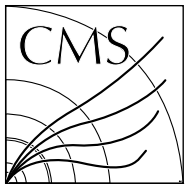
Available on CMS information server

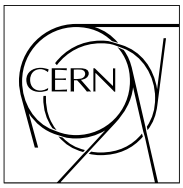
CMS AN -2008/110



The Compact Muon Solenoid Experiment

Analysis Note

The content of this note is intended for CMS internal use and distribution only



11 December 2008

Electron identification in the CMS experiment based on a likelihood algorithm

Emanuele Di Marco [a)], Paolo Meridiani [b)], Chiara Rovelli [a)], Carole Weydert [c)]

Abstract

We describe a likelihood-based algorithm to perform the identification of electrons with the CMS experiment. The observables used in the likelihood function are presented, focusing on the discrimination between real and fake electron candidates coming from mis-identified jets. We describe the control samples and the strategy to define the probability density functions on data. Finally we evaluate the performances of the electron identification in terms of efficiency and mis-identification rate as a function of the kinematics of the electrons.

[a)] U. di Roma "La Sapienza" and INFN Roma

[b)] *CERN PH*, 1211 Geneve, Switzerland

[c)] *CERN summer student*, now with U. Grenoble "Joseph Fourier" and LPSC Grenoble

Contents

1 Introduction

In this note we present a likelihood approach to the electron identification in CMS. We put the emphasis on the identification of electrons in the momentum range $5 < p_T < 80$ GeV/c, which is the range relevant for Standard Model Higgs boson searches (i.e. $H \rightarrow ZZ^* \rightarrow e^+e^-e^+e^-$ and $H \rightarrow WW^* \rightarrow e^+\nu e^-\nu$) and for some measurements with the first LHC data. Examples are the measurement of the cross section of Z+jets and the more challenging W+jets.

The electron identification variables we use as input of the likelihood function have been extensively described elsewhere [1] and widely used in CMS analysis. The identification also profits by sub-dividing the electrons in classes according to the fraction of energy lost in the passage through the tracker material. The possibility of large bremsstrahlung emissions introduces non-Gaussian fluctuations of the calorimetry and tracking measurements, so different classes can have different ECAL-tracker patterns. We consider these differences in the likelihood function.

We present the strategy to determine the probability density functions of the electron identification variables on data control samples with the first recorded data, for both signal and background hypotheses.

We finally discuss the performances of the identification in terms of efficiency on electrons from $W \rightarrow e\nu$ decays and of background mis-identification probability (from jets).

Some results on the effects of ECAL mis-calibrations and tracker misalignment in the LHC start up conditions are discussed, based on Monte Carlo events of the `CSA07` production.

2 Monte Carlo Datasets

The datasets used in this study come from different Monte Carlo samples. The events produced with different generators are passed through the full simulation of the CMS detector response, that relies on the on standard CMSSW software. For most of the results presented here the `Summer08` Monte Carlo production was used. The used samples are:

- W+jets (*MadGraph* matrix-element generator)
- Z+jets (*MadGraph* matrix-element generator)
- $t\bar{t}$+jets (*MadGraph* matrix-element generator)
- QCD jets, enriched in e.m. fraction, for the p_T bins of the leading parton:
  - $20 < p_T < 30$ GeV/c,
  - $30 < p_T < 80$ GeV/c,
  - $80 < p_T < 170$ GeV/c

At the moment of the preparation of this note, the samples are reconstructed with the CMS software release `CMSSW_2_1_8` with ideal conditions, while re-reconstruction with different mis-alignment and mis-calibration scenarios are not yet available. For this reason, the effect of the dispersion of the inter-calibration and alignment constants expected for integrated luminosities of $\sim$10 and 100 pb^{-1} are estimated using the `CSA07` Monte Carlo production reconstructed with release `CMSSW_1_6_7`.

3 Input Variables

This section describes the electron identification variables entering the likelihood function, and the classification used to sub-divide electrons in categories with different characteristics (and different purities). The set of variables used in this algorithm is the same described in the note [1], which are well established ones. Large attention is paid to the ECAL cluster shape variables, which are very correlated among them since they all describe the width of the e.m. deposit of the electrons. Only a limited set of the existing ones has been chosen.

3.1 Electron Classification

The population of the reconstructed electrons is divided into distinct classes, taking into account the amount of the bremsstrahlung and the energy loss in the passage of the electron through the tracker material. This classification is used to account for non-Gaussian sources of fluctuations of the ECAL supercluster energy and tracker momentum measurement, and it also results suitable to distinguish the different track-supercluster patterns with consequent different performances of the electron identification. The four, mutually exclusive, electron classes are described in the note [1]. Here we give only a brief description of the properties of each class:

- *golden electrons:* this class represents the most precisely measured electrons, which are least affected by bremsstrahlung and have a good track-supercluster match. The pattern in the ECAL is characterized by a single "seed" cluster.
- *big brem electrons:* this class contains the non-golden electrons characterized by a single "seed" cluster in ECAL, but with a large fraction of the initial energy radiated very early or very late in the tracker, resulting in the simple energy deposition in the ECAL.
- *narrow electrons:* this intermediate class contains electrons which still have a single "seed" cluster in ECAL, lower bremsstrahlung that the ones belonging to the big brem ones, but have a relaxed track-supercluster geometrical match.
- *showering electrons:* this class contains the electrons which are badly measured, due to an early radiation of a high amount of the electron energy, resulting in a supercluster made of multi sub-clusters.

The fraction of electrons in a given category is estimated on a sample of electrons from W decays having a transverse momentum $20 < p_{\mathrm{T}} < 50$ GeV/c. We also estimate these fractions for jets mis-reconstructed as electrons in a sample of W+jets with the jet in the same momentum range. The results are shown in Table 1. We

	electrons	*jets*
golden	17%	7%
big brem	5%	0.5%
narrow	8%	0.5%
showering	70%	92%

Table 1: Population of the four electron classes for real electrons coming from W decays, having a transverse momentum $20 < p_{\mathrm{T}} < 50$ GeV/c and for fake electrons in W+jets sample in the same momentum range.

did not performed a dedicated study of the electrons falling in the ECAL inter-module cracks and in the larger crack between the ECAL barrel and endcap and we treat them as showering electrons.

Most of the jets mis-reconstructed as electrons are classified as showering, while very few populate the big brem and narrow classes. This makes difficult to model the probability density functions (*PDFs*) of variables for the two intermediate classes using the first hundreds pb^{-1} of integrated luminosity. For this reason we merge the first three classes in the list above, characterized by having a single cluster in the ECAL, in a unique class, which we define as *non-showering*. In Fig. 1 we show the relative population of each class for real electrons and jets as a function of the candidate η and p_T. For real electrons the fraction of showerings is proportional to the amount of tracker material crossed by the electron, and therefore presents strong variations along η; it is quite constant with the transverse momentum, at least for $p_T > 15$ GeV/c. The fractional radiation length x/X_0 as a function of η is shown in Fig. 2 for the different sub-detectors in front of ECAL [2].

3.2 Electron Identification Variables

In this section we discuss the variables used to establish the compatibility of the reconstructed electron candidate with the track and supercluster pattern expected from a single real electron. The distributions of these variables (separately for barrel and endcap, non-showering and showering candidates) are used as inputs for the likelihood algorithm. Since the cross-correlation between the variables plays an important role in the performances of the algorithm we checked that the proposed set has correlations small enough both for signal and background.

The variables used in the likelihood algorithm coincide with the ones described in [1]. A cut-based electron identification exploiting the same variables has been extensively used in many analyses (ex. W/Z+jets ratio [**?**] and $H \rightarrow WW^*$, [4]), both performed on the CSA07 Monte Carlo samples). They are:

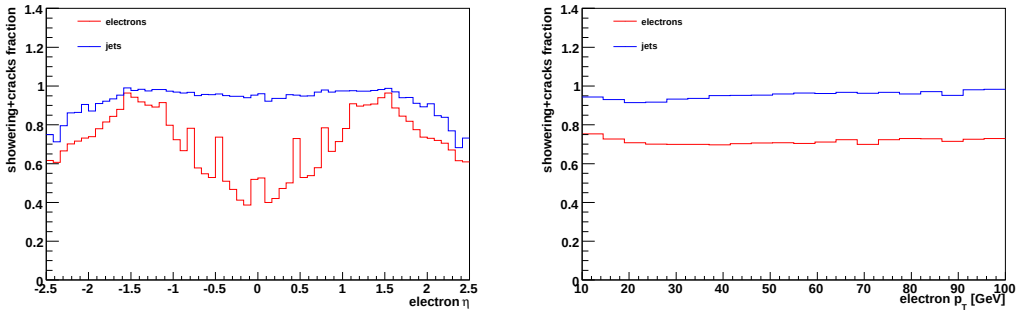


Figure 1: Fraction of candidates classified as showering or cracks for true electrons from Z or for mis-reconstructed jets as a function of η (left) and p_T.

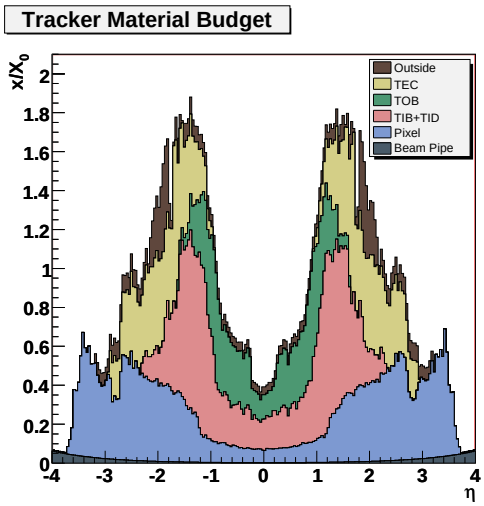


Figure 2: The fractional radiation length x/X_0 as a function of η for the different sub-detectors in front of ECAL.

- ratio of the energy of the supercluster seed over the track momentum at the last tracker layer, $E_{\rm seed}/p_{\rm out}$
- geometrical matching between the track parameters at the interaction vertex extrapolated to the super cluster and the measured super cluster position, $\mid \Delta\eta_{\rm in} \mid \; = \; \mid \eta_{\rm SC} - \eta_{\rm in}^{\rm extrap} \mid$ and $\mid \Delta\phi_{\rm in} \mid \; = \; \mid \phi_{\rm SC} - \phi_{\rm in}^{\rm extrap} \mid$
- ratio of the energy deposited in the HCAL towers in a cone of radius $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} = 0.1$ centered on the electromagnetic supercluster position over the supercluster energy, H/E
- ratio of the energy sums over the 3×3 and 5×5 matrices centred on the highest energy crystal of the seed cluster, $\sum_9 / \sum_{25}$
- the width of the ECAL cluster along the η direction: $\sigma_{\eta\eta}^2 \; = \; \sum_{\rm crystal} (\eta_i - \eta_{\rm seed})^2 \frac{E_i}{E_{\rm seed}}$. We didn't applied the correction to account different crystal geometry in the endcaps[1)].

The distributions of the identification variables are made in the signal case with candidate electrons coming from the decay of a Z boson (*probe*), when the other electron of the Z is selected as a good electron (*tag*). We describe in Sec. 4.2 the selection of the tag and probe objects and the strategy to obtain the signal PDFs from that dataset. As a background we consider the hadrons ($\pi^{\pm}$, $K^{\pm}$,...) inside jets which are reconstructed as electron candidates due to a some energy release in the ECAL of the hadron itself, or of the neighbouring particles inside the jet, (*fake electrons*). Jets are produced with high rate in QCD processes and constitute one of the main backgrounds for many analyses. Again, we describe a control sample on data with which characterizing the jet PDFs in Sec. 4.2.

[1)] We plan to apply corrections for endcap when we make the PDFs with the new samples of the Summer08 Monte Carlo production.

The possible correlation between electron identification and the *isolation* variables has to be accounted for and the study of the performances of the electron identification cannot be independent by the isolation criteria applied in the electron selection. As an example, H/E can depend on the jet multiplicity in the event. We therefore consider only loose isolated candidate electrons, both for signal and background.

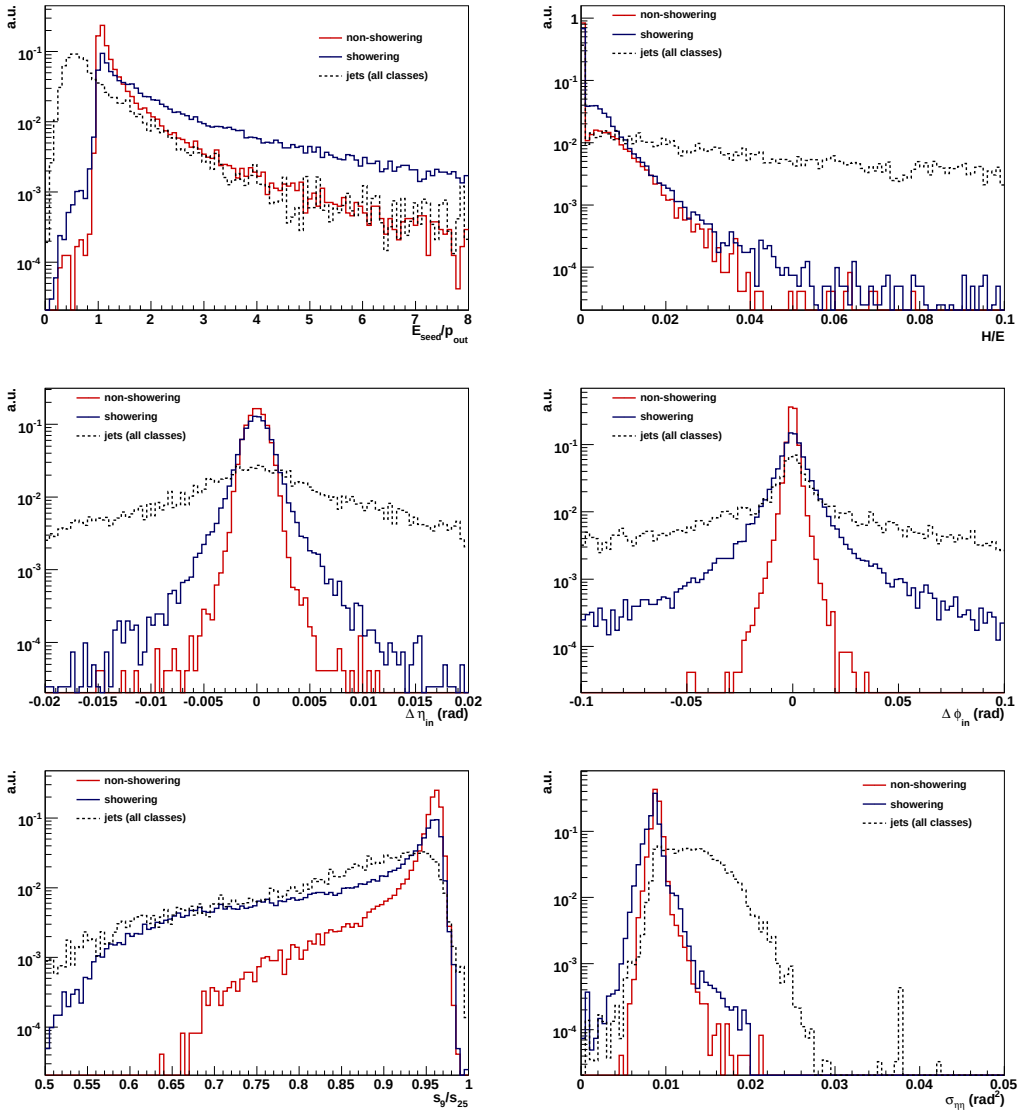


Figure 3: Distribution, normalized to unity, of the electron identification variables used as input in the likelihood for barrel. The signal distributions are split according the classification described, the background ones are unsplit, as they enter the likelihood. Top: $E_{\rm seed}/p_{\rm out}$(left), H/E(right). Middle: $\Delta\eta_{\rm in}$(left), $\Delta\phi_{\rm in}$(right). Bottom: $\sum_9 / \sum_{25}$(left), $\sigma_{\eta\eta}$(right).

Fig. 3 and Fig. 4 show the discriminating variables used as input in the likelihood algorithm for barrel and endcap electrons, respectively. For simplicity, we give only the variables for the kinematic bin $p_T > 15$ GeV/c. The double peak in the $\Delta\eta_{\rm in}$variable for endcap electrons is supposed to be correlated with the opposite tilt of crystals in the two ECAL endcaps. The electrons with $\eta > 0$ have mainly $\Delta\eta_{\rm in} > 0$ and vice-versa.

4 Extraction of PDFs from data control samples

In order not to rely too much on Monte Carlo description of the electron identification variables, the distributions can be extracted from data control samples both for signal and for background. In the following we describe the control samples that can be used on data and the strategy to

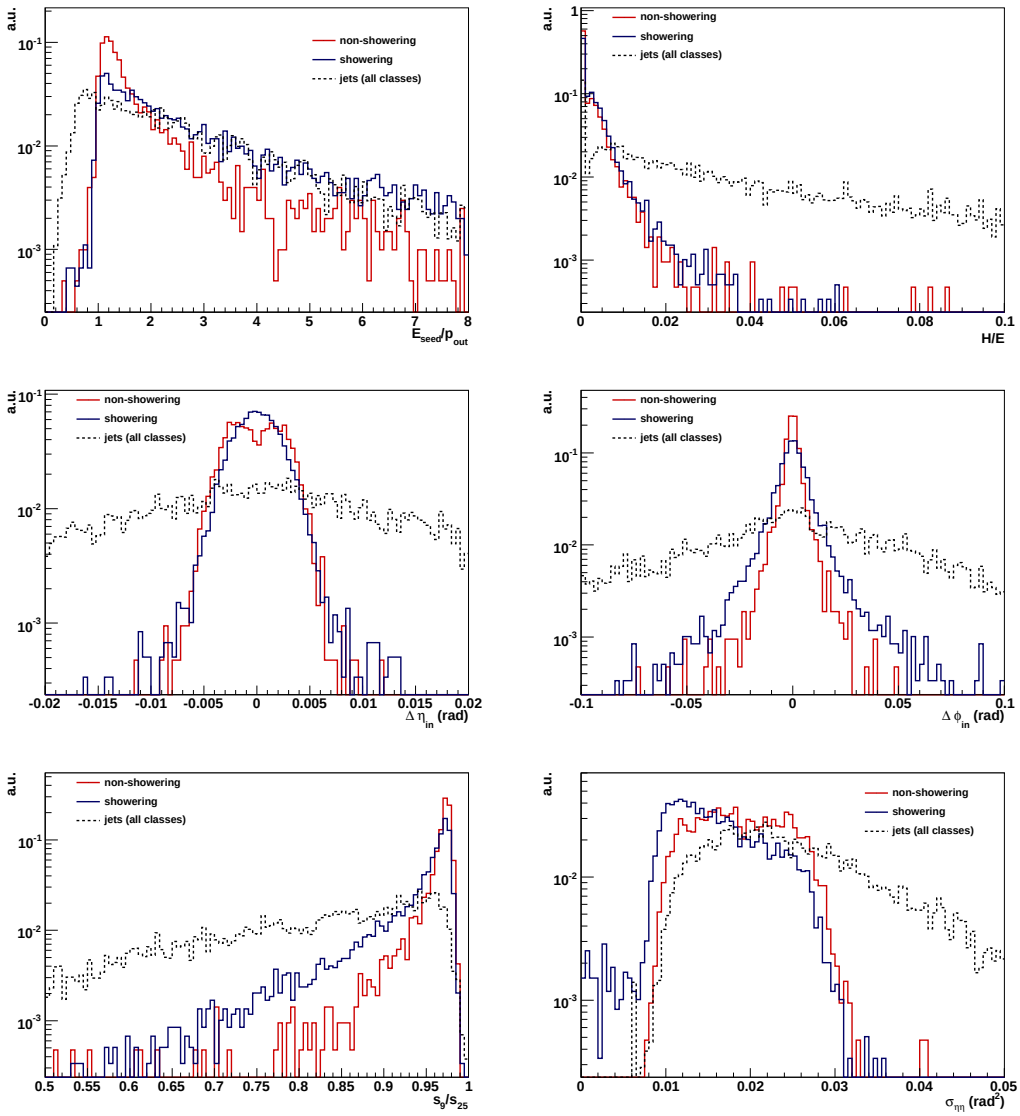


Figure 4: Distribution, normalized to unity, of the electron identification variables used as input in the likelihood for endcap. The signal distributions are split according the classification described, the background ones are unsplit, as they enter the likelihood. Top: $E_{\rm seed}/p_{\rm out}$(left), H/E(right). Middle: $\Delta\eta_{\rm in}$(left), $\Delta\phi_{\rm in}$(right). Bottom: $\sum_9 / \sum_{25}$(left), $\sigma_{\eta\eta}$(right).

4.1 Electron PDFs from Z decays

A quite clean electron control sample can be extracted by the Z decays. The production of the Z boson can be associated to the production of jets. The cross section of this process decreases roughly as a power of α_s with the jet multiplicity. We use the Z+jets Monte Carlo sample produced with MadGraph matrix-element calculator.

To extract a clean sample of electrons to model the PDFs of the electron identification variables we use the *tag and probe* method. This method is also used to estimate the electron reconstruction and identification efficiencies [5].

This method consists of selecting an electron with some quality criteria applied (*tag*) and look for another one in the event which, combined with the tag, form an invariant mass close to the one of the Z (*probe*). On the probe electron we study the electron identification variables. In order to have the largest electron sample, if the electron selected as a probe also fulfill the quality criteria for a tag, the roles are inverted and the first electron is used as a probe.

We require at least two reconstructed electrons (*pixelMatchGsfElectron*) in the event[2] with:

- $|\eta| < 2.5$
- $p_T > 5$ GeV/c

If more than 2 electrons are selected in the event, we choose the two which give the invariant mass closest to the Z nominal mass [6]. To clean up the sample we apply loose identification and isolation criteria on the tag electron:

- loose category based electron identification, defined in [7]
- loose tracker isolation: $\sum p_T/p_T^{electron} < 0.20$ in a cone of $\Delta R < 0.4$ around the electron track

Since in a typical analysis the electron identification is used together with isolation criteria, we also apply the same loose isolation on the probe electron. This requirement does not affect too much the electron identification variables on true electrons, while we expect to have more effect on background (mainly in H/E).

Even if the background under the Z mass peak is expected to be small, still some contamination can arise and distort the shapes of the signal PDFs. We apply a statistical background subtraction which makes use of the full Z lineshape extracted on data. In order to do this, we consider a loose requirement on the di-electron invariant mass:

- $40 < m_{e^+e^-} < 110$ GeV/c^2

and we assign to any event the probability to be signal or background through a maximum likelihood fit to the di-electron invariant mass.

We model the invariant mass for signal with a Cruijff function [3], defined as:

$$f(x; m, \sigma_L, \sigma_R, \alpha_L, \alpha_R) = N \times \exp\left[-\frac{(x-m)^2}{2\sigma_{L/R}^2 + \alpha_{L/R}(x-m)^2}\right] \quad (1)$$

where the σ_L and α_L (σ_R and α_R) corresponds to resolution and tail parameters of the distribution for $x - m < 0$ ($x - m > 0$). The use of this function allows to describe the tail in the distribution,induced by the mis-reconstruction of the energy of the electrons, due to possible leakage in the calorimeter or to large bremsstrahlung emission in the tracker material[3)].

We apply the selection to the Z+jets Monte Carlo sample, correspondent to 300 pb^{-1}of integrated luminosity. If both electrons are selected as tag, then the two probes enter the dataset with the same invariant mass.

We apply the same selection on the samples that can be considered as the most relevant for the Z+jets process:

- W+jets
- $t\bar{t}$+jets
- QCD di-jets

For W+jets and $t\bar{t}$+jets we apply the selection only on a subset correspondent to a luminosity of 300 pb^{-1}, as for the signal. Instead we use all the available Monte Carlo statistics for QCD di-jets. We consider the selected events in these samples as a unique background, and we parameterize the di-electron invariant mass as a second order polynomial. The distribution of the invariant mass for the tag-probe pairs is shown on Fig. 5, with the result of the fit superimposed, both for signal and background components.

In order to apply our strategy on a sample similar to the one that is selected on data, we merge together the signal and background events to get 300 pb^{-1}equivalent data. We then perform an unbinned maximum likelihood fit to this dataset fixing the background shape to Monte Carlo, while leaving the signal lineshape floating as well as the signal and background yields.

As output of the fit we get the signal and background yields, as well as the signal Cruijff function parameters, consistent with the expected values. The fit to the data-like sample is shown in Fig. 6.

[2)] In the Monte Carlo samples we used the trigger bits are not saved, so we didn't required any trigger to be fired. The plan, as soon as the information is saved, is to require the single electron trigger path.

[3)] The mean and the $\sigma_{L,R}$ can be slightly biased due to the choice of the best electron pair, but we checked that this does not shrink the background to peak near the Z mass peak.

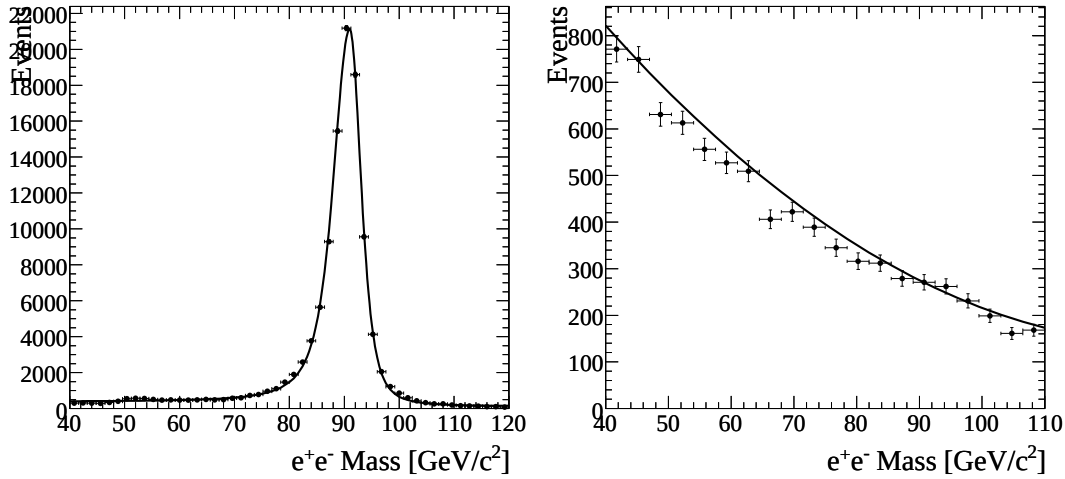


Figure 5: Distribution of the tag and probe electrons invariant mass for Z+jets events (left) and W+jets, $t\bar{t}$+jets events, QCD events (right) with the result of the fit superimposed. The drops in the e^+e^- invariant mass in the right distribution are due to the generation thresholds applied to the different Monte Carlo samples composing the background.

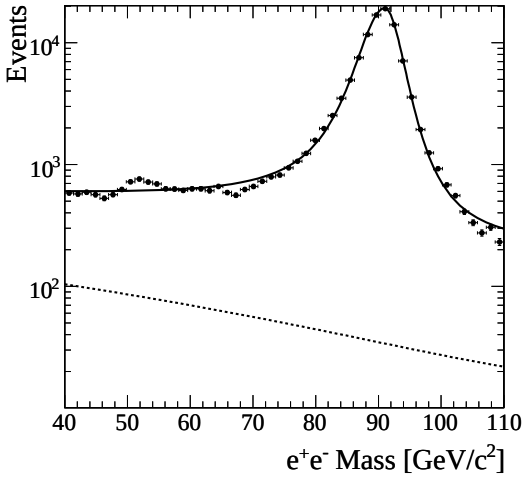


Figure 6: Distribution of the tag and probe electrons invariant mass for a data-like merged sample formed by Z+jets, W+jets, $t\bar{t}$+jets and QCD events, for an integrated luminosity of 300 pb^{-1}, with the result of the fit superimposed. The solid (dashed) line represents the projection of the signal+background (background only) likelihood, as obtained from the fit result. The bump around 50 $\mathrm{GeV/c}^2$is due by the Z generation threshold.

The value of the likelihood is then used to compute the signal ${}_sWeight$ [8], which is proportional to the probability for that event of being signal[4)]. We form the distributions of electron identification variables weighting each event with its signal ${}_sWeight$. In this way we are able to extract the signal PDFs directly on data with a statistical subtraction of background which fully exploit all the di-electron invariant mass shape. We also do not lose any statistics as if we would have done a background subtraction from the sidebands of the Z mass peak. We call these distributions ${}_sPlots$. They have the characteristic that the integral of the distribution correspond to the fitted number of signal events in the dataset.

We compare in Fig. 7 the distributions of the electron identification variables, for simplicity for barrel and endcap together, and low and high p_T bins together, as extracted from pure Z+jets Monte Carlo sample and as extracted on data-like sample after the background subtraction with the ${}_sPlots$ averaging technique. The proof is limited by

[4)] The value of signal and background ${}_sWeight$ can be negative to account for statistical fluctuations of the background component.

the fact that we used the same sample to parameterize the signal di-electron mass shape and to form the data-like sample, but it shows that the subtraction of the background is satisfying. The signal PDFs can be then defined on data with the first hundreds of $\mathrm{pb}^{-1}$of integrated luminosity[5)].

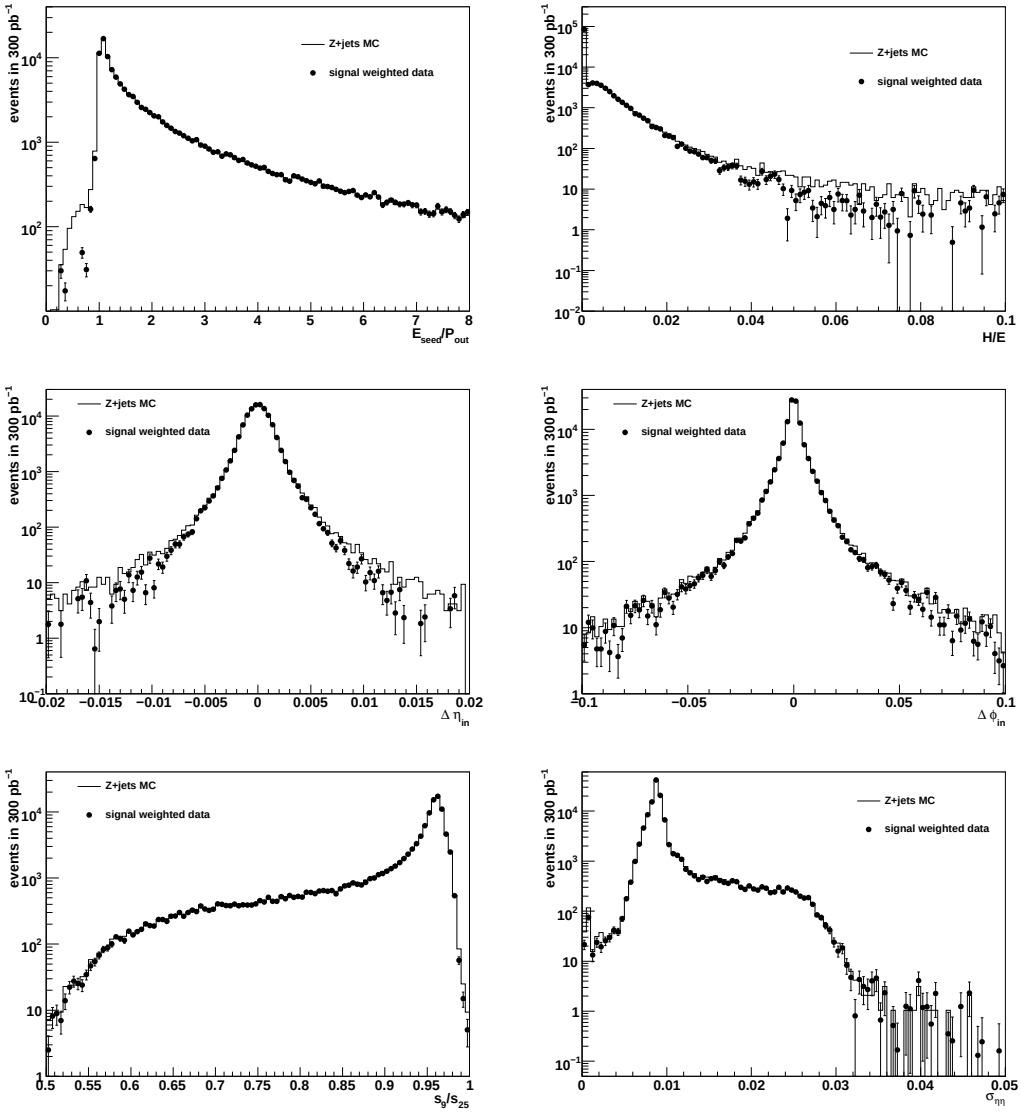


Figure 7: Distribution, normalized to expected events in 300 pb^{-1}, of the electron identification variables used as input in the likelihood for barrel and endcap together. The solid histogram represents the probe electrons from signal Monte Carlo Z+jets events, the dots are the signal ${}_sPlots$ as evaluated from a maximum likelihood fit to a merged sample of signal and background equivalent to 300 $\mathrm{pb}^{-1}$of integrated luminosity. Top: $E_{\mathrm{seed}}/p_{\mathrm{out}}$(left), H/E(right). Middle: $\Delta\eta_{\mathrm{in}}$(left), $\Delta\phi_{\mathrm{in}}$(right). Bottom: $\sum_9 / \sum_{25}$(left), $\sigma_{\eta\eta}$(right).

4.2 Jet PDFs from QCD di-jets and W+jets

In the LHC environment jets are produced with a huge cross section, being driven by strong coupling, with respect the electro-weak processes. The main process is the production of di-jets events. It can be used as a high statistics control sample to estimate the PDFs of the electron identification variables for electron fake candidates coming from jets.

The recording of di-jets events in the CMS data acquisition is driven by a dedicated jet trigger, which is prescaled to satisfy the HLT rate requirements due to the very high rate. The cross section of QCD jet production with p_T

[5)] The uncertainties on them are statistical only.

of the leading parton between $20 < p_T < 170$ GeV/c is about 0.5 mb. The productions of W or Z bosons with associated production of jets are electroweak processes, therefore the cross sections are much lower (respectively 18 nb and 1.5 nb). Assuming that the jet trigger efficiency is roughly similar for a QCD di-jet event and for a $W(Z)$+jets event, then the contamination of electrons from vector bosons in the jet triggered sample is of the order of 0.01%. In order to model the PDFs for the electron identification, at least one electron has to be reconstructed and this requirement enhances the pollution of real electrons from V+jets, V being a vector boson. The electron reconstruction efficiency is about 95%, while the probability for a jet to be reconstructed as an electron is order of 10% in the worst case (jets of $p_T > 20$ GeV/c), as will be discussed in Sec. 6.2. The PDFs are built also requiring that the reconstructed electron is loose isolated in the tracker. This request has an efficiency of about 95% on the signal, while reducing the QCD jets about of a factor 50%. It therefore further lowers the average purity of the jet trigger sample to the order of fraction of percent.

The V+jets contamination depends on the fake electron spectrum, since i.e. the p_T spectrum of jets falls quicker than that of real electrons. In the following we look for the leading jet according to its p_T (which plays the role of the tag jet) and we define the probe as a reconstructed electron which is back-to-back to the tag jet in the transverse plane. If more than one probe electron is reconstructed, we choose the furthest one from the leading jet. The tag jet is reconstructed with the IterativeCone algorithm with ΔR=0.5; it's required to have uncorrected $p_T > 30$ GeV/c and $|\eta| < 25$, to roughly emulate the trigger. In Fig. 8 we show the contamination of electrons from W+jets and Z+jets events in the selected QCD sample as a function of the fake electron η and p_T. Such contamination is defined as the number of probe electrons matching a real electron in $W(Z)$+jets events over the total number of probes reconstructed in W+jets, Z+jets and QCD events. The background contamination in the QCD control

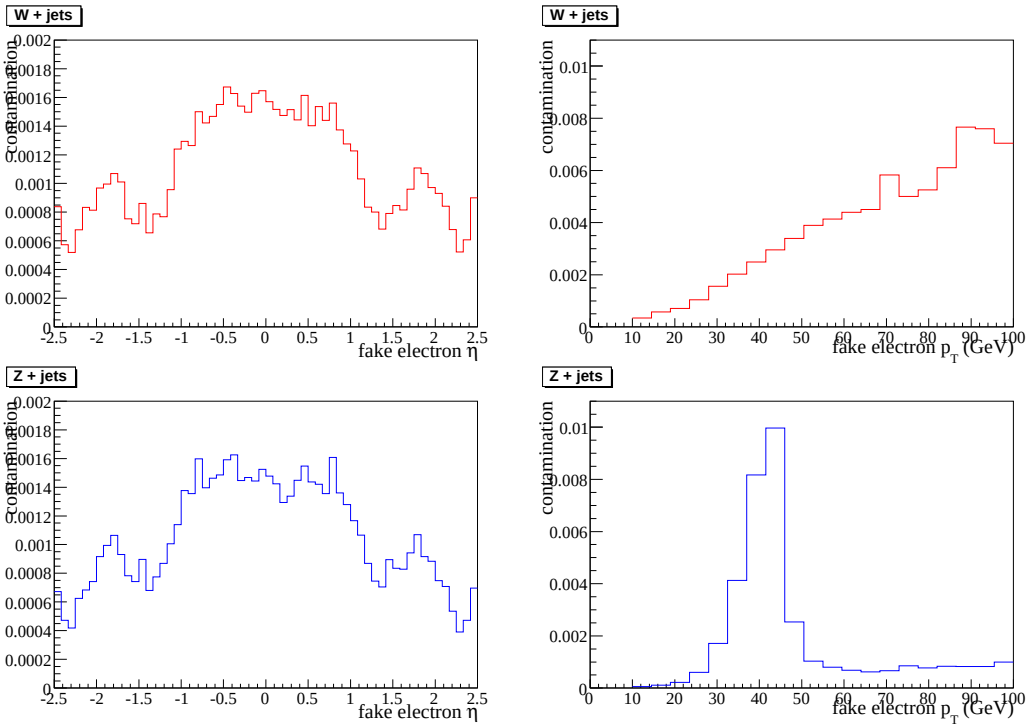


Figure 8: W+jets (top) and Z+jets (bottom) contamination in the proposed background control sample. The contamination is given as a function of the fake electron candidate η (left) and p_T (right).

sample is better than 1% in the full range. The purity can be further enhanced exploiting the kinematics of the events. As an example the following requirements can be applied:

- the uncorrected missing transverse energy of the event is required to be $MET < 20$ GeV
- the angle $\Delta\phi = |\phi_{tagjet} - \phi_{electron}|$ between the tag jet and the probe electron should be close to π (ex. $\Delta\phi < 2.5$) for a back-to-back di-jet event, while the directions are less correlated in the W+jets events due to the $W \rightarrow e\nu$ decay.
- the invariant mass between the tag jet and the probe electron is $m_{jet-electron} < 60$ GeV/c^2

The first and the second requirements can be used to suppress the amount of W+jets events. The proposed threshold on missing transverse energy is chosen as the complementary one for the W+jets selection in [3]. The efficiency of this requirement is about 75% for QCD jets events, while it is about 10% for W+jets events. With this selection considered, the pollution of W+jets events is of the order of 3×10^{-4}. The Z+jets sample is dominated by the Z+0 jets where the two electrons are back-to-back and since an electron is always reconstructed as a jet the $\Delta\phi$ distribution is highly peaked. These events can be anyway suppressed to a negligible level applying the invariant mass based criterion.

The distributions of missing transverse energy, $\Delta\phi$ and di-electron invariant mass for the QCD di-jet, W+jets and Z+jets are shown in Fig. 9.

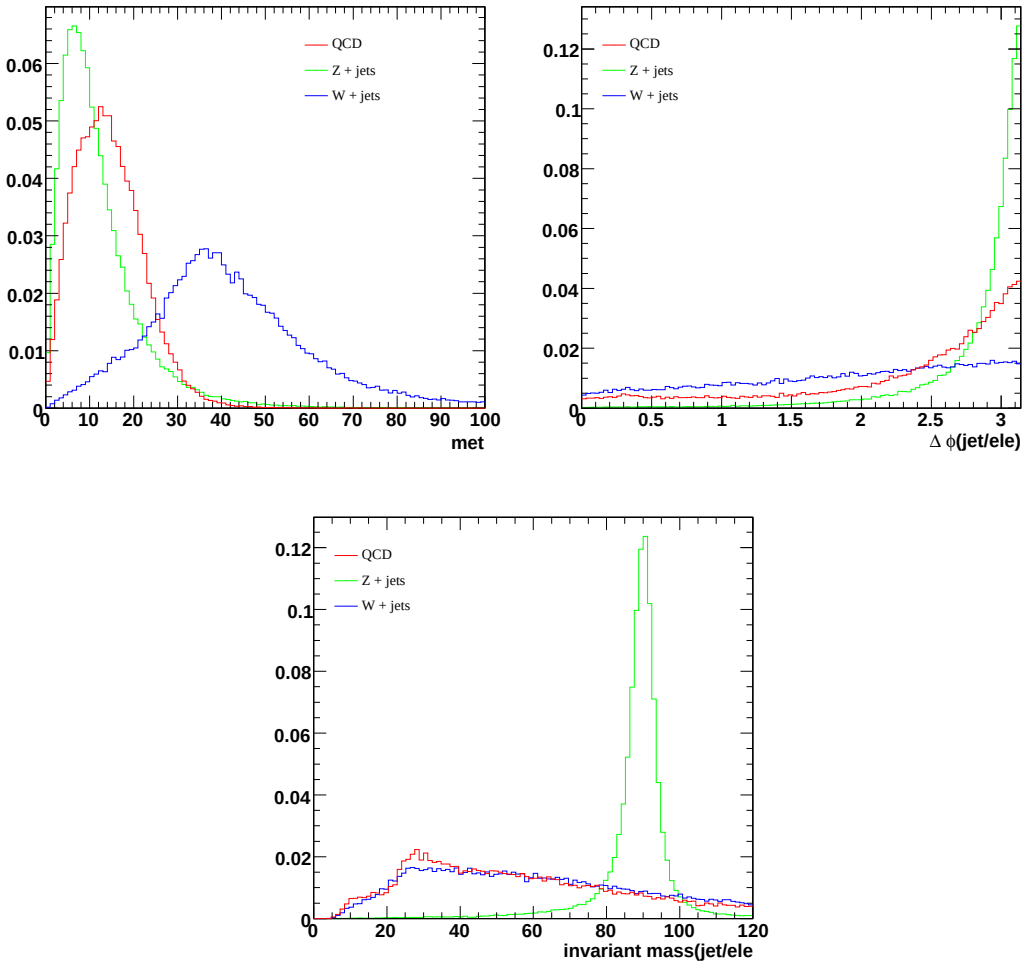


Figure 9: Distribution of the missing transverse energy (top, left), $\Delta\phi$ (top, right) and di-electron invariant mass (bottom) for the QCD di-jet, W+jets and Z+jets.

Examples of processes which produces fake electrons are QCD di-jets (fake background for W+jets) and W+jets (for dibosons, ex. WW and $H \rightarrow WW^*$). We compare the electron identification variables for fake electrons reconstructed in the QCD di-jet samples and the ones reconstructed in the W+jets samples in Fig. 10 and 11 for barrel and endcap, respectively. No significant difference is seen in most of the variables for the two samples, only the cluster shape variables seem to be slightly affected. We will investigate if the differences in the cluster shape variables are due to the quark content of the jets in further studies.

5 Likelihood Function Definition

We described the observables that can be used to discriminate between real electrons and hadrons in Sec. 3.2. Probability Density Functions (PDFs) are constructed for each of them from control samples on data, as described

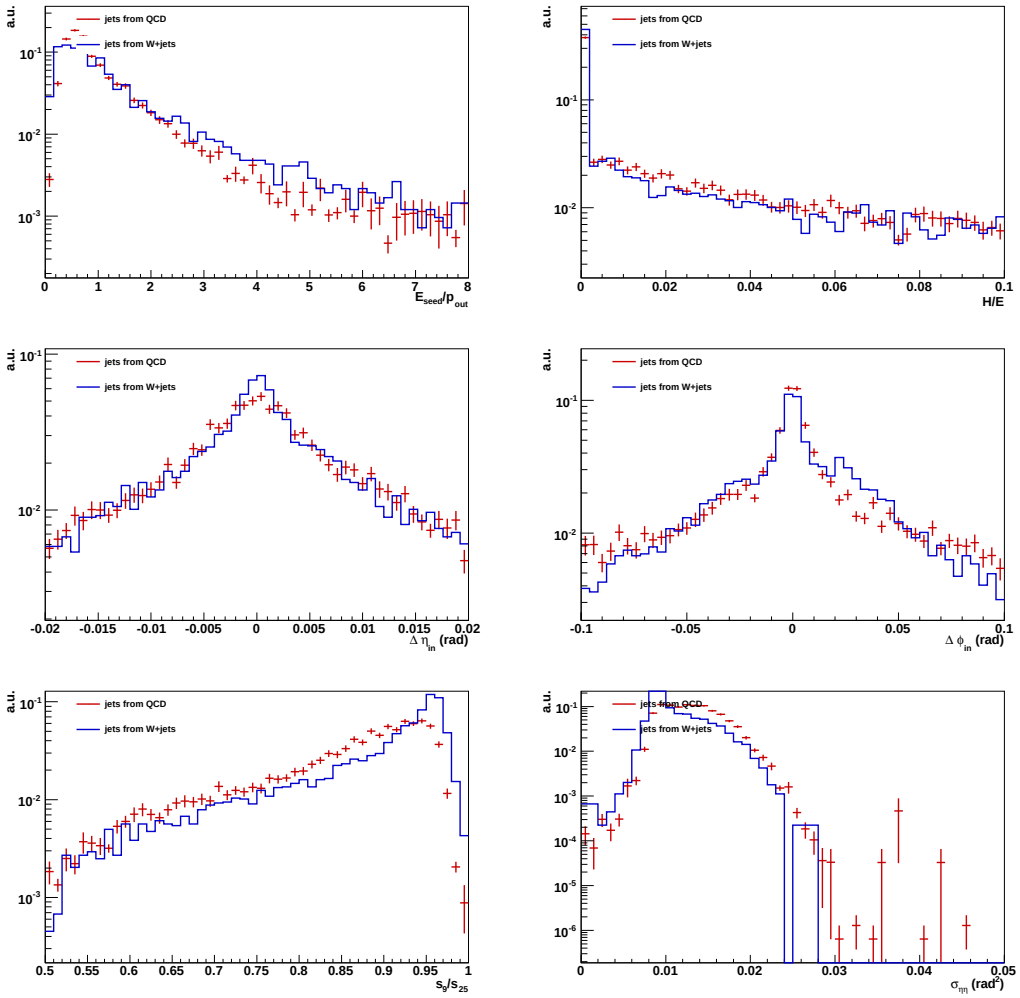


Figure 10: Distribution, normalized to unity, of the electron identification variables used as input in the likelihood for fake electrons in barrel reconstructed in a sample of W+jets (blue line) and QCD di-jets (red dots). The distributions are for all the electron classes together. Top: $E_{\rm seed}/p_{\rm out}$(left), H/E(right). Middle: $\Delta\eta_{\rm in}$(left), $\Delta\phi_{\rm in}$(right). Bottom: $\sum_9 / \sum_{25}$(left), $\sigma_{\eta\eta}$(right).

in Sec. 4.2. Under the assumption of indepedent measurements of these variables, they are combined to compute the likelihood $L_{k,c}(\xi)$ for:

- two particle hypotesis $\xi = \{e, jet\}$,
- 4 kinematic bins
  $k = \{(p_T < 15{\rm GeV}/c; {\rm b}arrel), (p_T > 15{\rm GeV}/c; {\rm b}arrel), (p_T < 15{\rm GeV}/c; endcap), (p_T > 15{\rm GeV}/c; endcap)\}$,
- 2 electron classes
  $c = \{{\rm non} - showering, showering\}$:

The likelihood function is defined as the product of the single variable PDF ($\mathcal{P}_{k,c}(x;\xi)$):

$$\begin{aligned} L_{k,c}(\xi) \quad = \quad & \mathcal{P}_{k,c}(E_{\rm seed}/p_{\rm out};\xi) \cdot \mathcal{P}_{k,c}(H/E;\xi) \cdot \mathcal{P}_{k,c}(\Delta\eta_{\rm in};\xi) \cdot \\ & \mathcal{P}_{k,c}(\Delta\phi_{\rm in};\xi) \cdot \mathcal{P}_{k,c}(\textstyle\sum_9 / \sum_{25};\xi) \cdot \mathcal{P}_{k,c}(\sigma_{\eta\eta};\xi) \cdot \end{aligned} \tag{2}$$

Weighting the individual likelihoods with their *a priori* probabilities p_ξ, we define the likelihood ratio as:

$$r = \frac{p_e L(e)}{p_e L(e) + p_{jet} L(jet)} \tag{3}$$

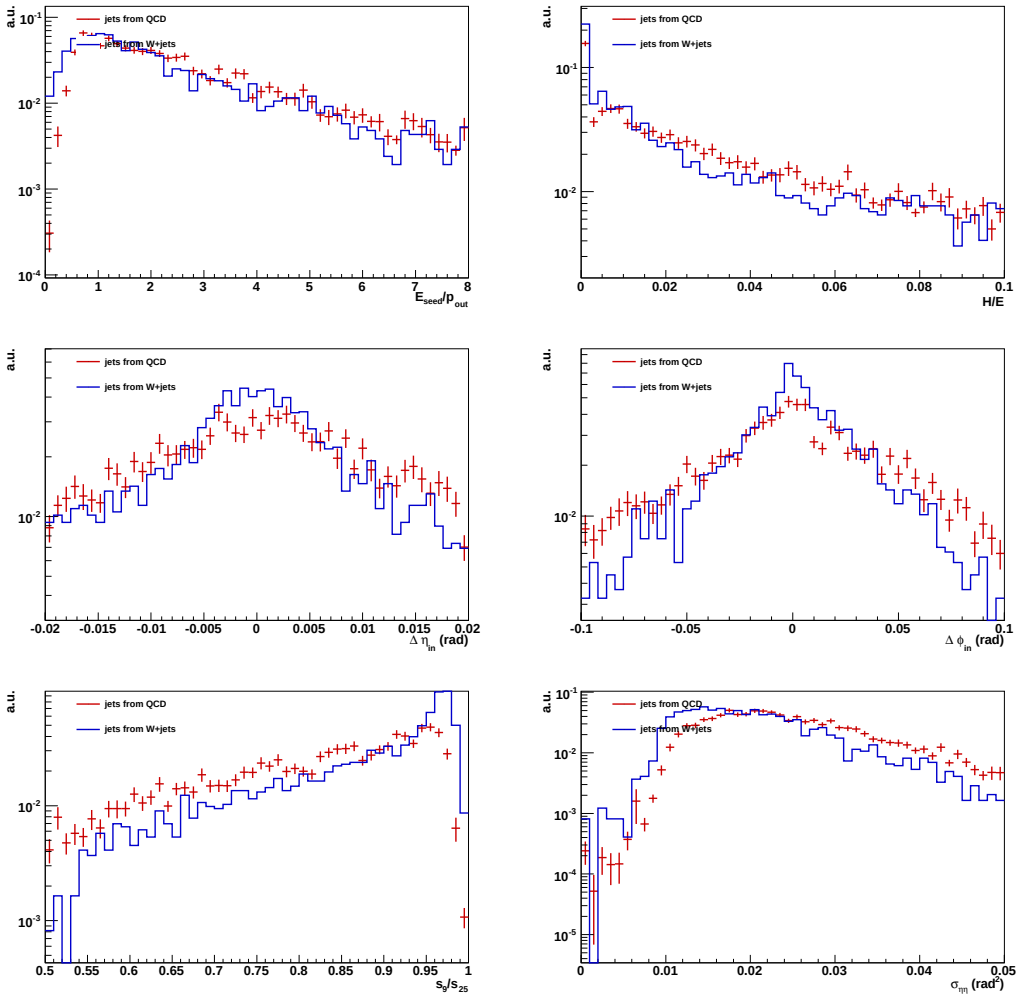


Figure 11: Distribution, normalized to unity, of the electron identification variables used as input in the likelihood for fake electrons in endcap reconstructed in a sample of W+jets (blue line) and QCD di-jets (red dots). The distributions are for all the electron classes together. Top: $E_{\rm seed}/p_{\rm out}$(left), H/E(right). Middle: $\Delta\eta_{\rm in}$(left), $\Delta\phi_{\rm in}$(right). Bottom: $\sum_9 / \sum_{25}$(left), $\sigma_{\eta\eta}$(right).

Since the a priori probabilities depend on the trigger settings, and these are not yet defined, we set them all equal to 1, i.e. assuming no a priori knowledge.

This likelihood ratio can be used as an electron identification variable asking a reconstructed electron to satisfy a given threshold on r, which may vary between 0 and 1.

The variables that enter the likelihood definition are the ones used in the cut based identification described in [1]. The product of the single PDFs in Eq. 2 can be strictly interpreted as a probability only in the hypotesis of the variables being uncorrelated. We found that the correlation among the used electron identification variables is at the percent level for both the showering and non-showering electrons (exceptions are: about 6% between $E_{\rm seed}/p_{\rm out}$ and $\sum_9 / \sum_{25}$for the non-showering electrons and 15% between $E_{\rm seed}/p_{\rm out}$and $\sigma_{\eta\eta}$for the showering ones). Given this level of correlation, we decided to still use all the variables of [1] in the likelihood.

In order to estimate the likelihood algorithm performances, we define two standard thresholds, one giving about 97% overall efficiency on $W \rightarrow e\nu$ events (*loose*) and another with 64% efficiency on the same sample (*tight*). These efficiencies have been defined to be the same of the standard cut based electron identification on the same events [7].

The performances are estimated in terms of efficiency on true electrons and rejection of fake candidates. These quantities are shown in terms of η and p_T of the electrons or of the faking jet, respectively.

6 Algorithm Performances

We study the electron efficiency and mis-identification of the likelihood algorithm on the Monte Carlo samples in the Summer08 production, using the PDFs currently available in the software release `CMSSW_2_1_12`, which were done using the `CSA07` Monte Carlo production. The PDFs are slightly changed with respect those samples, mostly in H/E variable, since in the newest release the zero-suppression in HCAL is applied, thus removing the tail at negative values in H/E. This can bring to a sub-optimal performance of the algorithm. More important, this is also a proof of the stability of the algorithm with respect to small changes of the input variables.

6.1 Electron identification efficiency

We evaluate electron identification efficiency on W+jets events, with $W \rightarrow e\nu$ prompt decay. We define the efficiency as the number of reconstructed and identified PixelMatchGsfElectrons with respect the true electrons coming from the prompt decay of the W which are generated in $-2.5 < \eta < 2.5$ and have a $p_T > 10$ GeV/c. A reconstructed electron matches the generated one if the angular distance between the direction of the true one and the reconstructed GSF track extrapolated at vertex $\Delta R <$ is lower than 0.3. The efficiency is estimated as a function of η and p_T of the true electron.

The thresholds having an overall efficiency correspondent to the loose and tight standard cut-based identification are listed in Table 2. The categorization used in the likelihood definition described in Sec. 3.2 (and in [1]) is not the same of the category based cut electron identification described in [7], but the overlap of the populations of the different categories is very large and they can be considered equivalent for our purposes.

Identification	efficiency
Loose: $r > 0.06$	97%
Tight: $r > 0.81$	80%

Table 2: Efficiency of electron identification on reconstructed electrons in a $W(e\nu)$+jets sample for two chosen thresholds on the likelihood ratio r. The thresholds have been chose to reproduce the efficiencies of the standard category cut based identification on $W \rightarrow e\nu$ events.

The efficiency of reconstruction only, loose category cut-based identification, loose and tight likelihood identification as a function of η and p_T of the true electron are shown in Fig. 12.

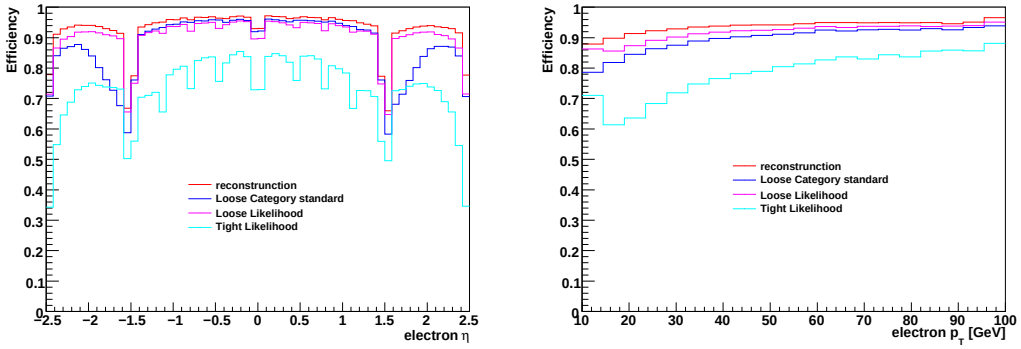


Figure 12: Efficiency of electron reconstruction (red line), loose category-based identification (blue line), loose and tight likelihood-based identification (pink and cyan lines, respectively). The efficiency of identification is cumulative with reconstruction.

It is clear that, when requiring tight identification, the electrons in the ECAL inter-module cracks are suppressed. This can be due to the fact that they are here considered as showering electrons. A more detailed study of the electron identification variables for crack electrons is needed. Outside cracks the overall efficiency is quite uniform both in barrel and endcap. The combined efficiency of electron reconstruction and identification is 85% at $p_T \sim 20$ GeV/$c$and reaches 90% at $p_T = 40$ GeV/c.

6.2 Electron mis-identification

We estimate the electron mis-identification rate as the probability for an object which is reconstructed as a calorimeter deposit with a pointing track to be reconstructed and identified as an electron. In this work calorimetric jets (reconstructed with IterativeCone with $\Delta R = 0.5$) are chosen as these loose objects. Since the samples of QCD jets available in Summer08 production so far are enriched of electro-magnetic component, they are not representative of the general QCD population[6)]. We therefore use a jets sample coming from W+jets events, where we discard the jet matching the generated electron from W.

The mis-identification rate is defined as the ratio

$$f(\text{jet} \rightarrow \text{fake electron}) = \frac{\#\text{jets matching electron } (l)}{\#\text{reconstructed jets } (d)} \ , \tag{4}$$

The numerator (l) is defined as the number of reconstructed jets that match a reconstructed (and eventually, identified) lepton with $p_T > 10$ GeV/c (called l-objects). The denominator (d) is the number of reconstructed jets (called d-objects). The jets are required to lie inside ECAL acceptance ($\mid \eta \mid < 2.5$) and have a transverse momentum $p_T > 10$ GeV/c. The electron is considered matched with the jet if the angular distance between the two is $\Delta R < 0.3$.

We show the electron mis-identification rate as a function of η and p_T (uncorrected) of the jet closest to the fake electron (in terms of ΔR) in Fig. 13. This probability is described as a function of the fakeable object to obtain from data the number of mis-reconstructed electrons in a given analysis. The latter can be evaluated as the product of the number of reconstructed jets times the probability of being identified as an electron, in the specific (η, p_T) region.

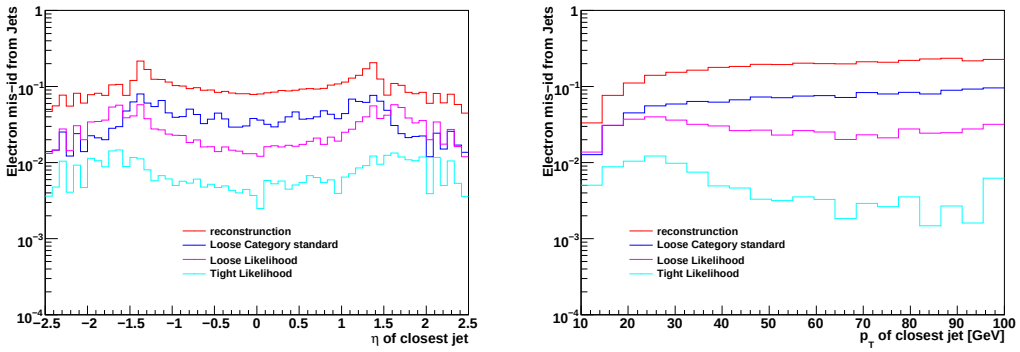


Figure 13: Electron mis-identification probability as a function of η (left) and p_T (right) of the closest jet to the fake electron. The probability is estimated after the reconstruction (red line), loose category-based identification (blue line), loose and tight likelihood-based identification (pink and cyan lines, respectively).

With the same identification efficiency on electrons, the likelihood algorithm provides about half fake identified electrons than the cut-based approach at $p_T \sim 20$ GeV/c, reaching a factor 3 of reduction for $p_T > 50$ GeV/c. Equal or worse fake rejection is obtained in the low p_T region. This can be attributed to the stronger energy dependence of the electron identification variables for low p_T electrons. A better description of the PDFs for low momentum electrons is then needed.

The average probability for a jet with uncorrected $p_T > 20$ GeV/c to be reconstructed and identified as an electron with the loose likelihood criterion is about 3%. This reduces to 0.4% if the tight likelihood identification is applied.

There is a strong dependence of the mis-identification probability on η of the closest jet, being much higher in the transition region between the barrel-endcap ECAL cracks.

A full electron identification should consider also electron isolation that provide a further order of magnitude reduction of fakes from jets. The isolation and identification criteria should be optimized analysis by analysis (ex. [**?**]). The proposed loose and tight thresholds are given only as an example.

[6)] More general samples of QCD jets are being produced with MadGraph matrix-element generator.

7 Miscalibration and Misalignment Effects

In the CSA07 Monte Carlo production, the same processes of W+jets and Z+jets have been reconstructed with ideal detector conditions and with mis-alignment and mis-calibration scenarios correspondent to the detector knowledge after 10 pb^{-1}or 100 $\mathrm{pb}^{-1}$of integrated luminosity. We have worsened the 10 pb^{-1} scenario convolving the electron identification distributions with a further Gaussian smearing. We evaluated the efficiency on prompt electrons from W+jets events using this worst case scenario PDFs and compared with ideal conditions. The efficiency variation is about 0.1% with $p_T > 15$ GeV/c(while is larger for lower p_T, 0.5%), showing the robustness of the algorithm for sufficiently high momentum electrons.

We plan to further study these effects, as well as the effects of the limited knowledge of the tracker material budget, when the Monte Carlo samples of the Summer08 production are re-reconstructed with different scenarios.

8 Conclusions

We described the definition of a likelihood based electron identification, defining the input variables and the way they are combined in the likelihood ratio.

We described a strategy to estimate the PDFs of the electron identification variables both for electron particle hypotesis and jet hypotesis on data control samples. These distributions can be extracted on the first hundreds pb^{-1}, considering the statistical error only.

We estimated the performances of the proposed method in terms of efficiency on prompt electrons from $W \rightarrow e\nu$ decays and of the mis-identification probability of a jet as an electron. The performances are given as a function of the pseudo-rapidity and the transverse momentum.

A loose and tight selection on the likelihood ratio have been proposed, only as examples, with efficiencies of about 97% and 80%, respectively, on $W \rightarrow e\nu$ electrons. The correspondent mis-identification probability, including electron reconstruction and identification for a jet is about 3% and 0.4% in average for electrons with $p_T > 15$ GeV/c, for the loose and tight identifications respectively.

We checked the robustness of the algorithm with respect the mis-calibration and mis-alignment scenarios with the CSA07 Monte Carlo production. Further studies has to be performed on the effect of limited knowledge of the CMS detector with the first recorded data when the Summer08 Monte Carlo samples considered in this analysis will be re-reconstructed using different conditions with respect the ideal one.

9 Acknowledgments

We wish to thank our 2007 Summer Student Ozana Celan for the discussions on the background control sample for electron identification variables. We also thank the VecBos group for the useful discussions and for the sharing of the production effort.

References

[1] CMS NOTE-2006/040, “Electron reconstruction in CMS”, S. Baffioni *et al.*.

[2] R. Ranieri, **CMS CR-2008/007**

[3] E. Di Marco *et al.*, **CMS Analysis Note 2008/XXX**.

[4] C. Charlot *et al.*, **CMS Analysis Note 2008/039**.

[5] G. Daskalakis *et al.*, **CMS Analysis Note 2007/019**

[6] C. Amsler *et al.* [Particle Data Group], Phys. Lett. B **667**, 1 (2008).

[7] J. Branson *et al.* **CMS Analysis Note 2008/082**

[8] M. Pivk and F. R. Le Diberder, Nucl. Instrum. Meth. A **555** (2005) 356 [arXiv:physics/0402083].