Advanced Lab Course: Particle Physics

Multivariate Analysis

Leander Flottau
leander.flottau@tu-dortmund.de

Ajeesh Garg
smajgarg@tu-dortmund.de

2. July 2024

TU Dortmund University – Department of Physics

Contents

1 Introduction

The Standard Model (SM) of particle physics is a highly successful theoretical framework that has been able to explain the most fundamental interactions of nature. Over the past century, it has been rigorously tested and its accuracy of predictions that have been confirmed through numerous experiments.

Despite its successes, the Standard Model is not a complete theory. Several fundamental questions still remain unanswered, such as the nature of dark matter and dark energy as well as the observed matter-antimatter asymmetry in the universe. This matter-antimatter asymmetry is particularly intriguing because the Big Bang should have produced equal amounts of matter and antimatter, yet our observable universe is overwhelmingly composed of matter.

Under Sakharov conditions, a crucial aspect of this asymmetry is CP-violation, which refers to the violation of the combined symmetries of charge conjugation (C) and parity (P). CP-violation has been observed in the decays of certain particles and is incorporated into the SM through the complex phase of the Cabibbo-Kobayashi-Maskawa (CKM) matrix. However, the extent of CP-violation predicted by the SM is insufficient to account for the baryon asymmetry observed in the universe, suggesting the presence of new physics beyond the Standard Model.

This mechanism of CP Violation was further verified experimentally in the weak B Meson decays by the Barbar (1999) and Belle experiments (2010) but the amount of CPV observed is still too small to account for matter-antimatter asymmetry of the Universe. The large Hadron collider beauty (LHCb) experiment is designed specifically to measure this CPV more precisely in the case of rare decays of hadrons containing bottom and charm quarks.

This report focuses on the measurement of CP-violation in the decays of B mesons (particles containing a bottom quark) using data collected by the LHCb experiment. By comparing the decay rates of B^+ mesons and their antiparticles (B^-) into three hadrons, we aim to quantify the CP-asymmetry. Specifically, we analyze decays producing three kaons, which are less affected by background noise compared to those involving pions.

The results from this analysis will contribute to the broader effort to uncover the mechanisms behind CP-violation and explore potential new physics that could explain the observed matter-antimatter asymmetry in the universe.

2 Large Hadron Collider

The Large Hadron Collider (LHC) at CERN, near Geneva, is a particle accelerator that collides proton beams at a centre-of-mass energy of 7 TeV. Each beam contains approximately 2000 bunches, each with about 1.1×10^{11} protons, colliding at an average rate of 15 MHz, resulting in numerous inelastic interactions. The main goal of the LHC experiments is to test the Standard Model (SM) of particle physics and to search for physics beyond SM.

The LHCb experiment is one of the four main experiments at the LHC, specifically

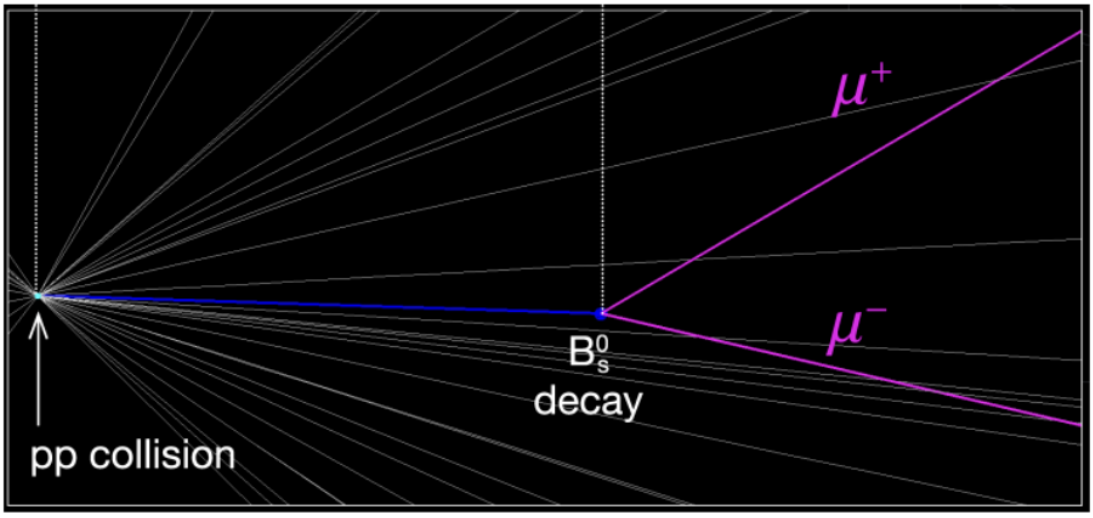

Figure 1: Schematic depiction of a proton-proton interaction followed by a B-decay [1]

designed to study CP-violation and the rare decays of hadrons containing b and c quarks. The detector is positioned at the LHC's interaction point 8 (IP8) and features a single-arm forward spectrometer. This design is optimized to capture particles produced in the forward direction, where b quarks are predominantly emitted following proton-proton collisions.

3 Construction of LHCb

3.1 Vertex Locator (VELO)

The Vertex Locator (VELO) is crucial for reconstructing the primary vertex, where the initial proton-proton collisions occur, and secondary vertices, where short-lived particles decay. The VELO consists of silicon strip detectors arranged in close proximity to the interaction point. These detectors provide high-resolution measurements of particle trajectories, allowing precise determination of vertex positions.

3.2 Tracking System

The tracking system of LHCb includes a dipole magnet and several tracking stations. The dipole magnet creates a magnetic field that bends the paths of charged particles. By measuring the curvature of these tracks, the momenta of the particles can be calculated. The tracking stations, positioned before and after the magnet, consist of silicon strip detectors and straw drift tubes, which detect the passage of charged particles and record their trajectories.

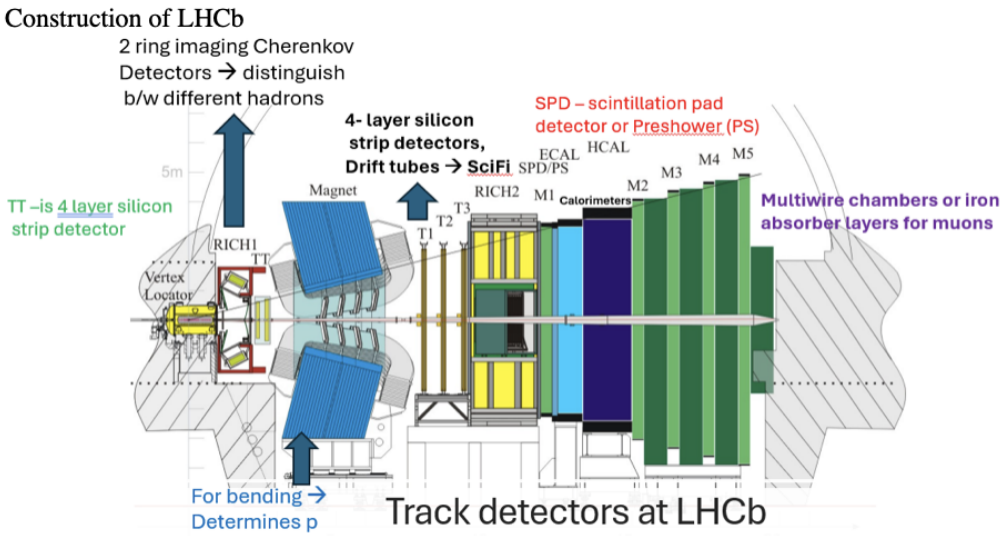

Figure 2: Schematic depiction of the Lhcb detector [1]

3.3 Calorimeter System

The calorimeter system of LHCb is designed to measure the energy of photons, electrons, and hadrons. It consists of the Scintillating Pad Detector (SPD), the Preshower Detector (PS), the Electromagnetic Calorimeter (ECAL), and the Hadronic Calorimeter (HCAL). The ECAL absorbs the electromagnetic particles while the HCAL absorbs the hadrons and hence, their energy is determined. The ECAL uses lead as the material while HCAL uses iron as the material.

3.4 Particle Identification Detectors (RICH1 and RICH2)

Two Ring-Imaging Cherenkov (RICH) detectors are used for particle identification. RICH1 is located upstream of the magnet, and RICH2 is downstream. These detectors exploit the Cherenkov effect, where charged particles traveling faster than the speed of light in a medium emit light at a characteristic angle. By measuring this angle, the velocity and, consequently, the mass and type of the particle can be determined.

3.5 Muon System

The muon system consists of alternating layers of iron and multi-wire proportional chambers. Muons, unlike most other particles, can penetrate the dense iron layers, allowing them to be identified and tracked. This system is essential for studying decays that produce muons in the final state.

3.6 Online and Offline Data Processing

The LHCb detector produces a vast amount of data, which is filtered and processed in real-time by an online trigger system. This system selects events of interest and reduces the data volume for further analysis. Selected events are then sent to an offline computing grid, where detailed analyses are performed.

4 Analysis

4.1 Aim and Available Samples

4.1.1 Aim

The main aim of this experiment is to identify and analyze the decay events of ($B_s^0 \rightarrow \psi(2S)K_S^0$) from the given data set. For this purpose, we are provided with three different datasets:

4.1.2 Lhcb Measurement Data

This is the real data obtained from the second run of the LHCb experiment after heavy trigger preselections. It contains a mixture of both signal and background events, but a large part of this data is the background. Being able to see the B_s^0 peak from this dataset is the main objective of this exercise. The distribution of the data can be seen in Figure 1. A peak can be seen in the dataset around 5279 MeV. This peak doesn't correspond

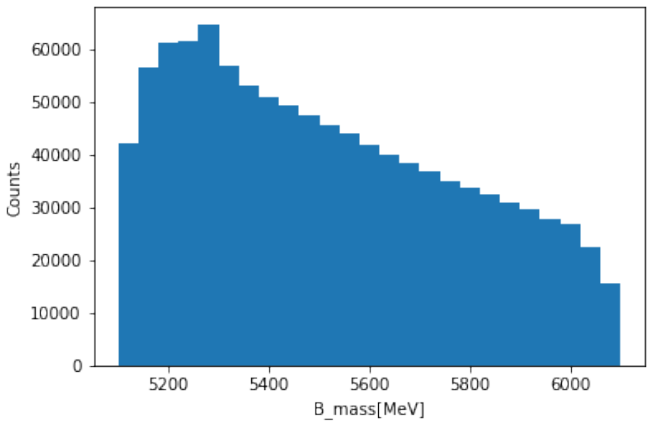

Figure 3: Mass distribution of the measurement data

to our actual signal, as our signal is hidden somewhere inside the dataset, there are far more background events than signal events. Instead the visible peak in the data is caused by the B^0, which is a cinematically similar process. Finding the hidden signal's peak is the ultimate aim of this analysis. The distribution of the measured particle masses can be seen in Figure 3. From this data the background sample can be taken in the form of the upper-sideband of the B_s^0-decay, which includes all events with a higher mass than that of the B_s^0. This subsample should theoretically be composed purely of combinatorial background.

4.1.3 Signal simulation

This sample contains simulated signal events for our decay ($B_s^0 \rightarrow \psi(2S)K_S^0$). This sample mimics the characteristics of the real signal decay and hence is used for training

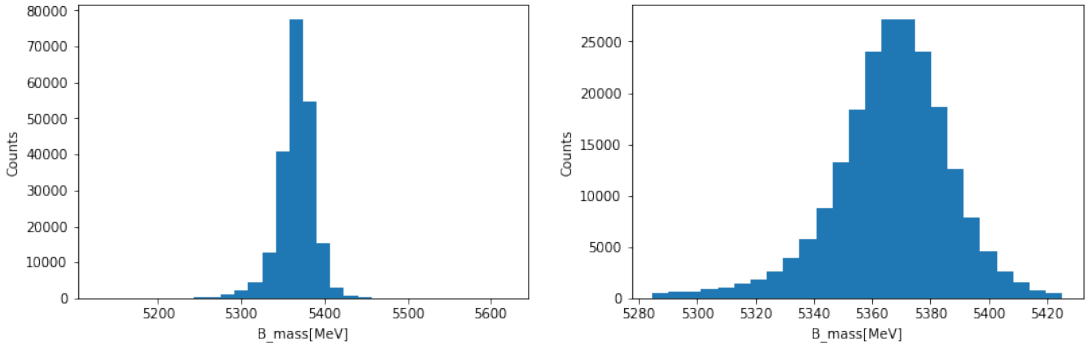

Figure 4: Mass distribution of the whole signal simulation (left) and the signal window (right)

our model. The need for this dataset gives us a way to compare our results and, in our case, actually train the algorithm without utilizing the useful data for training purposes. From this the signal window, defined as the shortest intervall containing 99% of all signal events, can be computed to use as a training sample for the classifier.

The peak of this signal's simulation can be seen around 55366.88 MeV [3].

4.1.4 Control Simulation Sample

This sample includes simulated events for the decay $B^0 \to \psi(2S)K^0_S$. This decay is kinematically very similar to our signal decay, so it passes our selection mechanisms and ends up in our data. As this decay has much higher branching ratio than the actual signal, making it a perfect candidate to use a simulation to control the similarity between simulated and measured events. The mass distribution of the control simulation

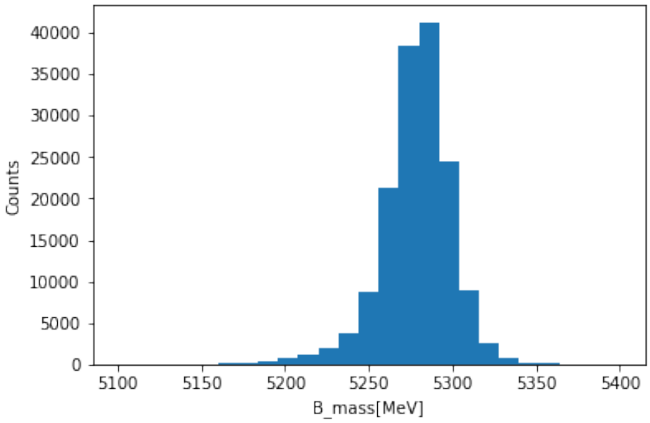

Figure 5: Mass distribution of the measurement data

is displayed in **??**.

4.2 Feature Selection

For our signal training, simulated data helps us understand the typical behavior of true signal events. However, simulations are often imperfect due to the limitations of theoretical models and the accuracy of detector simulations.

To address these issues, we use kinematic weights stored in the kinematic weights variable. These weights correct some of the inaccuracies in the simulated decay kinematics, ensuring our signal sample closely resembles what we expect from real data.

Despite these corrections, other properties may still be mis-modeled in the simulation. To further ensure our classifier accurately distinguishes between signal and background, we use the decay $B^0 \to \psi(2S)K^0_S$. This decay is similar to the B^0_s decay, particularly in kinematic variables crucial for removing combinatorial background. The B^0 decay is also abundant in our data set, providing a useful control sample.

We use sWeights, specifically $sWeights\_sig$, which give us pure B^0 events from the data sample. These sWeighted data distributions are reliable representations of the true distribution, provided the variable in question is uncorrelated with the invariant B mass.

Furthermore it has to be ensured, that the selected features descriminate well beween signal and background. To control both criteria the similarity measure:

$$\sup_n |F^1_n - F^2_n| \tag{1}$$

is used. Based on this measure 20 features are selected which descriminate well between data and background while also showing a good compatibility of simulation and real data. All of those features are additionally required to have a low correlation to the invariant mass.

4.3 Classification

To classify the measured events into signal and background decision trees are trained using a boosting method. Boosted decision trees are a special form of ensemble classifiers. Unlike a random forest, which trains decision trees simultaneously the boosting method trains sequentially and assigns a higher weight to events which were misclassified in previous iterations. This method of training performs well but the sequentiall training of classifiers takes up more computation time than other methods.

The upper-sideband used as a background sample for training purposes contains significantly more events than the signal simulation. To avoid constructing a biased classifier this descrepancy has to be fixed by assigning events a class-based weight to give a higher significance to signal events effectively counterbalancing the biased training sample

4.4 Cross Validation

To effectively evaluate a classification algorithm with a limited amount of training and testing data cross-validation is a popular method. In this procedure the data is split into k-subsets using one subset for training and the other $k-1$ sets for training. This

procedure can be repeated until every subset has been used once for training. This ensures a good generelization performance of the classifier and allows for early detection, if the classifier is too dependend on certain types of training data. In this case 3-fold cross-validation is used to evaluate the classifier.

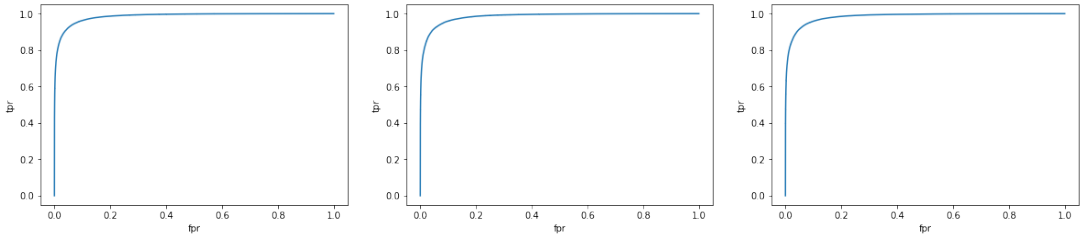

Figure 6: Receiver-operating-characteristic of the classifier for the three folds

4.5 ROC-curve

The roc-curve is generated by plotting the true-positive rate (tpr) versus the false-positive rate (fpr) for different threshholds. The area under the curve indicates the quality of the classifier with a diagonal line representing a random guess. All three folds show very good results on the test sets indicating good generelization and well selected features. As seen here cross-validation is a very effective method to account for potential biases in the training procedure, but due to the neccessity to train one classifier for every fold it can prove to be very runtime intensive. [2]

4.5.1 Threshold optimization

Another important component of constructing a classifier is the threshold optimization. To optimize this value a quality-measure is needed, to evaluate every individual thresholds performance. In this case this measure is:

$$FOM = \frac{\epsilon_{sig}}{5/2 + \sqrt{n_{bkg}}} \tag{2}$$

Where ϵ_{sig} represents the signal efficiency of the classifier which can easily be computed using the signal labels from the simulation. The value n_{bkg} represents the number of background events in the signal region. This value can not be directly calculated because the signal region of the data is not evaluated in testing to avoid additional biases. It therefore has to be estimated under the assumption, that the efficiency of the classifier on the background is constant over the mass distribution. With this efficiency estimated from the sideband and the total number of background events in the signal region prior to the selection estimated assuming they strongly outweigh the signal events before applying the classification n_{bkg} can be calculated. The quality measure in Figure 7 clearly shows, that an extremely high classification threshold is needed to effectively search for the signal because of the extremely low number of signal events relative to the background. The optimal computet value is $T_{opt} = 0.9981$.

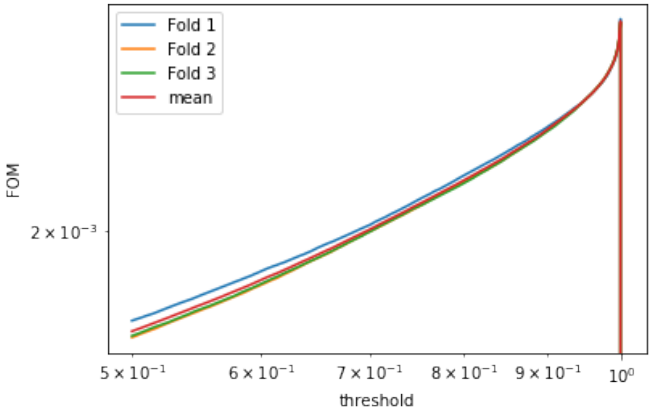

Figure 7: The FOM based on the threshhold for the three individual folds as well as the mean

4.6 Signal Fitting and significance

After application of the classifier trained on the whole training sample and using the threshold computed in the previous section The mass ditribution looks as displayed in 8.

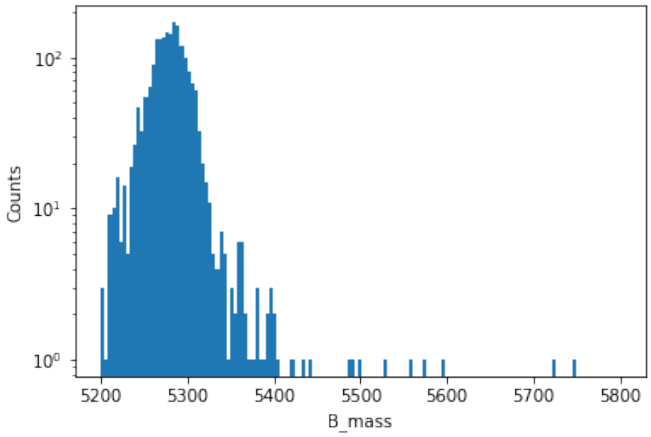

Figure 8: Mass distribution after applying the previously constructed classifier

It has a very dominant peak corresponding to the B^0 decay, which cannot be filtered out by the classifier as it works based on kinematics, as well as a second peak around the expected mass of the B_s^0. The third component of the resulting distribution is made up of the remaining combinatorial background.

The shape of the first two components is first fitted by using the distributions of the simulated data. The underlying function fitted to the peaks is a linear combination of two different gaussian peaks.

To get the distribution of the final fit the shapes of the previously determined peaks are

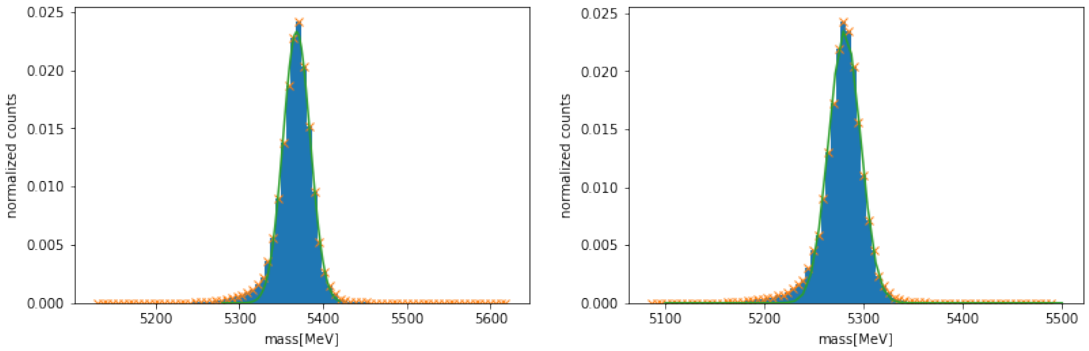

Figure 9: The shape-fitted distributions of the signal and control events based on the simulation

fitted using individual scaling factors as well as the background component. When using

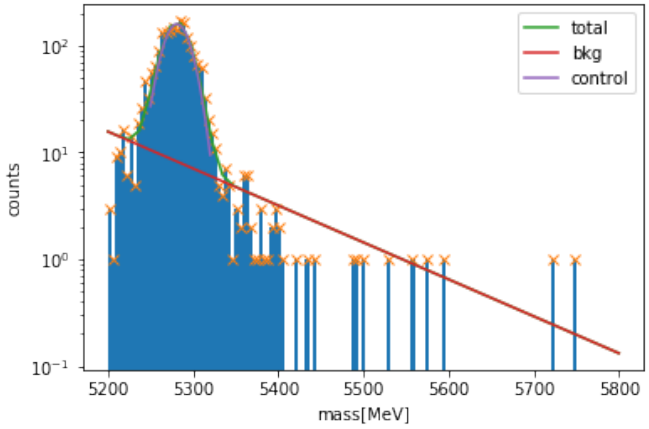

Figure 10: Fit for the total distribution using an exponential background

a decreasing exponential function for the background as displayed in 10 the background is so dominant compared to the signal peak that the peak is not recognizable. The scaling factors of the two components differ by approximately eight orders of magnitude making a significant measurement of the signal impossible.

The significance of the measured signal can be calculated by integrating the fitted components over the signal window thus determining the number of signal events n_{sig} and the number of background events n_{bkg}. Those can be used to estimate a proxy for the significance:

$$m = \frac{n_{sig}}{\sqrt{n_{sig} + n_{bkg}}} \tag{3}$$

In the case of the exponential background the resulting significance is $3.15 * 10^{-7}$. When assuming a constant instead of an exponential background as displayed in 11 the results are more significant with a corresponding proxy of 0.65.

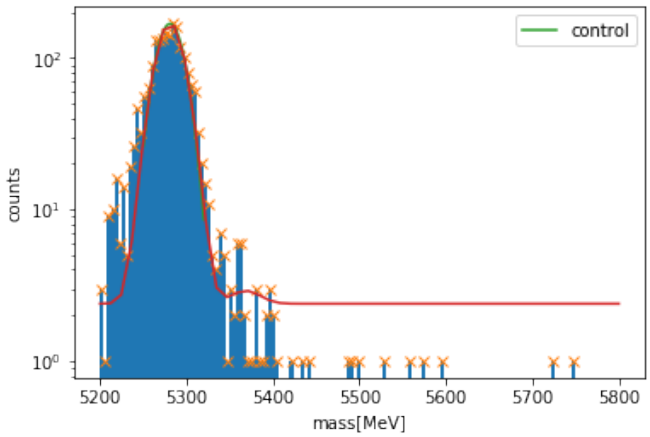

Figure 11: Fit for the total distribution using an exponential background

5 Conclusion

To select events corresponding to the B_s^0 first features are selected and tested using cross-validation. The features selected are all of kinematic nature implying plausible results. The classifier trained based on these features shows very promising results regarding the receiver-operating characteristic independent of the training set used.

Despite the good performance of the classifier the resulting fits show no relevant signal with a significance of the order of magnitude 10^{-7} when assuming an exponential background. Although there seems to be a visible peak in the data the exponential background seems overtakes the signal peak making it almost unrecognizable. When assuming a constant background over the viewed region the resulting significance of 0.65 is better but still far outside the realm of anything important.

A possible reason for this is that the threshold has been chosen too high resulting in a low number of remaining events making the assumption of an exponential background problematic. Another possible reason is a flawed selection of features. Also the distribution of the B^0 might be too broad and thus make the second peak unrecognizable. The sweighted data suggests, that the B^0 peak in the data might be significantly broader than the simulated one leading to problems with the fit. Because the control peak is estimated to be too narrow the remaing events have to be covered by the background making it very dominant. The reason for this descrepancy between signal and data is unknown.

References

[1] *Measurements of matter-antimatter asymmetries with the LHVb experiment*. Fakultät Physik, TU Dortmund. 2022.

[2] *Selection of $B_s^0 \to \Psi(2S)K_s^0$ events*. Fakultät Physik, TU Dortmund. 2020.

[3] R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.