Machine learning in HEP

Likhomanenko Tatiana

Summer school on Machine Learning in High Energy Physics

sPlot technique: solution for what?

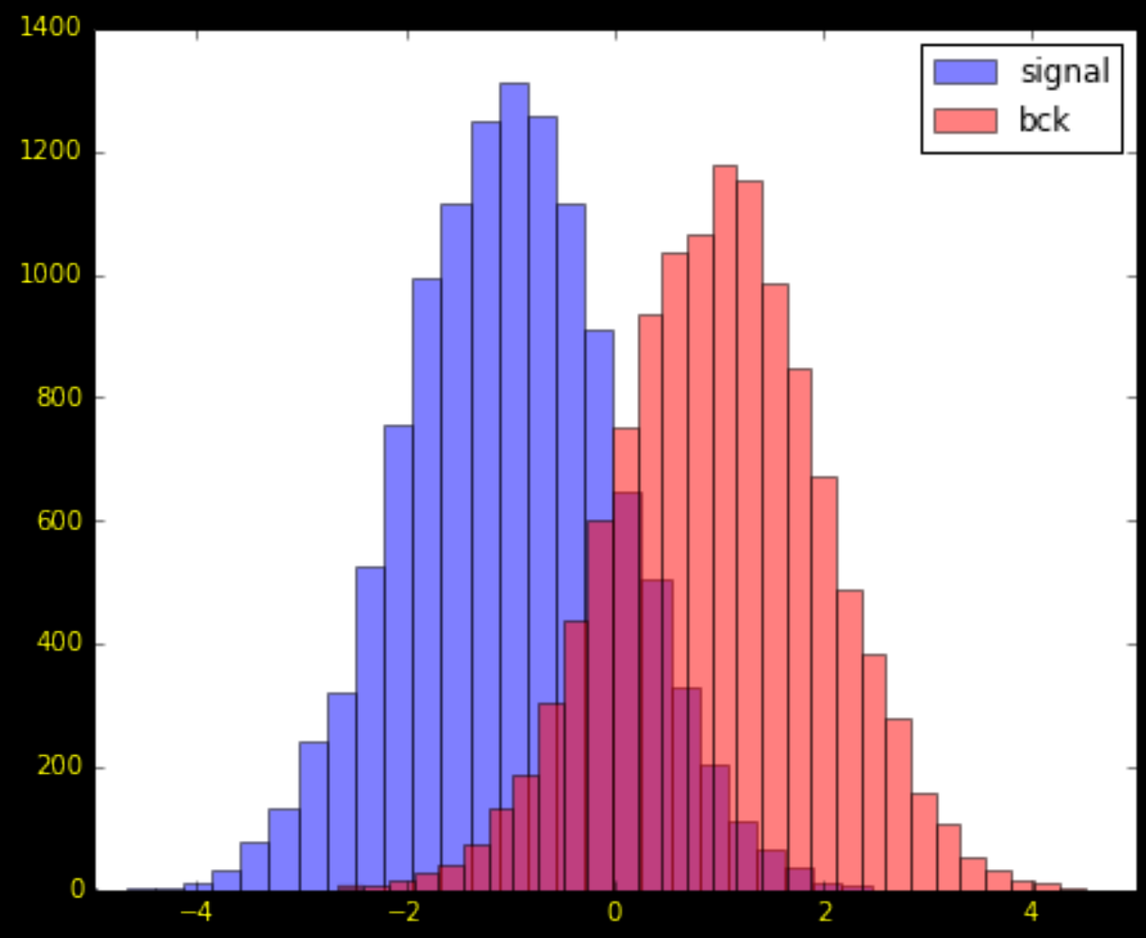
- We need to use real signal data from the control channel (but it is unlabeled).
- For some tasks we cannot simulate data (complicated or MC too differs from real data).

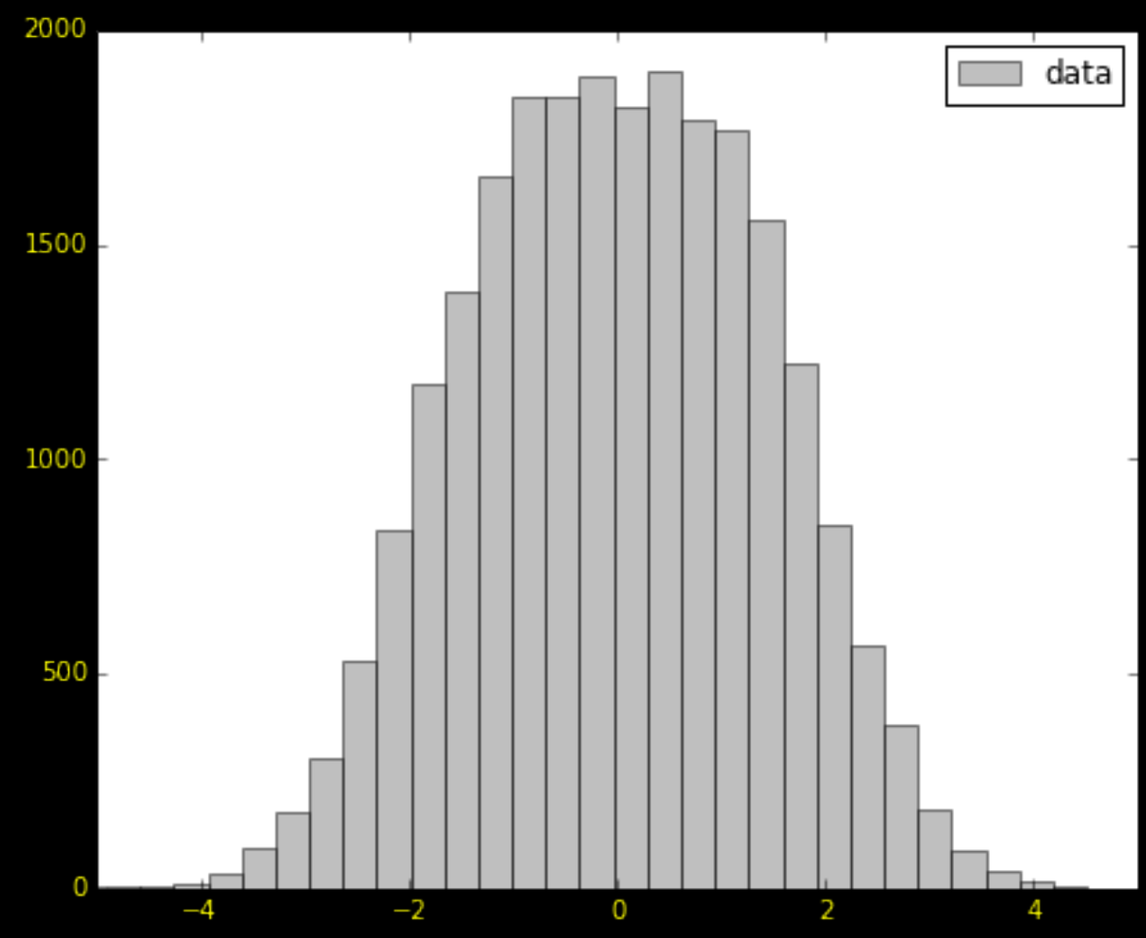
We need some solution to label real data, or to be more precise, we want to restore for features their distributions for the signal and background data.

Our main knowledge is the mass distribution for real data from which we can extract (using some physics) the mass pdfs for signal and background.

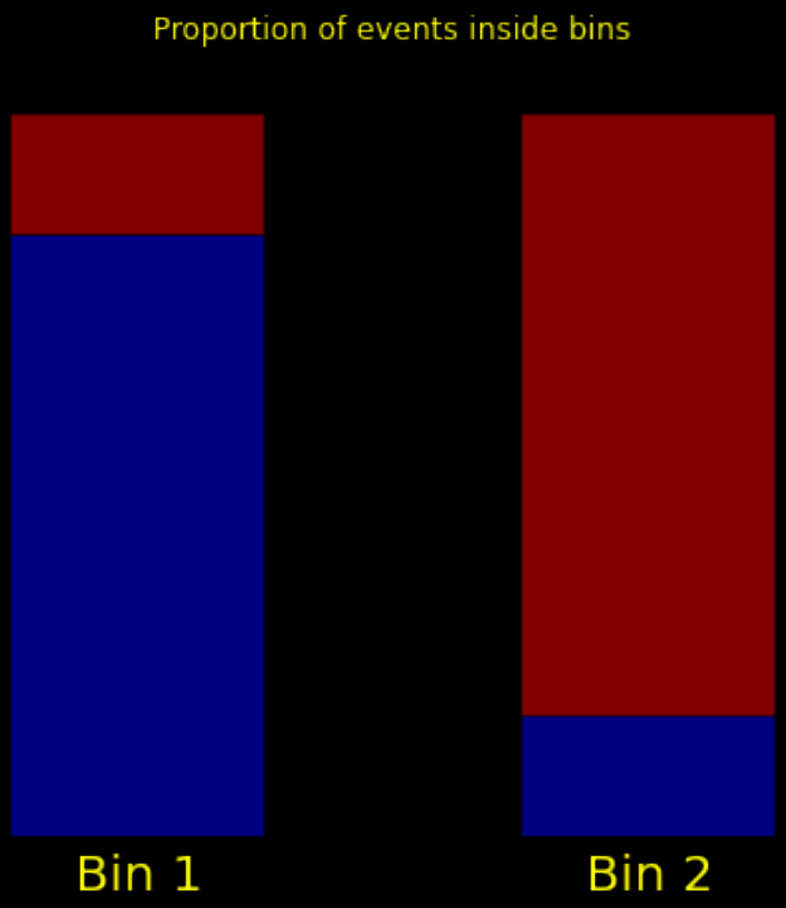
- The problem is how to restore signal/bck pdfs for other features (reconstructed features) if we know the mass pdf for each of the classes?

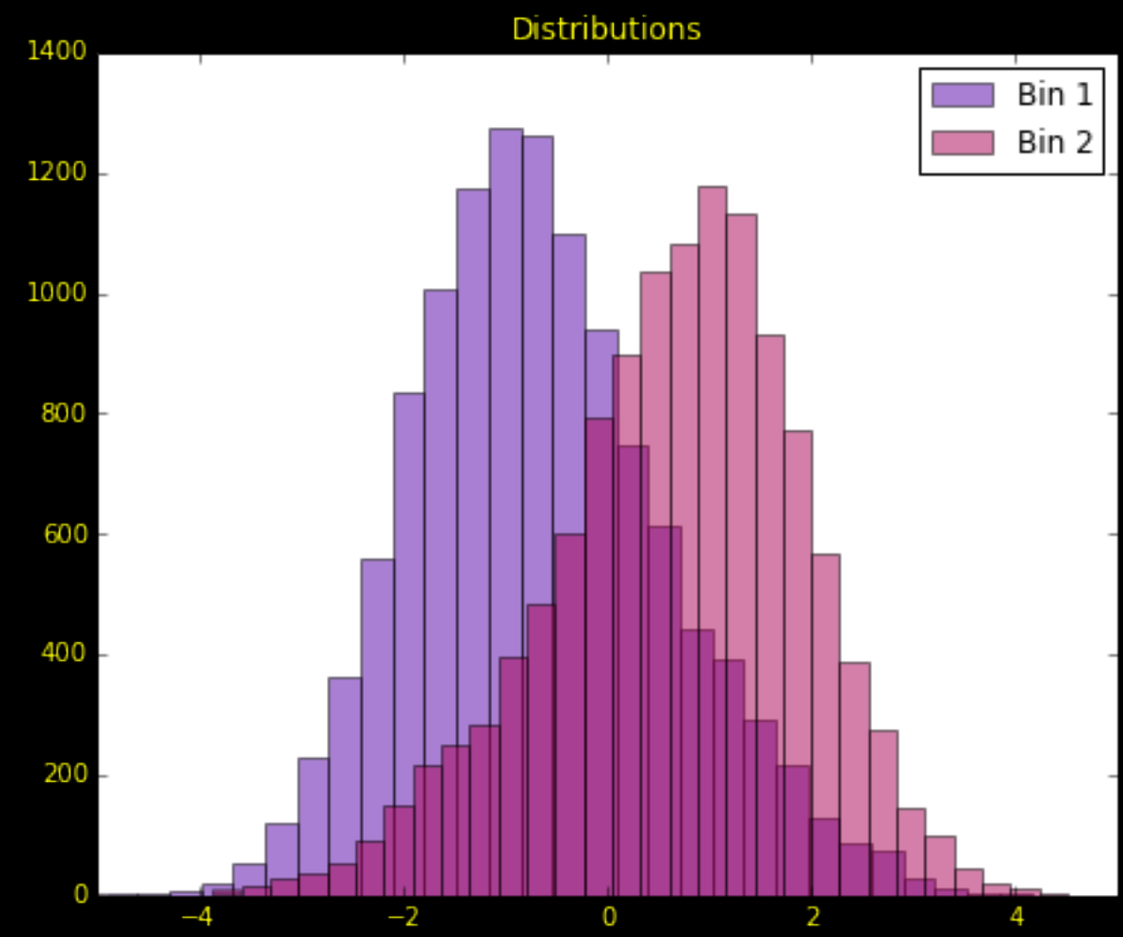
Feature initial distribution
for signal events and bck events





Two mass bins

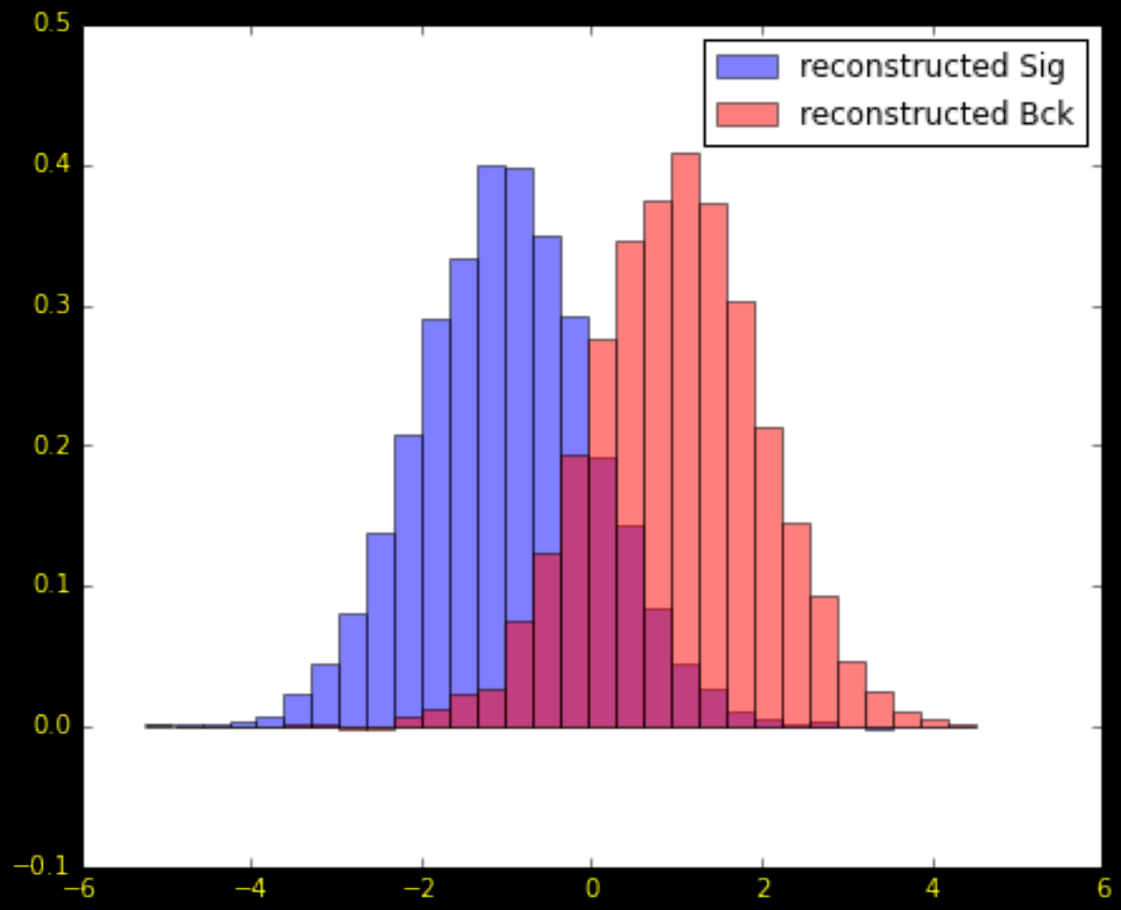


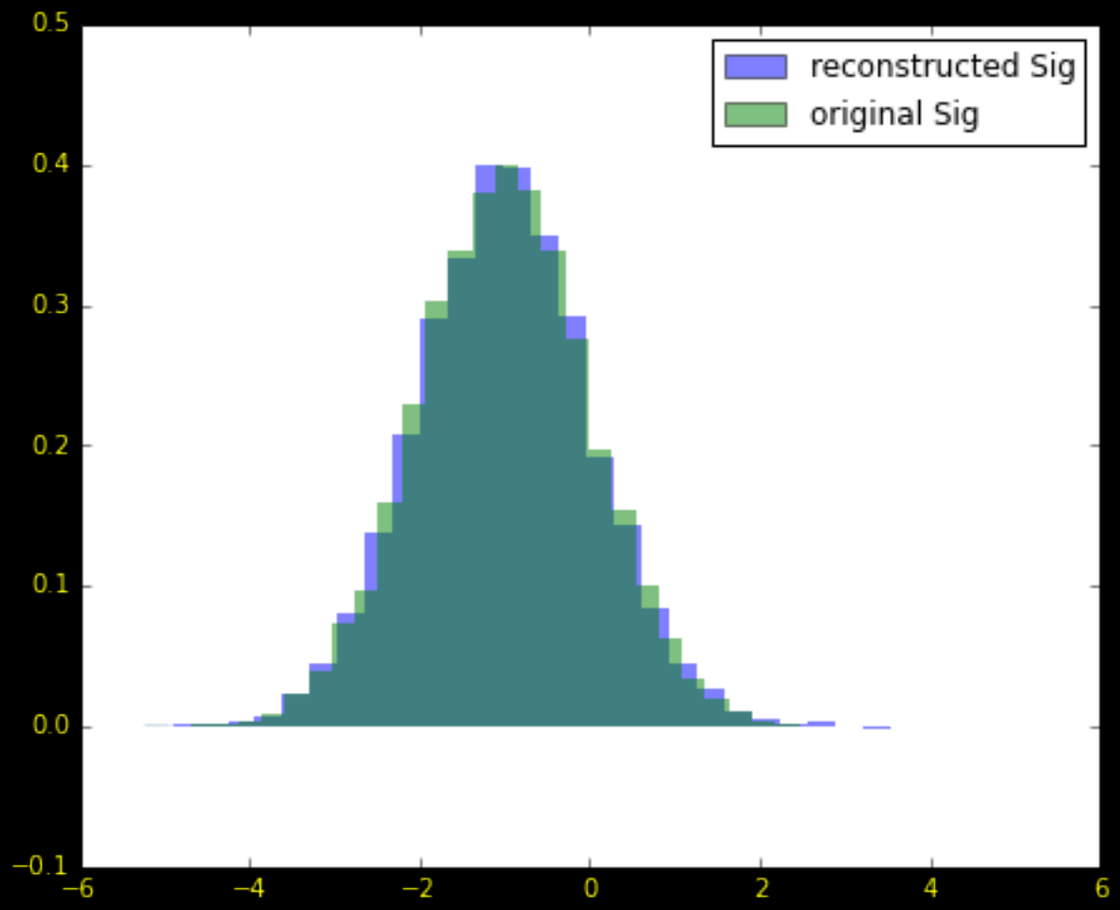


$$Bin1 : w_{b_1} f_b + w_{s_1} f_s \quad *w_{b_2}$$
$$Bin2 : w_{b_2} f_b + w_{s_2} f_s \quad *(-w_{b_1}) \quad +$$

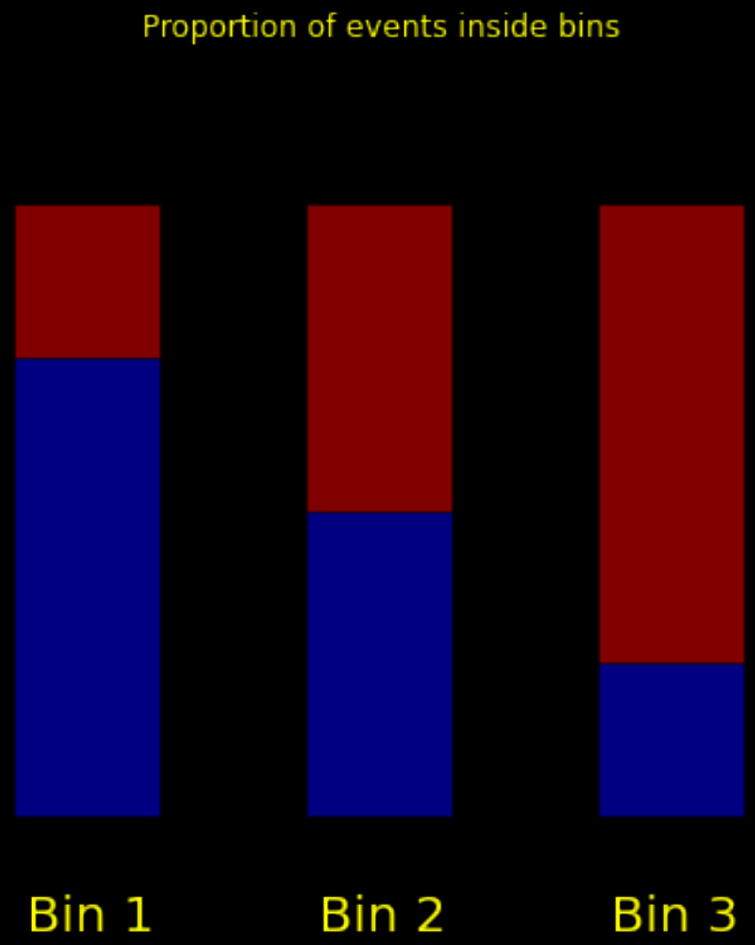
will obtain initial signal distribution

Reconstruction





More bins in mass



How to reweight this?

sPlot theory

After the mass maximum likelihood fitting we know for each event y

$$p_s(y), p_b(y)$$

which are probabilities of an event to be signal and background.

Will reconstruct number of *signal* events in a *particular* histogram bin for the reconstructed feature. Introduce unknown probability that the signal/bck event will be in a particular bin:

$$p_s, p_b$$

The total amount of signal/bck events obtained from the fit:

$$N_s, N_b$$

sPlot theory

The random variable, number of signal events in the bin:

$$X = \sum_y w_s(y) \mathbb{I}_{y \in bin}$$

where $w_s(y)$ are sPlot weights and are a subject to find.

Property: an estimation should be unbiased:

$$\mathbb{E}X = p_s N_s$$

Corollary:

$$p_s N_s = \mathbb{E}X = \sum_y w_s(y) \mathbb{E}\mathbb{I}_{y \in bin} =$$

$$= \sum_y w_s(y)(p_s p_s(y) + p_b p_b(y))$$

sPlot theory

Since the previous equation should hold for all possible p_s, p_b, we get two equalities:

$$p_s N_s = \sum_y w_s(y) p_s p_s(y)$$

$$0 = \sum_y w_s(y) p_b p_b(y)$$

after reduction:

$$N_s = \sum_y w_s(y) p_s(y)$$

$$0 = \sum_y w_s(y) p_b(y)$$

Then we can guarantee that mean input of background are 0 (the expectation is zero, but observed number will not be zero due to the deviation)

sPlot theory

Assumption of the linearity:

$$w_s(y) = a_1 p_b(y) + a_2 p_s(y)$$

Then:

$$\begin{pmatrix} V_{bb} & V_{bs} \\ V_{sb} & V_{ss} \end{pmatrix} * \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ N_s \end{pmatrix}$$

where $V_{ij} = \sum_y p_i(y) * p_j(y)$

The assumption of the linearity is correct because apart from having correct mean, we should also minimize the variation of the reconstructed variable.

$$\mathbb{V}X = \sum_y w_s^2(y) \mathbb{VI}_{y \in bin} \leq \sum_y w_s^2(y)$$

Minimization problem

$$\sum_y w_s^2(y) \to min$$

$$\sum_y w_s(y) p_b(y) = 0$$

$$\sum_y w_s(y) p_s(y) = N_s$$

Lagrangian:

$$\mathcal{L} = \sum_y w_s^2(y) + \lambda_1 \left(\sum_y w_s(y) p_b(y) \right) + \lambda_2 \left(\sum_y w_s(y) p_s(y) - N_s \right)$$

$$0 = \frac{\partial \mathcal{L}}{\partial w_s(y)} = 2 w_s(y) + \lambda_1 p_b(y) + \lambda_2 p_s(y)$$

It holds for each event, thus we are getting needed the linear dependence

Main assumption

The mass must be uncorrelated with the reconstructed feature:

$$p_s(mass, feature) = p_s(y, feature) = p_s(y)p_s(feature)$$

Then it holds for a particular bin for the reconstructed feature:

$$p_s(mass, bin) = p_s(y, feature) = p_s(y)p_s$$

Control channel

- Fit the mass pdf to get the pdfs for signal and bck
- Apply sPlot to obtain weights
- Now pdfs with weights for reconstructed features will be signal pdfs
- We can compare them with MC data
- But weights can be negative!
- Need some KS extension for the negative weights

KS with negative weights

- KS statistic cannot be extended on the pdfs with negative weights!
- KS metric can be calculated using FPR and TRP:
  - the ROC works with any weights
  - our weights will compensate background events, but some fluctuations can stay
- Hypothesis testing (two-samples test) can be done only by generating the KS metric pdf.