

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

Master 2 internship report

Signal and background discrimination in γ+jet events, recorded by the CMS experiment at the LHC.

Author :
Maxime GIRAUD

Supervisor :
Viola SORDINI

Version 0.1 du
July 3, 2018

Contents

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

At the Compact Muon Solenoid are produced at high-energy, proton-proton collision. At these energy scale quarks and gluons interact to form collimated jet of hadrons, called hadronics jets This phenomenon allow us to probe the QCD and the proton structure but is very complexe to analyze.
One way to measure jets energy is to study γ+jet events, on fig (1) the photon is prompt and balance the jet energy.

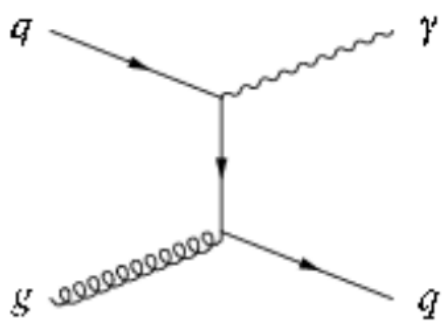


Figure 1: Feynman diagram of a quark-gluon interaction, giving in output a high-energy quark and a prompt photon.

This report describes the analysis of γ+jet events, in the first section will be described the CMS detector and the hadronic jets. Then will be introduced the data that has been used for the analysis part. For the analysis part will be implemented a multivariate analysis for photon identification this will be used for measuring the γ+jet purity in the data.

γ+jet event classification in LHC collisions

1.1 CMS experiment at LHC

The Compact Muon Solenoid (CMS) fig (1.2) is a particle physics detector built at one of the collision points of the Large Hadron Collider (LHC) at CERN in switzerland and France. The goal of the CMS experiment is to investigate the physics of the Standard Model and beyond. 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 range of the LHC particle accelerator.
CMS is made of multiple particle detectors designed to measure the energy and momentum of products of the collisions. The innermost layer called the "Tracker" reconstruct the paths of charged particles comming from the collision or from the decay of of 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.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 for detecting muons.

1.2 Hadronic jets in proton-proton collisions

Jets are the experimental signatures of high-energy quarks and gluons produced in high-energy processes.
These particles having a colour charge, they cannot exist freely due to coulour-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 transverse momentum and resolution is of crucial importance for many physics analyses.

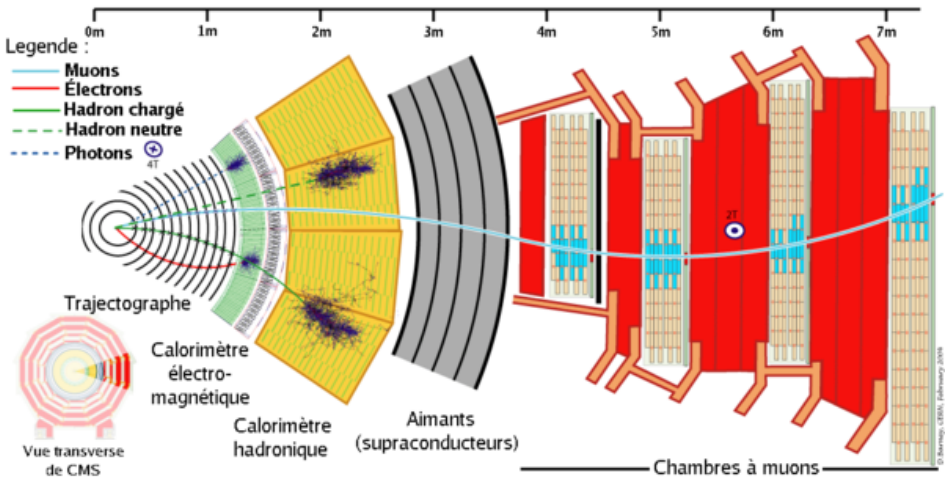


Figure 1.2: Diagram of a slice of the CMS detector, from left to right : tracker, electromagnetic calorimeter, hadronic calorimeter, superconductive solenoid and muon chamber. Lines represent track of particles.

Collision data

In this chapter will be described the data and simulations used for this study. Section 2.3 describes the input variables that have used for photon identification purpose.

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

We are interested in prompt photons comming from the hard scattering, for identifying thes photons we need to perform a multivariate analysis. For this work we use multiple variables representing various aspects of reconstructed photons :

Isolation variables represent additional objects (photons,charged hadron and neutral hadron) reconstructed in a ΔR 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) : $I_{\gamma} = \sum_{\gamma_i}^{\Delta R} p_{T,\gamma_i}$
γ_i corresponds to reconstructed photons, the sum doesn't account for the p_T of the processed photon. (parler du pile-up avec ρ ?)

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 ω_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 φdirection.

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

η_{width} γ: Shower width in η

φ_{width} γ: Shower width in φ

R_9 γ: 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

ρ: 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, they describe various aspect of photons and will be used to distinguish between prompt and fragmentation photon.
To perform classification a multivariate analysis will be implemented, but MVA training can be time consuming and the "curse of dimensionality" [1] forces us to select the shortest possible input set.

Variables with most differences of shape for background and signal will be the most relevant for the MVA classification.

The MVA will be trained with MC simulation for the signal sample and with the real data for the background sample. Indeed we trust MC simulation for the signal sample (γ+jet events) but on the contrary MC backgound (multi-jet) may not be accurate (by not taking into account ...) and gave us low statistics.
For this reason real data will be used for the background sample and so a control region enriched in background event has to be defined (sideband). In order to do that we need to perform a data-driven background estimation using a low-correlated variable for this sideband definition.

3.1 Background vs Signal discrimination

The choice of discriminating variables is done by looking at their shape for background and signal, processed from MC simulation.
Since the background is extracted from a data control region for the final analysis, a cross-check of the variables shape has to be done between Data and MC to validate this control region.
Fig. (3.3) shows an example of MC simulation and data comparison for *neutral hadron isolation* variable.

[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.

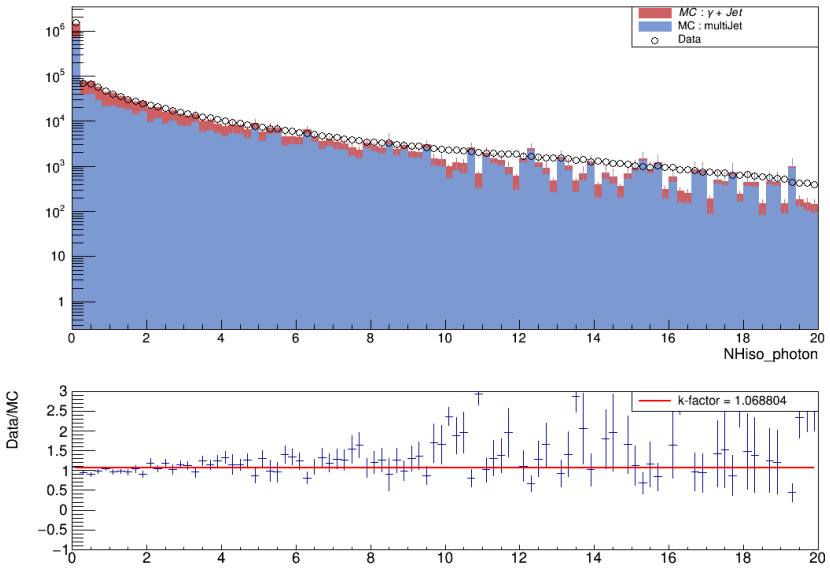


Figure 3.3: Top plot : Neutral hadron isolation variable for background (blue histogram) and signal MC (red histogram) and real data superimposed (empty circles). Normalized to integrated luminosity of $36fb^{-1}$
Bottom plot : Ratio of total expected events from MC (background+signal) over real data (blue cross) fitted by a constant (red line).

3.2 Variable correlations

Because we use distribution of the data in the control region as a proxy for their distribution in the signal region, we need to make sure that the variable for the sideband definition has low correlations with the other one. By looking at the correlation matrix fig. (3.4) we can see that *charged hadron isolation* is one good candidate and so will be used next for the sideband definition.

3.3 Data driven background estimation

An 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.*Charged hadron isolation* fig. (3.5) has been chosen and the sideband defined in order to find a good compromise between background purity and number of events on data.

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

Background purity = 95.00 %

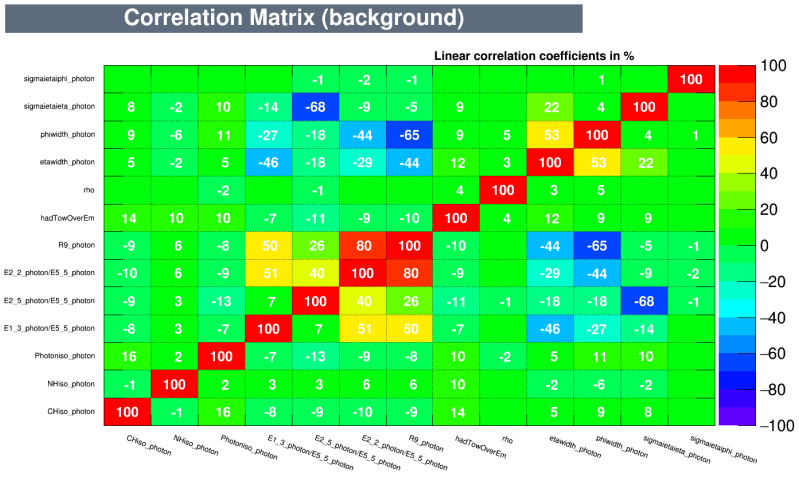


Figure 3.4: Correlation matrix for background MC, each line or column represent a variable.

$$\text{Number of events} = 7.59 * 10^5$$

For cross-check, we compare the variables shape for background MC in the signal region and DATA in the sideband region. Fig. (3.6) shows an example of a comparison between *neutral hadron isolation* for data in the sideband region and background Monte-Carlo. We can see a good agreement for MC and real data, except for a small trend in the high energy range probably due to the low statistics.

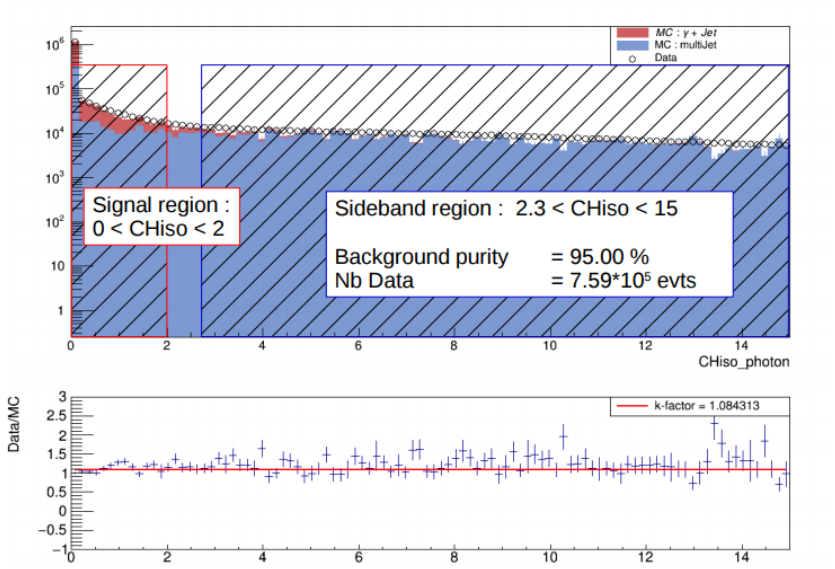


Figure 3.5: Charged hadron isolation for background MC (blue histogram), signal MC (red histogram) and real data superimposed (empty circles). Normalized to integrated luminosity of $36fb^{-1}$
On top of this is the sideband definition (red shaded area) and the signal region definition (blue shaded area)
Bottom plot : Ratio of total expected events from MC (background+signal) over real data (blue cross) fitted by a constant (red line).

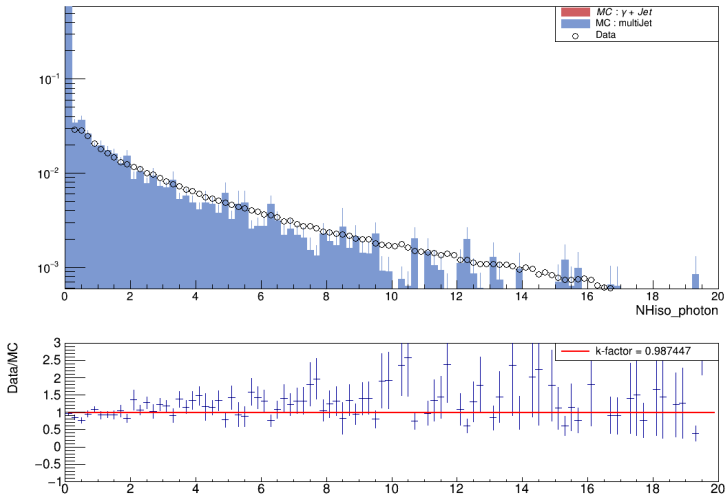
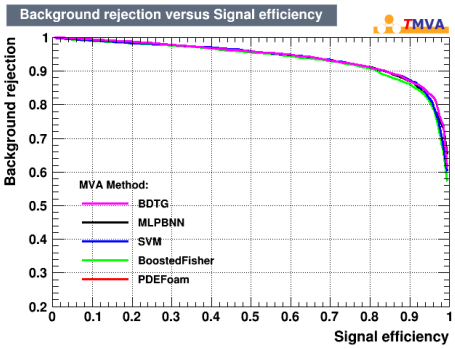


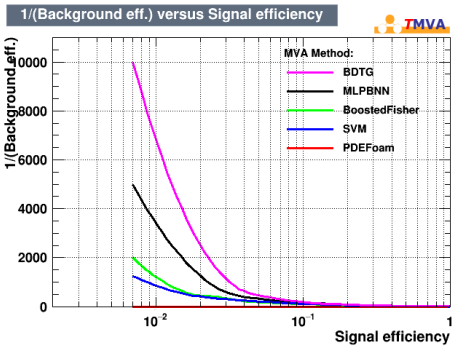
Figure 3.6: Neutral hadron isolation for background MC (blue histogram) and real data in the sideband superimposed (empty circles). The integral of both distribution are normalized to unity.
Bottom plot : Ratio of background MC over real data (blue cross) fitted by a constant (red line).

MultiVariate Analysis

Now that we get background and signal samples we can perform the MVA for classification. For this purpose the TMVA framework from ROOT was used. Multiple MVA techniques were tested fig. (4.7a) with default configuration then the 2 bests were selected for the tuning of their parameters : the ANN and BDT



(a) ROC curves of 5 bests MVA, these are almost overlapping.



(b) Inverse ROC curve for the 5 bests MVA, on this plot BDT and ANN(MLPBNN) are clearly the two bests.

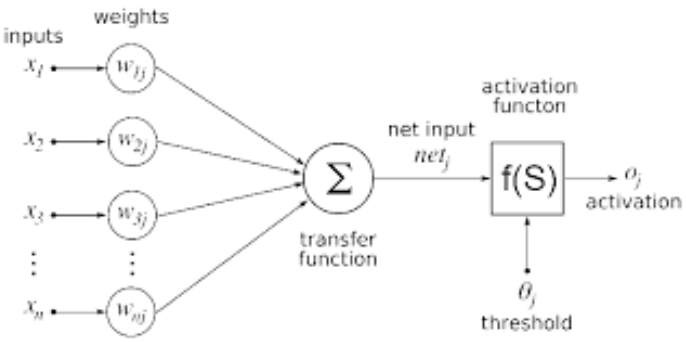
Figure 4.7: ROC curve for the 5 best MVA that has been tested. Receiver Operating Characteristic (ROC) curve reflects the discrimination power of a classifier. It is constructed by plotting the ratio of background rejection versus signal efficiency by varying a threshold on the MVA output.

4.1 Artificial Neural Network

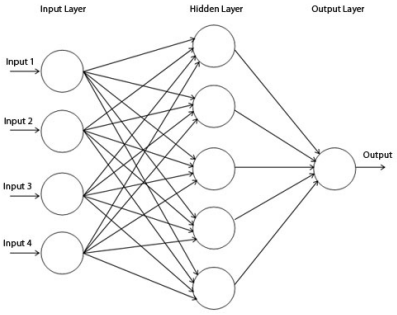
An ANN is a multilayer perceptron with fully interconnected layers fig. (4.8b). This ANN is used for classification, it is a function mapping an input vector $\vec{x_0}$ (input variables) to a scalar y with $y \in [0; 1]$ (classification category).
Fig. (4.10) shows the output y of the ANN that has been trained. Data have been divided in two, one training sample and one testing sample.

A neuron is referenced by his position in the network, a neuron $h_{i,j}(\vec{x_{j-1}}) \rightarrow h_{i,j}$ represent

the i-th neuron of the j-th layer.
Such neuron 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. (4.8a).



(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 variables 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.

4.1.1 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 is the sum of the products of signal and background response in each bin $OC = \sum_{i=1}^{N} signal_i * background_i$ with $N :=$ number of bins , $signal_i :=$ number of signal events in bin number i $background_i :=$ number of background events in bin number i. Good classifiers show low value for this estimator.

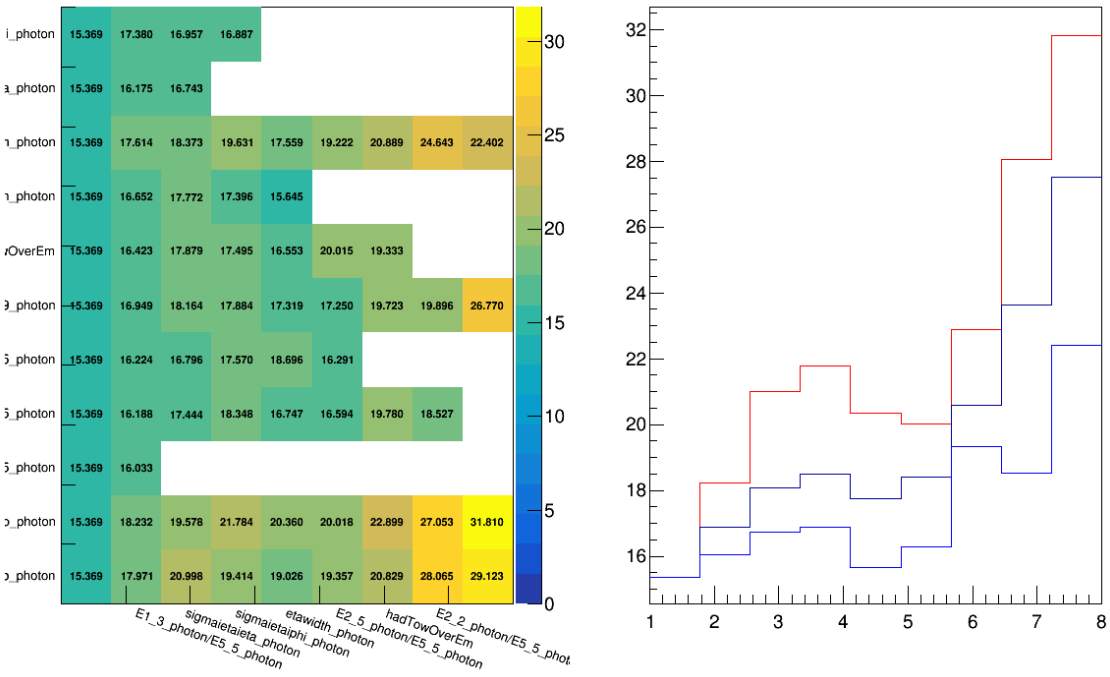


Figure 4.9: On the left : Input variable set optimization results overview for the "overlapping criteria" estimator. Each bin is the estimator value for one MVA. Except for the 1st column that represents the MVA trained with the whole input seti (step 1), 2nd column represent MVA's that has been trained after removing one variable at a time (step 2), following columns are the iterations of step 2.
On the right : overview of the estimator value for each column (step 2). maximum value (red solid line), average value (dark blue solid line) and lowest value (light blue solid line).

The optimization results in fig. (4.9) shows that keeping all the variables lead to the best MVA. So the whole input set will be used for the training. The ANN fig. (4.10) used 11 input variables : neutral hadron isolation, photon isolation, $\sigma_{i\eta i\eta}$, $\sigma_{i\eta i\phi}$, η_{width}, ϕ_{width}, R_9, Had/Em, E_{1x3}/E_{5x5}, E_{2x2}/E_{5x5} and E_{2x5}/E_{5x5}.

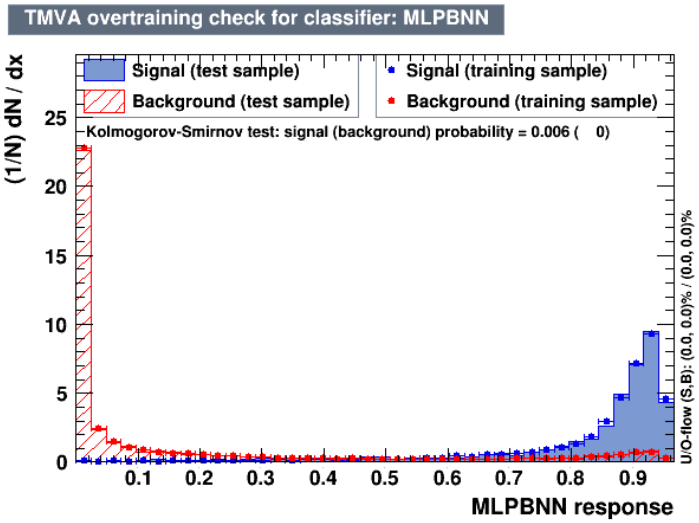


Figure 4.10: Artificial Neural Network response with signal from test sample (blue histogram), background from test sample (red shaded histogram), signal from training sample (blue dots) and background from training sample (red dots). The good agreement between training and testing sample shows no overfitting (in the case where these sample are representative of the data).

4.2 Boosted Decision Tree

Being the best MVA method a BDT has been trained also for the next part of the analysis fig. (4.11). Multiple learning method has been used, the Gradient Boost Method was the most efficient one.

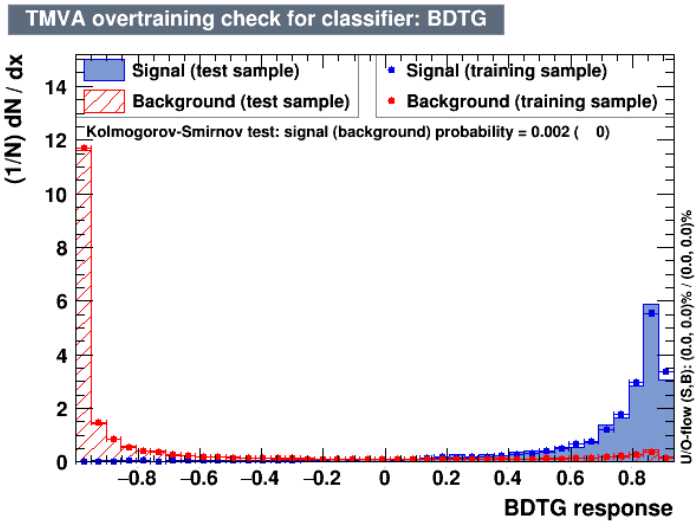


Figure 4.11: Boosted Decision Tree response.

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.
For the "Gradient Boosting" method, the classification is done by combining together weak classifiers in an iteratively way. A finir!!!

Signal extraction on DATA

In this section the results of the MVA training obtained in section 4 will be used to extract γ+jet event purity on real data via a maximum-likelihood fit. First we must establish PDF for signal (MC simulation) and background (real data in sideband), the analysis will be performed on the p_T^γ range [40 GeV ; 3000 GeV] divided in 12 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 MVA response for data in the signal region is expressed as :

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

With :

- $MVA(Data_{signal})$:= the MVA 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.

5.2 Fit on Data

With the PDF established in the previous section we want to extract values of the parameters a and b representing signal and background proportion in the sample. For this analysis we will perform a maximum likelihood estimation for each p_T^γ range that has been defined. (fig 5.12) show for example the fit perfomed for $p_T^\gamma \in [75GeV; 230GeV]$

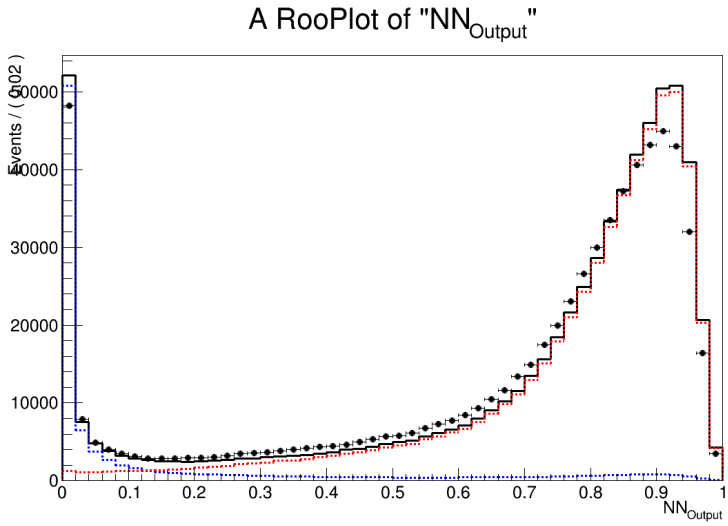


Figure 5.12: Example of a maximum likelihood fit performed for $p_T^\gamma \in [75GeV; 230GeV]$, showing background PDF (blue dotted line) signal PDF (red dotted line) and fit result (solid black line) superimposed with real data (black dot).

5.2.1 Pulls distribution cross-check

In order to perform a cross-ckeck of the maximum-likelihood we generate a pull distribution for each p_T^γ range.

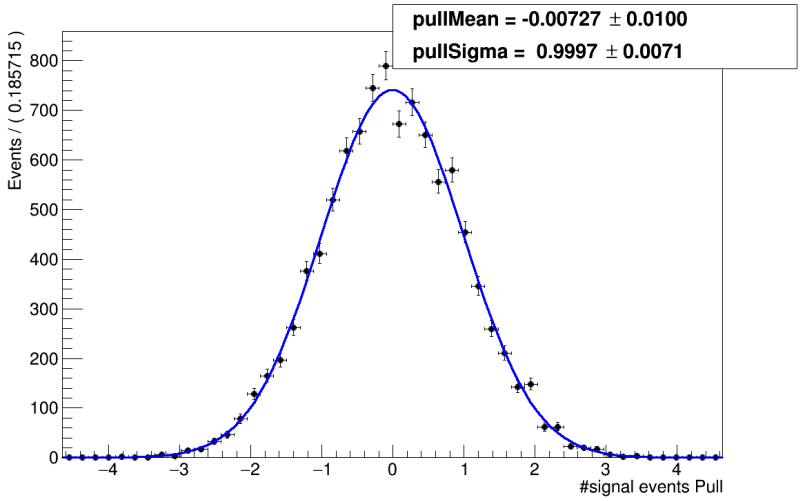


Figure 5.13: Example of a pull plot performed for $p_T^\gamma \in [75GeV; 230GeV]$, showing background PDF (dotted blue line) signal PDF (dotted red line) and fit result (solid black line) superimposed with real data (black dot).

5.2.2 γ+jet events purity

Finally the estimated parameters representing background and signal proportions are used to construct the γ+jet events (signal) purity function of the p_T^γ (fig. 5.14).

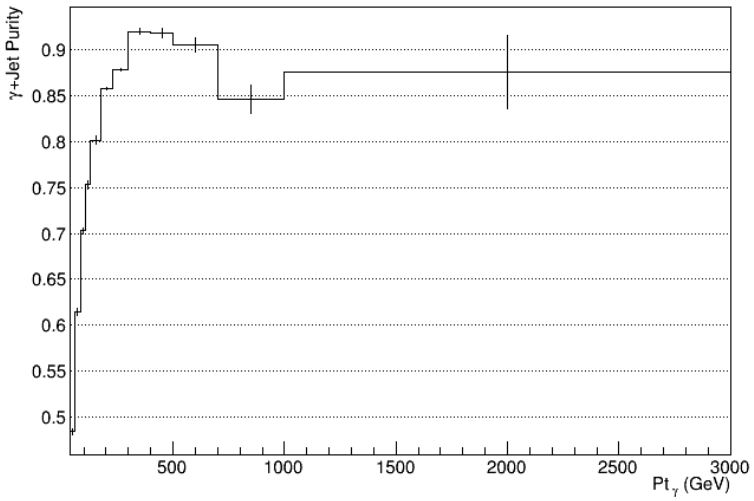


Figure 5.14: γ+jet purity function of $p_T^\gamma \in [40GeV; 3000GeV]$ evaluated with the ANN, showing background PDF (dotted blue line) signal PDF (dotted red line) and fit result (solid black line) superimposed with real data (black dot).

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.