# Statistical Tools in Collider Experiments
# Multivariate analysis in high energy physics
# **Exercises**

Nicolas Chanon - ETH Zürich

February 9, 2012

*Goal of these exercises:* be able to estimate the sensitivity of a search for a small peak over a steeply falling background, using multivariate methods. The following problem is inspired by Higgs searches in $H \to \gamma\gamma$ channel at LHC.

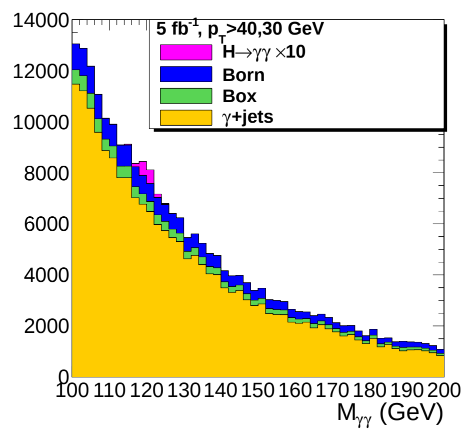


Figure 1: $\gamma\gamma$ invariant mass before photon identification (multijet not generated due to high number of events needed)

*Samples provided :*

| **Process** | **Number of events** | **weight (for 1fb$^{-1}$)** |
|---|---|---|
| Pythia $gg \to H \to \gamma\gamma$ $m_H = 120$ GeV | 100k | 19.0*0.00223/100000. |
| Pythia Born | 100k | 2.0*22.37/100000. |
| Pythia Box | 100k | 2.0*12.37/100000. |
| Pythia $\gamma$+jet | 20M | 1.3*19220./20000000. |

# 1 Exercise 1 : TMVA basics

In this exercise, we will install root and start to use the TMVA package.
To install ROOT (for these exercises, ROOT 5.28 was used), download the binaries from :
`http://root.cern.ch/drupal/content/downloading-root`
Just unpack it and go in the *bin* directory. To setup ROOT, do

```
source thisroot.sh
```

Download the ROOT samples for the exercise from
`http://www.phys.ethz.ch/~pheno/Lectures2012_StatisticalTools/`
Then go in the *tmva/test* directory and copy the samples there.
To look at the samples, first launch root

```
root -l
```

Open the browser with *TBrowser b* and open the Sample.root that you want.

You can check the diphoton invariant mass distribution just by running :

```
root -l PlotHggVariables.C
```

You can modify the selection in the macro *PlotHggVariables.C* provided with the exercises.

Running the training of different classifiers can be done inside ROOT with:

```
.x TMVAClassification.C
```

In the file TMVAClassification.C, different classifiers can be tried (rectangular cuts, likelihood, MLP, BDT, SVM....). See the first lines of the code.
You can choose what variable to use with *factory->AddVariable*.
Lines with *factory->AddSpectator* can be commented
Replace the relevant lines to use the samples provided for the exercises :

```
TFile *inputS = TFile::Open( "Signal.root" );
TFile *inputB = TFile::Open( "Background.root" );

 // --- Register the training and test trees
 TTree *signal     = (TTree*)inputS->Get("Tree");
 TTree *background = (TTree*)inputB->Get("Tree");
```

You can comment *factory->SetBackgroundWeightExpression( "weight" );*
You can specify preselection cuts with

```
TCut mycuts = ""; // for example: TCut mycuts = "abs(var1)<0.5 && abs(var2-0.5)<1";
TCut mycutb = ""; // for example: TCut mycutb = "abs(var1)<0.5";
```

Once a classifier trained, results can be investigated by using the GUI:
*.x TMVAGui.C*
To look at the background rejection versus signal efficiency plot, one can also directly do:

```
root -l TMVA.root
TH1F* histo = _file0->Get("Method_MLP/MLP/MVA_MLP_rejBvsS");
histo->Draw();
```

Other things can be checked in TMVA.root, produced by the traning.

# 2 Exercise 2 : Training a MVA for photon identification

- Using Born sample as signal and $\gamma$+jet as background, train a MLP and a BDT for photon identification. Compare the results with rectangular cuts method. The photon identification variables are : *brem, r9, sumiso03, sumiso04*. You can use the trailing photon ($\gamma$+jet trailing photon is a jet most of the time)
- How are correlated the variables ? What if applying a preselection on the isolation variables ? Any remark on the *brem* variable ?
- How is improved the performance by vetoing events where the trailing photon is in fact a neutral meson ? ($pdgId! = 22$) How is changed the MVA output ?
- By the $N-1$ procedure described in the lecture, find what is the optimal set of variables
- Show if the classifiers are overtrained or not.
- Optimize the architecture of the classifiers
- What is the best background rejection for 90% signal efficiency ?

# 3 Exercise 3 : Application of the multivariate methods in the analysis

*(Needs some knowledge about ROOT)*

- Modifying *TMVAClassificationApplication.C*, produce the samples for $gg \to H \to \gamma\gamma$, Born, Box and $\gamma$+jet including the output of the classifier for both of the photons

- Using the weights for the events in the 4 samples, plot the invariant mass of the diphoton system (you can use the macro *PlotHggVariable.C* provided with the exercises). Compute the significance for Higgs boson observation (can be taken as $S/\sqrt{B}$ computed in a $\pm 5$ GeV window around the Higgs boson mass) as a function of the cut value on the classifier output. Find the value of the cut maximizing the significance.

- Show the invariant mass distribution before and after applying this cut. What is the purity of each process before and after ?

# 4 To go further...

- Train a regression to improve the resolution on the photon energy, using $\eta$, *r9*, *brem*, $p_T$ variables. Apply it to determine the new photon energy and recompute the 4-momentum.

- Train a kinematic classifier with and without invariant mass. Compare the results.

- You can also include as input to this classifier the output of the photon identification.

- Use RooStat package (`https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome`) to compute the exclusion limits and p-values.