Institut de Physique Nucléaire de Lyon
Internship carried out from 2018/03/12 to 2018/07/13

Master 2 internship report

---

# Signal vs background discrimination in $\gamma$+jet events, recorded by the CMS experiment at LHC.

---

*Author :*
Maxime Giraud

*Supervisor :*
Viola Sordini

Version 0.1 du
July 2, 2018

# Contents

# List of Figures

# Acronyms and abbreviations

| | |
|---|---|
| **IPNL** | Institut de Physique Nucléaire de Lyon |
| **CERN** | Centre Européen pour la Recherche Nucléaire |
| **LHC** | Large Hadron Collider |
| **CMS** | Compact Muon Solenoid |
| **MC** | Monte-Carlo |
| **MVA** | MultiVariate Analysis |
| **ANN** | Artificial Neural Network |

# Introduction

First the CMS experiment at CERN will be described then the data that has been used will be described then the technics that has been used will be described then the exploitation will be described

# $\gamma$+jet event classification in LHC collisions

## 1.1 CMS experiment at LHC

The Compact Muon Solenoid (CMS) is a particle physics detector built on the Large Hadron Collider (LHC) at CERN in switzerland and France. The goal of CMS experiment is to investigate the physics beyond the Standard Model. CMS is designed as a general-purpose detector, capable of studying many aspects of proton collisions at 0.9-13 TeV, the center-of-mass energy of the LHC particle accelerator.
It is made of multiple particle detectors designed to measure the energy and momentum of products of the collisions. The first layer called the "Tracker" reconstruct the paths of high-energy muons, electrons and hadrons as well as see tracks coming from the decay of very short-lived particles.
Next the "Electromagnetic Calorimeter" is designed to measure with high accuracy the energies of electrons and photons.
The Hadronic Calorimeter measures the energy of hadrons and provides indirect measurement of the presence of non-interacting, uncharged particles such as neutrinos.
Theses layers all fit inside a large solenoid magnet of 3.8 Tesla, this allows the charge/mass ratio of particles to be determined from the curved track that they follow in the magnetic field. Finally the "Muon detectors and return yoke" are placed outside of the solenoid.

## 1.2 Hadronic jets in proton-proton collisions

In particle physics, jets are the experimental signatures of quarks and gluons produced in high-energy processes.
These particles having a net colour charge cannot exist freely due to couloour-confinement, thereby they are not directly observed in nature. Instead, they come together to form colour-neutral hadrons by a process called hadronisation that leads to a collimated spray of hadrons called a jet. The detailed understanding of both the jet energy scale and of the transverse momentum resolution is of crucial importance for many physics analyses.

# Collision data

In this chapter will be described the various sources of data, and input variable that has been used during this work.

## 2.1 Monte-Carlo simulation

?

## 2.2 CMS data

Run 2 at $\sqrt{s} = 13TeV$ for an integrated luminosity of $36fb^{-1}$ number of events ? slide de hugues ?

## 2.3 MVA variables

In order to perform a multivariate analysis we used multiple variables representing various aspects of reconstructed photons :

**Isolation variables** represent additional objects (photons,charged hadron and neutral hadron) reconstructed in a $\Delta R$ radius cone around the processed photon. These variables permit to discriminate between isolated prompt photons and neutral pions within a jet.

**Charged Hadron isolation (CHiso) :** $I_{cha} = \sum_{cha_i}^{\Delta R} p_{T,cha_i}$
$cha_i$ corresponds to reconstructed charged hadron.

**Neutral Hadron isolation (NHiso) :** $I_{neu} = \sum_{neu_i}^{\Delta R} p_{T,neu_i}$
$neu_i$ corresponds to reconstructed neutral hadron.

**Photon isolation (Photoniso) :** $I_{\gamma} = \sum_{\gamma_i}^{\Delta R} p_{T,\gamma_i}$
$\gamma_i$ corresponds to reconstructed photons, the sum doesn't account for the $p_T$ of the processed photon. (parler du pile-up avec $\rho$ ?)

**Shape variables** represent deposited energy shape in the ECAL.

**$\sigma_{i\eta i\eta}$ :** Energy weighted spread within the 5x5 crystal matrix centred on the crystal with the largest energy deposit in the supercluster. Obtained by measuring position by countinig crystals.
$\sigma_{i\eta i\eta} = \sqrt{\frac{\sum_j^{5x5} \omega_j (i\eta_j - i\eta_{seed})^2}{\sum_j^{5x5} \omega_j}}$
$i\eta$ is the crystal index at position $\eta$and $\omega_i$ is a weight representing the expected energy deposit measured.
$\omega_i = b + ln(\frac{E_i}{E_{5x5}})$

**$\sigma_{i\varphi i\varphi}$ :** same variable as $\sigma_{i\eta i\eta}$ but computed in the $\varphi$direction.

**$\sigma_{i\eta i\varphi}$ :** is the covariance between $\sigma_{i\eta i\eta}$ and $\sigma_{i\phi i\phi}$

**$\eta_{width}$ $\gamma$:** Shower width in $\eta$

**$\varphi_{width}$ $\gamma$:** Shower width in $\varphi$

**$R_9$ $\gamma$:** Energy sum of the 3x3 crystals centred on the most energetic crystal in the supercluster divided by the supercluster's energy. Lower values of $R_9$ for converted photons than those of unconverted photons.

**Had/Em :** Hadronic calorimeter energy deposit over Electromagnetic calorimeter energy deposit

**$E_{nxm}/E_{5x5}$ :** Energy of most energetic nxm crystal set over energy of 5x5 crystal set

**$\rho$:** Pile-up energy, median of the transverse energy density per unit area.

# Input variable analysis

A large set of variables is available from CMS data but MVA training can be time consumming and the "curse of dimensionality" [1] forces us to select only a few of them based on two main criteria :

**Background vs Signal discrimination :** Variables with most differences of shape for background and signal will be picked.

**Low correlation between variables :** Needed in order to reduce redundancy of input data and thus will permit to reduce MVA complexity (for example number of hidden neurons in ANN).

The MVA will be trained with MC simulation for the signal sample and with the real data for the background sample. In order to do that we need to perform a data-driven background estimation using a low-correlated variable for a sideband definition.

## 3.1 Background vs Signal discrimination

It is necessary to pick the smallest set of input variable for the MVA. This selection is done by looking at variable shape for background and signal data from MC simulation.
Then a cross-check of the variables shape has to be done between Data and MC.
Here an example of MC simulation and data comparison for *neutral hadron isolation* variable.(fig. 3.1)

## 3.2 Variable correlations

Training data needed-quantity increases with network complexity, so correlation between variables must be avoided in order to get the minimum redundancy.
One of these variables has to be used for the data-driven background estimation, by looking at the correlation matrix (fig. 3.2) we can see that *charged hadron isolation* is one good candidate and so will be used next for the sideband definition.//

[1] Curse of dimensionality refers to problems that commonly arise when analyzing high-dimensionality data. Increasing dimensionality lead to an increase of volume and so tends to scatter data points.

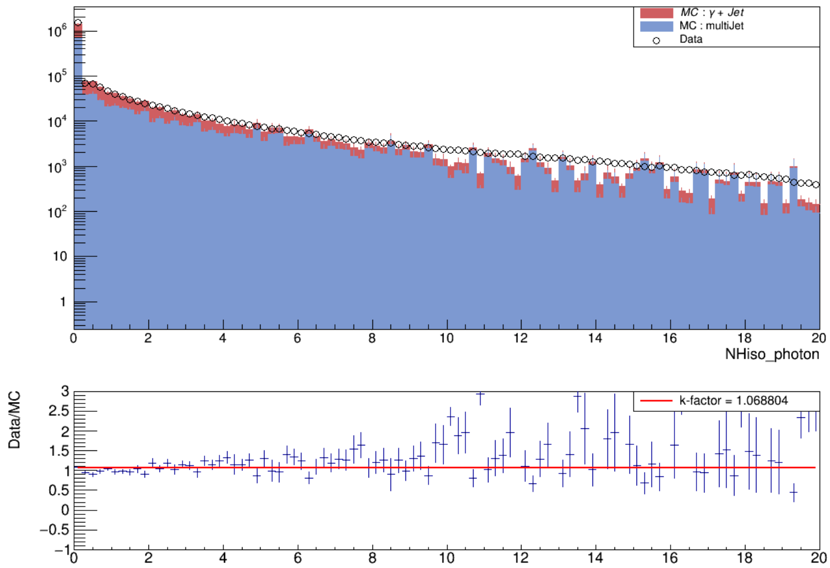



Figure 3.1: Neutral hadron isolation for background and signal MC

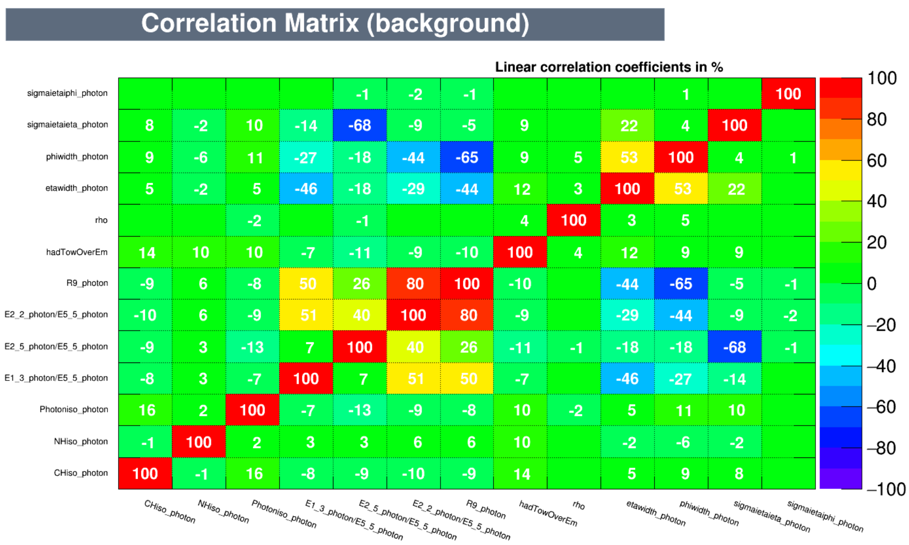



Figure 3.2: Correlation matrix for background MC

## 3.3 Data driven background estimation

MVA will be performed with real data for the background, thereby a sideband (background enriched region in the data sample) has to be defined on a low-correlated variable (fig. 3.3) that won't be used in the MVA. A cut has been applied on *charged hadron isolation* in order to get the best ratio of background purity over number of events.

**Sideband definition** $2.325 < \textit{Charge hadron isolation} < 15.$

Background purity $= 95.00$ %

Number of events $= 7.59 * 10^5$

Then a signal enriched region has been defined on the same variable, this cut will be applied on the signal data sample.

**Signal region definition** $\textit{Charge hadron isolation} < 2.$

Signal purity = ?? %

Number of events = ??

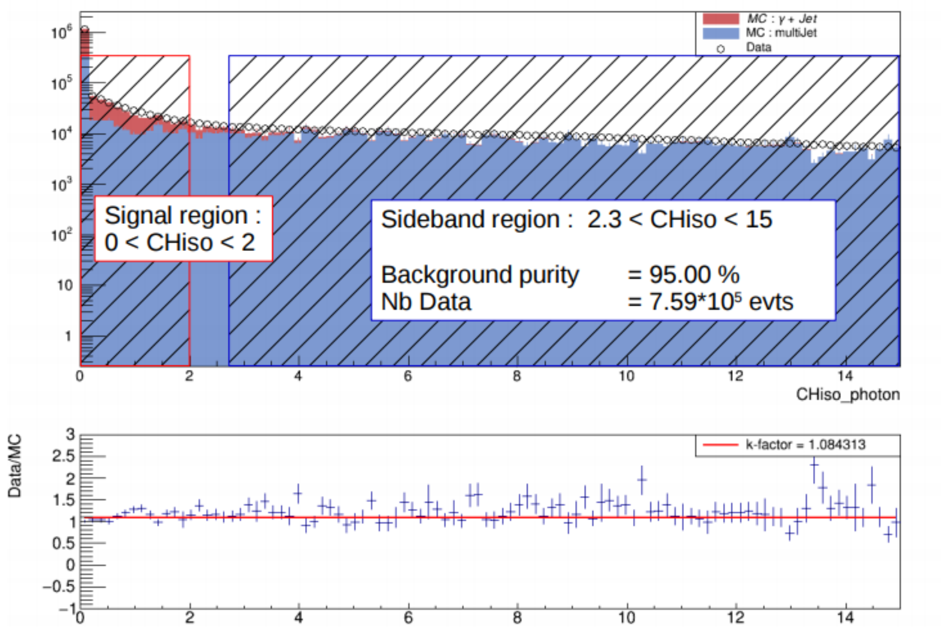



Figure 3.3: Charged hadron isolation for background and signal MC (histograms), signal region (red shaded area) and sideband (blue shaded area).

For cross-checking, we compare variables shape for background MC and DATA in the sideband region. Here you can see for example a comparison between *neutral hadron isolation* for data in the sideband region and background Monte-Carlo.(fig. 3.4)

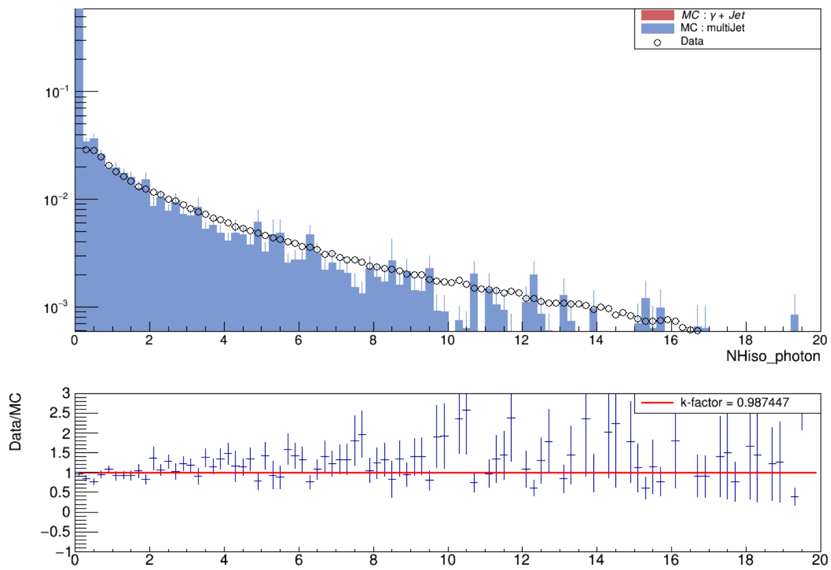
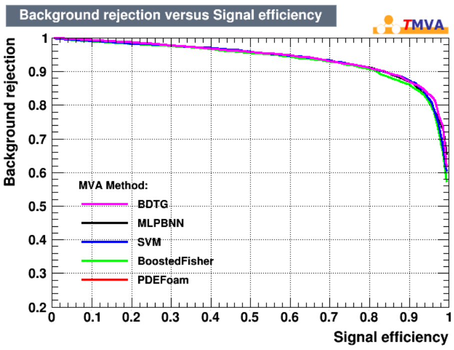


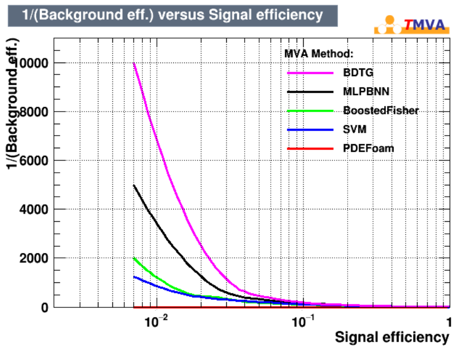

Figure 3.4: Neutral hadron isolation for background MC and DATA in the sideband region

# MultiVariate Analysis

Now that we get background and signal samples we can perform the MVA for classification. For this work the TMVA framework from ROOT was used. Multiple MVA were tested (fig. 4.5a) with default configuration then the 2 bests were selected for the tuning of their parameters.




(a) ROC curve for the 5 bests MVA that has been tested.




(b) Inverse ROC curve for the 5 bests MVA that has been tested.

## 4.1 Artificial Neural Network

An ANN is a multilayer perceptron with fully interconnected layers (fig. 4.7b). This ANN is used for classification, it is a function mapping an input vector $\vec{x_0}$ to a scalar $y$ with $y \in [0;1]$ Here is the output of the ANN that has been trained for the next part of the analysis (fig. 4.6).

### 4.1.1 Theory

A neuron is referenced by his position in the network, a neuron $h_{i,j}(\vec{x_{j-1}}) \to h_{i,j}$ represent the i-th neuron of the j-th layer.
It sums all neuron's output in the (j-1)-th layer, weighted by their connection weight. This net sum is then evaluated through the activation function ( sigmoid, logistic, heaviside, linear, etc) (fig. 4.7a).

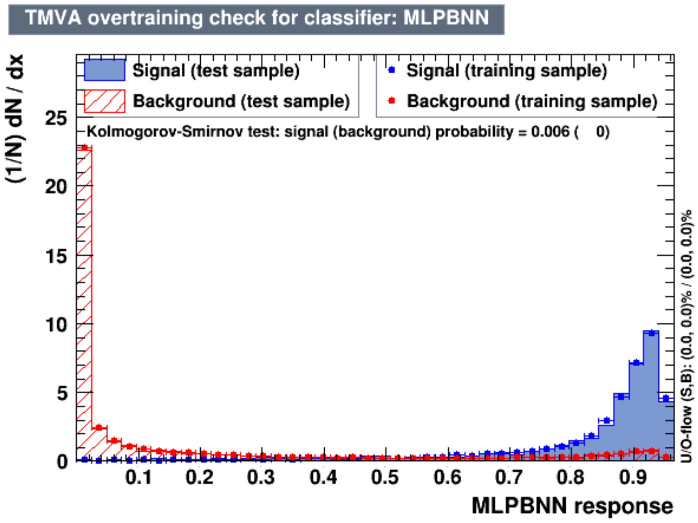
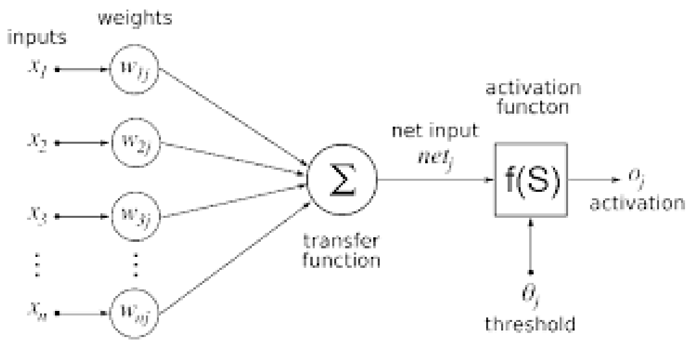
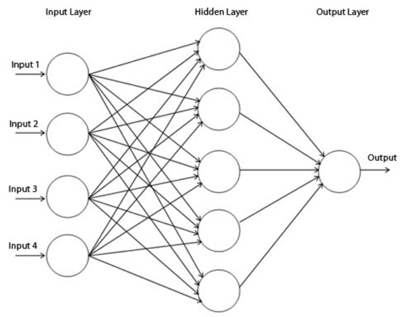



Figure 4.6: Artificial Neural Network response.




(a) Diagram of a single neuron algorithm.




(b) Architecture of an artificial neural network with 4 input variables, one hidden layer, and one output neuron.

A lot of parameters are available for tuning :

**Input variable** Choice of input variable set, number of variables, choice of a Pre-processing method, etc.

**ANN architecture** number of hidden layers, number of neurons per layer, choice of an activation function, etc.

**Learning algorithm parameter** Choice of a learning method, choice of a regulator, value of learning rate, step size, weight decay rate, etc.

All of these cannot be optimize at the same time, so a choice has to be made.
The first parameter to be tune is the input variable set, a compromise has to be made in order to have the smallest input set but containing the most relevant information for classification.

### 4.1.2 Input set optimization

For this part an iterative process of optimization will be performed :

**step 1** Train MVA with full input variable set

**step 2** Train N MVA removing one variable at a time

- **step 2.1** The MVA that succeed the best despite of having removed one variable, tells us that this variable wasn't revelant.
- **step 2.2** Remove this variable permanently, reiterate step 2 until no variable is left.

**final step** keep the input variable set of the best MVA

For evaluating the ANN multiple estimators has been tested :

**Mean Square Estimator (MSE)** $MSE(\hat{\theta}) = E_{\hat{\theta}}[(\hat{\theta} - \theta)^2] = Var_{\hat{\theta}} + Bias(\hat{\theta}, \theta)^2$

**Cross-Entropy (CE)** $H(T, q) = -\sum_{i=1}^{N} \frac{1}{N} log_2 q(x_i)$

**Overlapping criteria** $= \sum_{i=1}^{N} signal * background$

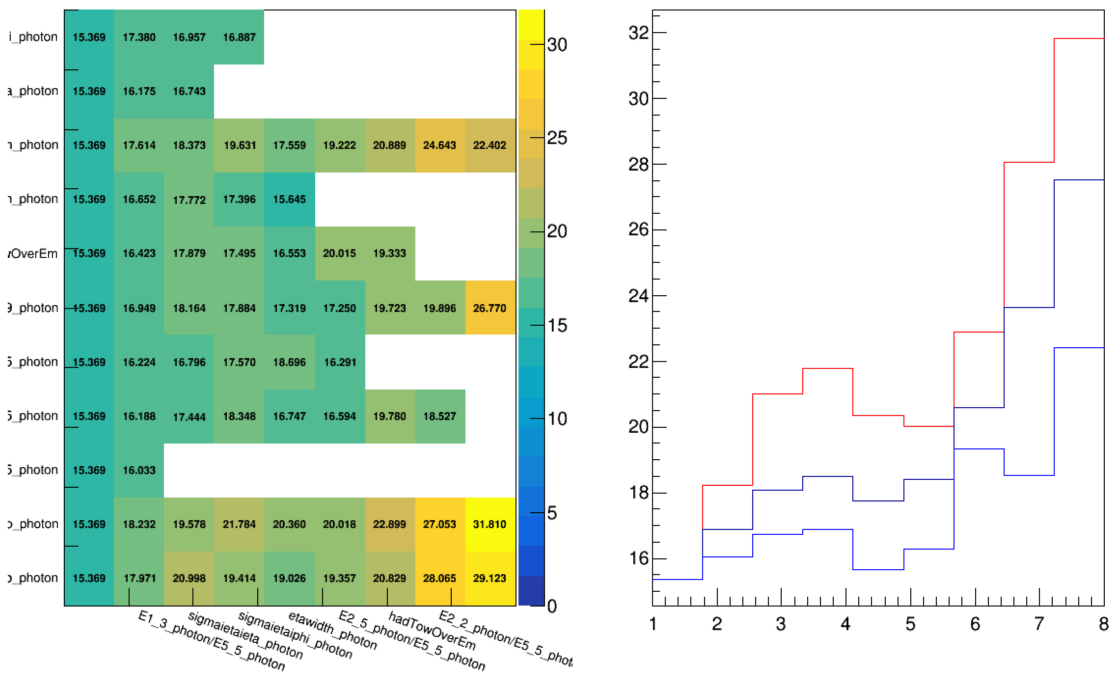



Figure 4.8: Input variable set optimization.

## 4.2 Boosted Decision Tree

Being the best MVA method a BDT has been trained also for the next part of the analysis (fig. 4.9). Multiple learning method has been used, the Gradient Boost Method was the most efficient one.

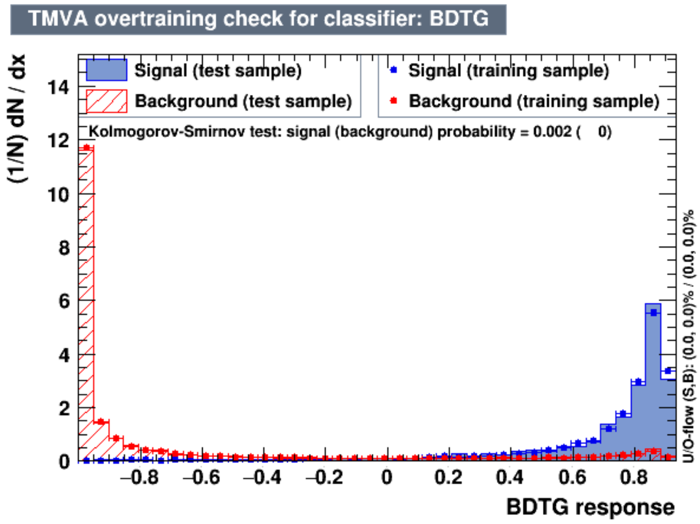



Figure 4.9: Boosted Decision Tree response.

### 4.2.1 Theory

BDT uses a decision tree in order to map from input variables to the event category (signal or background). For this kind of classification tree, branches represent relations between variable or cuts on variables that lead to leaf representing category of the event. The learning method was the "Gradient Boosting", the classification is done by combining together weak classifiers in an iteratively way.

# Signal extracted on DATA

In this section will be extracted $\gamma$+jet event purity on real data. First we must establish PDF for signal (MC simulation) and background (real data in sideband), the analysis will be performed on the $p_T^\gamma$ range [ ? ; ? ] divided in 11? bins

## 5.1 Probability Density Function parametrization

PDF are established using the ROOFit framework of ROOT using MC simulation for signal and real data in the sideband for the background. Then ANN response for data in the signal region is expressed as :

$$ANN(Data_{signal}) = a * PDF(MC_{signal}) + b * PDF(Data_{sideband}) \tag{5.1}$$

With :

- $ANN(Data_{signal})$ := the ANN response for Data in the signal region.
- $PDF(MC_{signal})$ := PDF for MC in the signal region.
- $PDF(Data_{sideband})$ := PDF for Data in the sideband.
- $a$ := number of signal events.
- $b$ := number of background events.


Figure 5.10: Example of PDF parametrization for $p_T^\gamma \in [?;?]$

## 5.2 Fit on Data

### 5.2.1 Pull-plot cross-check

### 5.2.2 $\gamma$+jet events purity

# Conclusion and future outlook

reference [Collaboration 2015].

# Bibliography

[Collaboration 2015] CMS Collaboration. *Performance of Photon Reconstruction and Identification with the CMS Detector in Proton-Proton Collisions at sqrt(s) = 8 TeV.* In JINST 10, 2015.

# Appendix A

# MC vs data comparison

# Appendix B

# Variable signal vs background discrimination

Appendix C

# Learning algorithms

## C.1 Back-Propagation

## C.2 Broyden-Fletcher-Goldfarb-Shanno (BFGS)

## C.3 Bayesian Regulator

---

**Résumé** — Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue. Praesent egestas leo in pede. Praesent blandit odio eu enim. Pellentesque sed dui ut augue blandit sodales. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam nibh. Mauris ac mauris sed pede pellentesque fermentum. Maecenas adipiscing ante non diam sodales hendrerit. Ut velit mauris, egestas sed, gravida nec, ornare ut, mi. Aenean ut orci vel massa suscipit pulvinar. Nulla sollicitudin. Fusce varius, ligula non tempus aliquam, nunc turpis ullamcorper nibh, in tempus sapien eros vitae ligula. Pellentesque rhoncus nunc et augue. Integer id felis.

**Mots clés :** Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor.