Department of Physics and Astronomy
University of Heidelberg

Bachelor Thesis in Physics
submitted by

Lennart Henning Uecker

born in St. Wendel (Germany)

2020

Towards a test of lepton universality in $B^0 \to K^*(892)^0 e^+ e^-$ at high q^2.

This Bachelor Thesis has been carried out by Lennart H. Uecker at the
Physikalisches Institut in Heidelberg
under the supervision of
Prof. Dr. Stephanie Hansmann-Menzemer.

Abstract

The rare decay $B^0 \to K^{*0} e^+ e^-$ is studied in proton-proton collisions recorded by the LHCb experiment at CERN. The data set corresponds to an integrated luminosity of about 6 fb^{-1} at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. Specifically, the previously unexplored region of high squared-dilepton-mass $q^2 > 14$ GeV$^2/c^4$ is studied here. In order to separate signal decays from background, a multivariate analysis is performed, using a data set of $B^0 \to K^{*0} \mu^\pm e^\mp$ candidates as proxy for the background. A new set of observables is optimised to target signal decays in the high q^2-region. The Monte Carlo simulated signal events are corrected to describe the recorded data, with respect to this new set of observables using a boosting technique developed on the control channel $B^0 \to K^{*0} J/\psi (\to e^+ e^-)$. A preliminary fit to the B^0 mass is performed to estimate the yield of signal decays and an improvement over previous selection methods is found.

Zusammenfassung

Der seltene Zerfall $B^0 \to K^{*0} e^+ e^-$ wird bei hohen q^2-Werten ($q^2 > 14$ GeV$^2/c^4$) studiert, wobei q^2 das Quadrat der invarianten Masse der beiden Leptonen des Zerfalls ist. Dafür werden Daten des LHCb-Experimentes am CERN verwendet, das Proton-Proton Kollisionen bei einer Schwerpunktsenergie von $\sqrt{s} = 13$ TeV aufzeichnet. Die verwendeten Daten entsprechen einer integrierten Luminosität von etwa 6 fb^{-1}. Um die Singal-Zerfälle vom Hintergrundereignissen zu separieren wird eine multivariate Analyse ausgeführt, die als Ersatz für den kombinatorischen Hintergrund gemessene $B^0 \to K^{*0} \mu^\pm e^\mp$ Kandidaten benutzt. Dazu werden spezielle Observablen ausgewählt. Die simulierten Signal Zerfälle werden mit einer Boosting Technik korrigiert um bestmöglich die Daten des LHCb-Experiments zu beschreiben. Hierbei wird der Zerfall $B^0 \to K^{*0} J/\psi (\to e^+ e^-)$ als Referenz benutzt. Abschließend wird ein vorläufiger Fit der B^0-Masse durchgeführt und so die Anzahl der Signalereignisse abgeschätzt. Dabei wird eine Verbesserung im Vergleich zu früheren Methoden gefunden.

Contents

Chapter 1

Introduction

The advancement of particle physics has led to new experiments on ever larger energy and quantity scales. At the LHCb experiment collisions can be observed at rates of up to 40 MHz. Even after reduction through an elaborate trigger system, 750 MB of data are written to storage every second. As historical reference, CERN's hugely impactful Gargamelle experiment, which discovered weak neutral currents, took 4.8 million photos in about 8 years. A similar amount of events is recorded in a matter of minutes at LHC detectors.

Experiments that produce such large amounts of data are constructed in order to test the limits of the Standard Model and potentially discover physics beyond it. One important area in this search for physics beyond the Standard Model are rare decays of Standard Model particles. It is expected that new particles could influence the branching fractions of these decays or parameters such as the angular distributions. An exciting place to look for physics beyond the Standard Model are the rare B^0 decays $B^0 \to K^*(892)^0 \ell^+ \ell^-$, which have produced promising deviations from the Standard Model in earlier analysis.

Since these decays are very rare, even at the high production rate at LHCb, the separation of these decays from the background presents a large challenge. In this thesis machine learning methods are applied to improve the signal extraction. This approach has enjoyed great success as it is able to take advantage of the large amount of data that is recorded for every event much more effectively and efficiently than more traditional methods. The improved selection will be used to study the angular distribution as well as lepton universality at high values of the lepton-invariant mass square q^2. This will complement previous analysis in providing information on potential contributions of physics beyond the Standard Model in this channel.

Chapter 2

Theory

In the following Chapter a brief summary of the Standard Model will be given, followed by a short introduction to $b \to s$ decays. In addition, a dedicated machine learning technique, called *Gradient Boosted Trees*, is introduced, as well as two methods for comparing different data sets.

2.1 The Standard Model

The Standard Model of particle physics (SM) is, in its most modern formulation, a gauge quantum field theory that contains all known fundamental particles and their interactions. The Standard Model has been finalised in its current form since the existence of quarks has been shown experimentally in 1969 [1] and went on to make a lot of successful predictions. Most notably, the SM predicts the existence of the Higgs boson, which was experimentally confirmed in 2012 [2]. Nonetheless it is well understood that the Standard Model fails to incorporate General Relativity and thus breaks down at energies, where a proposed "graviton" would emerge. As such, it is referred to as an *effective* field theory and expected to eventually be expanded or superseded by a more general field theory. This introduction largely follows [3].

2.1.1 Fermions

The Standard Model includes twelve[1] spin-1/2 particles, so-called *fermions*, which can be sorted into two different groups: *quarks* and *leptons*. Each group consists of six particles sorted into three generations, as seen in Table 2.1.

Electrically charged particles of the second and third generation are not stable and decay into the lighter and stable particles of the first generation. For all these particles there are also anti-particles with the same mass but opposite charges. Only neutrinos are electrically neutral and do not interact electromagnetically. All particles interact via the weak force, while only quarks couple to the strong force. Quarks also don't exist as free particles but only in colour neutral bound states called *hadrons*. Hadrons are classified as *mesons* or *baryons* dependent on the number of quarks involved, two and three respectively.

[1] not counting anti-particles

	leptons				quarks			
	particle		q/e	mass [GeV/c^2]	particle		q/e	mass [GeV/c^2]
first	electron	e^-	-1	0.0005	down	d	$-1/3$	0.003
generation	-neutrino	ν_e	0	$< 10^{-9}$	up	u	2/3	0.005
second	muon	μ^-	-1	0.106	strange	s	$-1/3$	0.1
generation	-neutrino	ν_μ	0	$< 10^{-9}$	charm	c	2/3	1.3
third	tau	τ^-	-1	1.78	bottom	b	$-1/3$	4.5
generation	-neutrino	ν_τ	0	$< 10^{-9}$	top	t	2/3	174

Table 2.1: Matter particles sorted into three generation of quarks and leptons. Electric charges are expressed as fraction of the fundamental charge e. The neutrino mass boundaries are based on the sum of all flavours.

force	strength	gauge boson		spin	boson mass [GeV/c^2]
strong	1	gluon	g	1	0
electromagnetic	10^{-3}	photon	γ	1	0
weak	10^{-8}	boson	W	1	80.4
		Z^0 boson	Z	1	91.2

Table 2.2: List of gauge bosons and their respective strengths, normed on the strong force.

2.1.2 Bosons

The SM describes interaction between the aforementioned particles via exchange-particles, called *bosons*. The three forces correspond to quantum-field theories describing the exchange of *gauge bosons*, as listed in Table 2.2.

In addition to these, the Higgs boson was discovered as predicted by the SM as the only scalar boson, i.e. spin = 0, with a mass of $m_{Higgs} \approx$125 GeV/c^2. The Higgs boson is thought to mediate the process, which is responsible for the mass of an elementary SM particle.

2.1.3 The electromagnetic force

The electromagnetic force is described by quantum electrodynamics (QED) and only particles with electric charge participate. The charge is conserved in all interactions. The exchange particle, the photon, is massless. The interaction is not limited spatially, which is the reason, why electromagnetic interactions are most prominent in daily life, however its strength decreases with the inverse of distance.

2.1.4 The weak force

All fundamental fermions interact via the weak force, which is best described in combination with the electromagnetic force in electroweak theory. Its exchange particles $W^\pm$ and Z^0 are massive and therefore limit the range of this interaction severely. A common example of the weak force in action, the β decay, is based upon the unique ability of the weak force to change quark flavour.

2.1.5 The strong force

The strong force is described by quantum chromodynamics (QCD) and acts only upon particles carrying colour charge, i.e. quarks and gluons. There a 9 gluons, 8 of which carry colour charge and one neutral one, which therefore cannot participate in strong interactions. There are three different colour charges, commonly referred to by the colours blue (b), red (r) and green (g), as well as the respective anti-charges ($\overline{b}$, $\overline{r}$, $\overline{g}$). However these colour-charged particles cannot exist freely, but must be confined in hadrons, which are colour-neutral, i.e. composed of colour and anti-colour or three different colours.

2.2 Quark mixing and $b \to s$ transitions

The weak force has the ability to change quark flavour. A common example is the β^- decay as shown in Figure 2.1. Here a down quark changes into an up quark. The SM description of flavour changes involves the unitary Cabibbo-Kobayashi-Maskawa (CKM) matrix, where the square of the absolute values $|V_{ij}|^2$ are proportional to the probability of a transition from quark i to quark j.

$$V_{ij} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cd} & V_{cb} \\ V_{rd} & V_{ts} & tV_{tb} \end{pmatrix}$$

The diagonal entries are very close to one, as a transition within a generation is most likely. The size of the CKM entries decrease with the distance from the diagonal, meaning transitions between the third and the first family are the rarest. The CKM Matrix only describes transitions with a change of charge. Charge neutral flavour-changing transitions are therefore prohibited at tree-level and are very suppressed in the SM.

The decay $b \to s\ell^+\ell^-$ is a Flavour Changing Neutral Current (FCNC) decay, which in the Standard Model, is forbidden at tree level, but is allowed through loop and box diagrams. Prominent examples are penguin- and W box-diagrams, as shown in Figure 2.2.

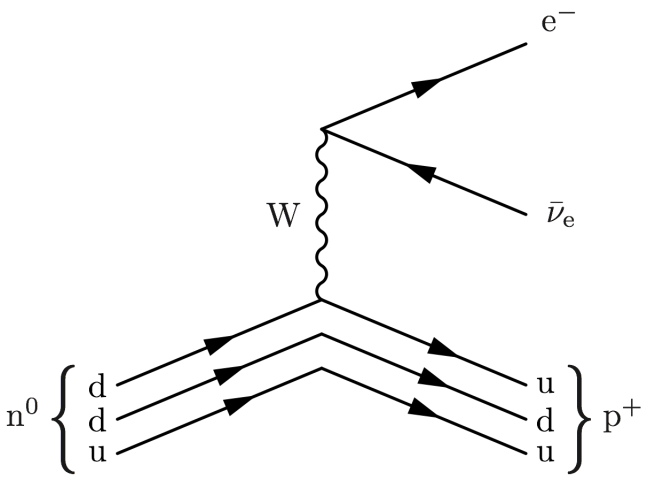

Figure 2.1: β^- decay, an example of a quark flavour changing weak interaction.

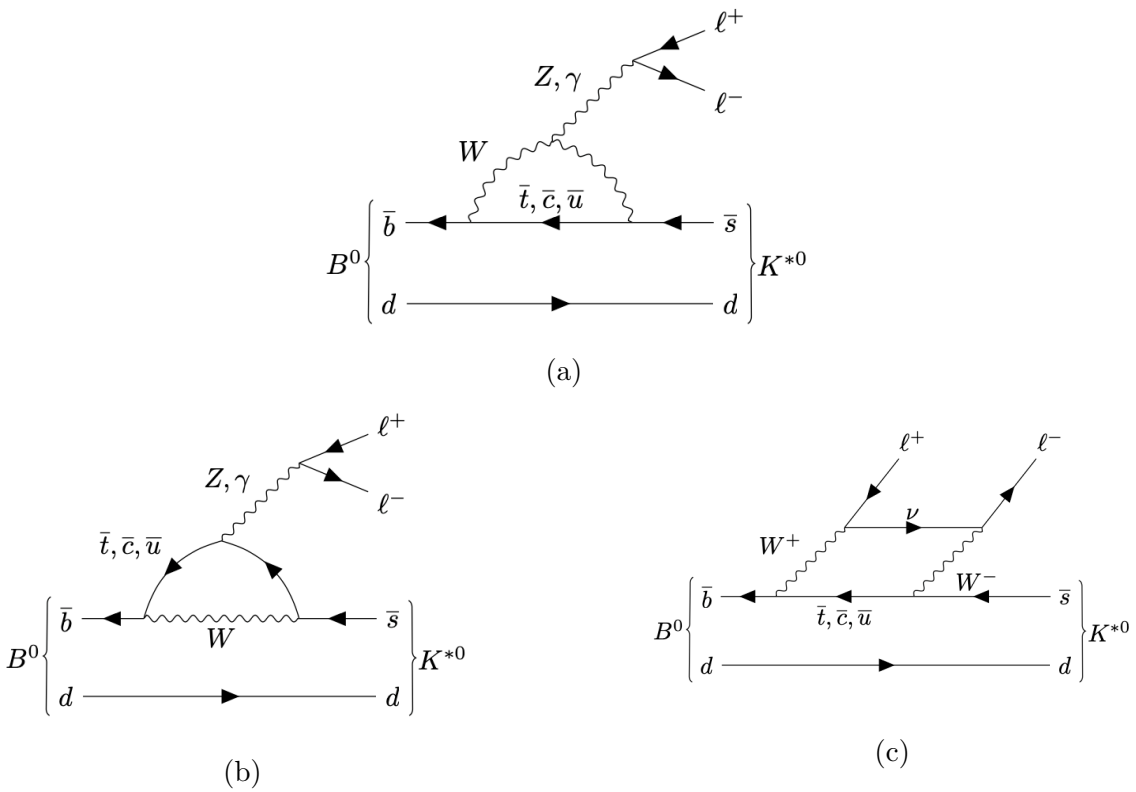

Figure 2.2: Loop-level diagrams of $B^0 \to K^{*0}\ell^+\ell^-$ decays involving $b \to s\ell^+\ell^-$ FCNC. a), b) are so called penguin diagrams; c) is a W box diagram.

The branching fractions of these decays are very small, typically below 10^{-6}. As such, they are often referred to as rare b decays. Virtual exotic particles in loops could change branching rations or angular distributions. These contributions are expected at the same level as rare decays, thus loop-dominated processes are an excellent laboratory to search for physics beyond the Standard Model, often called New Physics. They are well understood theoretically and the LHCb detector observes large enough numbers of B meson decays, that sensitivity to physics beyond the Standard Model could be expected in rare decays.
One prominent example for such a decay is $B^0 \to K^*(892)^0\ell^+\ell^-$ (in the following the resonance $K^*(892)^0$ will be referred to as K^{*0}), which has been used to search for New Physics in the angular distributions as well as to test lepton universality [4, 5]. Lepton universality is the assumption of the Standard Model, that different lepton generations are equal in all properties except mass. As such the measurement of the branching fraction ratio:

$$R_{K^{(*)}} = \frac{\int \frac{\mathrm{d}\Gamma(B^0 \to K^{*0}\mu^+\mu^-)}{\mathrm{d}q^2}\mathrm{d}q^2}{\int \frac{\mathrm{d}\Gamma(B^0 \to K^{*0}e^+e^-)}{\mathrm{d}q^2}\mathrm{d}q^2} \tag{2.1}$$

is expected to be 1 in the SM, with small and theoretically well understood phase space corrections related to the different masses of electrons and muons. However, some deviations have already been observed and are explored further.

2.3 Gradient boosted trees

This thesis relies heavily on a machine learning technique called gradient boosted trees (GBT), therefore a short introduction is given that is based on [6, 7]. Gradient boosted trees are a multivariate analysis (MVA) tool which is a much broader term for statistical analysis of multidimensional data.

2.3.1 Supervised machine learning

GBTs are applied to supervised learning problems, which means that training data $x_{ji} \in \mathbb{R}$ is used to predict a target variable $y_i \in \mathbb{R}$. Here i enumerates the data points (events), while j runs over the different entries (observables) of each data point. Which values y_i can assume depends on the task that the algorithm is supposed to solve, which may include regression, ranking or classification. The task of classifying data in two subsamples is most important for this thesis. In this case $y_i \in \{0, 1\}$ is a binary variable also known as label. The algorithm is defined by parameters $\vec{\theta}$, so that the training task is to find the best parameters $\vec{\theta}$ to match x_i and y_i, i.e. $\vec{\theta}$ is what is being learned. For this a objective function $\mathcal{L}(\vec{\theta})$ is introduced:

$$\mathcal{L}(\vec{\theta}) = L(\vec{\theta}) + \Omega(\vec{\theta}) \tag{2.2}$$

Here L is the training loss function, and Ω the regularisation function. $\hat{y}_i$ is the predication made by the model for the data point x_i, so that the loss function becomes a function of prediction $\hat{y}_i$ and true value y_i. For a simple regression task the mean squared error would be used $L(\vec{\theta}) = \sum_i (y_i - \hat{y}_i)^2$. For the XGBoost[2] [7] classifier that will be used in Chapters 5 and 6, a logistic loss function is used:

$$L(\vec{\theta}) = \sum_i [y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right)] \tag{2.3}$$

The regularisation term $\Omega(\vec{\theta})$ is used to limit the complexity of the model. This step is crucial to avoid overfitting, often also called overtraining. An overfitted model has lost its predictive power, since it has learned to recognise the training sample directly instead of picking up on the underlying pattern[3].

2.3.2 Tree ensembles

A classification and regression tree (CART) is the base unit of gradient boosted trees. Each tree starts at a node, where a binary decision is made, e.g. $(x_j)_i > 0$. The data is split into two subsets (branches), which can either create a new node and therefore a new decision, or create

[2] XGBoost is a specify implementation of gradient boosted Trees

[3] An examples would be a model trained to recognise faces that instead learned to use random noise, such as the picture background, for its predictions. It might predict all training samples perfectly, but will not make a correct identification from a different picture.

a leaf, i.e. end there. Each leaf is assigned a real value[4] w, called score. Several of these trees can be build and be combined to an ensemble. In an ensemble of N trees $f_n : \mathbb{R}^d \to \mathbb{R}$ and d observables per event, where $n = 1, ..., N$, the prediction is:

$$\hat{y}_i = \sum_{n=1}^{N} f_n(\vec{x}_i) \tag{2.4}$$

The space of all CARTs f is called $\mathcal{F}$. For the XGBoost model the regularisation term is defined as:

$$\Omega(\vec{\theta}) = \sum_{n=1}^{N} \Omega(f_n) = \sum_{n=1}^{N} (\gamma T_n + \frac{1}{2}\lambda \sum_{k=1}^{T_n} w_{n,k}^2) \tag{2.5}$$

Here T_n is the number of leaves in the n-th tree, $w_{n,k}$ is the weight assigned to the k-th leaf in the n-th tree. The variables γ and λ are model variables, so called hyperparameters, that adjust the strength of these regularisation terms and are set independently before training. Other hyperparameters for tree ensembles include constraints on the depth of a tree, the number of leaves or subsampling methods, e.g. a tree may be trained with only a fraction of the training data.

2.3.3 Tree boosting

So far no mention is made of how the training works as a procedure, however, the objective function $\mathcal{L}(\vec{\theta})$ is already introduced. This objective function is to be optimised, but as the space of $\vec{\theta}$ is essentially equivalent to $\mathcal{F}^N$, with $\mathcal{F}$ containing tree structure and leaf weights, traditional optimisation strategies are not applicable. Instead an additive model is chosen, where trees are build in succession and only the newest tree is optimised, while all prior trees remain fixed.

$$\hat{y}_i^{(0)} = 0 \tag{2.6}$$

$$\hat{y}_i^{(n)} = \sum_{j=1}^{n} f_j(\vec{x}_i) = \hat{y}_i^{(n-1)} + f_n(\vec{x}_i) \tag{2.7}$$

By inserting this recursive formula 2.7 into equation 2.2, the objective function of the k-th tree can be expressed based on the predictions made by all prior trees:

$$\begin{aligned} \mathcal{L}^{(k)} &= \sum_i [L(y_i, \hat{y}_i^{(k)})] + \sum_{j=1}^{k} \Omega(f_j) \\ &= \sum_i [L(y_i, \hat{y}_i^{(k-1)} + f_k(\vec{x}_i))] + \Omega(f_k) + const. \end{aligned} \tag{2.8}$$

Here i enumerates all events in the training sample. The loss function is then Taylor expanded to the second order at which point the complete objective is of quadratic order and can be minimised. The exact minimisation process in the space $\mathcal{F}$ is described in [7], it happens by

[4] This is the technical difference of CARTs to Decisions Trees, where only binary decisions are assigned to the leafs. However, this distinction is often ignored.

enumerating all[5] possible trees and assigning a gain value.

2.3.4 GBT as classifier

To use a gradient boosted tree ensemble for a classification task, a training data set is prepared. In this thesis, a Monte Carlo simulated sample of the signal decay is used, where $\vec{x}_i$ is called an event. It contains multiple observables x_{ji}, such as momentum or track quality of a decay particle of the signal decay. The signal sample also contains a label $y_i = 1$. The index i runs over the size of the sample $i \in \{1, 2...N-1, N\}$. The background training sample is build up similarly, with $\vec{x}_i$ taken from data. The label $y_i = 0$ is assigned and the index $i \in \{N+1, N+2, ..., M-1, M\}$ runs over the background sample. The loss function is the logistic function as seen in equation 2.3. The XGBoost algorithm has the added ability to not only return a binary decision, but also a probability for the event to be a signal event.

2.4 Comparison of two data sets

In this thesis large data sets are used and a focus is put on detecting and quantifying how much two samples differ, especially when each event is described by many observables. Two different methods that are used are introduced in this section.

2.4.1 The Kolmogorov-Smirnov two-sample test

The Kolmogorov-Smirnov two-sample test for goodness of fit [8] is a one-dimensional test that can be adapted to work for multidimensional samples by performing the test on just one entry of each event, e.g. compare how similar two samples are in their momentum distribution. The null hypothesis, both samples are taken from the same distribution, is tested. The test returns the so-called statistic, a real positive number, which will be referred to as KS-score in this thesis. The KS-Score can be used to reject the null hypothesis based on different certainty values. In this case it is used as a metric for difference in the samples. A score of zero would suggest, that the null hypothesis cannot be rejected. This test only observes the one-dimensional differences and can therefore not be used to make statements about multidimensional distributions. One intuitive example of failure of this test are two bivariate normal distribution with different non-diagonal entries in the covariance matrix. These would be classified as identical.

2.4.2 Multivariate analysis

The best detection of differences is achieved via multivariate analysis. Here a classifier is trained to detect the differences between subsamples of both samples, i.e. one as signal the other as background. Then the classifier is evaluated on the withheld data set and the Receiver Operator Characteristic (ROC) is recorded. The ROC is a curve of the true positive rate over the false

[5] or a well selected subsample if the training sample is too large

positive rate and characterises the diagnostic abilities of a binary classification tool. The area under the curve (ROC-AUC) is a simple metric, which can be used to quantify the performance of the classifier. Here a ROC-AUC score of 1.0 corresponds to perfect classification, while 0.5 indicates performance equivalent to random chance. Meaning one can be confident in the similarity of two samples, if the ROC-AUC is close to 0.5, as the MVA is not able to leverage the differences in the samples. A higher ROC-AUC indicates the existence of more pronounced differences.

Chapter 3

The LHCb experiment

The LHCb [9] is one of seven experiments connected to the Large Hadron Collider (LHC), from which it derives its name LHC-*beauty* (LHCb). It is dedicated to precision measurements in b and charm hadron decays.
In the following chapter the LHC is described as well as the LHCb detector and its trigger system.

3.1 The Large Hadron Collider

The Large Hadron Collider [10] is the world largest and most powerful particle collider. Situated at the CERN facility it utilises the roughly 27 km long underground tunnel, which used to house the Large Electron-Positron collider, to accelerate and collide protons at centre-of-mass energies of up to 13 TeV or lead ions at up to 5 TeV. Pre-accelerated protons at energies of 450 MeV are injected into two separate ultra-high vacuum beam pipes in which the LHC will further accelerate them to their final energies of 6.5 TeV (per beam). In total 10,000 superconducting magnets are installed. These dipole magnets keep the protons on their circular path and higher-order multipole magnets focus the beam and adjust the form of the proton bunches. The acceleration from the injection energy to the final 6.5 TeV is done in 16 radiofrequency (RF) cavities. This process takes about 20 minutes. The oscillating nature of the field used in this process also creates and maintains the bunches. An ideally timed proton will not be accelerated by the RF cavities, while slower or faster protons will be accelerated to match the nominal speed. Once a centre-of-mass energy of 6.5 TeV has been reached, the beams, travelling in opposite directions, collide at 4 different interaction points, each containing one of the four large experiments: The large general purpose detectors ATLAS and CMS, the heavy-ion focused detector ALICE and the b-physics focused spectrometer LHCb. In addition there are 3 smaller experiments TOTEM, LHCf and MoEDAL, with a fourth FASER to join them in 2021.

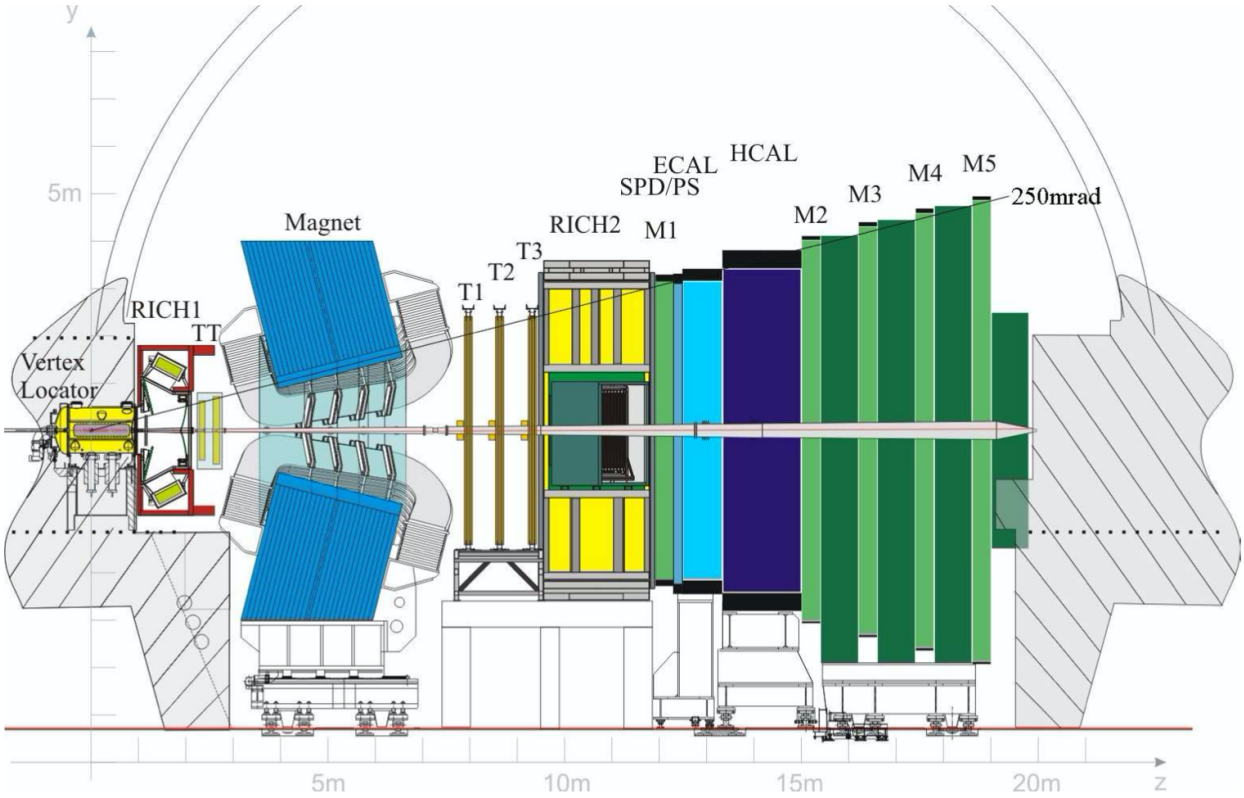

Figure 3.1: Cross section drawing of the LHCb detector in the non-bending plane, from [9].

3.2 The LHCb detector

The LHCb is a single-arm forward spectrometer, meaning it is designed to detect particles with pseudorapidity[1] $1.8 < \eta < 4.9$[2]. This choice is justified, since at LHC energies $b\bar{b}$ pairs are predominately produced in this forward (backward) cone. Approximately 25% of all $b, \bar{b}$ particles are within the acceptance of LHCb.
In Figure 3.1 the multitude of components, which make up the detector, can be seen. They are explored in more detail in this section.

3.2.1 The vertex detector

Due to the short lifetimes of heavy mesons, good particle tracking close to the interaction point is important, in order to accurately reconstruct the primary and secondary vertices. LHCb developed the VErtex LOcator (VELO) for this purpose, located around and downstream of the interaction point. Its 42 sensor modules are perpendicular to the beam and get as close as 8 mm to the beam axis. The geometry of the VELO is such that each particle within the acceptance of LHCb passes through at least 3 sensors. These modules then record the radial and polar position of a charged particle passing through.

[1] $\eta = \ln(\tan\theta/2)$ where θ is the polar angle with respect to the beam.
[2] Equivalent to a coverage of 10 mrad to 250 mrad.

3.2.2 The magnet

After particles leave the front part of the detector, they pass through a dipole magnet, with an integrated magnetic field of 4 Tm over a 10 m track. By calculating the curvature of a particle moving through the magnetic field, it is possible to reconstruct the momentum of the particles. The magnetic field is designed in such a way as not to interfere with the sensors located nearby.

3.2.3 The tracking system

The VELO is complemented by further tracking stations. The Trigger Tracker (TT) is located in front of the magnet and covers the full angular acceptance of the detector. The tracking stations T1, T2 and T3 are situated downstream of the magnet. They are split into inner and outer trackers[3] utilising different technologies. The Trigger Tracker and the inner tracker use silicon microstrip sensors, while the outer tracker consist of drift-tube detectors. The Trigger Tracker is important for the measurement of low momentum charged particles which are bent away by the magnet, while T1-T3 allow for precise momentum measurements over a large acceptance area. These precise momentum measurements are important for the reconstruction of the invariant mass of b hadrons.

3.2.4 The RICH detectors

The LHCb detector utilises two ring-imaging Cherenkov detectors (RICH) for particle identification (PID). RICH 1 is located upstream of the magnet and covers low momentum (1-60 GeV/c) charged particles within the full angular acceptance. RICH 2 is located downstream of T3 and covers the high momentum range ($\gtrsim$ 15 GeV/c). The angular acceptance of RICH 2 is smaller than of RICH 1, but includes the production region of these high momentum particles. When particles pass through a medium with a speed higher than the speed of light in the medium, it emits Cherenkov radiation. The emission angle $\cos\theta_c = c/(nv)$ is related to the velocity of the particle v as well as the refractive index n. From the measurement of θ_c the velocity can be calculated and, together with the measurement of the momentum in the tracking system, the mass can be reconstructed. To account for the different target momentum ranges, materials of different refractive indices are used in RICH1 and RICH2. The RICH detectors at LHCb are essential to differentiate, for example, pions from kaons in the observed B meson decays.

3.2.5 The calorimeters

Further parts of the PID system are an electromagnetic calorimeter (ECAL) followed downstream by a hadronic calorimeter (HCAL). They help identify electrons, photons and hadrons by measuring their energy deposition in the various layers of the calorimeter. These measurements are important for flavour tagging and are also used in the first trigger level.
In front of the ECAL, a scintillator pad detector (SPD) detects charged particles followed by a

[3] referring to the distance from the beam pipe.

pre-shower detector. The ECAL employs the shashlik technology, a sampling scintillator and lead structure with a total depth of 25 radiation lengths X_0. As in T1-T3 the structure changes to accommodate the different hit densities near the beam pipe.
One large challenge in the reconstruction of the energy of electrons is their frequent emission of bremsstrahlung photons upstream of the magnet due to interaction with the detector material. When this happens, the electron momentum is underestimated. ECAL energy deposits compatible with background photons coming from the electron track are searched for and their energy is added back to the electron to recover its momentum before bremsstrahlung emission. However, this recovery procedure often fails and the bremsstrahlung energy is lost. This effect is interesting for this thesis as the loss due to bremsstrahlung may cause an underestimate of the reconstructed dielectron and $K^{*0}e^+e^-$ invariant masses.
The HCAL is constructed of iron and scintillating tiles, for absorption and detection respectively. The tiles are oriented parallel to the beam axis. The HCAL contains about 5.6 hadronic interaction lengths.

3.2.6 The muon system

The muon system is composed of five stations, M1 in front of the calorimeters, M2-M5 downstream as the last pieces of the detector. The system provides the identification of muons. The first station M1 is placed apart from the others to improve the resolution of the transverse momentum, p_T. Iron absorbers are placed between the detectors to select penetrating muons, so that only muons with a momentum larger than 6 GeV/c cross all stations. All stations provide location information and the p_T measurements are used for triggering. As for T1-T3 and the calorimeters the structure changes with the distance from the beam pipe. The granularity and thus the spatial resolution decreases in the outer layers.

3.3 The trigger system

The trigger system at LHCb [11] is redesigned for Run 2, to allow for data taking with minimal dead time at LHC maximal bunch crossing rate of 40 MHz. Since the subdetectors can be read out at a maximum rate of 1 MHz, a hardware trigger, the Level 0 (L0) trigger, determines which events are kept based on information from ECAL, HCAL and the muon stations. It reduces the relevant event rate to approximately 1 MHz.
The next stage of data processing is the High Level trigger (HLT), which is divided in two stages HLT1 and HLT2, these are software triggers. HLT1 selects events based on partial reconstruction, reducing the event rate to 110 kHz. These events are then buffered and fully reconstructed in HLT2 after the detector is calibrated and aligned. HLT2 reduces the event rate to 12.5 kHz. All events passing HLT2 are then written to storage for later analysis.

Chapter 4

Signal selection

This thesis uses data collected at LHCb during Run 2 in the years 2015 through 2018. This corresponds to an integrated luminosity of about 6 fb^{-1} of proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 13$ TeV.

4.1 Data samples

Recorded data from three different decay channels is being used.

- $B^0 \to K^{*0} e^+ e^-$ is the signal to be studied in this thesis, the full Run 2 data is used.
- $B^0 \to K^{*0} J/\psi (\to e^+ e^-)$ is used as a reference channel for threshold selection studies as well as a control channel for discrepancies between recorded data and simulation. Due to availability only the 2015 and 2016 data is included.
- Fake $B^0 \to K^{*0} e^\pm \mu^\mp$ candidates from recorded data are used as background sample for the multivariate analysis. Since it is a forbidden decay in the SM, the data is made up of combinatorial background as well as partially reconstructed double-semileptonic decays. Only 2015 and 2016 data is used.

Additionally to the data samples the following Monte Carlo simulated events (MC) are used.

- $B^0 \to K^{*0} e^+ e^-$ signal events, for the multivariate analysis training.
- $B^0 \to K^{*0} J/\psi (\to e^+ e^-)$ as reference channel.

These samples are produced using the Gauss simulation program.

4.2 Stripping and preselection

After the events are reconstructed a first selection is applied, this so-called **stripping** sorts the data based on their decay topology. For this study data is stripped using the line `BuLLK`. In this stripping cuts are applied to the momentum and transverse momentum of the final state

category	requirements
misc.	number SPD Hits < 450 at least one PV reconstructed
ee	m < 5500 MeV/c^2 χ^2_{Vertex}/ndof < 9 separation from Vertex $\chi^2 > 16$ track distance from PV $\chi^2 > 9$
π, K, K^{*+},	$p_T > 400$ MeV/c m < 2600 MeV/c^2 track distance from PV $\chi^2 > 9$
K^{*0}	abs(m-m_{PDG}) < 300 MeV/c^2 $p_T > 500$ MeV/c Track Isolation $\chi^2 > 30$
B	abs(m-m_{PDG}) < 1500 MeV/c^2 χ^2_{Vertex}/ndof < 9 $\chi^2_{\text{IP}} < 25$ DIRA >0.9995 PV separation $\chi^2 > 100$

Table 4.1: Partial stripping requirements of `BuLLKee_line`: This list does not reflect the full amount of the selections made in the stripping, only the criteria relevant for this analysis are shown. A full documentation can be found in the technical documentation [12].

particles, as well as requirements on the quality of track, primary vertex (PV) reconstruction and particle identification. A partial list of these requirements can be seen in Table 4.1.
In addition to the stripping, further **preselection** criteria are imposed, these rectangular cuts are designed to reduce the amount of background. The chosen cuts are listed in Table 4.2. The cuts on the electrons are applied symmetrically on both electrons, and are analogously defined for use on the $B^0 \to K^{*0} e^{\pm} \mu^{\mp}$ background channel.
In effect, a pair of tracks with opposite charge, which are identified to be electrons, and a K^{*0} candidate form a signal candidate. The K^{*0} candidate is formed by two oppositely charged tracks identified as a π^- and a K^+. It is required for all of these tracks to have a significant impact parameter to all primary vertices in the event. From these tracks, a common vertex is reconstructed, which in turn is required to be of good quality. This so-called secondary vertex is where the B^0 candidate decays. The B^0 mesons momentum is required to have a small impact parameter to one of the primary vertices, but still to be significantly displaced from that same PV. The direction angle θ_{DIRA} describes the deviation from the B^0 momentum reconstructed from its decay products and the vector connecting the PV to the B^0 decay vertex and is required to be small, i.e. $\cos\theta_{\text{DIRA}} > 0.9995$. In addition, the invariant mass of the B^0 meson is reconstructed from its decay products. It is required to be close to its nominal value as quoted in the particle data group m_{PDG}. A similar, although more strict, mass requirement is also posed to the reconstructed invariant mass of the K^{*0} candidates.

particle	cut	explanation
K^{*0}	abs(m-m_{PDG}) < 100 MeV/c^2	Difference between reconstructed mass and literature value.
K	`GhostProb` < 0.4	Probability that track is a ghost track.
	χ^2_{Track}/ndof < 3	Quality of the track reconstruction.
	`hasRich` == 1	RICH data is recorded.
	p_T > 250 MeV/c	Transverse momentum.
	`PIDK` > 0	Identification as an kaon.
	`ProbNNk`*(1-`ProbNNp`) > 0.05	Neural network based PID probability kaon and not proton.
	χ^2_{IP} > 9	Significant impact parameter to all PV's.
π	`GhostProb` < 0.4	Probability that track is a ghost track.
	χ^2_{Track}/ndof < 3	Quality of the track reconstruction.
	`hasRich` == 1	RICH data is recorded.
	p_T > 250 MeV/c	Transverse momentum.
	`ProbNNpi` * (1-`ProbNNk`) *(1-`ProbNNp`) > 0.1	Neural network based PID probability pi and not kaon or proton.
	χ^2_{IP} > 9	Significant impact parameter to all PV's.
e	`GhostProb` < 0.4	Probability that track is a ghost track.
	χ^2_{Track}/ndof < 3	Quality of the track reconstruction.
	`hasRich`(`CALO`) == 1	RICH (CALO) data is recorded.
	p_T > 500 MeV/c	Transverse momentum.
	p > 3000 MeV/c	Total momentum.
	`LOCalo_ECAL_region` ≥ 0 (or) `LOCalo_ECAL_x(y)` projection > 363.3 (282.6)	Cut on ECAL hit location.
	`PIDe` > 2	Identification as an electron.
	`ProbNNe` > 0.2	Neural network based PID probability.
	χ^2_{IP} > 9	Significant impact parameter to all PV's.
	p_T no Brem > 200	Transverse momentum without added bremsstrahlung corrections.
	`InAccEcal` = 1	The extrapolation of the track reaches the ECAL.
J/ψ	p_T	Transverse momentum of an intermediate J/ψ.
misc.	Vetoes preselection	Vetoes against particle misidentification.
	Trigger preselection	Cuts based on triggers that selected the event.

Table 4.2: Preselection cuts, which were constructed by the Bd2KstEE-Angular analysis group at LHCb.

4.2.1 Backgrounds sources

Peaking background refers to B decays, which are falsely reconstructed in the detector as a signal decay and which results in an invariant mass close to the mass of the signal. This effect is exacerbated by the broad bremsstrahlung tails, which exist in decays with electrons, thus the signal mass region is larger. One potential source of these backgrounds are decays $B^0 \to K^{*0} J/\psi$ or $B^0 \to K^{*0}\psi(2S)$, where the J/ψ and $\psi(2S)$ decays into an electron pair. These decays can be vetoed by excluding the q^2-regions corresponding to the according masses. However, misidentification of tracks, such as an electron or positron identified as daughter particle of the K^{*0} and vice versa would skew the q^2 value so that the q^2-cut no longer catches the decay. Therefore preselection cuts are introduced, which are designed to veto decays, where the hadrons could be misidentified as electrons and vice versa. Additional tighter cuts on $m(K^+\pi^-)$ to be consistent with the nominal mass of K^{*0} are applied to veto such candidates. After these vetoes are applied the most important remaining peaking backgrounds are:

- $B^+ \to K^+\pi^+\pi^- e^+ e^-$, where π^+ is missed, therefore the B^+ is reconstructed as B^0 with missing mass.
- $B^0 \to K^{*0}\psi(2S)(\to e^+e^-)$, charmonium resonance decay, since the momentum reconstruction is relatively poor, not all of these decays are removed by the q^2-cut.
- $B^+ \to K^+\pi^+\pi^-\psi(2S)(\to e^+e^-)$ tree level decay, where π^+ is missed and B^+ is reconstructed as B^0 with lower mass.

Combinatorial background refers to the combination of particles from decays, that are misidentified to belong to a single signal decay. The larger part of this thesis is focused on a multivariate analysis to reduce contributions from this effect.

4.3 Dilepton invariant mass squared (q^2) range

The signal decay can be cut into several subsamples, dependent on the invariant mass squared of the final state electrons. The $B \to K^{*0}\ell\ell$ differential decay rate $\frac{\mathrm{d}\Gamma}{\mathrm{d}q^2}$ varies as a function of q^2, as sketched in Figure 4.1. The charmonium resonances included in the figure stem from the tree level decay $B^0 \to K^{*0} J/\psi(\to e^+e^-)$ and $B^0 \to K^{*0}\psi(2S)(\to e^+e^-)$. These $b \to sc\bar{c}$ decays dominate the $b \to s\ell\ell$ decays of our signal in the corresponding q^2 regions. The q^2 region above the $\psi(2S)$ resonance has so far been excluded from analysis of this decay. It is the focus of this thesis to explore that region.

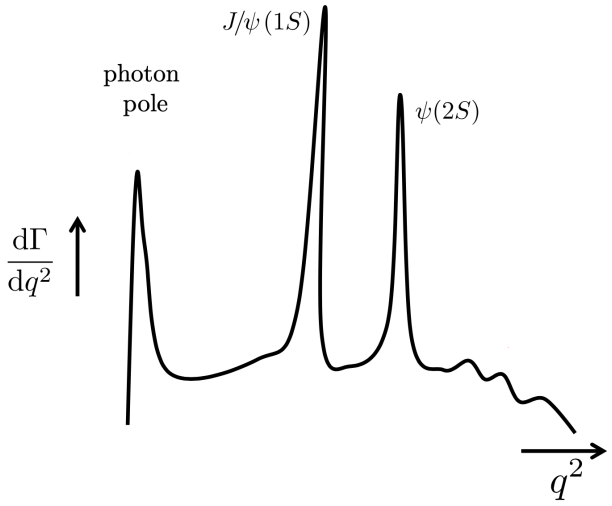

Figure 4.1: Differential decay rate of $B \to K^{*0}\ell^+\ell^-$. Adapted from [13].

4.4 Choice of observables

For the multivariate analysis to be performed in this thesis a set of observables has to be selected from all the available information collected about a decay. The goal is to select observables that are most suited to separate signal decays from combinatorial background. For the selection of observables there are a few criteria, which need to be considered. The observables need to be well reproduced in the Monte Carlo simulation, this can be tested by comparing the data and simulation for a reference decay, such as $B^0 \to K^{*0} J/\psi(\to e^+e^-)$. The observables need to have high separating power, meaning that the samples have to be significantly different between signal and background, when the observable is included in the data. The observables need to be uncorrelated to important analysis quantities, such as the angles used for the angular analysis, the reconstructed B^0 mass $m(K^{*0}e^+e^-)$ and q^2. Also the observables should not contain the same information twice, i.e. they should be only weakly correlated with each other. In addition, observables have to be vetoed when they are able to separate e from μ, since one sample used to train the MVA contains a muon in its final state. There are also several incentives to use as few observables as possible, such as computational time and protection against overfitting. The Kolmogorov-Smirnov two-sample test score is calculated over all continuously distributed observables and a large set of them is selected, based on their KS-scores. At this stages vetoes are applied, which reject observables not well reproduced in the simulation or influenced by lepton flavour. Next the correlation between all observables is calculated and highly correlated ones are removed, as, for the purpose of the MVA, they contain the same information. In a last step the observables are rejected, if they do not contribute to the trees build in the MVA. Leftover are 20 observables, which is a sufficiently small set to work with. They are listed and a short explanation of their meaning is given in Table 4.3. These observables are used for the assessment of the simulation corrections as well as the training and application of the MVA used to select signal decays from signal candidates and veto combinatorial background.

observable	explanation
B_DIRA_OWNPV	Direction angle of the B meson
B_ENDVERTEX_CHI2	B decay vertex χ^2
B_FDCHI2_OWNPV	Flight distance significance in units of χ^2 for the B meson
B_FD_OWNPV	Flight distance of the B meson with respect to the PV
B_IPCHI2_OWNPV	Impact parameter χ^2 of the B meson w.r.t. the PV
B_PVandBDTF_B_PT	Transverse momentum of the B meson
B_PVandBDTF_B_THETA	Angle between momentum of the B meson and beam axis (equiv. to η)
B_PVandBDTF_Jpsi_ETA	Pseudorapidity of the J/ψ meson
B_PVandBDTF_Jpsi_PT	Transverse momentum of the J/ψ meson
B_PVandBDTF_K_PT	Transverse momentum of K^+ meson
H_Max_Eta	Highest hadronic pseudorapidity
H_Max_PT	Highest hadronic transverse momentum
H_Min_IPCHI2_OWNPV	χ^2 of the smallest hadronic impact parameter
Jpsi_ENDVERTEX_CHI2	χ^2 of the J/ψ decay vertex
Jpsi_FDCHI2_OWNPV	Flight distance significance in units of χ^2 for the J/ψ meson
Kstar_ENDVERTEX_CHI2	χ^2 of the K^{*0} decay vertex
Kstar_IPCHI2_OWNPV	χ^2 of the K^{*0} impact parameter w.r.t. its vertex
Kstar_PT	K^{*0} transverse momentum
Kstar_eta	K^{*0} pseudorapidity
L_Max_PT	Maximum leptonic transverse momentum

Table 4.3: List of observables used in the MVA. The Prefix B_PVandBDTF, means that the variable was calculated using the decay tree fitter tool, with constraints on the PV and the B mass.

Chapter 5

Corrections of the simulated events

In this thesis a classifier is trained, using a simulation of the signal decay. The simulated events deviate from the measured events as the interaction of the particles with the detector cannot be simulated perfectly. Weights are added to events in order to correct these effects, as they could manifest later in the analysis, e.g. in a non-optimal or even biased signal selection.

5.1 Method of reweighting

In this thesis a Gradient Boosted Reweighting (GBReweighter) method [14] is used. This method is based on gradient boosted trees introduced in Section 2.3. The loss function is adapted, so that each node tries to maximise the differences in distribution between Monte Carlo simulated events (MC) and measured events (data) in the leaves. To each leaf a correction factor is assigned as score w, which corrects the differences in distribution between measured events and MC events. The next iteration of the boosting process then focuses on a region, which shows large differences, when including weights of the previous tree. In this way, $\hat{y}$ is no longer a prediction, and the loss function no longer penalises deviations from $\hat{y}$ and y. $\hat{y}_i$ is the weight assigned by the reweighting process to the event x_i. Identically to the GBTs the GBReweighter should be trained on a different sample than it is applied to later.

5.2 Reweighting

As no pure enough sample of signal decays is available to use directly in the reweighting, $B^0 \to K^{*0} J/\psi(\to e^+e^-)$ is used as a reference decay. In the reconstruction of the decay chain both electrons are constraint to the J/ψ mass. This gives a handle on the energy loss due to bremsstrahlung, which will result in a narrow B^0 mass peak. This is useful for background subtraction via the ${}_s\mathcal{P}lot$ [15] technique. The mass fit used for this background subtraction is shown in Figure 5.1. Furthermore the reference channel shares most of the kinematics with the signal decay. This means that GBReweighter is trained and evaluated using this decay, before it is applied to signal MC samples.
Reweighting all the observables used in the classification is not advisable, as this process might

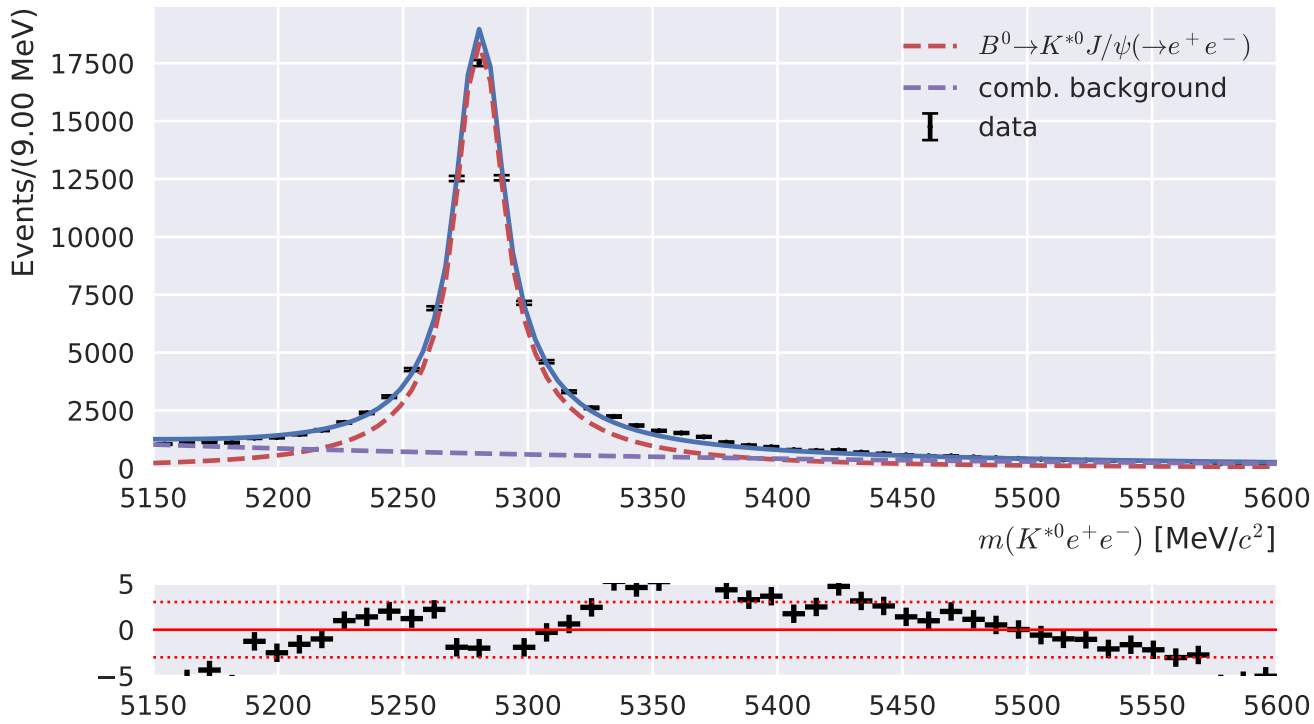

Figure 5.1: Mass fit used to subtract combinatorial background from the reference sample $B^0 \to K^{*0} J/\psi(\to e^+ e^-)$.

become unstable and the different kinematics of the reference channel also have to be respected. The observables to be reweighted are selected based on two criteria, firstly the observables show the differences between the data sample and the MC sample and secondly, selected observables should not show differences between the signal and reference decay, so that the reweighting can be applied to the signal MC. Thus observables that are selected pertain to the B meson and have the highest Kolmogorov-Smirnov two-sample test scores between data and MC, as listed in Table 5.1. The data and MC samples are each split into `A` and `B` subsamples of the same size.

5.2.1 Implementation of the Gradient Boosted Reweighter

The reweighting process can be subdivided into 4 distinct steps, where the GBReweighter is trained, evaluated and applied. These steps are listed in Table 5.2.

In the initial evaluation, like discussed in subsection 2.4.2, a classifier is used to identify differences between events in recorded data and the MC sample. The classifier is a XGBoost instance, with hyperparameters very similar to those chosen in Chapter 6 for the main classification task. All observables selected in Chapter 4 are used for training. The `A` subsamples are used for training and the classifier is evaluated on the `B` subsamples. The blue ROC-curve is generated in Figure 5.2. This ROC-curve shows that the classifier is able to pick up on some differences.

The GBReweighter is trained on the `B` subsamples and is then applied to the `A` subsample of the MC sample. The hyperparameters for the training are listed and explained in Table 5.3. A

observable	explanation	KS-score
$B:\ p_T$ (PV and BDTF)	Transverse momentum of the B meson, reconstructed with the B Decay Tree Fitter and restraint to the primary vertex.	0.020
$B:\ \chi^2_{ENDVERTEX}$	χ^2 of the fit to the position of the decay vertex for the B meson.	0.017
$B:\ DIRA$ (OWNPV)	The direction angle of the B meson, i.e. the angle between the line from primary vertex to the end vertex and the sum of the 4-momentum vectors of the decay products.	0.028
$B:\ \chi^2_{IP}$ (OWNPV)	χ^2 of the fit of the position the primary vertex for the B meson.	0.052
$B:\ \chi^2_{FD}$ (OWNPV)	χ^2 of the flight distance for the B meson.	0.033

Table 5.1: Observables used in the reweighting process.

step	tool	training	application	output
Initial evaluation	classifier	MC:A data: A	MC: B data: B	ROC-curve (blue)
Training reweighter	GBReweighter	MC: B data: B	MC: A	MC: A(r) (reweighted)
Evaluation	classifier	MC: A(r) data: A	MC: B data: B	ROC-curve (green)
Application	GBReweighter	-	signal MC	reweighted signal MC

Table 5.2: The four steps of the reweighting process, as well as the samples used in each step and the output that is generated. The ROC-curves are drawn in Figure 5.2.

comparison of the marginal density distributions for data sample, MC sample and reweighted MC sample are plotted in Figures 5.3 and 5.4. Only the first five observables are used in the reweighting process.

In the next evaluation step, a new XGBoost instance is trained on the reweighted MC `A(r)` subsample and the data `A` subsample. This classifier is also evaluated over the `B` subsamples and the green ROC-curve in Figure 5.2 is produced. The important fact here is that the reweighted (green) line is consistently lower than the blue line, which was created without reweighting. Both are calculated on the same sample, thus the differences are not statistical fluctuations. This indicates that GBReweighter has indeed reduced the differences between data and MC. It is also important to note that both lines are close to the red line, which implies that both classifiers do not perform much better than random guessing and therefore the difference is small. In addition, it has to be noted that both lines dip into negative values. This is an artefact of the negative weights assigned by the ${}_s\mathcal{P}lot$ technique to the data sample and does not reflect on the quality of the reweighting.

Since the GBReweighter produced satisfactory results for the training and evaluations samples of the reference channel, it is now applied to the signal MC sample. The GBReweighter calculates weights for the signal MC sample without further training. The effect of this reweighting can be seen in the marginal density distributions drawn in Figures 5.5 and 5.6. The most prominent changes are seen in χ^2_{FD} of the B meson, indicating that the flight distance significance of the B meson is larger in data, than it is originally simulated. The same effect is seen for the J/ψ χ^2_{FD}. The K^{*0} χ^2_{IP} is changed to contain less events with a low impact parameter significance. The marginal distribution of the B direction angle is changed to include more events with a value very close to one. The marginal density distributions of the remaining observables retain their shape in the reweighting. As an additional check on the stability of the algorithm a density distribution of the weights is plotted in Figure 5.7. The distribution indicates that a small number of events are assigned weights close to zero, therefore being effectively removed from the MC sample. The majority of events is assigned weights between 1 and 1.5. This is healthy behaviour for such an algorithm, and shows that the results can be applied in the analysis.

parameter	explanation	setting
`n_estimators`	The number of trees that are created. Higher values lead to better results, but can also lead to overfitting.	400
`learning_rate`	Step size shrinkage. This parameter limits the impact each additional tree has. Interacts with `n_estimators`.	0.02
`max_depth`	Maximum depth of a tree. Increases the complexity of model. Large values increase risk of overfitting	3
`min_samples_leaf`	Minimum number of events that have to remain in each leaf. Large values decrease risk of overfitting.	500
`loss_regularization`	Adds a regularisation term to the loss function.	10
`subsample`	The fraction of data, that is randomly chosen to train an individual tree. Lower values prevent overfitting	0.8

Table 5.3: Parameters used for GBReweighter.

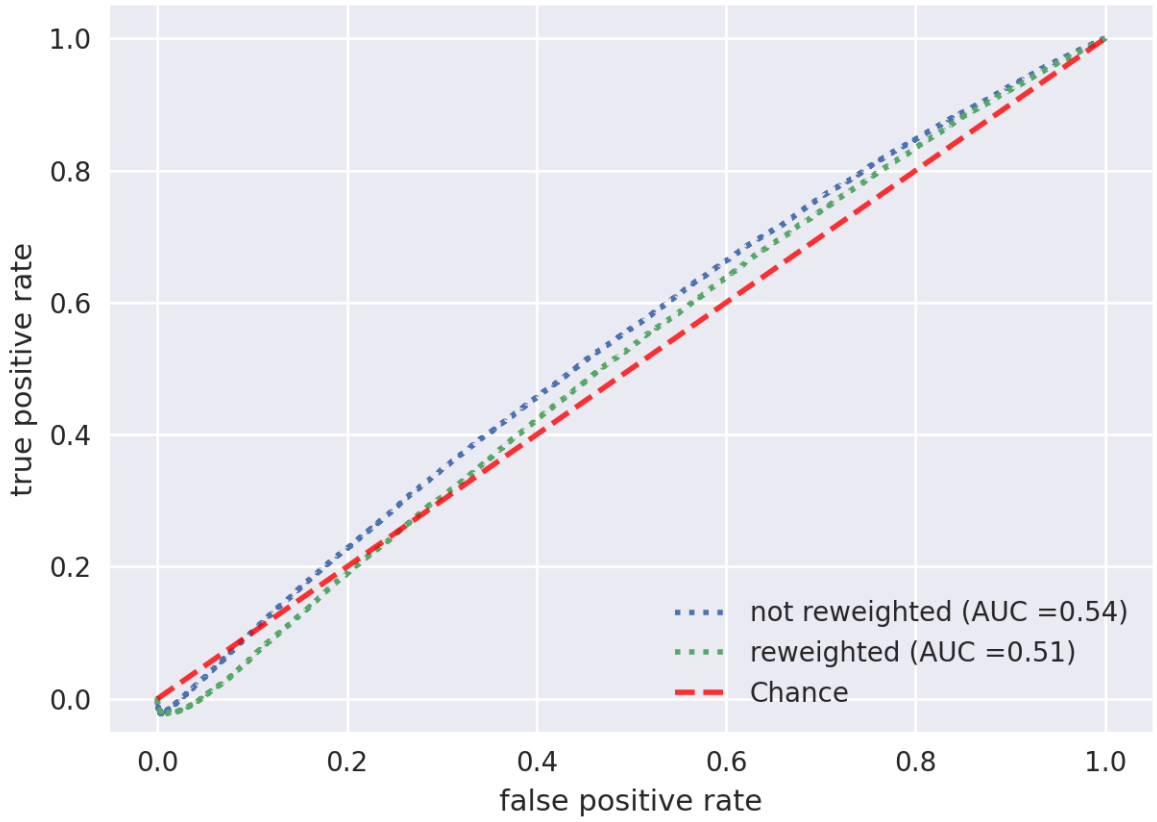

Figure 5.2: Receiver Operating Characteristic (ROC) of the two XGBoost instances, trained over the `A` samples, evaluated over the `B` samples. One instance was trained using the reweighted simulation `A(r)`, the other the uncorrected simulation.

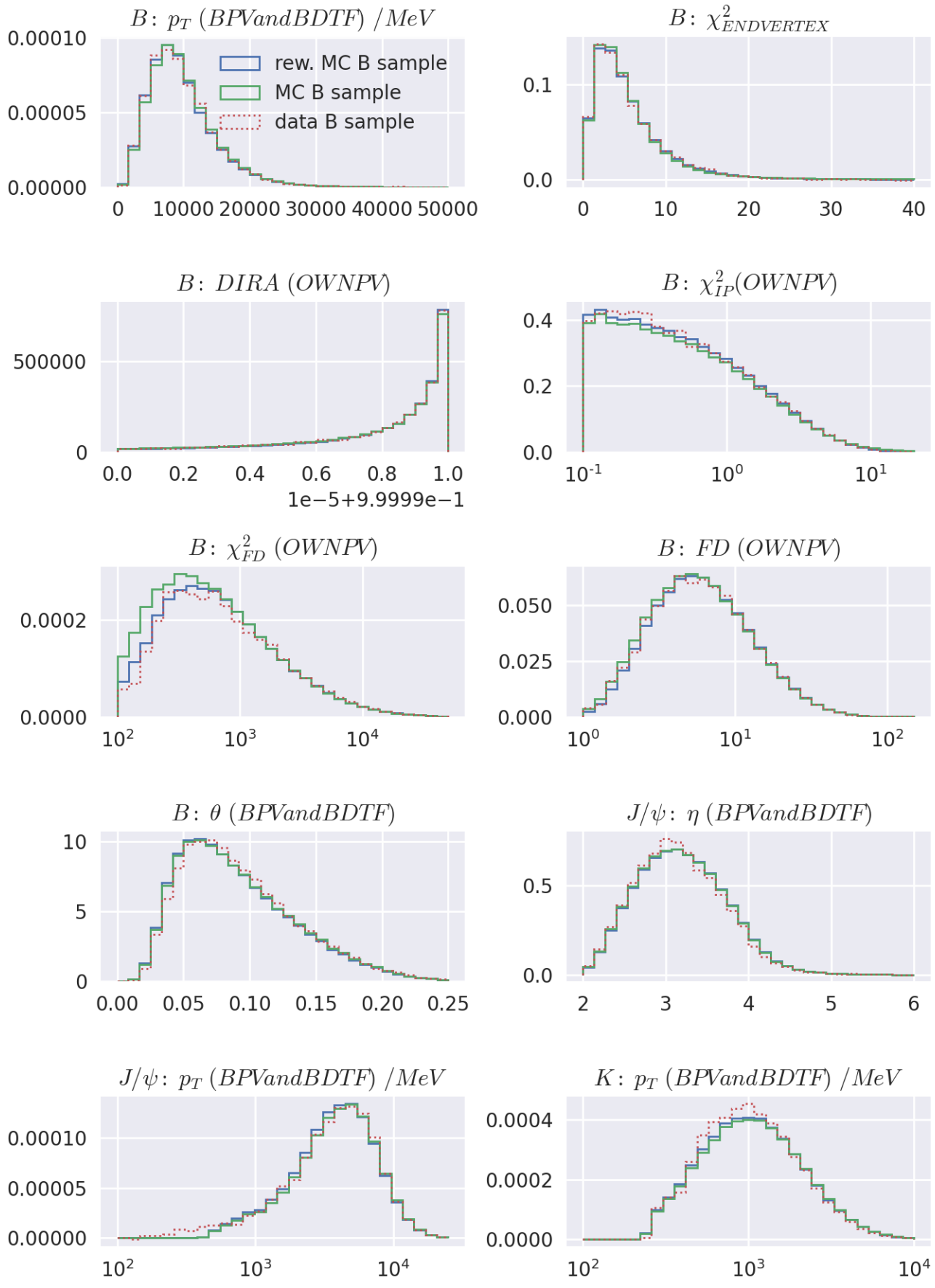

Figure 5.3: Marginal normalised distribution of MC and data of $B^0 \to K^{*0} J/\psi(\to e^+e^-)$ for all observables used in classification before and after reweighting (1/2). The legend in the first subplot applies to all.

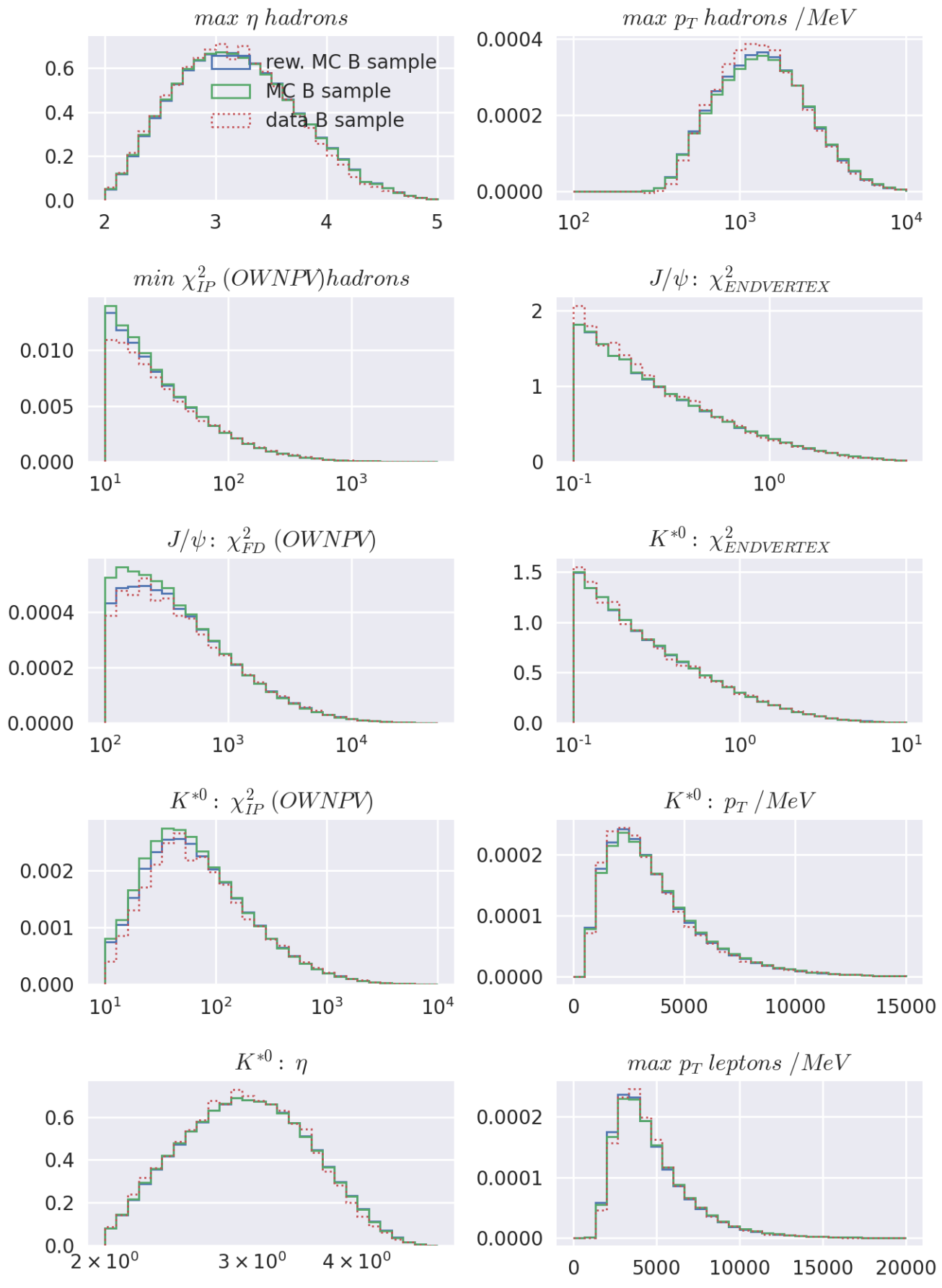

Figure 5.4: Marginal normalised distribution of MC and data of $B^0 \to K^{*0} J/\psi(\to e^+e^-)$ for all observables used in classification before and after reweighting (2/2). The legend in the first subplot applies to all.

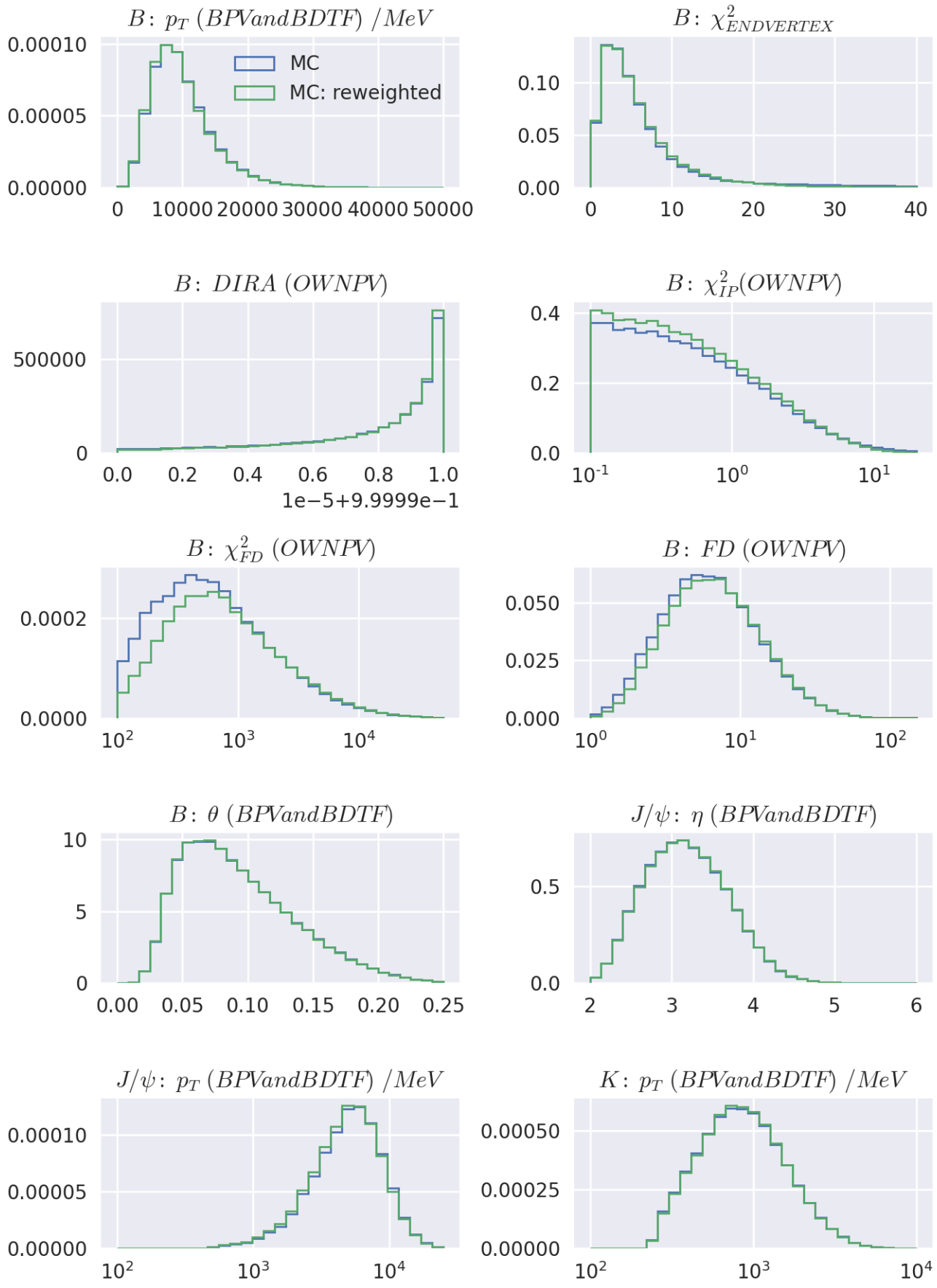

Figure 5.5: Marginal density distribution of the $B^0 \to K^{*0}e^+e^-$ simulation with and without weights (1/2). The legend in the first subplot applies to all.

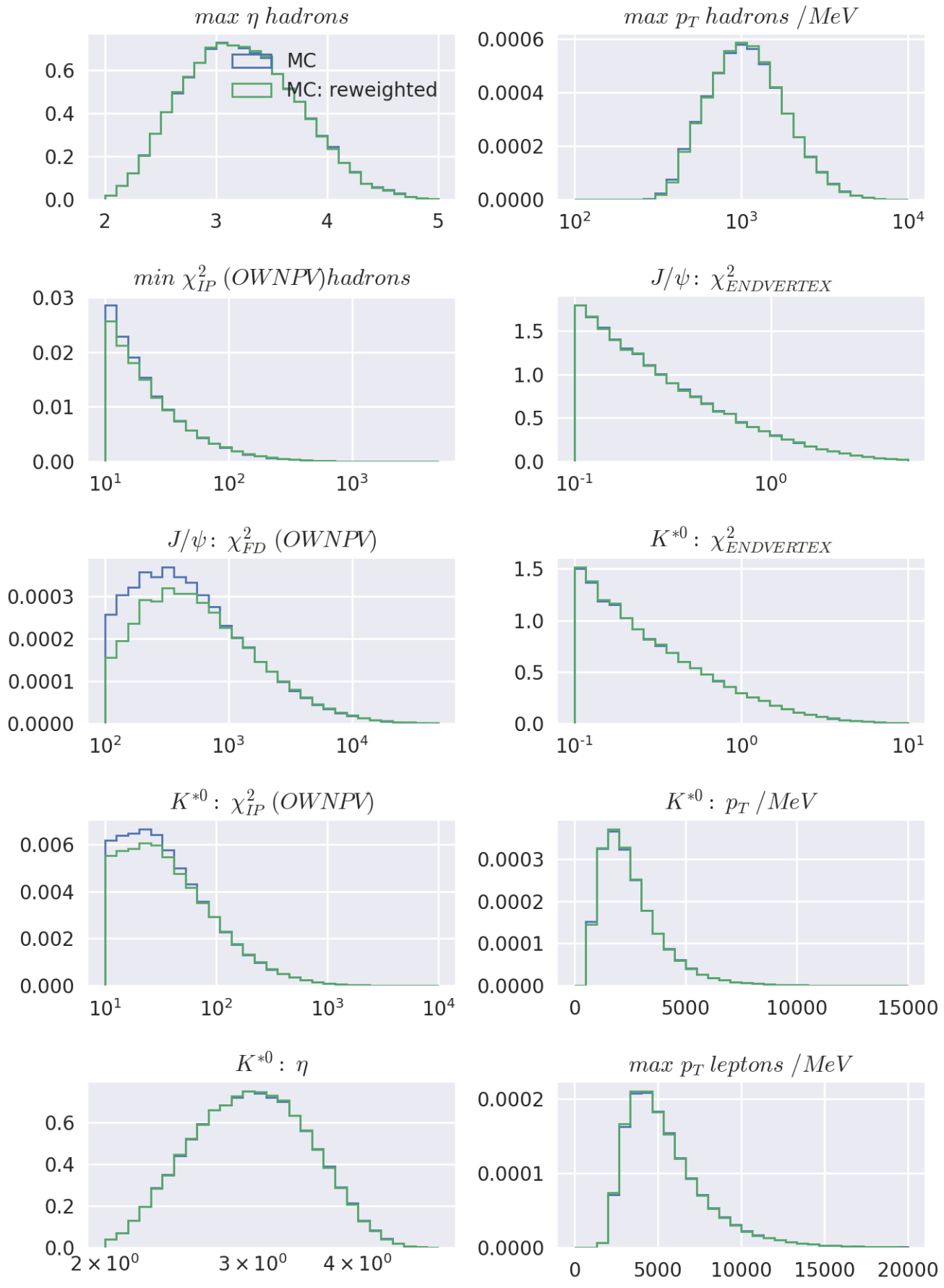

Figure 5.6: Marginal density distribution of the $B^0 \to K^{*0}e^+e^-$ simulation with and without weights (2/2). The legend in the first subplot applies to all.

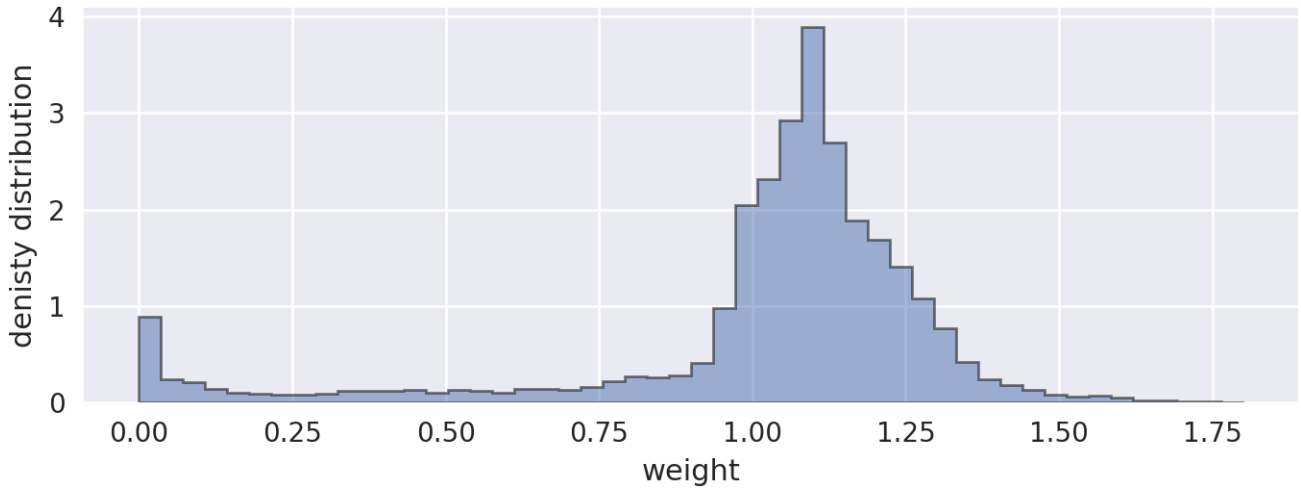

Figure 5.7: Density distribution of the weights assigned by the GBReweighter to the signal MC sample.

Chapter 6

Selection of $B^0 \to K^{*0} e^+ e^-$ decays at high q^2 values

To select the signal decay effectively from combinatorial and double semileptonic background, a multivariate analysis is performed. It is expected that a higher signal yield with a comparable background can be achieved with an MVA when compared to other methods.

6.1 Theory and intention

Similar analysis [4] on the same decay in different q^2 regions also use a MVA to select signal decays from background. The training sample for the background used in these studies is the upper B^0 mass $m(K^{*0}e^+e^-)$ sideband. This sample consists of combinatorial background and is a good approximation of the combinatorial background in the signal region.
However this approach is not feasible in the high q^2 region, since the upper B^0 invariant mass sideband contains insufficient amounts of background candidates. Instead it is proposed to use the data sample $B^0 \to K^{*0} e^\pm \mu^\mp$ in the same q^2 and the nominal B^0 mass regions, which includes combinatorial background as well as double semileptonic backgrounds.
The corrected simulated events (see Chapter 5) are used as a signal sample for the training of a classifier.

6.1.1 Choice of classification algorithm

A whole host of supervised machine learning algorithms is available for use in binary classification tasks. A selection of models has been tried and optimised, such as AdaBoost, multilayer perceptron artificial neural network and random forest models. The training and testing can be roughly described in the following steps. Each algorithm comes with its own set of hyper-parameters, over which a grid of potential values is spanned. This grid is then checked for the best values by training and testing over stratified k-folds[1], introducing some averaging as well.

[1] A stratified k-fold creates N subsets, where each set is of the same size, with the same ratio of signal to background. The training set is the complement to this subset, where the subset is then used for testing. This method allows for N unique test sets to avoid any bias.

The ROC-AUC is used as performance metric. Based on these results an in-depth training using these settings is performed over five folds. Further diagnostics are performed such as overfitting checks and selection efficiency checks over the interesting observables (B mass, q^2 and the observables used in the angular analysis [4]). Based on these checks, hyperparameters are adjusted manually, if needed.
XGBoost, as introduced in Section 2.3, performs as well as or better than all other tested algorithms. Additionally it offers some additional advantages, such as not being affected by the scale of the input, its robustness against highly correlated inputs and its very good computational performance. The last point stands out especially against artificial neural networks, which perform very similarly in terms of signal selection and background rejection but are almost one order of magnitude slower, than the XGBoost implementation.

6.2 Classification

The evaluation of the performance of the trained XGBoost algorithm is done for the least restrictive q^2-cut that is reasonably above the $\psi(2S)$ resonance, as the exact q^2 cut may be subject to further optimisation. Therefore the cut, "q2_PVDTF[2]" > 14 GeV$^2/c^4$ is chosen. This q^2 variable is reconstructed offline with a constraint on the primary vertex. It also includes a correction for bremsstrahlung that is recovered by the detector. Table 6.1 states the hyperparameters that are chosen, as well as a short description of their effect. The XGBoost implementation is trained using the training samples described in Section 6.1.

While the performance of the classifier is evaluated over five folds, a single classifier is trained over the complete data set and then saved, so it can be used later. Other options are to choose one of the classifiers associated to a fold at random or to implement a voting function. However, in machine learning competitions the approach to train a new classifier over the complete sample has shown to be marginally more effective. In Figure 6.1 the performance of this classifier is compared on the different folds, as well as to the classifier trained with the upper B^0 invariant mass sideband (XGBUnbiasedLabelScore) in the lower q^2 regions, on the ROC. First, it is important to note, that the different folds do not deviate significantly from each other. This means that the process is stable and therefore allows us to train a new classifier over the full sample and trust in its performance. Secondly the comparison to the XGBUnbiasedLabelScore might suggest, that no noteworthy improvement over the other classifier is achieved here. However, this can not be deduced from this metric, as we have to keep in mind that different observables are used in these MVAs. The observables used in training XGBUnbiasedLabelScore include observables that are sensitive to the muonic nature of the background sample. Thus it is likely that this metric overestimates the performance of XGBUnbiasedLabelScore. For a fairer comparison one might reweight the muonic observables before the calculation of the XGBUnbiasedLabelScore. However, since it is ultimately not

[2]The analysis at lower q^2 values bases cuts on the variable "q2_PVandBDTF", which includes a restraint on the invariant B mass. This is useful since the invariant B mass upper sideband is used in training, therefore the restraint effectively compensates for the additional mass in this combinatorial sample.

hyperparameter	value	explanation
`max-depth`	7	Maximum depth of a tree, deeper trees build more complex models, are more likely to overfit.
`n_estimators`	80	Number of trees that are built. Too many trees lead to overfitting. Strongly related to `eta`.
`eta`	0.1	Learning rate of the algorithm. Each new iteration adds to the predicted label. These additions are scaled according to this factor.
`colsample_bytree`	0.5	Each tree only uses half the available observables, this prevents overfitting.
`subsample`	0.5	Each tree only uses half of the available data, this prevents overfitting.
`reg_alpha`	1	Penalises complexity in the model, this prevents overfitting.
`gamma`	1	Regularisation parameter γ. Penalises complexity.
`random_state`	43	Random number seed. Needed for reproducibility.

Table 6.1: Hyperparameters used in XGBoost training. All parameters not named here are set to their default value or not needed for reproducibility.

important how the classifiers perform on the training data, but how they perform on the signal data sample, a preliminary mass fit can be performed to estimate the yield instead.

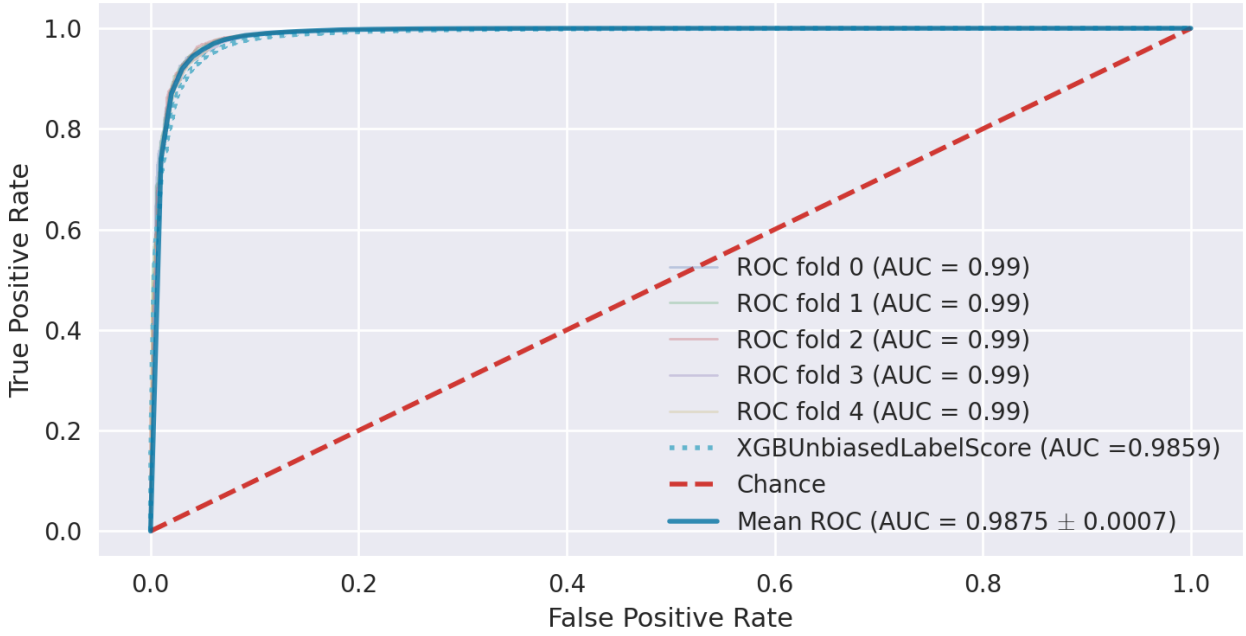

Figure 6.1: Receiver operating characteristic over all five folds, used for evaluation of the classification. Also included is the mean of all folds and as comparison the XGBUnbiasedLabelScore, trained for lower q^2 regions.

Chapter 7

Signal yield

In this section, a preliminary fit to the $K^{*0}e^+e^-$ invariant mass is performed. For the actual analysis a multidimensional fit of the mass, as well as the angular variables will be performed. Before this can be done, the backgrounds have to be studied in further detail, as such the following fit only aims to show the potential gains to be achieved by the selection performed in this thesis. For practical purposes a smaller q^2-region is chosen for the following calculations. This avoids some of the peaking and combinatorial background contributions that come with the larger a q^2-region. They are avoided at this point, because the consideration needed to estimate these backgrounds is outside of the scope of this thesis.

7.1 Threshold selection

The XGBoost classifier predicts a probability-like score for a given event to be a signal event. This allows for fine-tuning of the threshold an event has to meet in order to count as signal event, i.e. pass the selection. For an analysis like this, a simple figure of merit is introduced, which is then maximised as a function of the threshold. The figure of merit

$$\frac{S}{\sqrt{S+B}} \tag{7.1}$$

is used, where S is the expected signal yield and B is the expected background yield.
These two quantities are determined as a function of the threshold, t.
S, the **expected signal yield**, is accessible via the reference decay $B^0 \to K^{*0}J/\psi(\to e^+e^-)$, which was already used earlier.

$$S(K^{*0}e^+e^-)(t) = S(K^{*0}J/\psi(\to e^+e^-)) \times \frac{\epsilon^{sel}_{MC}(K^{*0}e^+e^-)(t)}{\epsilon^{sel}_{MC}(K^{*0}J/\psi(\to e^+e^-))} \times \frac{\mathcal{B}(K^{*0}e^+e^-)}{\mathcal{B}(K^{*0}J/\psi(\to e^+e^-))} \tag{7.2}$$

Here the factors ϵ^{sel}_{MC} are the selection efficiencies determined using the respective Monte Carlo simulations. This includes in the case of $\epsilon^{sel}_{MC}(K^{*0}e^+e^-)(t)$ the selection efficiency of the classifier. $\mathcal{B}$ are the branching fractions of the decays and $S(K^{*0}J/\psi(\to e^+e^-))$ is the signal yield of the reference decay scaled to account for the full integrated luminosity of Run 2. In the

following these factors are explained in more depth.
The selection efficiencies are determined in several stages, the preselection cuts, the q^2 cut and a reconstructed invariant mass $m(K^{*0}e^+e^-)$ cut are applied. The mass cut is applied only to the signal decay, more specifically a cut on the 0.9-quantile of the mass distribution[1], this is also done for the background. This mass cuts is chosen, because the figure of merit is only of interest when evaluated over the signal window, not in the complete $m(K^{*0}e^+e^-)$ window. The efficiencies of the cuts are listed in Table 7.1. The selection efficiency of the MVA as a function of the threshold is also included in this factor. However, it is easier to deal with these in a separate step.
The branching fraction of the channel $B^0 \to K^{*0}J/\psi(\to e^+e^-)$ is the product of the branching fractions of the sub decays $B^0 \to K^{*0}J/\psi$ and $J/\psi \to e^+e^-$ [16].

$$\mathcal{B}(K^{*0}J/\psi(\to e^+e^-)) = (7.58 \pm 0.30) \times 10^{-5} \tag{7.3}$$

It is more complicated for the signal decay, as the measured branching fraction quoted in the PDG does not include the full q^2 region, but the simulation used to determine the efficiency does. However, a value for the branching fraction for $q^2 > 13\ \mathrm{GeV}^2/c^4$ can be calculated using the Flavio phenomenology software [17]. For this reason the selection efficiency for the signal decay was normalised to the same Monte Carlo true q^2 region.

$$\mathcal{B}(K^{*0}e^+e^-)_{q^2>13\ \mathrm{GeV}^2/c^4} = (4.0 \pm 0.4) \times 10^{-7} \tag{7.4}$$

The signal yield of the reference channel $S(K^{*0}J/\psi(\to e^+e^-))$ is obtained via the same mass fit (see Fig.5.1), which was used to obtain the data for the reweighting in Chapter 5. While this mass fit includes corrections for combinatorial background, no studies or corrections on additional sources of background were included. However, this source of uncertainty is expected to be negligible, compared to the uncertainties of the branching fractions.

$$S(K^{*0}J/\psi(\to e^+e^-)) = (85100 \pm 400)\ \text{decays} \tag{7.5}$$

A preliminary estimate of the expected signal yield prior to the MVA application can be calculated as:

$$S(t=0) = (308 \pm 33)\ \text{decays} \tag{7.6}$$

B, the **expected background yield**, is calculated more easily, by accessing the data and performing all the same cuts as explained before, including the tight mass cut. Although, prior to the MVA cut this sample is expected to be largely dominated by background, the expected $B^0 \to K^{*0}e^+e^-$ signal contribution $S(t=0)$ is subtracted. This method is a good estimate for the full background in the data sample, but as can be seen in the mass fit later, the background also has peaking contributions, to which the MVA will respond differently. However, the peaking backgrounds can be neglected here as their yield is expected to be smaller than the signal yield,

[1] Mass values of the cut are found in Table 7.1

cut on $B^0 \to K^{*0}e^+e^-$	efficiency
Preselection	0.06632 ± 0.00022
$q^2 > 16$ GeV$^2/c^4$	0.2243 ± 0.017
5093 MeV/$c < m(K^{*0}e^+e^-) <$ 5411 MeV/c	0.900 ± 0.009
total	0.01339 ± 0.0010

cut on $B^0 \to K^{*0}J/\psi(\to e^+e^-)$	efficiency
Preselection	0.05957 ± 0.0008
8 GeV$^2c^{-4} < q^2 <$ 11 GeV$^2/c^4$	0.9683 ± 0.0008
total	0.05768 ± 0.00008

Table 7.1: Selection efficiencies ϵ^{sel}_{MC} of the signal and reference decay.

which in turn is much smaller than the background yield.

$$B = (2340 \pm 60) \text{ decays} \tag{7.7}$$

The calculation of the **MVA selection efficiency** is done by evaluating the training sets[2]. The MVA selection efficiency on the background sample is obtained from the $B^0 \to K^{*0}\mu^\pm e^\mp$ sample, which is expected to have no contributions from signal or peaking backgrounds. In an additional step, which is supposed to make the process more intuitive, the MVA probability predictions of the signal MC are monotonically transformed to a uniform distribution over the interval $[0, 1]$. The same transformation is applied to the background response. As such the signal probabilities are distributed uniformly and the background probabilities are mostly near zero, with a tail distributed over the full space. The threshold dependent component of the selection efficiency is calculated numerically as as function of the transformed probabilities. The figure of merit, as described by eq. 7.1, is calculated and drawn as a function of the threshold t. The optimal threshold leading to the numerical maximum is calculated as $t_{opt} = 0.13$. In addition, the expected signal and background yield, again for the 0.9-quantile $m(K^{*0}e^+e^-)$-region, are calculated:

$$\text{expected signal yield: } S^{opt}_{predicted} = (268 \pm 29) \text{ decays} \tag{7.8}$$

$$\text{expected background yield: } B^{opt}_{predicted} = (38 \pm 5) \text{ decays} \tag{7.9}$$

The figure of merit and its numerical maximum is depicted in Figure 7.1.

[2]Due to the k-folding technique, the training sets are assigned a score without training bias.

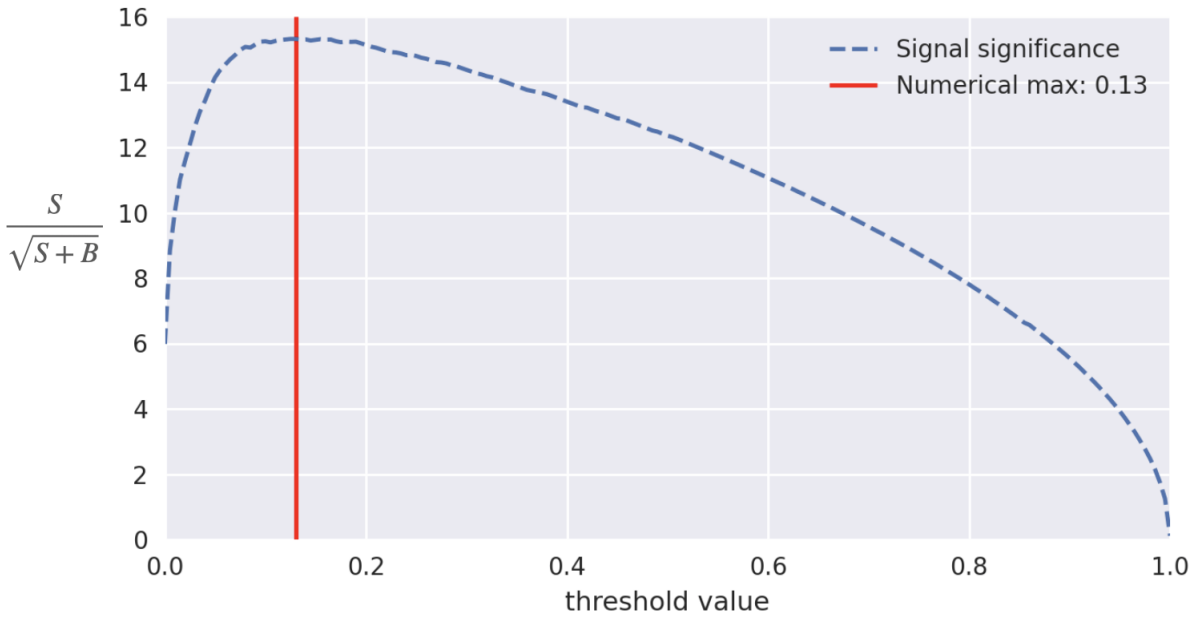

Figure 7.1: Figure of merit for the threshold selection, the numerical maximum is indicated.

7.2 Mass fit of the $B^0 \to K^{*0}e^+e^-$ distribution

In the mass fit the different components of the data, which pass all selections, are fitted against the signal. An approach is chosen, where a model is created for each components that is then fitted to the respective sample. The following different components are fitted:

- $B^0 \to K^{*0}e^+e^-$
- $B^0 \to K^{*0}\psi(2S)(\to e^+e^-)$
- $B^+ \to K^+\pi^+\pi^-e^+e^-$
- $B^+ \to K^+\pi^+\pi^-\psi(2S)(\to e^+e^-)$
- combinatorial background

The simulation of these decays, or for the combinatorial background the $B^0 \to K^{*0}\mu^\pm e^\mp$ data sample, passed the same stripping, preselection and MVA selection as the signal data. The Monte Carlo simulations are individually fitted to either single or double Crystal Ball functions, which consists of a Gaussian distribution with exponential tails. A Gaussian distribution is fitted to the combinatorial background sample, since this background is suppressed so strongly in the MVA selection that more complex models are unstable. The shape of these models is then fixed and their sum is fitted against the signal data, with the yields as remaining free parameters. The fitting is performed unbinned using extended models in zfit [18]. This last step is only done, after one is satisfied with the fits to the simulations, as adjustments based on the signal data fit will potentially lead to non-physical results. For evaluation purposes two different models will be compared here, the MVA selection presented in this thesis (Fig. 7.2)

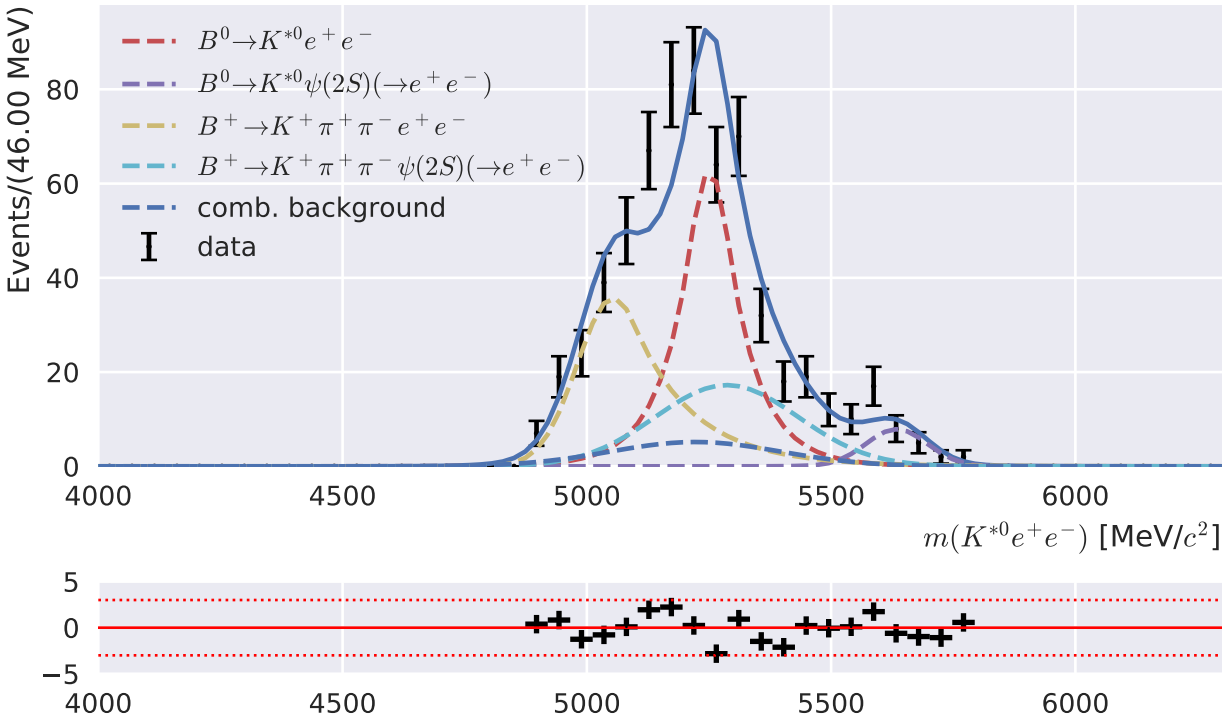

Figure 7.2: Mass fit to the recorded data selected by the MVA developed in this thesis at a threshold of $t = 0.13$. Data depicted with a bin width of 46 MeV.

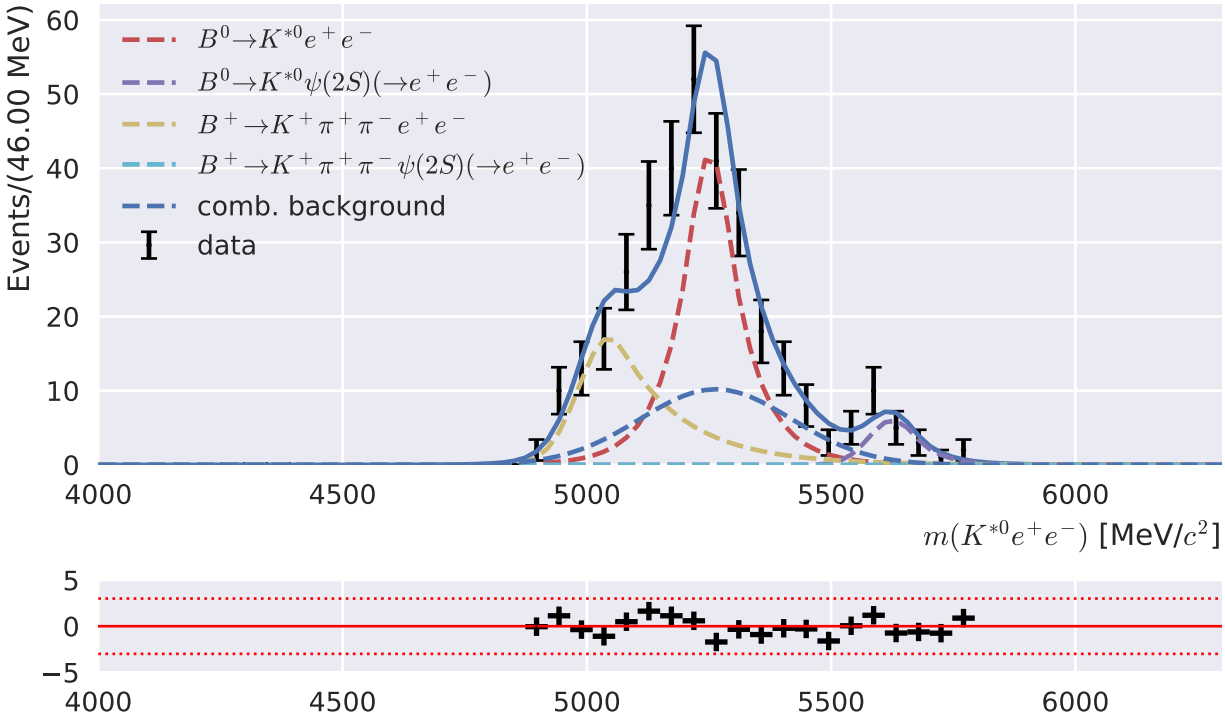

Figure 7.3: Mass fit to the recorded data selected by a MVA trained for lower q^2. Data depicted with a bin width of 46 MeV.

component	new MVA decays			low q^2 MVA decays		
$K^{*0}e^+e^-$	230	-40	+40	148	-27	+28
$K^{*0}\psi(2S)(\to e^+e^-)$	27	-6	+7	17	-5	+5
$K^+\pi^+\pi^-e^+e^-$	137	-50	+40	83	-17	17
$K^+\pi^+\pi^-\psi(2S)(\to e^+e^-)$	137	-50	+40	0	-0	+6
combinatorial background	50	-0	+70	84	-30	+30

Table 7.2: Comparison of the yields for the different signal components for both MVA selections. The first is the classifier trained in this thesis, and the second a MVA trained with the upper B mass sideband.

	Signal	$K^{*0}\psi(2S)$	$K^+\pi^+\pi^-e^+e^-$	$K^+\pi^+\pi^-\psi(2S)$	comb. bg
Signal	1	0.2	0.0	-0.8	0.0
$K^{*0}\psi(2S)$	0.2	1	0.0	-0.3	0.0
$K^+\pi^+\pi^-e^+e^-$	0.0	0.0	1	-0.3	-0.1
$K^+\pi^+\pi^-\psi(2S)$	-0.8	-0.3	-0.3	1	-0.1
comb. bg	0.0	0.0	-0.1	-0.1	1

Table 7.3: Correlations of the fitted yields, for the classifier trained in this thesis (Fig. 7.2), rounded to the first decimal place.

and a MVA selection designed for use at lower q^2 (Fig. 7.3), that was trained using the upper invariant B mass sideband. The resulting yields can be found in Table 7.2.

A large problem with this fit is an instability regarding the combinatorial background and the peaking background $K^+\pi^+\pi^-\psi(2S)(\to e^+e^-)$, which have very similar form and positions and therefore highly correlated yields. The yield correlations can be seen in Table 7.3. This shows that signal and $K^+\pi^+\pi^-\psi(2S)(\to e^+e^-)$ yields have a strong negative correlation. Further investigations and vetoes are needed to deal with the $K^+\pi^+\pi^-\psi(2S)(\to e^+e^-)$ contributions. A quick check of the threshold selection can still be done by calculating the signal significance. For this purpose the signal yield and the total background yield are integrated over the $m(K^{*0}e^+e^-)$ signal window.

$$\frac{S}{\sqrt{S+B}}(t = 0.13) = 10.1 \pm 1.8 \tag{7.10}$$

While this value is smaller than the value of 15.3 ± 1.0 (see Fig. 7.1) predicted earlier, it still represents a good signal significance. As a comparison, the same calculations have been completed for the previous MVA and give:

$$\frac{S}{\sqrt{S+B}}_{\text{low } q^2} = 8.7 \pm 1.7 \tag{7.11}$$

A more thorough treatment of this peaking background could again lead to improvements in the yield and its significance.

Chapter 8

Discussion

The aim of this thesis is to optimise the selection of $B^0 \to K^{*0}e^+e^-$ events for high values of the dilepton invariant mass squared q^2. The standard procedure of selection optimisation uses the events in the upper invariant B^0 mass sideband in recorded data as proxy for combinatorial background. However, this approach is not possible here. It turns out that there is not a sufficient amount of events in the $B^0 \to K^{*0}e^+e^-$ upper B^0 mass sideband at high values of q^2. Instead the recorded sample $B^0 \to K^{*0}\mu^\pm e^\mp$ in a large B^0 mass window is used. These types of decays are highly suppressed in the Standard Model, thus the reconstructed events must be combinatorial background events. Several changes are made to adapt the selection to this decay channel.
The properties of the signal are taken from simulated samples. To ensure that the simulation accurately describes the data, the agreement between simulation and data is evaluated exploiting multivariate analysis tools in the reference channel $B^0 \to K^{*0}J/\psi(\to e^+e^-)$. Weights are calculated using a boosting technique to correct potential differences. After evaluation on the reference channel, weights are calculated for the simulated signal sample $B^0 \to K^{*0}e^+e^-$.
Finally the selection itself is optimised on the simulated and reweighted $B^0 \to K^{*0}e^+e^-$ sample and the on the $B^0 \to K^{*0}\mu^\pm e^\mp$ sample, which serves as a background proxy. For the selection the state of the art algorithm XGBoost is chosen, which is one of the leading machine learning algorithms.
A preliminary fit of the B^0 mass distribution in data is performed, using the selection developed in this thesis and for comparison the previous MVA optimised for use at lower q^2-values. The signal significance is calculated to be $S/\sqrt{S+B} = 10.1 \pm 1.7$, which compares favourably to the older selection at $S/\sqrt{S+B}_{\text{previous}} = 8.7 \pm 1.7$.
Due to missing energy related to the loss of bremsstrahlung from the electrons in the final state, the signal mass distribution is rather broad. The next step is to understand and remove the peaking backgrounds in this large signal mass window.
This analysis focuses on Run 2 data. Improvements of similar size are expected with Run 1 as well. Furthermore improvements are expected by optimising the definition of the high q^2 region.

The work done in this thesis will be input for two analysis of the LHCb collaboration,

namely the test of lepton universality in the decays $B^0 \to K^{*0} e^+ e^-$ and $B^0 \to K^{*0} \mu^+ \mu^-$, and the angular analysis of the decays $B^0 \to K^{*0} e^+ e^-$. Both analysis currently show tensions of two to three σ with respect to the Standard Model. The improvement in selection presented here might have a decisive impact in the upcoming updates of these analysis.

Bibliography

[1] Martin Breidenbach et al. “Observed Behavior of Highly Inelastic electron-Proton Scattering”. In: *Phys. Rev. Lett.* 23 (1969), pp. 935–939. DOI: 10.1103/PhysRevLett.23.935.

[2] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[3] Mark Thomson. “Introduction”. In: *Modern Particle Physics*. Cambridge University Press, 2013, pp. 1–29. DOI: 10.1017/CBO9781139525367.003.

[4] R. Aaij et al. “Angular analysis of the B0 -> K*0e+e- decay in the low-q2 region”. In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)064. URL: http://dx.doi.org/10.1007/JHEP04(2015)064.

[5] Simone Bifani et al. “Review of lepton universality tests in B decays”. In: *Journal of Physics G: Nuclear and Particle Physics* 46.2 (Dec. 2018), p. 023001. ISSN: 1361-6471. DOI: 10.1088/1361-6471/aaf5de. URL: http://dx.doi.org/10.1088/1361-6471/aaf5de.

[6] Trevor Hastie, Robert Tibhirani, and Jerome Friedman. “Boosting and Additive Trees”. In: *The Elements of Statistical Learning*. Springer - Verlag, 2009, pp. 337–384. DOI: 10.1017/CBO9781139525367.003.

[7] Tianqi Chen and Carlos Guestrin. “XGBoost”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016). DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

[8] John L. Hodges. “The significance probability of the smirnov two-sample test”. In: *Arkiv för Matematik* 3 (1958), pp. 469–486.

[9] A Augusto et al. Alves. “The LHCb Detector at the LHC”. In: *JINST* 3.LHCb-DP-2008-001. CERN-LHCb-DP-2008-001 (2008). Also published by CERN Geneva in 2010, S08005. DOI: 10.1088/1748-0221/3/08/S08005. URL: https://cds.cern.ch/record/1129809.

[10] Lyndon R Evans and Philip Bryant. “LHC Machine”. In: *JINST* 3 (2008). This report is an abridged version of the LHC Design Report (CERN-2004-003), S08001. 164 p. DOI: 10.1088/1748-0221/3/08/S08001.

[11] Roel Aaij et al. “Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC.” In: *JINST* 14.arXiv:1812.10790. 04 (Dec. 2018), P04013. 43 p. DOI: 10.1088/1748-0221/14/04/P04013. URL: https://cds.cern.ch/record/2652801.

[12] Patrick Koppenburg et al. *StrippingBu2LLK_eeLine*. 2018. URL: http://lhcbdoc.web.cern.ch/lhcbdoc/stripping/config/stripping28r2/leptonic/strippingbu2llk_eeline.html.

[13] Thomas Blake. *Artistic sketch of the differntial decay rate of B0 ->Kstar0 l+l- as function of q2*. 2014.

[14] Alex Rogozhnikov. “Reweighting with Boosted Decision Trees”. In: *Journal of Physics: Conference Series* 762 (Oct. 2016), p. 012036. ISSN: 1742-6596. DOI: 10.1088/1742-6596/762/1/012036. URL: http://dx.doi.org/10.1088/1742-6596/762/1/012036.

[15] M. Pivk and F.R. Le Diberder. “: 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 (Dec. 2005), pp. 356–369. ISSN: 0168-9002. DOI: 10.1016/j.nima.2005.08.106. URL: http://dx.doi.org/10.1016/j.nima.2005.08.106.

[16] M. Tanabashi et al. “Review of Particle Physics”. In: *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[17] David M. Straub. *flavio: a Python package for flavour and precision phenomenology in the Standard Model and beyond*. 2018. arXiv: 1810.08132 [hep-ph].

[18] Jonas Eschle et al. “zfit: Scalable pythonic fitting”. In: *SoftwareX* 11 (Jan. 2020), p. 100508. ISSN: 2352-7110. DOI: 10.1016/j.softx.2020.100508. URL: http://dx.doi.org/10.1016/j.softx.2020.100508.

Acknowledgements

I would like to express my gratitude to Prof. Dr. Hansmann-Menzemer for giving me the opportunity to study and work in her group. I thoroughly enjoyed getting an insight into modern particle physics and the warm and supportive group she leads.
I would like to thank Dr. Martino Borsato for giving me an exciting and relevant problem to work on, for his support at every corner and for giving me plenty of free space while working on the more challenging parts.
Further I would like to thank Jiangquiao Hu for helping me out with a plethora of challenging programming problems and for sharing his insights and code snippets with me.
I would like to thank the staff and management of the Physikalisches Institut for prioritising the health and safety of all employees and students during the COVID-19 public health crisis.
My gratitude goes to the CERN- and LHCb-collaborations, which provide not only unrivalled accelerators and detectors, but also great learning opportunities and support to students.
I would like to thank Benedikt Schosser and Maja Rüth for proof reading this thesis and for their valuable input.
Lastly I would like to thank the Studienstiftung des deutschen Volkes for their continued support of my studies.

Erklärung

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 20.07.2020,