Statistical Methods for Particle Physics

Tutorial on multivariate methods

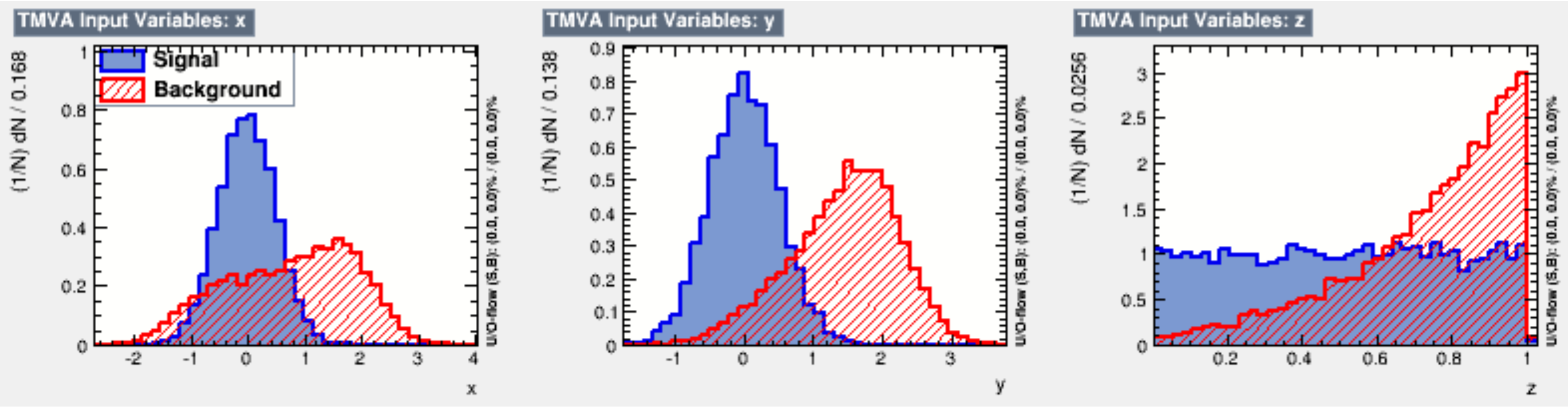
`http://indico.ihep.ac.cn/event/4902/`

iSTEP 2015
Shandong University, Jinan
August 11-19, 2015

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

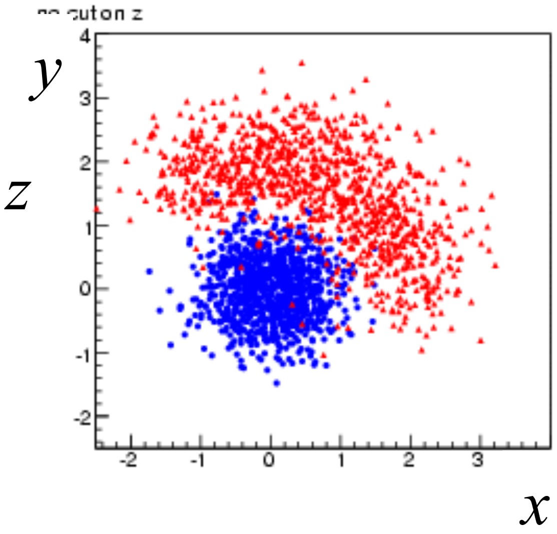
Test example with TMVA

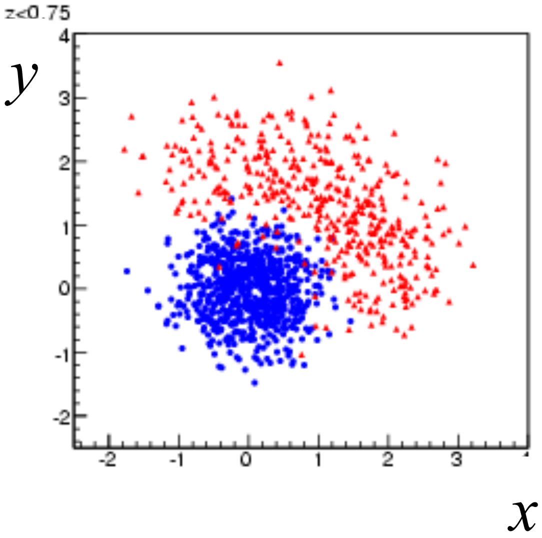
Suppose signal (blue) and background (red) events are each characterized by 3 variables, x, y, z:



Test example (x, y, z)

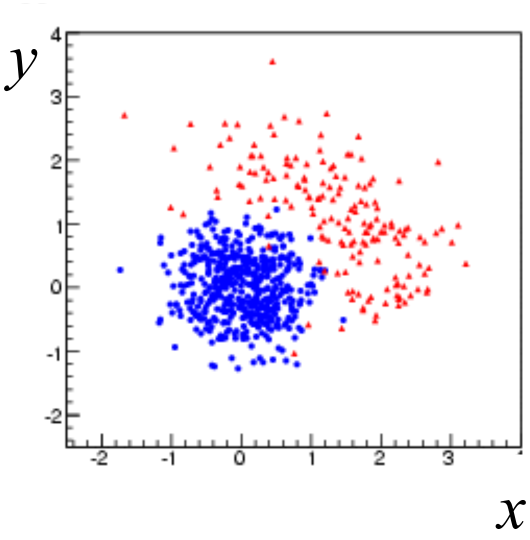
no cut on z

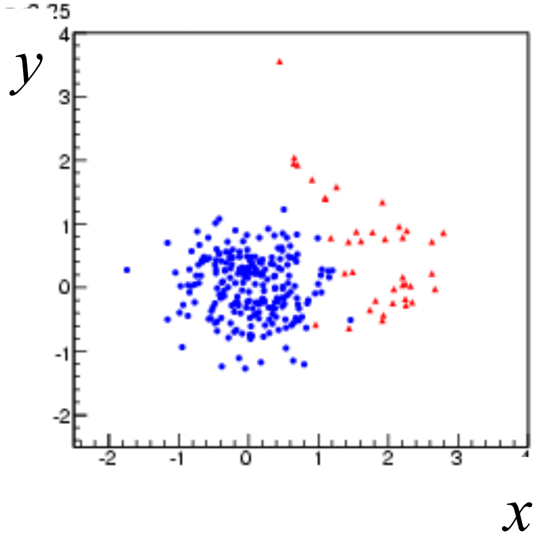




$z < 0.75$

$z < 0.5$





$z < 0.25$

Goal of test example: discovery of signal

We want to define a region of (x,y,z) space using a test statistic t and to search for evidence of the signal process by counting events in this region.

The number of events n that we observe we observe will follow a Poisson distribution with mean $s + b$, where s and b are the expected number of signal and background events. Goal is to maximize the expected significance for test of $s = 0$ hypothesis if signal is present.

We will see (tomorrow) that the expected discovery significance can be estimated using

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

(For $s << b$ this is approximately $s/\sqrt{b}$.)

Code for tutorial

The code for the tutorial is on whale12.hepg.sdu.edu.cn here:

```
/users/hepg/cowan/tutorial/
```

Copy the file `istep2015tmva.tar` to your own working directory and unpack using

```
tar -xvf istep2015tmva.tar
```

This will create some subdirectories, including

```
generate
train
test
analyze
inc
```

Generate the data

cd into the subdirectory generate, build and run the program generateData by typing

```
make
./generateData
```

This creates two data files: `trainingData.root` and `testData.root` an each of which contains a TTree (n-tuple) for signal and background events.

Train the classifier

cd into the subdirectory train, build and run the program generateData by typing

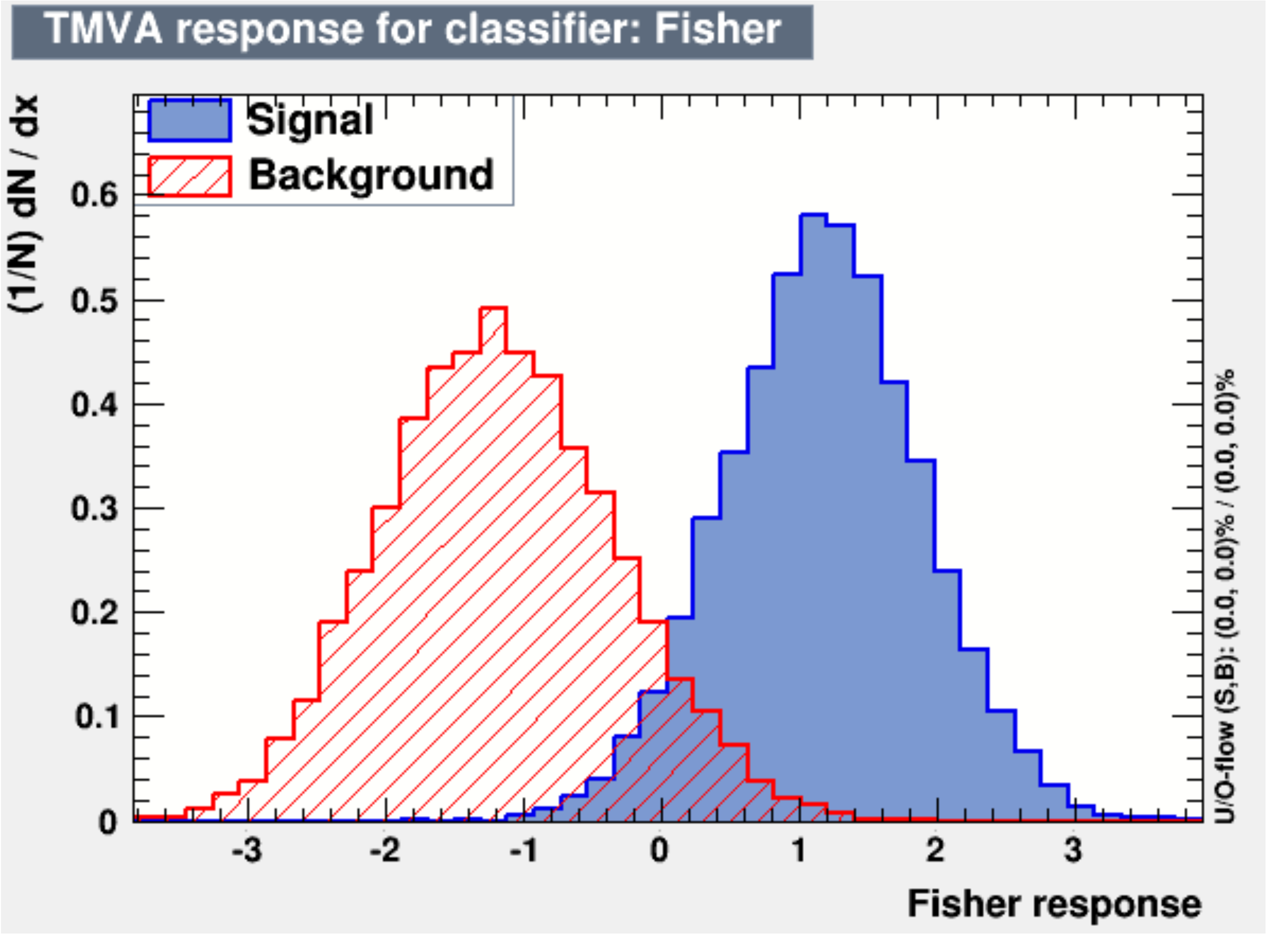
```
make
./tmvaTrain
```

This uses the data in `trainingData.root` to train a Fisher discriminant and writes the coefficients to a file `tmvaTest_Fisher.weights.xml`.

It also creates a file `TMVA.root`. Copy this file into the subdirectory test, and then you call look at various diagnostic histograms using the macros there. Try e.g.

```
.x plotall.C
```

Distribution of classifier output (Fisher)



Analyze the data

cd into the subdirectory analyze, build and run the program analyzeData by typing

```
make
./analyzeData
```

This reads in the data from `testData.root` and selects events with values of the Fisher statistic t greater than a given threshold t_{cut} (set initially to zero). The program counts the number of events passing the threshold and from these estimates the efficiencies

$$\begin{aligned} \varepsilon_{\rm s} &= P(t \geq t_{\rm cut} | {\rm s}) \, , \\ \varepsilon_{\rm b} &= P(t \geq t_{\rm cut} | {\rm b}) \, . \end{aligned}$$

Discovery significance

Suppose the data sample we have corresponds to an integrated luminosity of $L = 20\ \text{fb}^{-1}$, and that the cross sections of the signal and background processes are $\sigma_\text{s} = 0.2$ fb and $\sigma_\text{b} = 10$ fb.

The program then calculates the expected number of signal and background events after the cut using

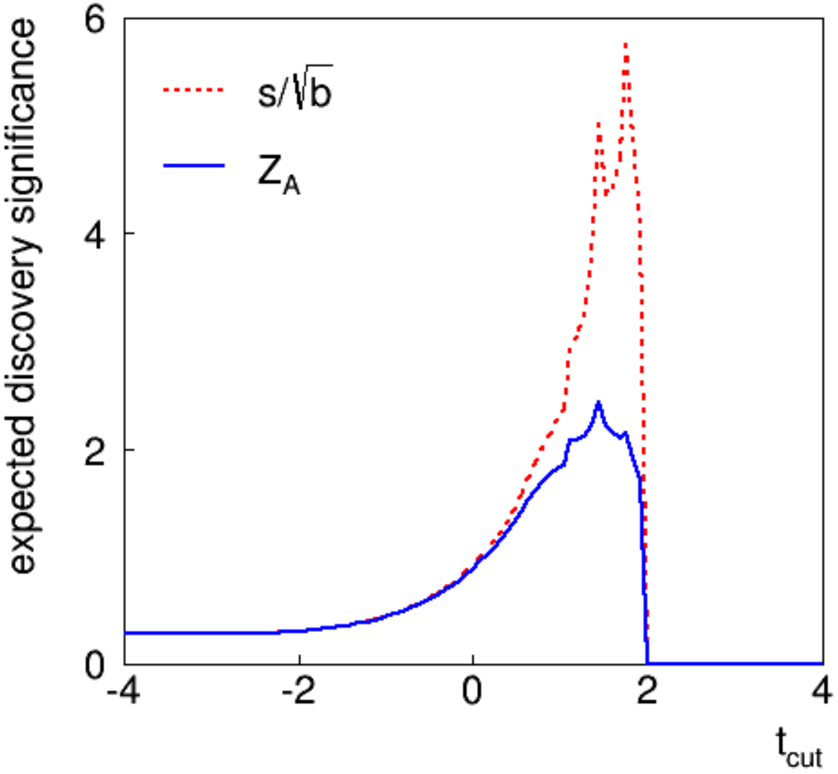
$$\begin{aligned} s &= \sigma_\text{s} L \varepsilon_\text{s} \\ b &= \sigma_\text{b} L \varepsilon_\text{b} \end{aligned}$$

From these numbers we can compute the expected discovery significance using the "Asimov" formula:

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

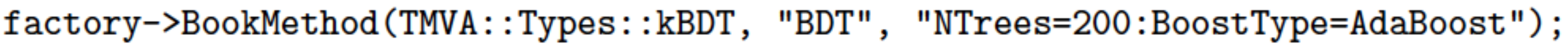
Extending the example

We would like to optimize the value of t_{cut} to obtain the maximum expected discovery significance. To do this, modify the code in `analyzeData.cc` to loop over a range of values of t_{cut}. You should find a plot something like



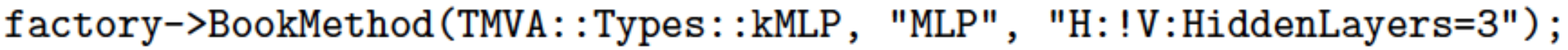
Using different classifiers

The Fisher discriminant is a linear classifier and is not expected to give the optimal performance for this problem. You can try adding classifiers to tmvaTrain.cc by adding the lines



```
factory->BookMethod(TMVA::Types::kBDT, "BDT", "NTrees=200:BoostType=AdaBoost");
```

(adds Boosted Decision Tree with 200 boosting iterations)



```
factory->BookMethod(TMVA::Types::kMLP, "MLP", "H:!V:HiddenLayers=3");
```

(adds a Multilayer Perceptron with a single hidden later contain 3 nodes)

Try using these classifiers to find the maximum expected discovery significance.

You can also try modifying the architecture of the classifiers or use different classifiers (see TMVA manual at **tmva.sourceforge.net**).

The Higgs Machine Learning Challenge

An open competition (similar to our previous example) took pace from May to September 2014 using simulated data from the ATLAS experiment. Information can be found

```
opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014
```

800k simulated ATLAS events for signal ($H \to \tau\tau$) and background (ttbar and $Z \to \tau\tau$) now publicly available.

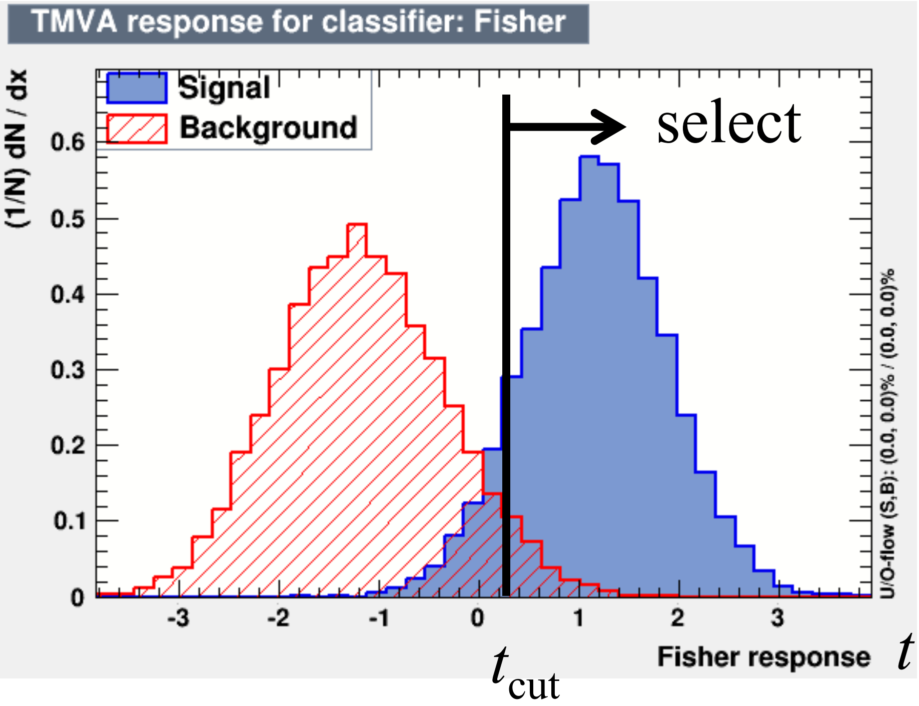
Each event characterized by 30 kinematic variables and a weight. Weights defined so that their sum gives expected number of events for 20 fb^{-1}.

Some (provisional, draft, buggy) code using TMVA is here (for now please only look at train and test; ignore analyze):

```
/users/hepg/cowan/higgsml/
```

Extension of TMVA Project

For the TMVA Project, you defined a test statistic *t* to separate between signal and background events.



You selected events with $t > t_{\mathrm{cut}}$, calculated *s* and *b*, and estimated the expected discovery significance.

This is OK for a start, but does not use all of the available information from each events value of the statistic *t*.

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 $n_1, \ldots, n_N$

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

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)

Discovery sensitivity

First one should (if there is time) write a toy Monte Carlo program to see if the distribution of q_0 follows the asymptotic "half-chi-square" form. For now let us assume this holds, so we can use asymptotic the 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":

$$n_i \to s_i + b_i$$

You can use the Asimov data set to evaluate q_0 and use this with the formula $Z = \sqrt{q_0}$ to estimate the median discovery significance.