EM Objects Identification in DØ Using Multivariate Techniques

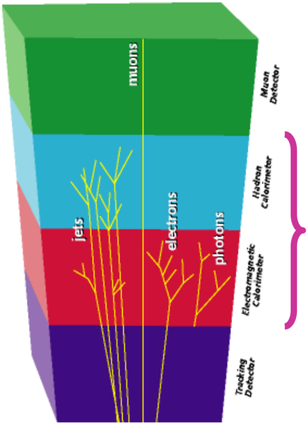
Status and plans

E.Chapon

CEA Saclay / Irfu / SPP

DØ France, May 30-31, 2011

- Object identification aims at tagging an object of being of a certain nature.
- We want to discriminate electrons and photons against jets.
  - Discriminating variables split up into two main categories: shower shape variables (such as the electromagnetic fraction) and track variables (such as the probability of the track match).

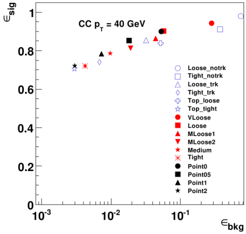


EMID group

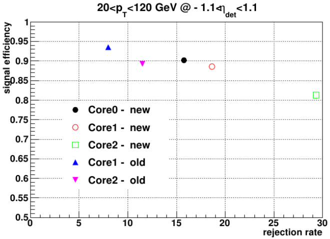
Electron and photon identification is developed inside the EMID group (Konstantinos PETRIDIS, Xuebing BU), which a sub-group of the CALGO group (Leo BELLANTONI, Christophe ROYON).

State of the art

- Current EM object identification is cut-based.
- Four definitions are available for electrons (Point0, Point05, Point1, Point2) and three for photons (Core0, Core1, Core2).
  - These definitions are divided into CC and EC.
- The certification for these is available for RunIIb4 included (`emid_eff v09-00-08`).



Signal versus background efficiency for electron definitions in CC (black).



Signal versus background efficiency for photon definitions in CC.

Motivation for multivariate identification

- Current identification is based on rectangular cuts.
- To improve "real / fake" discrimination, the obvious next step is to use a multivariate technique.
  - This is somehow already used: current definitions cut on some multivariate variables, such as the electron likelihood (lhood8) or some artificial neural networks (ANN).

Goal

Cut on the output of a single discriminant, instead of a set of EM variables.

- More freedom to the user.
- Better discrimination (take advantage of correlations, ...).
- Tim Head has been working on multivariate electron identification for about one year and half.
- I have been working on multivariate photon identification for about 6 months.

Basic idea

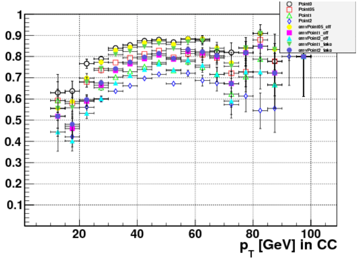
Use a BDT (called EMV), trained on data only, to identify real electrons against fakes (mainly jets). It uses standard EM variables, as well as shower shape and track related variables.

- Use the tag and probe method to select a pure sample of electrons in data (RunIIb1-3 2EMhighpt):
  - require one "tag" electron with tight quality requirements,
  - the "probe" electron has very loose quality requirements (but must be track matched and with $p_T > 15\,\text{GeV}$),
  - the invariant mass of the two electrons must be close to M_Z.
- The fake-enriched sample (from RunIIb1-3 EMinclusive) is selected in di-jet events:
  - require one jet with $p_T > 15\,\text{GeV}$,
  - the fake electron is an EM cluster back-to-back to this jet with the same quality cuts as used for the "probe" in the real electron sample,
  - the "invariant mass" of the EM cluster and any track in the event should be $< 60\,\text{GeV}$ and MET $< 10\,\text{GeV}$.
- The training is split into four regions: CC and EC, $\mathcal{L} < 180.10^{30}$ and $\mathcal{L} \geq 180.10^{30}\,\text{cm}^{-2}$.

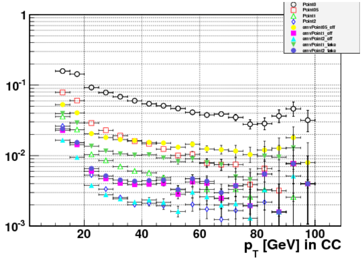
Efficiencies

Five working points (i.e. recommended cuts on the EMV BDT) have been defined.

- emvPoint05_eff, emvPoint1_eff and emvPoint2_eff match the signal (electron) efficiency of the corresponding definitions.
- emvPoint1_fake and emvPoint2_fake match the corresponding fake rates.

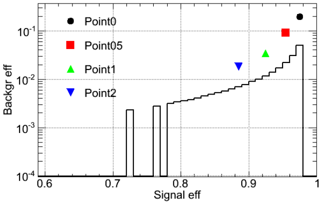


Efficiency on data probe electrons in the CC

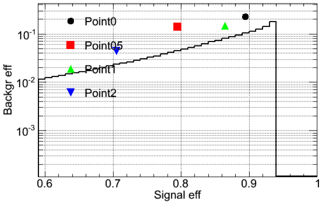


Efficiency on data fakes in the CC

Signal efficiency vs. background efficiency



Signal vs. background efficiency in the CC



Signal vs. background efficiency in the EC

Find more plots in Kostas' talk (May 13rd 2011, Conveners Meeting)

Status

Certification

Scale factors are available for RunIIb4 and the 5 working points.

The code for multivariate identification is in CVS:

- `tmb_tree p21-br-93`
- `emid_cuts p21-br-26`
- `emid-eff v09-00-08`
- `caf_util p21-br-150`
- See `https://plone4.fnal.gov/P1/D0Wiki/object-id/emid/emdev/EMV/Setup_v1.0` for instructions on how to use it.

Physics analysis

Multivariate electron identification can be used now in any analysis using electrons (example: $H \to WW$).

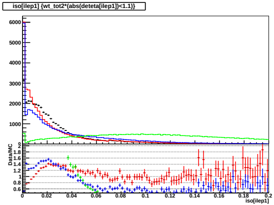
Specificities

There is one major difference between electron and photons: we cannot have a large pure sample of data photons, because we cannot use the tag and probe method to select them like for electrons.

Drawback

We are obliged to train our BDT on MC. This is a major difficulty, because many (most?) variables of interest for the photon identification are not very well modeled in our MC.

- This propagates to unreasonable scale factors if we do not pay a careful attention to the choice of variables.



Calorimeter isolation in the CC (electron data in black, electron MC in red)

Input

Two BDTs have been developed: one is trained against jets ("jet/gamma BDT"), the other one against electrons ("e/gamma BDT").

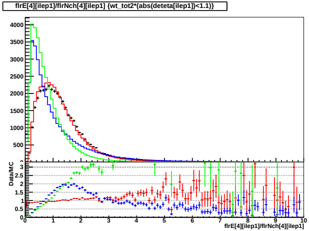
e/gamma BDT

The e/gamma BDT is not yet finalized. For now we focus on the jet/gamma BDT.

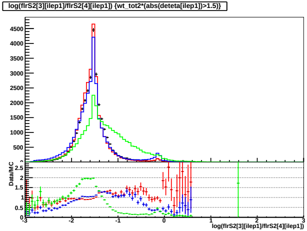
- I will only speak about the jet/gamma BDT in the remaining slides.

- The signal sample is QCD di-photon RunIIb1 MC, the background is EM-like jet RunIIb1 MC.
- No track variables as input of the BDT: we want to use electrons to compute scale factors.
- Optimizing the list of variables for performance only turned out to be a bad idea. Current variables are selected on a strict requirement on the data/MC agreement (on electrons).

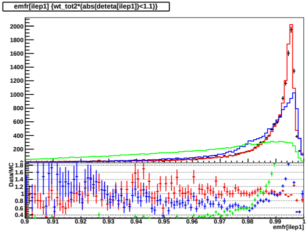
Some input variables



Average cell energy in the EM4 layer (CC).



log (flrS2[3]/flrS2[4]) (EC)

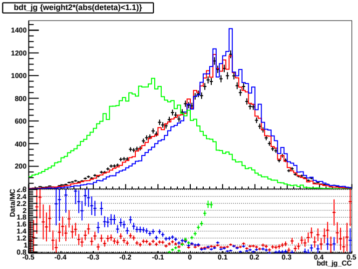


EM fraction (CC)

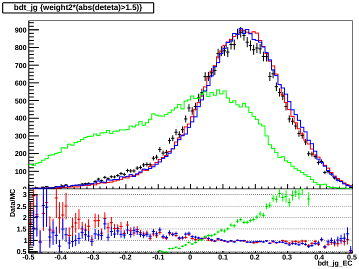
Legend:

- Electron data
- Electron MC
- Photon MC
- Jet MC

Output distribution



BDT output in the CC



BDT output in the EC

For now we want to cut on the BDT together with some loose cuts on standard EM variables (such as the EM fraction or the calorimeter isolation). Therefore the BDT is trained and evaluated on already signal-like candidates, which is why the apparent discriminant power looks poor.

Conclusion

- MV electron ID is finalized and certified.
    - It is ready for use in physics analyses.
    - For instance we are trying it in $H \rightarrow WW \rightarrow e\nu e\nu$!
- MV photon ID is under finalization.
    - It is foreseen for Lepton Photon 2011 and winter conferences.