-\alpha$ for each θ. Thus, c_θ should be the largest value of c for which

$$\min_\eta P_{\theta,\eta}[\Lambda_\theta(X) \geq c] \geq 1-\alpha. \tag{2}$$

For a fixed x, the confidence interval is then $\mathcal{C}(x) = \{\theta : \Lambda_\theta(x) \geq c_\theta\}$, and its coverage probability

$$P_{\theta,\eta}[\theta \in \mathcal{C}(X)] = P_{\theta,\eta}[\Lambda_\theta(X) \geq c_\theta] \geq 1-\alpha, \tag{3}$$

by construction. Being likelihood based, unified confidence intervals are generally reliable, even optimal, in large samples – for example [8] – but not necessarily so in small samples, and unified confidence intervals have been criticized in that context – e.g. [9] and [10].

In some simple cases, it is possible to compute c_θ analytically. This is illustrated in Section 2 and 3. In other cases, one can in principle proceed by numerical calculation. This requires computing $P_{\theta,\eta}[\Lambda(X) \geq c]$ over a grid of (θ, η, c) values, either by Monte-Carlo or numerical integration, and then finding the c_θ by inspection, replacing the minimum in (2) by the minimum over the grid. This is feasible if η is known or absent and was done by Feldman and Cousins in two important examples. But if η is present and unknown, then numerical calculations become unwieldy, especially if η is a vector.

One way to circumvent the unwieldy numerical problems, when η is present, is to use the chi-squared approximation to the distribution of Λ_θ, as in [11], or a chi-squared approximation supplemented by a Bartlett correction. Another is to use the hybrid resampling method of Chuang and Lai, [1] and [2]. Generate and random variable X^* from $P_{\theta,\hat{\eta}_\theta}$ and let $c_\theta^+ = c_\theta^+(x)$ be the largest values of c for which $P_{\theta,\hat{\eta}_\theta}[\Lambda(X^*) \geq c] \geq 1-\alpha$. Then the hybrid confidence intervals consist of θ for which $\Lambda_\theta(x) \geq c_\theta^+$. This requires computation over a grid of θ values, but not over η for fixed θ. Unfortunately, relation (3) cannot be asserted for the hybrid intervals, but Chuang and Lai argue both theoretically and by example that it should be approximately true. In some cases the calculations can be done by numerical integration, but they can always be done by simulation. For a given x, generate independent $X_1^*, \cdots, X_N^*$ (pseudo) random numbers from the density $f_{\theta,\hat{\eta}_\theta}$; compute $\Lambda_\theta(X_k^*)$ from (1) with x replaced by X_k^*; and let c_θ^* be the largest value of c for which

$$\frac{\#\{k \leq N : \Lambda_\theta(X_k^*) \geq c\}}{N} \geq 1 - \alpha. \tag{4}$$

Here the left side of (4) provides a Monte Carlo Estimate for $P_{\theta,\hat{\eta}_\theta}[\Lambda_\theta(X^*) \geq c]$, and c_θ^* provides an estimate of c_θ^+.

The hybrid method resembles Efron's bootstrap resampling method, but differs in one important respect. For computing (2) for fixed θ, θ and η are replaced by θ and $\hat{\eta}_\theta$, as opposed to $\hat{\theta}$ and $\hat{\eta}$. This is the origin of the term "hybrid". Evidence that the hybrid method is reliable – that is, that (3) is approximately true comes from two sources, asymptotic approximations and simulations. These are reported in [1] and [2] and include some dramatic successes. Here the method is applied to three examples of interest to astronomers and physicists: estimating a non-negative normal mean, estimating the angle when the mean (vector) of a bivariate normal distribution is expressed in polar coordinates, and estimating the Poisson mean in the presence of background. The hybrid method has (independently) been suggested in the physics literature by Feldman (see [6]).

2 The Normal Case

Suppose that $X = (Y, W)$, where Y and W are independent, Y is normally distributed with mean $\theta \geq 0$ and variance σ^2, and W/σ^2 has a chi-squared distribution with r degrees of freedom. For example, if data originally consists of a sample $Y_i = \theta + \epsilon_i,\ i = 1, \cdots, n$, where ϵ_i's are independent and identically distributed $N(0, \sigma^2)$, then one can let $Y = \bar{Y}$ and $W = (n-1)V^2/n$ where $\bar{Y}$ and V^2 denote the sample mean and variance of $Y_1, \cdots, Y_n$. The unknown parameters here are $\theta \geq 0$ and $\sigma^2 > 0$. Thus, the likelihood function is

$$L(\theta, \sigma^2 | y, w) = \frac{1}{\sqrt{2^{r+1}\pi}\Gamma(r/2)} \frac{w^{\frac{1}{2}r-1}}{\sigma^{r+1}} \exp\left\{-\frac{1}{2\sigma^2}[(y-\theta)^2 + w]\right\}.$$

For a given θ, L is maximized by

$$\hat{\sigma}_\theta^2 = \frac{1}{r+1}[w + (y-\theta)^2];$$

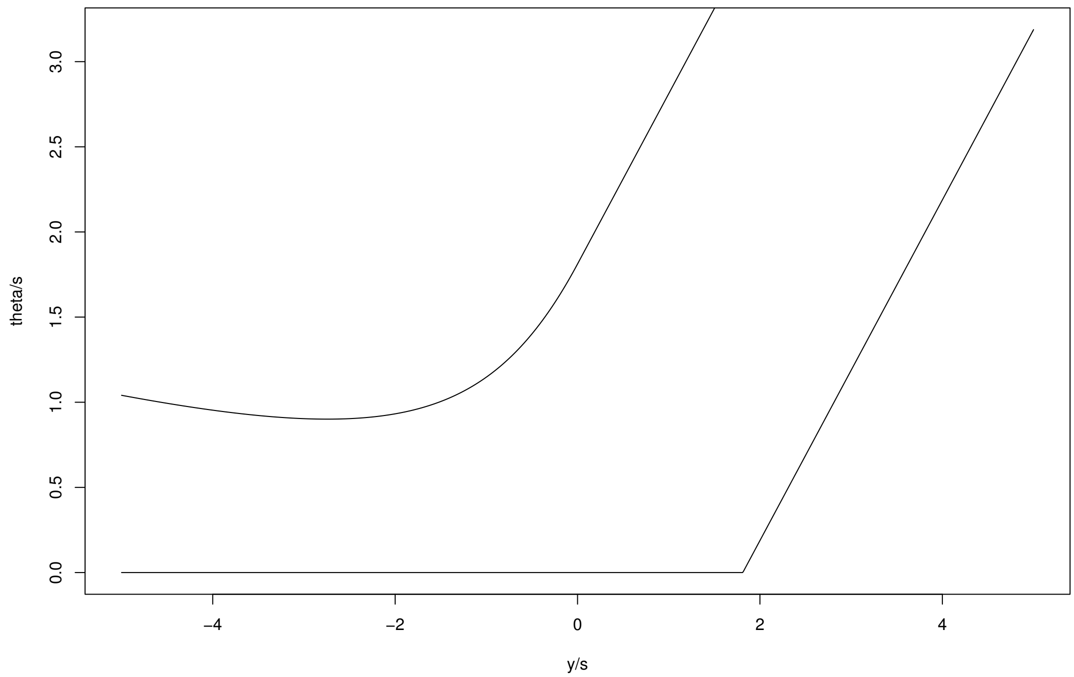

Figure 1: Confidence limits for θ/s as a function of y/s when $r = 10$ and $\alpha = .1$

and L is maximized with respect to θ and σ^2 jointly by

$$\hat{\theta} = \max[0, y] = y_+,$$

say, and

$$\hat{\sigma}^2 = \frac{1}{r+1}[w + y_-^2],$$

where $y_- = -\min[0, y]$. After some simple algebra,

$$\log[\Lambda_\theta] = -\frac{1}{2}(r+1)\log(\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}^2}) = -\frac{1}{2}(r+1)\log\left[\frac{W + (Y-\theta)^2}{W + Y_-^2}\right].$$

Let

$$U = \frac{W}{\sigma^2} \qquad \text{and} \qquad Z = \frac{Y-\theta}{\sigma}.$$

Then U and Z are independent random variables for which $U \sim \chi_r^2$ and $Z \sim \text{Normal}(0,1)$, and

$$\log[\Lambda_\theta] = -\frac{1}{2}(r+1)\log\left[\frac{U + Z^2}{U + (Z + \theta/\sigma)_-^2}\right].$$

This is an increasing function of σ for each $\theta > 0$. So, since the joint distribution of U and Z does not depend on parameters,

$$\min_{\sigma>0} P_{\theta,\sigma}[\Lambda_\theta \geq c] = \lim_{\sigma\to 0} P_{\theta,\sigma}[\Lambda_\theta \geq c] = P\left[-\frac{1}{2}(r+1)\log\left(1 + \frac{T^2}{r}\right) \geq -\log(c)\right],$$

where

$$T = \frac{Z}{\sqrt{U/r}}$$

has t-distribution with r degrees of freedom. Thus the desired c is

$$c = \exp\Big\{ -\frac{1}{2}(r+1)\log\big[1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}\big]\Big\},$$

where $t_{r,1-\frac{1}{2}\alpha}$ is the $1-\frac{1}{2}\alpha$ percentile of the latter distribution and is independent of θ.

To find the confidence intervals, one must solve the inequality $\Lambda_\theta \geq c$ for θ. Letting $s^2 = W/r$, this may be written

$$\frac{1 + (y-\theta)^2/(rs^2)}{1 + y_-^2/(rs^2)} \leq 1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}$$

or

$$[y - bs]_+ \leq \theta \leq y + bs, \tag{5}$$

where

$$b = \sqrt{t^2_{r,1-\frac{1}{2}\alpha} + \frac{y_-^2}{s^2}\left(1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}\right)}. \tag{6}$$

Thus, if $y > 0$, then the unified intervals are just the usual t-intervals, truncated to non-negative values; and if $y > bs$, then they are symmetric about y. This differs from the case of known σ, where the intervals are (slightly) asymmetric, even for large y. There is a more dramatic difference with the case of known σ for $y < 0$. Observe that for $y < 0$,

$$y + bs \geq s\sqrt{\frac{y^2}{s^2}\left(1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}\right)} - \frac{|y|}{s} = |y|\sqrt{1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}} - 1.$$

So the upper confidence limit approaches $+\infty$ as $y \to -\infty$, unlike the case of known σ where it approaches 0. Mandelkern [7] found the latter behavior non-intuitive. If we let $r \to \infty$ and $s^2 \to \sigma^2$, then we do *not* recover the intervals of Feldman and Cousins with known σ^2. Rather, we get the interval (5) with the t-percentile replaced by the corresponding normal percentile.

Observe that the confidence limits for θ may be written as $[y/s - b]_+ \leq \theta/s \leq y/s + b$. Figure 1 shows shows these upper and lower confidence limits for θ/s as a function of y/s for $r = 10$ and $\alpha = .1$. For a specific example, suppose that $r = 10$, $s = 1$, $y = -.3$, and $\alpha = .90$. Then $b = \sqrt{(1.812)^2 + (.3)^2(1 + (.1812)^2)} = 1.84$, and the interval is $0 \leq \theta \leq 1.54$. The hybrid method yields $0 \leq \theta \leq 1.14$ in this example. The details are omitted here, but an example using the hybrid method is included in Section 4.

3 Angles

Suppose that $X = (X_1, X_2)$ where X_1 and X_2 are normally distributed random variables with unknown means μ_1 and μ_2 and known variance σ^2. Write μ_1 and μ_2 in polar coordinates, $\mu_1 = \rho\cos(\theta)$ and $\mu_2 = \rho\sin(\theta)$, where $-\pi < \theta \le \pi$, and consider confidence intervals for θ. In this example, the likelihood function,

$$L(\theta,\rho|x) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}\left[(x_1-\rho\cos(\theta))^2+(x_2-\rho\sin(\theta))^2\right]\right\},$$

is maximized for a fixed θ by

$$\hat{\rho}_\theta = \max[0, x_1\cos(\theta) + x_2\sin(\theta)]$$

and unconditionally by $\hat{\rho}$ and $\hat{\theta}$, where $x_1 = \hat{\rho}\cos(\hat{\theta})$ and $x_2 = \hat{\rho}\sin(\hat{\theta})$. Then $L(\hat{\theta},\hat{\rho}|x) = 1/(2\pi\sigma^2)$, and

$$\Lambda_\theta = \exp\big[-\frac{1}{2\sigma^2}(\hat{\rho}^2 - \hat{\rho}_\theta^2)\big].$$

Let

$$Z_1 = \frac{1}{\sigma}[\cos(\theta)X_1 + \sin(\theta)X_2 - \rho],$$
$$Z_2 = \frac{1}{\sigma}[\sin(\theta)X_1 - \cos(\theta)X_2].$$

Then Z_1 and Z_2 are independent normal variables (both) with the same mean 0 and unit variance, and

$$\Lambda_\theta = \exp\big\{-\frac{1}{2}[(Z_1+\rho)_-^2 + Z_2^2]\big\},$$

where (recall) $z_- = -\min[0, z]$, after some simple algebra. Thus, Λ_θ is an increasing function of ρ for fixed Z_1, Z_2, and θ. So, since the joint distribution of Z_1 and Z_1 does not depend on parameters

$$\min_\rho P_{\theta,\rho}[\Lambda_\theta \ge c] = \lim_{\rho\to 0} P_{\theta,\rho}[\Lambda_\theta \ge c].$$

Letting $b = -2\log(c)$, this is just

$$P[Z_{1,-}^2 + Z_2^2 \le b] = P[Z_1 \le 0, Z_1^2 + Z_2^2 \le b] + P[Z_1 > 0, Z_2^2 \le b]$$
$$= \frac{1}{2}P[\chi_1^2 \le b] + \frac{1}{2}P[\chi_2^2 \le b].$$

So, $c = e^{-2b}$, where b solves $\frac{1}{2}P[\chi_1^2 \le b] + \frac{1}{2}P[\chi_2^2 \le b] = 1-\alpha$. For example, when $\alpha = .90$, $b = 3.808$.

Unified confidence intervals for θ then consist of θ for which $\hat{\rho}^2 - \hat{\rho}_\theta^2 \le b\sigma^2$, or equivalently $\hat{\rho}_\theta^2 \ge \hat{\rho}^2 - b\sigma^2$. Thus, if $\hat{\rho}^2 \le b\sigma^2$, then the interval consists of all values $-\pi < \theta \le \pi$. On one hand, this simply reflects the (obvious) fact that if $\hat{\rho}$ is small, then there is no reliable information for estimating θ, but it also admits the following amusing paraphrase: One is $100(1-\alpha)$ percent confident of something that is certain. If $\hat{\rho}^2 > b\sigma^2$, then the intervals consist of θ for which $\hat{\rho}\cos(\theta - \hat{\theta}) \ge \sqrt{\hat{\rho}^2 - b\sigma^2}$; that is

$$\hat{\theta} - \arccos(\sqrt{1 - \frac{b\sigma^2}{\hat{\rho}^2}}) \le \theta \le \hat{\theta} + \arccos(\sqrt{1 - \frac{b\sigma^2}{\hat{\rho}^2}}),$$

where $\arccos(y)$ is the unique ω for which $0 \le \omega \le \pi$ and $\cos(\omega) = y$ and addition is understood modulo π. Thus, there is a discontinuity in the length of the intervals as $\hat{\rho}$ passes through $b\sigma^2$: It decreases from 2π to something less than π.

4 Counts with Background

Suppose that $X = (W, Y)$ where W and Y are independent, W has the Poisson distribution with mean mb, and Y has the Poisson distribution with mean $b + \theta$. Here b and θ are unknown; m is assumed known and large values of m are of interest. In this case, the likelihood function and score functions are

$$L(\theta, b|w, y) = f_{\theta,b}(w, y) = \frac{(mb)^w}{w!}e^{-mb} \times \frac{(\theta+b)^y}{y!}e^{-(\theta+b)},$$

$$\frac{\partial \log(L)}{\partial \theta} = \frac{y}{b+\theta} - 1,$$

and

$$\frac{\partial \log(L)}{\partial b} = \frac{w}{b} + \frac{y}{\theta+b} - (m+1).$$

Consider $\hat{b}_\theta$ for a fixed θ. If $w = 0$, then L is maximized when $b = [y/(m+1) - \theta]_+$; and if $w > 0$ it is maximized at the (positive) solution to $\partial\log(L)/\partial b = 0$, i.e.,

$$\hat{b}_\theta = \frac{[(w+y) - (m+1)\theta] + \sqrt{[(m+1)\theta - (w+y)]^2 + 4(m+1)w\theta}}{2(m+1)}; \tag{7}$$

and fortuitously, (7) also gives the correct answer when $w = 0$. The unconstrained maximum likelihood estimators may then be found as $\hat{\theta}$ and $\hat{b} = \hat{b}_{\hat{\theta}}$, where $\hat{\theta}$ maximizes the profile likelihood function $L(\theta, \hat{b}_\theta|w, y)$. Considering the cases $y \le w/m$ and $y > y/m$ separately, shows that

$$\hat{\theta} = (y - \frac{w}{m})_+$$

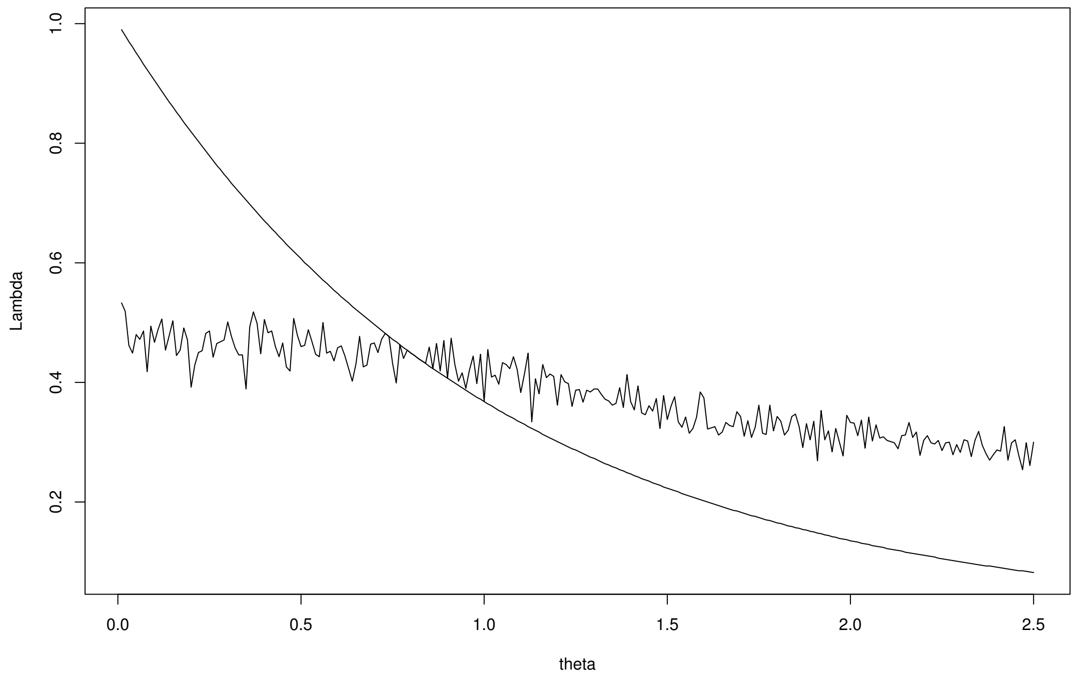

Figure 2: Λ_θ (smooth line) and c_θ (jagged line).

and

$$\hat{b} = \frac{w + y - \hat{\theta}}{m + 1}.$$

So,

$$\Lambda_\theta(y, w) = (\frac{\hat{b}_\theta}{\hat{b}})^w (\frac{\theta + \hat{b}_\theta}{\hat{\theta} + \hat{b}})^y \exp\left[(w + y) - (m + 1)\hat{b}_\theta - \theta\right],$$

after some simple algebra.

We have been unable to find the minimizing value in (2) and, so, will use the Hybrid Resampling Method. This is best illustrated by an example. Figure 2 below shows Λ_θ and c_θ for an example in which $m = 6,\ w = 23$, and $y = 0$. This is patterned after the original KARMEN report [4], but with a larger value of $\hat{b}$ and more variability in $\hat{b}$. The c_θ^* was computed by Monte Carlo on the grid $\theta = 0, .1, .2, \cdots, 10$ using $N = 10,000$ in (4). The right end-point of the interval can be read from the graph.

By construction, the hybrid-unified method always delivers non-degenerate subinterval of $[0, \infty)$, even when $y = 0$, and, thus, avoids the types of problems reported in [11]. It does not avoid the problems inherent in the use of the unified method without nuisance parameters, however – for example, dependence of the interval on $\hat{b}$ when $y = 0$. We believe that the interval $[0, 2.31]$ is a more reasonable statement of the uncertainty in this example, for reasons explained in [9].

References

[1] Chuang, C. and Lai, T.L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika*, **85**, 317-332.

[2] Chuang, C. and Lai, T.L. (2000). Hybrid resampling methods for confidence intervals. *Stat. Sinica*, **10**, 1-50.

[3] Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood for incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **3**9, 1-38.

[4] Eitel, K. and Zeitnitz, B. (1998). The search for neutrino oscillations $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ with KARMEN. Available at arXiv/hepex/9809007.

[5] Feldman, G. and Cousins, R. (1998). Unified approach to the classical statistical analysis of small signal.

[6] Feldman, G (2000). Multiple Measurements and Parameters in the Unified Approach. Talk at the FermiLab Workshop on Confidence Limits.

[7] Mandelkern, M. (2002). Setting Confidence Intervals for Bounded Parameters. *Stat. Sci.*, **17**, 149-172.

[8] Rice, John (1995). *Mathematical Statistics and Data Analysis.* Duxbury Press, California.

[9] Roe, B. and Woodroofe, M. (1998). Improved probability method for estimation the signal in the presence of background. *Phys. Rev. D*, **60**, 053009.

[10] Roe, B. and Woodroofe, M. (2000). Setting confidence belts. *Physical Review D*, **63**, 013009.

[11] Rolke, W., López, A. and Conrad, J. (2005). Limits and Confidence Intervals in the Presence of Muisance Parameters. *Nucl.Instrum.Meth.*, A551, 493-503 (arXiv:physics/0403059 v4 7 July 2005).