

Separating Signal from Background in $D^0 \rightarrow K_S^0 \pi^+ \pi^-$

Raymond Tat

August 23, 2018

Abstract

The decay of $D^0 \rightarrow K_S^0 \pi^+ \pi^-$ provides a unique opportunity for precise measurements of CP violation and mixing in the charm sector. We analyze $D^0 \rightarrow K_S^0 \pi^+ \pi^-$ candidates from LHCb run 2 (2016-2017) using the sPlot technique and multivariate analysis (MVA) in order to improve signal significance. After investigating a variety of different MVA algorithms, we conclude that a multilayer perceptron has the best performance in terms of signal significance, but uniform gradient boosting may reduce systematic uncertainties. In this process, we also find the signal and background yields for 2016 and 2017 for future sensitivity studies.

1 Introduction

Cosmological observations indicate that the universe is composed almost entirely of matter. The preponderance of matter over antimatter implies that there must exist processes which violate the combined charge-conjugation and parity (CP) symmetry. While the Standard Model predicts CP violation, the amount it predicts is insufficient to account for the observed matter-antimatter asymmetry. For this reason, searching for CP violation beyond the Standard Model is one of the primary goals of the LHCb experiment, as such searches will be a sensitive probe of new particles and interactions.

The $D^0 \rightarrow K_S^0 \pi^+ \pi^-$ decay is particularly useful for studying CP violation for two reasons. First, while CP violation has been observed in both the strange sector and the beauty sector, to date no CP violation has been observed in the charm sector. Second, this particular decay is unique in that it allows for measurements of all four CP violation and mixing parameters in the charm sector. An analysis of decay time versus Dalitz coordinates will lead to experimental constraints on all four of these parameters.

2 Neutral Meson Mixing and CP Violation

Neutral mesons can transition into their antiparticles in a process called mixing. This occurs because the physical eigenstates are a superposition of the flavour eigenstates. Mathematically, this can be expressed as

$$|D_{1,2}\rangle = p|D^0\rangle + q|\bar{D}^0\rangle \quad (1)$$

where $|D_{1,2}\rangle$ are the physical eigenstates and $|p|^2 + |q|^2 = 1$. CP violation occurs if $q/p \neq 1$. This leads to two real-valued parameters which describe CP violation: $|q/p|$ and $\arg(q/p) = \phi$. The effective Hamiltonian describing mixing is $\mathcal{H} = M - i\Gamma$, where M and Γ are the mass and decay matrices. The eigenvalues of this Hamiltonian can be expressed as $\lambda_{1,2} \equiv m_{1,2} - i\Gamma_{1,2}/2$, where $m_{1,2}$ and $\Gamma_{1,2}$ are the masses and decay widths, respectively. The time evolution of the physical eigenstates is then given by

$$|D_{1,2}(t)\rangle = e^{-i\lambda_{1,2}t} |D^0(0)\rangle \quad (2)$$

$$= e^{-im_{1,2}t} e^{-\frac{\Gamma_{1,2}}{2}t} |D^0(0)\rangle \quad (3)$$

In the charm sector, it is useful to define the dimensionless mixing parameters x and y as follows:

$$x = \frac{m_1 - m_2}{\Gamma} \quad (4)$$

$$y = \frac{\Gamma_1 - \Gamma_2}{2\Gamma} \quad (5)$$

$$\text{where } \Gamma = \frac{\Gamma_1 + \Gamma_2}{2} \quad (6)$$

There are thus two mixing parameters, x and y , as well as two CP violation parameters, $|q/p|$, and ϕ . By analyzing decay time versus the Dalitz coordinates, we can use the $D^0 \rightarrow K_S^0 \pi^+ \pi^-$ to measure all four of these parameters.

3 Flavour Tagging and Backgrounds

In order to distinguish between D^0 and $\bar{D}^0$, it is necessary to tag the production flavour of the D^0 meson. In our analysis, we do so using the decay $D^{*+} \rightarrow D^0 \pi^+$. The charge of the soft pion determines the production flavour of the D^0 meson. Flavour tagging in this manner introduces backgrounds, however, as unrelated pions can be incorrectly paired with the D^0 meson during reconstruction. These incorrectly reconstructed events, along with various other combinatoric backgrounds which mimic the $D^0 \rightarrow K_S^0 \pi^+ \pi^-$ decay, contribute to a background that appears as a threshold function below a sharp signal peak in the distribution of the difference between the reconstructed invariant masses of the D^* and D^0 , as shown in figures 1a and 1b. This mass difference is referred to as the delta mass.

4 Improving Signal Significance with sPlot and Multivariate Analysis

The goal of this project is to increase the signal significance, defined as $S/\sqrt{S+B}$, where S and B are the signal and background yields, respectively. This will minimize statistical uncertainty on the measurement. We will use the sPlot technique to determine signal and background distributions of the variables in the input dataset. We will then train a multivariate analysis (MVA) algorithm to discriminate between signal and background using these variables.

4.1 Multivariate Analysis (MVA)

Multivariate analysis is the process of combining information from many weakly discriminating variables into a more powerful discriminator. There are a large variety of MVA algorithms, a sample of which were tested in this project. Two MVA methods in particular - boosted decision trees and multilayer perceptrons - were particularly effective in distinguishing signal and background in the given dataset.

4.1.1 Boosted Decision Trees (BDT)

A decision tree is a learning algorithm which works by repeatedly splitting input data on a single variable until a minimum data size or maximum depth is reached. Usually, the variable and value at which to split the data are chosen so as to maximize the signal and background separation after the split. Normally, decision trees tend to be weak learners. The discriminating power of decision trees can be greatly improved with a process called boosting. With boosting, one trains a series of shallow decision trees. Each decision tree is trained on a reweighted data sample, where the new weights are chosen so that each successive decision tree prioritizes samples which were misclassified by previous decision trees.

4.1.2 Multilayer Perceptrons (MLP)

A multilayer perceptron is a type of neural network which consists of nodes organized into a series of layers, where a matrix of weights connects each layer of nodes to its successor. Input variables are then fed into

the first layer, and the output is evaluated at the final layer. The goal in training a MLP is to find values of weights such that when input variables are fed into the first layer, the output value most closely mimics the correct classification. There are a number of ways to train neural networks, but in this project we use the AdaDelta algorithm as implemented in `hep_ml`.

5 Analysis Overview

This analysis can be divided into six steps.

1. Fit signal and background distributions to the delta mass distribution.
2. Extract signal and background yields from the fit results.
3. Using `sPlot`, find the signal and background distributions for other variables in the dataset.
4. Based on the distributions output by `sPlot`, choose some subset of the variables in the dataset to be used as inputs for MVA.
5. Train multiple MVA methods on the `sWeighted` dataset. Each event is fed into each MVA method twice: once as signal with signal `sWeight`, and once as background with background `sWeight`.
6. Evaluate MVA methods to determine how much each method increases the signal significance.

In the following sections, these steps will be discussed in greater detail.

6 Fitting Signal and Background Distributions

Using `RooFit`, we performed an extended likelihood fit of the delta mass distribution of 2016 LL and 2017 LL datasets. The “LL” (long-long) designation refers to events in which the K_S^0 decays inside the Vertex Locator. The purpose of this fit is twofold. First, an extended likelihood gives the signal and background yields of a given dataset. These yields will be used in future sensitivity studies and extrapolations. Furthermore, the results of this fit will be used as input to the `sPlot` technique, as `sPlot` requires knowledge of the signal and background distributions for at least one variable in the dataset.

6.1 Signal and Background Yields

Figure 1 displays the signal and background distributions for the raw 2016LL and 2017LL datasets collected during LHCb run 2. The data here represent the complete sample of events collected using trigger requirements designed to identify this decay channel. After testing a variety of different signal and background models, it was found that a signal model consisting of two Crystal Ball functions and a Gaussian performed well. The background is modeled by the function $A(x-d)^{1/3} + B(x-d) + C(x-d)^{3/2}$, where x is the delta mass and A , B , C , and d are free parameters of the fit.

The corresponding yields (without any preselection cuts or MVA selection) are given in the table below.

Data sample	Signal yield	Background yield	Significance
2016LL	$3\,312\,322 \pm 8\,758$	$34\,015\,778 \pm 8\,882$	557.3 ± 1.4
2017LL	$3\,085\,761 \pm 3\,313$	$7\,146\,695 \pm 3\,881$	964.7 ± 0.9

Note that the 2016 dataset has higher background yields. This is to be expected, as the trigger requirements were tightened in 2017.

7 Preselection Cuts

Before applying `sPlot` or MVA, some loose preselection cuts are applied to the data. These are cuts which are known to have high signal efficiency and background rejection.

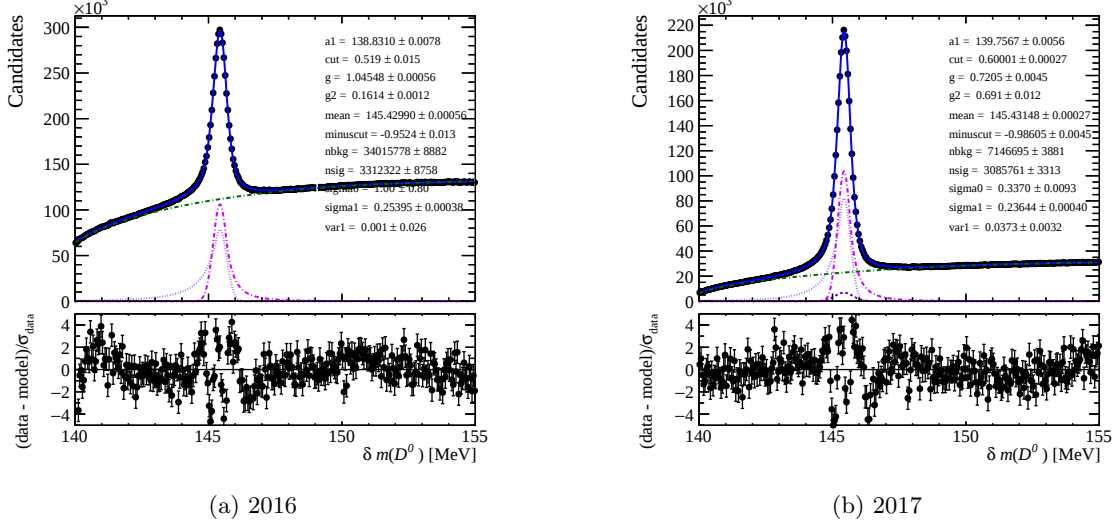


Figure 1: Raw yields and fit

7.1 Trigger Requirements

The following trigger requirements are enforced in the analysis that follows.

- D0_L0HadronDecision_TOS or D0_L0Global_TIS
- D0_Hlt1TrackMVADecision_TOS or D0_Hlt1TwoTrackMVADecision_TOS (only enforced for 2017LL as high level triggers were not saved in the available dataset)

7.2 Other Requirements

Along with the trigger requirements, a number of other preselection cuts are enforced.

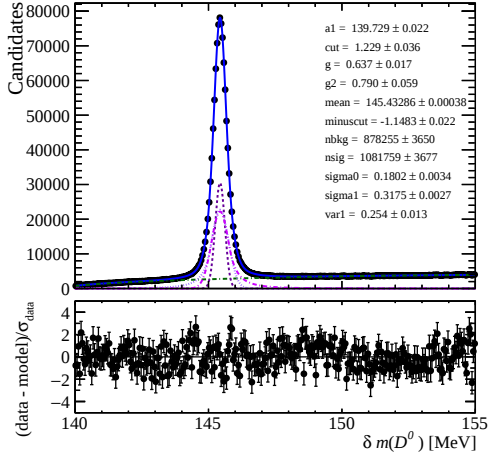
- $D0_pim_ProbNNghost < 0.8$ & $D0_pip_ProbNNghost < 0.8$ (π^+ and π^- ghost probability < 0.8)
- $D0_pim_ProbNNpi > 0.1$ & $D0_pip_ProbNNpi > 0.1$ (π^+ and π^- pion probability > 0.1)
- $1840 \text{ MeV} \leq m(D^0) \leq 1890 \text{ MeV}$
- $480 \text{ MeV} \leq m(K_S^0) \leq 515 \text{ MeV}$
- $D0_IPCHI2_OWNPV < 9$ ($IP\chi^2$ of D^0 primary vertex < 9)
- $D0_LoKi_DTF_CHI2 > 0$ (Decay tree fitter $\chi^2 > 0$)

7.3 Yields after preselection

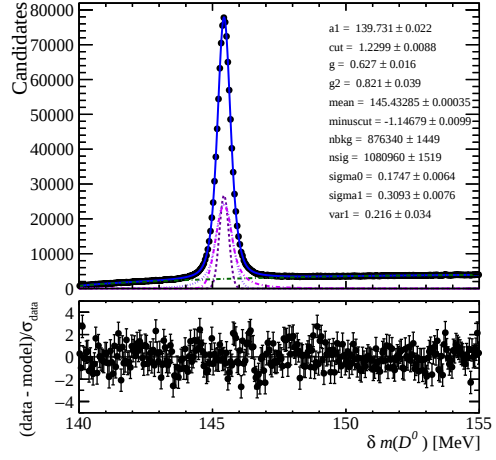
The fits after applying preselection cuts are shown in figure 2. At this point, data is split into even and odd numbered events in preparation for multivariate analysis.

The signal and background yields after preselection cuts are given in the following table (errors are added in quadrature).

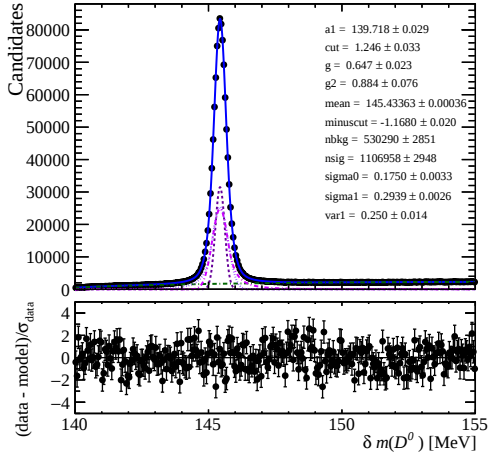
Data sample	Signal yield	Background yield	Signal Significance
2016LL	$2\,162\,719 \pm 3\,978$	$1\,754\,595 \pm 3\,927$	$1\,092.7 \pm 1.6$
2017LL	$2\,211\,127 \pm 3\,813$	$1\,060\,540 \pm 3\,661$	$1\,222.4 \pm 1.6$



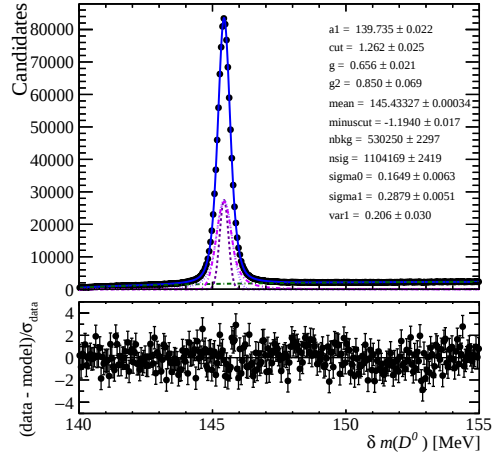
(a) 2016 odd events



(b) 2016 even events



(c) 2017 odd events



(d) 2017 even events

Figure 2: Fitted delta mass distributions after preselection cuts.

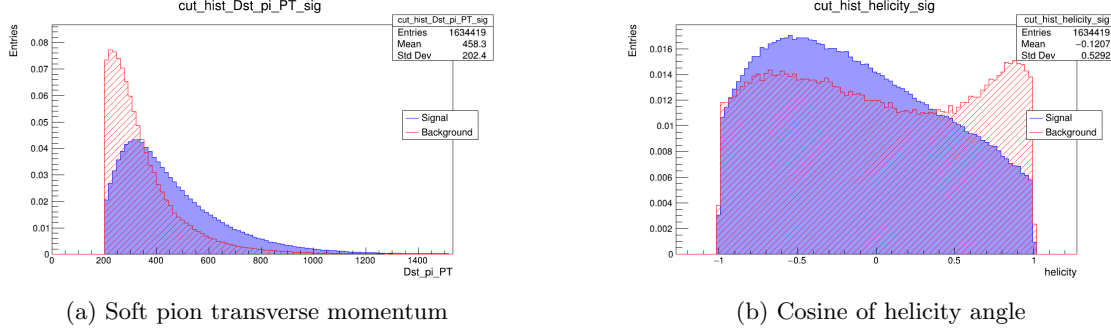


Figure 3: Signal (blue) and background (red) distributions of soft pion transverse momentum and cosine of the helicity angle.

8 Applying the sPlot technique

8.1 The sPlot Technique

sPlot [2] is a statistical tool that unfolds the contributions of multiple sources of events in a data sample. Given a dataset with multiple sources of events, where the distribution of at least one variable is known for each source of events, sPlot can find the distributions of other variables in the dataset for each source of events. For our purposes, the sources of events are signal and background, and so sPlot computes the signal and background distribution of variables in the dataset. In order to use sPlot, one first computes an extended likelihood fit in order to find the signal and background yields. Based on the results of the fit, sPlot assigns two weights to each event: a signal weight and a background weight. These weights are referred to as sWeights. sPlot then uses these sWeights to find the signal and background distributions of all other variables in the dataset. Sample signal and background distributions are shown for soft pion transverse momentum and cosine of the helicity angle in figure 3.

sPlot serves two purposes in this analysis. First, sPlot gives the signal and distribution of variables in the input data. This is useful as it allows us to select input variables for MVA which have large separation between their signal and background distributions. Second, the sWeights that sPlot computes can be used to train an MVA algorithm.

8.2 Selecting Input Variables

We use a number of criteria in order to select input variables for MVA. These criteria are listed below.

- Input variables should have signal and background separation. This is determined using the distributions output by sPlot.
- Input variables should be uncorrelated with the delta mass, as the sPlot technique distorts distributions for variables that are strongly correlated with the fit variable.
- Input variables should be uncorrelated with decay time or Dalitz variables, as correlations would distort the signal distribution in later analysis.
- Input variables should not introduce charge asymmetry. In practice, this means that variables pertaining to the π^+ meson should be paired with the corresponding variable pertaining to the π^- meson.

Using these criteria, the following eight variables were selected from the approximately 300 variables in the original dataset for use as inputs for training MVA methods.

- D^0 end vertex χ^2
- Sum of transverse momenta of π^+ and π^-
- Cosine of helicity angle

- Soft pion transverse momentum
- Soft pion primary vertex χ^2
- K_S^0 vertex z error
- D^0 momentum
- K_S^0 transverse momentum

9 MVA Training and Results

9.1 Training Process

As mentioned previously, the sPlot technique computes a signal sWeight and a background sWeight for each event. In order to train an MVA algorithm using the output of sPlot, each event is fed into the algorithm twice: once as signal with signal sWeight, and once as background with background sWeight. The algorithm then learns to distinguish between signal-like and background-like regions in phase space. This data-driven approach to MVA is described in *Advanced event reweighting using multivariate analysis* [1]. In order to counteract the effects of overfitting, the 2016LL and 2017LL events are further split into odd and even events, with a different classifier trained on each. After training, the classifiers trained on even data are evaluated on odd data, and vice versa.

9.2 Performance Evaluation

In order to help evaluate MVA performance, we use a graph called a receiver operating characteristic (ROC) curve. A ROC curve plots the signal efficiency against background rejection for a given MVA method. The area under the ROC curve can be used as a metric of MVA performance, where a larger area indicates greater discriminating power.

Of all the methods tested, the hep_ml MLP had the highest overall performance in terms of ROC curve area, with an integrated ROC curve area of 0.85 on 2016 data, and an area of 0.78 on 2017 data. The MLP had two hidden layers, one with 8 and one with 4 hidden layers. As the time needed to train an MLP appeared to scale poorly with data size, the MLP was trained on a random subample of 40000 events. The ROC curve for several MVA algorithms are shown in 4. Only ROC curves for algorithms trained on odd events are shown, as those for even events are nearly identical.

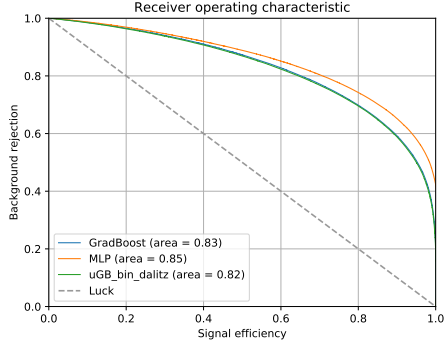
The fact that performance is lower for 2017 data is to be expected, as the trigger requirements are tighter on 2017 data, and therefore there is less improvement for MVA to make. After applying the MLP to 2016 and 2017 dataset, the amount of background is reduced, as shown in figures 5a and 5b.

The signal and background yields after MVA selection by the MLP are given in the following table (errors are added in quadrature).

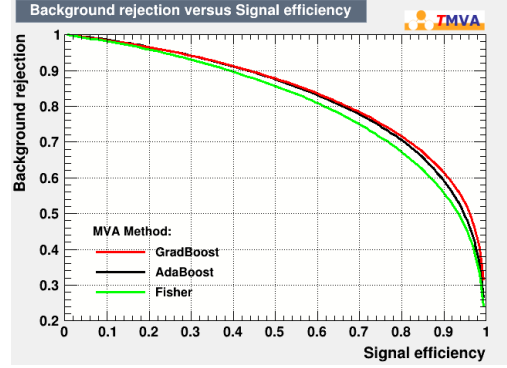
Data sample	Signal yield	Background yield	Significance
2016LL	$2\,039\,264 \pm 1\,876$	$899\,967 \pm 1\,545$	$1\,189.5 \pm 0.8$
2017LL	$2\,185\,377 \pm 2\,326$	$872\,708 \pm 2\,103$	$1\,249.7 \pm 1.0$

9.3 Uniformity

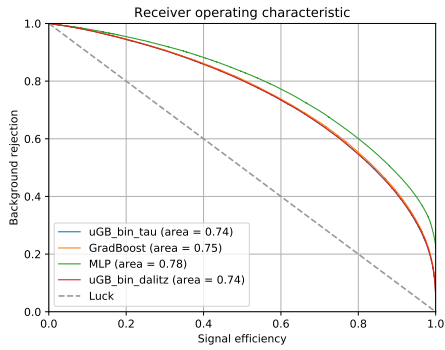
Signal efficiency and background rejection are not the only considerations in selecting an MVA algorithm, however. It is also important to consider whether MVA selection distorts the signal distribution in decay time or Dalitz variables, as this would introduce systematic uncertainties in future analysis. For this reason, uniform gradient boosting (uGB) was also explored as a possible MVA algorithm. Uniform gradient boosting is an algorithm which attempts to enforce signal efficiency in certain variables by incorporating uniformity into its loss function. As a result, uGB generally achieves lower performance than other algorithms, but in return the algorithm can achieve greater uniformity in key variables. In figure 6, the signal efficiency as



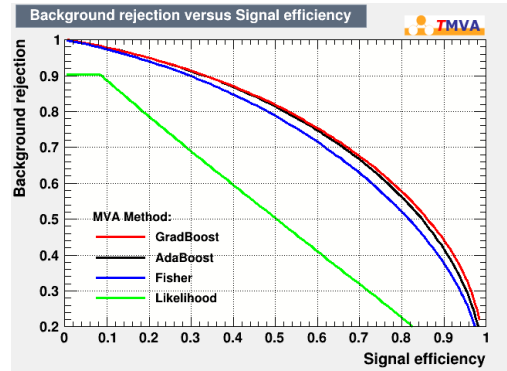
(a) sklearn and hep.ml algorithms trained on 2016LL



(b) TMVA algorithms trained on 2016LL

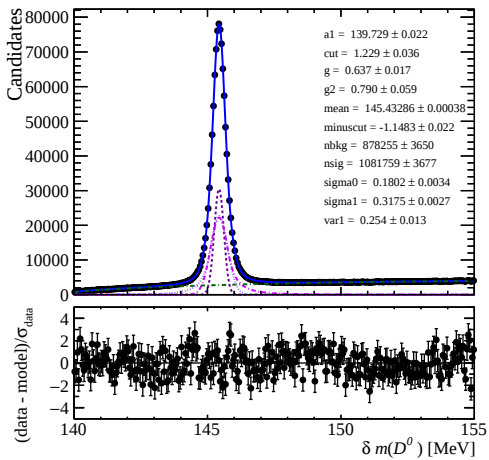


(c) sklearn and hep.ml algorithms trained on 2017LL

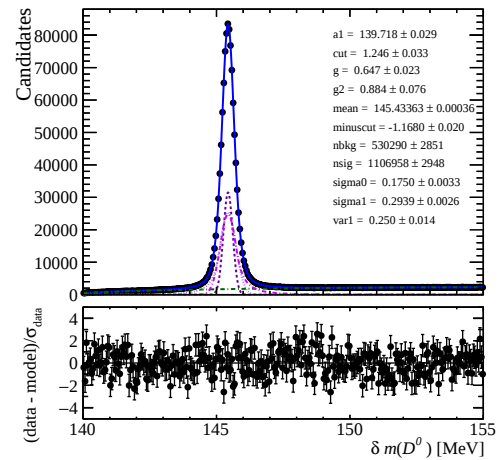


(d) TMVA algorithms trained on 2017LL

Figure 4: ROC curves for tested MVA algorithms



(a) 2016LL dataset



(b) 2017LL dataset

Figure 5: Delta mass distributions after selection with MLP (odd events only).

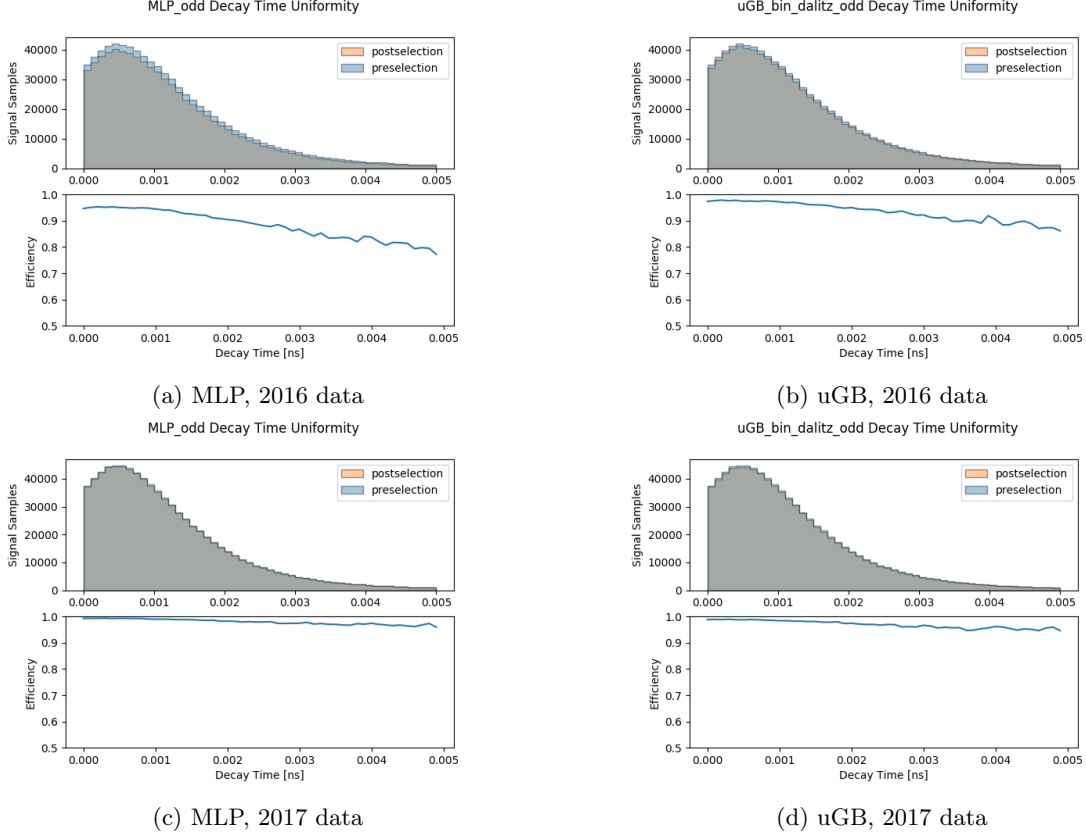


Figure 6: Signal efficiency with respect to decay time for MLP and uGB

a function of decay time are plotted for the MLP and uGB algorithms, both for 2016 and 2017. While it appears that uGB improves uniformity in the 2016 dataset, this effect is nonexistent for the 2017 dataset. This may be due to the fact that signal efficiency is close to unity for the 2017 dataset, and thus uGB can make no significant improvement to the uniformity. More research will need to be conducted in order to determine whether the increased uniformity is worth the loss in performance suffered by uGB.

10 Conclusion

To aid in the analysis of the $D^0 \rightarrow K_S^0 \pi^+ \pi^-$ decay, this project set out to improve the signal significance of the available datasets using multivariate analysis. To accomplish this without the need for simulations, the sPlot technique was used to determine the signal and background distributions for a variety of possible MVA input variables. This required testing a variety of signal and background models in order to achieve a good extended likelihood fit. Using the output of the sPlot technique, a number of MVA algorithms were tested to determine their effectiveness in improving signal significance. It was found that the multilayer perceptron was the most successful in achieving this goal. However, the multilayer perceptron also had nonuniform selection efficiency in decay time, particularly for the 2016 dataset. Another algorithm tested called uniform gradient boosting sacrifices signal significance, but in return achieves more uniform selection efficiency in decay time. Further research is needed in order to determine which of these two algorithms best balances signal significance against systematic uncertainties.

References

- [1] D. Martschei et al. “Advanced event reweighting using multivariate analysis”. In: *Journal of Physics. Conference* 368 (2012).
- [2] Muriel Pivk and Francois R. Le Diberder. “sPlot: a statistical tool to unfold data distributions”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 555.1-2 (Feb. 17, 2004), pp. 356–369. DOI: 10.1016/j.nima.2005.08.106. URL: <https://arxiv.org/abs/physics/0402083>.