Weighting Background-Subtracted Events

James T. Linnemann[1] *(presenter)*
Andrew J. Smith[2]
[1]Michigan State University, 3245 BPS Building, E. Lansing, Michigan 48823
[2]Department of Physics, University of Maryland, College Park, MD 20742

Abstract
Often a full maximum likelihood (ML) estimate is inconvenient for computational reasons (e.g., iteration over large data sets). If a variable x is a discriminating variable ($s(x) \neq b(x)$), a weight function can be found which allows estimation of the number of signal events with a variance approaching that of a ML estimate of the same quantity. We derive a formula and discuss it in the context of more general results on event weighting from earlier papers by Barlow and Tkachov, which also find weighting out-performs cutting.

1 Introduction

The origin of this talk lies within the Milagro cosmic ray experiment [1]. However, the results apply just as well within LHC experiments, because both have to subtract backgrounds in order to see signals. Milagro's physics goal is to look at structure with TeV gamma rays, which are outnumbered by charged cosmic rays by 10^3. Thus our analysis procedures must calculate backgrounds correctly to 1 ppt. We make background-subtracted sky maps of measured photon excess $m = n - \hat{B}$. To enhance our statistical significance, we seek discriminator variables x whose probability density distributions differ for signal $s(x)$ and background $b(x)$. We then consider the *background subtracted* excess $m(x) = n(x) - \hat{B}b(x)$.

What is the best way to combine (say) bins of $m(x)$ for a best overall estimate of the excess? The naive solution is just to sum all the bins. My colleague Andy Smith argued you could do better by weighting the bins by the ratio of expected signal and background contributions to each bin

$$w(x) = E[S(x)]/E[B(x)] = K\ s(x)/b(x) = K\ r(x) \tag{1}$$

where K is a constant (independent of x) which can be ignored for calculating relative weights of bins of x. My first reaction was that this was cheating, since you've already used the expected background in the subtraction that led to the observed $m(x)$. But Andy was right!

2 Event Weighting

The underlying hypothesis is that the signal distribution across x bins is governed by an overall intensity M, and the signal distribution $s(x)$. The naive estimate (from x bins i) of M and its variance would be:

$$\hat{M_1} = \sum m_i;\ \ V[M_1] = \sum V(m_i) = \sum V_i \tag{2}$$

But this is not using the information about the expected relationship between bins, $s(x)$, which allows each bin to independently estimate M:

$$E[m_i] = Ms_i;\ \ \hat{M_i} = m_i/s_i;\ \ V[\hat{M_i}] = V_i/s_i^2 \tag{3}$$

How should these independent estimates $\hat{M_i}$ be best combined? The Best Linear Unbiased Estimate (BLUE) method invokes the Gauss-Markov theorem (James [2] §7.4.4) for the solution, which is to weight the estimates by the inverse of their variance, so that smaller-variance estimates are more heavily

weighted; the result has the minimum variance among the class of unbiased linear estimators (of which $\hat{M}_1$ is an inferior member).

$$\hat{M} = \sum \hat{M}_i w_i / \sum w_i = \sum (m_i s_i / V_i) / \sum (s_i^2 / V_i); \tag{4}$$

This solution holds for *any* distribution of uncorrelated, unbiased, variables. But it does require that the actual variance be used (and fluctuations to low estimated variance are particularly damaging [3]). $\hat{M}$ is identical to the minimum M found for

$$\chi^2 = \sum (m_i - M s_i)^2 / V_i \; . \tag{5}$$

$\hat{M}$ can also be regarded as applying for each event in bin i a weight u_i as defined below:

$$\hat{M} = k \sum (m_i s_i / V_i) = k \sum m_i u_i; \;\; u_i = s_i / V_i; \;\; 1/k = \sum (s_i^2 / V_i) \tag{6}$$

Finally, we arrive at Andy's weight by recognizing that for a large Poisson background, $V_i \approx B_i = B b_i$, so that $u_i = K' s_i / b_i$, as in Eq. 1. Milagro deals with billions of events, so it is a considerable advantage to sum event weights, rather than minimizing for each pixel of sky.

We can calculate the variance of the BLUE solution (using the definitions of k, u in Eq. 6):

$$V[\hat{M}] = k^2 \sum V[m_i] u_i^2 = k^2 \sum V_i u_i^2 = 1 / \sum (s_i^2 / V_i) = k \tag{7}$$

For sufficient data, the BLUE solution approaches the Cramer-Rao minimum variance bound (James §7.4.5). Because we can formulate the BLUE solution as event weighting (often referred to as the method of moments) we find that despite the "suboptimal" reputation of the method of moments (James §8.2.2), in this case it is competitive with ML (assuming Gaussian uncertainties).

3 Sensitivity to Assumptions

It is worth clarifying here how the solution depends on the assumptions made. Since we have independently normalised the shapes s, b, their absolute normalisation S, B does not matter. But we are sensitively dependent on the shapes $s(x),\ b(x)$. In Milagro we determine $b(x)$ from the data and can use it to check the simulations. But s comes from the simulation, and depends on the input shower physics, and (if the variable x is correlated with energy) on the assumed source energy spectrum. We test by comparing the MC $s(x)$ distribution with data.

4 Barlow's Event Weighting

Was the good performance of event weighting a fluke of this particular problem? Remarkably, no! In a 1987 paper, Barlow [4] found the *best* event weight function to count signal events. He first wrote the expected weight $E[w(x)]$ in terms of $s,\ b$ as

$$E[w_d] = (M\mu_s + B\mu_b)/N = (a\mu_s + \mu_b)/(a+1); \quad a = M/B = M/(N-M) \tag{8}$$

where μ_s, μ_b are the expected mean weights for signal and background. Substituting the observed data weight $\overline{w_d} = \sum w(x_j)/N$ for the expected and solving for M gives:

$$\widehat{M_B} = \sum (w_j - \mu_b)/(\mu_s - \mu_b) \tag{9}$$

These two equations are independent of the specifics of w, though $\mu_s,\ \mu_b$ depend on the form of w and its parameters. Eq. 9 makes it crystal clear that $\widehat{M_B}$ is unchanged by multiplicative *or* additive x-independent constants in $w(x) \to Cw(x) + D$.

The method of moments is quite general: calculate any function of the data, then solve for parameters, considering expected moments as functions $f(u)$ of the parameters of the true pdf. Eq. 9 is an example. Typically one chooses power moments $w' = x^n$ and hopes for the best.

4.1 Barlow's Optimal Weight

But Barlow did (much) better: he calculated a completely general equation for the variance $V[\widehat{M_B}]$ and used the calculus of variations to minimise $V[\widehat{M_B}]$ with respect to the function $w(x)$. He found the variance with the optimal weight function approached the ML variance (and Cramer-Rao bound), but unlike the ML solution, required no iteration through all the events! Further, the variance is *less* than the variance resulting from cutting on the optimal weight variable, though fitting to the distribution of $w(x)$ is also close to optimal. The optimum weight function Barlow found (after choosing a suitable normalization) was

$$w(x) = a_o s(x)/(a_o s(x) + b(x)) \text{ or} \tag{10}$$

$$w(x) = a_o r(x)/(1 + a_o r(x)); \qquad r(x) = s(x)/b(x) \tag{11}$$

Clearly $w \in [0, 1]$ (though Eq. 9 reminds us $\hat{M} \neq \sum w$). This optimal weight function should look remarkably familiar. The Neyman-Pearson lemma (James §10.3.1) tells us that the best variable for testing the hypothesis of whether an event is signal or background is $r(x) = s(x)/b(x)$; and as a result the best Bayesian discriminant (ideal neural net output) is the posterior signal probability $d(s|x) = as/(as + b)$, where $a = \pi_s/(1 - \pi_s)$ is the prior odds ratio.

There is a mild catch: one has to make an initial guess at M_o/B (why we wrote a_o instead of a). But the optimum is quadratic, so close ($a_o \approx a$) is quite good. Further, guessing is actually advantageous: it relieves you of iterating through the data. Since $E[w]$ already has a near-optimal dependence on the pdf parameters, and all you lose by the guess is a bit of variance increase, not a bias. However, wrong $s,\ b$ functions still give a biased $\hat{M}$ since you are fitting normalisation to an incorrect shape and $\mu_s,\ \mu_b$.

4.2 Comparison with BLUE Weight

Barlow notes that knowing the expected Poisson mean B reduces $V[\hat{M}]$, but finds the same $w(x)$ is optimal. When the fraction a_o of signal events is small, as it is in Milagro, then the Eq. 10 weight becomes $w \approx K''s/b$, showing the subtraction weight of Eq. 1 to be near-optimal for large backgrounds.

5 Tkachov Weights and the ML Solution

Barlow solved the specific problem of the best weight for separating signal and background. But Tkachov [5] later solved the more general problem of choosing the optimal $w(x)$ ("generalized moment") to estimate *any* pdf parameter u with minimum variance, again using the calculus of variations. His result is both more general, and simpler! Having *fixed* w, one estimates $\hat{u}$ through the dependence $E[w] = f(u)$ of the expected moment on the pdf parameters u, solving $\overline{w_d} = f(\hat{u})$. Functional differentiation relates the variance of $V[\hat{u}]$ in first order to the moment variance $V[w]$. More functional differentiation minimises $V[\hat{u}]$ wrt w, giving the optimum $w(x)$ choice, intimately related to the ML solution:

$$w_{opt}(x) = C(u)\ \frac{\partial Ln[p(x;u)]}{\partial u} + D(u) \tag{12}$$

Specializing to our case, $u = a$, and seeking $\hat{a}$ with $p = (as + b)/(1 + a),\ C = a,\ D = a/(1 + a)$, and $a \to a_o$ (our pre-data guess)

$$w = s/(as + b) - 1/(1 + a) \to w = a_o s/(a_o s + b)\ , \tag{13}$$

matching Barlow's Eq. 10 weight. The expected data weight is then

$$E[w_d(a_o, a)] = \int w(x; a_o)\ p(x; a)\ dx = \int \left(\frac{a_o s}{a_o s + b}\right)\left(\frac{as + b}{a + 1}\right)dx = \frac{a_o + (a - a_o)\mu_s}{a + 1} \tag{14}$$

which can be compared with the iterative ML solution written in terms of the weight function:

$$\sum_j \frac{\partial Ln[p(x_j; a)]}{\partial a} = 0 \Rightarrow \ \sum_j \frac{w(x_j, a_o \to a)}{Na} = \frac{1}{a + 1} \tag{15}$$

Tkachov shows that with the optimal weight function choice, no matter what the parameter, the method of moments gives a variance approaching the best possible.

One final comment: the optimal weight function of Eq. 10 is strongly reminiscent of the Wiener optimal frequency filter [6] with squared amplitudes (absolute power) instead of pdf's. The derivation minimises the reconstructed variance wrt the true signal, again using the calculus of variations.

6 Summary

A near-optimal weight can achieve near-ML accuracy. Weighting methods are powerful and simple. There is a rational scheme leading to choice of optimal weight (moment) functions. And both Barlow and Tkachov show that weighting (or fitting to a weight distribution) is more accurate (lower variance) than making cuts in even an optimal weight variable. A longer version of this paper is in preparation for submission to the Astrophysical Journal (and the arxiv server).

7 Acknowledgement

JTL thanks Harrison Prosper for having brought to my attention Refs. [4, 5] (years ago by now); it's a pleasure to tie them together. JTL also appreciated conversations with Sekhar Chivukula and Neil Christensen of MSU on some mathematical points.

References

[1] R. Atkins *et al.*, "Observation of TeV Gamma Rays from the Crab Nebula with Milagro Using a New Background Rejection Technique" Astrophysical Journal 595 (2003) 803-811; A. Abdo *et al.*, "TeV Gamma-Ray Sources from a Survey of the Galactic Plane with Milagro", Astrophysical Journal Letters 664 (2007) L91-L94.

[2] F. James, "Statistical Methods in Experimental Physics", World Scientific, 2006, §7.4.4; and J. Rice, "Mathematical Statistics and Data Analysis", 2nd ed, Duxbury 1995, §14.8.pr5; the latter generalizes the Gauss-Markov theorem to variables with unequal variance by dividing each variable by the square root of its variance, arriving at weighted least squares.

[3] L. Lyons, "Statistics for nuclear and particle physicists", Cambridge 1986, §1.6.(ii)

[4] R. Barlow, "Event Classification Using Weighting Methods", J. Comp. Phys 72 (1987) p202. Barlow uses N_S, $\bar{w}$, A where we use M, μ, $1/a$. He uses individual events j instead of bins i.

[5] F. Tkachov, "Approaching the Parameter Estimation Quality of Maximum Likelihood via Generalized Moments" Part. Nucl. Lett. 111(2002) 28 or arXiv:physics/0001019; arXiv:hep-ph/0210116; arXiv:physics/0604127. Tkachov uses f, π, P where we use w, p, a. Derivatives of moments wrt pdf parameters a receive no contribution from $w(x, a_o)$ because the choice of a_o has no functional dependence on a, though one finds the optimal *value* of a_o is a. Tkachov prefers $E[w] = 0$, e.g. $C = 1,\ D = 0$ in Eq. 13 rather than $w \in [0, 1]$. A caution: the 2nd reference derives Eq 10 but is cavalier about the normalisation of $p(x)$, which would give the wrong ML solution.

[6] W. Press *et al.*, Numerical Recipes in C++, Cambridge 2002, §13.3. JTL thanks Prof. Igor Volobouev for pointing out the resemblance.