Final report

Study of a new selection strategy for $B^0 \to p\bar{p}K^+\pi^-$ decays

Name: Ganrong Wang

Email: ganrong.wang@cern.ch

Time: Summer ,2019

Supervisor: Giovanni Cavallero，

Frederic Teubert

Contents

Abstract: The study of a new selection strategy for $B^0 \to p\bar{p}K^+\pi^-$ decays is performed by using pp collision data recorded by the LHCb experiment at a centre-of-mass energy of 13 TeV in the 2016 data-taking year. Multivariate analysis (MVA) techniques are used to suppress the backgrounds. In order to introduce the particle identification (PID) variables in the MVA algorithms, known discrepancies between simulation and data must be taken into account. A procedure to correct the simulated PID distributions, using PID calibration samples from data, namely the PIDCorr procedure, is used. These corrected PID variables are used to train the MVA classifiers. The output of the MVA is optimised to obtain the maximum sensitivity. The results of the new selection strategy are compared with the previous analysis.

Key Words：B Physics；LHCb Detector；Monte Carlo Simulation; PIDCorr；

1. Introduction

1.1 Physics background and research motivation

Hadrons beyond conventional mesons and baryons have been foreseen since the formulation of the constituent quark model[1] in 1964. However, up to 2003, only conventional mesons ($q\bar{q}$) and baryons (qqq) have been observed. Since the discovery of the X(3872) state[2], several non-conventional states, also called exotic hadrons, have been observed. These candidates have either unexpected masses, decay properties or electric charge with respect to the conventional hadrons. Since the nature of these states is still not understood, it's important to search for new exotic candidates and for new decay modes of already observed non-conventional states.

In 2018, the LHCb collaboration reported an evidence for an $\eta_c(1S)\pi^-$ resonance studying $B^0 \to \eta_c(1S)(\to p\bar{p})K^+\pi^-$ [3] decays, using the data collected in 2011, 2012 and 2016 data-taking years. The corresponding Feynman diagrams are shown in Figure 1. The $\eta_c(1S)\pi^-$ mass spectrum obtained in the previous analysis is shown in Figure 2, showing an unexpected contribution with a mass around 4100 MeV and a width of approximatively 150 MeV.

This project is aiming to find possible improvements in the $B^0 \to p\bar{p}K^+\pi^-$ candidates selection, in order to study the feasibility to reach the observation level (if the Z_c^- state will be confirmed) when using the full LHCb dataset.

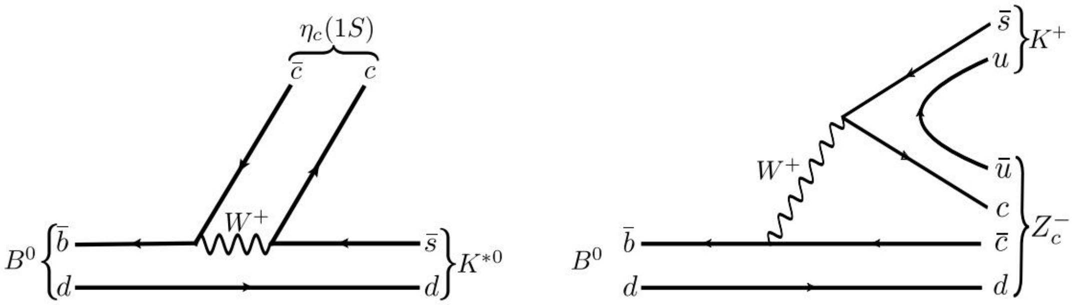


Figure 1 The Feynman diagram of the decay

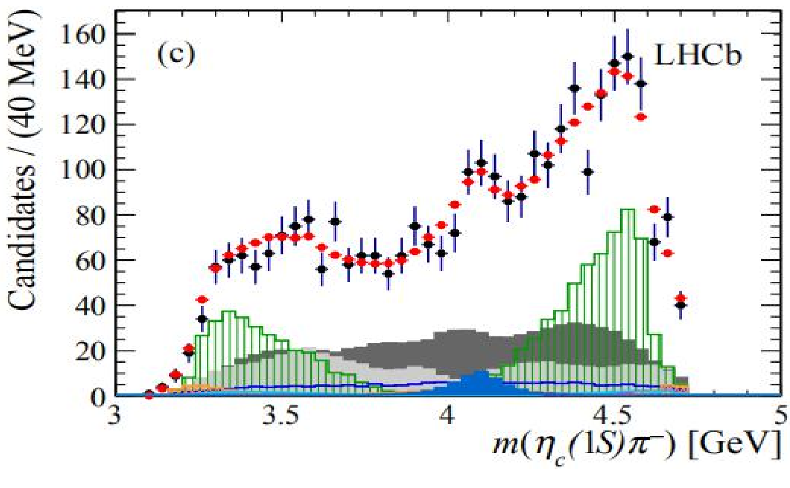


Figure 2 fit $m\left(\eta_c(1S)\pi^-\right)$

1.2 Research strategy

In the previous analysis, multivariate analysis (MVA) techniques have been used, exploiting kinematical and topological variables. Rectangular cuts have been used for particle identification (PID) variables: pProbNNp>0.25, kProbNNk>0.25 and piProbNNpi>0.2.

Since $B^0 \to \eta_c(1S)(\to p\overline{p})K^+\pi^-$ decays have a purely hadronic final state, the charged hadrons identification is very important to discriminate between signal and background. The objective of this project is to improve the selection efficiency by introducing the PID variables into the MVA classifiers.

The MVA variables used in the previous analysis are summarized in Sec. 2. The results of the correction of simulated PID variables are given in Sec. 3, together with the MVA results of the new selection strategy. The optimisation of the MVA output is reported in Sec. 4, and the comparison with the previous results is given in Sec. 5.

2. Multivariate Analysis

2.1 Introducing TMVA

The MVA classifiers[4] have significantly evolved in recent years. MVA classification methods based on machine learning techniques have become a fundamental ingredient in most analyses. Integrated into the analysis framework ROOT, TMVA is a toolkit which hosts a large variety of multivariate classification algorithms, such as boosted decision trees (BDT), multilayer perceptron (MLP), and boosted decision trees with gradient boost (BDTG). Classification and regression analyses in TMVA have training, testing and evaluation phases.

2.2 Data sample

The data used in this project includes collision data and Monte Carlo (MC) samples. This project use pp collision data which have been recorded in 2016 by the LHCb experiment at a centre-of-mass energy of 13TeV. The acquisition and processing of LHCb data includes two steps: online and offline. When the products of a pp collision meet certain trigger conditions, it will be recorded for offline processing and analysis. After the reconstruction, the events can be further selected by a process called stripping. This project uses the StrippingB2pphh_kpiLine and the stripping version is 28r1. The stripping line requirements are reported in Table 1.

Table 1 stripping selection

Variable	Run1	Run2
Kaons		
p	$> 1500\,\mathrm{MeV}$	
$p_\perp$	$> 300\,\mathrm{MeV}$	
Track χ^2/ndof	< 3.0	$< 4.0^{(*)}$
Track Ghost Prob	< 0.35	$< 0.4^{(*)}$
ProbNNk	> 0.05	
IP χ^2	> 4.0	> 5.0
Pions		
p	$> 1500\,\mathrm{MeV}$	
$p_\perp$	$> 300\,\mathrm{MeV}$	
Track χ^2/ndof	< 3.0	$< 4.0^{(*)}$
Track Ghost Prob	< 0.35	$< 0.4^{(*)}$
ProbNNπ	> 0.05	> 0.1
IP χ^2	> 6.0	> 8.0
Protons		
p	$> 1500\,\mathrm{MeV}$	
$p_\perp$	$> 300\,\mathrm{MeV}$	
Track χ^2/ndof	< 3.0	$< 4.0^{(*)}$
Track Ghost Prob	< 0.35	$< 0.4^{(*)}$
ProbNNp	> 0.05	
IP χ^2	> 2.0	> 3.0

Variable	Run1	Run2
$p\bar{p}$ combination		
$M_{p\bar{p}}$	$< 5000\,\mathrm{MeV}$	
$\sum p_\perp$	$> 750\,\mathrm{MeV}$	
$\sum p$	$> 700\,\mathrm{MeV}$	
Minimum $p_\perp$	$> 400\,\mathrm{MeV}$	
Minimum p	$> 4000\,\mathrm{MeV}$	
ProbNNp·ProbNN$\bar{p}$	> 0.05	> 0.01
χ^2 DOCA	< 20	
$p\bar{p}K^+$ combination		
$M_{p\bar{p}K^+}$	$< 5600\,\mathrm{MeV}$	
χ^2 DOCA	< 20	
B^0		
Max DOCA	$< 0.3\,\mathrm{mm}$	$< 0.25\,\mathrm{mm}$
χ^2 DOCA	< 20	
DIRA	> 0.9999	
Vertex χ^2	< 30	< 25
Min IP	$< 0.2\,\mathrm{mm}$	
$p_\perp$	$> 1000\,\mathrm{MeV}$	
$\sum p_\perp$ of the daughters	$> 3000\,\mathrm{MeV}$	
Mass range	[5.05,5.55] GeV	
Number of Long Tracks	< 200	

Collision events need to satisfy three trigger stages in sequence to be saved on disk. The first stage is the hardware level trigger (L0), which is used to select events containing particles with high transverse momentum and energy. The second and third stages are software level triggers, called High Level Trigger (HLT) 1 and 2. Each stage is divided into several trigger lines according to different online selection requirements. Table 2 shows the the trigger lines used in this analysis.

Table 2 trigger selection

	Run1	Run2
L0	Physics_TIS || Hadron_TOS	
Hlt1	TrackAllL0_TOS	(Track || TwoTrack)MVA_TOS
Hlt2	Topo(2 || 3 || 4)Body_TOS	

2.3 MVA discriminating variables of the previous analysis

The data samples used as the input to TMVA for classification training and testing, are simulated $B^0 \to p\bar{p}K^+\pi^-$ events for the signal and the upper sideband of the $m(p\bar{p}K^+\pi^-)$ distribution from data as the background sample. The specific sources of signals and background are shown in Table 3. Different kinematical and topological variables were used in the previous analysis to train the algorithms, shown in Table 4.

Table 3 dataset for TMVA Training

sample	**source**
signal sample	$B^0 \to p\bar{p}K^+\pi^-$ 2016 MC
background sample	candidates from the high-mass 2016 data sideband (5450 MeV< $m(p\bar{p}K\pi)$ <5550MeV)

Table 4 kinematical and topological variables

variables	**Definition**
B_LOKI_MIPCHI2DV_PRIMARY	minimum IP χ^2 of the B^0 to the PV
B_AMAXDOCA	maximum DOCA of the tracks to the B^0 vertex
log(B_BPVVDCHI2)	log of the χ^2 distance of the B^0 vertex from the PV
B_LOKI_POINTING	defined as $\dfrac{\lvert\vec{p}_B\rvert\sin\theta}{\lvert\vec{p}_B\rvert\sin\theta+\sum_i\lvert\vec{p}_i\rvert\sin\theta_i}$
B_LOKI_VFASPF_VCHI2	χ^2 of the B^0 vertex
log(B_pPTmin)	log of the minimum pT between the two protons
log(B_pPTmax)	log of the maximum pT between the two protons
log(B_pPmin)	log of the minimum p between the two protons
log(B_pPmax)	log of the maximum p between the two protons

There are many different analysis methods in the TMVA package. In this project, BDT, BDTG and MLP are used for consistency the MVA procedures are repeated in this projects for the set of variables shown in Table 4 . From the background rejection vs signal efficiency graphs of the three methods (Figure 3), it can be seen that the results of different methods are similar. Finally, BDT method is selected for application.

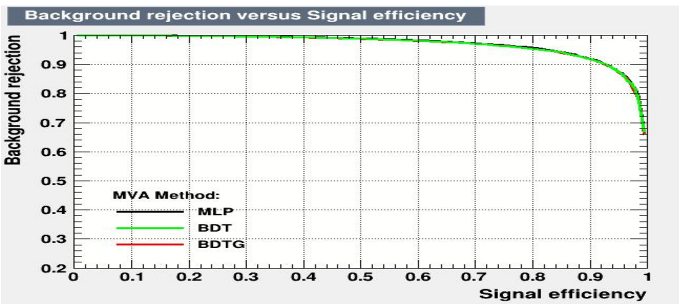


Figure 3 ROC curves for the BDT、BDTG、MLP algorithms used in the 2016 MVA analysis.

3. Correcting the PID variables

3.1 Correction of simulated PID variables

The main concept of the project is that adding PID variables for classification training a better discrimination between signal and background can be achieved, resulting in an improved selection efficiency. Because of the technical challenges in the simulation related to the conditions affecting the performance of the RICH detectors, such as the occupancy, temperature and pressure conditions, there are known discrepancies in the PID distributions between MC and data. For this reason it is necessary to correct the simulated PID variables before training the MVA classifiers.

The most widely used tool is PIDCalib[5] ,which uses the clean and high-statistics calibration samples $D^{*+} \to D^0\pi^+, \Lambda \to p\pi^-, \Lambda_c^+ \to pK^+\pi^-$ for hadrons, $J/\psi \to \mu^+\mu^-$ for muons, $B^+ \to J/\psi\left(\to e^+e^-\right)K^+$ for electrons).

Two methods are available within the PIDCalib package: PIDGen and PIDCorr. PIDGen uses a 4D distribution (PID,p,eta,nTracks) and generate a new PID variable matching data for any value of p, eta and nTracks. PIDCorr additionally preserves the correlations between different hypotheses (a track with a preferred hypothesis to be a proton (ProbNNp -> 1) has a disfavoured hypothesis to be a pion (ProbNNpi-> 0)). This project used the PIDCorr procedure, applying it to the ProbNN variables (MC15TuneV1) of all the tracks and for the proton, kaon and pion hypotheses.

The comparison between some corrected (red line) and uncorrected (blue line) PID distributions is reported in Figure 4.

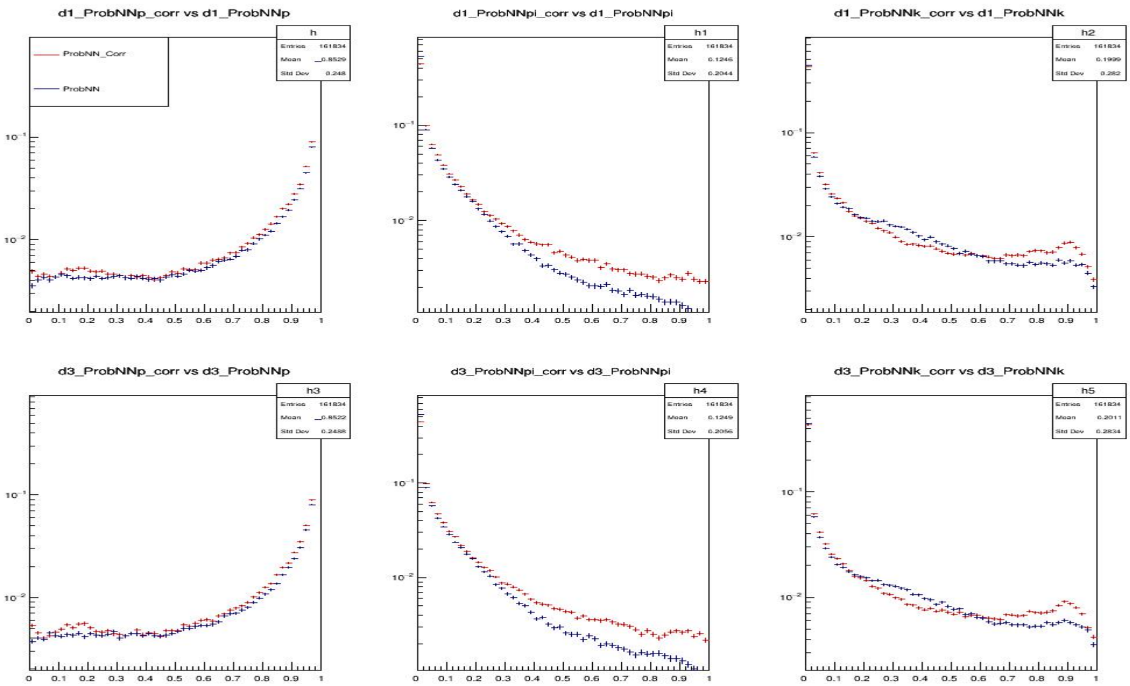


Figure 4 Compare the correct PID variable（red line） with the uncorrect PID variable (blue line)

3.2 Usage of the corrected PID variables

Three sets of discriminating variables have been used to train the MVA algorithms, and are reported in Table 5. The first one (in black) corresponds to the set used in the previous analysis, the others (in blue and green) are the additional PID variables that are used for classification in this project.

Table 5

Multivariate offline selection	
B_LOKI_MIPCHI2DV_PRIMARY	log(d1_MC15TuneV1_ProbNNp_corr)
B_AMAXDOCA	log(d3_MC15TuneV1_ProbNNp_corr)
log(B_BPVVDCHI2)	log(d2_MC15TuneV1_ProbNNk_corr)
B_LOKI_POINTING	log(d4_MC15TuneV1_ProbNNpi_corr)
B_LOKI_VFASPF_VCHI2	log(d1_MC15TuneV1_ProbNNk_corr)
log(B_pPTmin)	log(d3_MC15TuneV1_ProbNNk_corr)
log(B_pPTmax)	log(d2_MC15TuneV1_ProbNNp_corr)
log(B_pPmin)	log(d4_MC15TuneV1_ProbNNk_corr)
log(B_pPmax)	

Figure 5 reports the comparison between the PID variables in MC and data, showing that these variables are good for classification. In classification training, 9 variables of black were used for the first time, 13 variables of black and green were used for the second time, and 17 variables of black, green and blue were used for the third time. The results show the Receiver Operating Characteristic (ROC) curves of black, green and blue respectively (Figure 5).

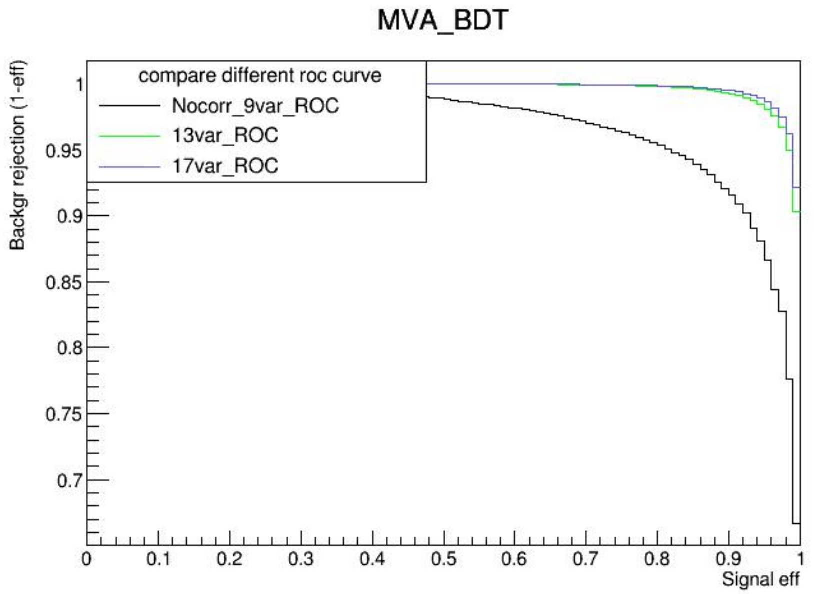


Figure 5 Comparison of the ROC curves obtained with the three sets of input variables

The blue line (17 variables) is closest to the upper right corner. The closer the curve approach this point, the more we can reject the background and receive the signal, the roc curve obtained by training with different variables are illustrate the benefits of using PID variables. The classification with 17 variables provides the best performance. The training results of 17 variables are good, but we need to check whether there is over-training in the process. From Figure 6, we can see that there is no evidence for over-training.

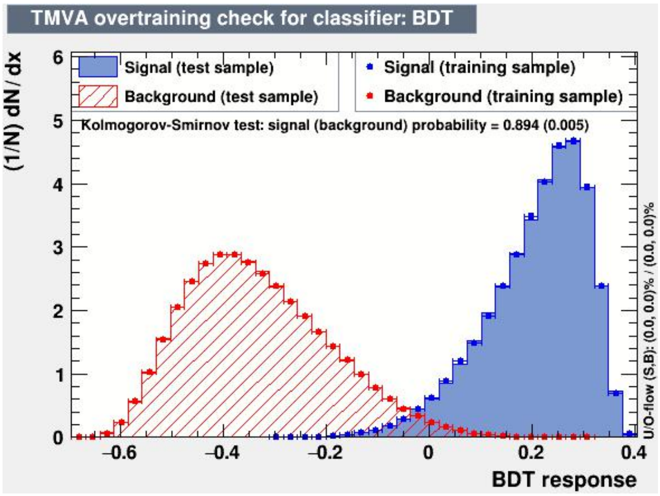


Figure 6 over-training check

4.Optimisation

To find the optimal BDT cut value that maximise the signal yield and the purity in the 2016 data sample a procedure to find the maximum value of the figure of merit defined as

$$F.O.M = \frac{S}{\sqrt{S+B}}$$

has been implemented and applied to data. First, a very loose cut of BDT>-0.1 is applied, in order to find a first estimate for the signal yield (S1) in the ± 3 sigma range around the B^0 peak, through an extended maximum likelihood fit to the $m\left(p\bar{p}K\pi\right)$ distribution in the range 5220< $m\left(p\bar{p}K\pi\right)$ <5340. The B0 signal is parametrised by the Hypatia probability density function, defined as

$$f(m|\mu,\sigma,l,\zeta,\beta,\alpha_1,\alpha_2,n_1,n_2) = \begin{cases} \frac{A}{(B+m-\mu)^{n_1}}, & \text{if } \frac{m-\mu}{\sigma} < -\alpha_1 \\ \frac{C}{(D+m-\mu)^{n_2}}, & \text{if } \frac{m-\mu}{\sigma} > \alpha_2 \\ ((m-\mu)^2+\delta^2)^{\frac{1}{2}l-\frac{1}{4}} \exp\left(\beta(x-\mu)\right) K_{l-\frac{1}{2}}(\alpha\sqrt{(x-\mu)^2+\delta^2}) & \text{otherwise} \end{cases}$$

$$\delta := \sigma\sqrt{\frac{\zeta K_l(\zeta)}{\zeta K_{l-1}(\zeta)}}, \quad \alpha := \frac{1}{\sigma}\sqrt{\frac{\zeta K_l(\zeta)}{\zeta K_{l-1}(\zeta)}}$$

The parameters related to the peak tails are fixed to the values obtained in the fit to the MC sample, as shown in Table 6 and Figure 7. The mass and resolution parameters are left floating in the fit to collision data.

Table 6 Parameter

parameter	value
μ	5281.138 ± 0.036
σ	14.766 ± 0.089
l	-3.315 ± 0.12
ξ	0.006458 ± 0.000050
α_1	2.184 ± 0.042
α_2	2.264 ± 0.058
n_1	2.03 ± 0.21
n_2	2.58 ± 0.36

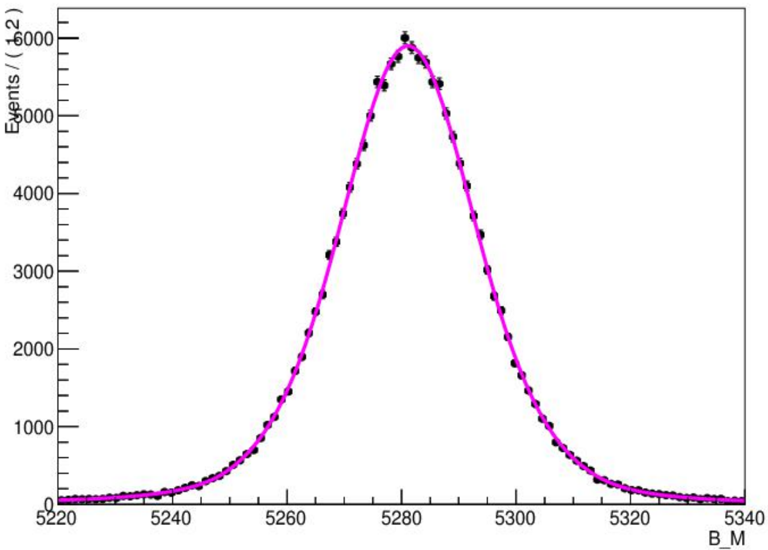


Figure 7 Fit B0 mass in MC sample

An exponential probability density function has been used to parametrise the background. The fit is repeated for different BDT cuts. It allows to obtain the number of background events (B) in the ± 3 sigma region around the B0 peak, as a function of the BDT. The number S of signal events is determined by $S = S_1 * \varepsilon_{MC}(BDT)$, where $\varepsilon_{MC}(BDT)$ is the ratio between the number of events in the MC sample with this BDT cut and the number of events in the MC sample with BDT>-0.1. The BDT cut is varied between -0.1 and 0.3 in steps of 0.02. The FOM is shown in Figure 8. The maximum value of the FOM is obtained for BDT>0.14.

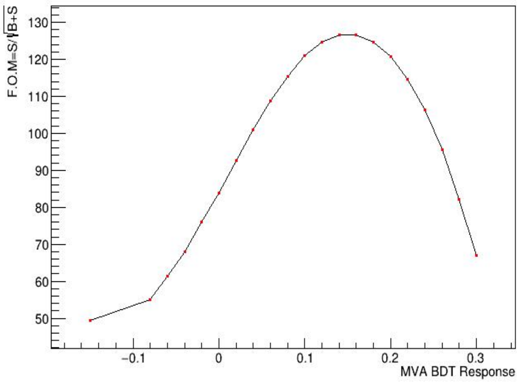


Figure 8 Optimisation

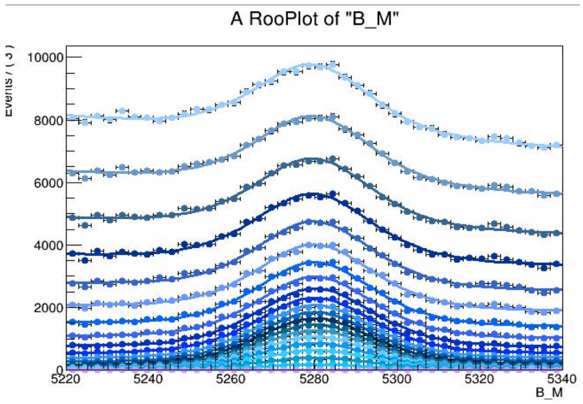


Figure 9 Fit B_M with different MVA_cut

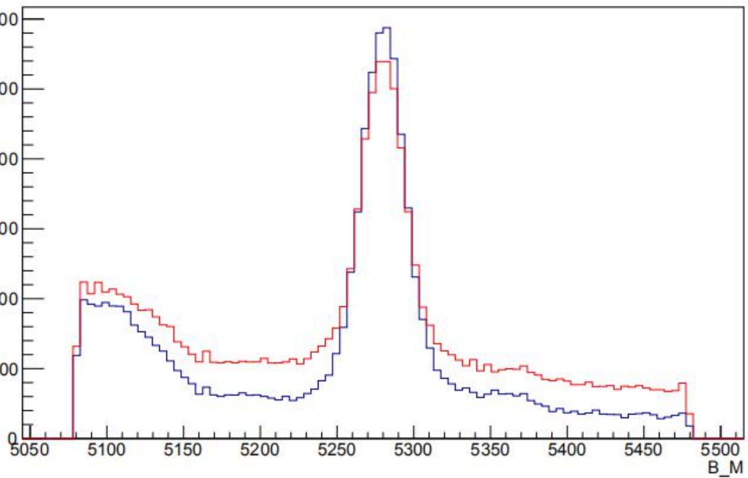
5.Results

The fit to the $m\left(p\bar{p}K\pi\right)$ distribution resulting from the previous selection strategy in the 5220< $m\left(p\bar{p}K\pi\right)$ <5340 gave 17046 ± 233 B^0 signal events and 13697 ± 225 background events. This corresponds to a purity of 1.55 in the ± 3 sigma region around the B^0 peak.

With the new selection strategy, introducing the PID variables in the MVA, the same fit gives 21393 ± 220 B^0 signal events and 7008 ± 185 background events, corresponding to a purity of 3.6 in the ± 3 sigma region around the B^0 peak. The fits are shown in Figure 9.

The improvements achieved with the new selection strategy can also be seen from Figure 10, comparing the old $m\left(p\bar{p}K\pi\right)$ distribution (red histogram), with the new one (blue histogram), in the 5080<$m\left(p\bar{p}K\pi\right)$<5180 range.

In conclusion, introducing the PID variables in the MVA allowed to increase the B0 signal yield of a factor 1.3, to reduce the combinatorial background of a factor 2, and hence to improve the purity of a factor 2.4 in the± 3 regin around the signal peak.



2

Figure 10 comparing the old $m\left(p\bar{p}K\pi\right)$ distribution (red histogram), with the new one (blue histogram)

References

[1] Gell-Mann M. A Schematic Model of Baryons and Mesons[J]. Physics Letters, 1964, Vol: 8(3).

[2] Belle collaboration, S.-K. Choi et al., Phys. Rev. Lett. 91 (2003) 262001

[3] LHCb collaboration, R. Aaij et al., Eur. Phys. J. C78 (2018) 1019

[4] Hoecker A, Voss K, Voss H, et al. TMVA-Toolkit for Multivariate Data Analysis with ROOT: Users guide[R]. 2007.

[5] Anderlini, L., Contu, A., Zhang, Y., Vazquez Gomez, R., Muller, D., Jones, C. R., ... & Otalora Goicochea, J. M. (2016). The PIDCalib package (No. LHCb-PUB-2016-021).

Acknowledgement

I am very happy to have a wonderful summer at CERN. This is a very memorable experience, so I must first thank CERN for giving me this opportunity. Here my supervisor Giovanni Cavallero gave me a lot of help. He is professional and patient. During my completion of my project, he gave me a lot of professional guidance and taught me a lot of skills and knowledge for data analysis, from which I benefit a lot.

At the same time, I want to thank Silvia Gambetta and Thierry Gys. They are very kind and gave me a lot of advice on work and life. Silvia is a very nice woman she will be a role model for me to learn in the future. Thierry knows a lot of things, he is very funny and humorous. I will always remember the plums in his garden that he gives to us. I will remmember the rainy evening when he invited us to eat pizza, the pizza was delicious, everyone was very happy.

Thanks to the other students and Michele Blago in our group, everyone was very friendly, although we come from different countries, we stll get along very well.