Track isolation algorithm studies for LHCb Upgrade

Christian Quaggio, supervisors: Basem Khanji and Fredrich Teubert

August 30th, 2019

1 Introduction

The Track Isolation is an algorithm which discriminates signal against partial reconstructed physics by checking that underlying tracks[1] in each event are not coming from a selected candidate.

Thus, the ensemble of tracks in each event is divided in three categories:

- **Selected tracks**, defined above
- **Isolating tracks**: any track that is not coming from the selected candidate. We will use this ensemble of tracks as signal.
- **Non-isolating tracks**: ensemble of tracks that are coming from the same vertex as the selected candidate. Those will be used as background.

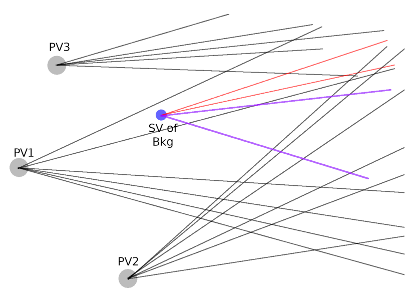

Figure 1: Example of a selected candidate (**red**), Isolating tracks (**black**) and Non-Isolating tracks (**violet**).

[1] each track in the event which does not belong to the selected candidate

A summary of the definitions of the different classes for tracks is shown in Fig.1
The Track Isolation is implemented through machine learning techniques, and is trained with geometric variables which help discriminate Isolating against Non-Isolating tracks.
Track Isolation has proven its prowess by increasing significance in many key measurements at LHCb, such as Lepton Flavour Universality[6] and rare decays searches[4].

2 Analysis

The analysis section of my project consisted in three main parts:

- An exploratory study of the Track Isolation features and outputs, in the context of a study on $B_s \to K\mu\nu$ topology using Run1 simulations samples, with these backgrounds:
  - $B_s\mu\nu \to D_s\mu\nu \to KK\pi\mu\nu$
  - $B^{\pm} \to J/\Psi K^{\pm} \to \mu\mu K^{\pm}$
  - $B_s\mu\nu \to K^{*+}\mu^{-}\nu \to K^{+}\pi^{0}\mu^{-}\nu$

  Selected candidates are coloured in red, while non-isolating tracks for each background are coloured in violet. Those will be the tracks that have to be identified as non-isolating in order to correctly classify that topology as background topology. Notice that in the $K^{+}\pi^{0}\mu^{-}\nu$ background there are no blue tracks, since the π^0 can't be reconstructed as a non-isolating track by the algorithm.

- An analysis of the performance of the Track Isolation for the upcoming LHCb Upgrade [5];

- An optimization of the algorithm for Upgrade MC samples (latest simulation: Beam7000GeV-Upgrade-Nu7.6.25ns), using different machine learning platforms focusing on Boosted Decision Trees.

2.1 Exploratory study

The Track Isolation processes each underlying track in an event, and it assings a value for each track, which ranges from -1 to 1 (score function values of the BDT). The closest this value is to -1, the more non-isolating the track is. So, for each event, if the minimum value of the whole ensemble of tracks is taken, then the most non-isolating track for that event is selected and its properties (geometric, PID, kinematic) are saved in a tuple. Distributions for this minimum (IsoMinBDT) value are shown in Fig.2.

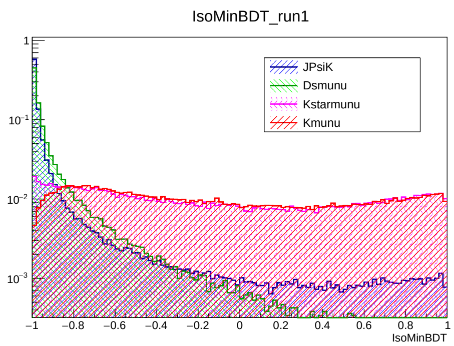

Figure 2: Distributions of the minimum of the Track Isolation for each sample

As expected, $D_s\mu\nu$ and $J/\Psi K$ distributions are peaking around -1, while $K^{*+}\mu^-\nu$ cannot be effectively separated from the signal, $B_s\mu\nu$. But, unexpectedly, $K^{*+}\mu^-\nu$ has a small peak around -1, while in principle it should not be separated from the signal at all. I investigated why that is, and it turns out that electrons are being picked up by the Track Isolation. These electrons are produced by pair production processes, from γ coming directly from π^0 decays. This was checked using MC IDs and Keys. Thus, $K^{*+}\mu^-\nu$ could correctly be classified as background if the Track Isolation was picking up those electrons. The comparison between distributions for electrons and other tracks the Track Isolation was picking up is shown in Fig.3.
Another result of this exploratory study is the dependence of the Track Isolation performance from the number of primary vertices in an event. As it can be expected, the Track Isolation's performance decreases as the number of PVs increases, due to the higher number of tracks in the event, which makes it more probable that the Track Isolation predicts a random negative value, affecting the selection of the non-isolaing track.

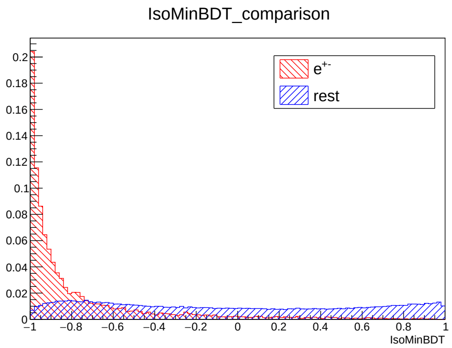

Figure 3: IsoMinBDT comparison between Non-isolating electrons and K^{*+} total IsoMinBDT distribution

2.2 Upcoming Upgrade performance analysis

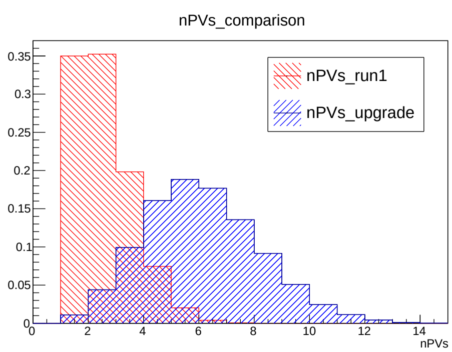

Figure 4: Comparison of number of primary vertices per event between Run1 and Upgrade

Upgrade scenarios portray events in which ~ 6 PVs will be present (Fig.4), which result in a drastic increase in the number of tracks that the Track Isolation has to scan per event. The focus of this part of the analysis is to study the impact of the Upgrade on the Track Isolation performance.

As it can be seen in Fig.5, there is a clear degradation of the algorithm's performance, using integral evaluation of ROC curves, we estimated the loss to be $\sim 20\%$ (depends on background topology). This is due to the added randomness in each event (more tracks, more values from -1 to 1), and to the particular way the non-isolating track is chosen, namely taking the *minimum* of all the values in each event.

However, in the Upgrade, not all tracks in an event will be saved at the trigger level; this is due to limited bandwidth. We investigated the most prominent scenarios in the Upgrade, where only tracks coming from the sig-

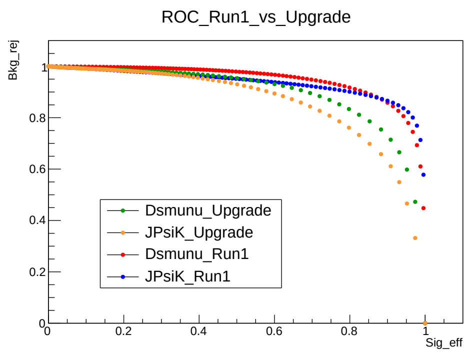

Figure 5: ROC curves comparison between Run1 and Upgrade samples

nal PV will be considered. We found that such a criterion for saving tracks would result in an additional decrease in the Track Isolation performance, due to the increase in the misassociation rate of the non-isolating tracks.
This loss in performance can be recovered, though, if, additionally to the signal PV criterion, tracks with IPChi2 > 6 are also saved. A plot evincing this behaviour is shown in Fig.6.

2.3 Optimization

In the previous studies in this work were done using Run1 Track Isolation. The natural step is to retrain the Track Isolation algorithm using Upgrade MC samples. This was done using two ML platforms, TMVA (from ROOT) and XGBOOST (from scikit-learn).
In both cases the followed procedure was the same:

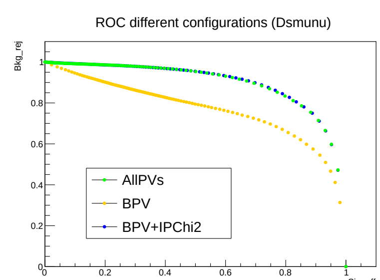

Figure 6: Comparison between ROCs for track-saving methods BPV(signal PV), BPV+IPChi2

- Firstly, the ensemble of the most discriminating variables was found. Examples of some of the discriminating variables are (more details in [2] and [3]):

- **PVDIS**: Distance between vertex of NisoTr and PV
- **SVDIS**: Distance between vertex of the mother of NisoTr and PV
- **DOCA**: Distance of closest approach between NisoTr and the muon/kaon
- **ANGLE**: Angle between NisoTr and the muon/kaon
- **FC**: kinematic variable using $\vec{P_{Niso}}, \vec{P_\mu}$ and the ANGLE
- **BPVZ**: Z position of the best PV associated with the particle
- **MinIPChi2**: T
- **VTXX**: X position of the reconstructed vertex between the track and the muon/kaon

• Secondly, the set of optimal hyperparameters for the machine learning classifier (Gradient boost DT) was found.

2.4 TMVA

In the TMVA[7] framework, the selected classifier was a Boosted Decision Tree. I started from a base of 5 variables, evaluated the ROC curve for that configuration, and then added one variable at a time. If the corresponding ROC curve was better, that variable was kept; otherwise it was excluded from further studies.
Also the optimal number of trees in the forest, the depth of each tree, and other parameters that define the BDT was also studied. The final classifier's output on Upgrade MC samples is shown in Fig.7

2.5 XGBOOST

Regarding XGBOOST [1] the same procedure was followed, training the classifier with the set of most discriminating variables and then finding the optimal configuration of the hyperparameters (with the help of GridSearchCV module). The final output is shown in Fig.8

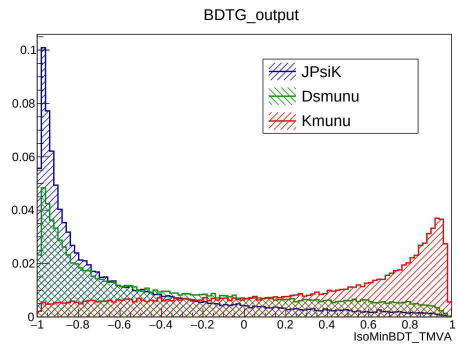

Figure 7: Output distributions for TMVA's BDT

2.6 Classifiers comparison

Eventually, a comparison between the three classifiers is made for $J/\Psi K$ background. Results are shown in Fig.9. The "_run1" classifier is the original Track Isolation (trained with Run1 MC samples) evaluated on Upgrade MC samples. As it is shown, "_run1" classifier is better on working points corresponding to higher signal efficiencies, while TMVA and XGBOOST classifiers are better at lower signal efficiencies.

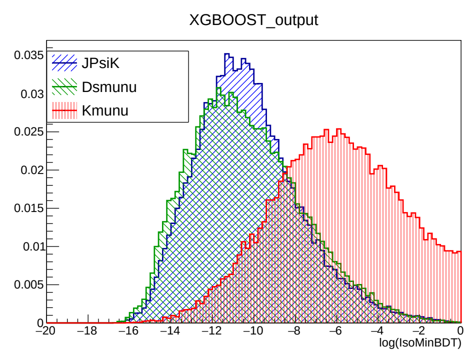

Figure 8: Output distributions for XGBOOST

3 Conclusions and future work

This project has presented a first draft for an optimization of the Charged Track Isolation for the LHCb Upgrade, proving that:

- The algorithm has some potential to remove neutral background through

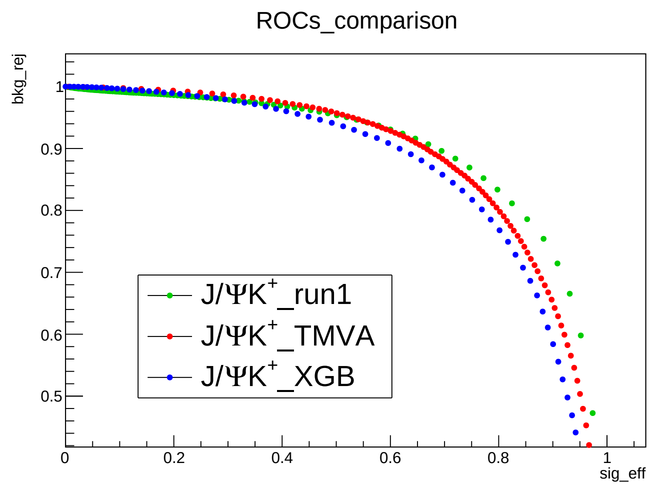

Figure 9: Comparison between all the classifiers used in the project

processes like pair production (and thus having small side effects on signal topologies with π^0),

- Its performance degrades with increasing number of tracks in an event, and it worsens even more if we save only tracks that are coming from the signal PV. In this scenario however, if we save also the tracks with IPChi2 $>$ 6 we get back that loss of performance.

Two machine learning frameworks were studied, with encouraging results. Given the low statistics of the current Upgrade simulations, the performance is still suboptimal, it can be improved using more statistics and tighter cuts on selected candidate particles. These two points will guarantee a superior performance for the Track Isolation algorithm for the Upgrade.

References

[1] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.

[2] Gavardi Laura. *Search for lepton flavour violation in τ decays at the LHCb experiment*, CERN-THESIS-2013-259

[3] G. Mancinelli, J. Serrano. *Study of Muon Isolation in the $B_s^0 \to \mu^+\mu^-$ Channel*, LHCb-INT-2010-011

[4] R. Aaij *et al.* [LHCb Collaboration], Phys. Rev. Lett. **118**, no. 19, 191801 (2017) doi:10.1103/PhysRevLett.118.191801 [arXiv:1703.05747 [hep-ex]].

[5] I. Bediaga *et al.* [LHCb Collaboration], CERN-LHCC-2012-007, LHCb-TDR-12.

[6] R. Aaij *et al.* [LHCb Collaboration], Phys. Rev. D **97**, no. 7, 072013 (2018) doi:10.1103/PhysRevD.97.072013 [arXiv:1711.02505 [hep-ex]].

[7] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," PoS A CAT 040 (2007) [physics/0703039].