sPlots

Adrian Bevan

email: a.j.bevan@qmul.ac.uk

Queen Mary
University of London

Visualising components of a fit

- The problem:
  - A multi-dimensional fit to data, with signal and one or more background components.
- How can we visualise what the signal really looks like in terms of the representation obtained as the optimal result of the fit?
  - The sPlot method is one way to approach this problem: The key concept is to reweight the data using information from the fit model and fit result. We compute sWeights., from which one can re-weight the data to make sPlots.
- Examples and pathologies.

Consider a fit to data

- Events:
  - N events in the data.
  - Each event is described by a multi-dimensional space of discriminating variables. e.g. mass, helicity, p_T, E_{miss}, Dalitz variables etc. For the example discussed here we assume 2 arbitrary variables x and y.
  - The set of discriminating variables y_e is used to compute an event sWeight for a given category in the fit model.
  - NOTE: y_e does not include the variable to be plotted. This means you need to refit the data for each sPlot variable you wish to display.
  - Using the sWeight one can plot the data re-weighted according to a given category of events for some other variable.

- Components:
  - There are N_s components in total, e.g.
    - Signal (j=1)
    - Background [one or more], j=2, 3, ... $N_{background}$

sWeights: ${}_s\mathcal{P}_n$

- Each event can be weighted according to the following:
  - y_e is the set of discriminating variables used in the fit.

$$ {}_s\mathcal{P}_n(y_e) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_{nj} f_j(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)} $$

sWeights: ${}_s\mathcal{P}_n$

- Each event can be weighted according to the following:

  - y_e is the set of discriminating variables used in the fit.

$$ {}_s\mathcal{P}_n(y_e) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_{nj} f_j(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)} $$

  - $f_{j/k}(y_e)$ is the PDF for the j^{th}/k^{th} component in the fit model.

sWeights: ${}_s\mathcal{P}_n$

- Each event can be weighted according to the following:
  - y_e is the set of discriminating variables used in the fit.

$$ {}_s\mathcal{P}_n(y_e) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_{nj} f_j(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)} $$

- V_{nj} is the covariance matrix obtained from fitting the data.
- $f_{j/k}(y_e)$ is the PDF for the j^{th}/k^{th} component in the fit model.

sWeights: ${}_s\mathcal{P}_n$

- Each event can be weighted according to the following:

  - y_e is the set of discriminating variables used in the fit.

$$ {}_s\mathcal{P}_n(y_e) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_{nj} f_j(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)} $$

  - N_k is the event yield fitted for the k^{th} component.
  - V_{nj} is the covariance matrix obtained from fitting the data.
  - $f_{j/k}(y_e)$ is the PDF for the j^{th}/k^{th} component in the fit model.

sPlots

- Having computed the sWeight for a set of events for a given discriminating variable set y_e, one can project out some variable $x \notin y_e$ by plotting a histogram containing all of the events in the data set, each weighted by ${}_s\mathcal{P}_n(y_e)$.

- Having excluded x from y_e, the projection ${}_s\mathcal{P}_n(y_e)$ is a representation of the data, weighted under the hypothesis that the fit has correctly extracted the n^{th} component from the fit.
  - If the PDF for x is known then the sPlot of the data for this component should look like the PDF if the fit has correctly determined that component.
  - Can use this method as a background subtraction algorithm.

Toy Example (for illustration)

- Define a 2 component fit, with 2 discriminating variables x and y.
  - Component 1: Signal:
    - $f_1(x) = G(x; -3.0, 2.0)$
    - $f_1(y) = G(y; -3.0, 2.0)$

    } $f_1 = f_1(x)\, f_1(y)$

  - Component 2: Background:
    - $f_2(x) = G(x; 5.0, 2.0)$
    - $f_2(y) = G(y; 3.0, 10.0)$

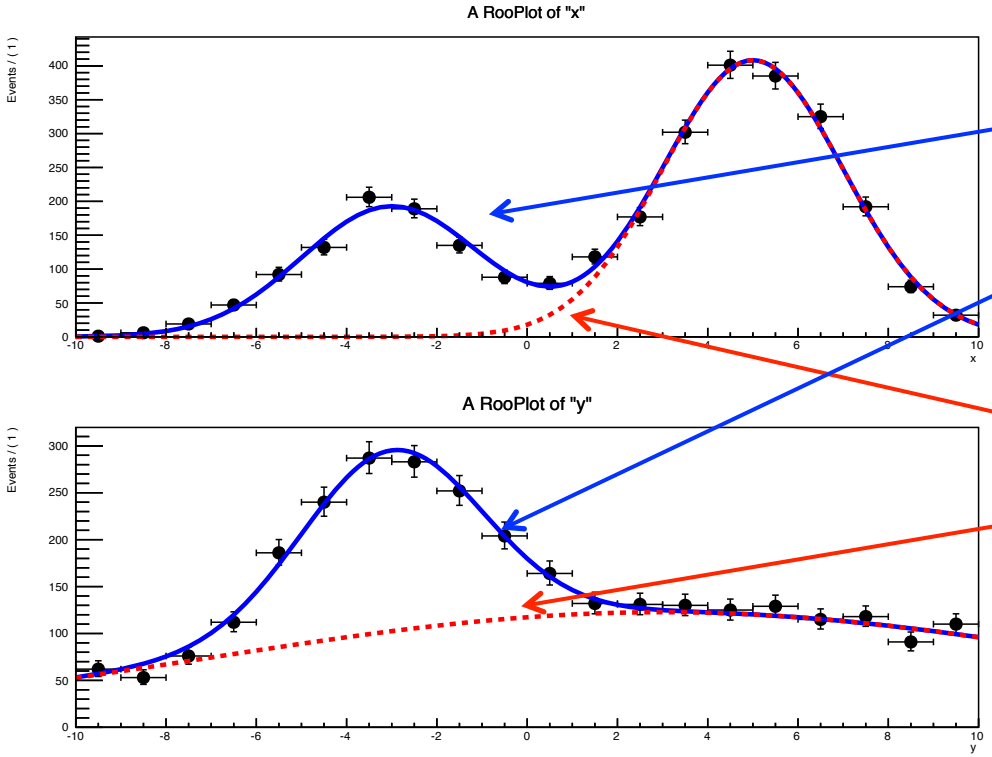
    } $f_2 = f_2(x)\, f_2(y)$

$$\mathcal{L} = \frac{e^{-(N_1+N_2)}}{N!} \prod_{i=1}^{N} N_1 f_1 + N_2 f_2$$

- Here we arbitrarily set the number of signal and background, generate a sample of events and can fit them back in the usual way.

Toy Example (for illustration)

- Projections of the generated data with the PDFs overlaid:



Total PDF

Background PDF

- Each projection marginalises (i.e. integrates) the PDF over other distributions.

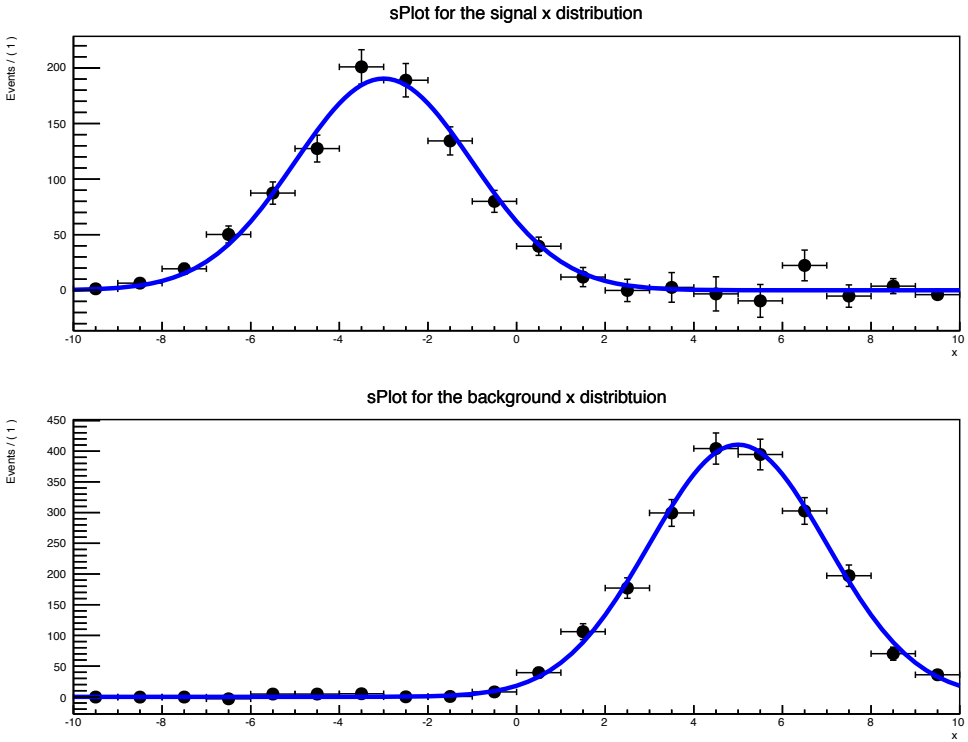
Toy Example (for illustration)

- The projection plot shows the sum of all components, where one projects the PDF representing a given fit component.
  - It is still possible for the total fit to represent the data, where one or more components are incorrectly extracted from the data as both x and y are used to compute the likelihood.

- The aim of the sPlot is to show a projection of the data in terms of the probability of a given event, summed over the data, for a given fit component.
  - Here we consider two components discussed previously, however to view y we weight the data according to the fit to x, and to view x we weight the data according to the fit to y.
  - If the extracted sPlots agree with the PDF for each component, then the fit is extracting each component correctly.

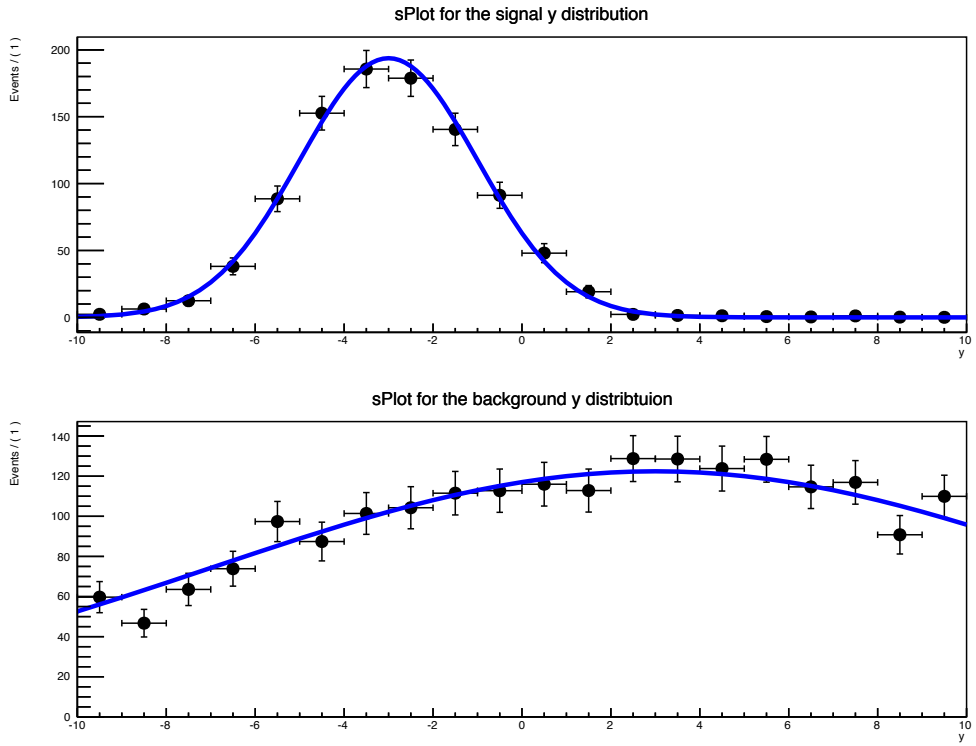
Toy Example (for illustration)

- Making the sPlot for x requires that
  - The data are fitted with the PDF for y, prior to computing the sWeights.



Toy Example (for illustration)

- Similarly making the sPlot for y requires that
  - The data are fitted with the PDF for x, prior to computing the sWeights.



sPlots

- If a negative yield is fitted for the n^{th} component, then this is also evident in the sPlot (just as it would be in a marginal distribution projected out from the likelihood).

- The method requires that the discriminating variable(s) of interest are excluded from the fit.
  - This reduces the statistical power of test statistic used to extract information about the component under study.

- The fit needs to converge with an accurate error matrix.
  - Otherwise the sWeight computation will lead to incorrect results, and hence an incorrect sPlot.

sPlots

- The method is independent of whether the discriminating variable you want to plot is in the nominal fit model or not.
  - Having computed the sWeights for a given event these are used to reweight that event.
  - The corollary of this is that any distribution related to the data can therefore be projected.
  - Example: Fit the B mass distribution for a decay into two spin 1 particles. Then make sPlots of the angular variables in some basis (e.g. the helicity or transversity basis). This enables one to fit the sPlot to determine f_L and other angular parameters without having to worry about the background model.

Uses of the sPlot technique

- sPlots can be used as a presentation technique to provide re-weighted event distributions that correspond to the sum of the probabilities fitted for a given fit component over the whole data sample.
  - You can think of this as a sophisticated background subtraction technique.

- Fit validation:
  - One can use sPlots of background components as a validation tool as part of a blind-analysis protocol. If the background sPlots are consistent with the background model hypothesis, then one can be confident that those background components are being treated sufficiently accurately in the fit model being used.
  - The implication of this is that one should be able to extract the signal accurately, unless there is a peaking background that is closer to the signal shape than the background components being projected.

Pathologies

- The re-weighting method can not do magic.
  - An accurate error matrix is required in order for the sWeights for an event to be correctly computed. Therefore the fit procedure must have converged properly in order to make an sPlot of the data.
  - If the fit is unable to robustly extract a fit component, then the sPlot of that component will almost certainly not be representative of the expected distribution.
    - This can happen when there is insufficient discrimination between two or more components in a fit, for example if the mass distribution of signal is close to that of a peaking background.
    - It can also happen if there are underlying correlations between discriminating variables in the data that have been neglected in the fit model.

Further reading

- Detailed reference on the method:
  - M. Pvik and F. Le Diberder, Nucl. Instrum. Meth A555, 356–369 (2005). doi:10.1016/j.nima.2005.08.106.physics/0402083.
- Examples of papers that have used the technique:
  - BaBar paper on B decays to charmless two-body final states: http://arxiv.org/abs/1206.3525.
- There are implementations of the sPlot algorithm in ROOT:
  - TSPlot
  - RooStats::SPlot
- See the example splot.cc for an illustration of how to use RooStats::SPlot.