Multivariant Analysis for the Selection of LHCb Data

by

Afaldy Hayeeteh
afaldy.hayeeteh@gmail.com

and

Md Rounak Jahan Raj
rounakjahanraj@gmail.com

Submitted to the Department of Physics

in fulfillment of Advanced Laboratory course for the degrees

of International Master on Advanced methods in Particle Physics

at

Technische Universität Dortmund

May 2025

Abstract

This study is aimed at revealing the signal decay of $B_s^0 \rightarrow \psi(2S)K_s^0$ which is hidden under a combinatorial background. The data used for the analysis were collected by the LHCb experiment during Run 2. On top of real data, simulations of the signal channel and of a kinematically similar control channel are used. XGBClassifier model is used to do the analysis, resulting in a proxy signal significance of 4.42, strongly suggesting that the model was rightly used to reveal the signal shape within the real data set.

Contents

Chapter 1

Introduction

LHCb is one of the four major experiments at Conseil Européen pour la Recherche Nucléaire (CERN) specialized in precision measurements and usually studies b- and c- quark hadrons. During run 2 (2015-2018), a large set of data has been collected after the initial selection process using the trigger system. However, the data has large amount of background, resulting in a shape of the invariant mass distribution where the peak of the $B_s^0 \to \psi(2S)K_s^0$ decay is hidden.

The tree level Feynman diagram of the decay is shown in Figure 1-1. This decay can be used to study all important CP violation. One example is the measurement of the time-dependent CP asymmetry in the decay.

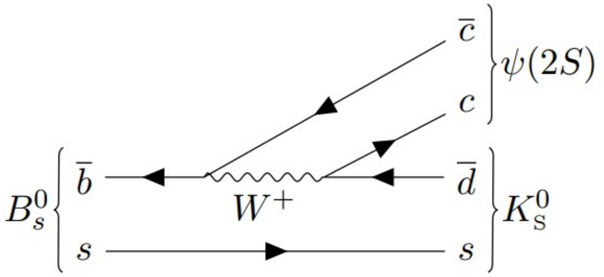


Figure 1-1: Leading order Feynman diagram of the decay $B_s^0 \to \psi(2S)K_s^0$ in the Standard Model.

Three datasets are used for this analysis:

1. The real data recorded by the LHCb experiment with a very rough selection.
2. A sample of the signal decay $B_s^0 \to \psi(2S)K_s^0$ obtained from Monte-Carlo simulation.
3. A sample of the (kinematically) very similar decay $B^0 \to \psi(2S)K_s^0$ obtained from Monte-Carlo simulation.

The LHCb experiment uses an elaborate trigger system to select likely interesting events and reduce the great amount of data to a size that is small enough to be stored. Often, the interesting events are are hidden in the real data sample such that some further selection is required. In order to study the physics of $B_s^0 \to \psi(2S)K_s^0$ events, a cleaner sample with better signal to background ratio is required. Obtaining such a sample, using a multi-variate analysis (MVA) to classify signal and background, is the aim of this study. A MVA classifier is a machine learning method used to distinguish between different categories or classes (e.g. signal vs. background) by analyzing multiple input variables simultaneously. The signal region should never be used in training the classifier to avoid bias. Otherwise, it might lead to false discovery.

Simulated data often differ from real data due to limitations in theoretical models and detector simulations. Improving accuracy would require significantly more computational resources, as simulating each event already takes several minutes and millions of events are needed per decay mode. To lower these discrepancies, kinematic weights are applied to simulation samples. These weights correct for differences between data and simulation in kinematic distributions, ensuring better alignment with real observations.

Since, the signal in the real dataset is hidden in the background, it is essential to use a loss function that needs to be minimized. The inverse of the loss function is called a figure of merit (FOM), which helps quantify the quality of a classifier. One important and common FOM in the high-energy physics (HEP) community is Punzi FOM [2]. It is defined as:

$$FOM = \frac{\epsilon_{sig}}{\frac{5}{2} - \sqrt{N_{bkg}}} \tag{1.1}$$

where N_{bkg} is the expected number of background events in the signal region and ϵ_{sig} is the classification efficiency of the signal. Again, N_{bkg} can be defined as:

$$N_{bkg} = \text{No. of background events in the background region} \times \frac{\text{width of signal region}}{\text{width of background region}}.$$

The background in the real dataset is mostly combinatorial. Combinatorial background consists of events that pass the selection purely due to coincidence. For example, two unrelated muons may accidentally appear to originate from a common vertex, prompting the reconstruction algorithm to falsely identify them as a $\psi(2S)$ candidate. If a nearby, randomly associated K_s^0 is also reconstructed, the algorithm may combine it with the $\psi(2S)$ to form a B candidate. However, since the $\psi(2S)$ and K_s^0 are uncorrelated in such cases, the resulting invariant mass of the reconstructed B candidate can span a wide range. In contrast, true $B_s^0 \to \psi(2S)K_s^0$ decays will cluster around the nominal B_s^0 mass, with well-defined kinematic and geometric properties. Because combinatorial background lacks such correlations, kinematic variables are particularly effective in distinguishing it from genuine signal events.

In this experiment, there are a lot of instances where distributions will be compared. A trivial way to do that is to get the largest distance between the cumulative probability distributions(CDF). It can be defined as:

$$\sup_n \left|F_n^1 - F_n^2\right|$$

where n is the number of bins in the histogram. This can be easily obtained using the Kolmogorov-Smirnov (KS) test from the scipy package in coding [1].

The final evaluation can be done using the significance which can be defined as:

$$m = \frac{N_{sig}}{\sqrt{N_{sig} + N_{bkg}}} \tag{1.2}$$

where N_{bkg} is the number of background candidates and N_{sig} is the number of signal candidates. This significance is a measure of how likely the result is produced by statistical calculations. A discovery is claimed if the significance is greater than 5. If the significance is larger than 3, then there is evidence of the signal.

Chapter 2

LHCb

The Large Hadron Collider (LHC) is a particle accelerator at CERN, the European organization for nuclear research, close to Geneva and LHCb is one of the four main experiments at the LHC. The data used in this analysis was taken during Run 2 of the proton-proton collisions and filtered through different triggering techniques. The schematic diagram of the LHCb is shown in Figure 2-1.

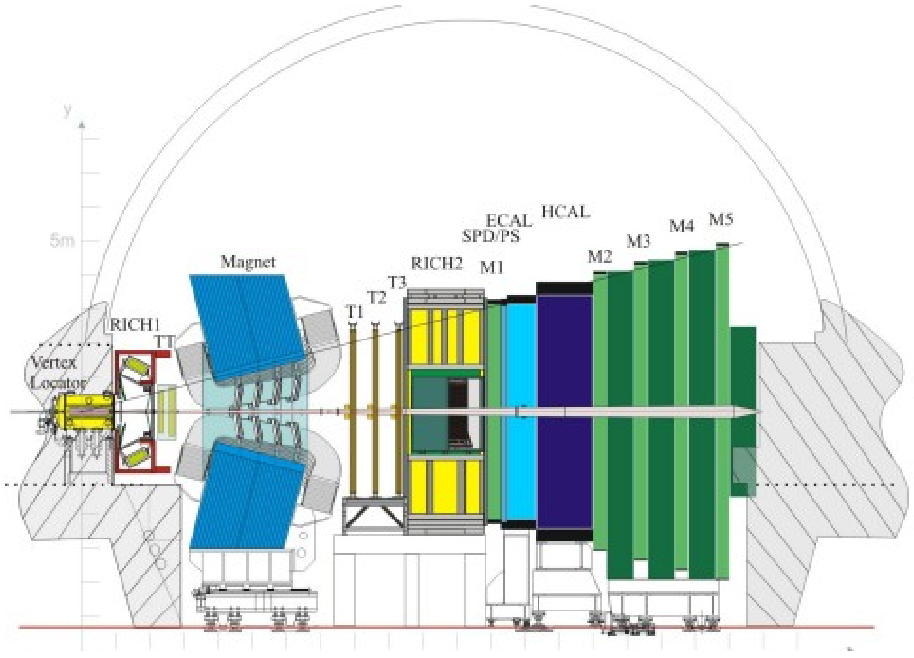


Figure 2-1: Schematic diagram of the LHCb detector.

Unlike the large, symmetric detectors like ATLAS and CMS, LHCb has a forward geometry, focusing on particles emitted at small angles relative to the beam line, where b-hadrons are most frequently produced. The detector is a forward spectrometer detector, about 20 meters long, consisting of several specialized components arranged in a sequence. Closest to the collision point is the Vertex Locator (VELO), which uses high-precision silicon sensors to detect where particles originate and helps distinguish the decay vertices of short-lived particles. Following VELO is the tracking system, which measures the momentum of charged particles as they curve in a magnetic field provided by a large dipole magnet.

To identify the type of each particle, LHCb uses two RICH (Ring Imaging CHerenkov) detectors that detect light emitted when charged particles travel faster than the speed of light in a medium. These detectors help distinguish between similar particles like pions, kaons, and protons. Further along the detector are the electromagnetic and hadronic calorimeters, which measure the energy of electrons, photons and hadrons by absorbing them. At the far end is the muon detection system, which identifies muons that pass through all other layers of the detector, as muons are crucial in identifying many rare decay processes. LHCb also features a sophisticated trigger system, which quickly filters out the vast majority of uninteresting collisions and keeps only the events likely to contain b-hadrons or other particles of interest. The combination of precision vertex detection, excellent momentum measurement, and robust particle identification makes LHCb exceptionally capable of studying rare decays, CP violation, and potential signs of new physics beyond the Standard Model.

As shown in Figure 1-1, B_s^0 decays to $\psi(2S)$ and K_s^0. Each of $\psi(2S)$ and K_s^0 are reconstructed from their decay products and are not observed directly in the detector. The first one has a higher branching fraction for strong decays, but for this study the electromagnetic decay into two pions ($\mu^+\mu^-$) is emphasized because identifying these with high precision is easier. Similarly, K_s^0 decays to two pions ($\pi^+\pi^-$). Moreover, B_s^0 can also be reconstructed from the nominal mass of it's decay products $\psi(2S)$ and K_s^0.

Chapter 3

Results and Analysis

The invariant mass distribution of the real data is shown in Figure 3-1. For simplicity, it was assumed that the invariant mass distribution shown in the figure is composed only of the decays originating from the signal and control channels and from the combinatorial background. The peak in Figure 3-1 is caused by $B^0 \to \psi(2S)K_s^0$ decays, while the signal decay ($B_s^0 \to \psi(2S)K_s^0$) cannot be observed as it is submerged by combinatorial background.

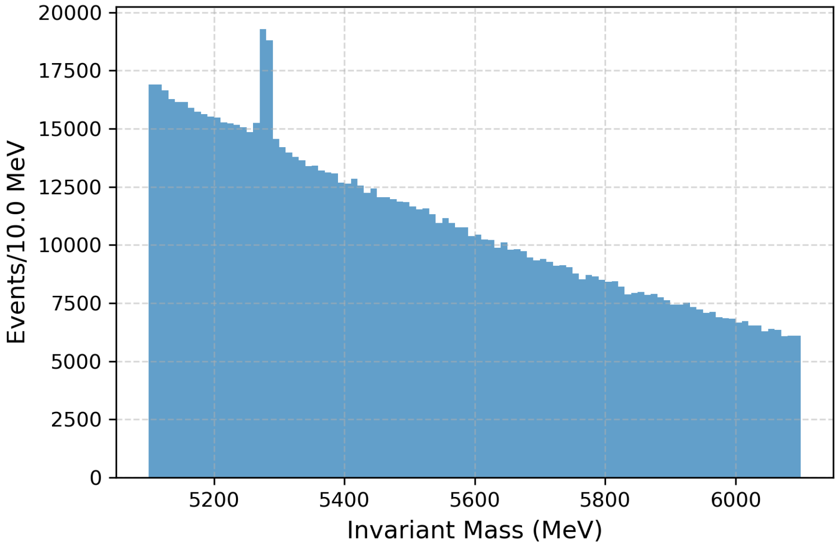


Figure 3-1: Reconstructed invariant mass distribution of the real data set.

3.1 Signal Region and Background Sample

In order to define the signal region in the real dataset, a selection was made in the real dataset was made with the help of the signal simulation dataset. First, a window containing 99% of the data was found in the signal dataset as shown Figure 3-2.

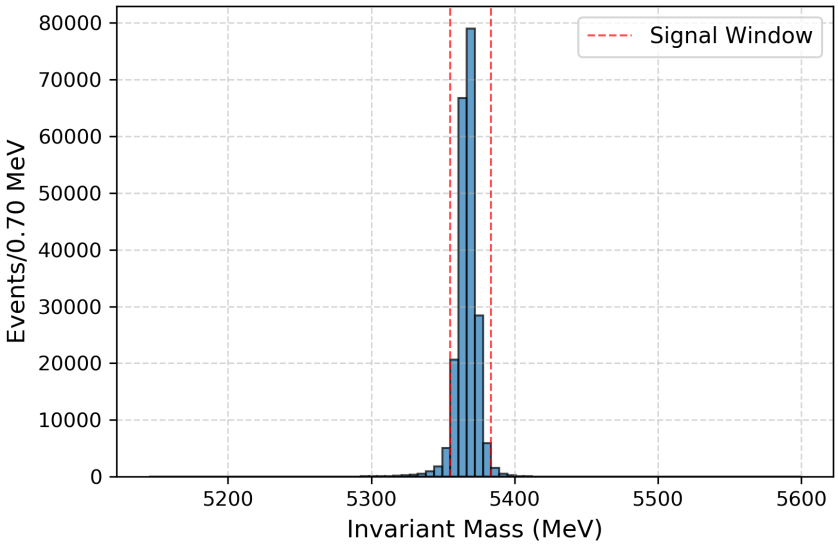


Figure 3-2: Reconstructed invariant mass distribution of the signal simulation.

Afterwards, this window is implemented in the invariant mass distribution of the real dataset as shown in Figure 3-3

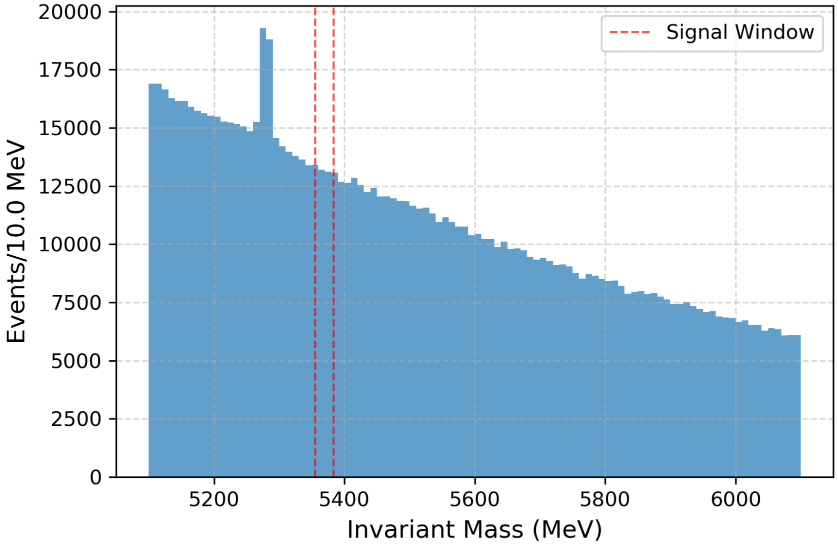


Figure 3-3: Implemented signal region in the real data set invariant mass distribution.

The two Monte Carlo simulations do not contain any background, so the background must be identified within the real data set. So, to train the MVA to recognize background, background from the real data sample shall be used. Hence, the upper side band (USB) of the real data set was taken as the background sample, i.e. all the data in the real data set that have invariant mass higher than the right side of the signal region. The signal simulation dataset and the background candidates obtained from the USB are used for the MVA training.

In the real data set, there is a column with weights, which if used gives the pure B^0 events. The figure below shows the distribution with and without these weights.

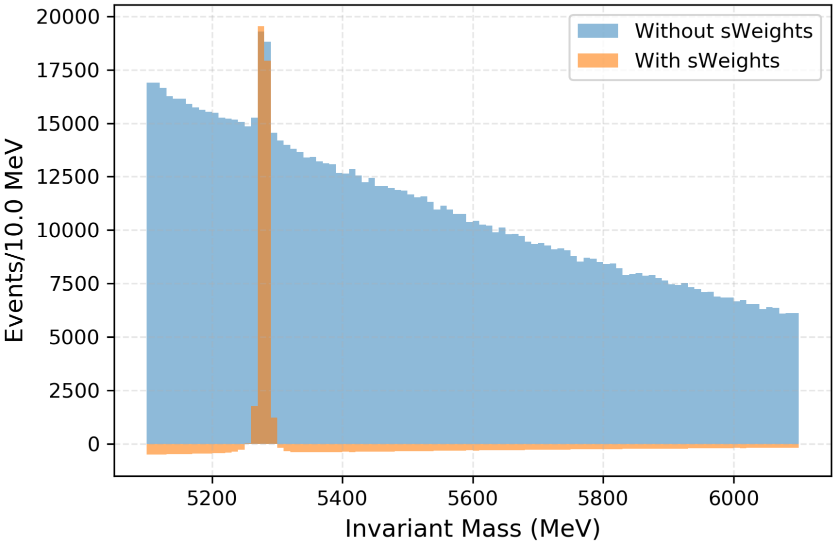


Figure 3-4: Invariant mass distribution as returned by the trigger line (blue) and invariant mass distribution of the B^0 candidates, highlighted using background subtraction (s-weights) whights (orange).

3.2 Feature Selection

Even after using the weights there are many differences in the distribution due to various reasons. A large number of features will make the model slow and also lead to overfitting whereas a small amount of features will make poor predictions. Not all of the features in the dataset contain useful information. In order to select suitable features, the KS test was used.

Hence, features which satisfies the following conditions must be used for training the model:

1. Reasonable agreement between data and simulation (otherwise the classifier only learns to distinguish data from simulation instead of signal from background).

2. Strong discrimination between signal (simulation) and background (USB).

3. Uncorrelated to the invariant mass of the B_s^0/B^0 candidate.

Using the KS test, 27 suitable features were found based on the above criteria. In order to find similarity between the B^0 simulation and B^0 data, a threshold KS < 0.05 had been used and to distinguish between the signal and the background a threshold of KS > 0.2 had been used. In this test, the KS metric had values [0, 1] where 0 means that the distributions are identical and 1 means that the distributions are completely different. The thresholds used in the analysis are achieved via exploratory analysis.

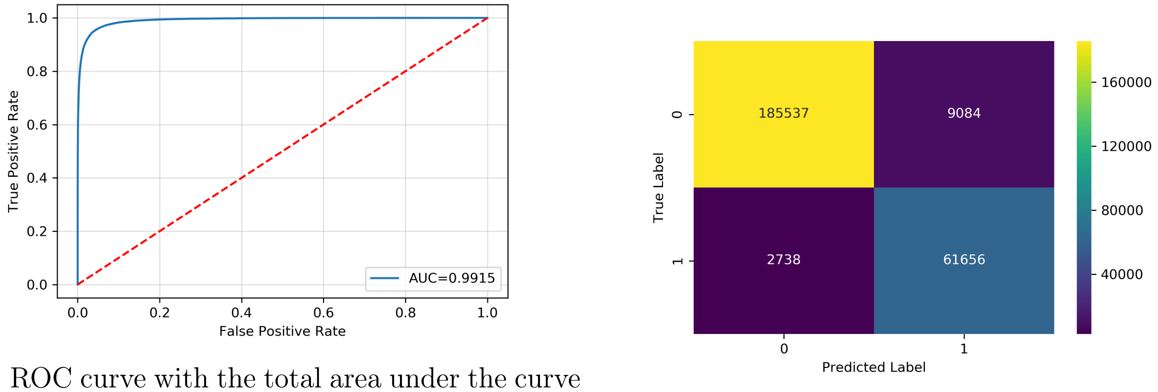
3.3 Training the Classifier and Optimization

The machine learning model used in the analysis was XGBClassifier, which is a supervised learning model. Since, the number of background data is very large compared to the number of signal data, weights has been used to make accurate predictions. Because, imbalanced data can lead to misleading or biased classifications. Afterwards, hyperparameter optimization was done to slightly improve the performance of the model. In this case, RandomizedSearchCV, which is a hyperparameter tuning technique provided by scikit-learn library, had been used. The best parameters obtained from the optimizations are shown in Table 3.1.

Parameters	Best Values
`n_estimators`	1000
`max_depth`	6
`learning_rate`	0.1
`lambda`	1

Table 3.1: Best parameters obtained from hyperparameter optimization.

The evaluation of the model can be done with the following receiver operating characteristic (ROC) curve and confusion matrix.



(a) ROC curve with the total area under the curve (AUC) value.

(b) Confusion matrix.

Figure 3-5: Evaluation of the BDT model.

Both the confusion matrix and AUC score suggest that the model was trained well. The AUC score is especially useful in this scenario, when the dataset is heavily dominated by background samples. Another parameter is the boosted decision tree (BDT) score, which is a numeric output of a boosted decision tree classifier such as XGBClassifier, that reflects how strongly the model classifies a given event as signal or background. The distribution of the BDT score is shown in Figure 3-6.

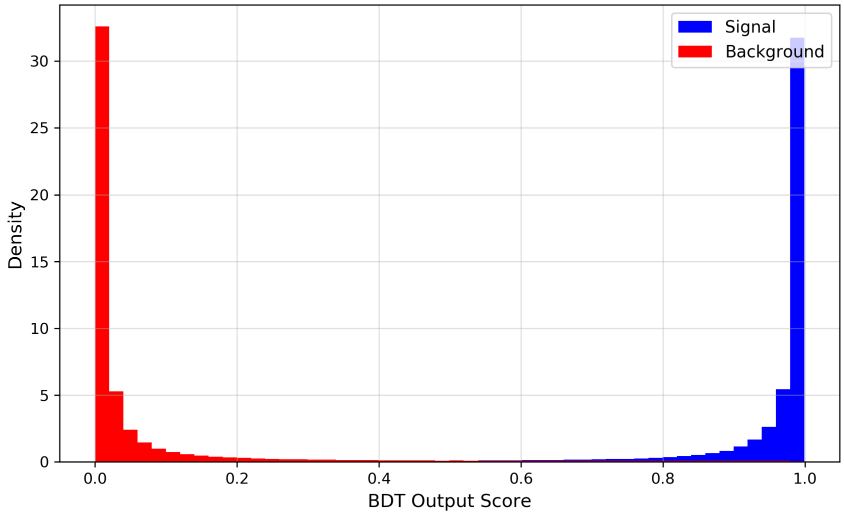


Figure 3-6: Distribution of the BDT output for background and signal samples.

From Figure 3-7, it can be easily found that the best cut value which maximizes the Punzi FOM is 0.9919. A plot of significances vs. threshold is shown in Figure 3-7.

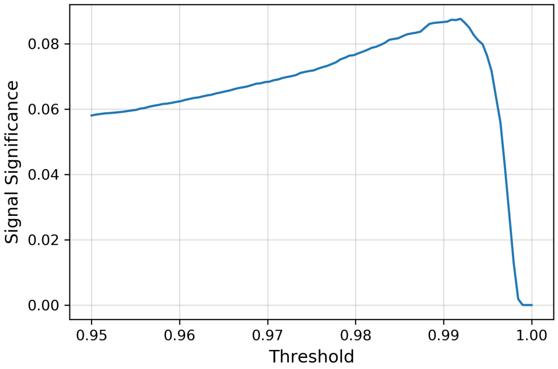


Figure 3-7: Plot of Punzi FOM as a function of threshold.

Applying this best cut, a purer plot of the real data sample can be obtained as shown in Figure 3-8.

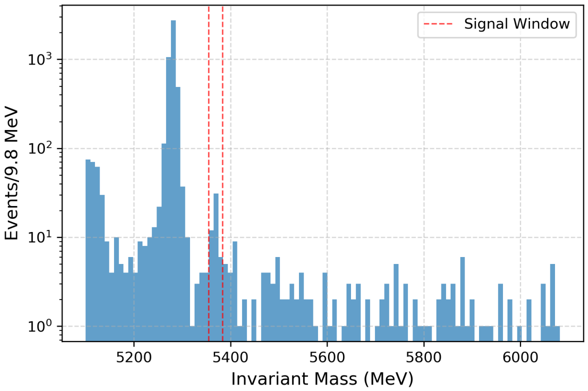


Figure 3-8: Selected mass distribution of the real data set.

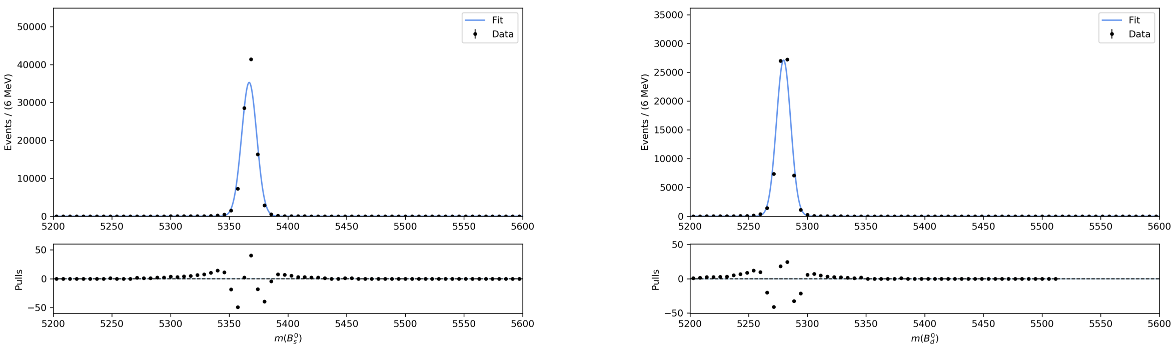
In Figure 3-8, two peaky structures can be observed in the regions where we expect the B_s and B^0 mass peaks.

3.4 Finding the Signal in the Data Sample

If peaks are fitted in the signal and control simulation using Gaussian function, then Figure 3-9a and 3-9b are obtained. A Gaussian function has been used to fit peaks in signal and control simulations because it models the symmetric distribution of reconstructed quantities around their true values due to detector resolution, which provides with simple and symmetric signal shape. The formula for a Gaussian function (also known as the normal distribution) is:

$$f(x) = A \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

where A is the height of the peak (amplitude), μ is the mean, σ is the standard deviation and e is the Euler's number.



(a) Signal simulation.

(b) Control simulation.

Figure 3-9: Fitted peak to the signal and control simulation.

When considering the full mass window of the real data sample, the shapes obtained from the preliminary fits on simulations along with an exponential background are shown in Figure 3-10. An exponential background refers to a type of background distribution in data that decays exponentially. This can be represented mathematically as:

$$f(x) = Ae^{-\lambda x} \tag{3.2}$$

where A is the normalization constant, λ is the decay constant and x is the observable.

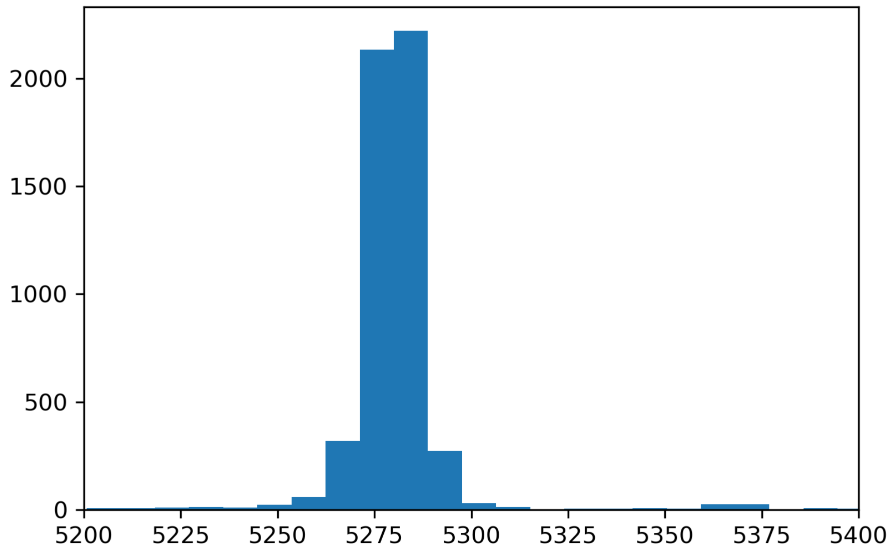


Figure 3-10: Full data fitted with fixed peak shapes and exponential background.

Figure 3-11 shows the invariant mass distribution of candidates from the real data sample which were selected via the classifier criteria. The results of the fit are overlaid.

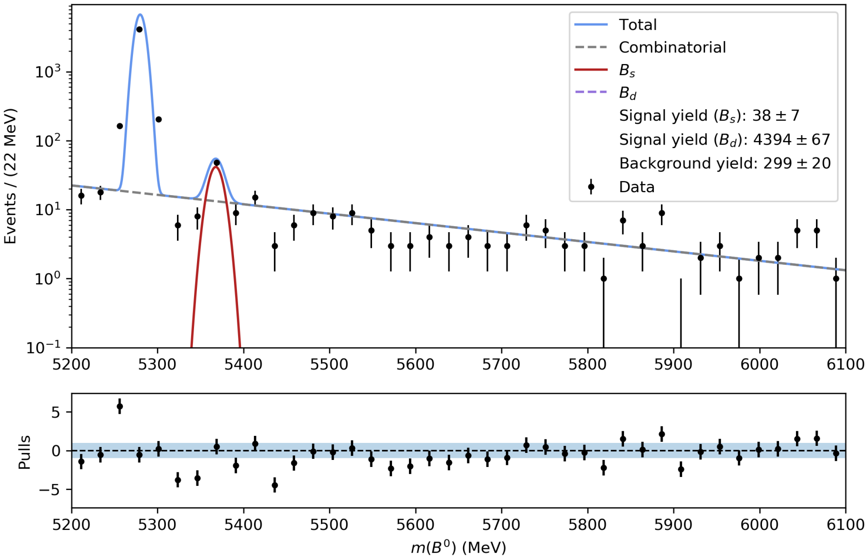


Figure 3-11: Full mass window of the real dataset using the shapes obtained from the preliminary fits on the simulations and an exponential background.

Using the values, obtained from the fit, the proxy significance can be calculated. The signal and background events in the signal region are $n_{sig} = 30$ and $n_{bkg} = 17$ respectively. Hence, according to Eqn.1.2, the signal significance is 4.42.

Chapter 4

Conclusions

The search for the decay $B_s^0 \to \psi(2S)K_s^0$ yields a number of signal and background candidates of 30 and 17, respectively. Signal is observed with a significance of 4.42. It indicates that signal has been found from the real dataset. The classifier training could have been improved with more hyperparameter optimizations. But, considering the limited scope of the analysis, it was kept limited. This might lead to limitations. The model used for the analysis is XGBClassifier, the opportunity to use other classification model is also open.

Both the simulations were produced using Monte-Carlo simulation technique. Even if specific weights are put in place to minimize the differences between data and simulations, these later ones can never be identical to real data in the behavior of all of their variables. All of these factors have affected the results of the analysis.

Literatures

[1] Cumulative distribution function, https://en.wikipedia.org/wiki/cumulative_distribution_function, 2025.

[2] Giovanni Punzi. Sensitivity of searches for new signals and its optimization, 2003.