Tau leptons identification via an impact parameter measurement in the ATLAS experiment at LHC

Carlo Pandini

Università degli Studi di Milano

18/03/2013

Introduction

All my past research experiences are within the ATLAS experiment, today I will present only a selected topic:

- **bachelor thesis**: *Tau leptons identification via an impact parameter measurement in the ATLAS experiment at LHC*

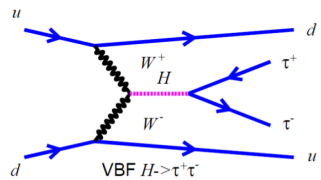
  $\Longrightarrow$ today talk!

- **summer internship** with the Chicago ATLAS group: *Optimization of patterns' bank generation in the Fast Track Trigger FTK for the ATLAS experiment at CERN*

  $\Longrightarrow$ very challenging and interesting topic but rather technical and hard to cover in $\sim$15 min

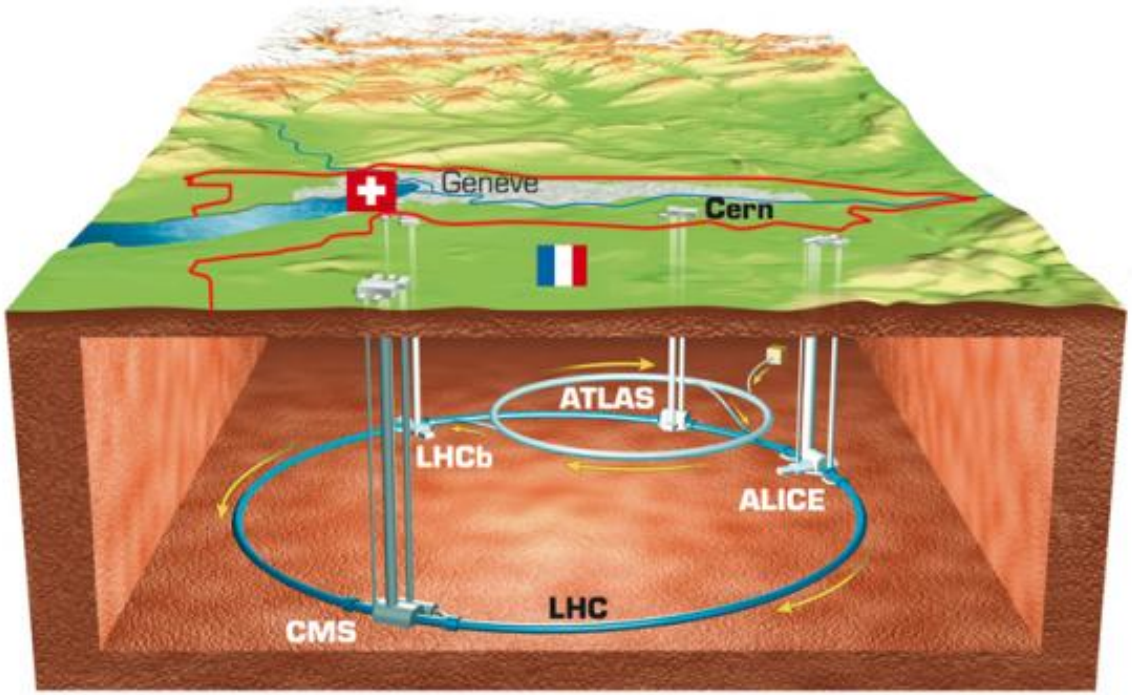
- **master thesis** (work in progress): *Search for the H$\rightarrow \tau\tau$ process with the ATLAS experiment at LHC*

  $\Longrightarrow$ very hot topic: needed to confirm that the new particle found at LHC is actually the source of mass generation $\longrightarrow$ the work is ongoing (preliminary results are still restricted)



If you are interested we can discuss these other works after the presentation.

Schematic overview of LHC and the ATLAS detector



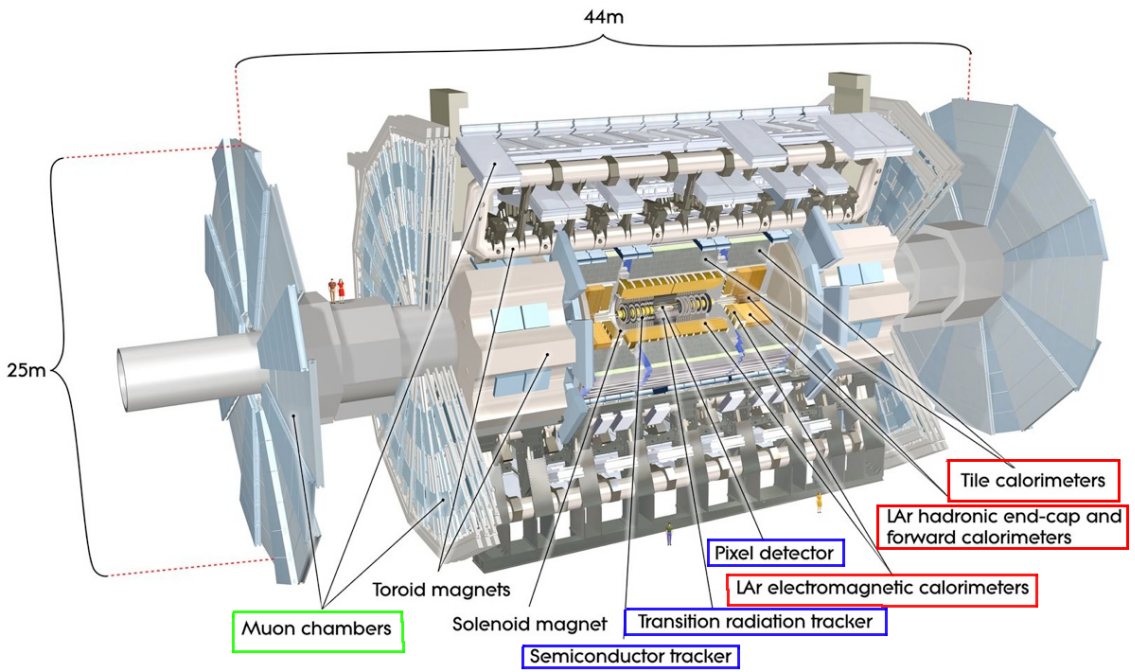
Proton-proton hadronic collider designed to achieve:

- CM energy $\sqrt{s} \approx 14$ TeV
- istantaneous luminosity $\mathcal{L} \approx 10^{34}\ cm^{-2}\ s^{-1}$

Today:

- CM energy $\sqrt{s} \approx 8$ TeV
- istantaneous luminosity $\mathcal{L} \sim 8 \times 10^{33}\ cm^{-2}\ s^{-1}$
- integrated luminosity L (delivered 2012) $\sim$ 29 fb^{-1}

ATLAS: A Toroidal LHC ApparatuS



ATLAS is a multi-purpose detector composed by:

- **Inner detector**
  - reconstruction of charged particles tracks
  - primary and secondary vertex reconstruction
- **Calorimeter system**
  - measurement of electron and jet energy
  - E_T^{miss} reconstruction
- **Muon spectrometer**
  - muon identification and reconstruction

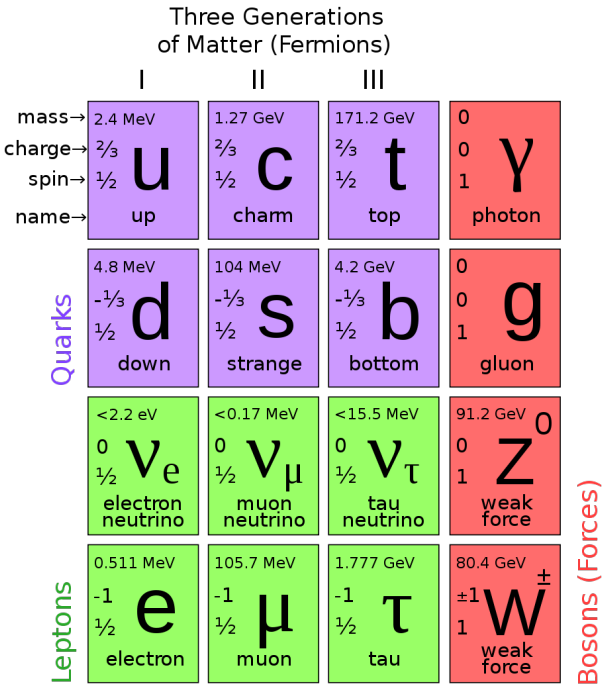
Tau leptons identification via an impact parameter measurement in the ATLAS experiment at LHC

Overview of the presentation:

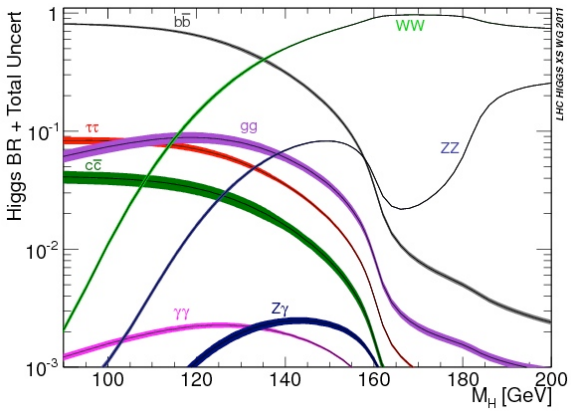
Goal: improvement of τ lepton identification using multidimensional analysis MVA and exploiting the impact parameters of the final state muons

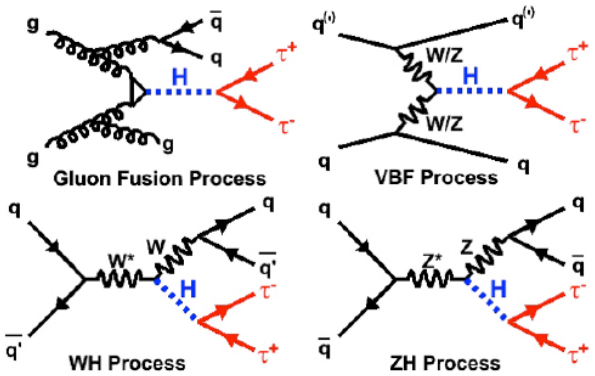
1. main motivations: SM and beyond SM physics
2. physics process studied for τ identification: $Z \rightarrow \tau\tau \rightarrow \mu\mu + \bar{\nu_\mu}\nu_\tau\nu_\mu\bar{\nu_\tau}$
3. discrimination between signal and SM backgrounds
4. estimate of systematic uncertainties
5. results achieved

Main motivations



- SM physics: study of $Z \rightarrow \tau\tau$ process
  - complete the study of the Z decay modes in the leptonic channels (ee, $\mu\mu$, $\tau\tau$)
  - study of an important background process for beyond SM physics
- beyond SM physics: study of $H \rightarrow \tau\tau$ process
  - confirm the discovery of the Higgs-like boson in the τ channel
  - probe the coupling of this new particle with fermions
  - search for the Higgs boson in supersymmetric extension of the Standard Model





Physics process: $Z \rightarrow \tau^+\tau^- \rightarrow \mu^+\mu^- + \bar{\nu}_\mu\ \nu_\tau\ \nu_\mu\ \bar{\nu}_\tau$

Well-known physics channel to study the τ lepton identification:

$$Z \rightarrow \tau^+\tau^- \rightarrow \mu^+\mu^- + \bar{\nu}_\mu\ \nu_\tau\ \nu_\mu\ \bar{\nu}_\tau$$

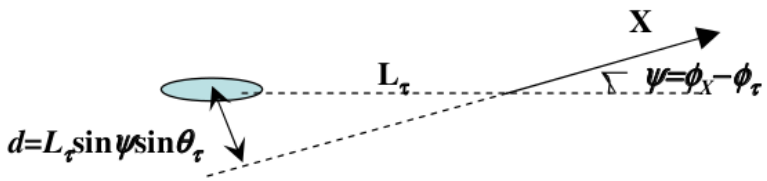
Main issues:

1. intermediate state with τ leptons decaying in a muonic final state $\Rightarrow$ no invariant mass peak to exploit for signal discrimination
2. irreducible background with a cross section higher than the signal's one: $Z \rightarrow \mu\mu$

Solutions:

$\Rightarrow$ powerful discriminating variable involving the τ leptons intermediate state: μ **impact parameter**
(τ mean lifetime $\simeq 0.29ps \Longrightarrow \mu$ with large impact parameter d_0)

$\Rightarrow$ cut based analysis has not enough discriminating power$\rightarrow$ comparative study of several multivariate analysis methods
$\rightarrow$ **best performing MVA**



Standard Model backgrounds & Signal preselection

Signal:

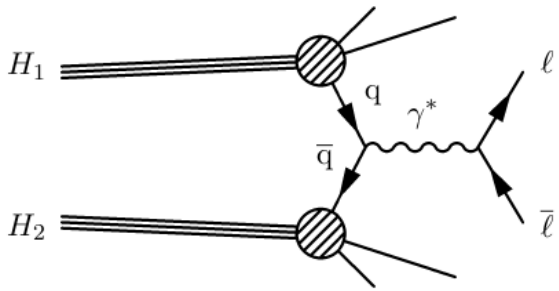
- **Signal process** $Z \rightarrow \tau\tau \rightarrow \mu\mu + 4\nu$

Backgrounds:

- **Drell-Yan process followed by** $Z/\gamma^* \rightarrow \mu\mu$ **decay**
  $\Rightarrow$ MonteCarlo samples
- **QCD multijet background**
  $\Rightarrow$ data-driven technique

Data:

- **Subsample of data collected during 2011 with ATLAS**: L = 638 $\pm$ 22 pb^{-1}



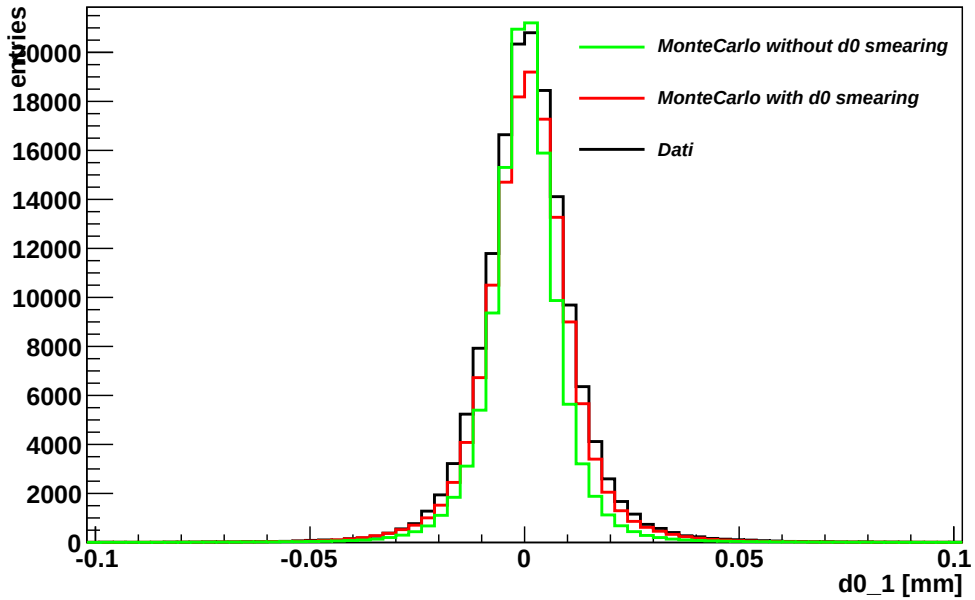
Signal Preselection	
muon trigger selection	
dilepton selection	2μ with opposite charge
transverse momentum cut	$p_T(\mu_1) > 15GeV$
	$p_T(\mu_2) > 10GeV$
invariant mass cut	$25GeV < m_{\mu\mu} < 65GeV$

Data-MC agreement	
N_{data}	27140
N_{signal}	1513
$N_{background_{EW}}$	25208
$N_{signal} + N_{background_{EW}}$	26712

MonteCarlo samples: corrections

The MonteCarlo samples have to be corrected to take into account mismodeling of the detector behaviour in the MonteCarlo simulation:

- impact parameter d(μ) smearing from $Z \rightarrow \mu\mu$ control region:
  $\Longrightarrow$ Control Region: $75GeV < m_{\mu\mu} < 105GeV$ (Z invariant mass peak)
  $\Longrightarrow d_0(MC)' = d_0(MC) + G(\mu, \sigma)$



- impact parameter d(μ) smearing from multiple scattering in the detector
- transverse momentum $p_T(\mu)$ smearing

Input variables for multivariate analysis

The variable choice is based on the signal events' features:

$$Z \rightarrow \tau^+\tau^- \rightarrow \mu^+\mu^- + \bar{\nu_\mu}\ \nu_\tau\ \nu_\mu\ \bar{\nu_\tau}$$

Kinematic variables:

- $\Delta\phi(\mu_1, \mu_2)$
- $\Delta\phi(\mu_1, E_T^{miss})$
- $\Delta p_T(\mu_1, \mu_2)$
- $d_0(\mu_1)$, $d_0(\mu_2)$

$\Rightarrow$ $M_Z(91GeV) \gg M_\tau(1,77GeV)$

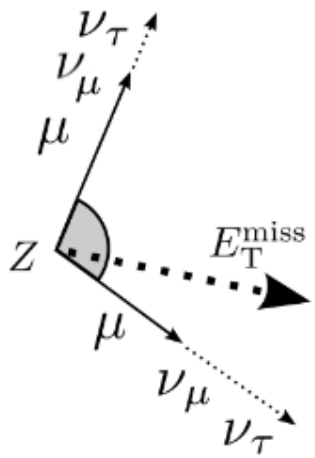
$\Rightarrow$ τ leptons boosted

$\Rightarrow$ τ decay products almost collinear

$\Rightarrow$ Z boson produced at low p_T

$\Rightarrow$ collinear τ decay products back-to-back

$\Delta\phi(\mu_1, \mu_2)$



Input variables for multivariate analysis

The variable choice is based on the signal events' features:

$$Z \rightarrow \tau^+\tau^- \rightarrow \mu^+\mu^- + \bar{\nu_\mu}\ \nu_\tau\ \nu_\mu\ \bar{\nu_\tau}$$

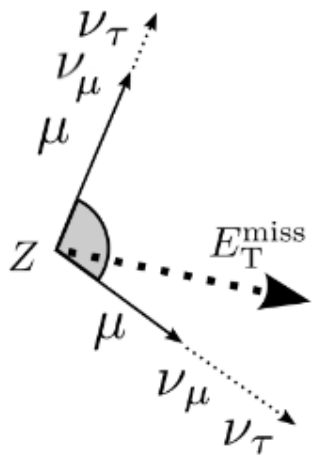
Kinematic variables:

- $\Delta\phi(\mu_1, \mu_2)$
- $\Delta\phi(\mu_1, E_T^{miss})$
- $\Delta p_T(\mu_1, \mu_2)$
- $d_0(\mu_1), d_0(\mu_2)$

$\Rightarrow$ E_T^{miss} dominated by the 4ν contributions

$\Rightarrow$ E_T^{miss} vector must lay between the two μ leptons in the transverse plane

$\Delta\phi(\mu_1, E_T^{miss})$



Input variables for multivariate analysis

The variable choice is based on the signal events' features:

$$Z \to \tau^+\tau^- \to \mu^+\mu^- + \bar{\nu_\mu}\ \nu_\tau\ \nu_\mu\ \bar{\nu_\tau}$$

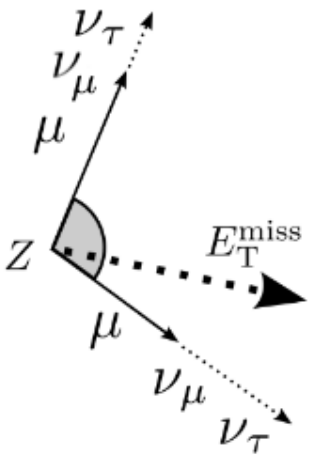
Kinematic variables:

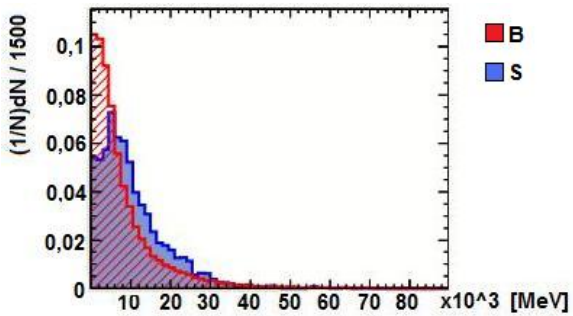
- $\Delta\phi(\mu_1, \mu_2)$
- $\Delta\phi(\mu_1, E_T^{miss})$
- $\Delta p_T(\mu_1, \mu_2)$
- $d_0(\mu_1), d_0(\mu_2)$

$\Rightarrow$ μ from $Z \to \mu\mu$ less boosted in the transverse plane wrt μ from $Z \to \tau\tau \to \mu\mu 4\nu$

$\Rightarrow$ μ from signal events have larger Δp_T wrt μ from background events

$\Delta p_T(\mu_1, \mu_2)$





Input variables for multivariate analysis

The variable choice is based on the signal events' features:

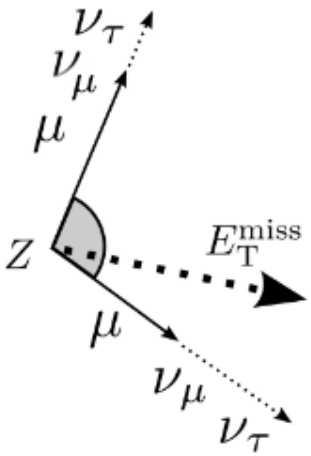
$$Z \rightarrow \tau^+\tau^- \rightarrow \mu^+\mu^- + \bar{\nu_\mu}\ \nu_\tau\ \nu_\mu\ \bar{\nu_\tau}$$

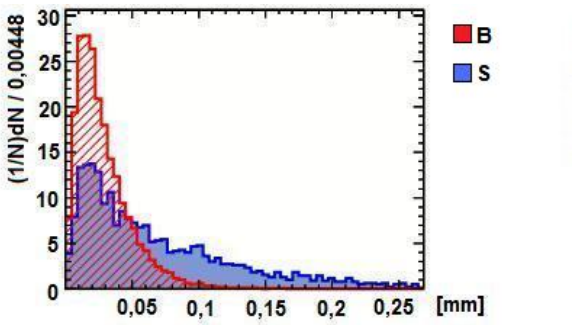
Kinematic variables:

- $\Delta\phi(\mu_1, \mu_2)$
- $\Delta\phi(\mu_1, E_T^{miss})$
- $\Delta p_T(\mu_1, \mu_2)$
- $d_0(\mu_1), d_0(\mu_2)$

$\Rightarrow$ I tested several combination of the μ impact parameters to find the variable with the best discriminating power:

$$d_0^{(1)} + d_0^{(2)} \qquad d_0^{(1)} - d_0^{(2)}$$
$$|d_0^{(1)}| + |d_0^{(2)}| \qquad |d_0^{(1)}| - |d_0^{(2)}|$$
$$|d_0^{(1)} + d_0^{(2)}| \qquad |d_0^{(1)} - d_0^{(2)}|$$





Signal/Background Multivariate discrimination

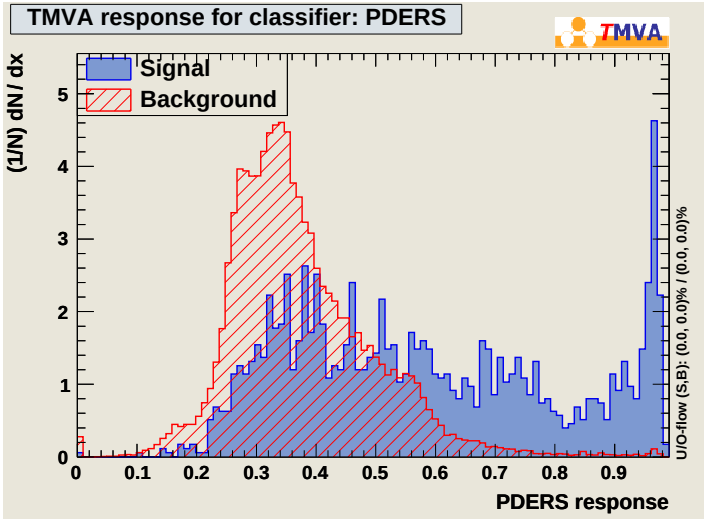
Multidimensional methods common features:

- several input variables $\rightarrow$ cut on a single output variable
- correct treatment of the correlation between variables
- **first step**: training with MonteCarlo samples that reproduce signal and background distributions
- **second step**: application on data samples to solve a classification/regression problem

I implemented several multidimensional methods using the TMVA software (ROOT environment), to find the most suitable for this analysis.

Multivariate analysis methods:

- Rectangular cut method
- One-dimensional Likelihood
- Multi-dimensional Likelihood $\Longrightarrow$
- Artificial Neural Network ANN
- Boosted Decision Trees BDT



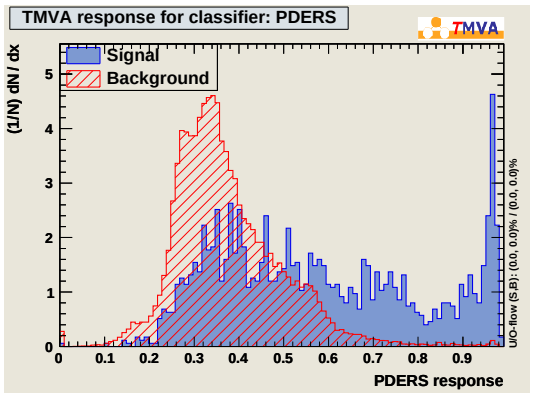
Efficiency study: how to choose the best working point?

How can I choose the best pair (multivariate method, impact parameter variable)?

Statistical significance $\longrightarrow$ $\frac{S}{\sqrt{B}}$

Efficiency study procedure:

- $\Rightarrow$ find the maximum of $\frac{S}{\sqrt{B}}$: working point
- $\Rightarrow$ compute signal efficiency (ϵ_S) and background rejection (r_B) at the working point
- $\Rightarrow$ rank (method, variable) pair



Results

Multivariate method: multi-dimensional Likelihood estimator **PDE-RS**

Impact parameter variable: $|d_0^{(1)}| + |d_0^{(2)}|$

Results achieved

PDE-RS - $\mid d_0^{(1)} \mid + \mid d_0^{(2)} \mid$	
ϵ_S [%]	29 ± 2
r_B [%]	$98{,}6 \pm 0{,}25$
Working point (cut value)	0,696
$\frac{S}{\sqrt{B}}$	5,29
$\frac{S}{\sqrt{S+B}}$	4,13
S	44
B	68
$S + B \pm \sqrt{S + B}$	112 ± 11
Data	91

$\Rightarrow$ high background rejection $\longrightarrow$ low signal efficiency

$\Rightarrow$ signal $\sim$ **5** times background statistical fluctuations

$\Rightarrow$ signal $\sim$ **4** times signal+background statistical fluctuations

$\Rightarrow$ (S+B) in agreement with Data within 2σ (stat.)

Systematic Uncertainties

The systematic uncertainties considered are associated to the discrepancies between data and MonteCarlo:

- resolution effects on muon impact parameter d_0
- multiple scattering effect on muon impact parameter d_0
- resolution effect on muon transverse momentum p_T

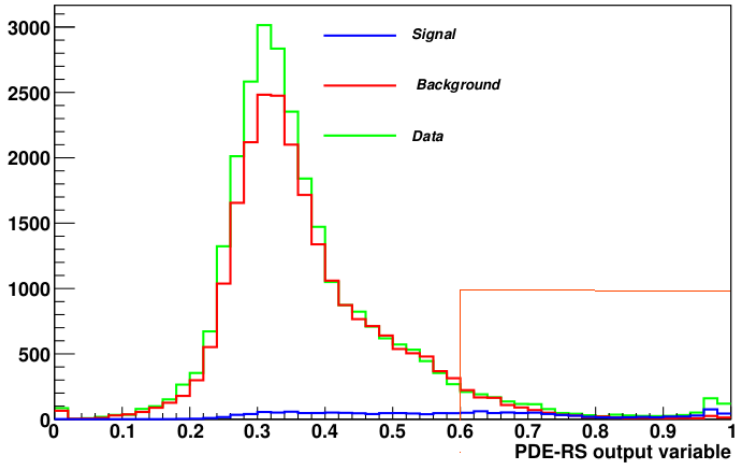
Systematic error for signal efficiency and background rejection:

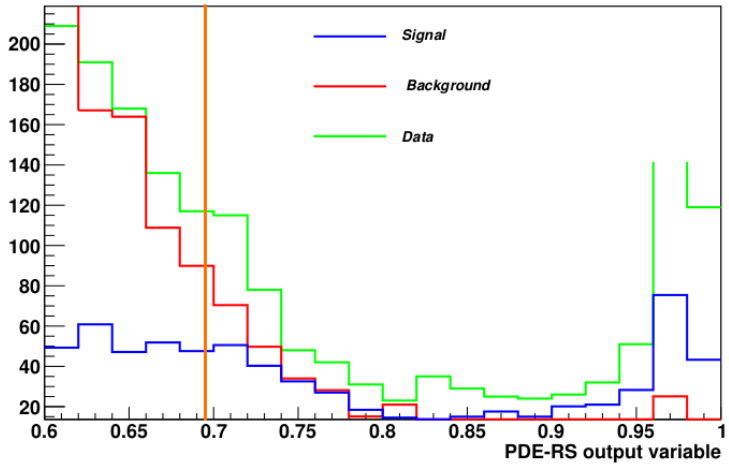
$$\sigma(\epsilon_S, r_B) = \sqrt{\sigma^2(MS) + \sigma^2(p_T) + \sigma^2(d_0)}$$

PDE-RS - $\mid d_0^{(1)} \mid + \mid d_0^{(2)} \mid$	
ϵ_S [%]	29 $\pm$ 2 (syst.)
r_B [%]	98,6 $\pm$ 0,25 (syst.)

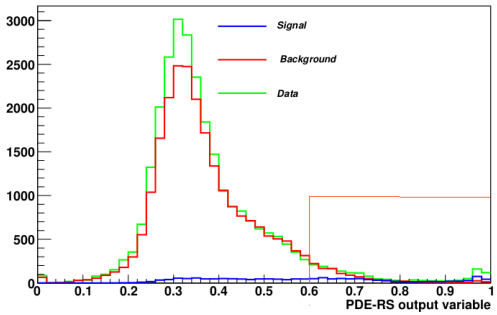
Results achieved

Events distribution vs output variable from multi-dimensional Likelihood method:



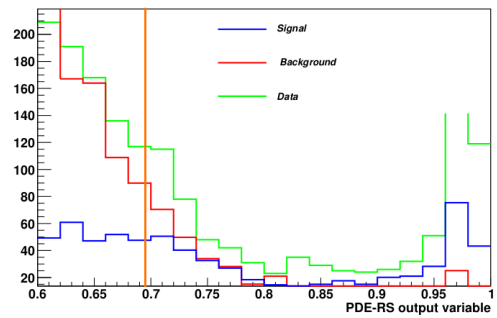


Results achieved



Results:

- best MVA method: PDE-RS multi-dimensional Likelihood
- best impact parameter variable: $| d_0^{(1)} | + | d_0^{(2)} |$
- signal efficiency $\epsilon_S \sim 30\%$
- background rejection $r_B \sim 99\%$

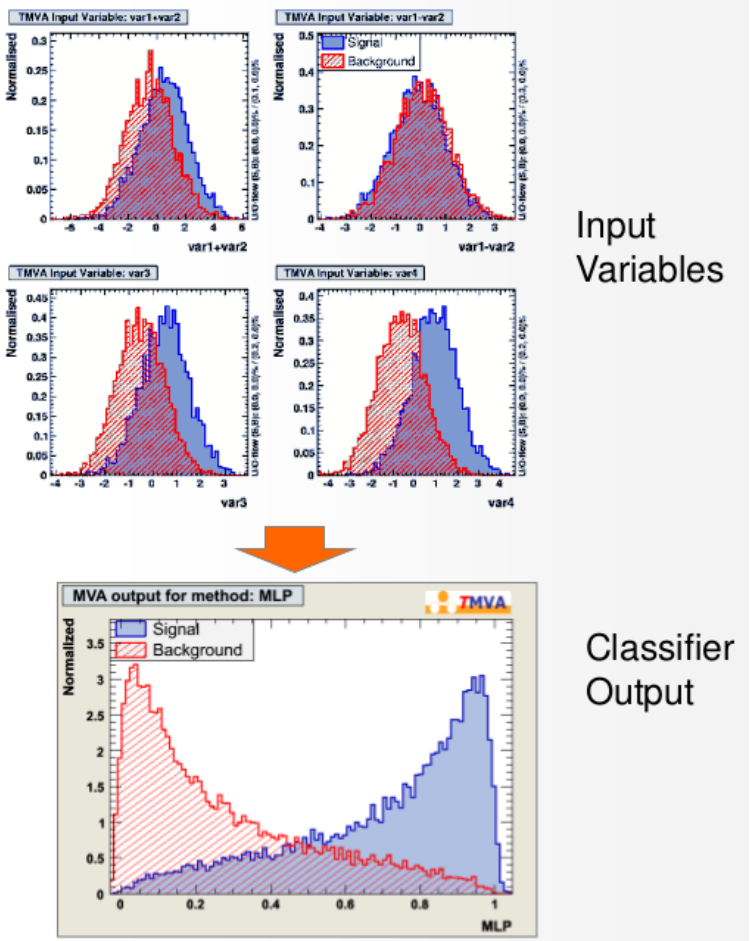


Possible improvements:

- include the SM electroweak and QCD backgrounds that have been neglected
- QCD multijet background: data-driven estimate not satisfying $\Rightarrow$ neglected
- repeat the analysis with full statistics and data from 2012

BACKUP SLIDES

Multivariate Analysis



⟶ multivariate method = multidimensional function
n_{var}**-dim** space $\Longrightarrow$ **1-dim** space

⟶ combine all input variables into one output variable

⟶ treat correctly the correlation between input variables

⟶ the output variable combines the discriminating power of all the input variables

⟶ Supervised learning: the method learns by example extracting patterns from training data:

1. training step (on MonteCarlo samples)
2. application step (on real data)

Multivariate Analysis: Overtraining

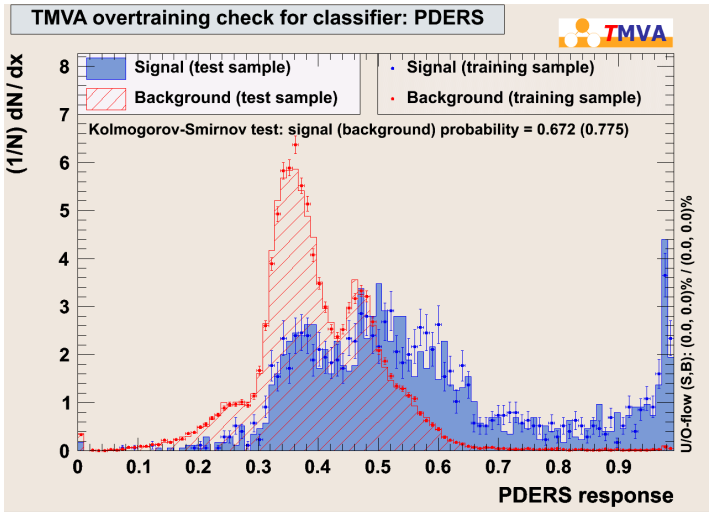
Overtraining $\longrightarrow$ issue for models with too few degrees of freedoms (free parameters):

- the training step is usually repeated until the method satisfy a convergence criteria
- the training can be repeated only a limited number of times before reaching a limit in the performance of the method
- after passing this limit the method is **overtrained**: his complexity increases but his efficiency is fixed

To avoid/check overtraining, the MonteCarlo samples are splitted in :

- training sample: used only for training
- testing sample: used to test the method after training

$\Longrightarrow$ If I get different results between training/test it means that the method is overtrained: it behaves very well on the training sample, but his performance is worse on any other sample (also data!)



Multivariate methods overview: Rectangular cut

Rectangular cut $\Rightarrow$ Set of cuts on the input variables:

- it's the only method that doesn't give a single output variable $\Rightarrow$ direct discrimination between signal and background
- **binary output**
- no correlation treatment for input variables
- no combination of input variables

Best set of cuts $\Longrightarrow$ choosen with a **Genetic Algorithm**

Multivariate methods overview: one-dimensional Likelihood

One-dimensional Likelihood $\Rightarrow$ 1-dim probability density functions (one per each input variable)

$\Rightarrow$ $p_{S(B),k}$ = pdf for the k-variable (S=signal, B=background) - pdfs' shape empirically approximated from training data

$\Rightarrow$ Likelihood function = $L_{S(B)}(i) = \prod_{k=1}^{nvar} p_{S(B),k}(x_k(i))$

Discriminating output variable:

$$y_{L(i)} = \frac{L_S(i)}{L_S(i) + L_B(i)}$$

No correlation treatment between input variables.

Multivariate methods overview: multi-dimensional Likelihood

Multi-dimensional Likelihood $\Rightarrow$ n_{var}-dimensional probability density function defined in the input variables' space

PDE-Range Search (**PDE-RS**) = pdf estimator that classifies each event after a local estimate of his probability density

PDE-RS estimator:

$$y_{PDE-RS}(i, V) = \frac{N_B}{n_B(i, V)} \cdot \frac{n_S(i, V)}{N_S}$$

- $N_{S,B}$ = total number of (signal, background) events
- $n_{S,B}(i, V)$ = number of (signal, background) events in the V volume around the i-event in the training sample
- $y_{PDE-RS}(i, V) \longrightarrow$ 1 signal-like events
- $y_{PDE-RS}(i, V) \longrightarrow$ 0 background-like events

Volume choice: adaptive volume

$\rightarrow$ little V in the regions with high population in the events space

$\rightarrow$ big V in the regions with low population in the events space

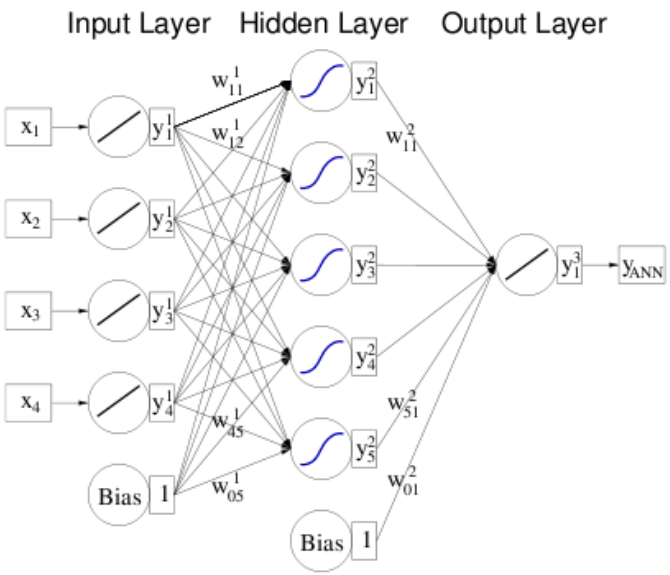
Events' weight: gaussian weight function

$\rightarrow$ events at the boudaries have small weight

$\rightarrow$ their inclusion/exclusion doesn't change too much the value of $y_{PDE-RS}(i, V)$

Multivariate methods overview: Artificial Neural Network

Neural Network ANN $\Rightarrow$ function from n_{var}-dim space to 1-dim space built with a set of interconnected neurons

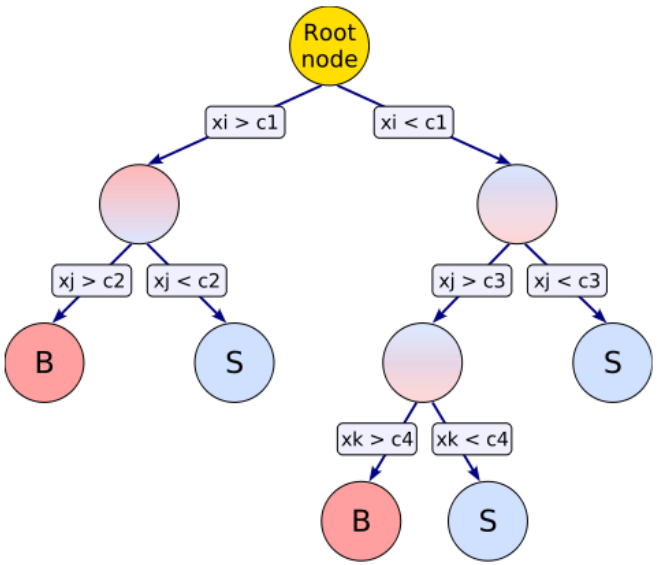


- each neuron is associated with a neuronal function composed by:
  - activation function: required for the neuron activation
  - synapsis function: it gives the output for each neuron set of inputs
- the output from each neuron is associated with a weight
- ANN training: best set of weight to discriminate between S and B

  $\Rightarrow$ **retropropagation algorithm** that minimizes the ANN error function

Multivariate methods overview: Boosted Decision Trees

Boosted Decision Trees BDT $\Rightarrow$ set of binary trees that splits the phase space in several regions, labeled as signal- or background-like by counting the number of events in the training samples



- boosted $\longrightarrow$ set of trees combined in a single classifier

  $\Longrightarrow$ better performances

  $\Longrightarrow$ more stability with respect to fluctuations in the event sample

- trees built from the same training sample with reweighted events

  $\Longrightarrow$ **Boosting** = reweightening procedure to increase statistical stability and discriminating power

  $\longrightarrow$ misidentified events with bigger weight

Impact parameter smearing - d_0 resolution

$Z \longrightarrow \mu\mu$ **Control Region**: $75GeV < M_{\mu\mu} < 105GeV$

- d_0 distribution for the leading muon in data and Montecarlo
- fit of the distributions with a double gaussian (inner and outer)

Smearing Gaussian Function:

$$\Rightarrow \quad \mu = (w_{i,data} \times \mu_{i,data} + w_{o,data} \times \mu_{o,data}) - (w_{i,MC} \times \mu_{i,MC} + w_{o,MC} \times \mu_{o,MC})$$

- $w_{i(o),data} = \dfrac{f(\mu)_{i(o),dati}}{f(\mu)_{i,data} + f(\mu)_{o,data}}$ (the same for MC)
- $f(\mu)_{i,dati}$ = inner gaussian for data distribution evaluated in the mean value

$$\Rightarrow \quad \sigma = \sqrt{\sigma^2_{dati} - \sigma^2_{MC}}$$

- inner σ from inner gaussian
- outer σ from outer gaussian
- $p = \dfrac{g(d_0)_{i,MC}}{g(d_0)_{i,MC} + g(d_0)_{o,MC}}$ probability for the inner σ