Glen Cowan
RHUL Physics
30 June, 2009
Revised: 7 April, 2010

Increasing the flexibility of a distribution to allow for systematic uncertainty

1 Introduction

Consider fitting a distribution that is a mixture of several components, e.g., signal plus several sources of background, and suppose that the shape and normalization of some of the background components are taken from Monte Carlo in the form of MC templates. The goal of the analysis could be to estimate the parameters of the signal component, e.g., its amplitude and shape parameters.

Even if the MC templates have negligible statistical uncertainty, one would still like to incorporate into the fit the systematic error due to an imperfect MC model. In this note we discuss a method for doing this that involves modifying the MC templates in a way that depends on new parameters that are included in the fit. The correlations between the parameters result in general in increased statistical uncertainties in the original parameters of interest.

In the fit using the extended model, penalty terms are included so that specified properties of the altered distribution (e.g., normalization, mean, width) do not depart excessively from their nominal values relative to their assigned uncertainties. In this way, systematic uncertainties in the MC template are effectively propagated into the statistical uncertainties on the other parameters in the fit. Furthermore, the procedure significantly reduces potential biases in the fitted parameters of interest.

Section 2 describes the method for altering a distribution by introducing additional parameters, and in Section 3 the method is applied to an example fit.

2 Altering a distribution

Suppose a continuous random variable y follows a given probability density function (pdf) $g(y)$. We would like to alter the shape of $g(y)$ using a transformation containing a continuous parameter that determines the magnitude of the change.

As a preliminary step, let us assume that the variable y has been scaled and translated so that its minimum and maximum values are 0 and 1, i.e., one makes the transformation

$$y \to \frac{y - y_{\mathrm{min}}}{y_{\mathrm{max}} - y_{\mathrm{min}}} \,. \tag{1}$$

After carrying out the procedure described below, one simply transforms back to the original minimum and maximum values by inverting Eq. (1).

Now transform the variable y according to a specified function $x(y)$ which has a single-valued inverse $y(x)$. Then x will follow a new pdf $f(x)$, which can be found from

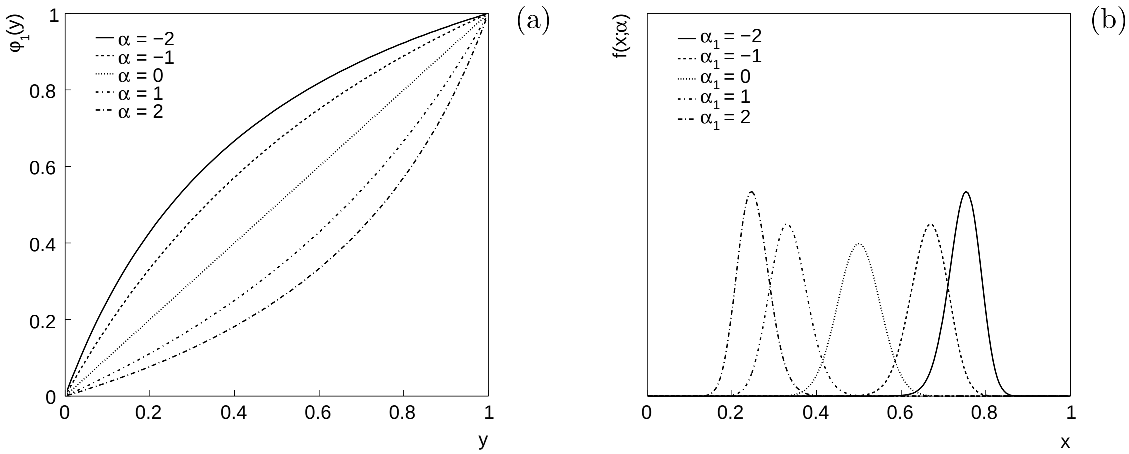

Figure 1: Plots of (a) the transformation function $x(y) = \varphi_1(y;\alpha)$ and (b) the effect of applying the transformation to a Gaussian pdf $g(y)$ for several values of the parameter α.

$$f(x) = g(y(x)) \left| \frac{dy}{dx} \right| \; . \tag{2}$$

By an appropriate choice of the transformation $x(y)$, we can, for example, shift the mean or width of the new pdf $f(x)$ relative to that of the original $g(y)$.

A simple choice for such a transformation that can shift the mean of a distribution is

$$x(y) \equiv \varphi_1(y;\alpha) = \begin{cases} \frac{y}{1+\alpha(1-y)} & \alpha \geq 0 \; , \\ \frac{(1-\alpha)y}{1-\alpha y} & \alpha < 0 \; . \end{cases} \tag{3}$$

If the parameter α is zero, the transformed variable is identical to the original; otherwise, the primary effect is to shift the pdf to the right or left. The function $x(y) = \varphi_1(y;\alpha)$ is shown in Fig. 1(a) for several values of α.

The inverse transformation $y(x)$ is

$$y(x) \equiv \psi_1(x;\alpha) = \varphi_1(x;-\alpha) = \begin{cases} \frac{(1+\alpha)x}{1+\alpha x} & \alpha \geq 0 \; , \\ \frac{x}{1-\alpha(1-x)} & \alpha < 0 \; . \end{cases} \tag{4}$$

To find the pdf $f(x)$ we need the derivative

$$\frac{dy}{dx} = \psi_1'(x;\alpha) = \begin{cases} \frac{1+\alpha}{(1+\alpha x)^2} & \alpha \geq 0 \; , \\ \frac{1-\alpha}{(1-\alpha(1-x))^2} & \alpha < 0 \; . \end{cases} \tag{5}$$

Applying the transformation to find $f(x)$ gives the pdfs shown in Fig. 1(b) for several values of the parameter α.

The procedure described above can easily be generalized to allow not only for a shift in the mean of a distribution, but also to stretch its width or introduce more complicated distortions. Suppose one takes for the mapping $x(y)$ the function

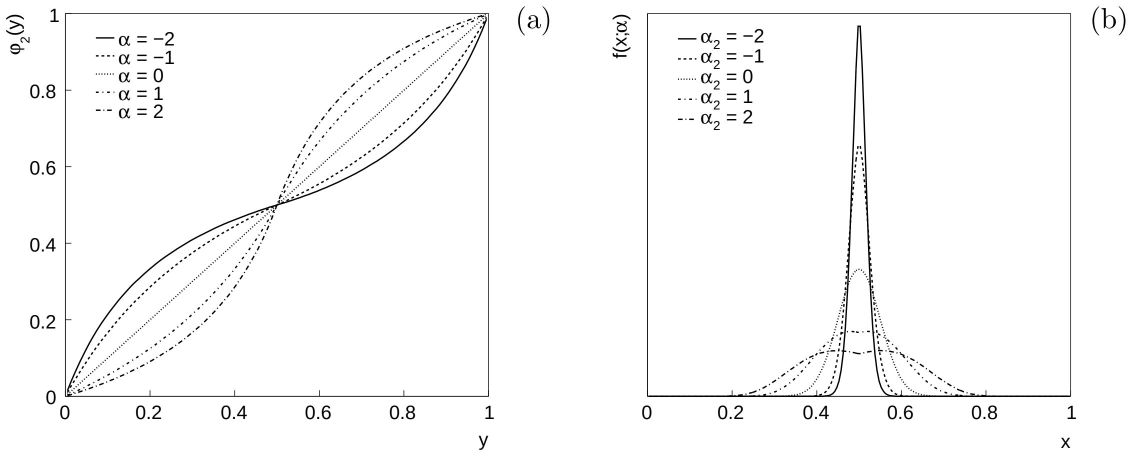

Figure 2: Plots of (a) the transformation function $x(y) = \varphi_2(y;\alpha)$ and (b) the effect of applying the transformation to a Gaussian pdf $g(y)$ for several values of the parameter α.

$$x(y;\alpha) \equiv \varphi_2(y;\alpha) = \begin{cases} \frac{1}{2}\varphi_1(2y;\alpha) & y \leq \frac{1}{2} \ , \\ \frac{1}{2}\left(1+\varphi_1(2y-1;-\alpha)\right) & y > \frac{1}{2} \ , \end{cases} \tag{6}$$

which has the inverse

$$y(x;\alpha) \equiv \psi_2(x;\alpha) = \varphi_2(x;-\alpha) = \begin{cases} \frac{1}{2}\psi_1(2x;\alpha) & x \leq \frac{1}{2} \ , \\ \frac{1}{2}\left(1+\psi_1(2x-1;-\alpha)\right) & x > \frac{1}{2} \ . \end{cases} \tag{7}$$

The function $x(y) = \varphi_2(y;\alpha)$ is shown in Fig. 2(a) and the effect of the transformation on a Gaussian distribution is shown in Fig. 2(b) for several values of α.

The transformations can be nested, so that one modifies both the original distribution's mean and width. Suppose, for example, that the original variable is z, which follows a pdf $h(z)$. This can be transformed to y using

$$y(z) = \varphi_1(z;\alpha_1) \ . \tag{8}$$

Then y can be transformed to x using

$$x(y) = \varphi_2(y;\alpha_2) \ . \tag{9}$$

The distribution of x is therefore

$$f(x) = h(z(y(x)))\left|\frac{dz}{dy}\frac{dy}{dx}\right| = h(z(y(x))\psi_1'(y;\alpha_1)\psi_2'(x;\alpha_2) \ . \tag{10}$$

where

$$z(y(x)) = \psi_1(\psi_2(x;\alpha_2);\alpha_1) \tag{11}$$

and

$$\psi_2'(x;\alpha) = \begin{cases} \psi_1'(2x;\alpha) & x \leq \frac{1}{2} \, , \\ \psi_1'(2x-1;-\alpha) & x > \frac{1}{2} \, . \end{cases} \tag{12}$$

It is straightforward to generalize the mapping $x(y)$ to include an arbitrary number of alternating elements each defined by $\varphi_1(y;\alpha)$.

In practice, the original distribution may be represented by a histogram. To apply the transformation described above however, one requires a continuous representation of the distribution. That is, even if one only needs the value of the transformed distribution for the variable x at the centre of a bin, the value $z(y(x))$ will not in general be at the centre. Furthermore one would like the changes introduced in the distribution to be continuous functions of the parameters. To circumvent the problems with binning, one can simply replace the histogram by a convenient continuous approximation such as a spline. One must ensure, however, that the distribution does not go negative, e.g., by assigning a value of zero if the spline is negative. In other cases it may be possible to first fit the histogram to a parametric distribution that by construction cannot go negative and to use this as the untransformed distribution.

3 An example fit

As an example of this procedure is shown in Figs. 3 and 4. The model consists of a Gaussian signal peak, a polynomial background, and a peaking background whose form is taken from an MC simulation, which here was also a Gaussian. The Gaussian signal was generated with a mean and standard deviation of $\mu_{\rm s} = 0.5$, $\sigma_{\rm s} = 0.1$, and the peaking background is generated with $\mu_{\rm b} = 0.5$, $\sigma_{\rm b} = 0.05$.

Suppose the goal of the analysis is to estimate the mean and width of the signal. In Fig. 3(a) this is done using a fitting model that is identical to the one used to generate the data. As expected, the fitted values of $\mu_{\rm s}$ and $\sigma_{\rm s}$ are close to the input values,

$$\hat{\mu}_{\rm s} = 0.50025 \pm 0.00232 \, , \tag{13}$$

$$\hat{\sigma}_{\rm s} = 0.10578 \pm 0.00325 \, , \tag{14}$$

and one finds $\chi^2 = 30.6$ with 44 degrees of freedom.

In Fig. 3(b), however, the MC template for the peaking background was systematically altered by using $\mu_{\rm b} = 0.45$, $\sigma_{\rm b} = 0.045$, compared to the values of 0.5 and 0.05 used to generate the data. Because of the incorrect modeling of the peaking background, the fitted values of $\mu_{\rm s}$ and $\sigma_{\rm s}$ for the signal peak come out to be significantly different from the true values, namely,

$$\hat{\mu}_{\rm s} = 0.51676 \pm 0.00226 \, , \tag{15}$$

$$\hat{\sigma}_{\rm s} = 0.08933 \pm 0.00308 \, . \tag{16}$$

Furthermore one has a poor goodness-of-fit, with $\chi^2 = 91.2$ for 44 degrees of freedom.

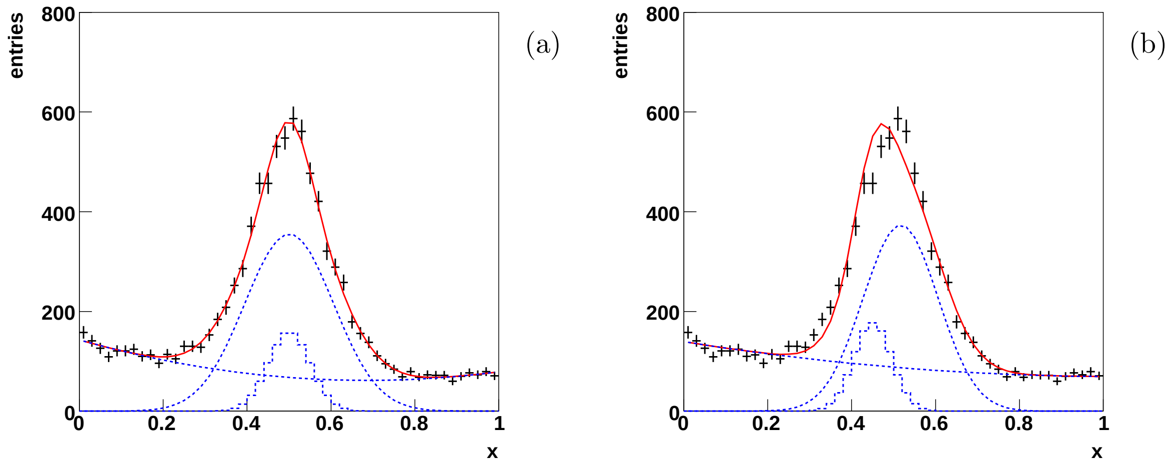

Figure 3: Plots of example fits using (a) the unaltered peaking background component (lower histogram) and (b) the altered version (see text).

Suppose now that one regards the peak position and width of the MC template to have given systematic uncertainties, e.g., $\sigma_{\mu_\mathrm{b}} = 0.05$ and $\sigma_{\sigma_\mathrm{b}} = 0.005$. That is, in the altered MC template, the mean and width are both off by one standard deviation. To incorporate this error into the fit, one can include in the chi-squared the terms

$$\left(\frac{\mu_\mathrm{b}(\boldsymbol{\alpha}) - \mu_\mathrm{b}(0)}{\sigma_{\mu_\mathrm{b}}}\right)^2 + \left(\frac{\sigma_\mathrm{b}(\boldsymbol{\alpha}) - \sigma_\mathrm{b}(0)}{\sigma_{\sigma_\mathrm{b}}}\right)^2 \,, \tag{17}$$

where $\mu_\mathrm{b}(\boldsymbol{\alpha})$ and $\sigma_\mathrm{b}(\boldsymbol{\alpha})$ are the mean and standard deviation of the MC template after application of the transformation with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $\mu_\mathrm{b}(0)$ and $\sigma_\mathrm{b}(0)$ are the original values, i.e., those corresponding to $\alpha_1 = \alpha_2 = 0$. The systematic uncertainties assigned to these quantities are σ_{μ_b} and $\sigma_{\sigma_\mathrm{b}}$, respectively. The fit results are shown in Fig. 4.

When using the systematically shifted MC template but including the adjustable α_1 and α_2, the fitted values of the mean and standard deviation of the signal are

$$\hat{\mu}_\mathrm{s} = 0.50014 \pm 0.00290 \,, \tag{18}$$

$$\hat{\sigma}_\mathrm{s} = 0.10582 \pm 0.00347 \,, \tag{19}$$

Now the goodness-of-fit is again very good, with $\chi^2 = 32.1$ for 44 degrees of freedom. Note that when including the two additional parameters, α_1 and α_2, one effectively treats the nominal mean and standard deviation of the peaking background, $\mu_\mathrm{b}(0)$ and $\sigma_\mathrm{b}(0)$, as measurements, and therefore the number of degrees of freedom does not change.

By including α_1 and α_2, the fit errors have increased to reflect the uncertainty in the MC template and the bias in the fitted values is greatly reduced, as can be seen by comparing the results from the original fit based on the correct model, (13) and (14), to the fitted values (18) and (19) obtained using the biased MC template but including the adjustable parameters α_1 and α_2.

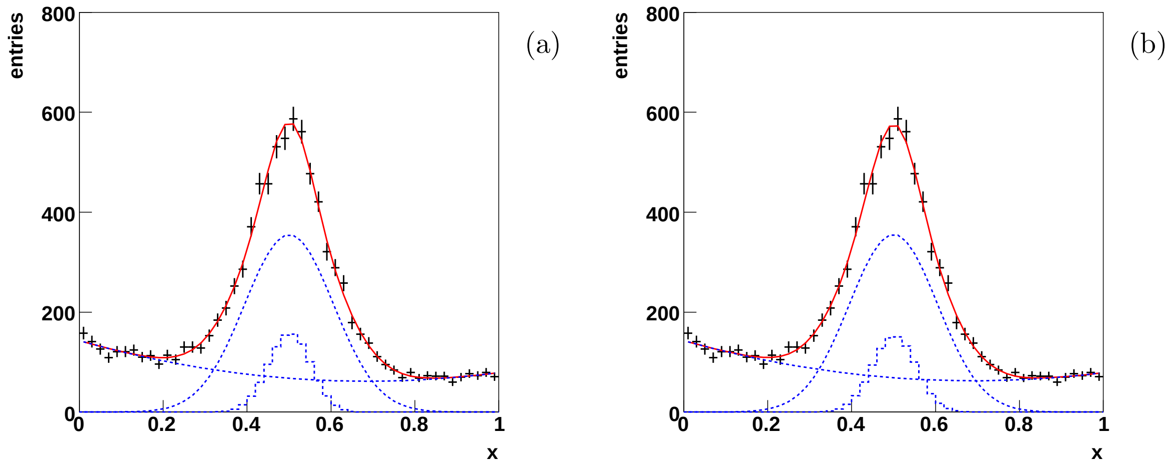

Figure 4: Plots of example fits allowing for variation of the mean and width of the peaking component through the parameters α_1 and α_2 using (a) a systematically altered peaking background component (lower histogram) and (b) an undistorted peaking background.

One obtains very similar results when one includes the variable α_1 and α_2 but without any systematic distortion of the MC template, namely,

$$\hat{\mu}_{\rm s} = 0.49981 \pm 0.00292 \; , \tag{20}$$

$$\hat{\sigma}_{\rm s} = 0.10498 \pm 0.00354 \; , \tag{21}$$

which has $\chi^2 = 30.0$ for 44 degrees of freedom.

One can regard the quadratic difference between the statistical errors with and without the additional adjustable parameters as the as the contribution from the systematic uncertainty in the MC template. Here this is

$$\sigma_{\hat{\mu},\rm sys} = \sqrt{0.00290^2 - 0.00226^2} = 0.00182 \; ,$$

$$\sigma_{\hat{\sigma},\rm sys} = \sqrt{0.00347^2 - 0.00308^2} = 0.00160 \; .$$

Formally, however, this part of the error has been converted to part of the statistical error of the fit, and it is probably most appropriate to report it as such. It is interesting to note that these systematic errors are in fact much smaller than the change one would find in the fitted values when using the correct and shifted MC templates. These are

$$\Delta\hat{\mu}_{\rm sys} = |0.50025 - 0.51676| = 0.01651 \; ,$$

$$\Delta\hat{\sigma}_{\rm sys} = |0.10578 - 0.08933| = 0.01645 \; .$$

This is not to say that one or the other way of estimating the systematic uncertainty is incorrect. Rather, in the case where one introduces additional adjustable parameters, the systematic error is in fact reduced, as can be seen by the greatly reduced bias in the fitted signal parameters.