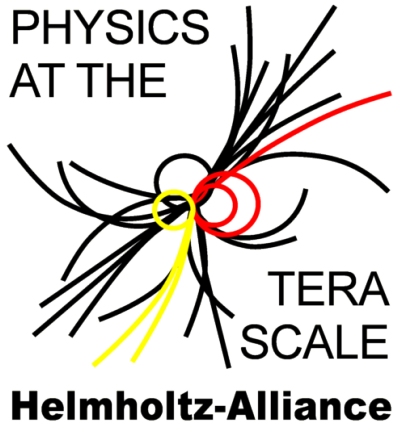
Energy Weighting for CMS-HCAL Upgrade

Matthias Stein

YIG Meeting

17th May 2010



→ Methods to estimate
The Electron fake

Vladimir Andreev, Kerstin Borras, Dirk Krücker, Isabell Melzer-Pellmann, Peter Schleper

Overview

(1) CiC = **C**uts **i**n **C**ategories = Cut based electron Identification

(2) kNN= **k**-**N**earest-**N**eighbour Technique

(3) Conversion Removal for Electron Fakes and Charge Misidentification

(4) Estimation of the fake electron background using Data-Driven techniques

CiC (Cuts in Categories)

- Set of Cuts, **optimized** to select electrons from W/Z-> ee and reject fakes from jets or conversions
- Split candidates into **categories**
- “We have found that many of the features of the electron ID problem in CMS can be dealt with by dividing the problem into categories”
- deals with the large amount of **radiation in tracker** material and the significant probability that the track will not be well measured

→ The cleaner the sample, the higher the cost in efficiency

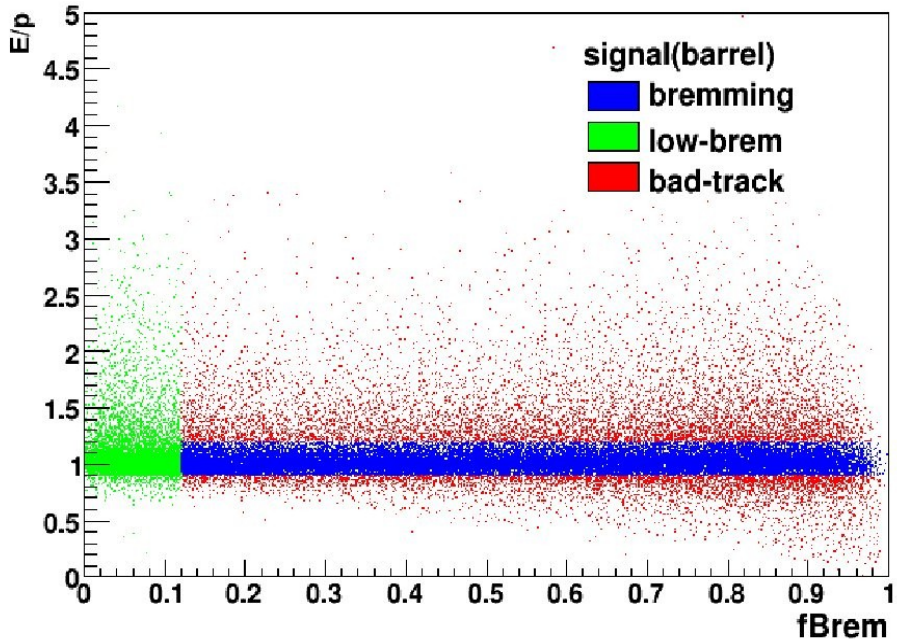
https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideCategoryBasedElectronID

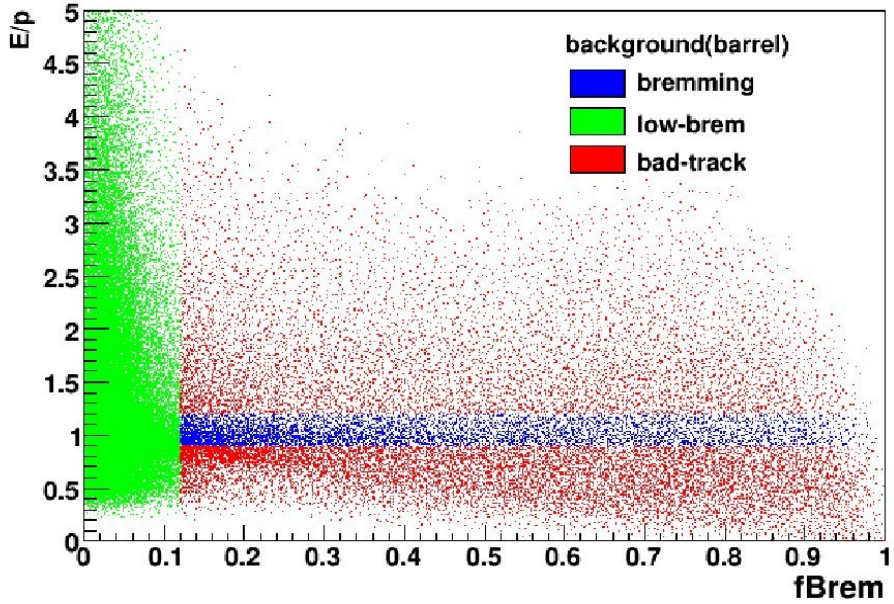
CiC (Cuts in Categories)

(1) Separate between endcap and barrel

(2) fbrem vs. E/p categorization:
- Low-fBrem (green)
- Bremming (blue)
- Bad-Track (red)

(3) Separate into 3 different ET-Regions due to the very high magnetic field:
- 12 < ET < 20
- 20 < ET < 30
- 30 < ET





(1) Cuts

The selection is performed with cuts on the following variables (one cut-value for each category):

- **Track match with ECAL:** delta_phi_in, delta_eta_in, e_seed/p_in
- **HCAL energy directly behind ECAL cluster:** H/E
- **Cluster Shape:** sigma_ieta_ieta

} Pure ID cuts

- **Conversion Rejection:** number of missing hits near beginning of track (also rejects really bad tracks)
- **Track vertex:** Impact Parameter w.r.t. reco vertex

- **Isolation:** Tracker Isolation (0.3), ECAL Isolation (jurasic 0.4), HCAL Isolation (0.4)

(2) Def. Of Brem-regions

- **Low-Brem (green):**
  - Barrel: 0.9 < E/pin < 1.2; fbrem < 0.12
  - Endcap: 0.82 < E/pin < 1.22 ; fbrem < 0.2
  - fake-like region with high population from both real and fake electrons
- **Bremming (blue):**
  - Barrel: 0.9 < E/pin < 1.2; fbrem > 0.12
  - Endcap: 0.82 < E/pin < 1.22; fbrem > 0.2
  - electrons-like region with little contamination from fakes
- **Bad-Track (red):**
  - remaining regions
  - region with not many real electrons, but too many just to cut out
- separates electrons with quite different measurement characteristics and with **very different S/B** (signal to background ratios)

9 defined severity levels of cuts

- For ET > 20 GeV (because efficiency too low for ET < 20)
- E.g.: Loose cuts might be used for di-electrons from Z, while HyperTight cuts might be appropriate for selecting single electrons without much help from MET
- **Each step decreases the fake rate by about a factor of 2**

VeryLoose
Loose
Medium
Tight
SuperTight
HyperTight1
HyperTight2
HyperTight3
HyperTight4

Definitions

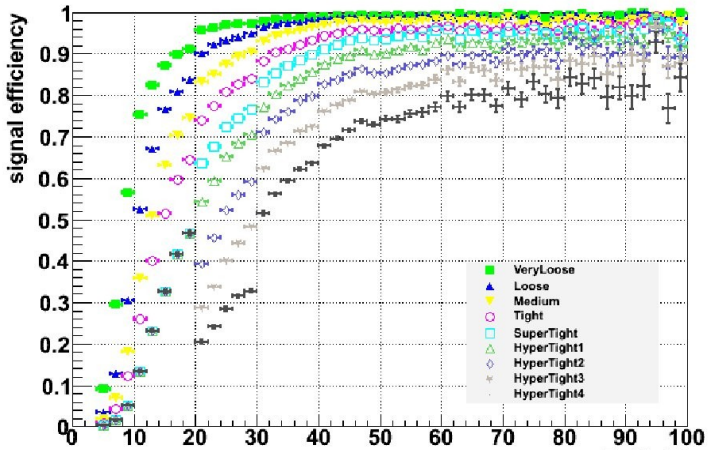
$$Fake\ Rate = \frac{signal\ leptons}{looser - defined\ leptons}$$

$$Efficiency = \frac{\dots}{\dots}$$

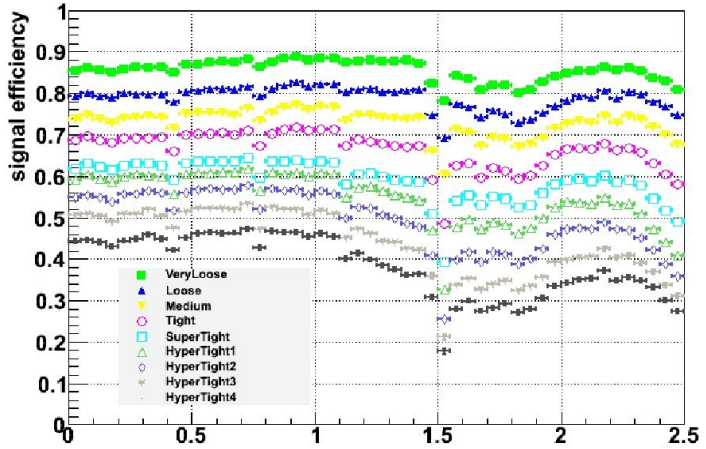
$$Purity = \frac{\dots}{\dots}$$

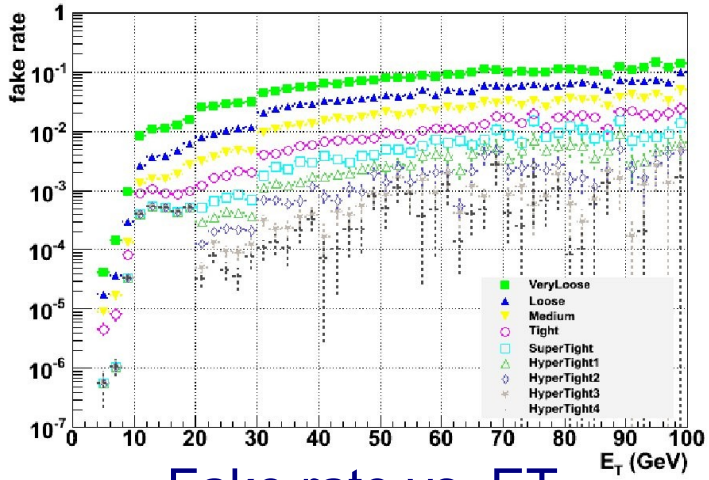
Electron Efficiency / Fake Rate

Eff. vs. ET

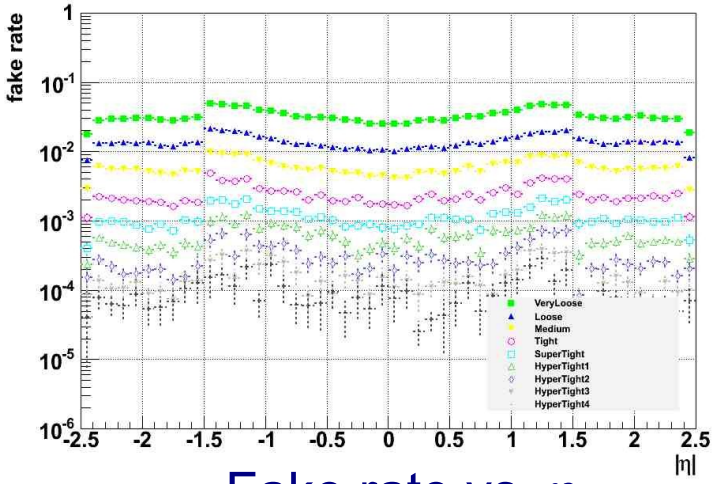


Eff. vs. η





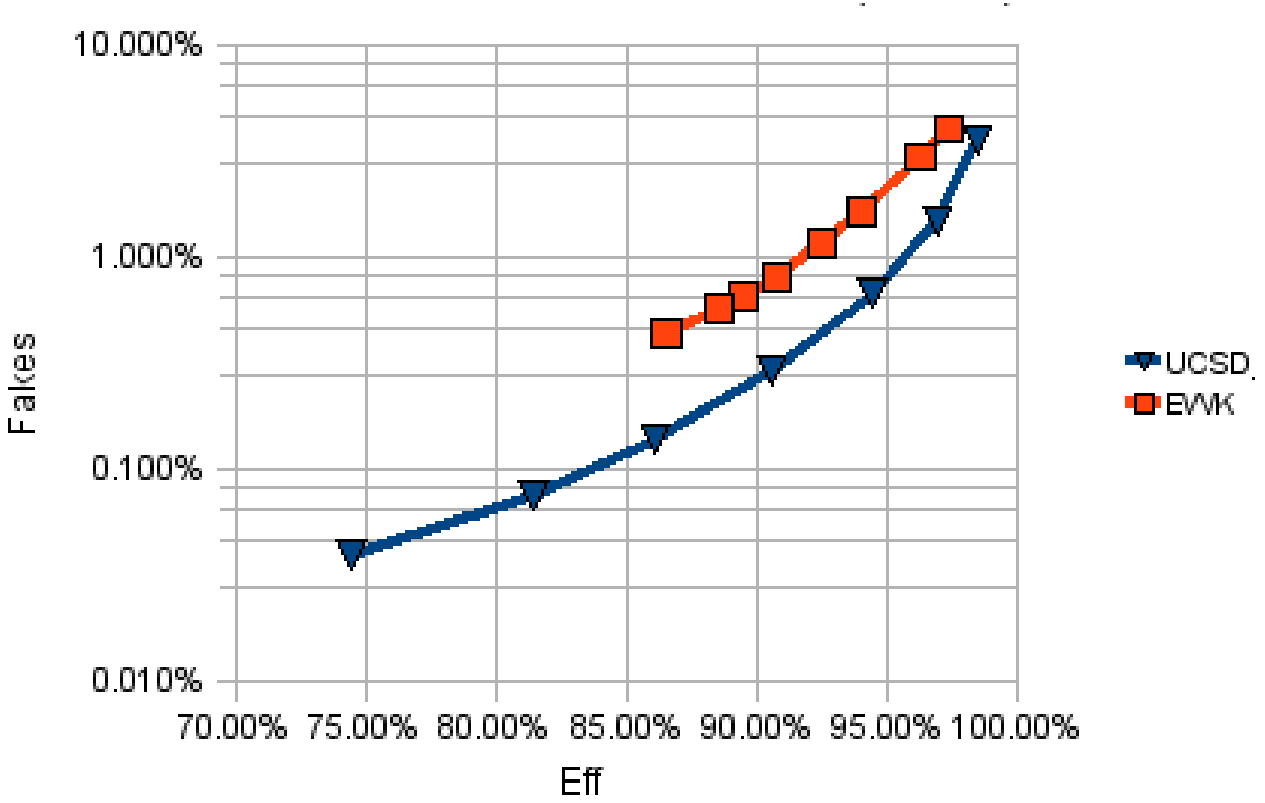
Fake rate vs. ET



Fake rate vs. η

- Some **small discontinuities** at boundaries of different ET bins (12, 20, 30)
- AT low ET efficiency drops for the tighter cuts (worse S/B)

Fake Rate vs. Efficiency



- CiC is more complicated
- Gives a large performance improvement
- This electron ID should be generally useful for electrons from W, Z, and top, in an ET range between 12 and perhaps 500 GeV.

k-Nearest Neighbor Technique

CMS AN AN-09-131

- Access fake backgrounds in **early measurements**
- **General Technique** which can easily be extended to other fake background and analysis

- Fakes result from processes that are difficult to model

- Estimate fake backgrounds by calculating empirical fake probabilities for electron objects that pass the **robust** ID cuts

- Predict:
  - real/fake electron composition of samples
  - Kinematic distributions of samples

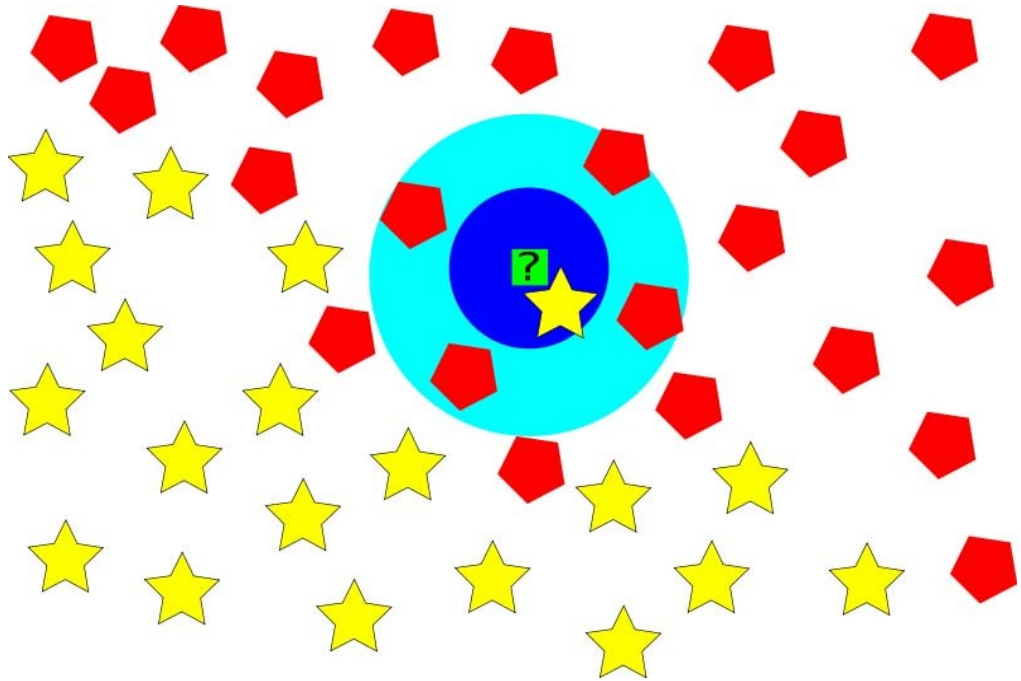
- Method is **intuitive**, easy to implement and should be relatively **insensitive** to composition differences between control and data samples

- Principle: “like things look alike”'
- Multivariate technique that is typically used for the purpose of classification

kNN

- k-nearest neighbors algorithm (k-NN) is a method for **classifying objects** based on closest training examples in the feature space (Merkmal-Raum)
- An object is classified **by a majority vote** of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors

Example:



k=1: light blue
k=5: deep blue

- Larger k reduce the effect of **noise** but make boundaries between classes **less distinct**
- The special case where the class is predicted to be the class of the closest training sample (i.e. when **k = 1**) is called the nearest neighbor algorithm

kNN

- One does not need prior knowledge of the underlying probability distribution functions
- It can be useful to **weight** the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones
- A common weighting scheme is to give each neighbor a weight of **1/d**, where d is the distance to the neighbor
- Distance via metric, e.g. **Euclidean metric**
- The training examples are vectors in a **multidimensional feature space**
- The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples

kNN – application to particle physics

X = m-dimensional feature vector
ω_i = class of object (true "nature of state")
Θ_i = label of object
i = fake, real
$d_{i,j}$ = m-dimensional distances
$A_{m,n}$ = metric (simplest choice: Euclidean)

- X[m] refers to measured values of individual ID variables
- Assign labels to objects $\vec{X} \rightarrow \Theta_i$ such that $\Theta_i = \omega_i$

- Mapping by comparing an object's feature vector X_{test}, to those objects in the training sample, $X_{training}$, for which ω_i are known

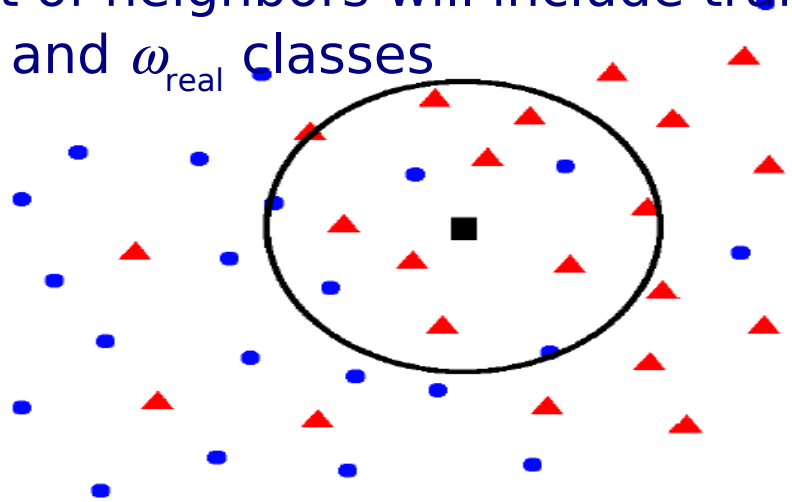
$$d_{i,j}^2 = \sum_{m,n} A_{m,n} \cdot \left(X_{i \in training}[m] - X_{j \in test}[n]\right)^2$$

$$A_{m,n} = \begin{cases} 1, & \text{if } m = n \\ 0, & \text{otherwise} \end{cases}$$

- $d_{i,j}$ is sensitive to differences in the scales of the individual variables considered in the calculation
- Normalize with a factor ΔX_{mn}

$$d_{i,j}^2 = \sum_{m,n} A_{m,n} \cdot \frac{(X_{i \in training}[m] - X_{j \in test}[n])^2}{\Delta X_{mn}^2}$$

- First calculate all $d_{i,j}$
- Sort training/ test pairs by decreasing distance
- Select the k training objects with the smallest $d_{i,j}$ → These are the k nearest neighbors for the object represented by X_{test}
- The selected set of neighbors will include training objects from both of the ω_{fake} and ω_{real} classes



- For large training samples, the conditional probability that the test object belongs to class ω_i is estimated from the proportion of objects of class ω_i found amongst its k nearest neighbors

$$P(\omega_i|\vec{X}_{test}) = \sum_{n_i=0}^{k_i} w_i/k$$

$$w_i = 1/d^2_{i,test}$$

- The sum is over the k_i objects of class ω_i included among the k nearest neighbors of the test object
- This factor allows closer neighbors to contribute with greater **weight** to the sum

$$P(\omega_{fake}|\vec{X}) = \sum_{n_i=0}^{k_{fake}} w_{fake}/k$$

$$P(\omega_{real}|\vec{X}) = 1 - P(\omega_{fake}|\vec{X})$$

kNN

- $P(\omega_i|X)$ are **empirical posterior probabilities** of class membership
- These estimates are proven to **converge to the true posteriors** in the limit of infinite training sample sizes
- In practice, one works with finite training samples of **equal size** to prevent biasing the probability estimates toward one or the other class
- Too small value can lead to estimates that are overwhelmed by **noise** while too large a value can result in a **loss of sensitivity** to local features in the data

Data-Driven techniques

CMS AN AN-10-043

- **Fake Rate (FR) Method: Assigns a probability** (or fake rate) that each jet will fake an electron depending on different properties of the jet
- determine the probability that an event will contain a fake electron

- **Isolation Extrapolation Method: Extrapolates** from a region dominated by fake electrons (poorly isolated electrons) to a region dominated by electrons

- Two sources of background:
  - A jet faking an electron
  - A heavy quark decaying to an electron
- Identify and estimate major backgrounds by separating into two scenarios:
  - Events with exactly one electron
  - Events with more electrons
- Different backgrounds dominate these two final states
- Events cuts differ, electron/ jet selection remain the same

Data-Driven techniques - Good objects

Electron:

- RECO GSF Electron, $p_T > 20$ GeV
- $|\eta| < 2.5$, and η not in ECAL gap ($1.47 < |\eta| < 1.567$)
- "robustLoose" identification [4]
- $\text{RelIso} = \frac{1}{p_T^{\text{ele}}} \sum_{\Delta R<0.4}^{\text{iso dep}} \left(E_T^{ECAL} + E_T^{HCAL} \right) < 0.1$ [2]

Jet:

- L2 (relative) and L3 (absolute) corrected RECO sisCone5 Jet [5]
- $p_T > 40$ GeV
- $|\eta| < 3.0$
- Electromagnetic Fraction (EMF) < 0.9
- Jets within $\Delta R < 0.3$ of an identified electron are not counted

Data-Driven techniques – Event selection

Our event selection for the single electron final state:

- **Trigger**: Passes single electron trigger HLT_Ele15_LW_L1R.
- **Electron**: Exactly one good electron.
- **Jets**: Four or more good jets.
- $\not{E}_T$: Uncorrected $\not{E}_T > 150$ GeV.

Our event selection for the two electron final state:

- **Trigger**: Passes single electron trigger HLT_Ele15_LW_L1R.
- **Electron**: Exactly two? good electrons.
- **Jets**: Two or more good jets.
- $\not{E}_T$: Uncorrected $\not{E}_T > 150$ GeV.

Data-Driven techniques – Def. Fake Rate

AO = Analysis objects
FO = Fakeable objects: jets with loose cuts
F = Fake rate = assigns probability that a FO is reconstructed as a good electron

$$F(p_T, \eta, ...) = \frac{AO(p_T, \eta, ...)}{FO(p_T, \eta, ...)}$$

$$\mathcal{F}(\mathrm{p_T}, \eta, \ldots) = \frac{\mathrm{jets}(\mathrm{p_T}, \eta, \ldots) \cdot \mathrm{matched\ to\ good\ electron}\ (\Delta R < 0.4)}{\mathrm{jets}(\mathrm{p_T}, \eta, \ldots)}$$

- Any inefficiency in matching would result in an underestimation of the number of fakes
- Using jets as fakeable objects has the advantage of **increased statistics** and reduced contamination from electrons, but it suffers from **increased systematics** introduced by varying jet properties that may not be correctly captured in the fake rate (e.g. different fake rates for quark and gluon jets)

Data-Driven techniques

- To measure the fake rate, one needs a **fairly pure sample** of fake electrons (look at jet triggered events; prescale doesn't matter)
- [**Trigger bias** may be introduced by including the leading jet (that fires the trigger) in the the fake rate calculation]
- There is a non-negligible **contamination** of the control sample due to electrons from W and Z decays → Cleaning cuts:
  - Remove events with ≥ 2 good electron to reduce contamination from Z events
  - Remove events with 1 good electron that have $65\ \text{GeV} < M_T < 80\ \text{GeV}$ to reduce contamination from *W* events

- Like to parameterize the fake rate as a **function of any jet variable** that might be important (e.g. pT, *h*, EMF, jet charge, etc.)
- But: The more **parameters** (or bins) that are used to separate the fake rate, the more one limits the statistics that enter into the fake rate calculation
  → focus on p_T and η.
- Sd
-

Data-Driven techniques

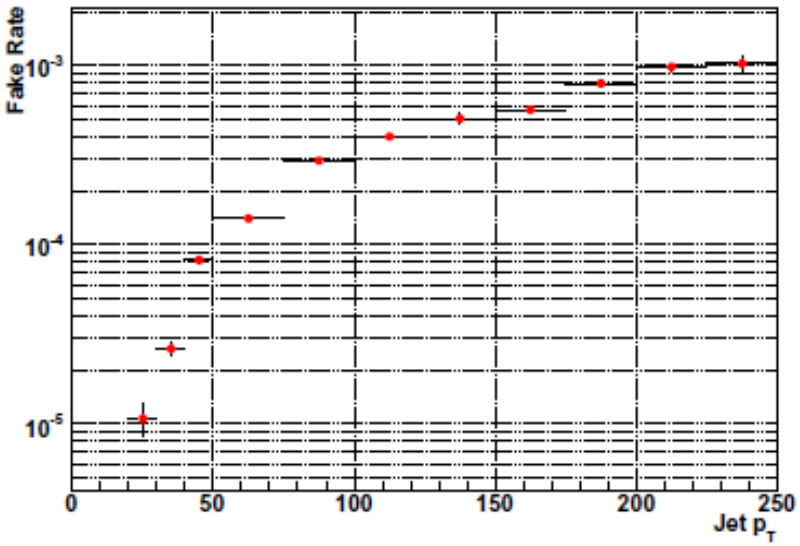


Figure 1: Fake rate as a function of the jet p_T when using all our MC samples combined with cleaning cuts. Binning was chosen to have reasonable statistics in each bin.

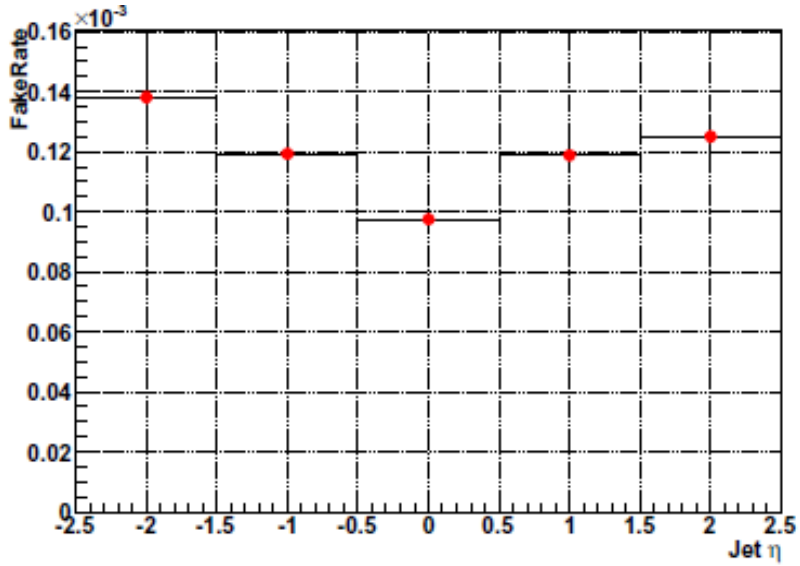


Figure 2: Fake rate as a function of the jet η when using all our MC samples combined with cleaning cuts. The plot illustrates that it is reasonable to use the $\pm\eta$ symmetry of the fake rate to increase statistics.

- Thus far, the fake rate was defined as the **probability** to find a fake electron per jet. The probability that an event (labelled as a) has one fake electron is given by

$$P_a^{1\ \text{fake}} = \sum_{j=1}^{N_{\text{jets}}} \mathcal{F}\left(p_T^j, \eta^j\right) \prod_{i \neq j}^{N_{\text{jets}}} \left(1 - \mathcal{F}\left(p_T^i, \eta^i\right)\right) \tag{4}$$

where the sum (and product) is defined over all jets (our FOs) in that event. Recall that we only run over events without any fake electrons. Therefore, we need to normalize the probability of (Eq. 4) by the probability that each event has no fakes

$$P_a^{0\ \text{fake}} = \prod_{j=1}^{N_{\text{jets}}} \left(1 - \mathcal{F}\left(p_T^j, \eta^j\right)\right) \tag{5}$$

From (Eq. 4) and (Eq. 5), we can construct an event weight

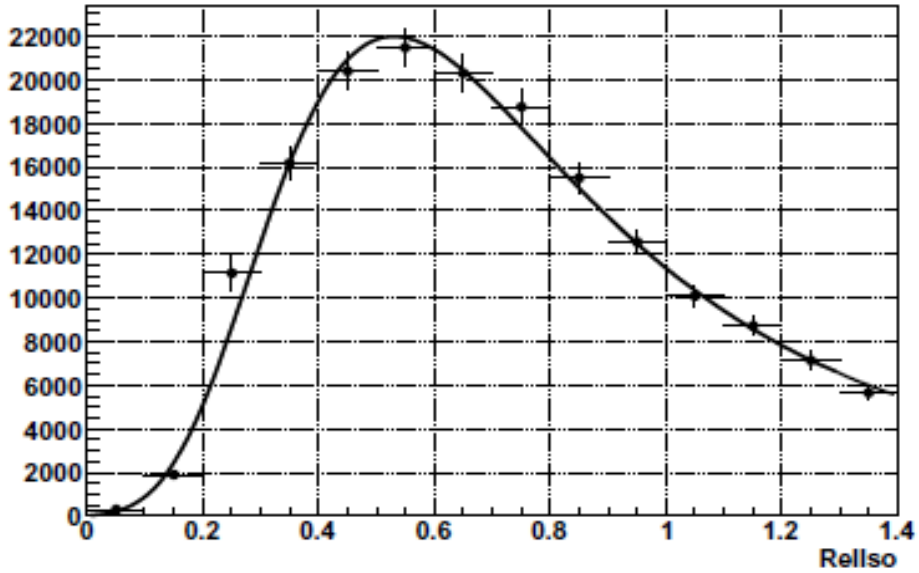
$$\mathcal{W}_a^{1\ \text{fake}} = \frac{P_a^{1\ \text{fake}}}{P_a^{0\ \text{fake}}} \tag{6}$$

which can be summed up to give an estimate of the number of events with 1 fake:

$$N_{\text{events}}^{1\ \text{fake}} = \sum_{a=1}^{N_{\text{events}}^{0\ \text{fake}}} \mathcal{W}_a^{1\ \text{fake}} \,. \tag{7}$$

Data-Driven Techniques – Isolation Extrapol.

- Works by extrapolating from a region dominated by the background to the signal region
- Use **relative calorimeter isolation** (RelIso) as a discriminator between electrons and fakes
- By fitting the relative isolation distribution in a control region, (which we assume is dominated by QCD (fakes)), we can **get the full shape** of this distribution and use it to predict the contribution from the QCD background in the signal region
- Expect RelIso to be the **combination of two shapes** from two different sources: electrons and fakes
- Electrons are expected to peak at low values of RelIso
- Focus on QCD events and require cuts
- Best described by a **Landau distribution**



Data-Driven Techniques – Isolation Extrapol.

- Parametrize Landau distribution by: Most probable value, width, normalization
- After fitting the RelIso shape to a Landau, **extrapolate** this fit to the signal region (**RelIso<0.1**)
- Integrating our fit function in the signal region gives us the **prediction** on the number of events with 1 fake electron
- The error on the prediction was found by propagating the error on each of the fit parameters to the error on the fit function in the signal region

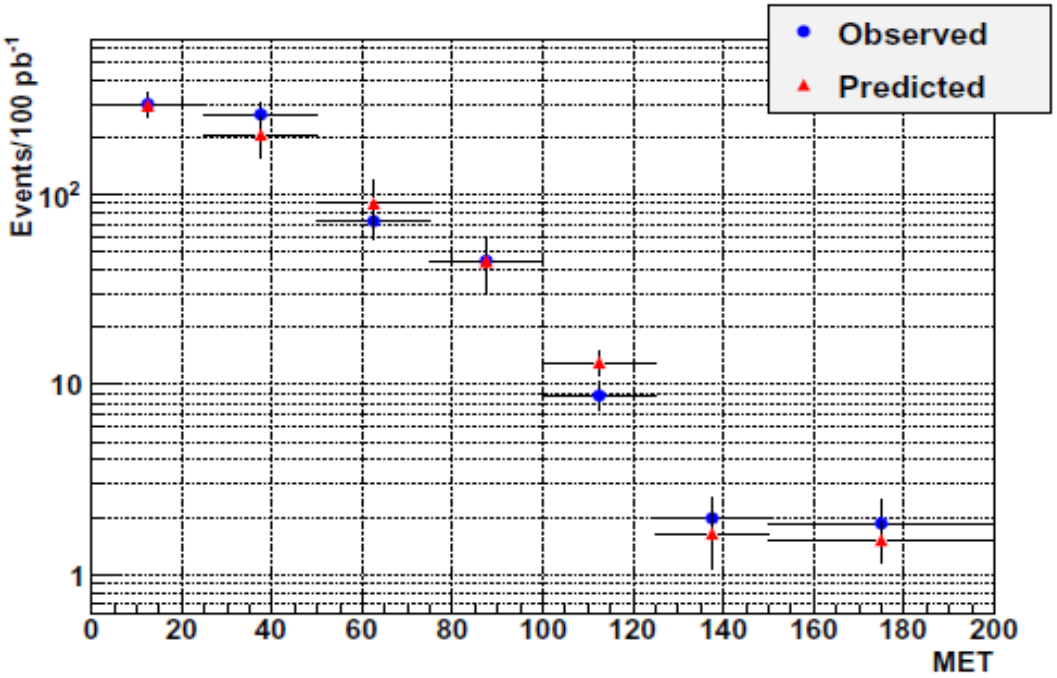


Figure 8: Comparing observed (from summing the histogram) and predicted (from the fit) number of events with 1 fake electron and at least 4 good jets, as a function of $\not\!\!E_T$ for the MadGraph QCD sample. Overall, the method performs well.

Data-Driven Techniques – Isolation Extrapol.

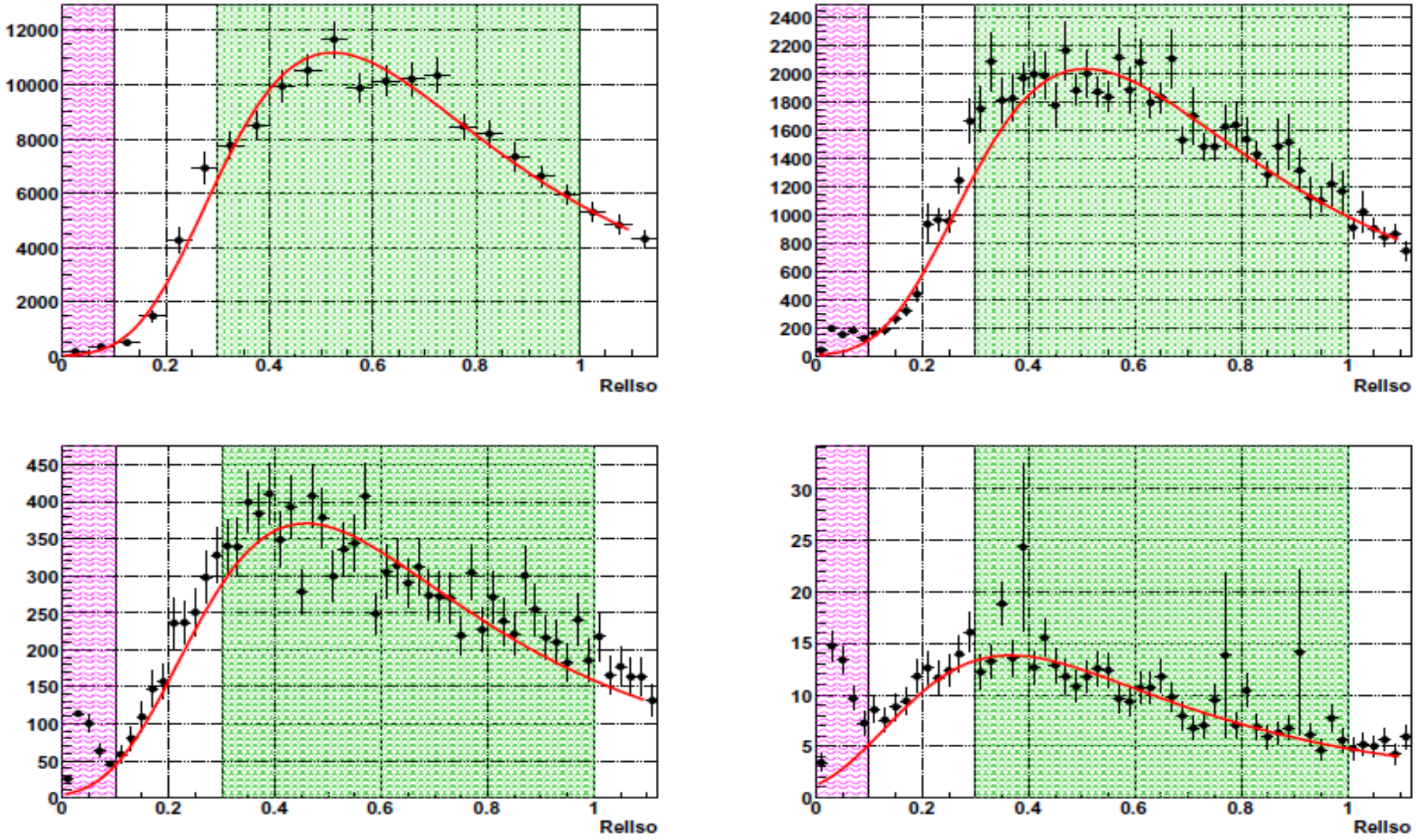


Figure 9: RelIso distribution for all of our MC samples combined after requiring that the electron trigger fired, $N_{jets} \geq 4$, at most one electron passing our full selecion, and a $\not\!\!E_T$ cut of $0 < \not\!\!E_T < 25$ GeV (top left), $25 < \not\!\!E_T < 50$ GeV (top right), $75 < \not\!\!E_T < 100$ GeV (bottom left), and $100 < \not\!\!E_T < 150$ GeV (bottom right). We chose the signal region (purple wavy line region) to be $0 <$ RelIso < 0.1 and the fit region (green cross-hatched region) to be $0.3 <$ RelIso < 1.0.

Conversion Removal

- Photon conversion removal using a vertex fit with a 3D conversion constraint
  - Conversion from:
    - $\pi^0 \rightarrow \gamma\gamma$
    - Prompt photons
- Can also be used to improve charge electron identification
- Steps:
  - Select collection of tracks
  - Apply vertex fit to all pairs of tracks passing a preselection
  - Make final selection of conversion candidates
  - Veto any matching electron candidates

CMS AN AN-09-173

- Standard CMS track reconstruction using a Combinatorial Kalman Fitter (CKF) (which is mathematically equivalent to a global least-squares minimazation)
- This “generalTracks” collection is the most inclusive available from the reconstruction, in particular for lower p_T tracks which do not reach the ECal
- Electron candidates use tracks reconstructed with a Gaussian Sum Filter (GSF) fit, which employs a weighted sum of Gaussian components to model the non-Gaussian energy loss distribution for electrons passing through the material
- Merging genaralTracks, two Ecal seeded Track collection and GSF Electrons results in the conversion candidates

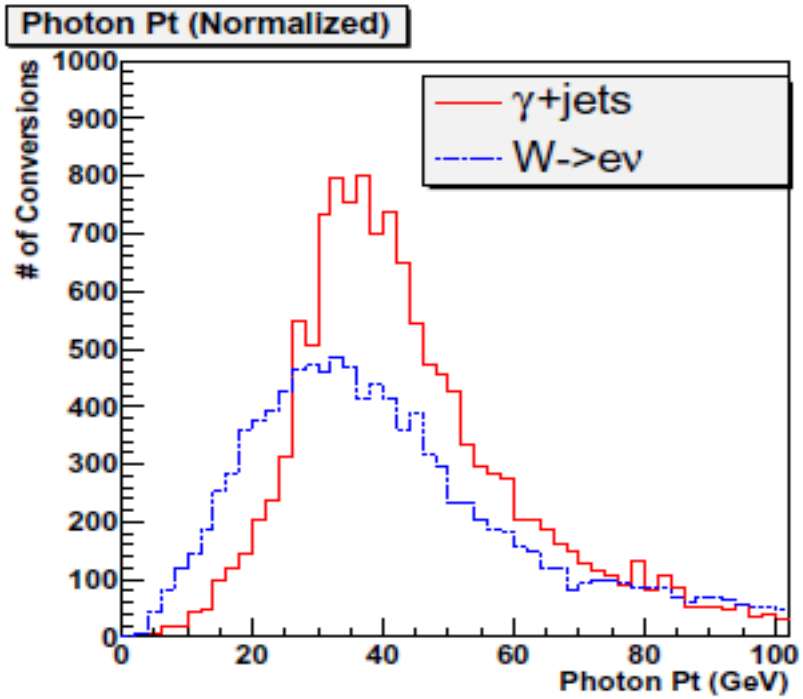
Conversion Removal

- Two (standard-)cuts in electron selection/ identification to highlight:
  - $d_0 < 0.025$ cm (transverse impact parameter cut on the track)
    → already removes many fakes from conversion
  - Charge identification requirement (charge (GSF electron) = charge(matching CKF track)
    → reduces charge misidentification before conversion veto

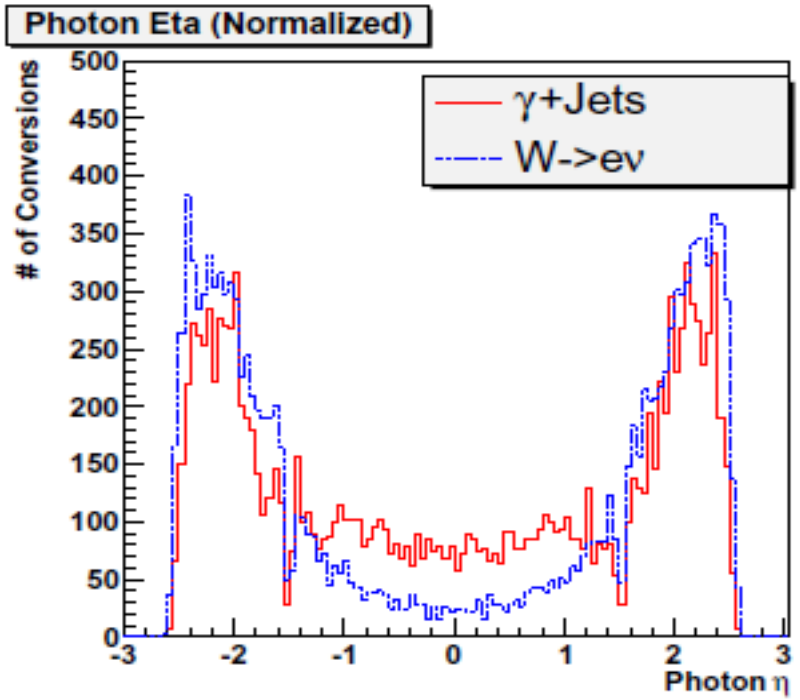
- Preselection to all opposite-sign pairs of tracks:
  - r > 0.9 cm (r = point of closed approach)
- Now: CTVMFT fitter applied for vertex fit
  → 3D fit: tracks are parallel at the vertex for both r-ϕ– and r-z-planes
  - Each fit which converges is stored as a conversion candidate
- Because track seeding for electron reconstruction requires hits in the pixel detector or TID, photons which convert early in the detector (in the pixels or inner-most layers of the tracker) are much more likely to fake electrons
- The remaining cases where real electrons are reconstructed as conversions are mainly those where a bremsstrahlung photon converts early in the tracker or pixel detector, and one of the conversion legs is incorrectly paired with the prompt electron track to make a conversion candidate
-

Conversion Removal

- Remove all electron candidates where the GSF track corresponds to one of the selected conversion candidates



(a) Conversion p_T



(b) Conversion η

Figure 5: Kinematics of reconstructed conversion candidates, normalized to area.

Selection	γ+jets	$W \rightarrow e\nu$ (all)	$W \rightarrow e\nu$ (Wrong charge)
Fake Removal	37%	1.0%	29%
Charge ID	49%	8.3%	70%

Conversion Removal

- The fraction of charge-misidentified electrons removed by the veto is 29%, reducing the charge misidentification rate from 0.53% to 0.38%. This correlation arises because a bremsstrahlung photon conversion close to the initial electron can cause the track reconstruction to include hits from one of the conversion legs, leading to a wrong charge assignment.
- We can minimize the electron charge misidentification rate by explicitly attempting to remove events with conversions very close to the initial electron where there is a higher probability of confusion in the track reconstruction
- In this case we remove 8.3% of the W electrons, but 70% of the wrong charge cases, reducing the charge misidentification rate from 0.53% to 0.18%
- We also manage to remove 49% of the conversion fakes from the +jets sample