

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 4, 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.

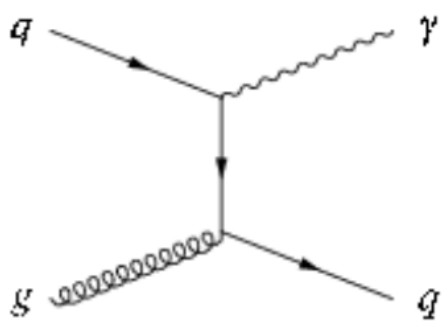


Figure 1: Feynman diagram of a quark-gluon interaction, giving in output a high-energy quark and a prompt photon.

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 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 generating in its inner part a magnetic field 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 magetic field flux return yoke, outside the solenoid, is instrumented with muon detectors.

1.2 Pile-up events

At the LHC protons are bunched together, into up to 2808 bunches composed of around 10^{11} protons each. When two bunches interact there is a non-negligeable probability that multiple events arise from this collision, these are called pile-up events and have to be taken into account for particle reconstruction.

1.3 CMS coordinate system

The absolute CMS coordinate system is defined with respect to the LHC ring. The X axis points towards the center of the ring, the Y axis points up and the Z axis is defined as per a right handed coordinate system.

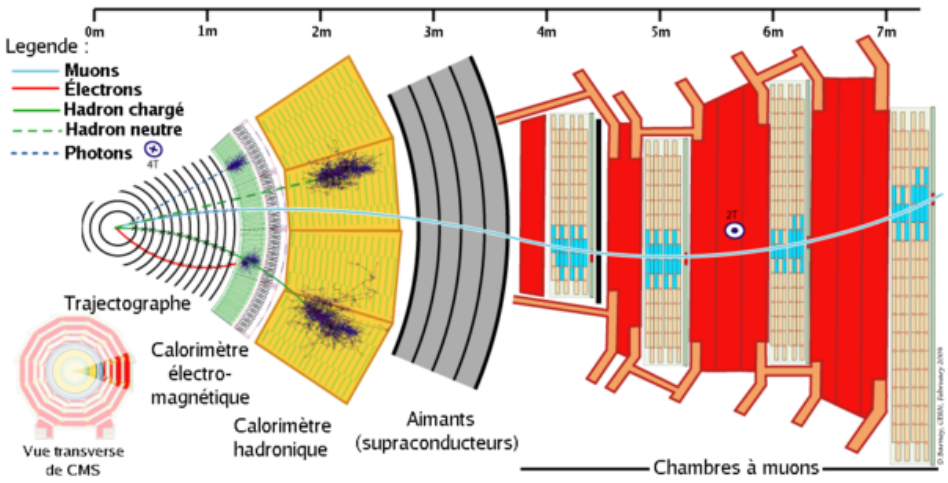


Figure 1.2: Diagram of a slice of the CMS detector, from left to right : tracker, electromagnetic calorimeter, hadronic calorimeter, superconductive solenoid and muon chamber. Lines represent track of particles.

The absolute azimuth φruns from the X axis (at 0 degree) towards the Y axis and $\eta$is the polar angle.
Initial energy of the particles is along the beam axis. Due to the energy conservation, the total energy in the plane transverse to the beam axis has to be zero. For this reason, in the following we will always consider the transverse momentum $\vec{p}_T$ (with $p_T = |\vec{p}_T|$).

1.4 Hadronic jets in proton-proton collisions

Jets are the experimental signatures of high-energy quarks and gluons produced in high-energy processes.
During proton-proton collisions high-energy quarks and gluons may be produced they will radiate high-energy gluons or decay into high-energy quark-antiquark pairs (for gluons). This process is called a parton shower, a lot of quarks and gluons are produced but these particles having a colour charge.
Thereby they cannot exist freely due to coulour-confinement and come together to form colour-neutral hadrons by a process called hadronisation that leads to a collimated spray of hadrons called a hadronic jet. The detailed understanding of both the jet transverse momentum scale and resolution is of crucial importance for many physics analyses.

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 been used for photon identification purpose.

2.1 Monte-Carlo simulation

We used simulated events for photon+jet and multijet production produced with the PYTHIA generator, interfaced to GEANT4 for the simulation of the CMS detector. A special care is taken to accurately simulate the pileup contribution to well describe the data. The samples used in this study are official CMS samples.

2.2 CMS data

This study uses the full dataset collected by CMS in 2016 for proton-proton collisions. These data were data recorded with good LHC conditions and with a fully-functioning detector only, at $\sqrt{s} = 13TeV$ for an integrated luminosity of $36fb^{-1}$.

2.3 MVA variables

We want to distinguish the so-called prompt photons, stemming for real photon+jet events, as the one from Fig. (1, from photons produced inside jets during the hadronisation process and following decays. To this end, we dispose of several variables, representing various aspects of reconstructed photons :

Isolation variables represent the total transverse momentum carried by additional objects (photons,charged hadron and neutral hadron) reconstructed in a fixed cone, around the processed photon, of radius ΔR. 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 ηand ω_i is a weight representing the expected energy deposit measured.

$i\eta_{seed}$ is the crystal with the largest energy deposit in the supercluster.

$\omega_i = b + ln(\frac{E_i}{E_{5x5}})$

$\sigma_{i\varphi i\varphi}$: 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: Electromagnetic shower width in η

φ_{width} γ: Electromagnetic 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 density, 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 considering all the contributing processes) and would give us low statistics leading to fluctuations. For this reason, a control region enriched in background multijet events (called sidebands in the following) has to be defined. Events in the sidebands will be used to extract the expected distribution of discriminant variables for background, and use them in the training of the multivariate analysis. The results will be then used in a signal region, in order to extract the signal directly from data.

3.1 Background vs Signal discrimination

The choice of discriminating variables is done by looking at their shape for background and signal, both taken from MC simulation.
Fig. (3.3) shows an example of MC background simulation and sidebands data comparison for *neutral hadron isolation* variable.

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 (in the sidebands) and MC to validate this control region.

[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.

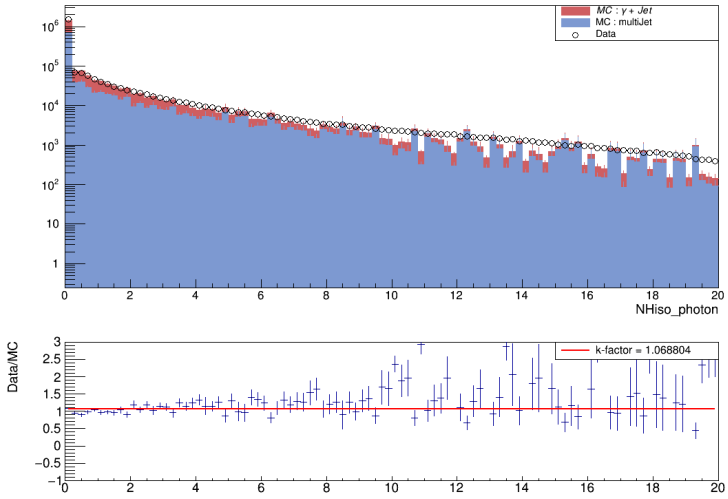


Figure 3.3: Top plot : Neutral hadron isolation variable for background (blue histogram) and signal MC (red histogram) and real data superimposed (empty circles). Normalized to integrated luminosity of $36fb^{-1}$
Bottom plot : Ratio of total expected events from MC (background+signal) over real data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

3.2 Variable correlations

It is interesting to look at the correlations between the variables considered in the multivariate analysis, which can also be a source of discrimination. Fig. (3.4) shows an example of the correlation values for the background MC.

3.3 Background-enriched control region definition

Because we use the 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 ones used in the multivariate analysis. By looking at the correlation matrix fig. (3.4) we can see that *charged hadron isolation* ($:= I_{CH}$) is one good candidate and so will be used next for the sideband definition.
The sideband defined in order to find a good compromise between background purity and number of events on data.

Sideband definition $2.3 < I_{CH} < 15$.

Background purity = 95.00 %

Number of events = $7.59 * 10^5$

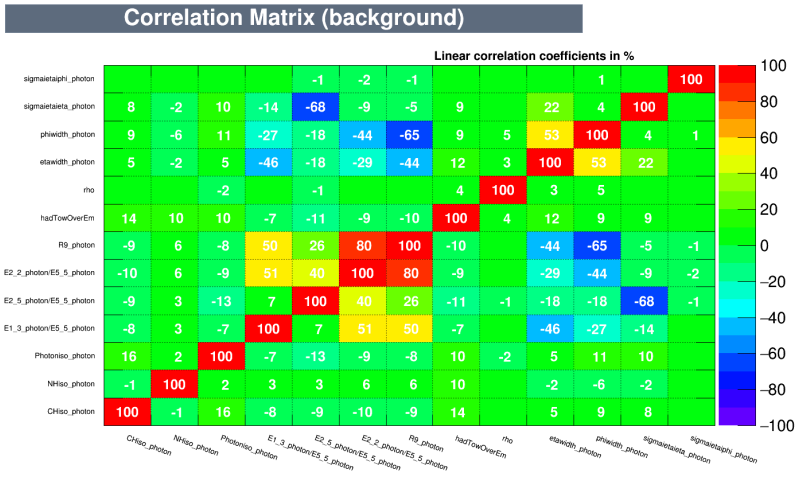


Figure 3.4: Correlation matrix for background MC, each line or column represent a variable.

Fig. (3.5) sketches the definition of the sidebands and the signal region.

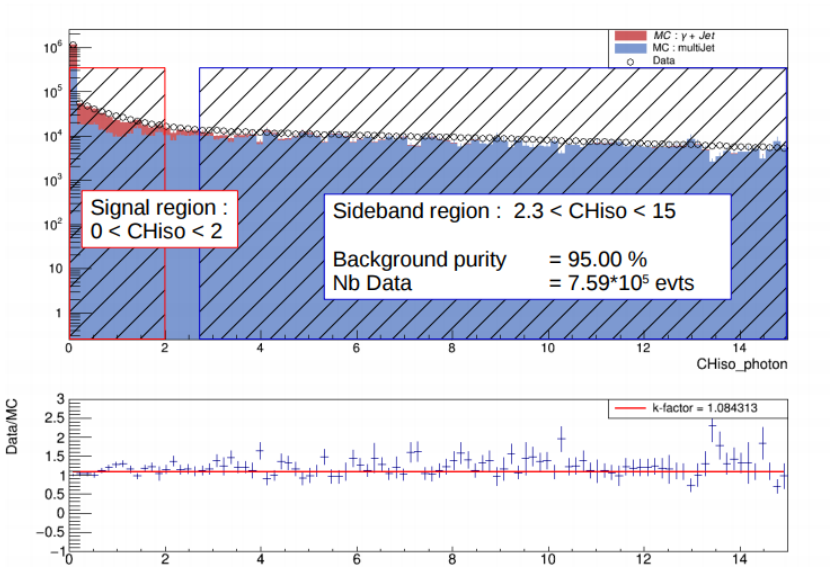


Figure 3.5: Charged hadron isolation for background MC (blue histogram), signal MC (red histogram) and real data superimposed (empty circles). Normalized to integrated luminosity of $36 fb^{-1}$
On top of this is the sideband definition (red shaded area) and the signal region definition (blue shaded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

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. It can also be noticed that using the sidebands sensibly increases the available background statistics, with respect to the simulation.

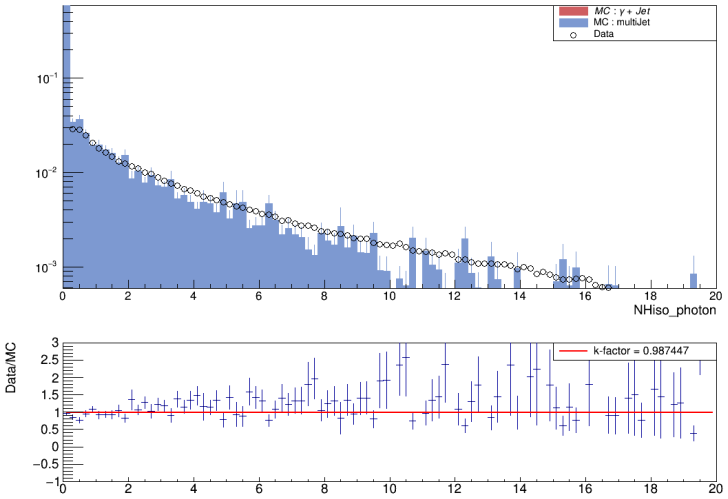
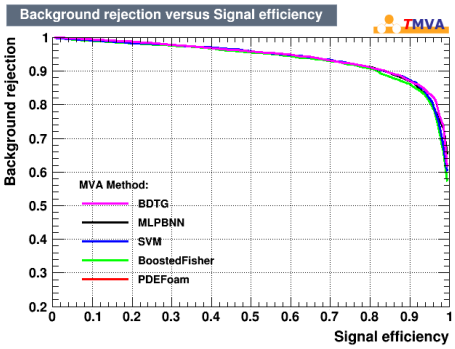


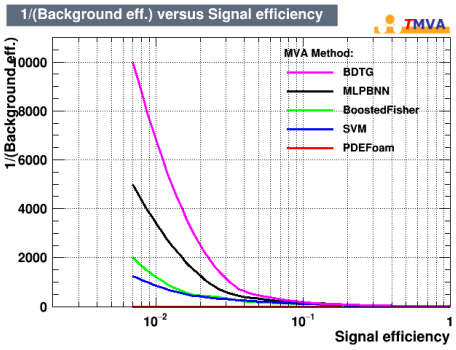
Figure 3.6: Neutral hadron isolation for background MC (blue histogram) and real data in the sideband superimposed (empty circles). The integral of both distribution are normalized to unity.
Bottom plot : Ratio of real data over background MC (blue cross) fitted by a constant (red line).

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.

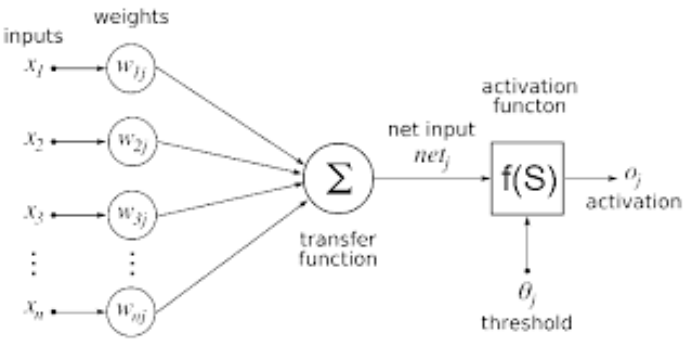
Figure 4.7: ROC curve for the 5 best MVA that has been tested. Receiver Operating Characteristic (ROC) curve reflects the discrimination power of a classifier. It is constructed by plotting the ratio of background rejection versus signal efficiency by varying a threshold on the MVA output.

4.1 Artificial Neural Network

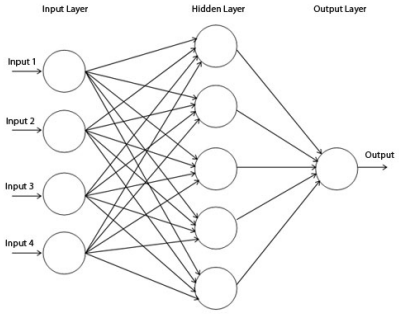
An ANN is a multilayer perceptron with fully interconnected layers fig. (4.8b). This ANN is used for classification, it is a function mapping an input vector $\vec{x_0}$ (input variables) to a scalar y with $y \in [0;1]$ (classification category).
Fig. (4.10) shows the output y of the ANN that has been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

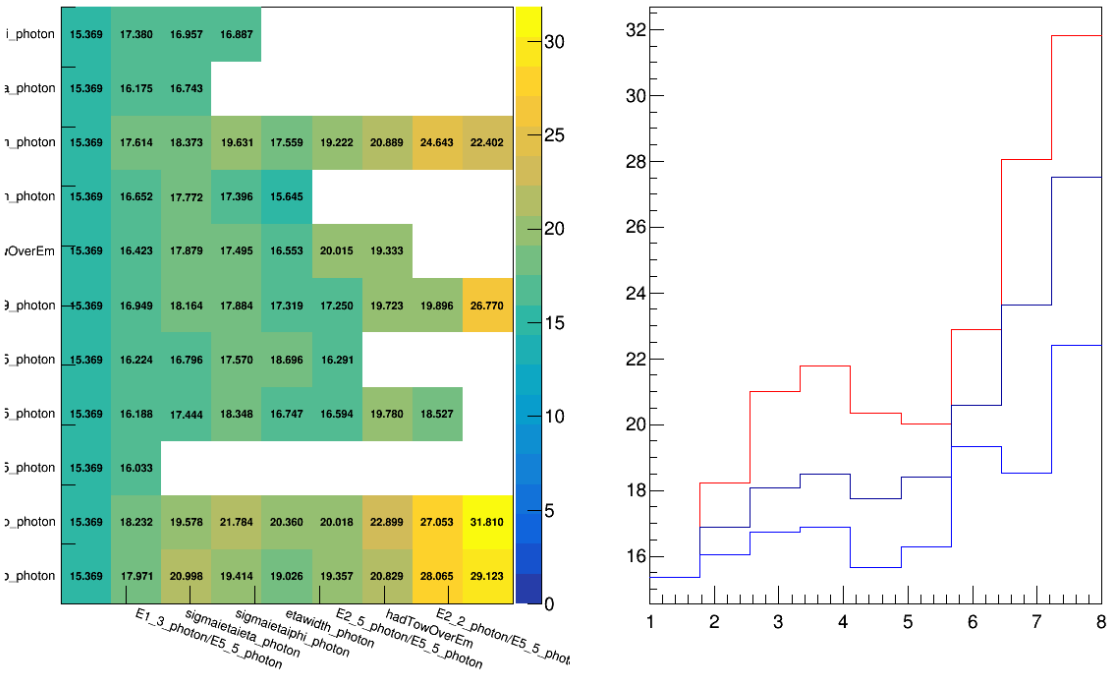


Figure 4.9: On the left : Input variable set optimization results overview for the "overlapping criteria" estimator. Each bin is the estimator value for one MVA. Except for the 1st column that represents the MVA trained with the whole input seti (step 1), 2nd column represent MVA's that has been trained 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}.

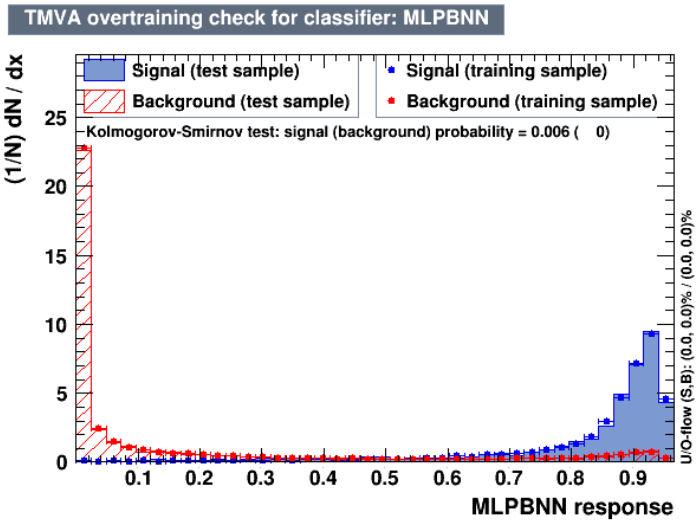


Figure 4.10: Artificial Neural Network response with signal from test sample (blue histogram), background from test sample (red shaded histogram), signal from training sample (blue dots) and background from training sample (red dots). The good agreement between training and testing sample shows no overfitting (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. (4.12). 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. Fig. (4.11) shows an example diagram of a small BDT classifying into 5 classes.

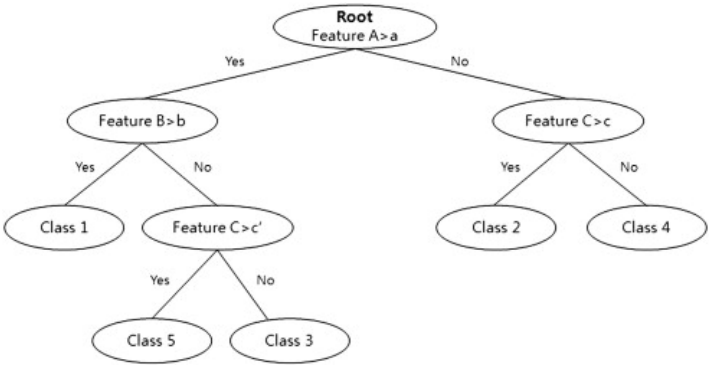


Figure 4.11: Boosted Decision Tree classifying in 5 classes diagram.

Multiple learning method has been used, the Gradient Boost Method was the most efficient one. With this method, the classification is done by combining together weak classifiers in

an iteratively way : multiple "weak classifier" are trained and their output are combined in a weighted sum giving the "big classifier" output, then at each iteration the "weak classifiers" weights are adjusted to minimize the error on classification.

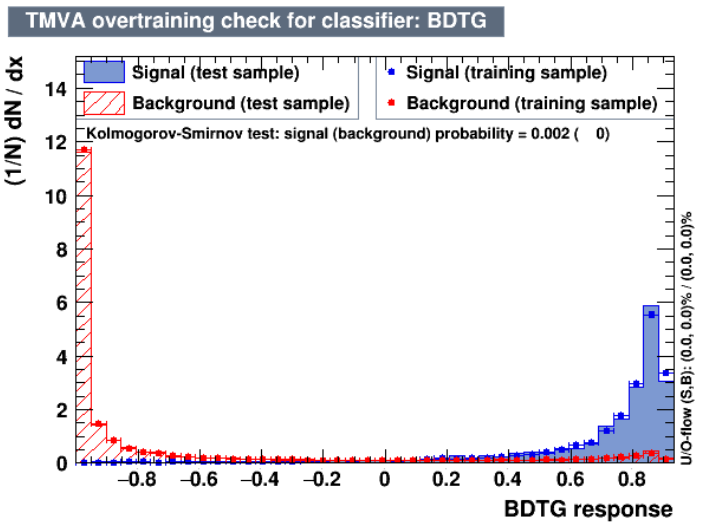


Figure 4.12: Boosted Decision Tree response with signal from test sample (blue histogram), background from test sample (red shaded histogram), signal from training sample (blue dots) and background from training sample (red dots).

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. The probability density function are determined on MC for the signal and sideband data for the background. The analysis is performed on the p_T^γ range [40 GeV ; 3000 GeV] divided in 12 bins. In each bin a fit is performed to the data distribution of the MVA to extract the number of signal and background event.

5.1 Probability Density Function parametrization

Maximum-likelihood fit is implemented using the ROOFit framework of ROOT. Then the MVA response for data in the signal region is expressed as :

$$F(MVA) = N_{signal} * f^{signal}(MVA) + N_{background} * f^{background}(MVA) \tag{5.1}$$

With :

- $F(MVA)$:= the MVA response for Data in the signal region.
- $f^{signal}(MVA)$:= PDF for MC in the signal region.
- $f^{background}(MVA)$:= PDF for Data in the sideband.
- N_{signal} := number of signal events.
- $N_{background}$:= 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 N_{signal} and $N_{background}$ representing signal and background number of events in the sample. For this analysis we will perform a maximum-likelihood estimation for each p_T^γ range that has been defined. Fig. (5.13) show for example the fit perfomed for $p_T^\gamma \in [175GeV; 230GeV]$

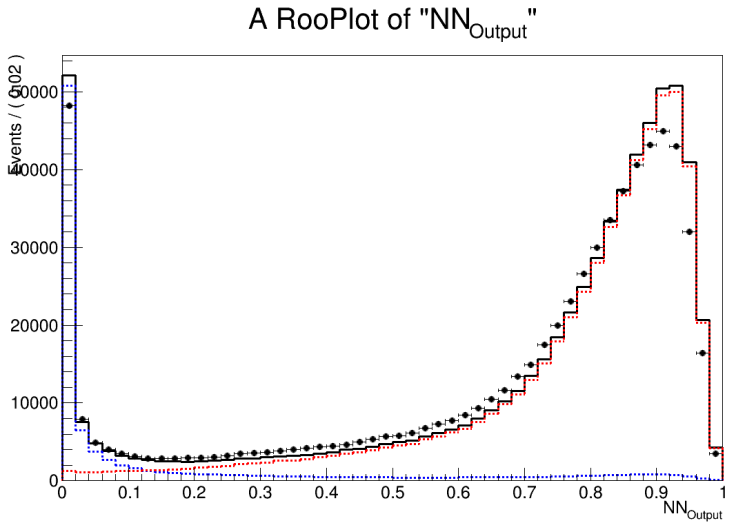


Figure 5.13: Example of a maximum likelihood fit performed for $p_T^\gamma \in [175GeV; 230GeV]$ on the ANN, showing background PDF (blue dotted line) signal PDF (red dotted line) and fit result (solid black line) superimposed with real data (black dot).

5.2.1 Pulls distribution cross-check

In order to validate the maximum-likelihood fit we generate a pull distribution for 10^4 toyMC experiments for each p_T^γ range. For each toyMC, a fake data sample is generated according to f^{signal}, $f^{background}$, N_{signal} and $N_{background}$.
The sample is then fitted with the same fit model to obtain $N^{fit}_{signal} \pm \sigma_{N_{signal}}$ and $N^{fit}_{background} \pm \sigma_{N_{background}}$ To ensure the fit has no intrinsic bias, we look at the pull distribution defined as :

$$Pull = \frac{N^{fit}_{signal} - N^{generated}_{signal}}{\sigma^{fit}_{N_{signal}}} \tag{5.2}$$

If the likelihood of the fitted variables is well described by a gaussian, we expect the mean of the pull distribution to be centered with zero and its RMS to be centered with one.

5.2.2 γ+jet events purity

Finally the estimated number of signal and background events are used to construct the γ+jet events (signal) purity function of the p_T^γ fig. (5.15), defined as :

$$Purity = \frac{N_{signal}}{N_{signal} + N_{background}} \tag{5.3}$$

We can see that the purity is at about 50% at low p_T^γ and goes up to 90% around p_T^γ 500GeV. There is a decreasing of the purity at 800 GeV, this could be due p_T^γ dependance that hasn't been taken in account in the MVA's.

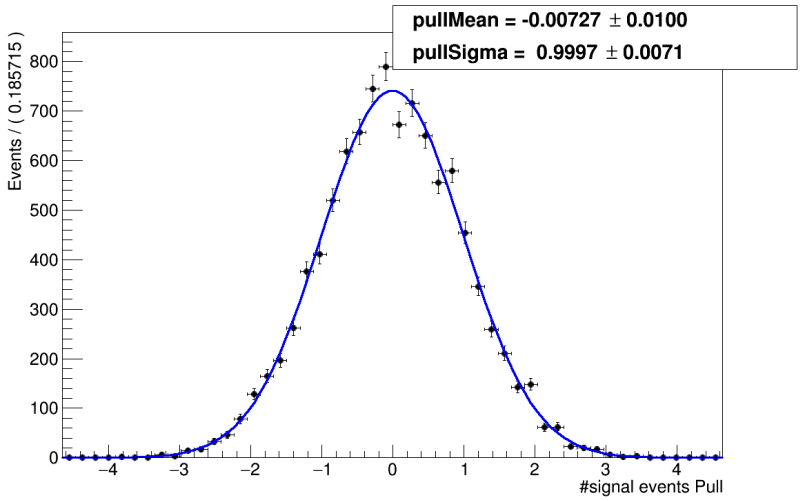


Figure 5.14: Example of a pull plot performed for $p_T^\gamma \in [175GeV; 230GeV]$ with the ANN, showing pull results (black dot) and a gaussian fit (blue line)

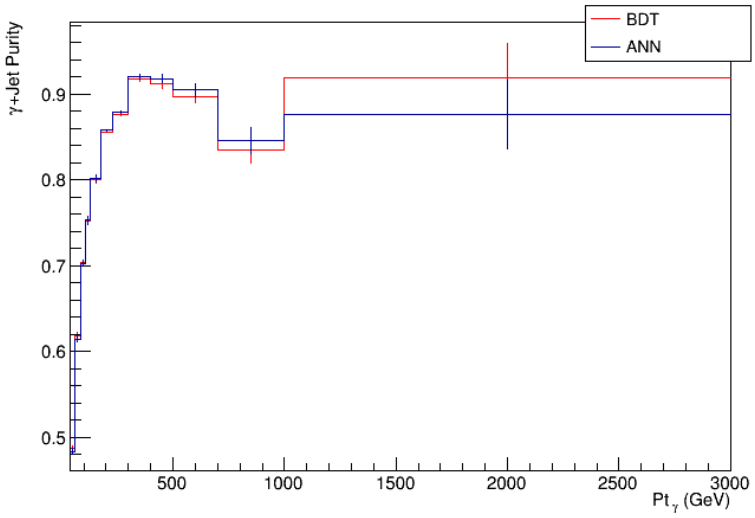


Figure 5.15: γ+jet purity function of $p_T^\gamma \in [40GeV; 3000GeV]$ evaluated with the BDT and the ANN.

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

In order to reduce the appendix size, only 4 variables for the MC (background+signal) vs real data are shown.

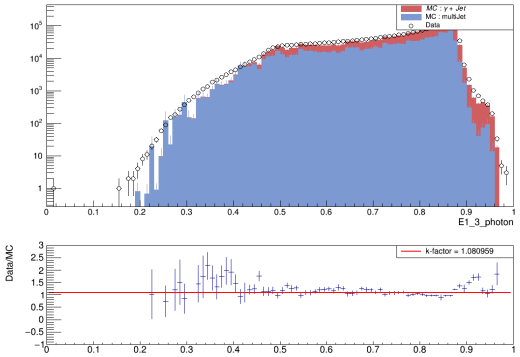


Figure A.1: E_{1x3}/E_{5x5}

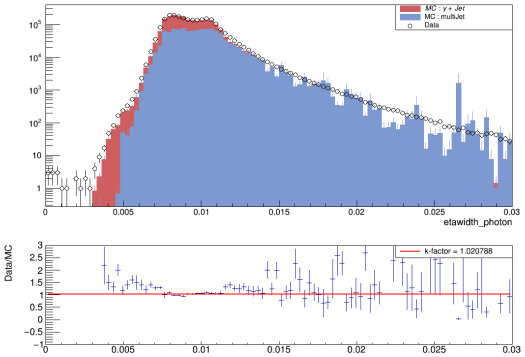


Figure A.2: η_{width}

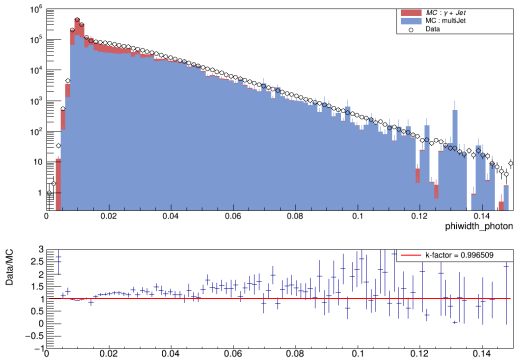


Figure A.3: ϕ_{width}

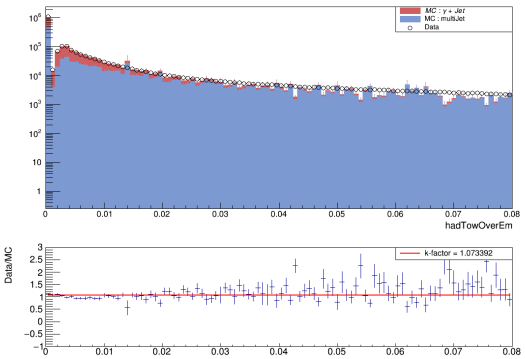


Figure A.4: Had/Em

Top plot : 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).

Appendix B

Maximum-likelihood fit

Maximum-likelihood fit performed with the ANN for all p_T^γ bins.

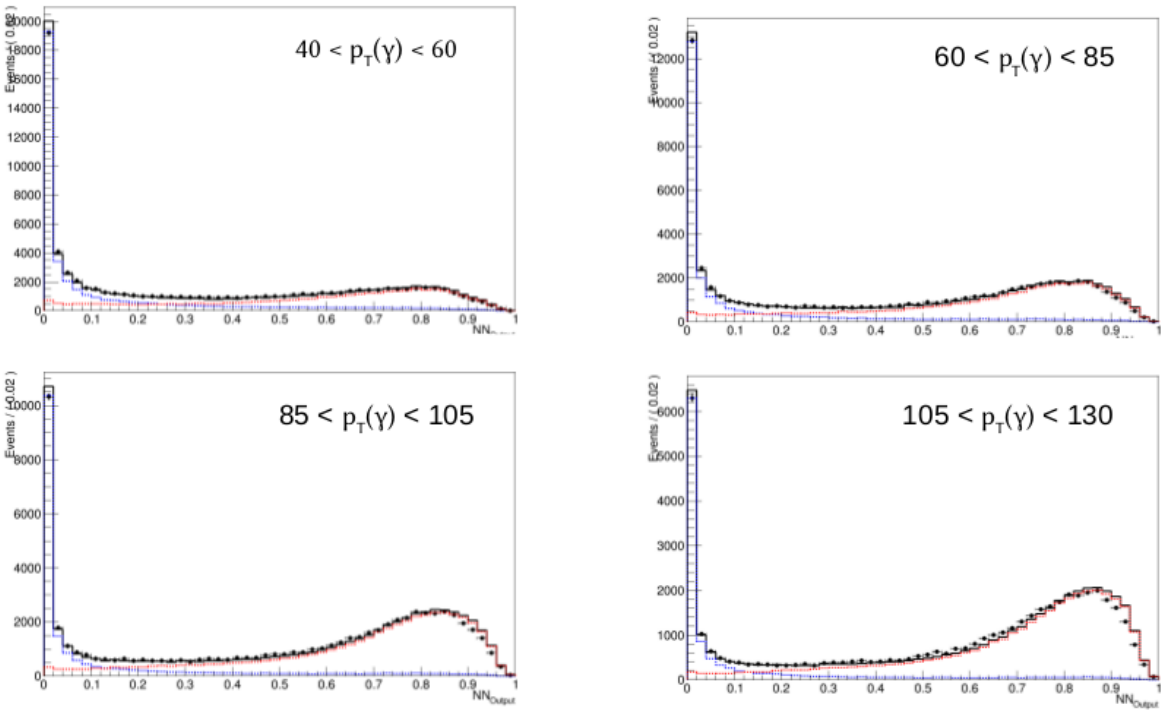


Figure B.1: Maximum-likelihood fit performed with the ANN for 4 bins in pT(γ).

Maximum-likelihood fit for the ANN, showing background PDF (blue dotted line) signal PDF (red dotted line) and fit result (solid black line) superimposed with real data (black dot).

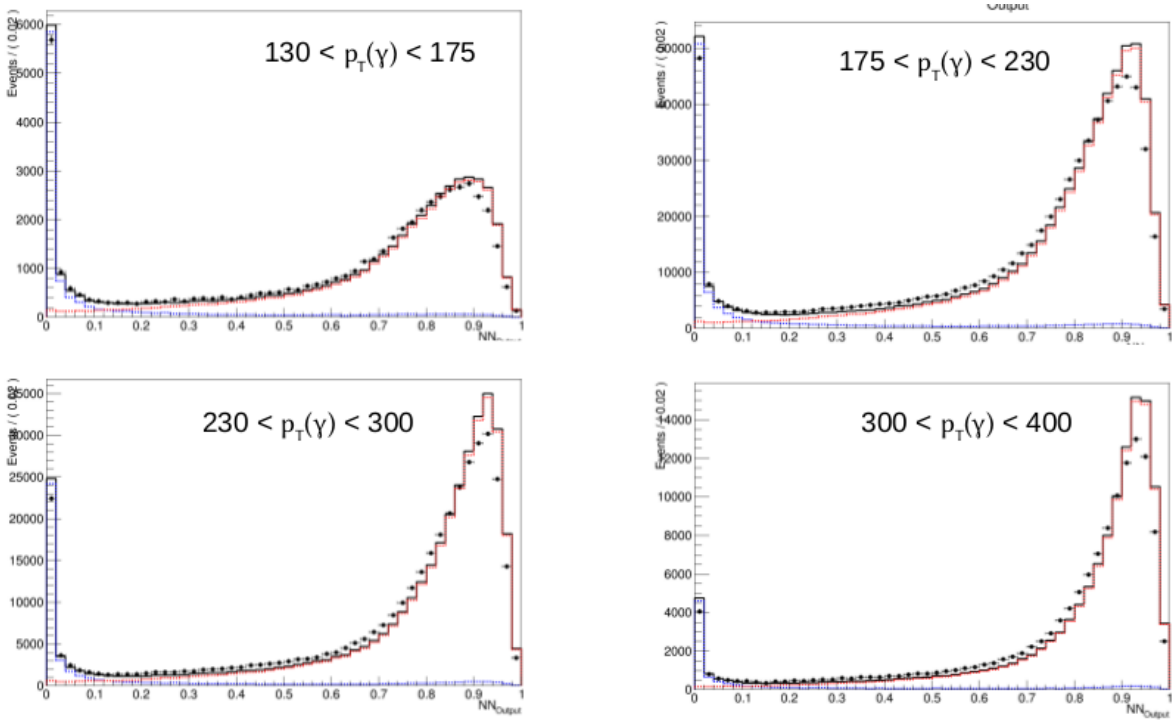


Figure B.2: Maximum-likelihood fit performed with the ANN for 4 bins in pT(γ).

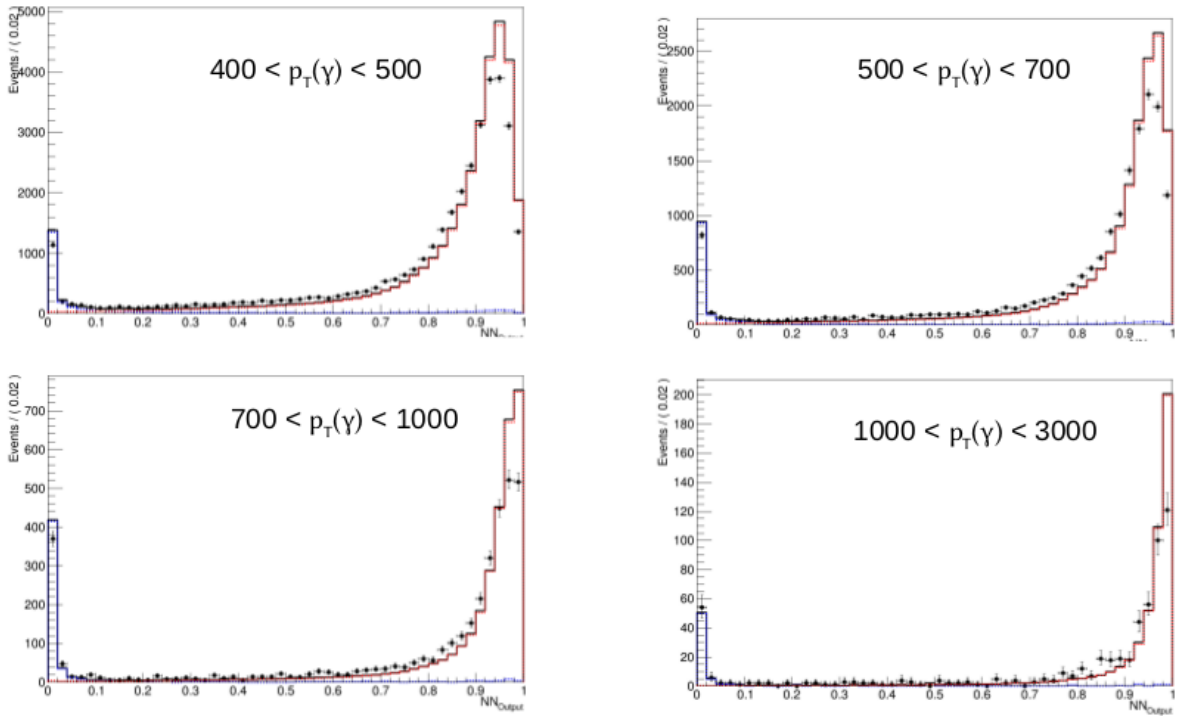


Figure B.3: Maximum-likelihood fit performed with the ANN for 4 bins in pT(γ).

Appendix C

Pull plot

Pull plots for the maximum-likelihood fit performed with 10^4 toyMC experiments for the ANN.

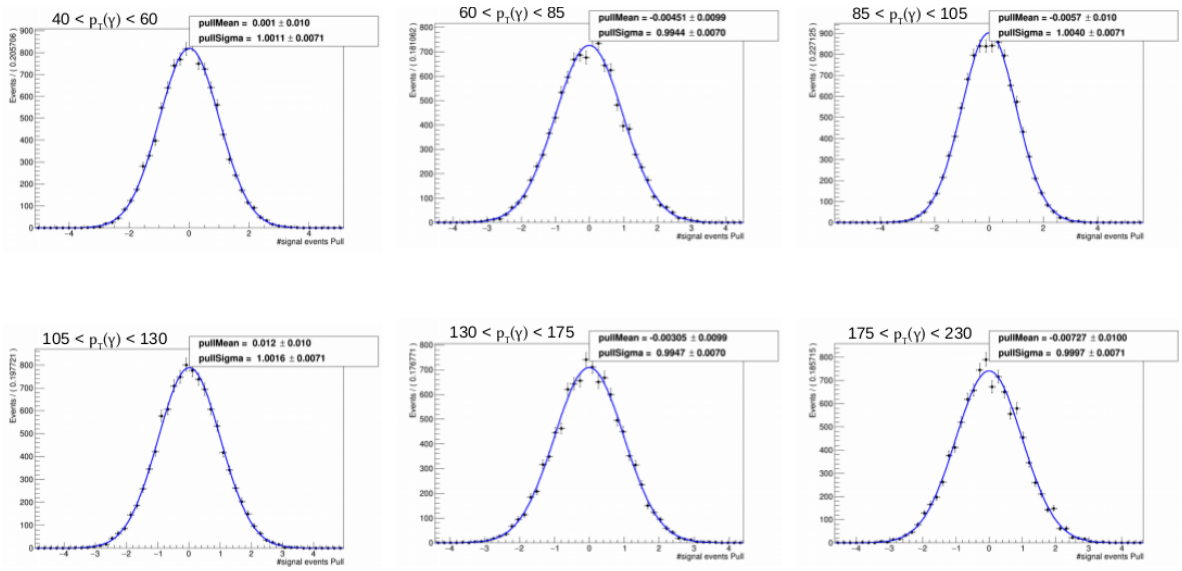


Figure C.1: Pull plots of the maximum-likelihood fit performed with the ANN for 6 bins in pT(γ), showing pull results (black dot) and a gaussian fit (blue line)

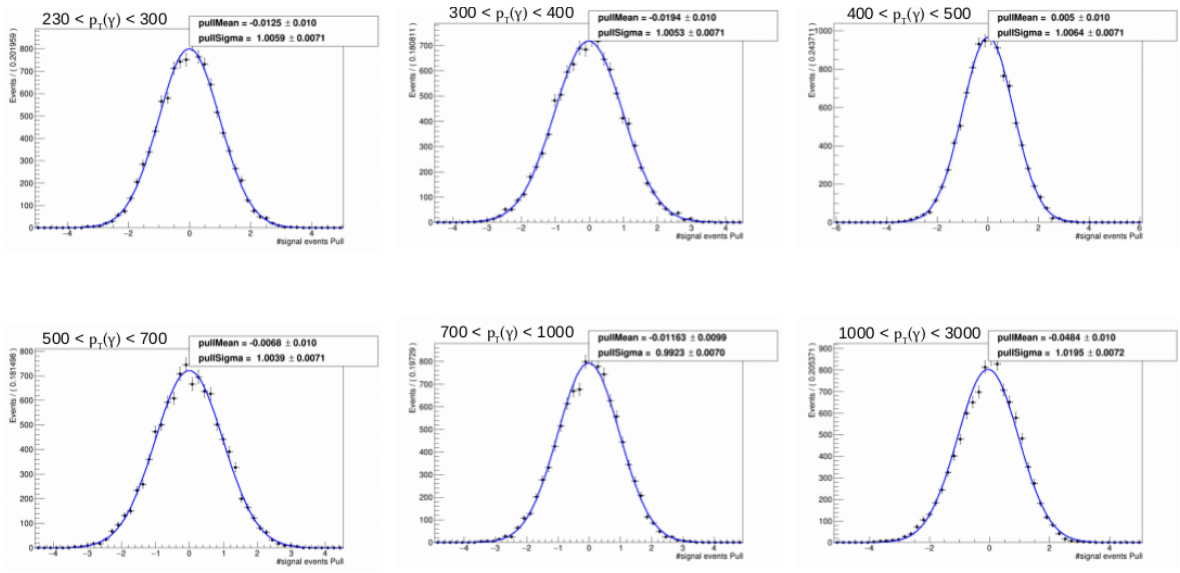


Figure C.2: Pull plots of the maximum-likelihood fit performed with the ANN for 6 bins in pT(γ), showing pull results (black dot) and a gaussian fit (blue line)