FITTING THEORY TO DATA IN THE PRESENCE OF BACKGROUND UNCERTAINTIES

BYRON ROE

ABSTRACT. When fitting theory to data in the presence of background uncertainties, the question of whether the spectral shape of the background happens to be similar to that of the theoretical model of physical interest has not generally been considered previously. These correlations in shape are considered in the present note and found to make important corrections to the calculations. The discussion is phrased in terms of $\chi^2$ fits, but the general considerations apply to any fits. If these new correlations are not included, the distribution usually does not have a $\chi^2$ behavior, the $\chi^2$ probabilities obtained are overestimated, and the confidence regions will be incorrect. Fake data studies, as used at present, will not be optimum. Problems will also occur in comparisons of related $\chi^2$, such as occur in the Maltoni-Schwetz theorem. Neutrino oscillations are used as examples, but the problems discussed here are general ones.

1. INTRODUCTION

When fitting theory to data in the presence of background uncertainties, the question of whether the spectral shape of the background is similar to that of the theoretical model of physical interest has not generally been considered previously. These correlations in shape are considered in the present note and found to make important corrections to the calculations.

Some causal correlations between background and the theoretical model are usually included at present. For example, beam normalization uncertainties affect both the theoretical model and the background. There are also some correlations when the size of the theoretical model affects the size of the background.

However, when the theoretical model parameters are allowed to vary in a fit, there is a new, qualitatively different, correlation that must be considered if the error matrix includes the effects of uncertainties in the backgrounds.

This correlation is not a causal correlation, but occurs simply due to a similarity between the spectrum of the theoretical model over the bins and the spectrum of the backgrounds. This similarity can be quantified using a regression analysis. These correlations should, perhaps, be distinguished by being called “regression correlations”.

1
If the corrections for these regression correlations are not included, the experimental \( \chi^2 \) often will not have a \( \chi^2 \) distribution. The best fit point will be incorrect and confidence regions will be incorrect. Even fake data studies will not give optimum results if these corrections are not included and the "effective number" of degrees of freedom obtained from fake data studies will be affected. Correcting for this new correlation makes the experimental \( \chi^2 \) distribution be more like the \( \chi^2 \) distribution with number of degrees of freedom equal to the number of bins minus the number of parameters fit.

The toy model considered in Section 3 describes the essence of the problem considered in this note.

2. Notation

Suppose data have been obtained for a histogram with \( N_{\text{bins}} \). The \( i \)th bin has \( N_{\text{data}}(i) \) events. The model used for fitting the data has background parameters and theory parameters.

Let \( B_j(i) \) be the true expected value of background \( j \) in bin \( i \). Let \( N_{\text{bkrd}} \) be the number of background parameters associated with these backgrounds. For any given data set, the number of background events will have a statistical spread with mean value \( B_j(i) \). It is assumed that the backgrounds have been evaluated in independent experiments previously, giving estimates for the mean value of each \( (B_m)_j(i) \) and the covariance matrix of the error in the estimated mean value \( \text{cov}(B_m)_j \). There are correlations from bin to bin and also correlations between the various backgrounds, so that the covariance matrix is not, in general, diagonal. The total estimated background expectation is \( (B_m)_{\text{tot}}(i) \).

For the initial discussion, it is assumed that there are a string of separate backgrounds, each with one parameter; the total background is then

\[
(B_m)_{\text{tot}}(i) = \sum_{j=1}^{N_{\text{bkrd}}} (B_m)_j(i). \tag{1}
\]

Each background has the form \( B_j(i) = b_j g_j(i) \), where \( i = 1, \cdots, N_{\text{bins}}, \) and \( j = 1, \cdots, N_{\text{bkrd}}. \) The \( b_j \) correspond to scale parameters defining the level of background. The bin dependence of each of the backgrounds \( g_j(i) \) is assumed known and fixed. In Section 5, the form of the backgrounds will be generalized.

The second kind of parameters are parameters of the theoretical model of physical interest and are initially unknown. There are \( N_t \) theory parameters, \( t_k, \ k = 1, \cdots, N_t. \) The theory model prediction for each bin, \( N_{\text{th}}(i), \) is a function of these parameters.

Define the "signal" as the data minus the background estimates. This is the part of the data that will be fitted to the theoretical model of interest.

\[
N_{\text{sig}}(i) \equiv N_{\text{data}}(i) - (B_m)_{\text{tot}}(i). \tag{2}
\]

Two common \( \chi^2 \) fitting methods are considered. For the first fitting method, the parameters assigned to the background(s) are held fixed in the fit at their measured values. The uncertainty in the background parameters enters into the covariance used in the fit.
The covariance matrix of the signal in bins $i, j$ is given as

$$ \text{cov}_{\text{sig}}(i, j) = \text{cov}_{\text{data}}(i, j) + \text{cov}(B_m)_{\text{tot}}(i, j). $$

$\text{cov}_{\text{data}}(i, j)$ is diagonal and is the variance expected for the data in each bin given the theory plus background prediction for the mean number of events in that bin.

The fit done is to minimize the $\chi^2$,

$$ \chi^2 = \sum_{i,j=1}^{N_{\text{bins}}} [N_{\text{sig}}(i) - N_{\text{th}}(i)][\text{cov}_{\text{sig}}^{-1}(i, j)][N_{\text{sig}}(j) - N_{\text{th}}(j)], $$

(4)

where $[\text{cov}_{\text{sig}}^{-1}(i, j)]$ is the inverse of the signal covariance matrix.

A second approach uses a constrained fit to account for backgrounds, constraining the backgrounds to be at their measured values within errors, when performing fits. The $\chi^2$ is written as a $\chi^2$ ignoring the background uncertainties and a sum of terms ("pulls") for each background. For simplicity, assume for the initial discussion that there are no correlations between backgrounds and that each background depends on a single variable. Let $\beta_k(i) = \partial(B_{\text{tot}}(i) + N_{\text{th}}(i)) \bigg|_{\delta_k = 0}$, where $\delta_k$ is the deviation of the true $k$th background or nuisance parameter from its measured value. The set of $b_k$ parameters is usually enlarged for this fit method to include nuisance parameters which can affect both the theory and the backgrounds, such as the overall normalization of the data. However, in this note the $\delta_k$ will refer only to the background parameters. The experimental quantities are $N_{\text{data}}(i)$ and $(B_m)_k(i)$, the latter obtained from previous experiments. $\beta_k(i)$ is a fixed quantity obtained from the spectrum of the backgrounds. (For the initial assumption that $g_k(i)$ is known, $\beta_k(i) = g_k(i)$.)

$N_{\text{th}}$ and the $\delta_k$ are quantities to be fitted. Let $\sigma^2_{b_k}$ be the variance of the estimate of background parameter $b_k$.

$$ \chi^2 = \sum_{i=1}^{N_{\text{bins}}} \frac{(N_{\text{data}}(i) - (N_{\text{th}}(i) + \sum_k ((B_m)_k(i) + \beta_k(i)\delta_k)))^2}{\sigma^2_{\text{data}}(i)} + \sum_k \frac{\delta^2_k}{\sigma^2_{b_k}}. $$

In the presence of correlations between backgrounds, the usual introduction of the covariance matrix is used. Since each pull term introduces an effective extra bin and an extra parameter to be fit, the number of degrees of freedom is unaffected. If one minimizes Equation 5 over the $\delta_k$, this approach reduces to the previous one [2].

$\chi^2$ fits are often done in an iterative fashion taking one step at a time looking for the minimum. The correlations are re-evaluated at each step.

3. A TOY MODEL

Suppose one is fitting a theoretical model parameter to data, where the theoretical model is of the form $N_{\text{th}}(i) = tf(i)$, where $t$ is a constant to be fit and $f(i)$ is a known function of the bin number $i$. Suppose, further, that there is a single background which is uncorrelated with the model parameter. $B(i) = bg(i)$ with $g(i)$ known. Let $b_{\text{m}}$ be the estimated mean of $b$ and $\sigma^2_{b_{\text{m}}}$ be the estimated variance of $b_{\text{m}}$. The fit is done by using the method of
Equations 3 and 4. The covariance of the background is $cov_{B_m}(i, j) = g(i)g(j)\sigma_{B_m}^2$. The overall covariance, $cov_{overall}$, is the sum of the statistical variance of the data, $\sigma_{data}^2$ and the covariance of the background, $cov_{overall}(i, j) = \sigma_{data}^2(i)\delta_{ij} + cov_{B_m}(i, j)$.

$$\chi^2 = \sum_{i=1}^{N_{bins}} [N_{data}(i) - b_mg(i) - tf(i)]\sigma_{overall}^{-1}(i, j)[N_{data}(j) - b_mg(j) - tf(j)]$$

where $cov_{overall}^{-1}(i, j)$ is the inverse of the covariance matrix. Let $t_0$ be the result of that fit. The uncertainty is assigned to the $t$ parameter by the usual $\Delta \chi^2 = \chi_{th(t)} - \chi_{best\ fit}$ method.

Now suppose that the background has the form $bf(i)$, where $b$ is a constant and $f(i)$ is the same known function that appeared for the theoretical model parameter. The model and background are now completely correlated, that is, they have the same shape as a function of bin number and differ only in normalization. The minimum $\chi^2$ is then independent of whatever value had been estimated for $b_m$, since if $b_m$ is changed, the fitted value of $t$ will change to compensate, because $f(i)$ is a common factor for theoretical model and background.

Now write the square bracketed terms in the numerator for the $\chi^2$ fit, using $b$ not $b_m$, as $[N_{data}(i) - (t + b)f(i)] = [N_{data}(i) - zf(i)]$, where $z \equiv t + b$. The background experiment has previously obtained an estimated mean and uncertainty for the parameter $b$. The fit here obtains a mean and uncertainty for the parameter $z = t + b$.

Assume, incorrectly, that the error matrix remains the same as for the uncorrelated fit. If $b$ is estimated by $(b_m)$, the new fit for $t$ is the same as the old one, $t_{fit} = t_0$.

However, the covariance matrix is not the same as it was for the previous fit. For this new fit using $z$, the background is part of the parameter $z$ being fitted. There is no mention of $b$; only $z$ appears in the fit. The $z$ fit can be done perfectly well if there is no prior knowledge of $b$. The $cov_{B_m}(i, j)$ is not included in the error matrix of the $\chi^2$ terms.

$$\chi^2 = \sum_{i=1}^{N_{bins}} \frac{[N_{data}(i) - zf(i)]^2}{\sigma_{data}^2}.$$
background error is sufficiently large, a good fit can be obtained even when the form of the data differs very considerably from $f(i)$.

This problem is seen more clearly using the constrained fit method of Section 2, the “pulls” method fit.

$$\chi^2 = \sum_{i=1}^{N_{\text{bins}}} \frac{[N_{\text{data}}(i) - (t + b_m + \delta_b)f(i)]^2}{\sigma_{\text{data}}^2(i)} + \frac{\delta_b^2}{\sigma_b^2},$$

since $\beta(i) = f(i)$ here. If there is complete correlation, the pull term will become zero in the fit, regardless of the size of the actual $\delta_b$. Any difference between $b$ and the estimated $b_m$ will be taken up by changing $t$ and leaving $\delta_b$ zero. This means that, effectively, one bin has been lost and the number of degrees of freedom is reduced by one. The probability obtained using the nominal degrees of freedom is then too high, just as seen for the first fitting method.

The odd results with both of the $t, b$ fits have the same root cause, which is seen most easily in the constrained fit method. The problem is that $t$ and $b$ are correlated parameters in the fit appearing only as $t + b$; $z$ is the only fitting parameter. The fitting methods discussed in Section 2 allow correlations between background parameters. The constrained fit method even, as mentioned in Section 2, allows causal correlations between background and signal parameters such as occur in overall normalization of the data. However, it cannot handle a non-causal accidental correlation between a background parameter and a theory parameter due solely to a similarity between the shape of the background and theory functions over the histogram bins, i.e., by a regression correlation. Fitting $t$ and $b$ without accounting for this correlation produces incorrect results. The fit with $z$ avoids this problem. The $z$ fit is a straightforward $\chi^2$ fit with the usual $\chi^2$ properties. As indicated above, the fit determines the estimate for $z = t + b$ and the uncertainties of that estimate, while the preliminary background experiment has obtained the same quantities for $b$. It is then completely straightforward to combine the two measurements to get an estimate for the value of $t$ and the uncertainties of that estimate.

The subtlety here is that both $t$ and $b$ are involved in the fit. If a fixed value of $t$ is used and there is no fit, the question of correlation between $t$ and $b$ does not enter and, for the first method, the $\chi^2$ distribution would be correct if $\text{cov}_{B_m}(i, j)$ is included.

The same correlation correction is also needed for the likelihood method. For this discussion, as a shorthand, $t, b$, and $z$ are taken to be the random variables for these parameters. The variables can be changed from $t, b$ to the independent variables $z, b$ with a Jacobian of one. Then the probability of the $b$ term can be integrated out and, as with the $\chi^2$ fits, only $L(\text{data}|z)$ remains

$$L(\text{data}|t, b)L(b) = L(\text{data}|t + b)L(b) = L(\text{data}|z)L(b),$$

$$\int L(\text{data}|z)L(b)db = L(\text{data}|z).$$

Another method to treat the correlation problem would be to make a combined fit for the background and the model parameters using the combined histograms for both the present
experiment and the background experiments. However, if there are many backgrounds, often determined by varied methods, the “curse of dimensionality” makes this impractical.

Consider the neutrino oscillation fit, used in MiniBooNE. In this experiment a beam of muon neutrinos is produced and sent to a detector. In the flight from production to interaction in the detector some muon neutrinos may have changed into electron neutrinos. The experiment looks for these electron neutrinos, and tries to determine the number of electron neutrinos obtained and their energy spectrum. The theoretical model for this process involves the product (not sum) of two parameters, one a function of $\Delta m^2$ and the other is $\sin^2 2\theta$. $\Delta m^2$ determines the shape of the energy spectrum and $\sin^2 2\theta$ is a scale parameter determining the size of the effect. A major background comes from muon neutrino interactions in the detector producing $\pi^0$, which decay into two photons and may make the event appear to be an electron neutrino event. For the MiniBooNE experiment, does the $\pi^0$ background resemble the results expected from the theoretical model? $\pi^0$ background numbers for the neutrino exposure are obtained from Table 6.5 of the MiniBooNE Technical Note 194 [3]. The neutrino spectrum and the neutrino quasi-elastic cross sections were read from Figures 2 and 5 from a MiniBooNE neutrino elastic scattering Physical Review article [4]. Results are shown in Figure 1. For $\Delta m^2 = 2$ or 1 eV$^2$, the mean and $\sigma$ of the $\pi^0$ background and the theoretical expectations are quite close and the correlations are large.

4. Partial correlations

In practice, regression correlations are almost never complete; it is necessary to consider partial correlations between the theoretical model and background. In this section a simple model will be considered. The model then will be generalized in the following section. The theory model here will be taken as $N_{th}(i) = t f(i)$, where $f(i)$ is a known function and $t$ is the fitting variable. There may be several backgrounds, but it is assumed here that only background $B_k$ has a significant correlation with the theoretical model shape. The background is taken as $B_k = b_k g_k(i)$, where $g_k(i)$ is a known function and $b_k$ was measured in a previous experiment with some uncertainty. Let $M_f$ and $\sigma_f$ be the mean value and spread over the histogram bins for $f(i)$ and $M_{g_k}$ and $\sigma_{g_k}$ be the mean value and spread of $g_k(i)$. The means and $\sigma$’s here refer to mean values and spreads over the bins of the histogram. They do not refer to measurement uncertainties. Let $f'(i) \equiv f(i) - M_f$ and $g'_k(i) \equiv g_k(i) - M_{g_k}$.

Start by considering the regression correlation between $g'_k(i)$ and $f'(i)$. Consider a plot of the background $g'_k(i)$ (y-axis) versus the theoretical model $f'(i)$ (x-axis), where the points are the values for each bin. A straight line regression fit of background on theoretical model is made. Let $x_{\text{temp}}^*(i) = a f'(i)$, where $a$ is a constant chosen to minimize $\sum_{i=1}^{N_{\text{bins}}} [g'_k(i) - x_{\text{temp}}^*(i)]^2$. The line $x_{\text{temp}}^*(i)$ represents the completely correlated part of $g'(i)$ in this linear approximation. The result is

$$\frac{x_{\text{temp}}^*(i)}{\sigma_{g_k}} = \rho \frac{f'(i)}{\sigma_f},$$

(9)
Figure 1. The top figure shows the energy spectrum expected for the $\pi^0$ background. The horizontal axis is the energy in GeV obtained if the $\pi^0$ events are mistakenly reconstructed as charged current quasi-elastic electron neutrino events. The mean value of the events/bin is 19.5 events, the standard deviation is 14.0 events and the ratio mean/standard deviation is 1.39. The dashed green line is the part of the $\pi^0$ background that is correlated with the theory if $\Delta m^2 = 2$ eV$^2$. The dotted blue line is the part of the $\pi^0$ background that is correlated with the theory if $\Delta m^2 = 1$ eV$^2$. The middle figure shows the energy spectrum expected for the neutrino oscillation signal if $\Delta m^2 = 2$ eV$^2$. The mean value of events/bin is 18.3, the standard deviation is 14.8, the ratio mean/standard deviation is 1.24, and the correlation with the $\pi^0$ spectrum is 0.718. The bottom figure shows the energy spectrum expected for the neutrino oscillation signal if $\Delta m^2 = 1$ eV$^2$. The mean value of events/bin is 13.0, the standard deviation is 11.8, the ratio mean/standard deviation is 1.1 and the correlation with the $\pi^0$ spectrum is 0.771.
where the regression correlation coefficient is given by

$$\rho = \frac{1}{N_{\text{bins}}} \sum_{i=1}^{N_{\text{bins}}} g'(i)f'(i)}{\sigma_g \sigma_f},$$

and

$$\sigma_f = \left[ \frac{1}{N_{\text{bins}}} \sum_{i=1}^{N_{\text{bins}}} (f'(i))^2 \right].$$

$\sigma_g$ is obtained similarly. This result is derived in [5] and is discussed in [6].

As seen in Equation 9, this straight line fit $x^{*\text{temp}}$ is linearly related to the theoretical model bin dependence $f(i)$. For a given theoretical model there is no uncertainty in the values of $M_f$ and $\sigma_f$. With the present assumption that $g(i)$ is a known function, there is no uncertainty in the values of $\rho$, $M_{gk}$ and the value of $\sigma_{gk}$. Note that Equations 9 and 10 are normalization independent because of the divisions by $\sigma$.

$$x^{*\text{temp}}(i) = \frac{\rho \sigma_{gk} f'(i)}{\sigma_f} = \rho_{\text{eff}} f'(i),$$

where $\rho_{\text{eff}} \equiv \rho \sigma_{gk} / \sigma_f$ is defined as the “effective” correlation coefficient.

$x^{*\text{temp}}(i)$ is appropriate for $f', g'$ not for the full $f, g$ which include the mean values. It does not approach the appropriate limit for full correlation for $f, g$. It is necessary to translate along the $f'$ axis. The size of the translation is uniquely set by asking that, including mean values, appropriate results are obtained for $\rho = 0$ and $\rho = 1$. The proper translation is to plot the points at $f' + \rho_{\text{eff}} M_f$. The translated $x^{*}$ is then

$$x^{*}(i) = x^{*\text{temp}}(i) + \rho_{\text{eff}} M_f = \rho_{\text{eff}} f'(i) + \rho_{\text{eff}} M_f = \rho_{\text{eff}} f(i).$$

Since $x^{*}(i)$ is proportional to $f(i)$, the regression correlation of $x^{*}$ with $f$ is one. It can easily be shown that the regression correlation of the residuals, $g(i) - x^{*}(i)$ is zero. $b_k x^{*}(i)$ can be considered to be the correlated part of the full background $k$ and the distances $b_k [g_k(i) - x^{*}(i)]$ will be taken to represent the uncorrelated part of the background. This division will be assumed to give a reasonable first order division of the backgrounds. Let $\Delta_k(i)$ be the uncorrelated part of background $k$.

$$\Delta_k(i) = b_k [g_k(i) - x^{*}(i)] = b_k [g_k(i) - \rho_{\text{eff}} f'(i)] + [M_g - \rho_{\text{eff}} M_f].$$

$\Delta_k(i)$ does go to appropriate limits for $\rho = 0$ and $\rho = 1$. If $\rho = 0$, then $\Delta_k(i) = b_k g_k(i)$ as expected. For $\rho = 1$, then $g'(i)$ is proportional to $f'(i)$, and $\rho_{\text{eff}} = 1 \times (\sigma_{gk} / \sigma_f)$ is the constant of proportionality. $\Delta_k(i) = b_k [M_g - (\sigma_{gk} / \sigma_f) M_f]$. For the toy model considered in the previous section, $g(i) = f(i)$, the means are the same, $\sigma_{gk} / \sigma_f = 1$ and $\Delta_k(i) = 0$, as expected. If the toy model is extended to be $g(i) = f(i) + C$, where $C$ is a constant, then the above expression is again correct.

Note that it can turn out that $\rho < 0$. It still will be the proper value to minimize the sum of the squares of the distances of the background from $x^{*}$. Depending on the mean values $M_f$ and $M_g$, the correlated part of background $k$ can have some bins with a value more than the experimental background points. This should be expected. Even if the means
were the same, when the fit to the correlated straight line is made, some background points will be above and some below that line. This is not generally a problem for the method as $b_k x^*(i) + \Delta_k(i)$ must equal the entire background $B_k(i)$.

The covariance matrix of $\Delta_k$ due to the uncertainty in the experimental values obtained for background from the experiments determining the background is given by:

$$\text{cov}_{\Delta_k(i,j)} = [g_k(i) - \rho_{\text{eff}} f(i)][g_k(j) - \rho_{\text{eff}} f(j)] \sigma_{b_k}^2.$$  (15)

The fitting function is:

$$N_{\text{fit}}(i) = tf(i) + b_k x^*(i) =zf(i),$$  (16)

where $z = t + b_k \rho_{\text{eff}}$.

$N_{\text{sig}}$ in Section 2 had subtracted from data the estimates of all background. $N_{\text{sig-corr}}$ adds to $N_{\text{sig}}$ the estimate of the correlated part of $B_k$. The total background, $B_{\text{tot}}(i) = \sum_{j=1}^{N_{\text{bkd}}}[B_j(i)]$, should now not include all of $B_k(i)$, but should include only the uncorrelated part of it.

$$N_{\text{sig-corr}}(i) = N_{\text{data}}(i) - \sum_{j \neq k} (B_m)_j(i) - \Delta_k(i).$$  (17)

$$\text{cov}_{\text{corr}}(i,j) = \text{cov}_{\text{data}}(i,j) + \text{cov}_{(\Delta_k + \sum_{j \neq k} B_j)(i,j)}.$$  (18)

For the first fitting method in Section 2

$$\chi^2 = \sum_{i,j=1}^{N_{\text{bins}}} \left[ N_{\text{sig-corr}}(i) - zf(i) \right][\text{cov}_{\text{corr}}^{-1}(i,j)][N_{\text{sig-corr}}(j) - zf(j)].$$  (19)

Because the square bracket terms in the $\chi^2$ still add up to data minus the sum of theory plus background, the only change in the $\chi^2$ calculation due to regression correlations comes in the covariance for this method.

For the constrained fit method, $\beta_k(i) = \frac{\partial[\Delta_k(i) + \sum_{j \neq k} (B_m)_j(i)]}{\partial b_k} |_{\delta b_k = 0}$, where $\delta b_k$ is the deviation of background parameter $b_k$ from its measured value. Only the uncorrelated part of the background is used in calculating $\beta_k(i)$. For background $k$, $\beta_k(i) = g_k(i) - \rho_{\text{eff}} f(i)$.

$$\chi^2 = \sum_{i=1}^{N_{\text{bins}}} \frac{[N_{\text{sig-corr}}(i) - zf(i) - \sum_{t=1}^{N_{\text{bkd}}} \beta_k(i) \delta_t]^2}{\sigma_{\text{data}}^2} + \sum_{t=1}^{N_{\text{bkd}}} \frac{\delta_t^2}{\sigma_{\text{b_k}}^2}.$$  (20)

After the fit the estimate for $t$ is found from

$$t_{\text{fit}} = zf - \rho_{\text{eff}} b_k,$$

where, for the first method $(b_m)_k$ is used and for the constrained fit method the fitted value of $b_k$ is used. The covariance of $t$ is given by

$$\text{cov}_t = \text{cov}_{zf} + \text{cov}_{\rho_{\text{eff}} b_k}.$$  (22)
5. Generalization of the model

The correlations considered in the last section can vary with the values of the theoretical model parameters. The object of the fit is to minimize $\chi^2$ given the correlations relevant to the fit point. This means that the correlations must be adjusted to follow the fitting. Often the fitting of complex data is done by successive approximations, such as occur in the popular CERN program Minuit \[7\]. Sometimes, the fitting is just done by calculating $\chi^2$ at points along an $N_t$ dimensional lattice. However the fit is done, at each step or point, the correlations must be recalculated given the model parameters at that step or point.

The assumptions in the last section that the theory could be represented by $t f(i)$ and the background $k$ by $b_k g_k(i)$ with $f(i)$ and $g_k(i)$ known functions were taken to simplify the initial discussion. Often, theory and background have more general functions. Furthermore, most problems do not have a single correlated background parameter and often have more than one theoretical fitting parameter. It is necessary to include the effects of multiple backgrounds and multiple theoretical fitting parameters. Redo the regression analysis of the last section, using the full theoretical model and the total background instead of $f$ and $g$. Let $M_{B_{tot}}$ and $M_{th}$ be the means over the bins of the total background ($B_{tot}$) and the total theory model ($N_{th}$). The $\sigma$’s are defined in an analogous manner. Let:

$$B_{tot}'(i) = B_{tot}(i) - M_{B_{tot}}; \quad N_{th}'(i) = N_{th}(i) - M_{th}. \tag{23}$$

Equation 13 becomes:

$$\frac{x^*(i)}{\sigma_{B_{tot}}} = \frac{N_{th}(i)}{\sigma_{th}}. \tag{24}$$

$$\rho = \frac{1}{N_{\text{bins}}} \frac{\sum_{i=1}^{N_{\text{tot}}} B_{tot}'(i) N_{th}'(i)}{\sigma_{B_{tot}} \sigma_{th}}. \tag{25}$$

Define $\rho_{eff} \equiv \rho \sigma_{B_{tot}} / \sigma_{th}$.

$$x^*(i) = \rho_{eff} N_{th}(i). \tag{26}$$

For independent backgrounds, $x^*(i)$, defined above, is just the sum of the $x^*_k(i)$ for the individual backgrounds $B_k(i)$. The uncorrelated part of the background is

$$\Delta(i) = B_{tot}(i) - x^*(i) = B_{tot}(i) - \rho_{eff} N_{th}(i). \tag{27}$$

$$N_{\text{sig-corr}}(i) = N_{\text{data}}(i) - \Delta(i). \tag{28}$$

$$\text{cov}_{\text{corr}}(i, j) = \text{cov}_{\text{data}}(i, j) + \text{cov}_{\text{corr}}(i, j). \tag{29}$$

If the only uncertainty in the background is the overall normalization $b$ of the total background, then $\text{cov}_{\Delta}(i, j)$ is particularly simple. For a normalization change, $B_{tot}(i)$, $M_{B_{tot}}$, and $\rho_{eff}$ are all proportional to the normalization $b$. $\rho_{eff}$ is proportional because of the $\sigma_{B_{tot}}$ in the definition. Therefore $\delta(\Delta(i)) = (\delta b / b) \Delta(i)$, leading to $\text{cov}_{\Delta}(i, j) = \Delta(i) \Delta(j) \sigma_b^2 / b^2$. If there are a sum of independent backgrounds, each of which has only a normalization uncertainty,
one can use a simple generalization of this calculation since, as was noted above, one can just use the sum of the individual backgrounds for \( x^* \).

Remembering that the bin dependence of \( x^*(i) \) is proportional to that of \( N_{th}(i) \) the fitting function can be written

\[
N_{fit}(i) = N_{th}(i|\vec{t}) + x^*(i) = N_{th}(i|\vec{t})(1 + \rho_{eff}),
\]

where \( N_{th}(i|\vec{t}) = N_{th}(i|t_1, t_2, \ldots, t_{N_i}) \) is the prediction for the mean number of theory model events in bin \( i \), given the values of the theoretical parameters \( t_1, t_2, \ldots, t_{N_i} \). The fitting function should now be viewed as \( N_{fit} \), not just \( N_{th} \). In the last section the theory and background parameters were simple scaling functions. The regression correlation considered \( f'(i) \) and \( g_k(i) \) which were independent of the scaling. \( \rho_{eff} \) and \( M_g \) were not functions of the parameters. The parameters \( t \) and \( b_k \) appeared explicitly in \( \Delta \) and \( N_{fit} \). This simple separation cannot be assumed for the general problem considered here and the regression correlation was calculated using the full theory and full background. The effect of background parameters and their uncertainties now appears in \( N_{fit} \) through \( \rho_{eff} \).

If in the last section, the regression had been done with \( tf \) and \( bg \) instead of \( f \) and \( g \), then, because \( \sigma_{bg} = \sigma_{tg} \), \( \rho_{eff} \) would have been proportional to \( b \). The \( \chi^2 \) uncertainty obtained from \( \Delta \chi^2 \) refers to the uncertainty in \( N_{fit} \). The uncertainties in \( N_{th} \) are obtained by appropriately convoluting the uncertainties coming from the \( \chi^2 \) and the background uncertainties. This is the generalization of Equation 22.

For the first fitting method discussed in Section 2,

\[
\chi^2 = \sum_{i,j=1}^{N_{bins}} \left[ N_{sig-corr}(i) - N_{th}(i|\vec{t})(1 + \rho_{eff}) \right] (\text{cov}_{corr}(i,j)) \left[ N_{sig-corr}(j) - N_{th}(j|\vec{t})(1 + \rho_{eff}) \right].
\]

(31)

\[
\delta_k = \frac{\partial \Delta(i)}{\partial b_k} \big|_{\delta_k = 0}, \text{ where } \delta_k \text{ is the deviation of background parameter } b_k \text{ from its measured value.}
\]

As in the previous section, only the uncorrelated background is used to find \( \beta_k(i) \).

\[
\chi^2 = \sum_{i=1}^{N_{bins}} \left[ N_{sig-corr}(i) - N_{th}(i|\vec{t})(1 + \rho_{eff}) - \sum_{k=1}^{N_{bkrd}} \beta_k(i) \delta_k \right]^2
\]

\[
\sigma_{data}^2 + \sum_{k,\ell=1}^{N_{bkrd}} \delta_k [\text{cov}^{-1}(k, \ell)] \delta_\ell,
\]

(32)

where \([\text{cov}^{-1}(k, \ell)]\) is the inverse of the covariance matrix between \( \delta_k \) and \( \delta_\ell \).

6. Incorporating These Correlations in Practice

It was noted in Section 3, that the fits for neutrino oscillation in the MiniBooNE experiment are not simple sums. In terms of the notation from Section 4, for \( tf(i) \), \( t \) is determined by \( \sin^2 2\theta \) and \( f(i) \) by a function of \( \Delta m^2 \). The two parameters work together
to produce a single spectrum. For a background which matches the shape for a given $\Delta m^2$, any mis-estimate of the scale of the background will appear in the fitted value of $\sin^2 2\theta$ and the correlated part of the background should be associated with that parameter.

The individual terms in the sum of model terms will sometimes be these composite terms. This occurs, for example in fits of neutrino data for sterile neutrino hypotheses. There will sometimes also be more complicated dependences than the simple ones here and it is necessary to examine the situation for each particular experiment as indicated in the last section.

For determining confidence regions, consider a theoretical model point A. If the absolute $\chi^2$ is to be used and A is a fixed point, not the result of a fit, then the regression correlations should not be included in the calculation of $\chi^2$. However, if there is a fit regression correlations should be used. The $\Delta \chi^2$ method, $(\Delta \chi^2 \equiv \chi^2_A - \chi^2_{\text{best fit}})$ is a comparison of two values obtained in a fit. The $\chi^2$ at A using the regression correlations at A minus the best fit $\chi^2$ using the regression correlations at the best fit should then be used. It is important to note that the $\chi^2$ returns the probability of $N_{\text{fit}}$. To find a confidence region for the theoretical parameters, the set of $t$'s, it is necessary to convolute the $\chi^2$ probability with the probability for the background uncertainties. See the discussion following Equation 30.

If a fake data study is used, the procedure is similar. For these studies the experiment is replicated a number of times with Monte Carlo events. The theoretical model is held fixed. However, for each repetition, the backgrounds are varied randomly within their uncertainties. These studies are useful for estimating the expected variance of the result, for examining the effects of individual backgrounds, and for examining the sensitivity to the values of the uncertainties of these backgrounds. In the presence of the correlations discussed in this note, there is no change to the usual procedure for choosing Monte Carlo events. The regression correlations do not enter and only the usual backgrounds are randomly varied.

What happens next depends upon the question asked. Suppose there is no fit and the question asked is, “What is the distribution of $\chi^2$ if the model parameters are fixed?” For this question, the regression correlations do not enter.

However, suppose there is a fit, and it is desired to find the probability that a specific theoretical model A would give at least as high a $\chi^2$ as obtained in the real data. A number of fake data samples are generated for model A and, for each, the $\chi^2$ for A and the $\chi^2$ values for the best fit are found using the regression correlations. Then, if the $\Delta \chi^2 = \chi^2(A) - \chi^2(\text{best fit})$ method is used, the likelihood of that model is estimated by the fraction of times that the Monte Carlo $\Delta \chi^2$ is greater or equal to the $\Delta \chi^2$ value obtained for the data. If an absolute $\chi^2$ method is used, then the likelihood is the fraction of time that the Monte Carlo result is larger than the data $\chi^2$.

If the model and background correlations vary with the model point and the regression correlations are ignored, the best fit point will be different. Furthermore, confidence regions will not be optimum even for the fake data studies.
Consider an analogy. Suppose one estimates the distance between two objects on a somewhat blurry image, obtaining an estimated distance and the uncertainty of that estimate. If a better focussed image is available, a new estimate can be made. Both estimates are correct given the images each has used, but the estimate from the more focussed image can be considered a more optimum estimate.

At present, the use of an “effective number of degrees of freedom” is frequently employed to give corrections to the nominal number. The effective number of degrees of freedom is found by fitting the $\chi^2$ curve found from fake data studies. That correction is far less precise than the procedure studied here, which may well be a major part of the departure of the effective number of degrees of freedom from the actual one. With this new procedure the effective number should be closer to the real number. However, because the procedure described here is an approximation and because not all uncertainties are Gaussian distributed and for possible other problems, it may still be useful to include the new effective number as a residual correction.

7. The Maltoni-Schwetz theorem

Suppose one has two experiments, each measuring some of the same parameters of a given theory model, and wishes to see if the two sets of data are compatible, assuming the model is correct. Suppose one finds the best fits for the model parameters for data set 1, data set 2, and for the combined data sets 1 plus 2. Let the number of parameters fit for the three fits be $N_1$, $N_2$ and $N_{1+2}$. The Maltoni-Schwetz theorem [1] then states that

$$\chi^2_{MS} \equiv \chi^2_{1+2} - \chi^2_1 - \chi^2_2.$$  

(33)

is a measure of the compatibility of the data sets. If they are compatible, $\chi^2_{MS}$ has a $\chi^2$ distribution with $N_{MS} = N_1 + N_2 - N_{1+2}$ degrees of freedom.

Consider, as an example, the question of explaining some possible neutrino experiment anomalies as being due to the presence of two sterile neutrinos. The model for this hypothesis assumes several $\sin^2 2\theta$-like variables, several $\Delta m^2$-like variables and a $CP$ phase. The first data set corresponds to the appearance of $\nu_e$ events from an originally $\nu_\mu$ beam and the second data set is for the disappearance of $\nu_\mu$ events from an originally $\nu_\mu$ beam.

Fits have been made using both $\nu_e$ appearance and $\nu_\mu$ disappearance experiments [8] [9]. Both kinds of experiments are fit reasonably well with this model, but, using the Maltoni-Schwetz formalism, tension is found between the appearance and the disappearance experiments.

For the J. M. Conrad et al. fits [8], the number of fitted variables for each of the three data sets was $N_{app} = 5; \ N_{dis} = 6; \ \text{and} \ N_{comb} = 7$. This leads to $N_{MS} = 5 + 6 - 7 = 4$. For the disappearance experiments there is no $\pi^0$ background. There is a $\pi^0$ background for some of the appearance experiments including the MiniBooNE experiment which has a large weight within the appearance experiment sample.

If the correlations are not taken into account for the appearance experiments, the covariance matrix is wrong. The distribution will be too broad, NOT a $\chi^2$ distribution with the stated number of d.o.f., and usually not a $\chi^2$ distribution at all. If the $\pi^0$ is the dominant
correlated error, then for appearance experiments, if the correlation effect is omitted, the resulting minimum $\chi^2$ will be too small, and since the correlation changes with fit position, the minimum will not be at the minimum position obtained with the correct covariance matrix. The disappearance experiments do not have this background. For the combined fit, the appearance part has too small a $\chi^2$ and the disappearance the correct $\chi^2$, distorting the net result and again leading to a different minimum and an incorrect overall $\chi^2$.

In addition, because of the sensitivity of the appearance experiment to the $\pi^0$ background, an error in the estimate of that background can have a disproportionate effect. For the combined data fit, if the $\pi^0$ background is larger than the value estimated, some $\sin^2 2\theta$ type parameters will want to be bigger than they should be for the appearance bins, but not for the disappearance bins, giving some tension within the combined data set, i.e., increasing the $\chi^2$. Furthermore the number of degrees of freedom for MS is only four, which makes the discrepancy turn into an extremely low probability. The MS method is especially sensitive to these errors. Even without the correlations considered here, much of the tension between the appearance and disappearance results goes away if it is assumed that the MiniBooNE estimate of the $\pi^0$ background is low by $1.4 \sigma$.

8. Summary

Methods are given for using the $\chi^2$ method when regression correlations between the shapes of backgrounds and theoretical models being fitted occur. These methods are appropriate whenever these correlations exist.

(1) If the background shape correlations with the theoretical model are not taken into account, $\chi^2$ fits will not give correct results. The experimental $\chi^2$ usually will not have a $\chi^2$ distribution.

(2) Fake data studies without including these correlations will not be optimum. The use of “effective number of degrees of freedom” will help the situation, but will not be as precise as the methodology introduced here.

(3) One must use caution in applying the Maltoni-Schwetz theorem to find the compatibility of two sets of data to a theoretical model hypothesis. The theorem may incorrectly indicate incompatibility if there are correlations between the backgrounds and the theoretical model and/or if there are problems with the estimations of background variables.

9. Acknowledgements

I wish to acknowledge the considerable help of Xuming-He, the H. C. Carver Professor of Statistics, University of Michigan, during several discussions which helped clarify the problem and correct several errors. Michael Shaevitz, Professor of Physics, Columbia University, MIT graduate student Christina Ignarra and LANL Staff Member Zarko Pavlovic went through the note in some detail in order to apply it to MiniBooNE, and in the process, helped to find ambiguities and errors. I wish to thank MIT Professor J. M. Conrad for also going through the note and for pointing out to me the effect of changing $\pi^0$ background estimates on the Maltoni-Schwetz probabilities for a two sterile neutrino picture.
References

[1] M. Maltoni and T. Schwetz, “Testing the statistical compatibility of independent data sets”, Phys. Rev. D 68, 033020 (2003).

[2] D. Stump et al., Phys. Rev. D 65, 014012 (2001).

[3] “Measurement of the Rate of NC p0 Production at MiniBooNE”, Z. Djuric et al., MiniBooNE Technical Report 194, 8/8/06, revised 2/10/07.

[4] A.A. Aguilar-Arevalo et al., “First Measurement of the Muon Neutrino Charged Current Quasielastic Double Differential Cross Section”, arXiv:1002:2680 [hep-ex], Phys. Rev. D81, 092005 (2010).

[5] Harald Cramer, “Mathematical Methods of Statistics, Sixth U.S. Printing, Chapter 23”, Princeton University Press, Princeton, New Jersey (1954); first published in Uppsala Sweden, Almqvist & Wiksells (1945).

[6] Byron P. Roe, “Probability and Statistics in Experimental Physics, 2nd ed., Chapter 10”, Springer-Verlag, New York, Inc. (2001).

[7] Fred James and Matthias Winkler, “Minuit User’s Guide”, CERN, Geneva 23, Switzerland (2004).

[8] J. M. Conrad, C. M. Ignarra, G. Karagiorgi, M. H. Shaevitz, and J. Spitz, [arXiv:1207.4765] [hep-ex] (2012).

[9] J. Kopp, P. A. N. Machado, M. Maltoni, and T. Schwetz, arXiv:1303:3011 [hep-ph] (2013).

[10] Janet M. Conrad, William C. Louis, and Michael H. Shaevitz, Annu. Rev. Nucl. Part. Sci. 63 (2013) 45-67.

University of Michigan, Ann Arbor, MI 48105