JOURNAL OF COMPUTATIONAL PHYSICS 72, 202–219 (1987)

Event Classification Using Weighting Methods

ROGER BARLOW

Department of Physics, Manchester University, Manchester, England

Received August 13, 1985; revised December 29, 1986

This paper considers the general method of estimating the numbers of different classes of events in a sample by use of a weighting technique, with particular reference to high energy physics experiments, and shows how to construct the optimal weight function to do this in any situation. Results using this function are always better than the imposition of a cut and can be as good as a maximum-likelihood technique. Various useful formulae are given.

1. INTRODUCTION

The problem of assigning members of a sample to different species, sometimes referred to as taxonomy, is common to many disciplines and was considered long ago by Fisher [1]. In high energy physics experiments it often arises in a form where one wishes to estimate the numbers of members of various species in a sample even though the classification of individual members ("events") is ambiguous; one then speaks of "statistical separation" rather than "separation on an event by event basis." This has lately become relevant in the study of the properties of b quarks produced in e^+e^- annihilation [2, 3, 4], and will be discussed in this context here, though the techniques can be applied to other separation problems such as particle identification and separating quark and gluon properties. Generally events are divided into two classes, the wanted "signal" and the unwanted "background."

This is illustrated in Fig. 1, which is typical of many figures found in experimental papers. The value of some discriminator variable x has been histogrammed for all events, and the signal emerges as a clear peak standing out from a sloping background. In such a situation, the traditional approach is to impose a cut at some value of x, say at $x = 0.6$, and the number of signal events is given by the number that survive the cut, after (small) corrections for the loss of signal events that lie outside the cut, and the contamination of background events that survive.

Unfortunately, for b quark signals and many other cases, the situation is more like that shown in Fig. 2; again the signal and background are visible, but they merge into each other. Any cut imposed must remove an appreciable number of signal events and/or include an appreciable number of background events; isolating a reasonably pure signal sample is impossible without unacceptably savage selection cuts.

0021-9991/87 $3.00

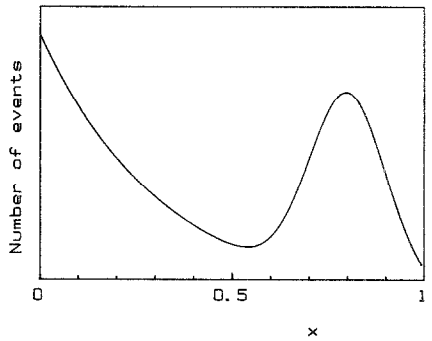

FIG. 1. A typical HEP figure. Signal events stand out from a background when some variable x is histogrammed, and can be separated by some cut in x.

Application of a cut in such a case seems undesirable, as it entails throwing away a large number of wanted events—data just outside the cut contain a sizable proportion of signal events—but accepting the data just inside, which contain a sizable amount of unwanted background. A more satisfactory approach is to apply some weight function $w(x)$ which is large in the regions where the signal is high, small in the regions where there is little signal, and of intermediate size in the region where both signal and background are important.

This note offers a complete prescription for reconstituting signals from the data, using such weighting techniques. Faced with a distribution like that of Fig. 2, from which the signal contribution is to be extracted, one would want to know:

What are the results of using a particular weight function?
What is the best weight function to use? How good is it?
Where is the best place to put a cut?
Which gives better results, the best weight or the best cut?
Is there any better technique?

This paper answers all these questions.

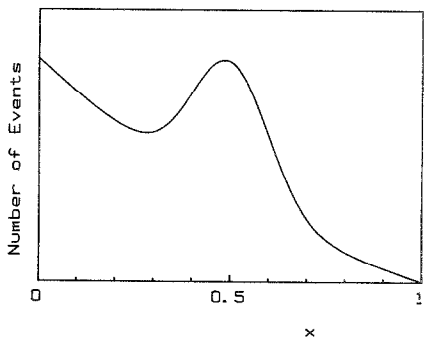

FIG. 2. A situation similar to Fig. 1, but a clean separation is not possible.

2. Estimation Using Weights

Consider the typical data shown in Fig. 2; x is a *discriminator* variable whose behaviour is well understood for both signal and background, though not in itself of any interest. Note that x can and probably will be a vector containing more than one variable (e.g., p_t and jet mass for b quarks [2]); it is presented here as a single variable for the sake of simplicity, but this need not be so. We suppose that the data in the plot consist of N_S signal and N_B background events, and that the problem is to estimate the expected value of N_S on the basis of the information in the plot. The figure might actually contain all the data taken, and N_S would then be the total number of signal events and could be used to give a cross section, or it could contain events which have already been histogrammed under some more interesting variable (e.g., for b quarks, the impact parameter or the polar angle [3, 4]) and consist of the contents of one bin of such a histogram.

Suppose a weight function $w(x)$ is chosen which enhances the signal. $w(x)=x$ obviously serves this purpose, as does $w(x)=(x-0.5)^2$; the latter looks to be more effective, but as yet we have no way of quantifying this. There are N_U events in the histogram, and when each is multiplied by its appropriate w, the weighted total is N_W. The x distributions for signal and background are presumed to be well understood, so the mean value of the weight for signal events, $\overline{w_S}$, and for background events, $\overline{w_B}$ can be calculated.

For a given N_S and N_B, the expected values for the unweighted and weighted totals are given by

$$N_U = N_S + N_B$$
$$N_W = \overline{w_S} N_S + \overline{w_B} N_B.$$

So if the data give some values for N_U and N_W, the estimate for the desired number N_S is

$$\hat{N}_S = \frac{N_W - \overline{w_B} N_U}{\overline{w_S} - \overline{w_B}} \tag{1}$$

which can also be written as

$$\frac{\sum_j (w_j - \overline{w_B})}{\overline{w_S} - \overline{w_B}},$$

where the sum is over all events in the sample.

$\hat{N}_S$ is unbiased, as can be seen by considering the expectation for

$$N_S - \frac{N_W - \overline{w_B} N_U}{\overline{w_S} - \overline{w_B}}$$

which is zero.

3. Statistical Errors from Using Weights

In finding the variance for the estimate of Eq. (1), one has to consider carefully the actual conditions of the experiment and what one is seeking to estimate. Different arrangements can give different variances, though the data they record are identical. We consider three such types of estimation: for (i) the actual number, (ii) the proportion, and (iii) the Poisson mean. To give an example, suppose that in 1 day an experiment collects 100 events, and when analysed the number of b quark events, $\hat{N}_S$, is 10. One can then say that (i) there are 10 b quark events in this sample, or (ii) 10 % of events contain b quarks, or (iii) 10 b quark events/day are produced under these conditions. All of these are valid, though an error should be quoted on the 3 measured quantities (10 events, 10 %, 10 events/day) and this error, the variance of the estimate $\hat{N}_S$, is different in the 3 cases.

(i) *The actual number.* If the numbers of signal and background events which one wishes to estimate are regarded as fixed, then the expected variance on $\sum_j w_j$ is due to the variances of w for the N_S signal and N_B background events

$$V\left(\sum_j w_j\right) = N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)$$

and the variance on the estimate for N_S is

$$V(\hat{N}_S) = \frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)}{(\overline{w_S} - \overline{w_B})^2}.$$

(ii) *The proportion.* Alternatively the total number of events may be fixed, but the signal and background numbers occur randomly, according to the binomial distribution. One is then measuring the proportion of signal events in the mixture. The variance in N_S is larger than for the first case, being given by the variance of the distribution of w for the whole sample, and this gives

$$V\left(\sum_j w_j\right) = N_U(\overline{w^2} - \bar{w}^2) = N_S\overline{w_S^2} + N_B\overline{w_B^2} - \frac{(N_S\overline{w_S} + N_B\overline{w_B})^2}{N_U}$$

$$V(\hat{N}_S) = \frac{N_S N_B}{N_U} + \frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)}{(\overline{w_S} - \overline{w_B})^2}.$$

(iii) *The Poisson mean.* Another possible experimental arrangement is to take data for a fixed period of time, collecting events which are generated randomly, according to Poisson distributions with some values for the mean numbers N_S and N_B, and it is these mean numbers which are to be estimated. Due to the additional uncertainty, the variance is again larger than that in the previous case. For a set of data generated according to the Poisson distribution, it can be shown

(see the Appendix) that the variance on the estimate of a weighted sum is the sum of the squares of the weights, so the variance of the estimate of N_S is given by

$$V(\hat{N}_S) = \frac{\sum_j (w_j - \overline{w_B})^2}{(\overline{w_S} - \overline{w_B})^2}. \tag{2}$$

In a particular situation the expected value of this, with a little rearranging, is

$$V(\hat{N}_S) = N_S + \frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)}{(\overline{w_S} - \overline{w_B})^2} \tag{3}$$

which we can rewrite as

$$V(\hat{N}_S) = N_S + X$$

with

$$X = \frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)}{(\overline{w_S} - \overline{w_B})^2}. \tag{4}$$

The difference between these three arrangements and their associated variances can be seen by considering the case of clear-cut separation, when the weight functions for signal and background are sharply peaked, so $\overline{w_S^2} - \overline{w_S}^2$ and $\overline{w_B^2} - \overline{w_B}^2$ are both zero and X vanishes. The variance for the fixed-number arrangement is zero, because you know exactly how many of your sample are signal. The variance in the second viewpoint is $N_S N_B / N_U$, which is the variance of the binomial distribution with proportions N_S/N_U and N_B/N_U. The variance for the fixed-time, Poisson, arrangement is N_S, because although you know the number in the sample exactly, this is still only an estimate of the mean of the Poisson distribution which generated it, and the error associated with a Poisson distribution is $\sqrt{N}$.

The third, Poisson, case describes the experimental arrangement appropriate for high energy physics experiments, which take data for some period of time, during which events are accumulated on a random, Poisson, basis. For this reason the rest of this paper will work in the framework of the fixed-time arrangement, and the variance given by Eq. (3). However, in all three cases the expression X given by Eq. (4) contains the excess on the variance due to the presence of a background which cannot be cleanly separated, and the problem of minimising the variance is the same: the "best" weight function is that which minimises X. It should be a function for which the signal and background each have a sharply peaked distribution, so that $\overline{w^2} \approx \bar{w}^2$, and also that these peaks are widely separated, so $\overline{w_S} \not\approx \overline{w_B}$.

Another useful number is the ratio:

$$\mathbf{F} = \sqrt{1 + ((\overline{w_S^2} - \overline{w_S}^2) + \mathbf{A}(\overline{w_B^2} - \overline{w_B}^2))/(\overline{w_B} - \overline{w_S})^2}, \tag{5}$$

where **A** is the noise/signal ratio N_B/N_S. **F** is the factor which multiplies $\sqrt{N_S}$ to give errors. If it is 1 then the separation is complete. Otherwise the effectiveness of the separation can be judged by how close **F** is to 1. Although this note does not address the problem of the choice of the discriminator variable(s) x (this has been considered first by Fisher [1] and recently, with specific applications to b quarks, by Marshall [2]), it does tell one how to use the variable(s) once chosen. The number **F**—Eq. (5)—provides a clear way of expressing how effective the choice of variable is and can be used to compare the merits of alternatives and express how well the separation is achieved.

Signal–Background Correlations

The values of the unweighted and weighted totals can also be combined to give the number of background events, N_B,

$$\hat{N}_B = \frac{N_W - \overline{w_S} N_U}{\overline{w_B} - \overline{w_S}}.$$

The variance of this is

$$V(\hat{N}_B) = N_B + \frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)}{(\overline{w_S} - \overline{w_B})^2}$$

$$= N_B + X$$

as is apparent from the symmetry of the algebra. If both N_S and N_B are evaluated then there is a correlation between the two results. This is negative and is given by

$$\operatorname{cov}(\hat{N}_S, \hat{N}_B) = -\frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B(\overline{w_B^2} - \overline{w_B}^2)}{(\overline{w_S} - \overline{w_B})^2}$$

$$= -X$$

This can be proved directly, by the same method as used in the Appendix to find the variance or by observing that the variance of N_U is given by

$$V(\hat{N}_U) = V(\hat{N}_S + \hat{N}_B) = V(\hat{N}_S) + V(\hat{N}_B) + 2\operatorname{cov}(\hat{N}_S, \hat{N}_B).$$

As the events are generated by a random (Poisson) mechanism, this variance on the total number of events, N_U, must be equal to N_U. The covariance term must therefore be such as to exactly cancel the second terms in the expressions for $V(\hat{N}_S)$ and $V(\hat{N}_B)$.

4. The Optimal Weight Function

An optimal weight function is a function $w(x)$ for which $V(\hat{N}_S)$ is minimum, given N_S, N_B, and the signal and background distributions in the discriminator

variable x. Denote these by $s(x)$ and $b(x)$—they are normalised to unity and are known, perhaps from theory (for example, if one is studying a resonance one might know that $s(x)$ is a Breit–Wigner with known parameters) or perhaps from the results of Monte Carlo programs.

In terms of the functions $s(x)$, $b(x)$, and $w(x)$, Eq. (4) becomes

$$X=\frac{N_S(\int w^2 s\,dx-(\int ws\,dx)^2)+N_B(\int w^2 b\,dx-(\int wb\,dx)^2)}{(\int ws\,dx-\int wb\,dx)^2}.$$

If $w(x)$ is to be optimal then X must be a minimum, i.e., if an arbitrary small function $\delta(x)$ is added to $w(x)$, the change in X must be zero.

The change resulting in, for example, $\overline{w_S}$ is

$$\Delta(\overline{w_S})=\int (w(x)+\delta(x))\,s(x)\,dx-\int w(x)\,s(x)\,dx=\int s(x)\,\delta(x)\,dx.$$

The change in the denominator of the above expression for X is

$$2(\overline{w_S}-\overline{w_B})\left(\int s\delta\,dx-\int b\delta\,dx\right)=2(\overline{w_S}-\overline{w_B})\int (s-b)\,\delta\,dx.$$

The change in the numerator is

$$N_S\left(\int 2ws\delta\,dx-2\overline{w_S}\int s\delta\,dx\right)+N_B\left(\int 2wb\delta\,dx-2\overline{w_B}\int b\delta\,dx\right)$$

which is

$$2\int (N_S(w-\overline{w_S})\,s+N_B(w-\overline{w_B})\,b)\,\delta\,dx.$$

Combining these two expressions to give the change in X, setting it to zero, and eliminating some common factors give the requirement

$$\int (\overline{w_S}-\overline{w_B})(N_S(w-\overline{w_S})\,s+N_B(w-\overline{w_B})\,b)\,\delta\,dx$$

$$=\int (N_S(\overline{w_S^2}-\overline{w_S}^2)+N_B(\overline{w_B^2}-\overline{w_B}^2))(s-b)\,\delta\,dx.$$

As it is required that this be true for an arbitrary function $\delta(x)$, the functions multiplying δ in the two integrands must be equal for all x,

$$(\overline{w_S}-\overline{w_B})(N_S(w-\overline{w_S})\,s+N_B(w-\overline{w_B})\,b)$$
$$=(N_S(\overline{w_S^2}-\overline{w_S}^2)+N_B(\overline{w_B^2}-\overline{w_B}^2))(s-b).$$

Rearranging this gives the requirement for $w(x)$:

$$\begin{aligned}(\overline{w_S}-\overline{w_B})(N_S s(x)+N_B b(x))\, w(x) \\ = ((N_S\overline{w_S^2}+N_B\overline{w_B^2})-\overline{w_B}\,(N_B\overline{w_B}+N_S\overline{w_S}))\, s(x) \\ -((N_S\overline{w_S^2}+N_B\overline{w_B^2})-\overline{w_S}\,(N_B\overline{w_B}+N_S\overline{w_S}))\, b(x). \end{aligned} \tag{6}$$

From this it can be seen that $w(x)$ has to have the form

$$w(x)=\frac{Ps(x)-Qb(x)}{N_S s(x)+N_B b(x)},$$

where the constants P and Q can be determined by inserting this form for $w(x)$ into the above requirement. When this is done the result is very simple; writing various quantities in terms of P and Q, and remembering $\int s(x)\,dx=\int b(x)\,dx=1$

$$N_B\overline{w_B}+N_S\overline{w_S}=\int (N_B wb+N_S ws)\,dx=\int (N_S s+N_B b)\,w\,dx=P-Q$$

$$N_S\overline{w_S^2}+N_B\overline{w_B^2}=\int (N_S w^2 s+N_B w^2 b)\,dx=\int w(Ps-Qb)\,dx=P\overline{w_S}-Q\overline{w_B}.$$

The right-hand side of Eq. (6) becomes

$$(P\overline{w_S}-Q\overline{w_B}-(P-Q)\,\overline{w_B})\,s(x)-(P\overline{w_S}-Q\overline{w_B}-(P-Q)\,\overline{w_S})\,b(x)$$

which is

$$P(\overline{w_S}-\overline{w_B})\,s(x)-Q(\overline{w_S}-\overline{w_B})\,b(x)$$

and this is equal to the left-hand side for any values of P and Q; the requirement is satisfied automatically (unless $P=-N_S Q/N_B$, when $\overline{w_S}=\overline{w_B}$ and X is infinite). This arbitrariness in $w(x)$ is to be expected. If any weight function is multiplied by a constant, or has a constant added to it, then this will not change its effect.

We choose to take $P=N_S$, and $Q=0$. This gives the optimal weight function

$$w(x)=\frac{N_S s(x)}{N_S s(x)+N_B b(x)}. \tag{7}$$

Choosing these values for P and Q gives $w(x)$ a nice physical interpretation: all weights lie between 0 and 1, and for an event in the data at a given value of x, it is the probability that that event originated from the signal rather than the background sample. It is zero in regions where $s(x)$ is zero—i.e., events that are definitely not signal have weight zero—and one if $b(x)$ is zero—events that are certainly signal have weight one.

If these optimal weights are used then various simple useful relations are true:

$$N_S(1-\overline{w_S}) = N_B\overline{w_B}$$
$$N_B\overline{w_B^2} + N_S\overline{w_S^2} = N_S\overline{w_S} \tag{8}$$
$$V(\hat{N}_S) = N_S\frac{1-\overline{w_B}}{\overline{w_S}-\overline{w_B}}.$$

Furthermore, having minimised $V(\hat{N}_S)$ by minimising the quantity X, one has also minimised $V(\hat{N}_B)$ and $\text{cov}(\hat{N}_S, \hat{N}_B)$. These weights are therefore optimal in this respect also.

Nearly Optimal Weights

The problem with the above recipe for optimal weights is that it involves N_S and N_B, which are unknown. However, this is not a fatal difficulty. Equation (7) can be rewritten as

$$w(x) = \frac{s(x)}{s(x)+\mathbf{A}b(x)}, \tag{9}$$

and the ratio $\mathbf{A}$ $(=N_B/N_S)$ is not known exactly. However, one probably has a reasonably good initial estimate; in a particular problem the relative size of the signal expected is usually roughly known to within a factor of 2 or so, and this is adequate. Provided the value of $\mathbf{A}$ used in the weight function is not very different from the true N_B: N_S ratio, $\mathbf{F}$ and $V(\hat{N}_S)$ are not significantly different from their ideal values.

This slow dependence can be seen as the second derivative at the optimum point is

$$\frac{d^2(\mathbf{F}^2)}{dA^2} = 2\,\frac{N_S(\overline{w_B^3}(\overline{w_B}-\overline{w_S}) - \overline{w_B^2}\,\overline{w_S} + \overline{w_B^2}(\overline{w_B}+\overline{w_S^2}-\overline{w_B^2}))}{N_B(\overline{w_S}-\overline{w_B})^3}.$$

(The first derivative is of course zero.) If the separation is working at all then $\overline{w_B}$ and $\overline{w_B^2}$ must be small and very small compared to $\overline{w_S}$, so this expression is at least *small*^{2} compared to $\mathbf{F}^2$.

The effects of this are shown in Fig. 3 for the data of the type of Fig. 2 which, in fact, consists of a Gaussian signal of mean 0.5 and standard deviation 0.1, on top of a background falling exponentially with slope 1 to zero at $x=1$. The value of $\mathbf{F}$ is shown as a function of the value of $\mathbf{A}$ used in the weight function, for various actual N_B: N_S ratios. It can be seen that if the guessed ratio is correct to within about a factor of 10, then the increase in $\mathbf{F}$ is insignificant.

It should be stressed that if the weight function does not use the correct value for N_B/N_S this does not invalidate the result obtained. This is given by Eq. (1), which is true for *any* weight function $w(x)$. The actual error will be given by Eq. (2), which is also correct. The only penalty paid for a wrong guess is that the error is (slightly) larger than it need be.

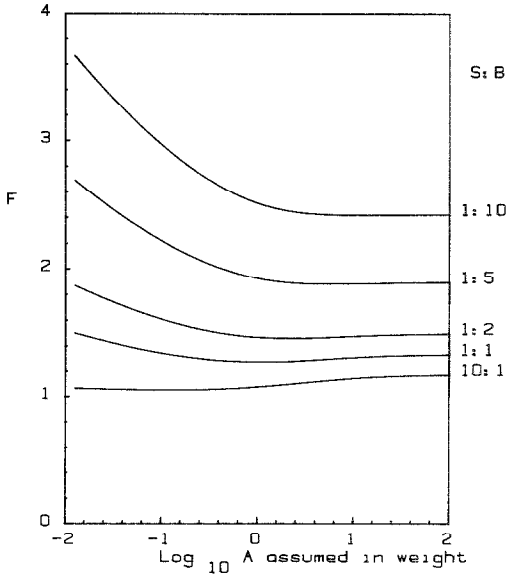

FIG. 3. The dependence of **F** on the value of **A** used in the weight function, for signal and background mixtures of the form of Fig. 2, for various signal: background ratios.

5. COMPARISON WITH OTHER METHODS

Comparison with the Use of Cuts: Optimal Cuts

The method of cuts can be considered as a weight function. If a cut is applied at x_c then one counts the number of events within the cut, this is equivalent to calculating N_W with

$$w(x) = 0 \qquad (x < x_c)$$
$$= 1 \qquad (x \geqslant x_c).$$

The effects of a cut can be conveniently described by the fraction of signal and background events which remain after it, C_S and C_B. Then $\overline{w_S} = \overline{w_S^2} = C_S$ and $\overline{w_B} = \overline{w_B^2} = C_B$. The estimate of N_S obtained from the number in the cut by correcting for loss of signal and inclusion of background is just Eq. (1), and the error is given by Eq. (3), which becomes

$$V(\hat{N}_S) = N_S + \frac{N_S C_S(1 - C_S) + N_B C_B(1 - C_B)}{(C_S - C_B)^2}. \tag{10}$$

The optimal cut is that which maximises this quantity. Again, reasonable guesses at N_S and N_B will give good results.

As such a weight distribution is not the optimal one, the error from this cut must be larger than the error from the optimal weights. However, a cut is simpler to use, and some would say safer, and if the benefit to be obtained from using weights is only small a cut may be preferred. This can be investigated by a simple analysis of the given problem. Figure 4 shows the results obtained from distributions of the gaussian on an exponential type of Fig. 2. The value of **F** is shown as a function of N_B: N_S for optimal weights and optimal cuts, and it can be seen that the weighting method offers a considerable advantage over the cuts, particularly when the signal is relatively small. In Fig. 4a both use the correct value of N_B/N_S for **A** and in Fig. 4b a value of 1.0 is used; the difference is small, as expected.

Comparison with Maximum-Likelihood Fitting

In this method $\hat{N}_S$ and $\hat{N}_B$ are found by maximising the likelihood:

$$\mathscr{L} = \sum_j \ln(N_S s(x_j) + N_B b(x_j)) - \int (N_S s(x) + N_B b(x))\, dx,$$

where the sum is taken over all events. The integral is included in accordance with the principle of extended maximum likelihood, [5] and ensures that the sum of the fitted values, $\hat{N}_S + \hat{N}_B$, is the total number of events. (It contains the log likelihood for all the values of x at which no event occurred.)

For a large number of events, the variance is given by the inverse of the matrix of expectation values of second derivatives [6]

$$V(\hat{N}_S) = E\left(\frac{\partial^2 \mathscr{L}}{\partial N_B{}^2}\right) \Big/ \left(E\left(\frac{\partial^2 \mathscr{L}}{\partial N_B{}^2}\right) E\left(\frac{\partial^2 \mathscr{L}}{\partial N_S{}^2}\right) - \left(E\left(\frac{\partial^2 \mathscr{L}}{\partial N_B\, \partial N_S}\right)\right)^2\right), \qquad (11)$$

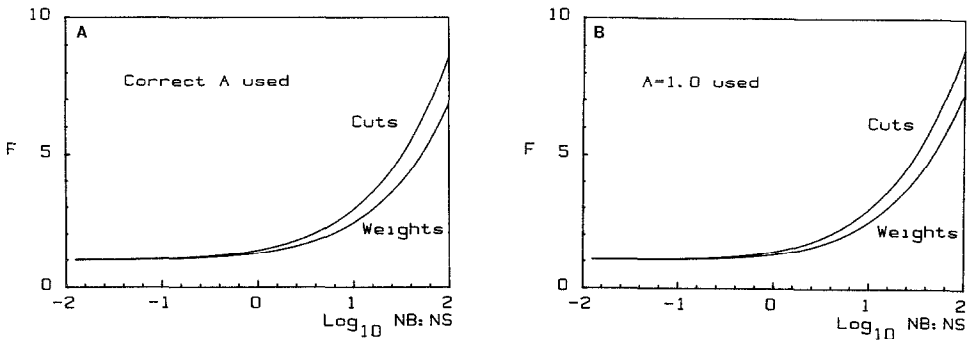

FIG. 4. (A) Dependence of **F** on the actual signal: background ratio for situations like that of Fig. 2, showing the **F** achieved by the optimal weight and by the optimal cut, assuming that the value of **A** used is the true one. (B) As Fig. 4A, except that the value of **A** used is 1.0 throughout.

where

$$E\left(\frac{\partial^2 \mathscr{L}}{\partial N_S{}^2}\right) = \int \frac{s(x)^2}{N_S s(x) + N_B b(x)}\, dx$$

$$E\left(\frac{\partial^2 \mathscr{L}}{\partial N_S\, \partial N_B}\right) = \int \frac{b(x)\, s(x)}{N_S s(x) + N_B b(x)}\, dx$$

$$E\left(\frac{\partial^2 \mathscr{L}}{\partial N_B{}^2}\right) = \int \frac{b(x)^2}{N_S s(x) + N_B b(x)}\, dx.$$

If now a function $w(x)$ is defined according to Eq. (7), then Eq. (11) can be simplified as follows: the first of the three expressions can be written

$$\int \frac{s(x)^2}{N_S s(x) + N_B b(x)}\, dx = \frac{1}{N_S} \int w(x)\, s(x)\, dx = \frac{1}{N_S}\, \overline{w_S},$$

remembering that $s(x)$ was introduced in Section 4 as the normalised x distribution for signal events. In the same way, the second expression is

$$\int \frac{s(x)\, b(x)}{N_S s(x) + N_B b(x)}\, dx = \frac{1}{N_S} \int w(x)\, b(x)\, dx = \frac{1}{N_S}\, \overline{w_B}.$$

The third expression is simplified by considering the normalisation condition for $b(x)$,

$$1 = \int b(x)\, dx = \int \frac{N_S s(x) + N_B b(x)}{N_S s(x) + N_B b(x)}\, b(x)\, dx$$

$$= \int w(x)\, b(x)\, dx + \int \frac{N_B b(x)^2}{N_S s(x) + N_B b(x)}\, dx$$

$$1 = \overline{w_B} + N_B \int \frac{b(x)^2}{N_S s(x) + N_B b(x)}\, dx$$

so the third expression is $(1 - \overline{w_B})/N_B$.

Making these 3 replacements in Eq. (11) gives

$$V(\hat{N}_S) = \frac{N_S{}^2 (1 - \overline{w_B})}{N_S \overline{w_S}(1 - \overline{w_B}) - N_B \overline{w_B}{}^2}.$$

This can be further simplified by the substitution for $N_B \overline{w_B}$ in the denominator

$$N_B \overline{w_B} = N_S (1 - \overline{w_S})$$

(which is obtainable by treating the normalisation condition for $s(x)$ in the same way as that for $b(x)$ was treated above). This gives

$$V(\hat{N}_S) = N_S \frac{1 - \overline{w_B}}{\overline{w_S} - \overline{w_B}}.$$

This holds for any N_S, N_B, and normalised $s(x)$ and $b(x)$, where $w(x)$ as given by Eq. (7) is a function formed from these quantities. The dependence of the result on the signal and background distributions is contained in their average values for w. If this function is also being used as a weight function, then the variance on the estimate of N_S is given by Eq. (8), and the two expressions are seen to be identical.

So for large numbers of events, the *maximum likelihood* (ML) and *optimal weighting* methods are equally effective, and either can be used. Only if a good guess at the N_B: N_S ratio is impossible will the ML method have a significantly better error. If such a guess is possible, then the power of the two methods is the same, and the weighting method is much easier to use—values and errors are given by simple formulae ((1) and (2), or (8)) and no fitting is needed. There are also advantages in the weighting method when it comes to determining the Monte Carlo distributions—a problem which is not discussed here, but can be quite tricky, particularly if x in fact consists of many variables ("many" ≡ more than 1). For the weighting method one needs to parametrise the ratio $w(x)$; for maximum likelihood one parametrises $s(x)$ and $b(x)$. The function $w(x)$ is smoother than $s(x)$ and $b(x)$ because these latter contain various lumps and bumps due to kinematic and other factors. Also, if the parametrisation of $w(x)$ is inaccurate then the validity of the method, as expressed in Eqs. (1) and (2), is not affected; the only penalty is that the weights are not quite optimal, whereas inaccuracies in the $s(x)$ and $b(x)$ functions may have serious systematic effects on the maximum likelihood fit.

For small numbers of events the variance on the maximum likelihood estimator is no longer given by the asymptotic form of Eq. (11). It is suggested [7, p. 214] that things are worse for small samples than the asymptotic theory indicates; if this is so then the optimal weights will perform better than the maximum likelihood fitting for small samples. Also, for small samples the bias in the result of the maximum likelihood fit is not negligible, and the weighting method may be preferred for this reason.

Use of More than One Weight Function

One might consider the use of two or more independent weight functions but, surprisingly, this does not give any better result. This can be seen as follows.

Use of extra weight functions can only be beneficial if they are independent, and the best one could possibly do would be to use an (infinite) set of independent weight functions forming a basis for functions of x—e.g., $\{1, x, x^2, x^3, \ldots\}$ or the Chebyshev polynomials. All such infinite basis sets must be equivalent for our pur-

pose as they can be transformed into each other. So we can consider only one, namely the set of delta functions

$$w(x; X_i) = \delta(x - X_i),$$

where $\{X_i\}$ is the (infinite) set of all points in the range of x. Each such function involves only one point, X_i, at which the number of events will be either one or zero. One therefore has to adjust $\hat{N}_S$ and $\hat{N}_B$ to satisfy, as best one can, the equations

$$\hat{N}_S s(X_i) + \hat{N}_B b(X_i) = 1 \qquad \text{for all } X_i \text{ with an event}$$
$$\hat{N}_S s(X_i) + \hat{N}_B b(X_i) = 0 \qquad \text{for all } X_i \text{ with no event.}$$

This is precisely the problem solved by the *method of extended maximum likelihood* [5]. Thus the use of extra weight functions leads to a method equivalent to the method of maximum likelihood and cannot offer a better result.

6. Further Considerations

The Case of a Known Background

In the situation described above it is assumed that there is no *a priori* knowledge of the size of the background. Indeed the terms "signal" and "background" are purely descriptive, it is a matter of choice which of the two species is regarded as which.

However, the case can arise that although the signal size is (of course) unknown, the expected size of the number of background events is known (and the actual number in the experiment is given by the appropriate Poisson distribution). For example, Fig. 2 could represent the production of a particle, with a background for which the cross section is known from theory or from other measurements.

Using weights, one again writes

$$N_W = \overline{w_S} N_S + \overline{w_B} N_B,$$

with the difference being that this time, N_B is the known expected number of background events. The equivalent of Eq. (1) is then

$$\hat{N}_S = \frac{N_W - \overline{w_B} N_B}{\overline{w_S}} \tag{12}$$

and Eq. (3) becomes

$$V(\hat{N}_S) = N_S + \frac{N_S(\overline{w_S^2} - \overline{w_S}^2) + N_B \overline{w_B^2}}{\overline{w_S}^2} \tag{13}$$

A similar analysis to that used for the variance given by Eq. (3) shows that this is also a minimum when the weight function defined by Eq. (7) is used. So the weight function which was optimal in the previous case is also optimal for this one.

The variance obtained is, however, different. It is smaller as more information is given. When the optimal weight function is used, Eq. (13) becomes, in analogy with Eq. (8),

$$V(\hat{N}_S) = N_S/\overline{w_S} \tag{14}$$

$$\mathbf{F} = 1/\sqrt{\overline{w_S}}. \tag{15}$$

The maximum-likelihood fit, for which there is now only one free parameter, gives the same error.

So if the expected background is known *a priori*, then the optimal weight function is still given by Eq. (7), and the method of weights is as good as a maximum-likelihood fit in this case also.

The Weighting Method for Several Species

It may be that the sample contains not merely 2 species but several, separable on the basis of their different distributions in x. If there are n such species, then their separation requires n independent weight functions, $w^{(i)}$, $i = 1,..., n$. If the weighted totals from the sample are $W^{(i)}$, then these are produced by the numbers of the various species present, $N^{(i)}$, according to

$$W^{(i)} = \sum_j M_{ij} N^{(j)},$$

where M_{ij} is the mean value of weight i for events belonging to species j (a generalisation of $\overline{w_S}$ and $\overline{w_B}$). Given the $W^{(i)}$ from the data and the elements of M_{ij} from previous study of pure samples, the numbers present belonging to each species are estimated by

$$\hat{N}^{(i)} = \sum_j M_{ij}^{-1} W^{(j)} = \sum_k \sum_j M_{ij}^{-1} w_k^{(j)},$$

where the sum over k is over all events, and $w_k^{(j)}$ is the value of weight j for the kth event. Equation (1) can be seen to be a special case of this, where the 2 weight functions are the function $w(x)$ discussed there and the constant weight of unity.

The variance is again given by the sum of the squares of the elements in the summation

$$V(\hat{N}^{(i)}) = \sum_k \left(\sum_j M_{ij}^{-1} w_k^{(j)} \right)^2.$$

The optimal weight functions are, by analogy with Eq. (5),

$$w^{(i)}(x) = \frac{N^{(i)} s^{(i)}(x)}{\sum_j N^{(j)} s^{(j)}(x)}$$

as each $w^{(i)}(x)$ is the optimal weight function for the ith species.

Weighting and the Method of Moments

The estimator used in the weighting method—Eq. (1)—can be regarded as an application of the method of moments: the numbers of species in the sample are found by comparing the zeroth and first moments of the function $w(x)$ for the sample with those of the pure signal and background. The use of the method of moments, of maximum likelihood, and other techniques, for the separation of species has been considered by many authors—for a review, see [7]—albeit not in a form readily accessible to most experimental physicists. Various comparisons of the method of moments and the maximum likelihood method have been made, however, the context of their conclusions has to be carefully considered. Thus, when Tubbs and Coberly [8, p. 1120] conclude that the method of moments is inferior to the other methods they study, it should be noted that (i) they consider only simulation of a few cases of samples, (ii) they consider only samples generated by the Normal, Gaussian distribution, (iii) the moments they use are the first- and second-order moments of the discriminator variable x—there is no attempt to find a better function of which to take the moments, and finally (iv) their criterion for judging the merits of the methods is not the size of the variance, but the sensitivity of the estimate to systematic changes in the species distribution functions. Similarly, the comparison of Odell and Basu [9, p. 1105], which prefers the maximum likelihood method to that of moments (i) refers to the method applied to the first 4 moments of the discriminator variable, not those of an optimised function, (ii) considers only mixtures of two normal distributions, and (iii) concerns the case where the individual species distributions are not known and have also to be obtained from the data. Thus our conclusion that the method of weights (or moments) is as good as the maximum likelihood method, for separation in the circumstances described here (which are those appropriate to high energy physics experiments), is not in conflict with statements to the contrary which can be found in the literature but apply to different circumstances and usually only to moments of the discriminator variable.

Using the Method: A Brief Note on Systematic Errors

In applying the method of weights to a data sample, the process falls into two parts: first, the choice of the weight function $w(x)$; and second, its application to the data in order to extract the number of signal events.

As has been discussed, some weight functions are better than others, the best being that given by Eq. (7); this choice requires an assumed value for the ratio N_S: N_B, but does not depend critically on it. It also requires parametrisation of the

x dependence of the signal and background distributions or, combining the two requirements, a parametrisation of the fraction of events that belong to the signal as a function of x.

In the second stage of the process the given function is applied to the data, and Eq. (1) is used to extract the desired number of signal events. In this extraction, $\overline{w_S}$ and $\overline{w_B}$ are used. These numbers, the mean weights for a pure signal and a pure background sample, are obtained by applying the chosen weight function to such samples, which are typically obtained from Monte Carlo simulations. Although it is essential that these values are accurate, all that is necessary for this is that the samples used are faithful representations of the real signal and real background data. No assumptions about $N_S: N_B$, and no parametrisations, are necessary. The estimate of N_S is unbiassed provided $\overline{w_S}$ and $\overline{w_B}$ are the correct mean weights for the weight function being used, whether or not this weight function is optimal.

Use of Monte Carlo programs to calculate the values of $\overline{w_S}$ and $\overline{w_B}$ is sometimes used as a ground for criticism for this approach, as being too dependent on Monte Carlo programs because any inaccuracies will give systematic effects. Against this criticism two points can be made. First, the variable x is unexciting and well understood, and the Monte Carlo programs or theoretical models used are presumably well established and known to describe the data accurately. Second, in a situation like this the results of any cut require large corrections to account for the signal lost by the cut and/or the background remaining in the selected sample. These corrections are also heavily dependent on the Monte Carlo programs used. So the use of cuts will also introduce large systematic errors if the Monte Carlo programs are inaccurate, just as with this method.

7. Conclusions

The weighting method for extracting signals, as embodied in (1), (2), and (7) of this paper, is easy to understand and use, superior to cuts as no data is wasted by throwing it away, and as good as the more complicated maximum-likelihood fitting technique. It should find many useful applications in the analysis of high energy physics data, and perhaps elsewhere.

APPENDIX: Variance of a Weighted Sum of Values Generated by the Poisson Distribution

This is a simple derivation, but does not appear in most textbooks which treat the variance of a weighted sum or mean only in the context of Guassian errors. Suppose a number of events are recorded in a certain time, each of which has some weight w, and the total $\sum w_i$ is formed. A practical example might be a compound radiactive source containing several sources with differing strengths and emitting

gamma rays of various energies which are (precisely) measured. It is desired to find the total energy deposited, W, and the associated error.

The best estimate of the total is clearly

$$\sum_{\text{all events}} w_i,$$

where w_i is the measured gamma ray energy of the ith event.

To find the variance on this, divide the gamma ray energies into bins, the jth bin having energy w_j and expected number of events λ_j. Then the sum can be rewritten

$$W = \sum_j w_j n_j,$$

where this sum runs over all bins. The number of events in each bin, n_j, are all independent, and the variance of W is the sum of their variances, appropriately weighted

$$V(W) = \sum_j w_j^2 V(n_j).$$

The n_j are generated by the Poisson distribution, so their variances are given by $V(n_j) = \lambda_j$; and as n_j is an unbiased estimate of λ_j, the estimate of the variance of W is

$$V(W) = \sum_j w_j^2 n_j$$

which can be rewritten in a form which makes no reference to the binning used and is, therefore, true in general:

$$V(W) = \sum_{\text{all events}} w_i^2.$$

REFERENCES

1 R. A. FISHER, *Ann. Eugen.* **79**, 179 (1936).
2. R. MARSHALL, *Z. Phys. C* **26**, 291 (1984).
3. W. BARTEL *et al.*, *Phys. Lett. B* **146**, 437 (1984).
4. W. BARTEL *et al.*, DESY Report 86-001 (1986); *Z. Phys. C* **31**, 349 (1986).
5. J. OREAR, University of California Report UCLR 68-8417, 1968 (unpublished).
6. A. G FRODESEN *et al.*, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Bergen–Oslo–Tromsø, 1979), p. 208, Eq. 9.25. M. G. KENDALL AND A. STUART, *The Advanced Theory of Statistics*, Vol. 2 (Griffin, London, 1961), p. 55, Eq. 18.60.
7. R. A. REDNER AND H. F. WALKER, *SIAM Rev.* **26**, 195 (1984).
8. J. D. TUBBS AND W. A. COBERLY, *Comm. Statist. Theor. Meth. A* **5**, 1115 (1976).
9. P. L. ODELL AND J. P BASU, *Comm. Statist. Theor. Meth. A* **5**, 1091 (1976).