TMVA Exercise

Cristóvão Beirão da Cruz e Silva

Instituto Superior Técnico,
Laboratório de Instrumentação e Partículas

cristovao.silva@ist.utl.pt

June 18, 2012

Exercise Outline

Steps of the exercise:

- Train a MVA method to distinguish H→ZZ→4l from SM background
- Run MVA on a soup containing signal + background
- Determine cross section/number of signal events in soup
- Study systematic effects and bias of result

Files

Files for the exercise provided by Pedro Silva.

Files:

- H_ZZ_reco.root → $\sigma_{MC_{Signal}} = 8.4\ fb$
- SM_ZZ_reco.root → $\sigma_{MC_{Background}} = 42\ fb$
- TheStoneSoup.root → $\mathcal{L} = 4.9\ fb^{-1}$

Pre-selection Cuts

Requirements on leptons:

- Isolated - Isolation flag from the datasets
- $P_T > 10\ GeV$
- $|\eta| < 2.5$

Pre-selection efficiency (calculated with the Clopper Pearson method):

$$\epsilon_{Signal} = 0.402601^{+0.002403}_{-0.002399}$$
$$\epsilon_{Background} = 0.587236^{+0.001076}_{-0.001077}$$

Pre-selection Cuts

Requirements on leptons:

- Isolated - Isolation flag from the datasets
- $P_T > 10\ GeV$
- $|\eta| < 2.5$

Pre-selection efficiency:

$\epsilon_{Signal} = 0.4026 \pm 0.0024$
$\epsilon_{Background} = 0.5872 \pm 0.0011$

Event Reconstruction

There are three sub-channels:

- 4 electrons → Order leptons by momentum, pair different charge leptons of highest momentum
- 4 muons → Order leptons by momentum, pair different charge leptons of highest momentum
- 2 electrons + 2 muons → Pair same generation leptons

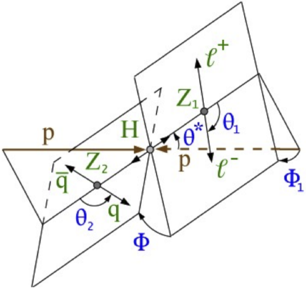
MultiVariate Analysis

Multivariate Analysis involves the analysis of more than one statistical variable at a time (hence the name).

By taking into account the effects of all variables, a better discriminant power (with respect to a cut based analysis) can be obtained.

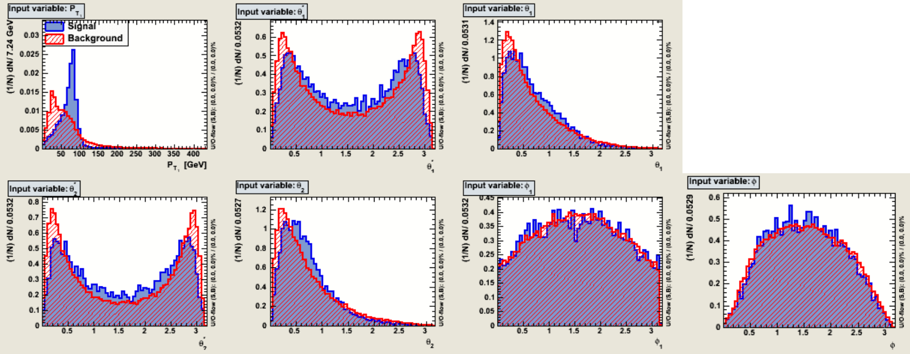
MVA Input Variables

The chosen input variables for the MVA were the P_T of the highest energy Z boson and several angles defined by the decay products.



Angles give insight to the physics process (arXiv)

MVA Input Variables



TMVA permits transformations on the input variables:

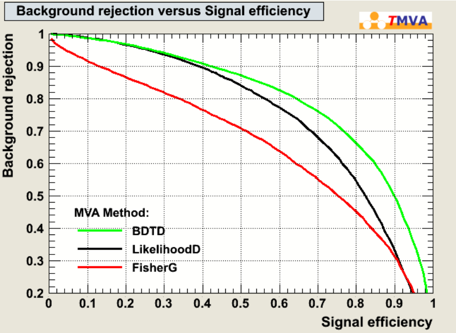
- Decorrelation
- Principal Component Analysis
- Gaussianization

MVA Method

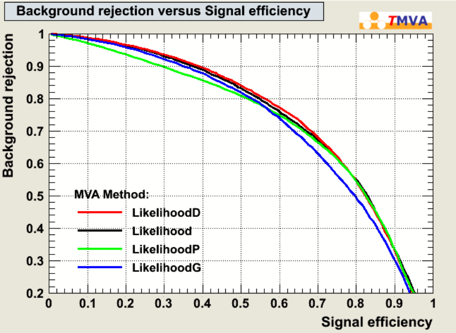
MVA methods:

- Likelihood
- Fisher Discriminant
- Boosted Decision Tree (BDT)

Character at the end describes transformation on input variables

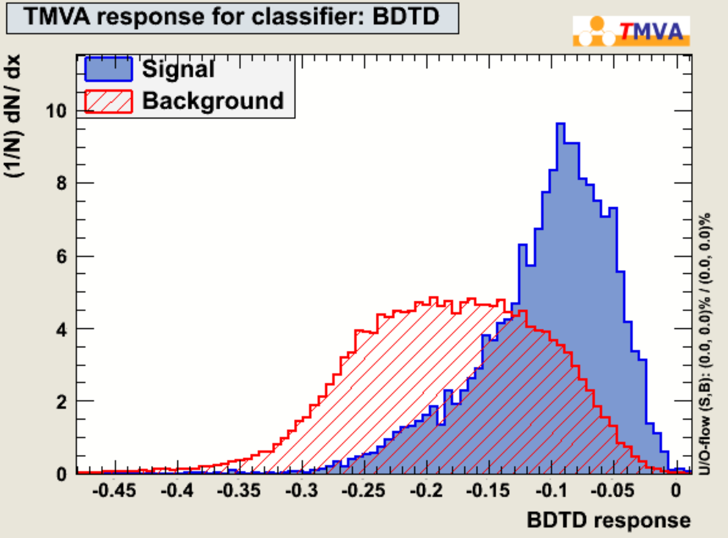


Receiver Operating Characteristic (ROC Curve)



- Illustrates the performance of a binary classifier
- Allows to evaluate performance independently from the working point

BDT Output Distributions



MVA Overtraining

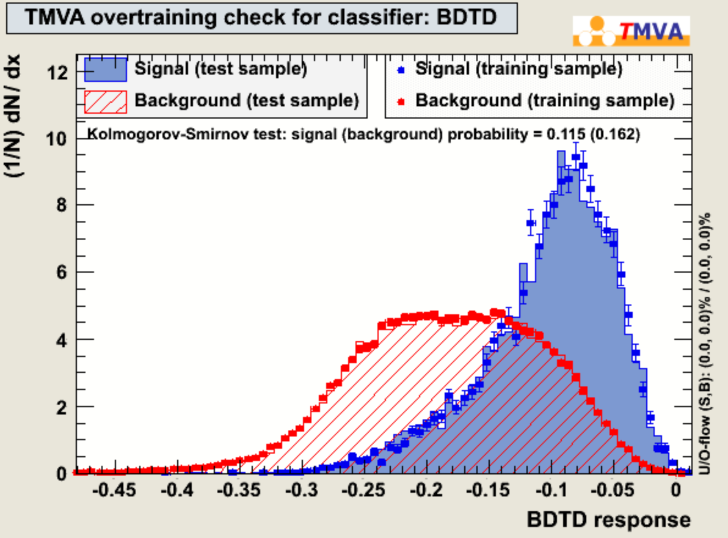
MVA methods are subject to overtraining (some methods more than others).

Overtraining means the algorithm "learned" the statistical fluctuations from the input data.

- The output of the algorithm will be different for different datasets (different performances)
- Hard to predict behavior and difficult to validate

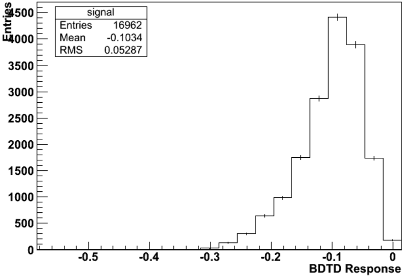
Monte-Carlo samples are split in two, half for training and the other half for validation.

Overtraining Check

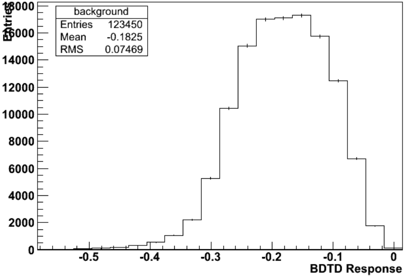


Monte-Carlo Templates

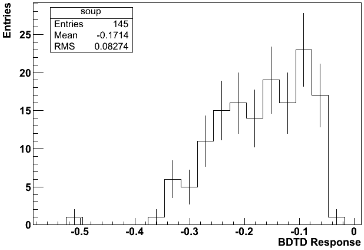
Signal

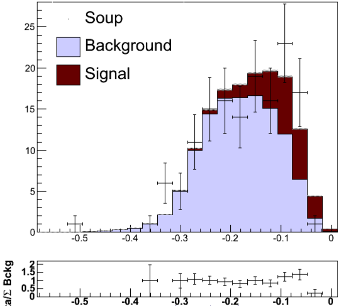


Background



Template Fitting





Template Fitting:

- Signal:
  26.6 ± 11.0 *events*
- Background:
  118.4 ± 14.5 *events*

Events in Soup & Cross Section

$$N_{fit_x} = N_{Soup_x}\,\epsilon_x \implies N_{Soup_x} = \frac{N_{fit_x}}{\epsilon_x}$$

$$N_{Soup_x} = \mathcal{L}\,\sigma_x \implies \sigma_x = \frac{N_{Soup_x}}{\mathcal{L}}$$

$$k = \frac{\sigma_x}{\sigma_{MC_x}}$$

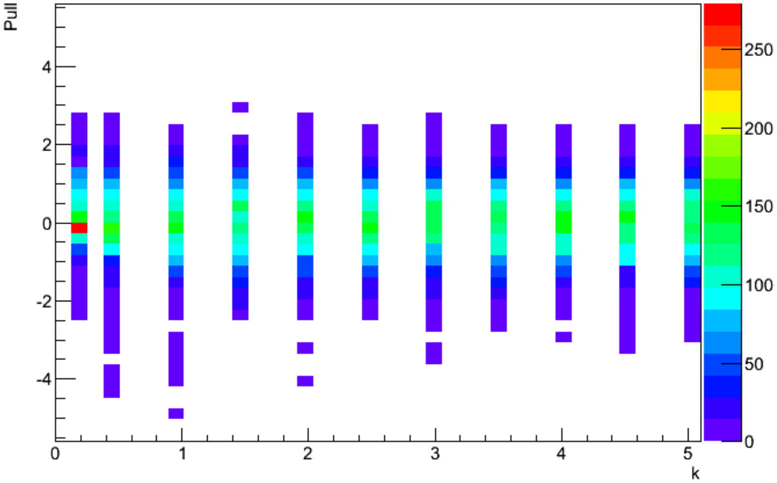
	N_{fit_x}	N_{soup_x}	σ_x (fb)	σ_{MC_x} (fb)	k
Signal	26.6 ± 11.0	66.1 ± 27.3	13.5 ± 5.5	8.4	1.61 ± 0.65
Background	118.4 ± 14.5	201.6 ± 24.6	41.1 ± 5.0	42	0.98 ± 0.12

Bias Study

Procedure:

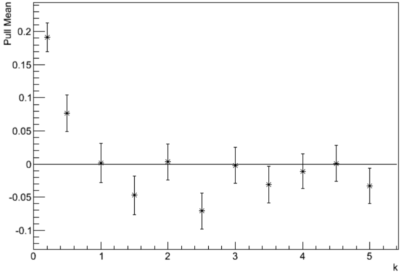
- Take several signal cross sections ($\sigma_{Signal} = k\,\sigma_{MC_{Signal}}$, $k = \{0.2, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}$)
- For each cross section
  - Calculate mean expected events ($\bar{N}_{Signal} = \mathcal{L} \times \sigma_{Signal}$) for signal and background
  - Throw 1000 "toys"
- For each "toy"
  - Sample number of signal events (N_{Signal}) and number of background events (Poisson distribution with mean $\bar{N}_x$)
  - Sample individual events from respective Monte-Carlo datasets (Bootstrapping)
  - Do the template fit to MVA output distribution
  - Calculate pull ($\frac{N_{Signal_{fit}} - N_{Signal}}{\sigma_{Signal_{fit}}}$)

Pull Distribution

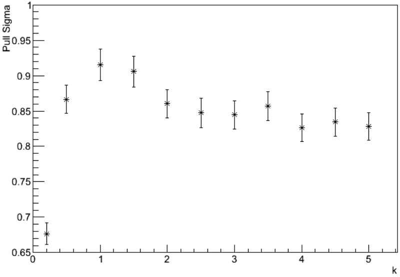


Pull Distribution Details

Pull Mean



Pull Sigma



Systematic Effects

Systematic Uncertainties:

- Lepton Energy Scale:
  - 1% for Muons
  - 2% for Electrons where $|\eta| < 1.442$
  - 3.5% for Electrons where $|\eta| > 1.442$
- 2.2% on Luminosity

Systematic Effects

Nuisance	Variation	$\frac{\epsilon_{Signal_{Nuisance}}-\epsilon_{Signal}}{\epsilon_{Signal}}$ (%)	$\sigma_{Signal_{Nuisance}}$	$\frac{\sigma_{Signal_{Nuisance}}-\sigma_{Signal}}{\sigma_{Signal}}$ (%)
$P_T(e)$	2% 3.5%	Up: 0.000 Down: -0.189	Up: 13.0 Down: 12.6	Up: -3.7 Down: -6.8
$P_T(\mu)$	1%	Up: 0.000 Down: -0.053	Up: 12.4 Down: 13.8	Up: -8.3 Down: 2.3
$\mathcal{L}$	2.2%	-	Up: 13.2 Down: 13.8	Up: -2.2 Down: 2.2
μ_{Pull} $(k = 1.5)$	-	-	-	4.7
Total	-	-	-	11.9

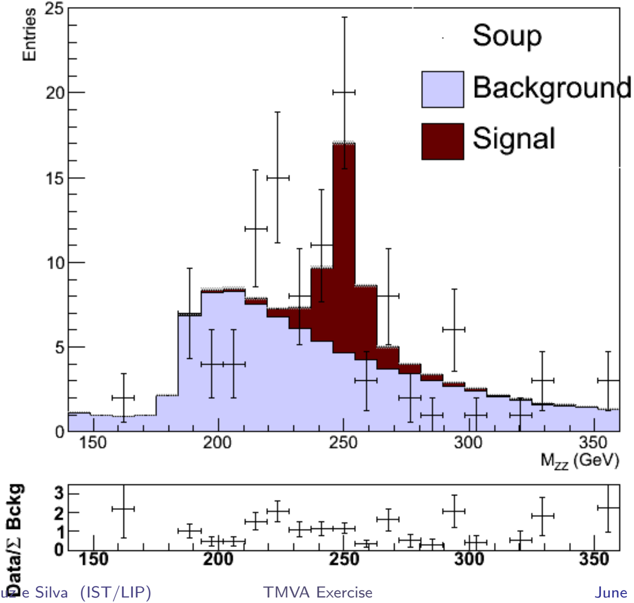
Results

Statistical error on measurement is corrected by the width of the pull distribution ($\sigma_{Pull}(k = 1.5) = 0.91$).
Bias of the pull distribution is considered a systematic error ($\mu_{Pull}(k = 1.5) = -0.047$).

$$\sigma_{H\rightarrow ZZ\rightarrow 4l} = 13.5 \pm 5.0\,(stat.) \pm 1.6\,(syst.)\ fb$$

$$\frac{\sigma_{H\rightarrow ZZ\rightarrow 4l}}{\sigma_{MC}} = 1.61 \pm 0.60(stat.) \pm 0.19(syst.)$$

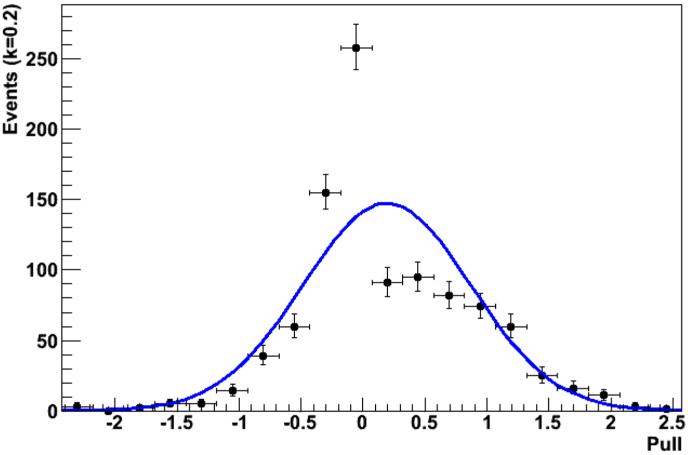
Results


Entries
25
20
15
10
5
0
150
200
250
300
350
M_{ZZ} (GeV)
Soup
Background
Signal
Data/Σ Bckg
3
2
1
0
150
200
250
300
350

Backup

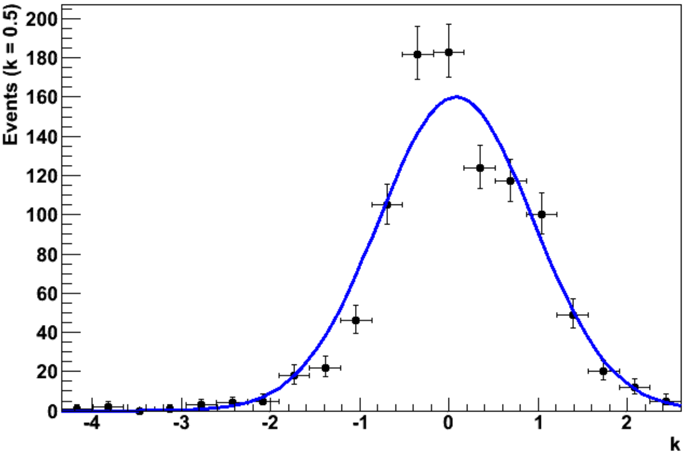
Pull Fits

$k = 0.2$



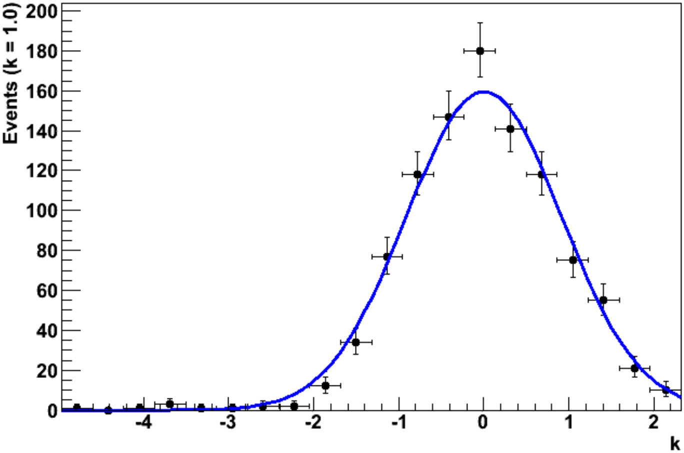
Pull Fits

$k = 0.5$



Pull Fits

$k = 1.0$ (for other values of k, the fits are similar to this one)


Events (k = 1.0)
k