

Multivariate methods for particle identification

Cosmin Vişan

(*CERN, cosmin.visan@cern.ch*)

September 10, 2013

Abstract

The purpose of this project was to evaluate several MultiVariate methods in order to determine which one, if any, offers better results in Particle Identification (PID) than a simple $n\sigma$ cut on the response of the ALICE PID detectors. The particles considered in the analysis were Pions, Kaons and Protons and the detectors used were TPC and TOF. When used with the same input $n\sigma$ variables, the results show similar performance between the Rectangular Cuts Optimization method and the simple $n\sigma$ cuts. The method MLP and BDT show poor results for certain ranges of momentum. The KNN method is the best performing, showing similar results for Pions and Protons as the Cuts method, and better results for Kaons. The extension of the methods to include additional input variables leads to poor results, related to instabilities still to be investigated.

1 The ALICE Experiment

A Large Ion Collider Experiment (ALICE) is one of the seven detector experiments at the Large Hadron Collider at CERN[1]. The other six are: ATLAS, CMS, TOTEM, LHCb, LHCf and MoEDAL. ALICE is optimized to study heavy ion collisions. Pb-Pb nuclei collisions are studied at a centre of mass energy of 2.76 TeV per nucleus. The resulting temperature and energy density are expected to be large enough to generate a quark-gluon plasma, a state of matter wherein quarks and gluons are deconfined[2].

1.1 TPC Detector

The ALICE Time Projection Chamber (TPC) is the main particle tracking device in ALICE. Charged particles crossing the gas of the TPC ionize the gas atoms along their path, liberating electrons that drift towards the end plates of the detector. An avalanche effect in the vicinity of the anode wires gives the necessary signal amplification. The positive ions created in the avalanche induce a positive current signal on the pad plane. The readout is done by the 557 568 pads that form the cathode plane of the multi-wire proportional chambers (MWPC) located at the end plates. This gives the r and ϕ coordinates. The last coordinate, z , is given by the drift time. The specific energy loss de/dx is also recorded with up to 159 samples per track.

1.2 TOF Detector

The time of flight detector (TOF) contributes to the identification of charged particles. For a given momentum p , heavier particles are slower and so take longer to reach the outer layers of the detector. For its TOF system ALICE uses detectors called Multigap Resistive Plate

Chambers (MRPC). There are approximately 160 000 MRPC pads distributed over the large surface of 150 square meters. The total time resolution is about 100 ps, the intrinsic MRPC resolution being smaller. Using the tracking information from other detectors, the tracks firing a sensor can be identified.

2 Aim of the Project

In Figure 1 the response of the TPC detector is shown. It can be seen that while at low momentum the different species are clearly separated and can be easily identified, after a certain momentum, about 1.8 GeV/c, all the lines merge and track-by-track identification becomes difficult. Similar considerations apply to the response of the TOF detector. The TPC and TOF detectors were the PID detectors used in this project.

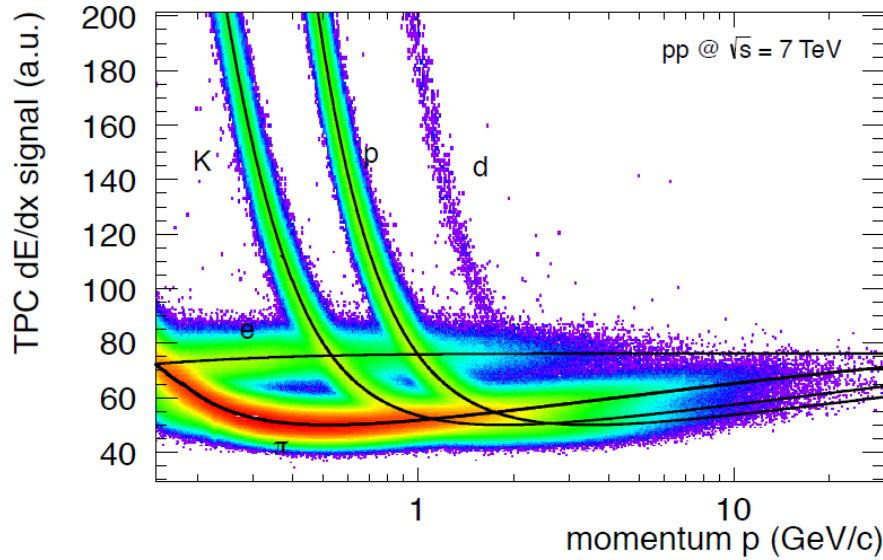


Figure 1: TPC Detector Response, figure from [3]

There are various techniques that can be used to do a proper particle identification. The simplest of them is to use the $n\sigma$ cuts on the (almost) gaussian response of the PID detectors, i.e. find the particles that are within $3n\sigma$ cuts from the expected response of a given detector. In order to improve PID in the regions of overlapping bands, it was decided to include other variables and to treat them with a multivariate analysis (MVA). These are methods known for giving an optimal response for a problem which depends on many variables.

The final goal was to determine if by using MVA instead of simple $n\sigma$ cuts, the particles can be better identified. In order to achieve this comparison, two quantities were used: efficiency and purity. They are defined as follows:

$$Efficiency = \frac{Correctly\ identified\ i}{All\ reconstructed\ i}, \text{ where } i = P, K \text{ or } \pi$$

$$Purity = \frac{Correctly\ identified\ i}{Identified\ as\ i}, \text{ where } i = P, K \text{ or } \pi$$

3 The TMVA Methods

MVA are methods which make optimal use of many variables for classification or regression problems. They are examples of the machine learning algorithms. In general, machine learning can be classified into supervised and unsupervised learning. In supervised learning, the machine is trained on a set of data which already contains the inputs and the desired outputs. In unsupervised learning, the available data is unlabelled and the machine is trying to determine the hidden structure of the data. We use ROOTs TMVA which is a toolkit of supervised learning methods. For a general introduction on MVA methods, see [5].

3.1 Initial Analysis

As the TMVA results needed to be benchmarked against a simple analysis technique, the method used as a reference in this project for finding efficiency and purity was the simple $n\sigma$ cut method. In this method, cuts of $n\sigma < 3$ were applied to the responses of the TPC and TOF detectors.

3.2 Rectangular Cuts

This method consists of applying rectangular cuts in a N-dimensional feature space, where N is the number of variables. The method is analogous to the simple $n\sigma$ cuts, the difference being that the cut is not applied for $n\sigma < 3$, but is determined automatically by the TMVA in such a way as to obtain the highest possible significance for a fixed value of the efficiency. For certain distributions, as shown in Figure 3, the data must first be rotated in order for the rectangular cuts to give the optimal results, rotation meaning removing correlations. TMVA implements several standard methods to remove correlations.

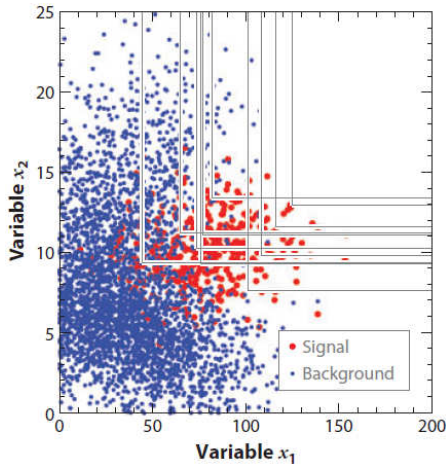


Figure 2: Cuts Method, figure from [4]

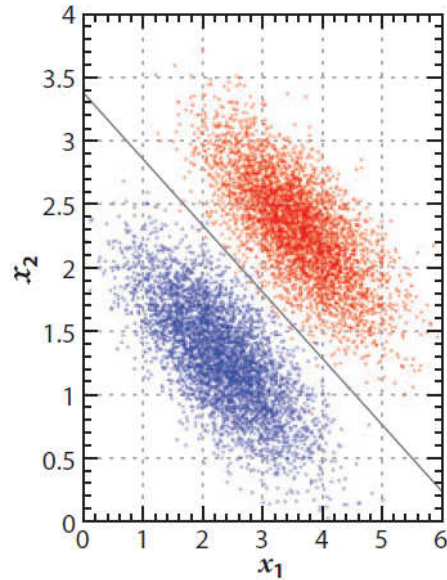


Figure 3: Cuts method - Data correlation, figure from [4]

3.3 K-Nearest Neighbours (*KNN*)

In this method, the search uses hyperspheres centred on points in the data, to find the k-nearest neighbours (see Figure 4). The output is a probability for the point to be a signal,

depending on the distribution of points around it. The search is done using a specific metric, the simplest one being the Euclidian distance. When the input variables have different units, a variable with a wider distribution contributes with a greater weight to the Euclidian metric. In this case, a weighted metric is more suitable to be used.

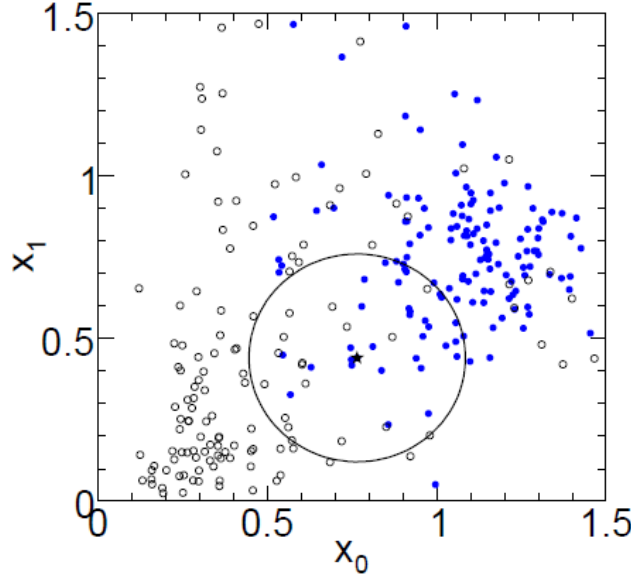


Figure 4: KNN Method, figure from [5]

3.4 Multi Layer Perceptron (*MLP*)

MLP is a type of neural network algorithm. The search is done using many layers of interconnected neurons. There is one layer representing the input variables, one layer of one neuron representing the output response and at least one hidden layer. A neural network is a generic way to implement non-linear function. In the example shown in Figure 6, the MLP method finds contours which wraps around the signal points. Based on the contours, an optimal cut can be chosen to be further used in the analysis.

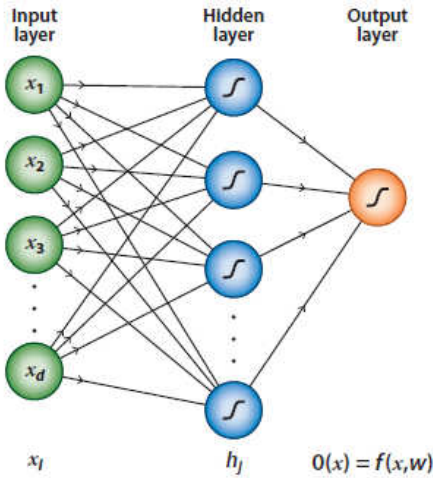


Figure 5: MLP structure, figure from [4]

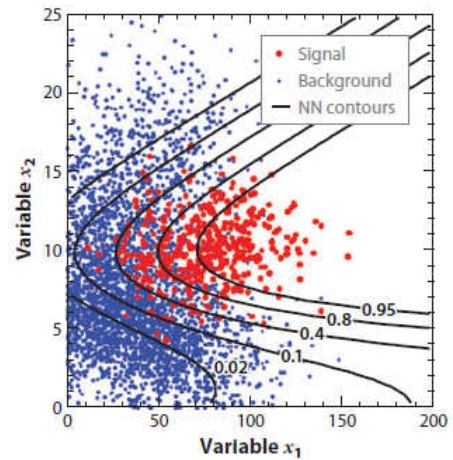


Figure 6: MLP output, figure from [4]

3.5 Boosted Decision Trees (BDT)

A decision tree consists of a series of rectangular cuts applied sequentially at each node (see Figure 7). The process stops when a given terminal condition is satisfied. The result are distinct regions in the variable space as in Figure 8, each region being classified as a signal or as a background.

The performance of a decision tree can be increased using boosting, which consists of using a sequence of classifiers that work progressively harder on increasingly difficult events, instead of simple rectangular cut. The classifiers chosen need not to be the most high-performing ones, but need to be chosen such that collectively offer a boosted performance. If a forest of decision trees is used together with the boosting principles, the MVA method is called Boosted Decision Trees (BDT).

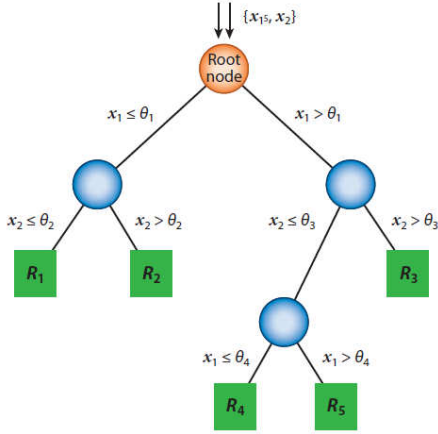


Figure 7: Decision Tree, figure from [4]

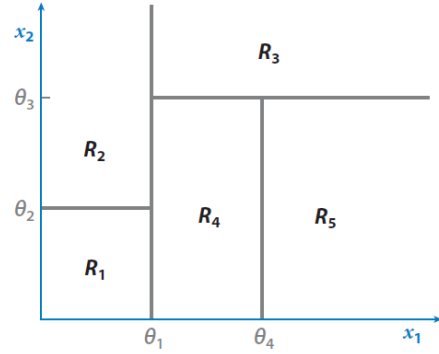


Figure 8: Decision Tree output, figure from [4]

4 TMVA Workflow

The workflow in TMVA is presented in figure 9. It consists of 2 parts: classification and application.

The classification is the part where the methods are trained, tested and evaluated. TMVA offers a Factory object which is the interface between the user and the methods. All the information needed for the training and testing of the methods, are given to the Factory, such as the input variables and the chosen method. TMVA requires input data for the signal and for the background. In this project, different set of data were provided to the Factory for each of the 2 categories, signal and background. Then, TMVA splits internally the data into data to be used for the training and data to be used for the testing. In the training phase, TMVA is training the method for the input set of data. Then the trained method is tested on another set of data. The results are stored in weight files which are then used in the Application phase.

The Application phase provides the user with a Reader object which takes up similar role to the Factory object. The same variables used in the Classification, need to be registered now to the Reader. Then after the method is selected, the outputs of the Classification are used for various analysis tasks. For the Application phase, a different set of data was used.

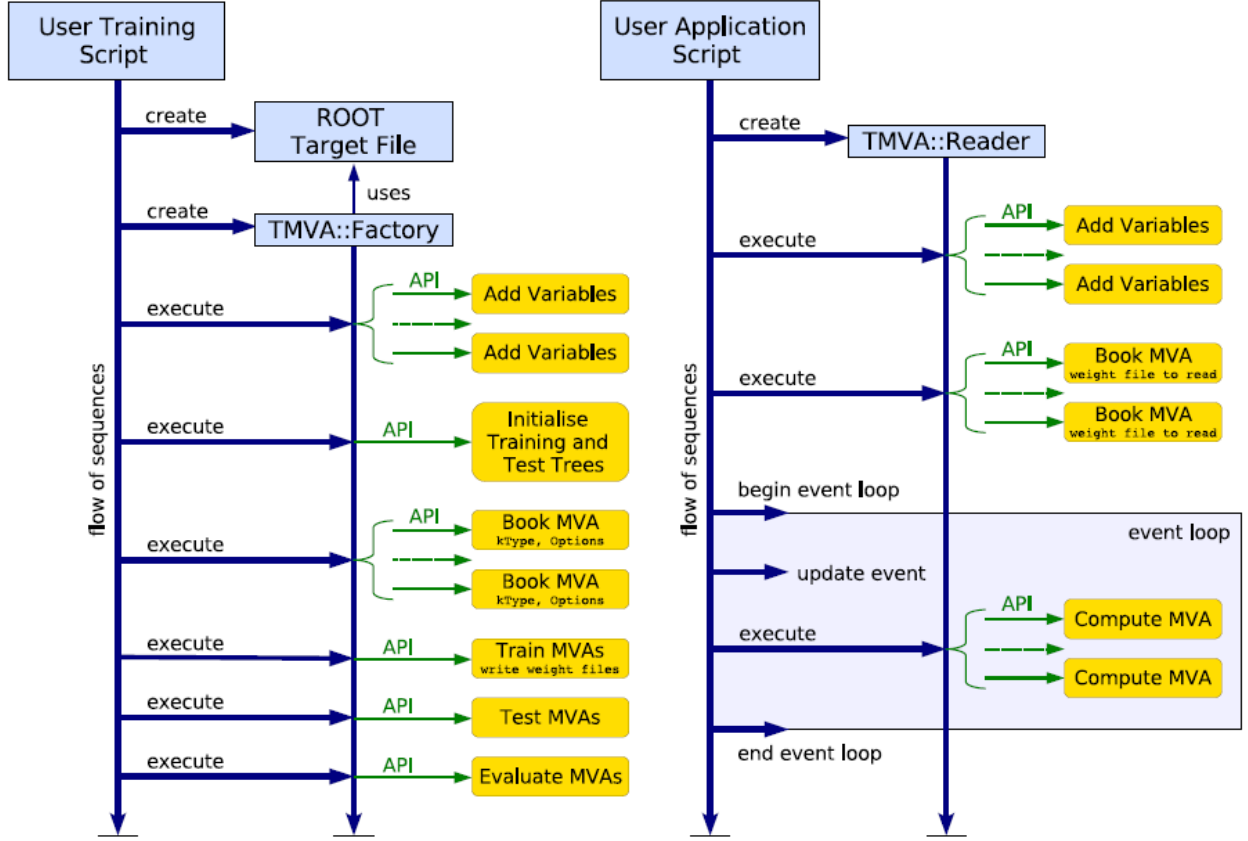


Figure 9: TMVA Workflow, figure from [5]

5 The Analysis

The data used in the analysis comes from MonteCarlo simulation. The runs used are: *LHC13b2_efix_p4_ESD_195483*, *LHC13b2_efix_p4_ESD_195568* and *LHC13b2_efix_p4_ESD_195644*. The identities of the particles are known by their PDG Code. The training and testing phase can be done by taking this into account. The particles used in the analysis were pions (PDG Code = 211), kaons (PDG Code = 321) and protons (PDG Code = 2212). Because the sizes of the data samples used were too big to be able to carry the analysis, I wrote an analysis task based on the alice framework (AliAnalysisTask) to preselect events, apply basic track quality cuts and store lightweight trees with reduced information. The analysis task was processed on the CERN Analysis Facility (CAF), one of the PROOF clusters used by the ALICE collaboration.

Because the detectors have different responses for different momentum bin, the classification was done separately for 20 momentum bins ranging from 0.2 to 5 GeV/c. For the application, 100 momentum bins were used.

The analysis consisted of 2 steps, as outlined below.

5.1 Step 1

In the first step, the discriminating variables used were $n\sigma_{TPC}$ and $n\sigma_{TOF}$. This choice was made so that the results of the TMVA methods can be compared to the results of the simple $n\sigma$ cuts method.

After the Classification is run, TMVA provides a number of plots that show relations between variables and the separation between signal and background as a function of the

classifier response. Using these graphs, a cut on the classifier response is chosen such that the optimal separation between signal and background is obtained. One example of how this cut is chosen is presented below. In Figure 10, the graphs used to choose the cut are presented. In this particular example, the method used was MLP and the particle designated as signal was proton. On the plots on the left, the momentum bin between 1.4 and 1.6 GeV/c is selected. On the plots on the right, the momentum bin between 2.8 and 3.1 GeV/c is selected. On the left plots the best cut is at 0.3, but it cannot be well determined on the right plots. Because it is not practical to choose a different cut for each momentum bin, one general cut was used in this first exercise for all momentum bins. As a result, the final purity graphs have gaps in them as is presented in Figure 11, which corresponds to the selection explained above.

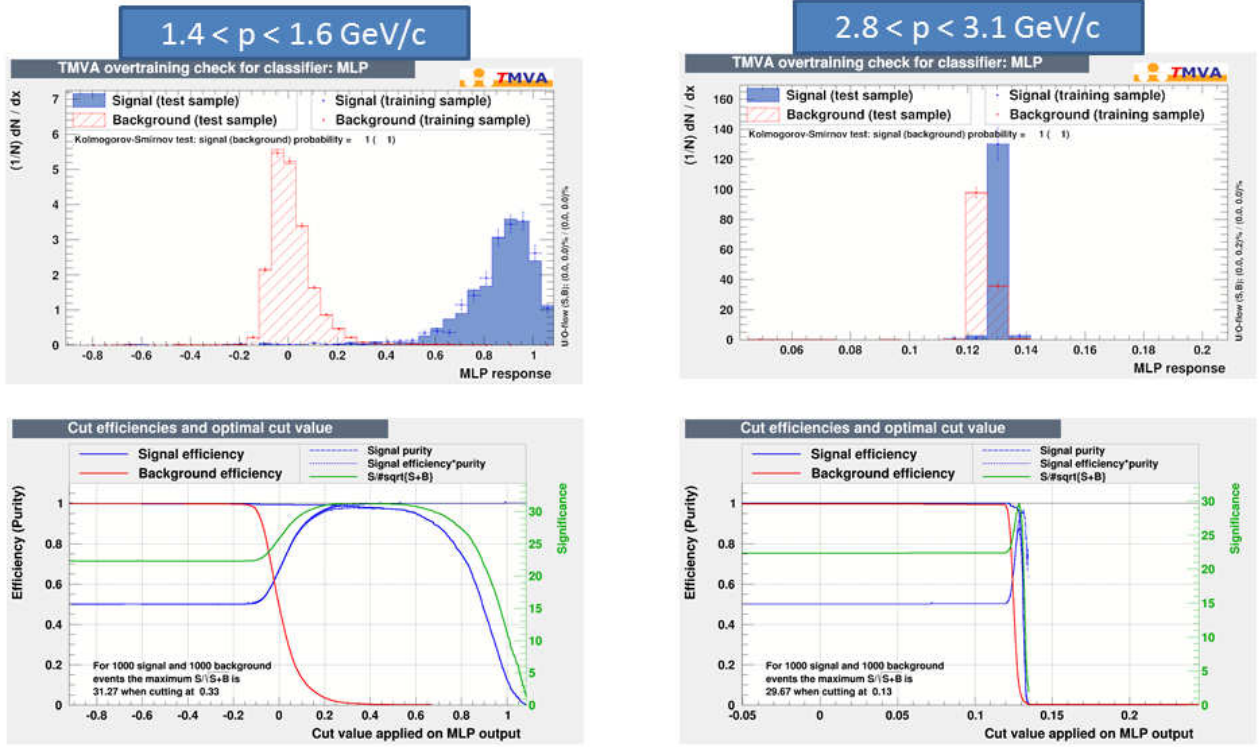


Figure 10: TMVAGui example

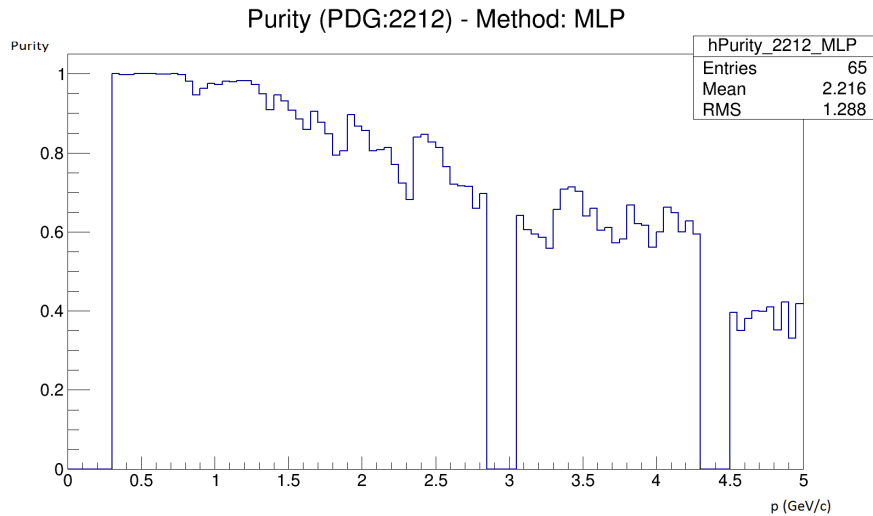


Figure 11: Poor cut for a particular p bin

5.2 Step 2

After the basic features of the different methods have been understood in step one, the analysis was extended to include other discriminating variables, to fully profit of the multivariate techniques. The variables chosen and their discriminating power are shown in Figure 12, where blue represents signal and red represents background. The variables used are: TOF - time of flight, dEdxTPC - energy loss, SigmaTPC - energy loss resolution, nClustersTPC - number of clusters, TPCNclsIter1 - number of clusters after the first iteration, nCrossedRowsTPC - number of crossed rows, chi2PerClustersTPC, TOFsignalDz, and TOFsignalDx. These variables are chosen because they are expected to be related to the identity of the particle, since they are related to interactions with the detector material.

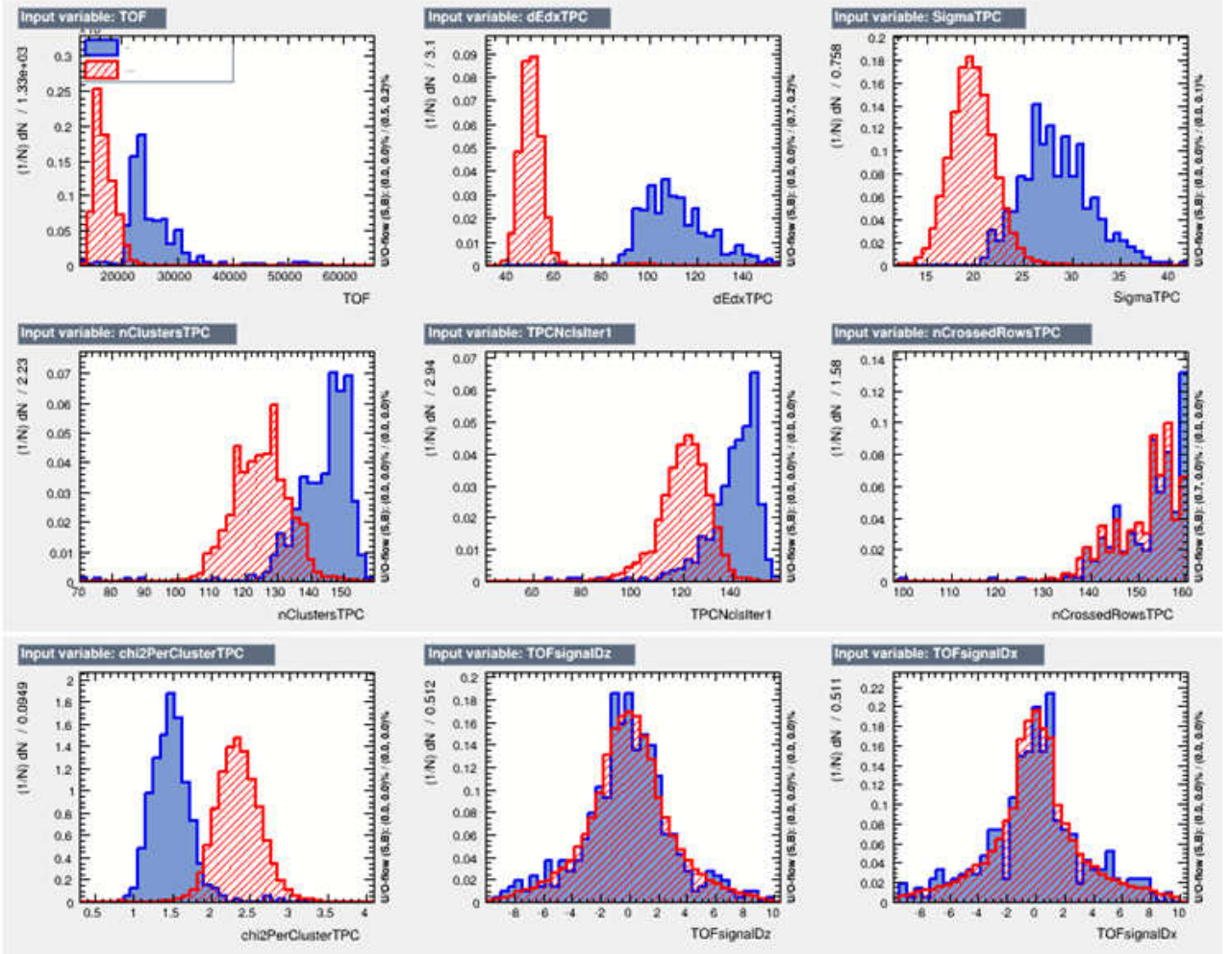


Figure 12: Discriminating variables used in Step 2.

6 Results

The final results, obtained applying the procedures described above, are shown below.

6.1 $n\sigma$ Cuts - not TMVA

The Figures 13, 14 and 15 show the results of the simple $n\sigma$ cut method. It can be seen that the obtained efficiency has a value around 0.95, which was then targeted when using TMVA.

The purity for pions shows very good results of almost 100% up until 1.8 GeV/c, and after that still maintains a good value of 80%. The purity for protons is the second best case, with a value of almost 100% maintaining up to the same critical value of 1.8 GeV/c, where the lines in the response of the TPC detector start to cross. Then the value slowly decreases until about 30%. The Kaons' case is the worst, the purity dropping rapidly after 1.5 GeV/c down to a value of 35%

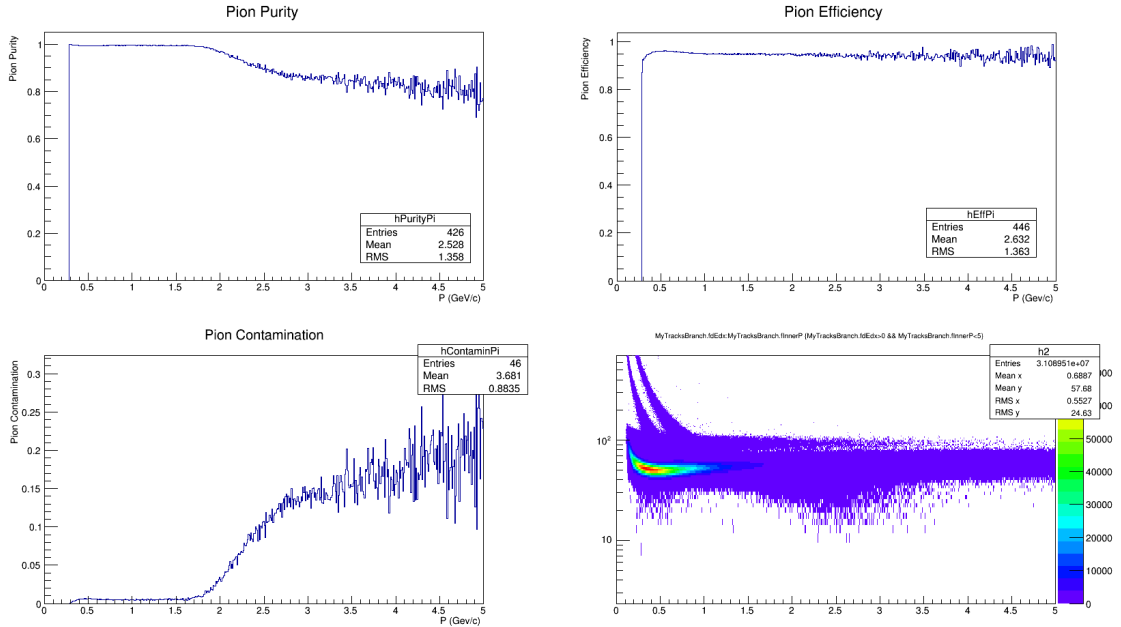


Figure 13: Simple nSigma Cut Method - Pions Graphs

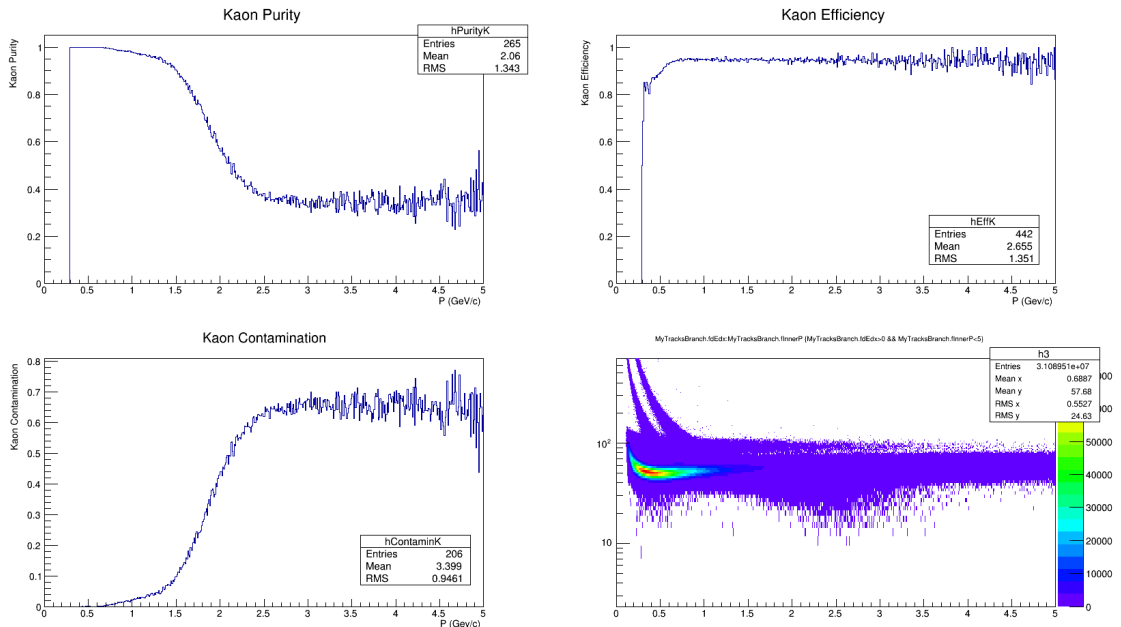


Figure 14: Simple nSigma Cut Method - Kaons Graphs

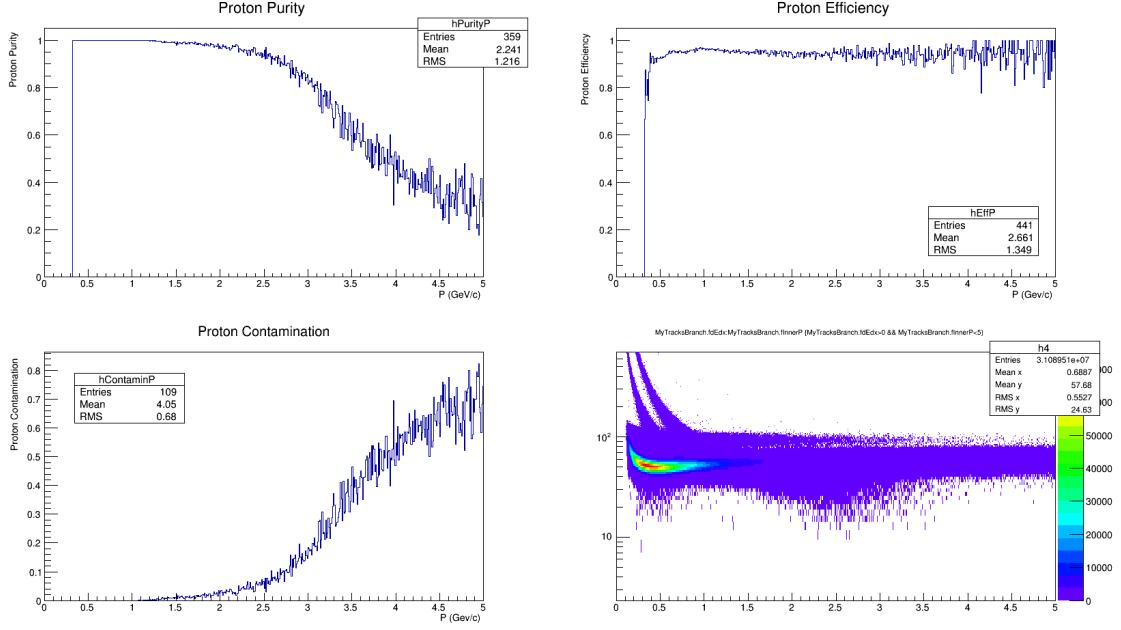


Figure 15: Simple nSigma Cut Method - Protons Graphs

6.2 TMVA Results

In the six figures shown below, the results of the TMVA analysis - the efficiency and purity - are presented. For the Rectangular Cut Optimization (Cuts) method, the plots resemble the plots obtained in the simple $n\sigma$ cuts, which is what was expected. The methods MLP and BDT show many gaps. These two methods need to be investigated further. The KNN method for Pions and Protons offer similar results as the Cuts method, but in the case of Kaons, it offers improved results for purity for momentum above 1.8 GeV/c. The increase is up to 20% as compared to the simple $n\sigma$ cuts.

The results from step 2 are not available, because the optimization likely converges to a local minimum and the results show poorer performance than in the case of step 1, where the method was applied with only two variables. Additional studies, in particular an optimization of the methods parameter or of the minimization algorithm will be needed.

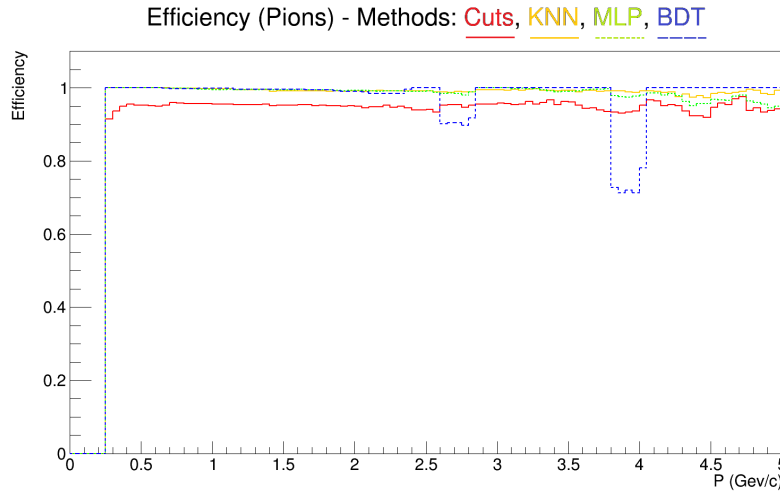


Figure 16: Pions Efficiency

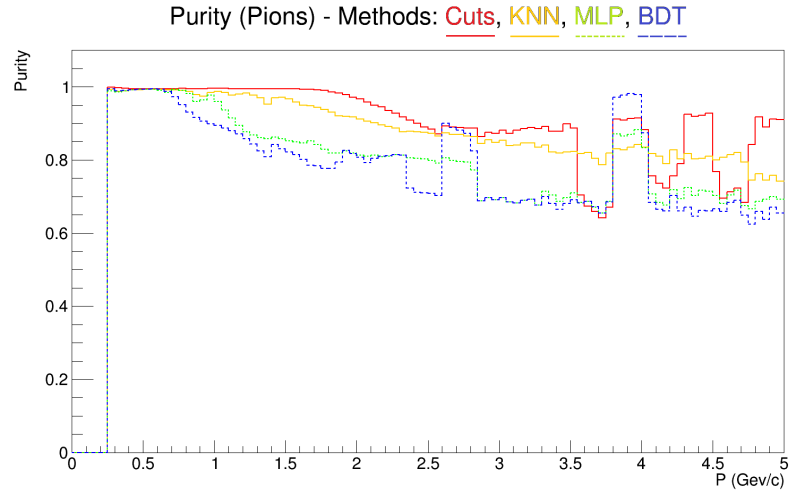


Figure 17: Pions Purity

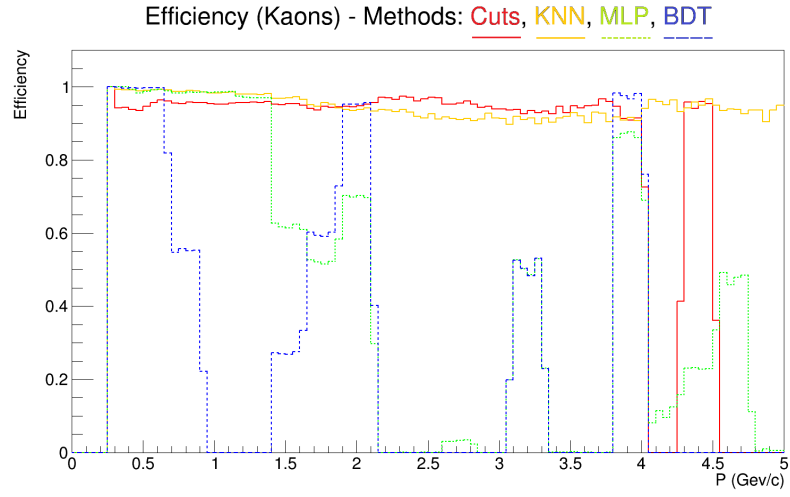


Figure 18: Kaons Efficiency

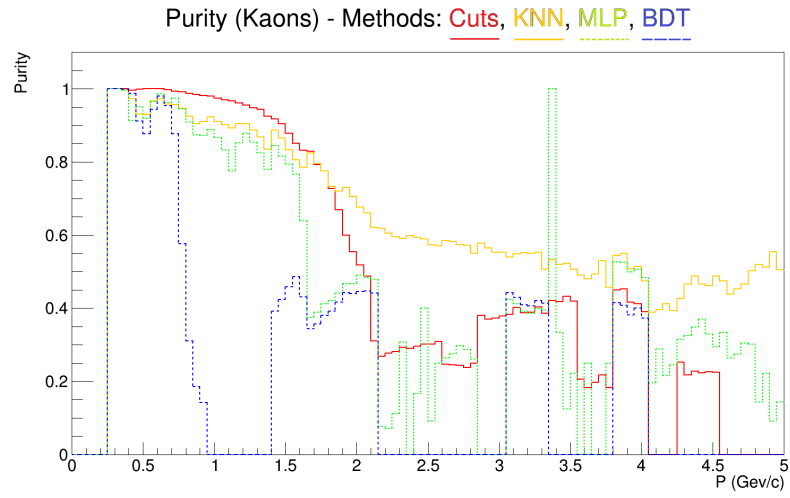


Figure 19: Kaons Purity

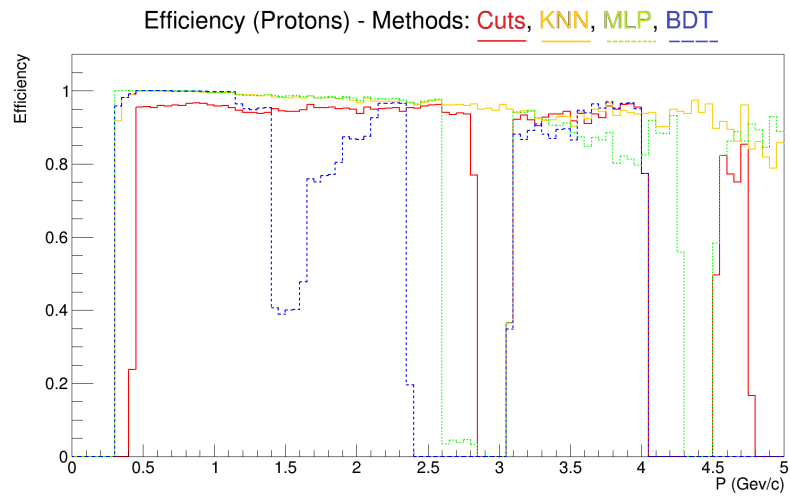


Figure 20: Protons Efficiency

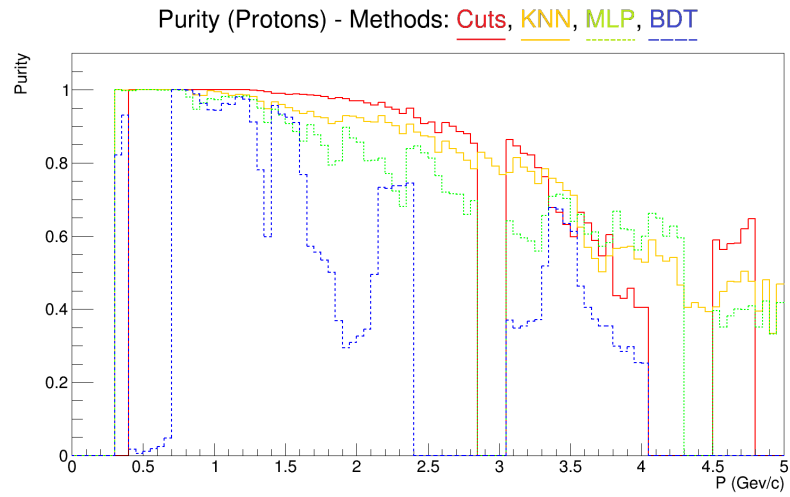


Figure 21: Protons Purity

7 Conclusions

The general results of using MultiVariate analysis in order to improve particle identification at momentum values above 1.8 GeV/c, are not yet promising, even though the results from step 2 would be needed for a firm conclusion. As discussed above, only the KNN method showed an improvement of 20% for purity in the case of kaons. There is still room for improvement. As discussed, the methods were used only with the default settings for the parameters. The parameters can be tuned in order to improve separation. During this project, only the TPC and TOF detectors were used. Using data from other detectors is another way by which the analysis can be improved. The gaps seen in the plots as a consequence of a failed classification phase can be eliminated by either increasing the number of momentum bins, the statistics or by choosing different cuts for each momentum bin. Even so, there might be some cases in which the signal is not well separated from the background, and no suitable cut can be chosen. A problem encountered when increasing statistics was that the running time of the analysis increased to impractical values.

References

- [1] 2008, JINST 3 S08002
- [2] The quest for the quark-gluon plasma, NATURE—Vol 448—19 July 2007—doi:10.1038/nature06080
- [3] Production of light flavour hadrons and anti-nuclei at the LHC, Doctoral Thesis by Alexander Kalweit
- [4] Annu. Rev. Nucl. Part. Sci. 2011. 61:281-321
- [5] TMVA Users Guide: arXiv:physics/0703039 [Data Analysis, Statistics and Probability]