Enhancing Lepton Identification in CLAS12 using Machine Learning Techniques

Mariana Tenorio-Pita
Old Dominion University

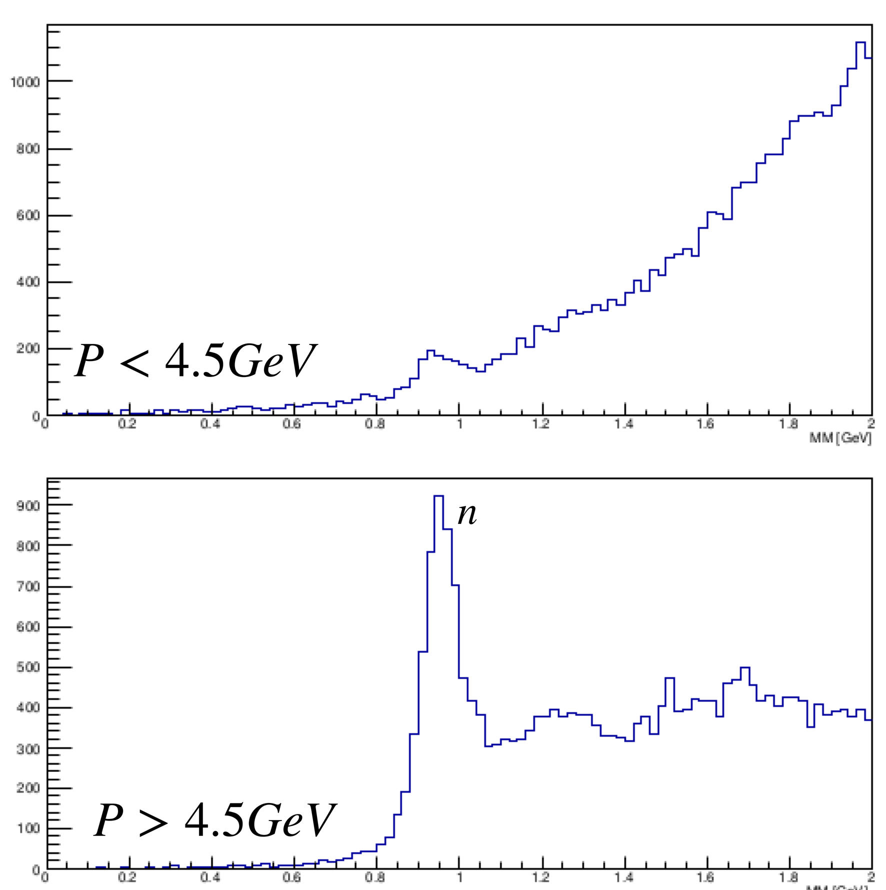
Motivation

- In CLAS12, the Event Builder (EB) is responsible for associating the detector responses to particles to execute particle identification (PID).

  PID $= 11 \rightarrow$ electron
  PID $= -11 \rightarrow$ positron

- One of the requirements to assign a PID=± 11 is a signal in the High Threshold Cherenkov Counter (HTCC).
- At P >4.9GeV, both leptons and charged pions produce a signal in HTCC.

This led to pion contamination in both experimental data and simulations.



For example, in the reaction $ep \rightarrow e^-\pi^+(X)$ in the missing mass distributions of X we should see a peak corresponding to the missing neutron. We select events with one e- and one particle with PID=-11 and assign mass of the pion to it to calculate the missing mass.

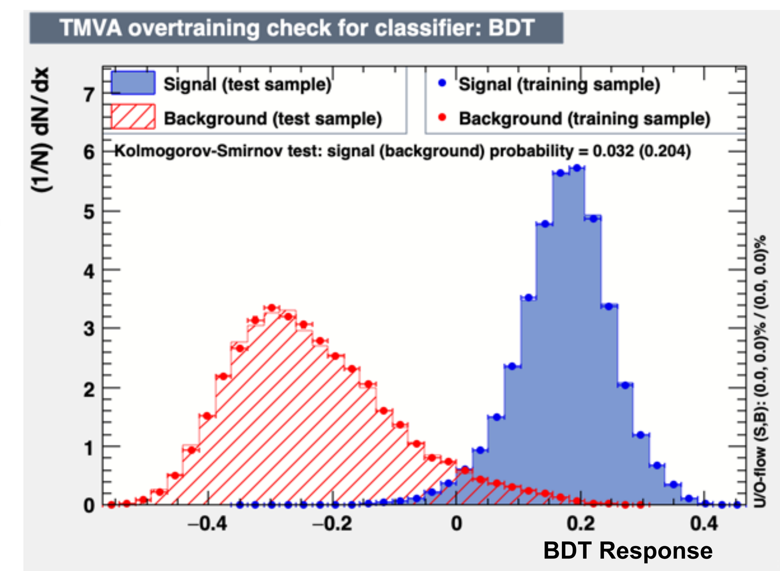
By looking in two different momentum range, we observe a clear peak in the neutron mass, due to the π^+ that has been identified as a positron by the EB.

TMVA Methods

- The Toolkit for Multivariate Analysis is a ROOT integrated environment for the processing, evaluation and application of multivariate classification and regression techniques belonging to the family of supervised learning algorithms.
- With TMVA packages, we can use several variables, such as the sampling fractions, to distinguish between leptons (signal) and pions (background).
- The techniques used for this work were Boosted Decision Trees (BDT) and Multilayer Perceptron (MLP).

Training

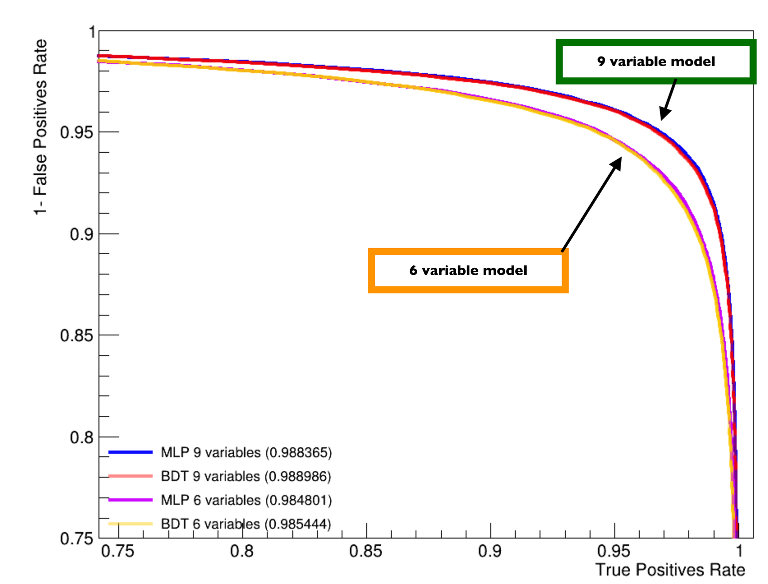
- The models were trained on simulation.
- The variables used where P, θ, ϕ, the sampling fractions and sum of second-moment (m2) of the calorimeters (PCAL, ECIN and ECOUT) and we tested 6 and 9 variable models.
- After the training we obtain for each model a weight file that containing the results of the training.
- The cut efficiencies and the optimal cut value are obtained for each method.



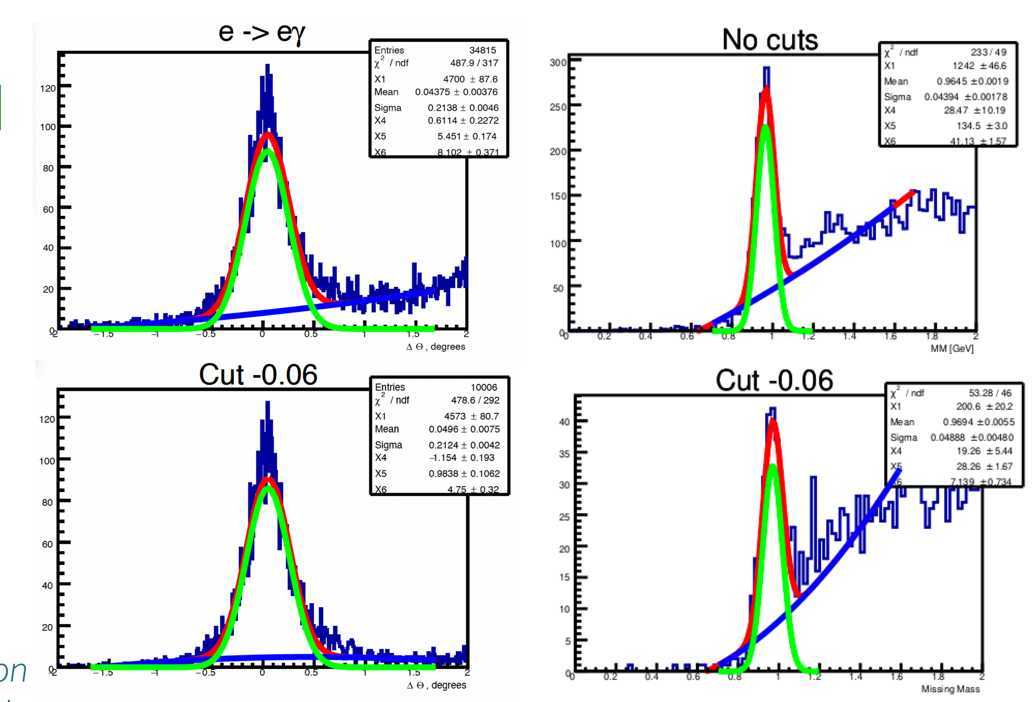
Distributions of signal and background events as a function of the classifier response. Model BDT 9.

Validation

- The validation of the models was done using simulations and data.
- **For simulations:** Using an independent Montecarlo data set, the response for each event was obtained and stored. From here, we have 4 possible outcomes: True Positive, False Positive, True Negative, False Negative.
- Using this information, we can calculate the signal efficiency and background rejection rate, and build the ROC curve.



ROC curve obtained from the validation on simulations. The results show that the 9 variable models are slightly more efficient.



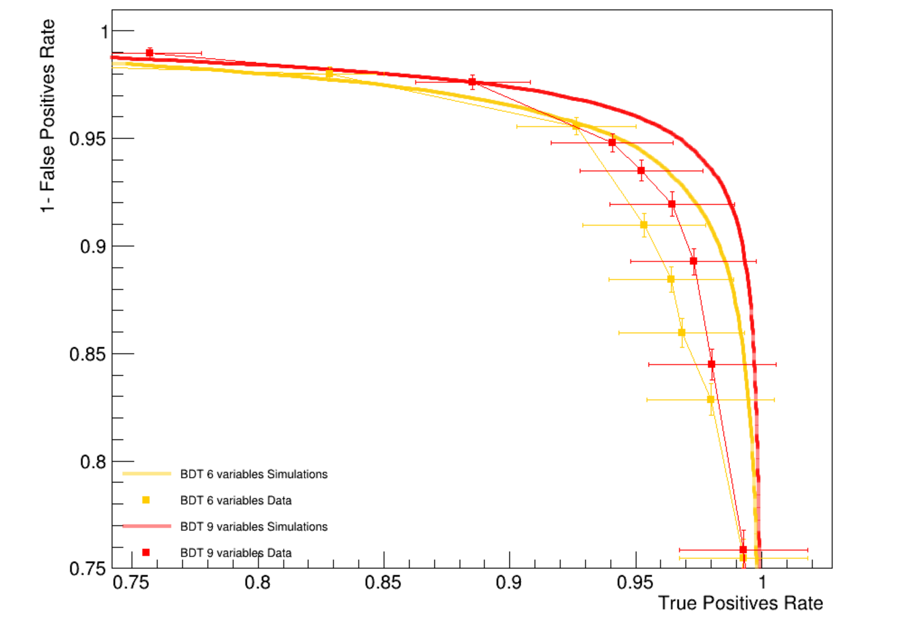
Signal (left) and background (right) samples obtained from the data. By applying a cut in the response at -0.06, we keep most of the data while reducing significantly the background.

- **For data:** A clean sample of leptons was obtained using radiated photons while they go through the CLAS12 target and the scattering chamber. The number of true leptons was obtained by fitting the peak at zero in $\Delta\theta_{e\gamma}$.
- For the background sample, we use the reaction $ep \rightarrow e^-\pi^+_{PID=-11}(X)$, where π^+ was identified as positron, and extracted the number of background events by fitting the missing neutron mass peak.
- For all samples, we look at the region $P > 4.5GeV$.

Results

BDT 9/ Spring 2019	Data	Simulation
Signal	97.30%	99.36%
Background	10.73%	10.70%

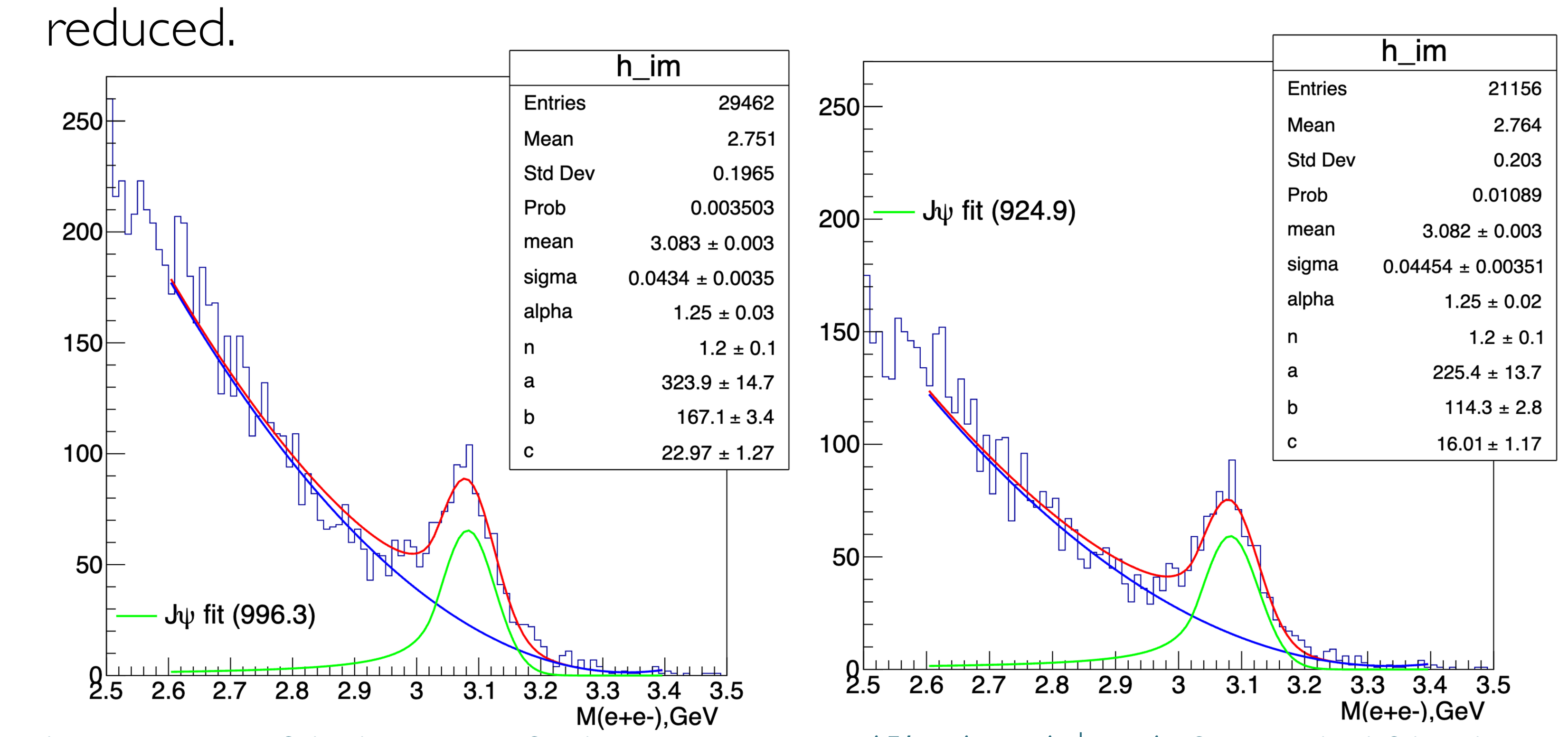
By having a cut in the response at -0.06, we observe that we are keeping most of the data while suppressing most of the background.



ROC curve of validation on simulation (solid) and data (dots) show consistency with each other.

Impact on Time-like Compton and J/ψ production studies

- The reaction of study is $ep \rightarrow e'J/\psi p' \rightarrow e'e^+e^-p'$
- The Time-like Compton scattering and the J/ψ production close to threshold are important reactions for studying nucleon Generalized Parton Distributions and the gluonic Gravitational form-factors.
- Fiducial cuts in momentum, energy and vertex time are applied to remove background and radiative energy loss correction is applied to the lepton pair.
- We can apply the BDT 9 model to the data, and we can see the effect on the invariant mass, where the background is clearly reduced.



Invariant mass of the lepton pair for the reaction $ep \rightarrow e'J/\psi p' \rightarrow e'e^+e^-p'$ after standard fiducial cuts and radiative energy loss correction. On the left we have the analysis without the machine learning lepton ID, on the right we have applied the models to the data. We can see that while the background is clearly reduced, the number of J/ψ events remains almost the same.

References

V. Ziegler, et al. NIM-A, 959:163472 (2020)
G. Asryan, et al. NIM-A, 959:163425 (2020)
A. Hoecker, et al. TMVA - toolkit for multivariate data analysis. (2009)
S. Brodsky, E. Chudakov, P. Hoyer, J. Laget, Phys. Lett. B. 498, 23 (2001)

marianat@jlab.org