Comment on Event Quality Variables for Multivariate Analyses

ATLAS Machine Learning Workshop
CERN, 29-31 March 2016

indico.cern.ch/event/483999/

Glen Cowan
Physics Department
Royal Holloway, University of London
www.pp.rhul.ac.uk/~cowan
g.cowan@rhul.ac.uk

Outline

First comment is to emphasize that variables related to the quality of an event's reconstruction can help in a multivariate analysis, even if that variable by itself gives no discrimination between event types.

See also Ben Sowden, 10 Sep 2015 Stat. Forum talk

Second comment (if there is time) is on use of the distribution of a classifier output in a search.

A simple example (2D)

Consider two variables, x_1 and x_2, and suppose we have formulas for the joint pdfs for both signal (s) and background (b) events (in real problems the formulas are usually not available).

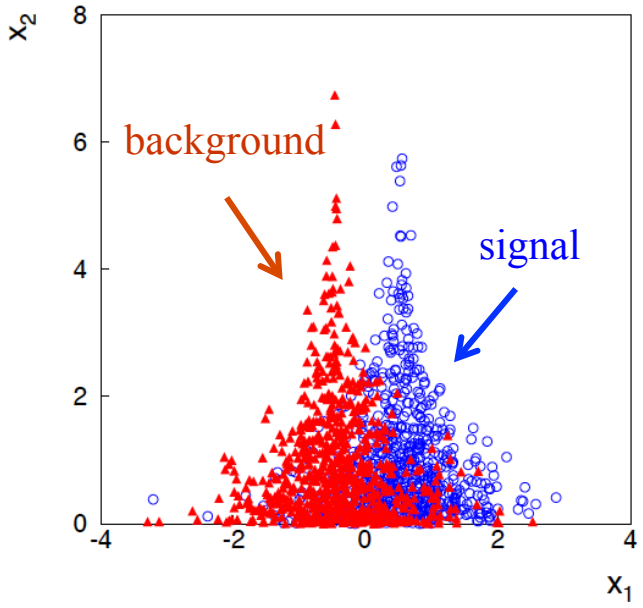
$f(x_1|x_2)$ ~ Gaussian, different means for s/b,
Gaussians have same σ, which depends on x_2,
$f(x_2)$ ~ exponential, same for both s and b,
$f(x_1, x_2) = f(x_1|x_2)\, f(x_2)$:

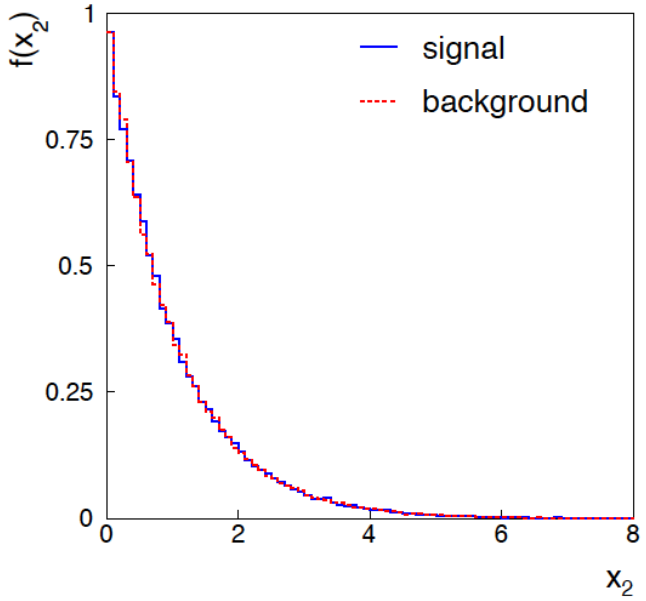
$$f(x_1, x_2|\mathrm{s}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1-\mu_\mathrm{s})^2/2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$

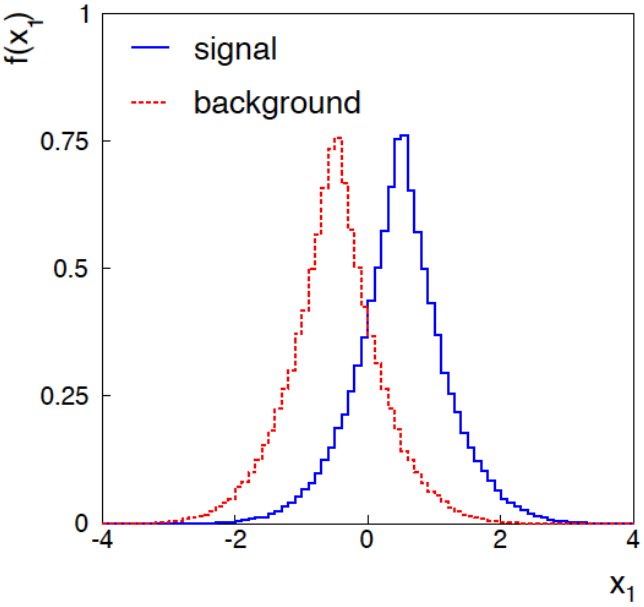
$$f(x_1, x_2|\mathrm{b}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1-\mu_\mathrm{b})^2/2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$

$$\sigma(x_2) = \sigma_0 e^{-x_2/\xi}$$

Joint and marginal distributions of x_1, x_2







Distribution $f(x_2)$ same for s, b.

So does x_2 help discriminate between the two event types?

Likelihood ratio for 2D example

Neyman-Pearson lemma says best critical region for classification is determined by the likelihood ratio:

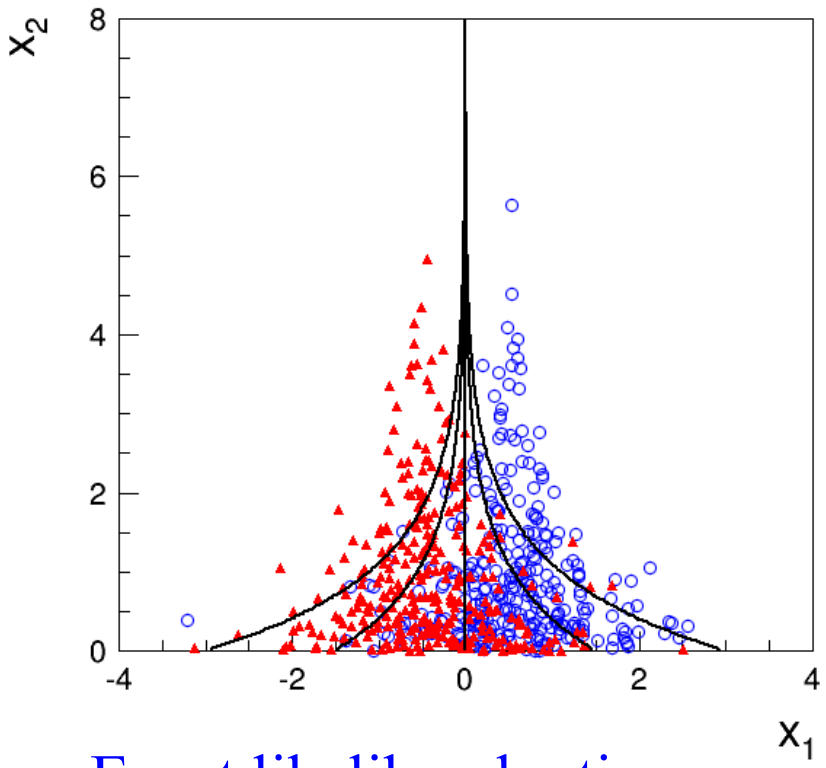
$$t(x_1, x_2) = \frac{f(x_1, x_2|\mathrm{s})}{f(x_1, x_2|\mathrm{b})}$$

Equivalently we can use any monotonic function of this as a test statistic, e.g.,

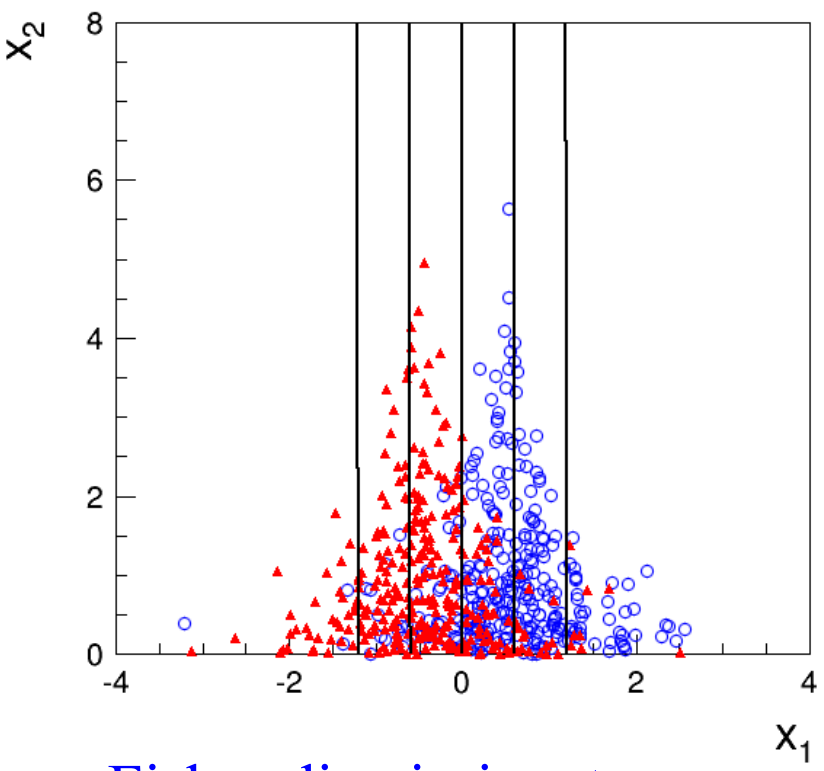
$$\ln t = \frac{\frac{1}{2}(\mu_\mathrm{b}^2 - \mu_\mathrm{s}^2) + (\mu_\mathrm{s} - \mu_\mathrm{b})x_1}{\sigma_0^2 e^{-2x_2/\xi}}$$

Boundary of optimal critical region will be curve of constant ln t, and this depends on x_2!

Contours of constant MVA output

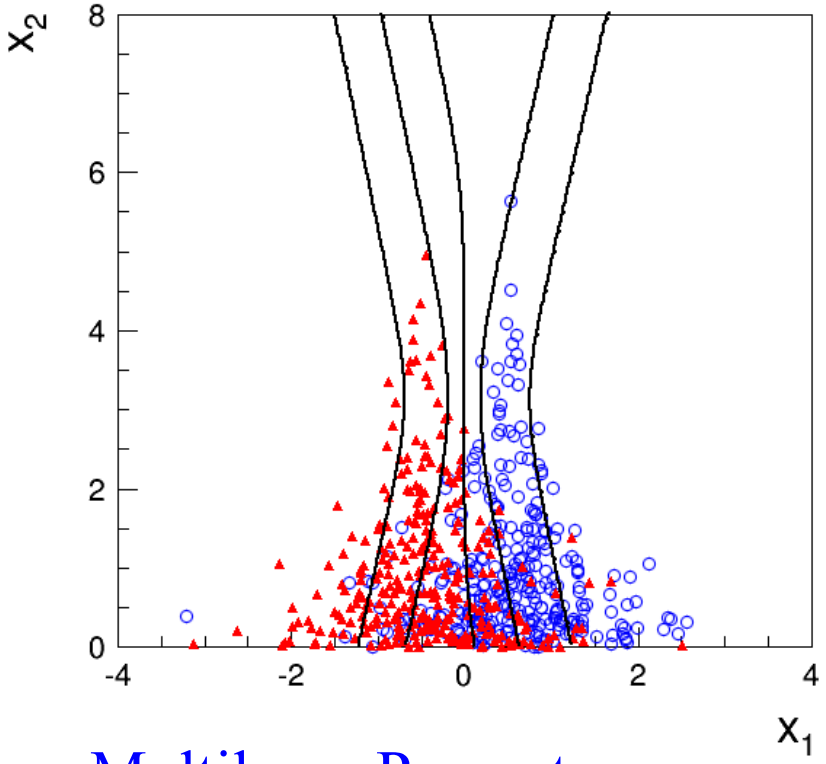


Exact likelihood ratio

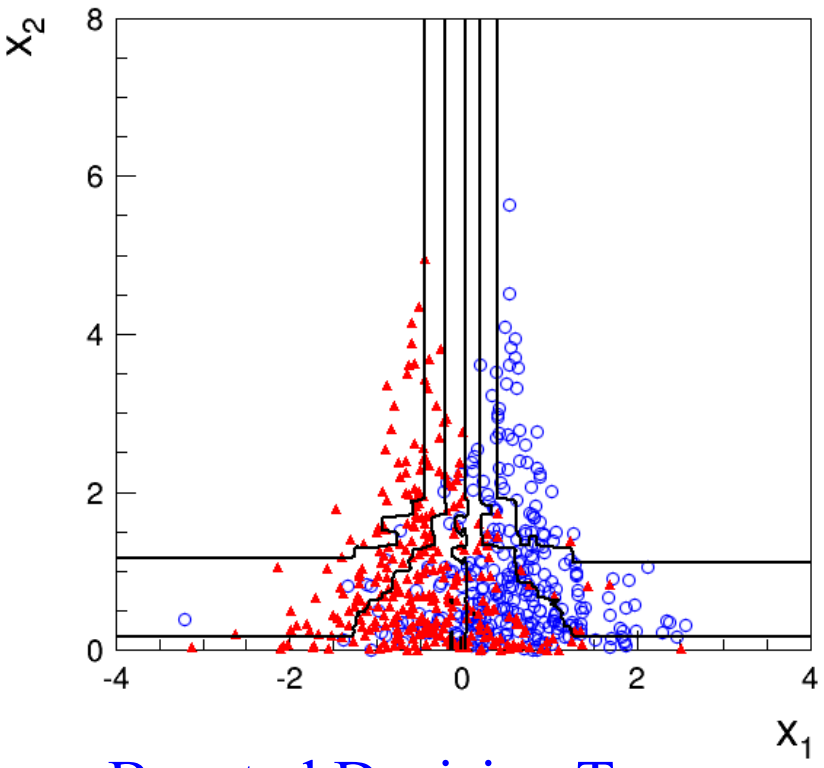


Fisher discriminant

Contours of constant MVA output



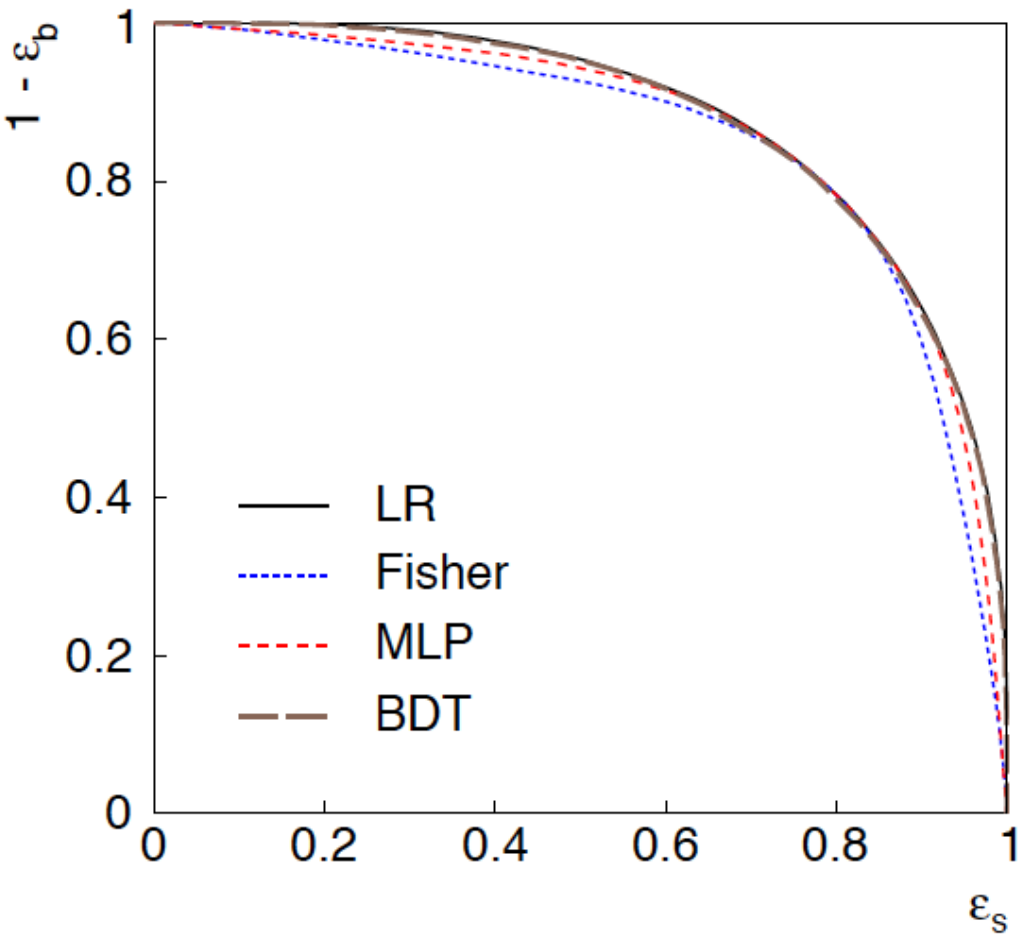
Multilayer Perceptron
1 hidden layer with 2 nodes



Boosted Decision Tree
200 iterations (AdaBoost)

Training samples: 10^5 signal and 10^5 background events

ROC curve



ROC = "receiver operating characteristic" (term from signal processing).

Shows (usually) background rejection ($1-\varepsilon_b$) versus signal efficiency ε_s.

Higher curve is better; usually analysis focused on a small part of the curve.

2D Example: discussion

Even though the distribution of x_2 is same for signal and background, x_1 and x_2 are not independent, so using x_2 as an input variable helps.

Here we can understand why: high values of x_2 correspond to a smaller σ for the Gaussian of x_1. So high x_2 means that the value of x_1 was well measured.

If we don't consider x_2, then all of the x_1 measurements are lumped together. Those with large σ (low x_2) "pollute" the well measured events with low σ (high x_2).

Often in HEP there may be variables that are characteristic of how well measured an event is (region of detector, number of pile-up vertices,...). Including these variables in a multivariate analysis preserves the information carried by the well-measured events, leading to improved performance.

Comment on use of classifier output

Often start with input variables $\boldsymbol{x}$, construct a classifier $y(\boldsymbol{x})$ using MC samples of (s/b) training data and then construct a statistic based on the distribution of y to test values of a parameter μ proportional to rate of signal process (e.g., $\mu = 1$ is background only, $\mu = 1$ is nominal signal model).

E.g., from a histogram of y with N bins, entries $(n_1,..., n_N)$, construct a Poisson likelihood $L(\mu) = P(n_1,..., n_N|\, \mu)$ with mean value in bin i, $E[n_i|\mu] = \mu s_i + b_i$.

If there are nuisance parameters ν corresponding to systematic uncertainties, then $s \rightarrow s(\nu)$, $b \rightarrow b(\nu)$, the likelihood becomes $L(\mu,\nu)$ and one uses e.g. profiling or Bayesian marginalization.

If one wants to exploit only the shape of the distribution of y and not the absolute numbers of events found, use Multinomial model (examples follow).

Example with ttH (Run I)

Background: inclusive tt from Powheg + Pythia8

Signal: ttH Pythia8, $m_H = 125$ GeV

After initial selection (4 b jets, 2 leptons), for 20 fb^{-1}:

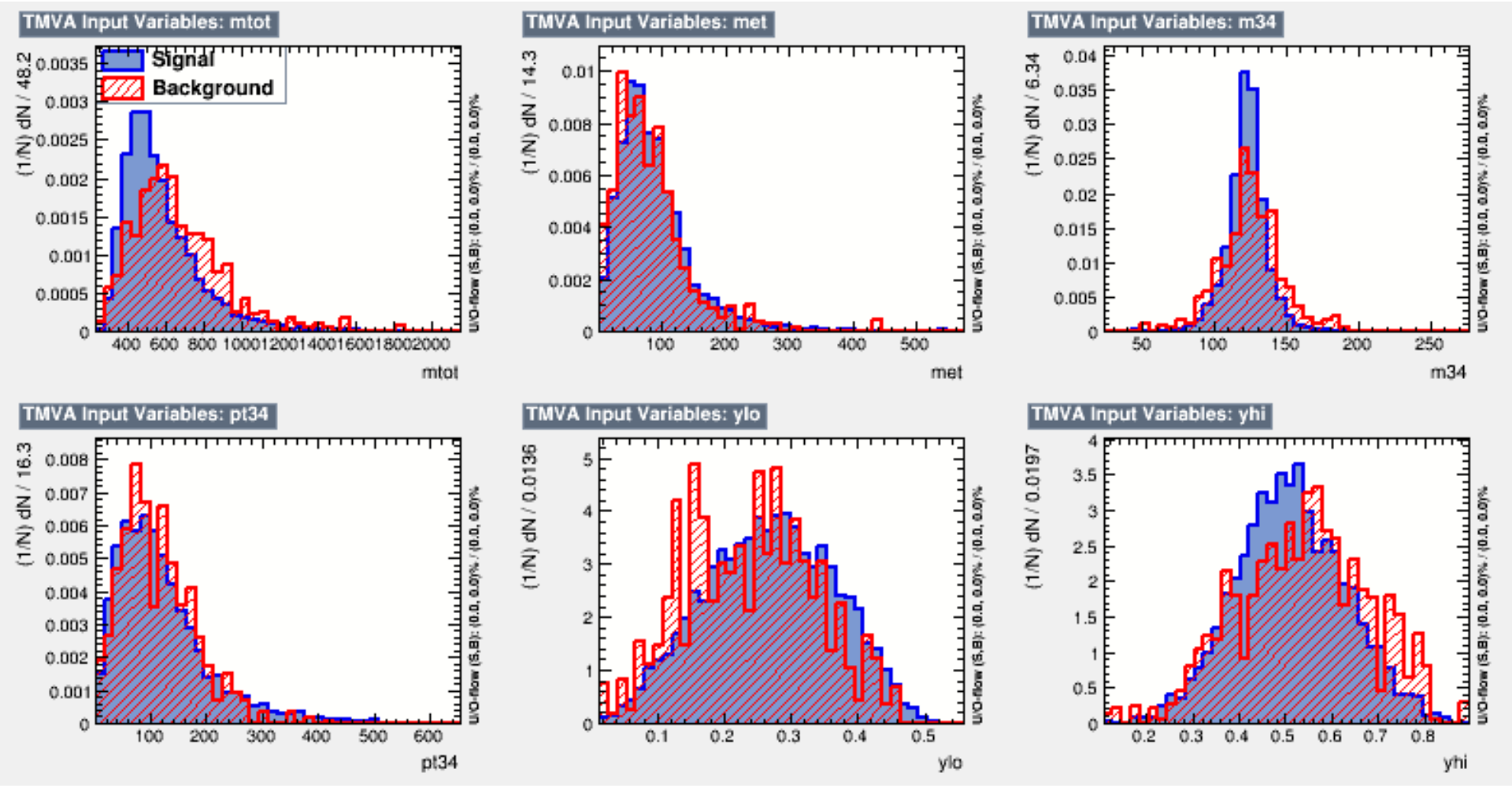
$$s_{tot} = 4.24$$
$$b_{tot} = 124.4$$

So without any further cuts, naively one has discovery sensitivity:

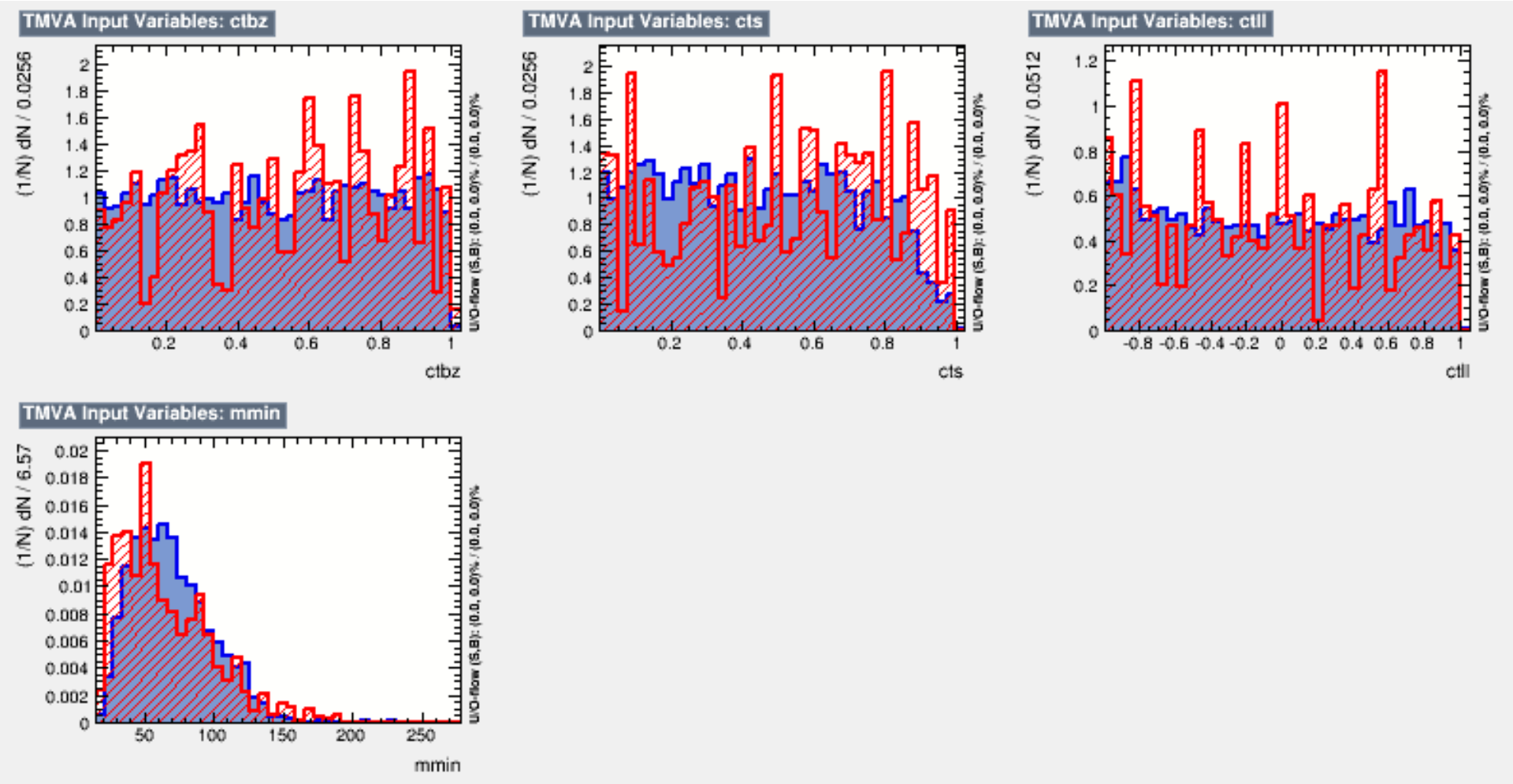
$$s/\sqrt{b} = 4.24/\sqrt{124.4} = 0.38$$

Then define10 variables for input variables to train classifier (shown on next slides; details not important here).

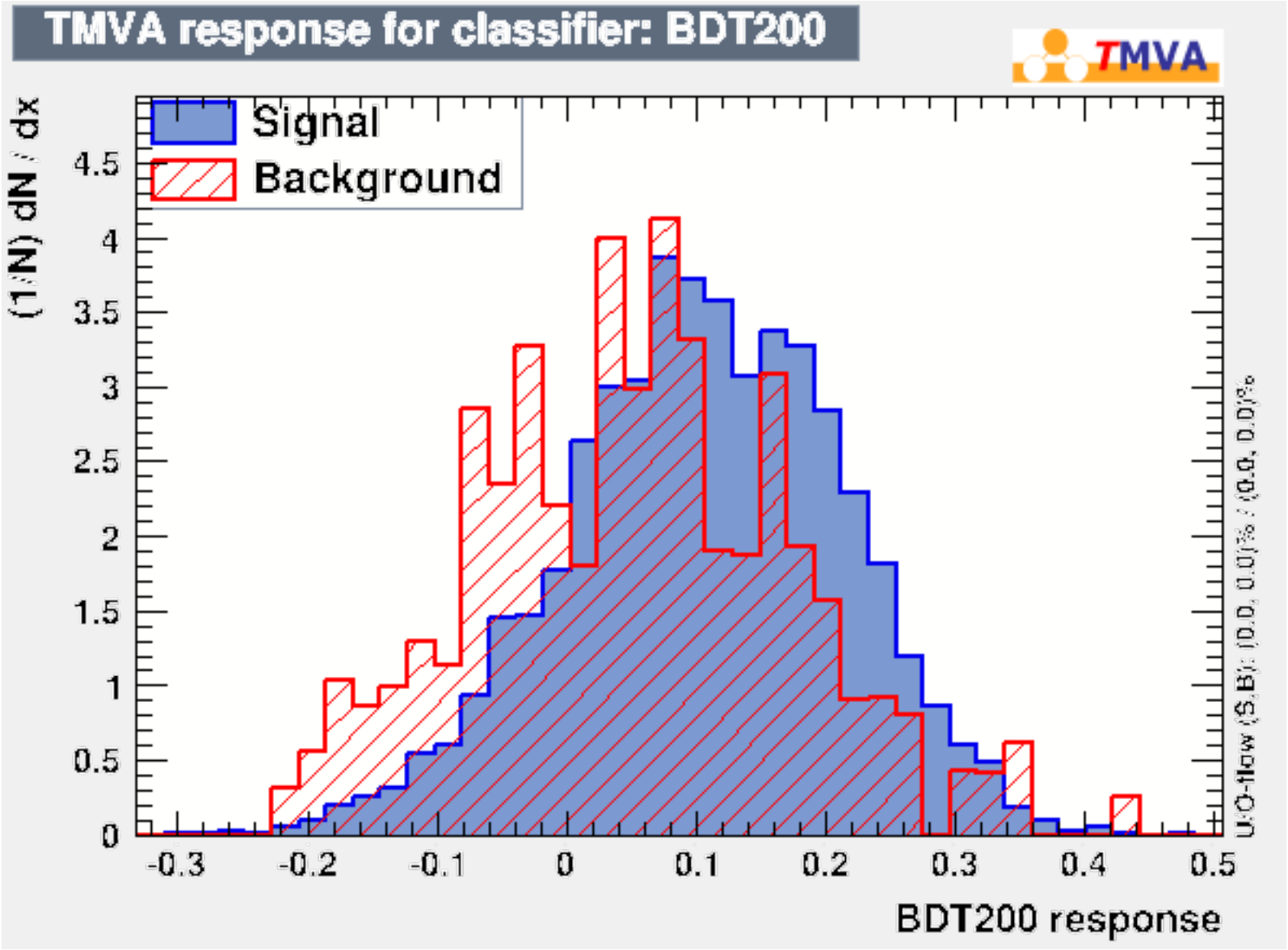
Distributions of input variables



Distributions of input variables



Distribution of MVA output (BDT)



Likelihood ratio statistic for discovery test

In bin i of test statistic t, expected numbers of signal/background:

$$s_i = s_{\rm tot} P(t \in \text{ bin } i|s) \qquad b_i = b_{\rm tot} P(t \in \text{ bin } i|b)$$

Likelihood function for strength parameter μ with data n_1,..., n_N

$$L(\mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}$$

Statistic for test of $\mu = 0$:

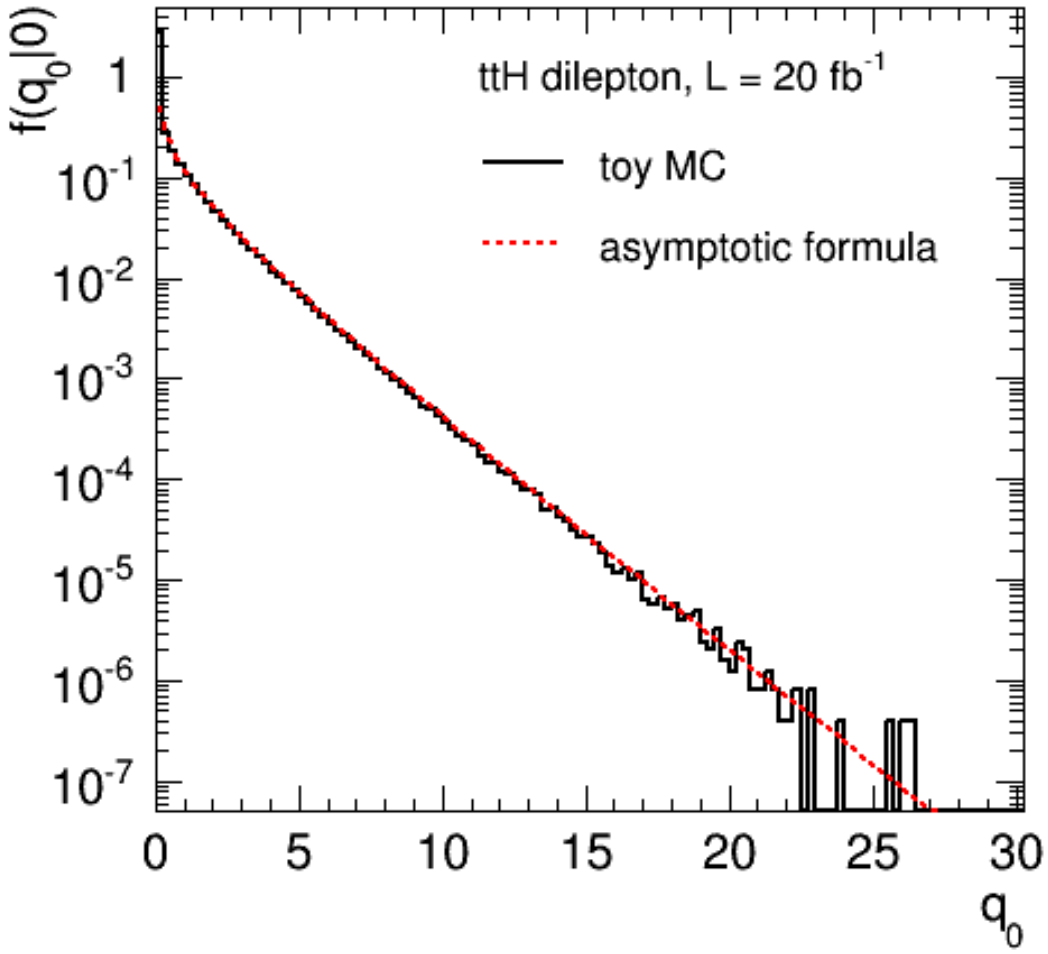
$$q_0 = \begin{cases} -2\ln(L(0)/L(\hat{\mu})) & \hat{\mu} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(Asimov Paper: CCGV EPJC 71 (2011) 1554; arXiv:1007.1727)

Background-only distribution of q_0

For background-only ($\mu = 0$) toy MC, generate $n_i \sim \text{Poisson}(b_i)$.

Large-sample asymptotic formula is "half-chi-square".



Discovery sensitivity

Good agreement between toy MC and large-sample formulae, so OK to use asymptotic formula for significance Z,

$$Z = \sqrt{q_0}$$

Median significance of test of background-only hypothesis under assumption of signal+background from "Asimov data set":

$$n_i \to s_i + b_i$$

gives

med[Z|s+b] = 0.59 (BDT)

med[Z|s+b] = 0.46 (Fisher)

(recall $s/\sqrt{b} = 0.38$)

Multinomial model

Nominal ATLAS analysis uses the information from the distribution of the MVA ouput ("shape information").

No info is taken from the total observed number of events (presumably because the systematic uncertainty on *b* is large).

This corresponds to using a multinomial model for the observed histogram of MVA output values:

$$L(\mu) = \frac{n!}{n_1! n_2! \cdots n_N!} \prod_{i=1}^{N} p_i^{n_i}$$

$$p_i = \frac{\mu s_i + b_i}{\mu s_{\text{tot}} + b_{\text{tot}}}$$

Multinomial model results

The multinomial model gives discovery sensitivities:

$$\text{med}[Z|\text{s+b}] = 0.45 \quad \text{(BDT)}$$

$$\text{med}[Z|\text{s+b}] = 0.27 \quad \text{(Fisher)}$$

Check (recall $s/\sqrt{b} = 0.38$):

$$\sqrt{(0.38^2 + 0.45^2)} = 0.59$$

$$\sqrt{(0.38^2 + 0.27^2)} = 0.47$$

I.e. the $s/\sqrt{b} = 0.38$ represents the contribution from event counting, the second term from the shape information.

Include variables related to event quality

If the number of pile-up-vertices is included as a variable, then

$$\text{med}[Z|\text{s+b}] = 0.447 \rightarrow 0.458 \quad \text{(multinomial, BDT)}$$

i.e., not much help in this case.

Including the Higgs mass resolution for individual events

$$\text{med}[Z|\text{s+b}] = 0.458 \rightarrow 0.530 \quad \text{(multinomial, BDT)}$$

So there does seem to be some scope for improvement.

Perhaps even better if jet resolution includes more info, e.g, EM fraction, presence of leptons, ...

Extra slides

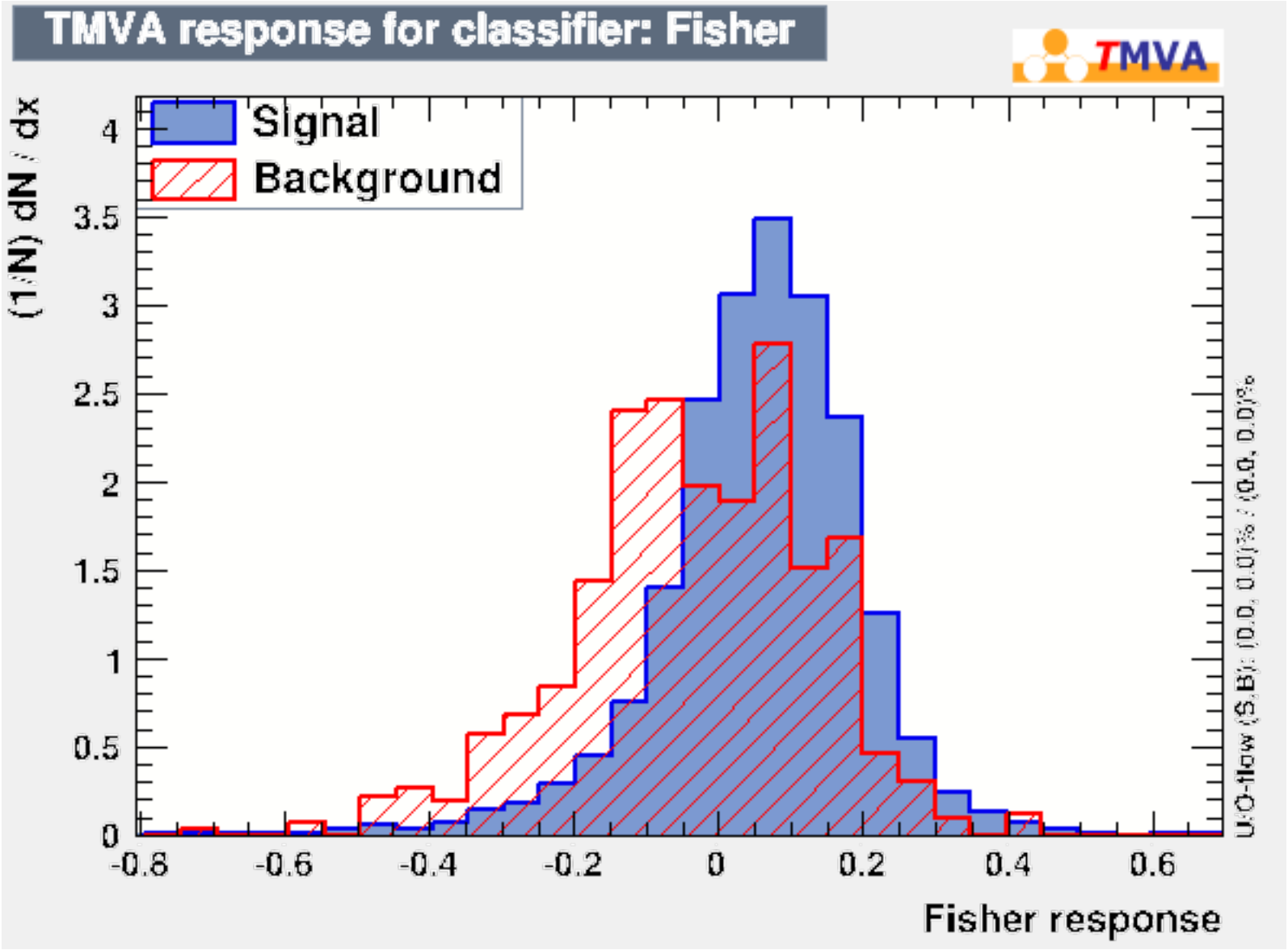
Comment on systematics

If the training data are imperfect, the statistic based on $y(\boldsymbol{x})$ will not give an optimal test (not maximum power wrt $\mu > 1$ for test of $\mu = 0$).

Once the function $y(\boldsymbol{x})$ is fixed, the question of systematics boils down to the uncertainty in distribution of y under the different hypotheses, e.g., $p(y|\mu=0)$ for a test of background-only hyp.

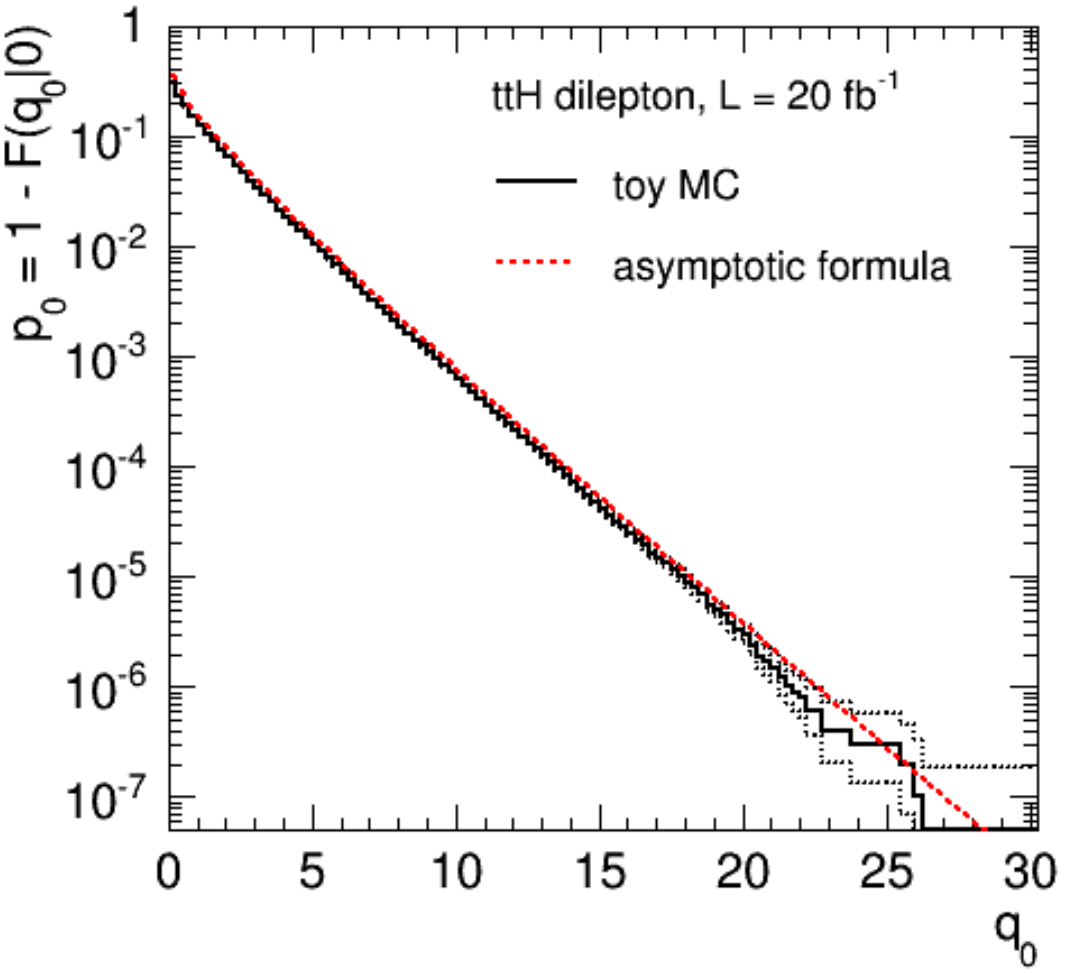
If e.g. $p(y|\mu=0)$ is well determined using a data control sample, imperfections in the MC used to design $y(\boldsymbol{x})$ do not lead to any further systematic error (only sub-optimality of test).

Distribution of MVA output (Fisher)



Background-only cumulative distribution of q_0

p-value is probability, assuming $\mu = 0$, to find q_0 even higher than the one observed (one minus cumulative distribution).



From *p*-value, equivalent significance:

$$Z = \Phi^{-1}(1 - p)$$

Φ^{-1} = standard normal quantile