Application of the Naive Bayes Classifier technique to the Fermi LAT Calorimeter data

M. Pesce-Rollins[1], L. Baldini[1], L. Latronico[1], and M. Tinivella[1]
on behalf of the Fermi-LAT Collaboration
[1]Istituto Nazionale di Fisica Nucleare, sezione di Pisa

A clustering stage in the calorimeter, followed by a moments analysis of the resulting clusters, is currently the first pass of the new event reconstruction being developed for the Fermi Large Area Telescope. One of the primary goals is to identify possible signals due to out-of-time, accidental coincidences in the detector and separate them from the genuine gamma-ray event. To this aim, we developed a multivariate classification technique based on a Naive Bayes Classifier (NBC) algorithm, to exploit the full topological information made available by the clustering and the moments analysis stages. The probability density functions (PDF) for the discriminating variables are stored in the form of one-dimensional histograms with approximately equally populated bins. The energy dependence of the event topology is accounted for by training different classifiers in suitable energy bins, with a weighting scheme being investigated in order to avoid possible artifacts at the bin edges. Preliminary tests on flight data and Monte Carlo simulations show that this approach indeed provides discrimination power between in-time and out-of-time events. It also features good separation power between electromagnetic and hadronic events that can be used in the following reconstruction steps and for background rejection.

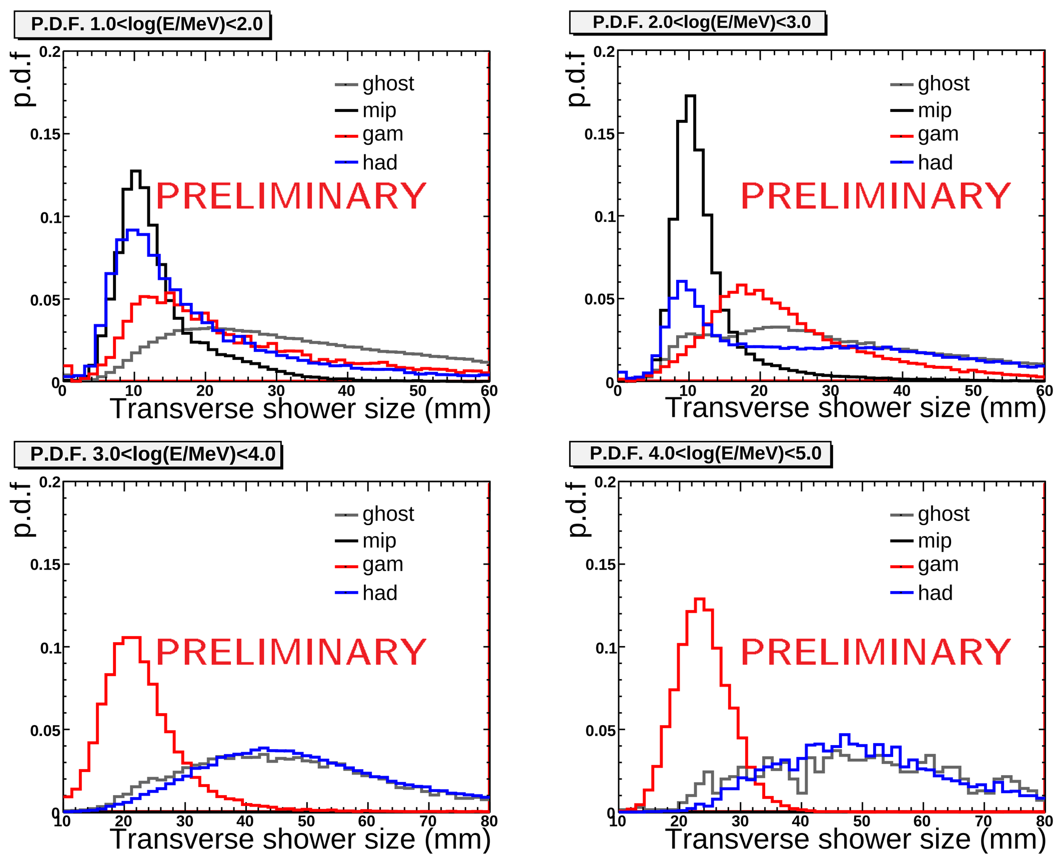
Naive Bayes Classifier

A Naive Bayes Classifier [1] is a simple probabilistic classifier based on Bayes' theorem. Given a set of m topological classes $c_1 \dots c_m$ using n variables $v_1 \dots v_n$ (assumed to be independent) and a known probability density function, $p_i^{c_j}(v_i)$, for each class c_j and variable v_i, Bayes' theorem can be used to invert the problem and calculate the probability that a generic event with $\tilde{v}_1 \dots \tilde{v}_n$ belongs to class c_j:

$$P(c_j) = \frac{\prod_{i=1}^{n} p_i^{c_j}(\tilde{v}_i)}{\sum_{j=1}^{m} \prod_{i=1}^{n} p_i^{c_j}(\tilde{v}_i)}$$

Where $\sum_{j=1} P(c_j) = 1$.

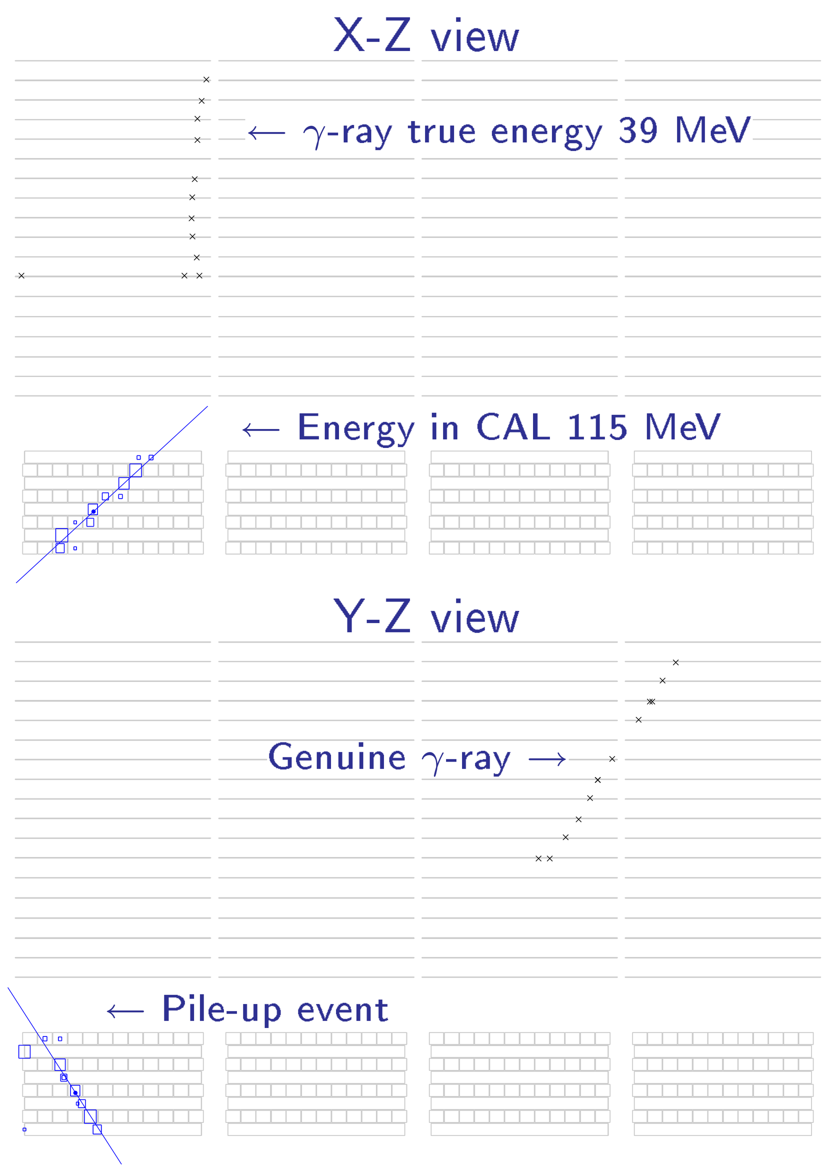
Probability Density Functions



For each each class type we:

- Apply a minimal selection cut on number of calorimeter crystals hit greater than 3.
- Build PDF's for each topological variable based only on calorimeter information including output from
  - Minimum Spanning Tree [2] variables
  - Moment analysis [3] variables
- Bin in energy (figure on the left illustrates an example for the transverse shower size in four different energy intervals)
- Use equally populated bins to avoid *holes* in the distributions

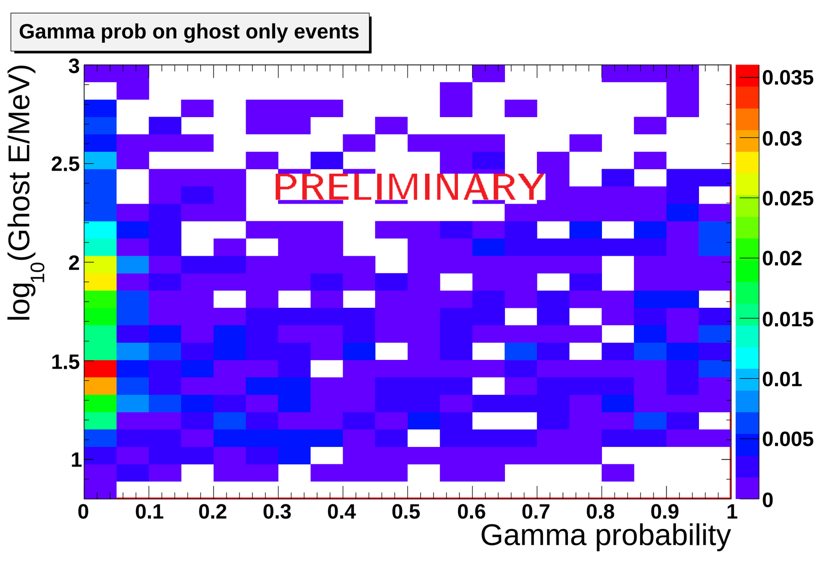
Pile-up events at low energy



Pile-up events increase the energy deposited in the calorimeter and at low energies (100's of MeV) clusters may be entirely due to the pile-up event. With the aid of an on orbit instrumental pile-up from periodic trigger events overlaid on simulated γ events we are able to test the capabilities of NBC in separating the two components. Below are the classification for the event to the left:

- Gam probability 3%
  - Pontentially recoverable event thanks to sensible energy measurement based on the multiple scattering in the tracker.
- Ghost probability 67%
- Had probability 30%

Gamma probability as a function of deposited energy when the only signal in the calorimeter is from the pile-up event.



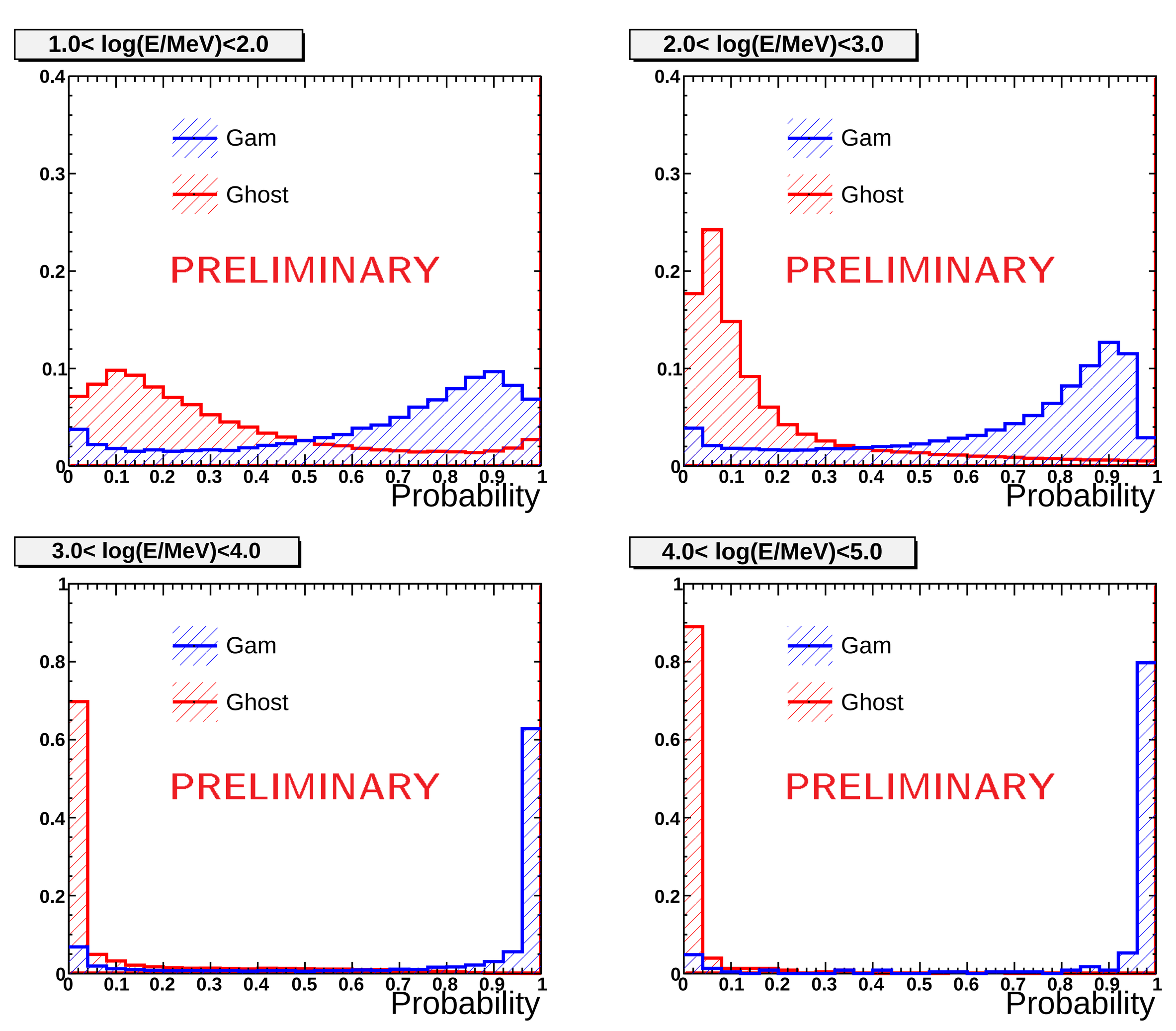
Event classes and datasets

We define four event classes

- Gamma (Electromagnetic showers)
- Pile-up events (Ghost)
- Hadron
- Minimum Ionizing Particle (MIP)

Our classifier is trained using dedicated Monte Carlo simulations of each class type. The *ghost* class is trained on a sample of periodic trigger events from flight data.
NBC performance is tested on dedicated calibration flight data sets. A detailed description of each calibration data set can be found in [4].

Classification on flight data



- NBC test on calibration flight γ-ray data set.
  - Bright source of γ-rays available in flight data thanks to the Earth limb.
- Probability distribution for gammas is peaked towards 1
- For enegies below $\sim$1 GeV, more gammas are misclassified as pile-up events. Room for improvement.

Future plans

- Implement energy weighting scheme to avoid possible artifacts at the bin edges.
- Investigate potential of NBC separation power in the background rejection stage through the Gamma, Hadron and MIP event class probabilities provided by the algorithm.

Conclusions

Preliminary tests on flight data show that the Naive Bayes Classifier technique is a useful tool for discriminating between gamma and pile-up events. In the 100's of MeV range, pile-up events can constitute most, if not all, of the signal in the calorimeter leading to potential loss of gamma-ray events. We find that the NBC algorithm is capable of identifying pile-up events in these particularly difficult cases. This analysis is still work in progress.

References

[1] H. Zhang, *The Optimality of Naive Bayes*, FLAIRS2004 conference (2004).
[2] C. Sgrò, et al., *A Minimum Spanning Tree clustering algorithm for the Fermi LAT Calorimeter*, Poster Fermi Symposium 2011
[3] L. Baldini, et al., *The Fermi LAT Calorimeter as a gamma-ray telescope*, Poster Fermi Symposium 2011.
[4] J. Bregeon, et al., *Calibration flight data sets for the Fermi LAT data analysis*, Poster Fermi Symposium 2011.