Statistical Methods for Particle Physics

Project on H→ττ and multivariate methods

http://indico.ihep.ac.cn/event/5966/

iSTEP 2016
Tsinghua University, Beijing
July 10-20, 2016

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

Extension of TMVA tutorial

The H→ττ Group Project is an extension of the TMVA tutorial.

In the tutorial, we looked at "toy" data where each event had 3 measured variables (x,y,z).

In the H→ττ project, we use 800K fully simulated ATLAS events:

signal: H→ττ

background: Z→ττ and ttbar

Each event is characterized by 30 kinematic variables

The goals of the project are to

1) define a multivariate classifier (MLP, BDT SVM,...) to separate signal from background;

2) define a search region based on the classifier and determine the expected discovery significance for $L = 20\ \text{fb}^{-1}$;

3) If time permits, extend the analysis to multiple bins.

The Higgs Machine Learning Challenge

The data are from a competition organized by ATLAS Physicists and Computer Scientists on `kaggle.com` from May to September 2014. Information can be found

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

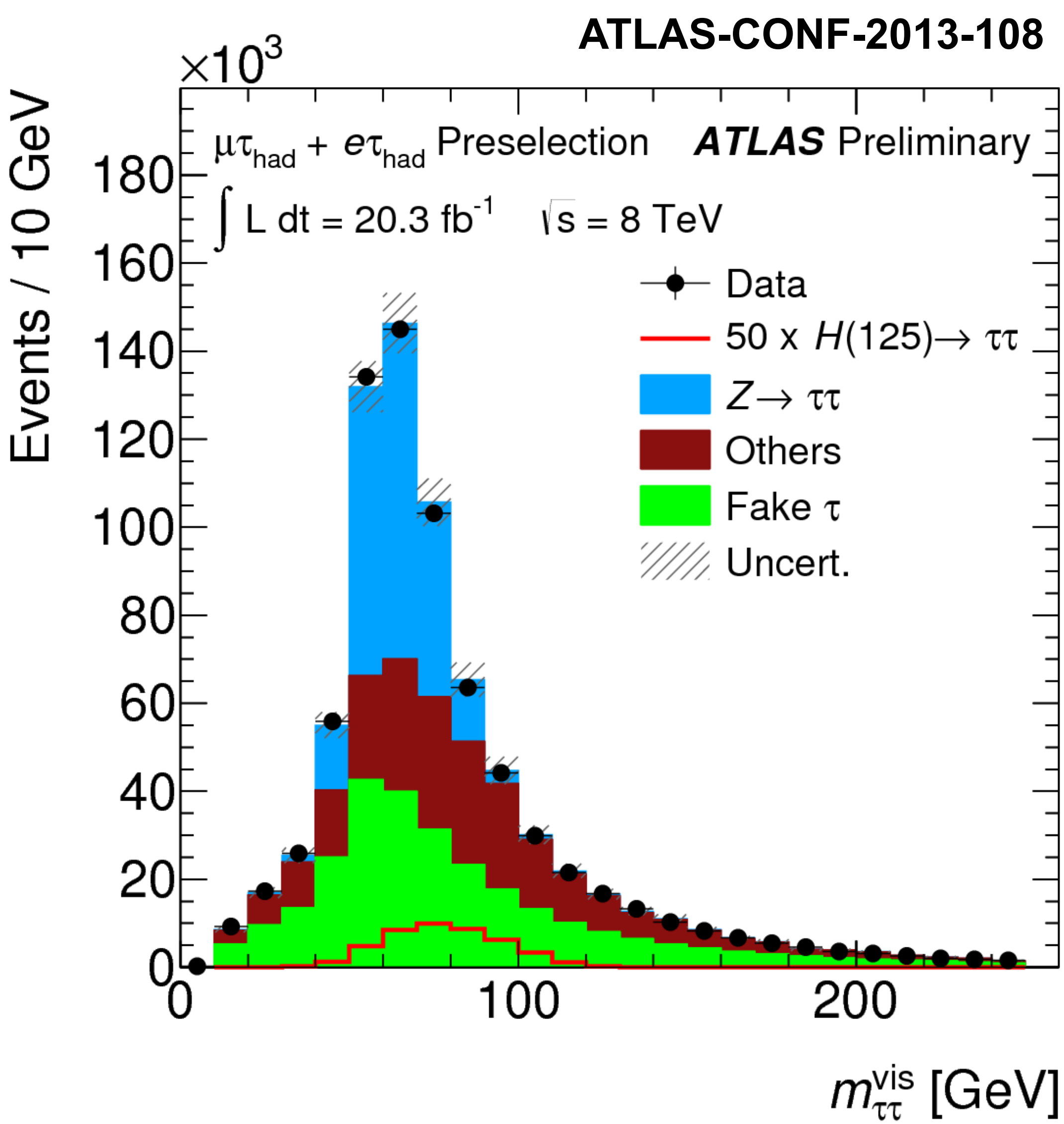
800k simulated ATLAS events for signal ($H \to \tau\tau$) and background (ttbar and $Z \to \tau\tau$) now publicly available.

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 code using TMVA is here (download and unpack as usual with tar –xvf):

`www.pp.rhul.ac.uk/~cowan/higgsml/tmvaHiggsML.tar`

The signal process: Higgs → $\tau^+\tau^-$



4.1 σ evidence

Now superseded by ATLAS paper: *Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector*, arXiv:1501.04943

ATLAS Monte Carlo Data

ASCII csv file converted here to root format with mixture of Higgs to $\tau\tau$ signal and corresponding background, from official GEANT4 ATLAS simulation:

30 variables (derived and "primitive")

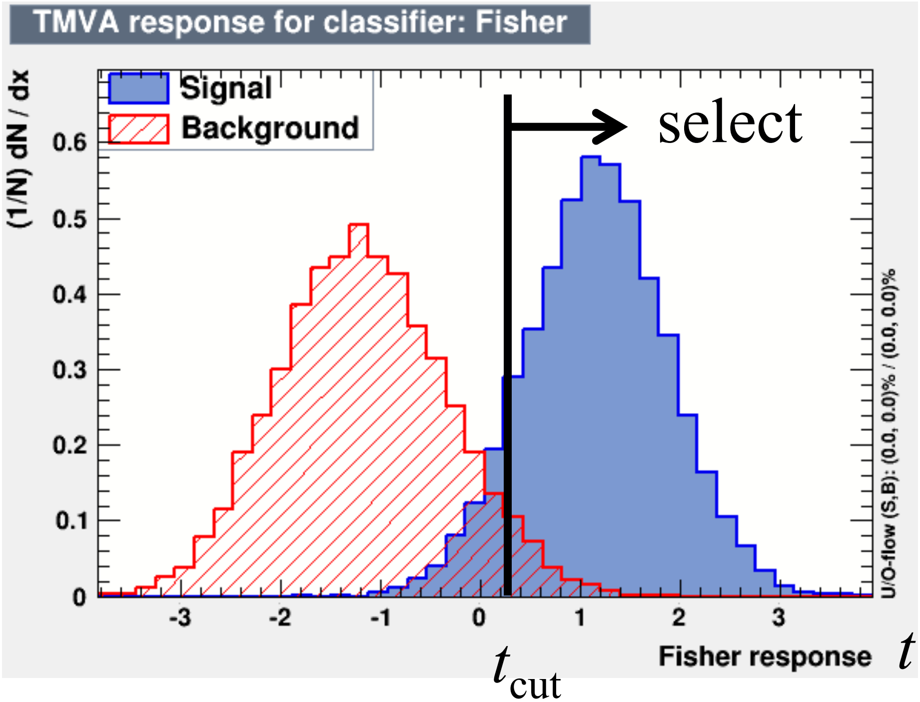
+ true class label (signal = 1, background = 0)

+ weight (sum of weights = expected number of events for 20 fb^{-1})

DER_mass_MMC	DER_pt_ratio_lep_tau	PRI_met_phi
DER_mass_transverse_met_lep	DER_met_phi_centrality	PRI_met_sumet
DER_mass_vis	DER_lep_eta_centrality	PRI_jet_num (0,1,2,3, capped at 3)
DER_pt_h	PRI_tau_pt	PRI_jet_leading_pt
DER_deltaeta_jet_jet	PRI_tau_eta	PRI_jet_leading_eta
DER_mass_jet_jet	PRI_tau_phi	PRI_jet_leading_phi
DER_prodeta_jet_jet	PRI_lep_pt	PRI_jet_subleading_pt
DER_deltar_tau_lep	PRI_lep_eta	PRI_jet_subleading_eta
DER_pt_tot	PRI_lep_phi	PRI_jet_subleading_phi
DER_sum_pt	PRI_met	PRI_jet_all_pt

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 $t > t_{cut}$, 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 event's value of the statistic t.

Binned analysis

Choose some number of bins (~20) for the histogram of the test statistic. In bin *i*, find the expected numbers of signal/background:

$$s_i = \sigma_s L \, P(t \in \text{bin } i | s) \qquad b_i = \sigma_b L \, P(t \in \text{bin } i | b)$$

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 and generate data sets $(n_1,\ldots, n_N)$ following the $\mu = 0$ hypothesis, i.e., $n_i \sim \text{Poisson}(b_i)$ for the $i = 1,\ldots, N$ bins of the histogram.

This can be done using the random number generator **`TRandom3`** (see **`generateData.cc`** for an example and use **`ran->Poisson(bi)`**.)

From each data set $(n_1,\ldots, n_N)$, evaluate q_0 and enter into a histogram. Repeat for at least 10^7 simulated experiments.

You should see that the distribution of q_0 follows the asymptotic "half-chi-square" form.

Hints for computing q_0

You should first show that ln $L(\mu)$ can be written

$$\ln L(\mu) = \sum_{i=1}^{N} [n_i \ln(\mu s_i + b_i) - (\mu s_i + b_i)] + C$$

where C represents terms that do not depend on μ.

Therefore, to find the estimator $\hat{\mu}$, you need to solve

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{N} \left[\frac{n_i s_i}{\mu s_i + b_i} - s_i \right] = 0$$

To do this numerically, you can use the routine **`fitPar.cc`** (header file **`fitPar.h`**). Put fitPar.h in the subdirectory **`inc`** and **`fitPar.cc`** in **`analyze`**. Modify **`GNUmakefile`** to have

```
SOURCES        = analyzeData.cc fitPar.cc
INCLFILES      = Event.h fitPar.h
```

To plot the histogram and superimpose

To “book” (initialize) a histogram:

```
TH1D* h = new TH1D(“h”, “my histogram”,
                        numBins, xMin, xMax);
```

To fill a value **x** into the histogram: `h->Fill(x);`

To display the histogram: `h->Draw();`

To superimpose ½ times the chi-square distribution curve on the histogram use:

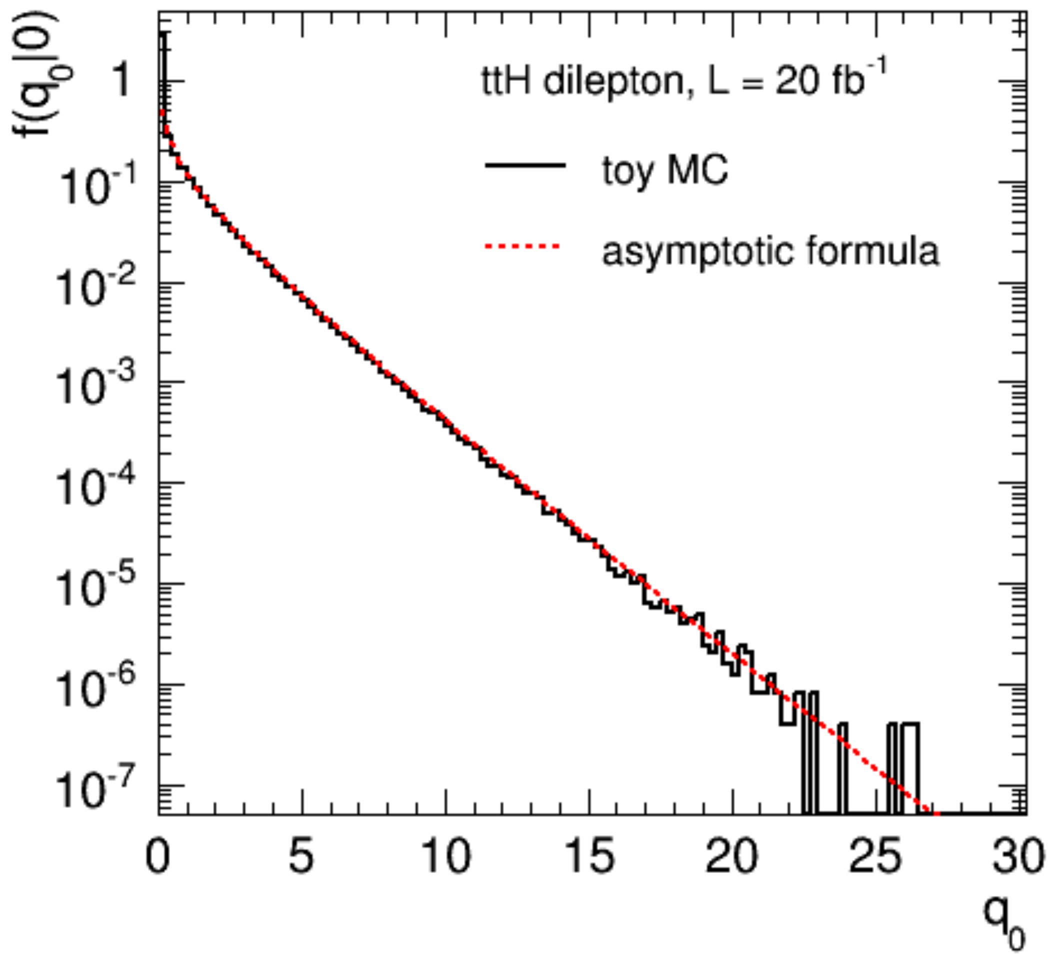
```
TF1* func = new TF1("func", ScaledChi2, 0., 50., 2);
func->SetParameter(0,1.0);       // degrees of freedom
func->SetParameter(1, 0.5);      // scale factor 0.5
func->Draw("same");
```

You can get the function `ScaledChi2.C`. from subdirectory `tools`.

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

Provided that the asymptotic approximation is valid, we can estimate the discovery significance (significance of test of μ= 0)) from the formula

$$Z = \sqrt{q_0}$$

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

$$n_i \rightarrow s_i + b_i$$

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.

This should give a higher significance that what was obtained from the analysis based on a single cut.