Advanced Classification Techniques

Omar Moreno

Santa Cruz Institute for Particle Physics

University of California, Santa Cruz

omoreno1@ucsc.edu

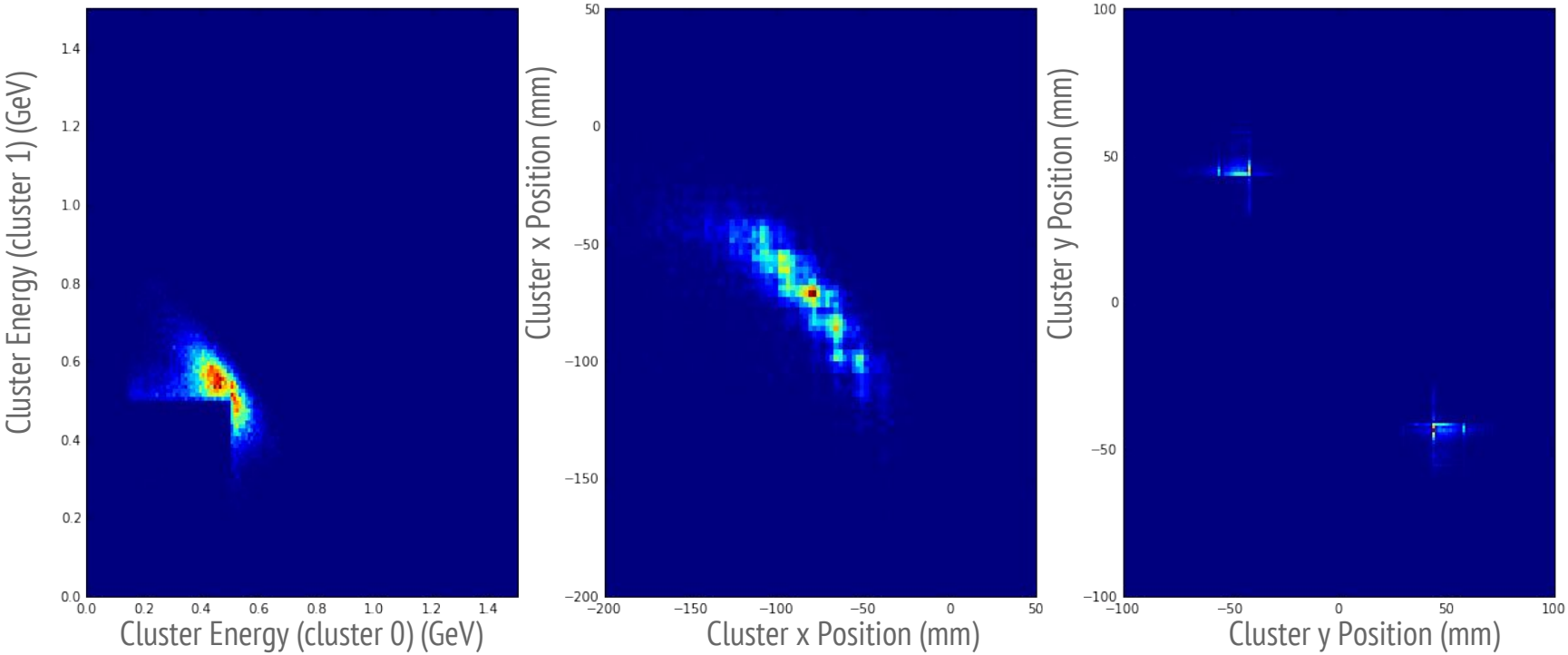
Heavy Photon Search Collaboration Meeting

October 26-28, 2015

Preliminaries

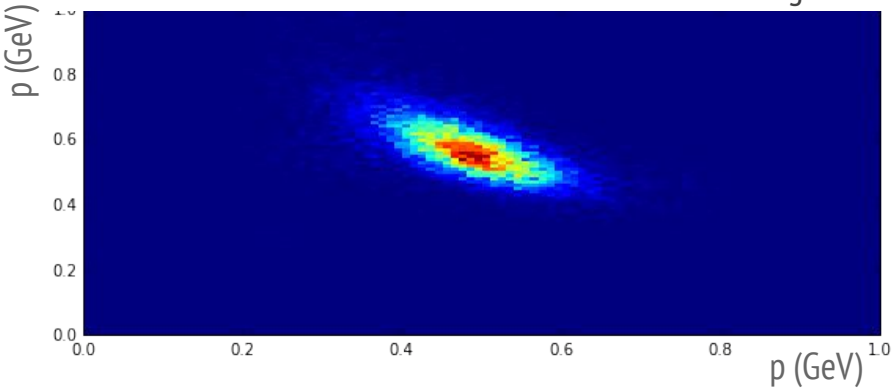
- Wanted to understand if classification of events using any one of the machine learning algorithms is more efficient than using a manual selection
- Started using TMVA but switched to scikit-learn
- As a proof of concept, begin by trying to classify Mollers using a simple decision tree and ensemble methods
- Train and validate all algorithms using the pass 3 MC Moller sample as the signal and MC beam-tri files, with some cuts, as the background
- Run over run 5772 in order to check that the algorithms are reasonable
- All files (MC and data) are preprocessed
  - Only look at events that have two clusters in opposite volumes and are within a time window of 1.6 ns
  - Require that both clusters have a track matched to them
- A flat ntuple containing the following variables is made out of the events that pass the cut:
  - Energies and position of Ecal cluster pair
  - Track momentum, parameters and charge of the tracks matched to the clusters

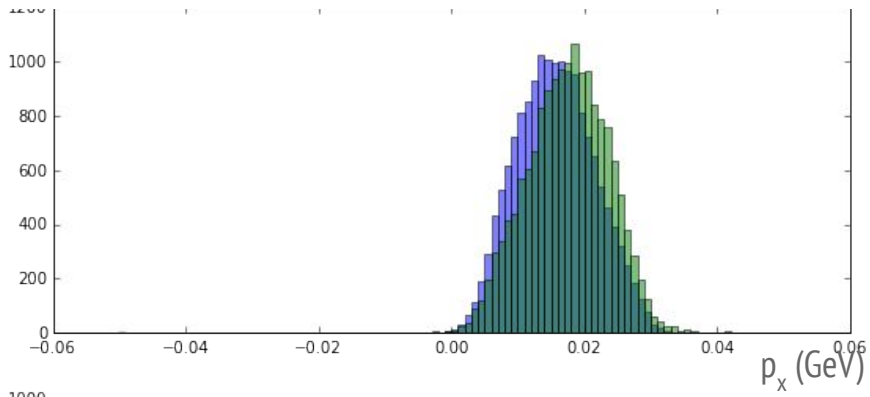
Signal Preprocessing

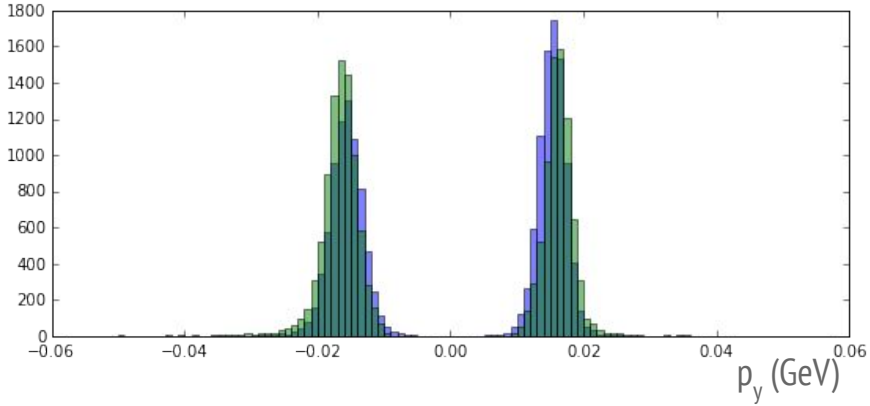


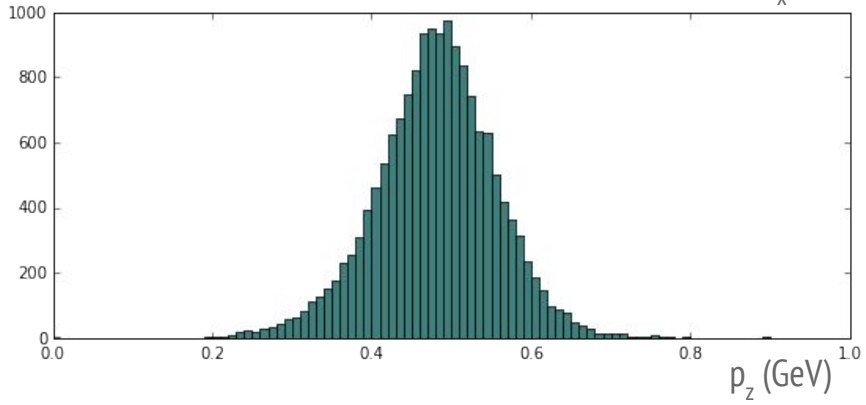
Signal Track Variables

blue: track 0 - track associated with highest energy clusters
green: track 1



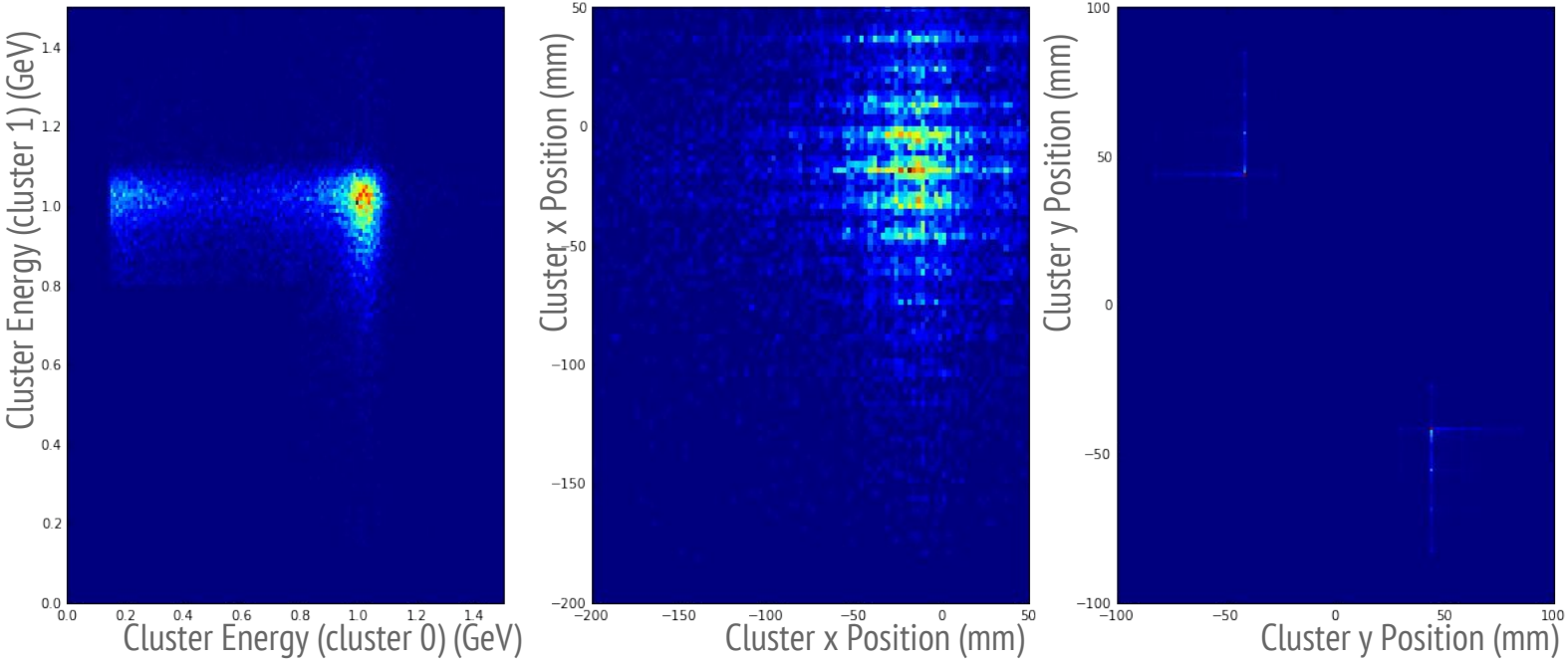




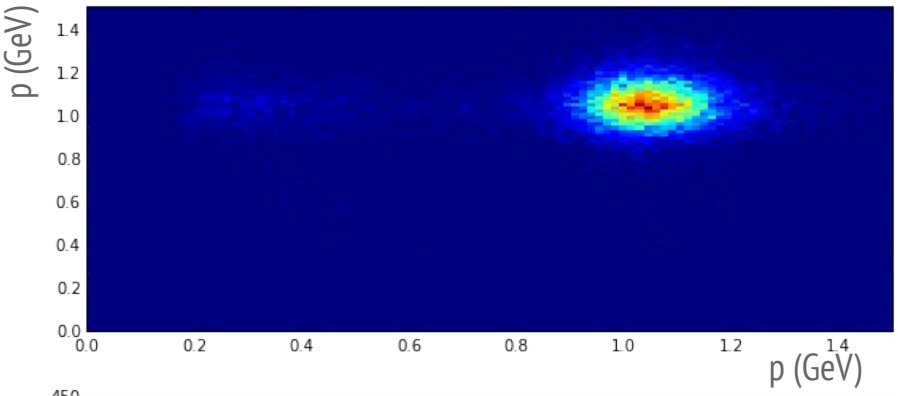


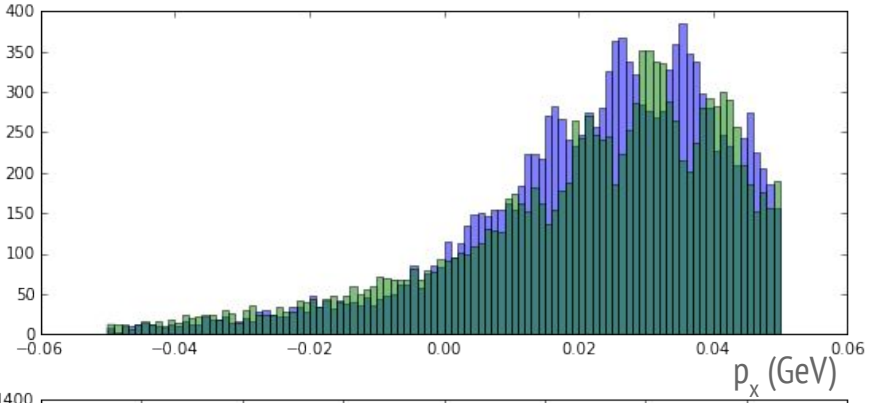
Background Preprocessing

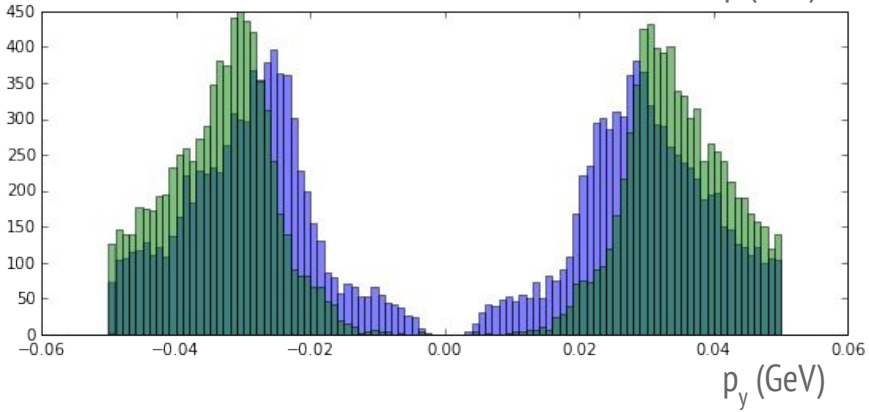
- Apply same requirements as signal but …
- Don't have background files without Mollers so require that one of the clusters in the pair have an energy > .8 GeV cut to remove possible moller candidates

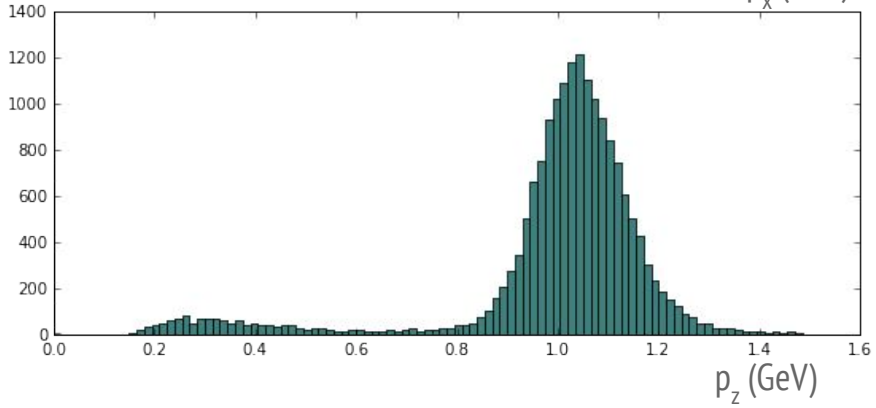


Background Track Variables



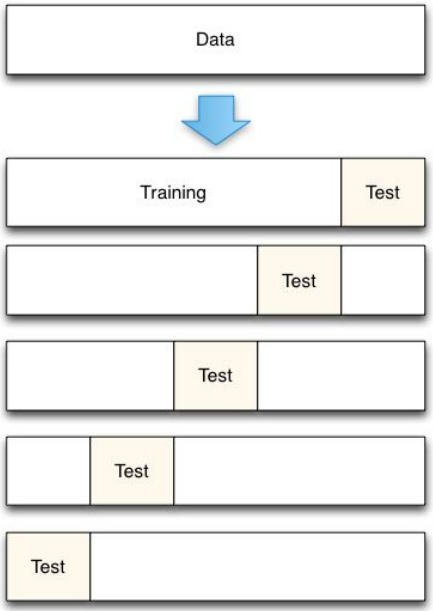






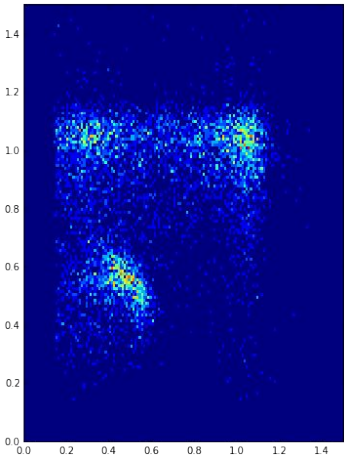
Cross Validation

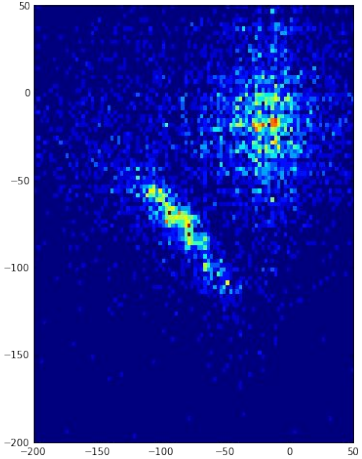
- Combine signal and background sample and then break them up into 20 equal samples
- Train the algorithms using 19 out of the 20 samples
- Calculate the efficiency to separate signal from background and their means as their score

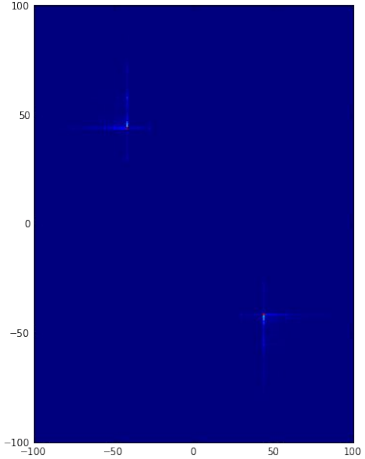


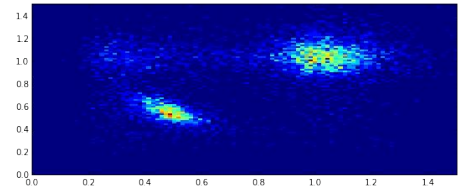
Classifier	Score
Decision Tree	0.999410359802
Random Forest	0.99929246999770327

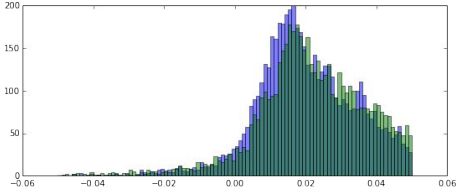
Using Real Data

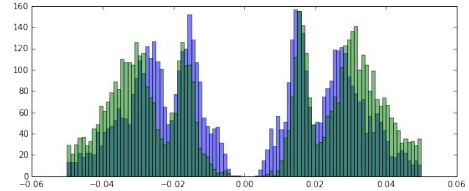


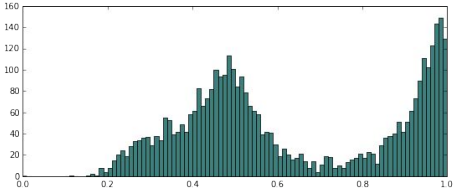




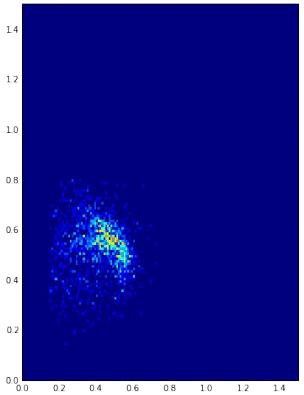


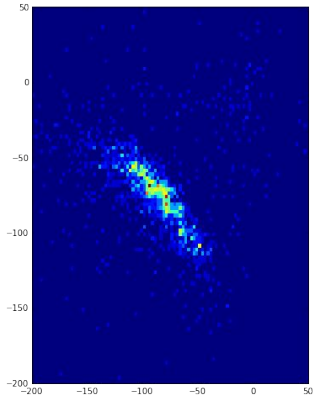


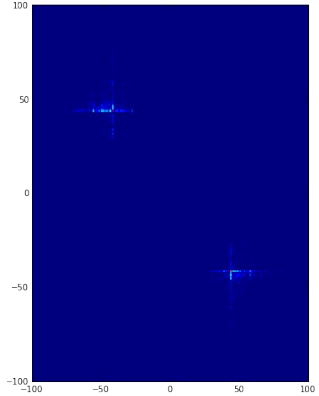


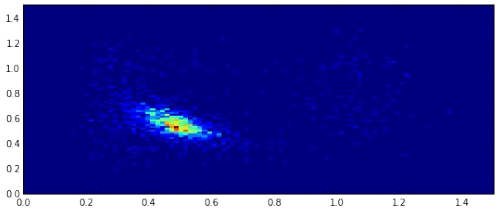


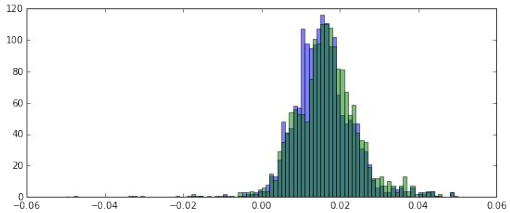
Decision Tree

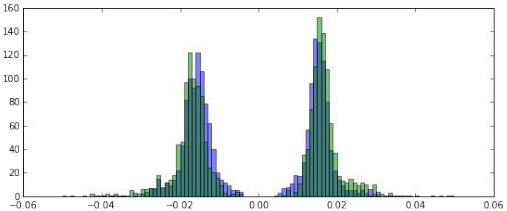


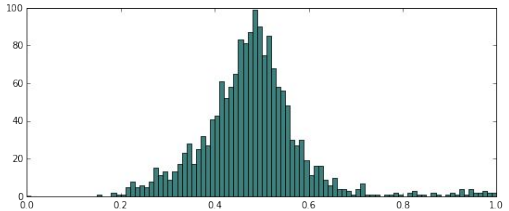












Random Forest