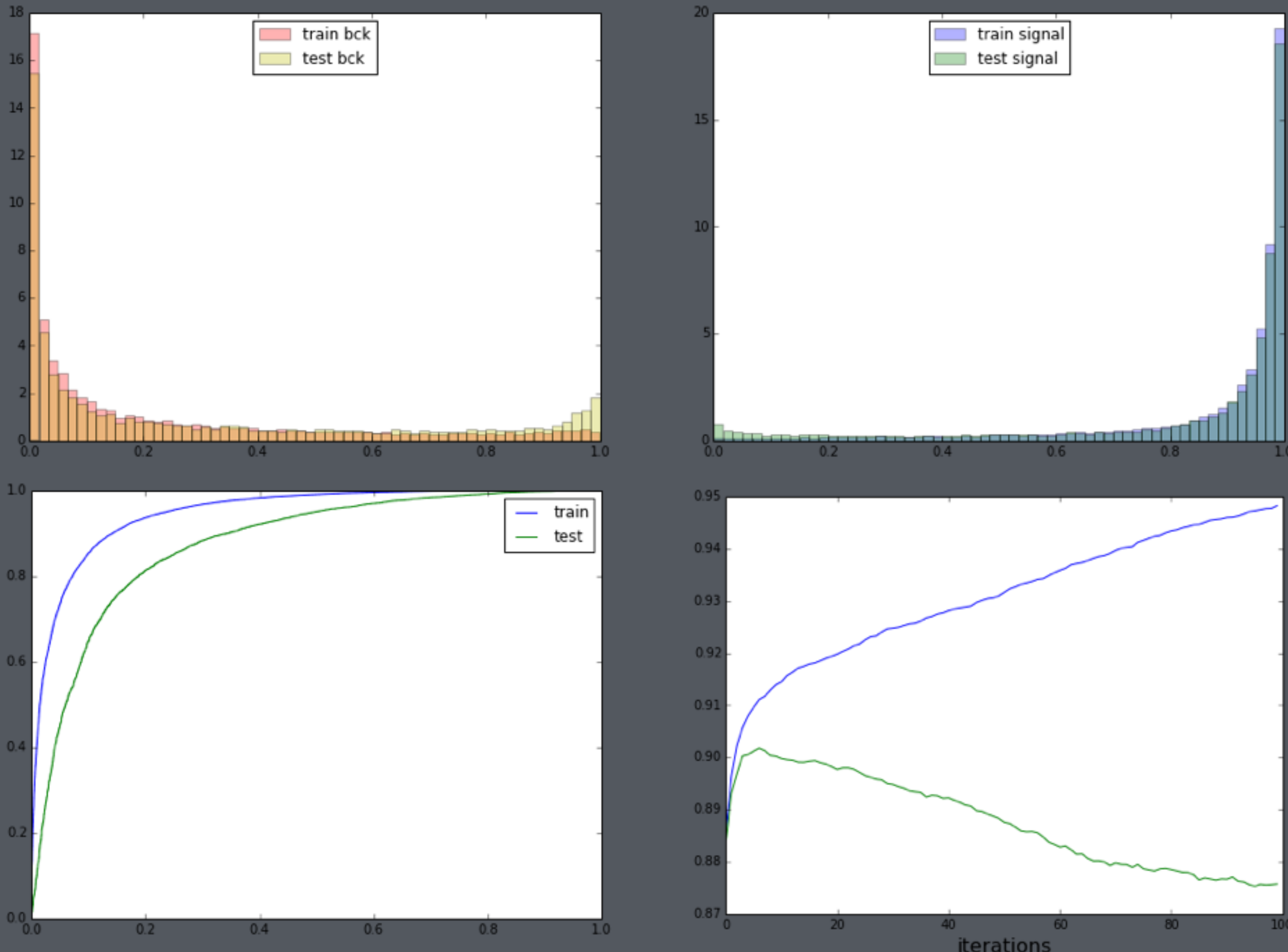
Machine learning in HEP

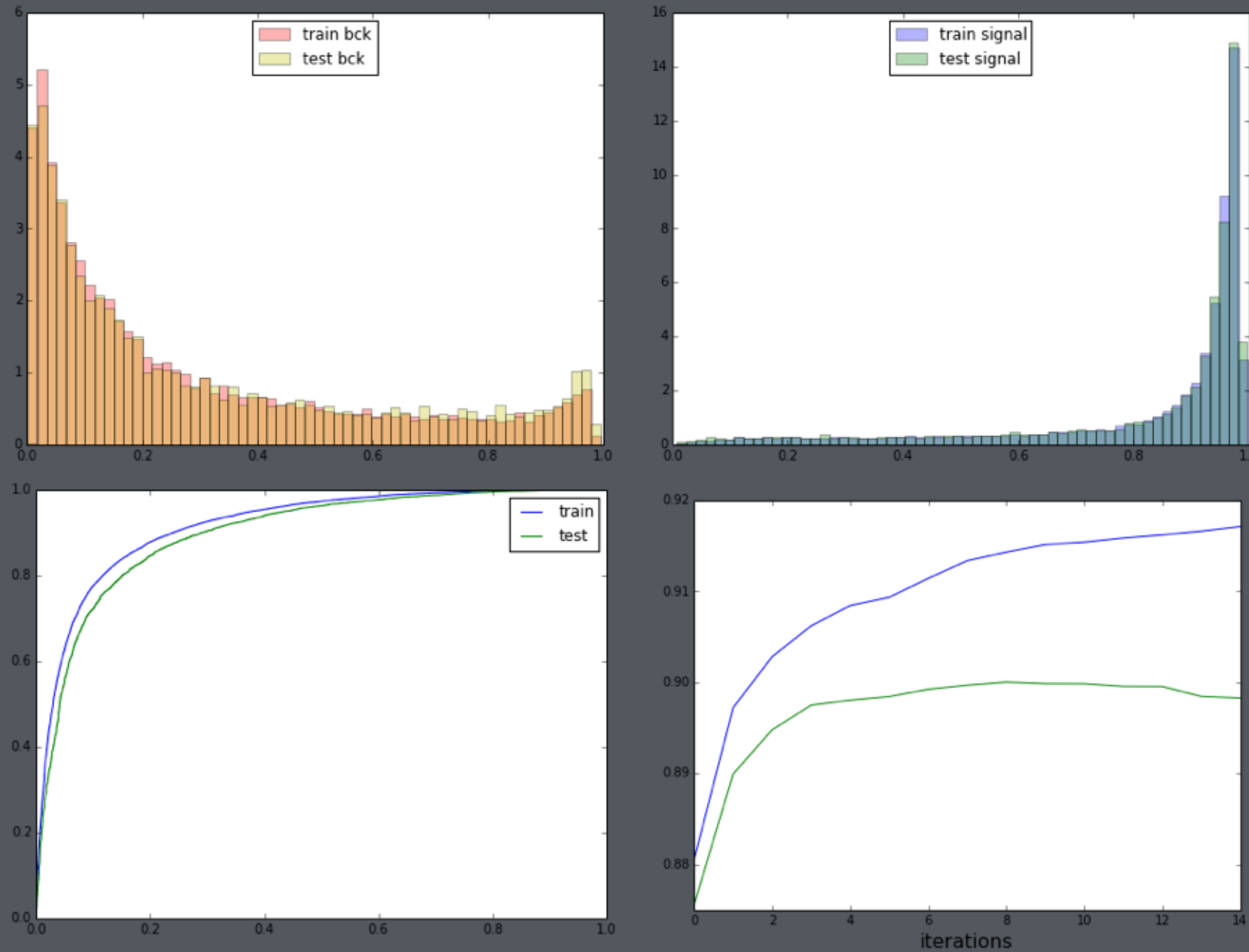
Likhomanenko Tatiana

Summer school on Machine Learning in High Energy Physics

Overfitting



train bck
test bck
train signal
test signal
train
test
iterations



train bck
test bck
train signal
test signal
train
test
iterations

Hypotheses testing

- We observe *n*-events. The value, observed number of event, is assumed to have the Poisson distribution.

$$H0 : n \sim Poiss(b) \text{ background only}$$
$$H1 : n \sim Poiss(s+b) \text{ signal + bck}$$

- We can define the optimal Neumann-Pearson statistic:

$$Q = \frac{L(X|s+b)}{L(X|b)}$$
$$\ln Q = -s + n \ln\left(1 + \frac{s}{b}\right)$$

- If we use several independent experiments:

$$\ln Q = -s_{tot} + \sum_i n_i \ln\left(1 + \frac{s_i}{b_i}\right)$$

- In our experiments $s_i << b_i$

$$\ln Q \sim q = -s_{tot} + \sum_i \frac{s_i}{b_i} n_i$$

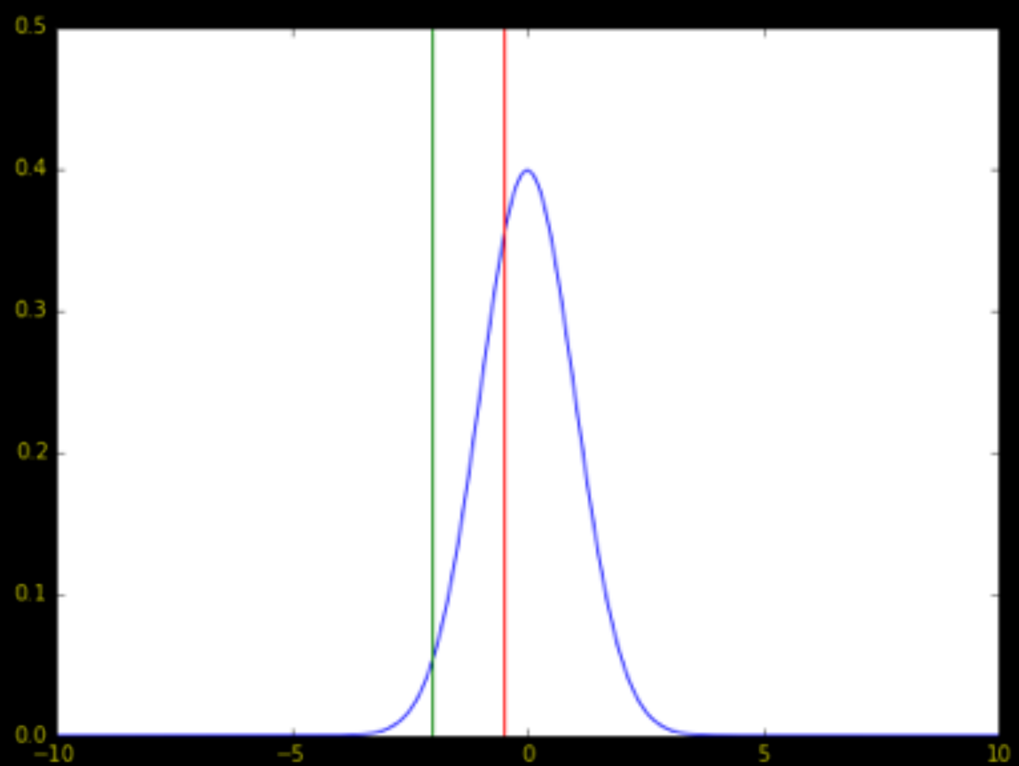
$$\mathbb{E}q = -s_{tot} + \sum_i \frac{s_i}{b_i} \mathbb{E}n_i$$

$$\mathbb{E}q|H0 = -s_{tot} + \sum_i \frac{s_i}{b_i} b_i$$

$$\mathbb{E}q|H1 = -s_{tot} + \sum_i \frac{s_i}{b_i} (s_i + b_i)$$

$$\mathbb{V}q|H0 = \sum_i \frac{s_i^2}{b_i}$$

$$\mathbb{V}q|H1 = \sum_i \frac{s_i^2}{b_i^2} (s_i + b_i) \sim \sum_i \frac{s_i^2}{b_i}$$



$$\eta = q|H1 - q|H0$$

$$\text{power} = P(\eta > 0)$$

$$P(\eta > 0) = P\left(\frac{\eta - \mathbb{E}\eta}{\sqrt{\mathbb{D}\eta}} > -\frac{\mathbb{E}\eta}{\sqrt{\mathbb{D}\eta}}\right)$$

$$\frac{\mathbb{E}\eta}{\sqrt{\mathbb{D}\eta}} \to max$$

- The power of the test (power to distinguish two hypotheses):

$$\text{power} \sim \frac{\mathbb{E}\eta}{\sqrt{\mathbb{D}\eta}} \sim \sqrt{2\sum_i \frac{s_i^2}{b_i}} \to max$$

- Thus, threshold for classifier’s selection should optimize:

$$\frac{s^2}{b} = \frac{TPR^2(threshold)}{FPR(threshold)}$$

- To calculate TPR we use Monte Carlo. Thus TPR will be overestimated. The control channel can give calibration multiplier:

$$\frac{[\alpha(threshold) * TPR(threshold)]^2}{FPR(threshold)}$$

- To make our model more sensitive, we can use several regions of the classifier's predictions (divide predictions into several bins). Each bin is assumed to be an independent experiment (is used in the Bs->μμ)

$$\sum_i \frac{[\alpha_i * s_i]^2}{b_i}$$

- If Monte Carlo and real data agree, then the classifier optimization is simpler

Hypotheses testing:
common case

- We observe *n*-events. The value, observed number of event, is assumed to have the Poisson distribution with the mean ($\mu s+b$), where s is a number of signal events, b is a number of bck events, μ is a signal power. And likelihood of our model:

$$L(\mu) = P(n|\mu)\prod_{i=1}^{n} f(x_i|\mu) = \frac{(\mu s + b)^n e^{-(\mu s+b)}}{n!}\prod_{i=1}^{n}\frac{\mu s f(x_i|s) + b f(x_i|b)}{(\mu s + b)}$$

$$L(\mu) = \frac{e^{-(\mu s+b)}}{n!}\prod_{i=1}^{n}[\mu s f(x_i|s) + b f(x_i|b)]$$

$$H0 : \mu = 0$$

$$H1 : \mu > 0$$

- Often in HEP one claims discovery when the p-value of the background-only hypothesis is found below 2.9 × 10^−7, corresponding to a 5-sigma effect.

- Construct the Neumann-Pearson statistic.

$$\ln Q = \ln \frac{L(0)}{L(\hat{\mu})} = \hat{\mu} s - \sum_{i=1}^{n} \ln \left(1 + \frac{\hat{\mu} s}{b} \frac{f(x_i|s)}{f(x_i|b)} \right),$$

where $\hat{\mu}$ is the maximum likelihood estimation of $L(\mu|H1)$

- x_i is a discriminative variable. To get the previous statistic x_i should be seen as a bin of predictions. Thus

$$f(x_i|s) = P(i-th \text{ event is in the bin}|s) \sim \frac{s_{bin}}{s}$$

$$f(x_i|b) = P(i-th \text{ event is in the bin}|b) \sim \frac{b_{bin}}{b}$$

$$\ln Q = \hat{\mu}s - \sum_{bin} n_{bin} \ln\left(1 + \frac{\hat{\mu}s_{bin}}{b_{bin}}\right)$$

Metrics

$$\frac{s}{\sqrt{b}} \qquad \frac{s}{\sqrt{s+b}} \qquad \frac{s}{\sqrt{b+b_{reg}}}$$

$$\frac{s}{\sqrt{b}+sigmas/2}$$

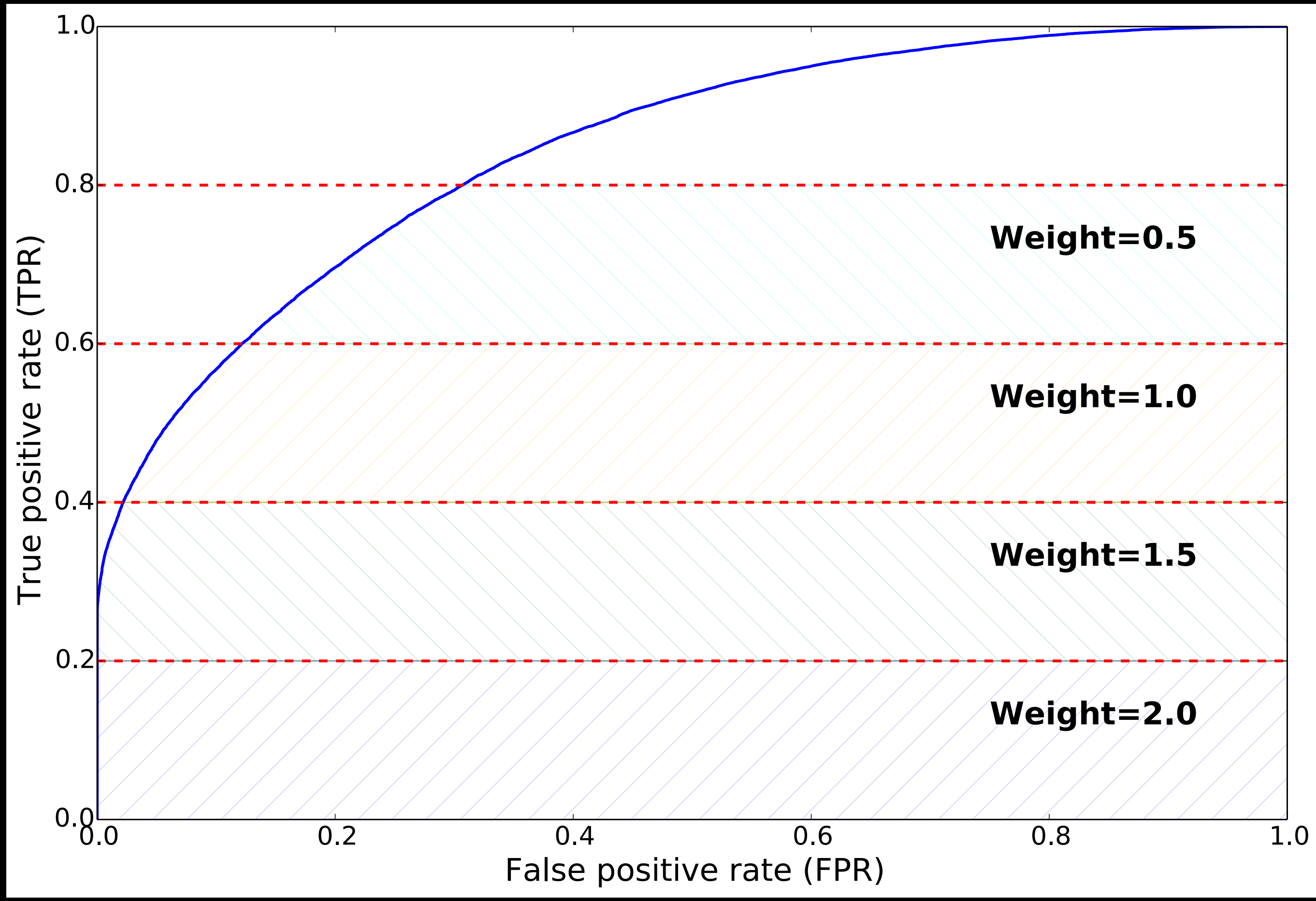
$$AMS = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

- Details about metrics can be found in G. Punzi paper (He also uses Gaussian approximation)

http://arxiv.org/abs/physics/0308063v2

Weighted ROC AUC

- The most sensitive bin is that what has the greatest predictions.
- For our hypotheses metric only part of the ROC curve is interesting
- Different parts of the ROC curve have different sensitivity in terms of our metric
- The solution for optimization of the classifiers is to take the weighted ROC AUC



ROC curve demonstration

Hypotheses testing:
ND pdfs

- Train classifier to distinguish our ND data (ND pdfs)
- If it is close to 0.5 than pdfs are similar, if close to 1. than they are different
- How to test hypotheses: H0-similar, H1-different? And how to compute p-value?

Mann-Whitney U-test

- It is a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis, especially that a particular population tends to have larger values than another.
- Assign numeric ranks to all observations:

0011101010110
sample 1, 1-labeled ranks: 3,4,5,7,9,11,12
sample 2, 0-labeled ranks: 1,2,6,8,10,13

$$U_i = R_i - \frac{(n_i + 1)n_i}{2}$$

$$m_U = \frac{n_1 n_2}{2}$$

$$U_1 + U_2 = n_1 n_2$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$z = \frac{U - m_U}{\sigma_U}$$

U-statistic and AUC

The U statistic is equivalent to AUC

$$AUC_1 = \frac{U_1}{n_1 n_2}$$

$$AUC_1 = \frac{U_1}{n_1 n_2} = 1 - \frac{U_2}{n_1 n_2} = 1 - AUC_2$$

$$AUC_2 = \frac{U_2}{n_1 n_2}$$