Masterhesis: Development of improved method for fitting differential decay rates, Classification of data and Reweighting

Supervisors: Prof. Dr. Nicola Serra, Dr. Marcin Chzsaszcz

Author: Oliver Dahme

Here comes the abstract

Contents

1 Introduction

The work presented in this thesis consits of three different parts:
First a new fitting routine has been implemnted, as an more transperent alternative to the broadly used routine Minuit [1].
Second a test of classifiers has been performed. Classifiers are a type of algorithm used to seperate signal from background. The test was performed to find the best classifier, for the analysis of the $B_0 \to K^* \mu^+ \mu^-$ decay.
Third data from the Run II of the LHCb detector was used to reweight the Monte Carlo simulation of the $B_0 \to K^* \mu^+ \mu^-$ decay. The reasons for choosing this decay are explained in section 4.1

1.1 New fitting routine

In particle physics the angular coefficients of a decay rate are obtained, by fitting (expained in section **??**) the decay rate distribution the experimentally obtained kinematic variabels like angles between final state particles and their energies.
This has been done with the broadly used algorithm MINUIT [1]. But like explained in section 2 MINUIT can make random fatal errors while fitting, without any obvious explanation. the second problem with MINUIT is the intransperency: One can not comprehend how the fit is performed or if it really has found the global minimum.
The new algorithm implemented as part of this thesis, is called RooMCMarkovChain. It is based on a Monte Carlo Markov Chain, where one link in the chain respresents a point on the log-likelihood function of the fit. The chain is designed to follow the global minimum and to jump out of local minima. This provides two things: A fitting parameter set, for example the angular coefficents of a decay rate, at the global minimum. And a scan of the log-likelihood function in the vicinity of the global minimum, which can be used for the error estimation and can be send along side papers containing fits to make it transperent for the reader how the fit performed.

1.2 Test of classifiers

Raw data from a particle detector contains all kind possible decays. The first challenge of an anlysis is to distinguish the one decay one wants to analyse from all the others. The common prcedure to reach that goal in our days the usage of Machine Learning algorithms called Classifiers. The need two things befor performing the analysis: A training dataset where signal and background is labeld accordingly. And a list of parameters that could be different in signal and background. While training the algorithm

will adapt thresholds on these parameters. With these thresholds the algorithm is able to distinguish signal from background in unlabeled data.
The test of classifiers presented in section 5 has the goal to find the best classifier to distinguish $B_0 \to K^* \mu^+ \mu^-$ decays in the raw data of the LHCb detector.

1.3 Reweighting

The infomation needed in an analysis is not all contained inside the data. For example efficiencies can not be recorded by a detector. But they can be simulated by a Monte Carlo simulation of the particle decay. The problem is that the simulation has to be the same as data to get the correct answers. Since resimulating everything with different parameters until one hits the data distribution would take too much resources. One can simply reweight the simulation to match the data. As an initial weight the sWeights see equation 37 are used. The parameters used to reweight the MC samples are applied in the following order: 'nTracks', 'B_0 p_T' and the quality of the $K\pi\mu\mu$ vertex. The new weights derived by the difference in data and simulation of these parameters are used to weight all the MC samples. For this analysis the MC and Data of the $K^* \to J/\Psi K^{*0}$ are used, because it is a very clean channel.

2 RooMCMarkovChain

RooMCMarkovChain is a new class for the ROOT Data Analysis Framework, which can be used to for fits and presents a alternative to the brodly used MINUIT algorithm. RooMCMarkovChain uses a metropolis algorithm as a minimizer for the negative log likelihood function. The metropolis algorithm is based on a Monte Carlo Markov Chain and can therfore easily be scaled to multidimensional parameter space and moreover, such kind of algorithms can easily be parallelized.

2.1 Monte Carlo Markov Chain (MCMC)

A Markov Chain is a random process which undergoes several states. From each state there is a probability ditribution to change into another state or to stay. Most important i the asumption that every next step just depends on the current state. The figure 1 illustrates the behavior of such a Markov Chain.

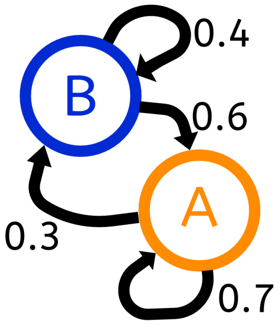


Figure 1: Illustration of states and the probability distribution to change or the stay in a state of a Monte Carlo Markov Chain. In state A there is a 0.3 probability to change into state B and a 0.7 probability to stay in state A.

2.2 Metropolis Hastings

Metropolis Hastings is a so called Monte Carlo updater. It updates the probabillity distribution when a change of states is proposed. That is very usefull in the cases where the states are not discrete like in figure 1 but continous. Suppose that a specified distribution has unnormalized density h. The Metopolis Hastings update does the following:

1. The current state x is proposed to move to another state y with a conditional probability given x denoted as $q(x, \cdot)$.
2. The Hastings ratio is calculated:
$$r(x,y) = \frac{h(y) \cdot q(y,x)}{h(x) \cdot q(x,y)} \tag{1}$$
3. The proposed move to y is accepted with the probability $a(x,y)$:
$$a(x,y) = min(1, r(x,y)) \tag{2}$$

This principle is used in the RooMCMarkovChain class to minimize the negative log likelihood function. In detail a robust adaptive Metrpolis hastings process is implemented. It is defined as:

1. Compute next proposed state Y_n as a random change from the current state X_{n-1}:
$$Y_n \equiv X_{n-1} + S_{n-1}U_n$$
, where U_n is an independent random vector.
2. With the probability α_n the proposal is accepted and $X_n = Y_n$. Otherwise the proposal is rejected and $X_n = X_{n-1}$
$$\alpha_n \equiv min\{1, r(X_{n-1}, Y_n)\}$$
, where r is the hastings ratio see equation 1.
3. compute the lower-diagonal matrix S_n with positive diagonal elements satisfying the equation:
$$S_n S_n^T = S_{n-1}\left(I + \eta_n(\alpha_n - \alpha_*)\frac{U_n Y_n^T}{||U_n||^2}\right) S_{n-1}^T \tag{3}$$
,where $I \in R^{dxd}$ is the identity matrix.

2.3 Features

The features of the RooMCMarkovChain class will be shown by fitting the following probability density function (pdf):

$$g(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + f \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \tag{4}$$

It is so called double gaus pdf, which is just the sum of two gaussian pdfs with a fractional parameter f. The result of the fit is compared with the result of the Minuit algorithm:

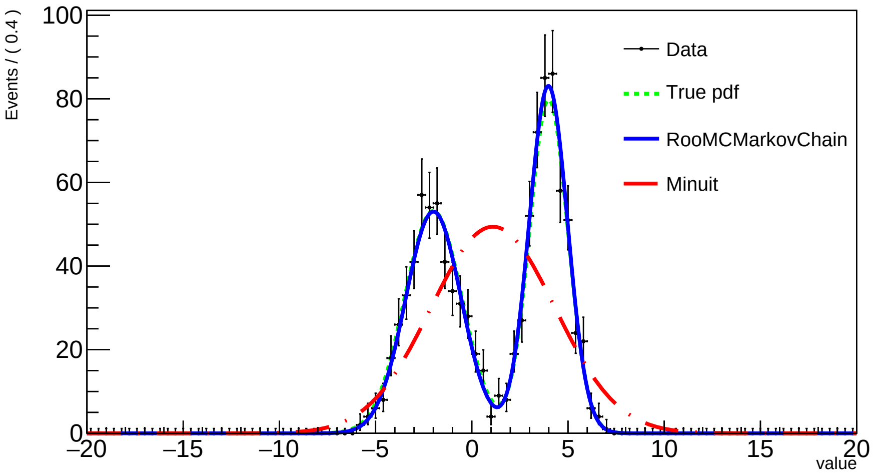
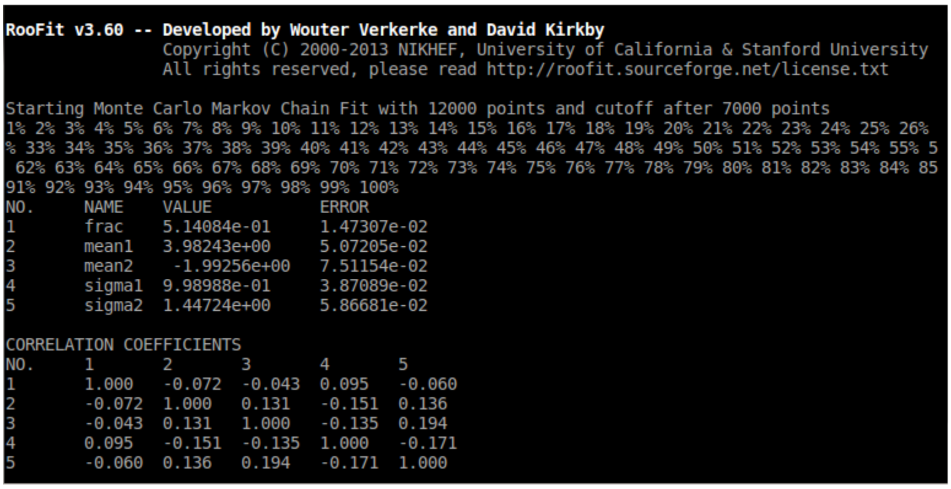


Figure 2: Fit of double gaus pdf (see equation 4) with RooMCMCarkovchain and with Minuit. The true pdf has the follwing values: $\mu_1 = 4$, $\sigma_1 = 1$, $\mu_2 = -2$, $\sigma_2 = 1.5$ and the fraction $f = 0.5$

The corresponding terminal output gives the estimated parameters with error interval and correlation coefficents of each parameter pair.



```
RooFit v3.60 -- Developed by Wouter Verkerke and David Kirkby
                Copyright (C) 2000-2013 NIKHEF, University of California & Stanford University
                All rights reserved, please read http://roofit.sourceforge.net/license.txt

Starting Monte Carlo Markov Chain Fit with 12000 points and cutoff after 7000 points
1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12% 13% 14% 15% 16% 17% 18% 19% 20% 21% 22% 23% 24% 25% 26%
% 33% 34% 35% 36% 37% 38% 39% 40% 41% 42% 43% 44% 45% 46% 47% 48% 49% 50% 51% 52% 53% 54% 55% 5
 62% 63% 64% 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77% 78% 79% 80% 81% 82% 83% 84% 85
91% 92% 93% 94% 95% 96% 97% 98% 99% 100%
NO.     NAME    VALUE           ERROR
1       frac    5.14084e-01     1.47307e-02
2       mean1   3.98243e+00     5.07205e-02
3       mean2    -1.99256e+00   7.51154e-02
4       sigma1  9.98988e-01     3.87089e-02
5       sigma2  1.44724e+00     5.86681e-02

CORRELATION COEFFICIENTS
NO.     1       2       3       4       5
1       1.000   -0.072  -0.043  0.095   -0.060
2       -0.072  1.000   0.131   -0.151  0.136
3       -0.043  0.131   1.000   -0.135  0.194
4       0.095   -0.151  -0.135  1.000   -0.171
5       -0.060  0.136   0.194   -0.171  1.000
```

Figure 3: Terminal output of the RooMCMarkovChain fit.

This gives a good initial overview of results after the fit has finished. The error calculation can be set to assume gaussian or non-gaussian errors.

In addition several other properties of the parameters can be obtainded:

1. The 1 dimensional profile of the negative log liklihood function for a given parameter.

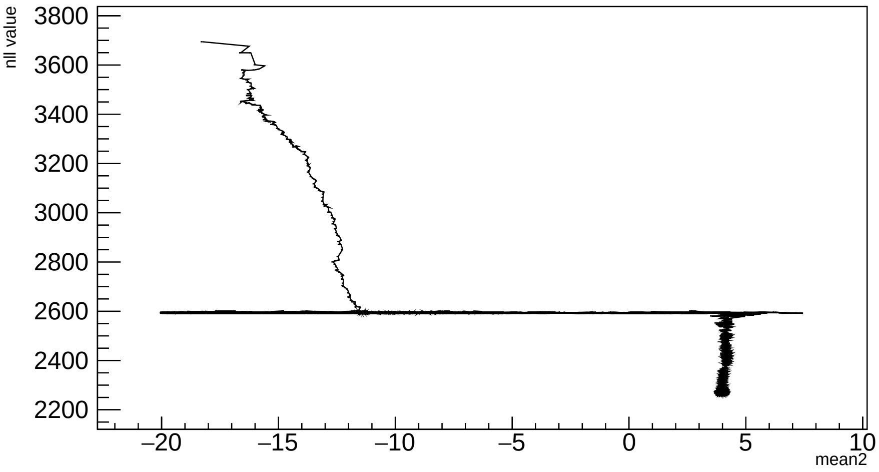


Figure 4: Profile of the negative log lilihood function for μ_2 in equation 4. There is a local minimum at the nll value of 2600.

From this plot on can see how the algorithm reached the minimum of the negative log likelihood function and if there are local minima.

2. The walk distribution of a given parameter.

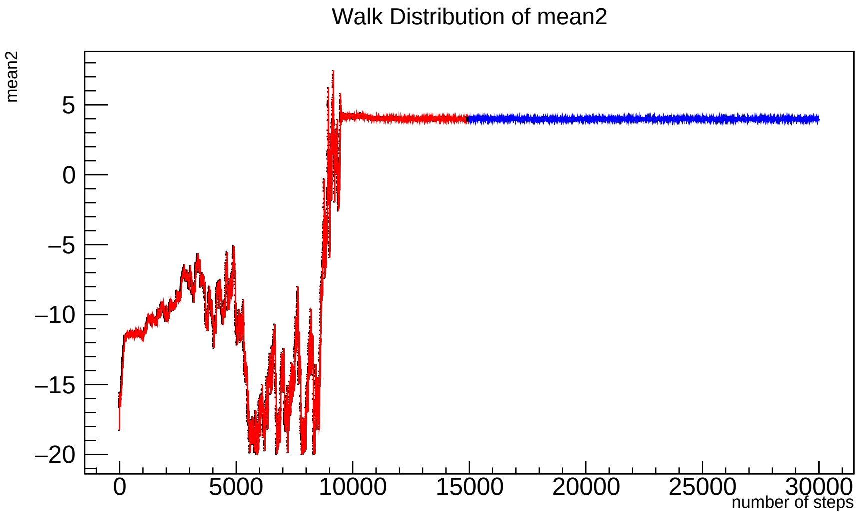


Figure 5: Walk distribution of μ_2 in equation 4. The red points are not considered for the error calculation.

This plot is very important for handling the RooMCMarkovChain class. Since the error calculation is based on the variance of the walk distribution, cutting of points in the begginning greatly reduces this variance. The user has to choose how many points are cut off. Plotting the walk distribution of all parameters helps choosing the right amount.

3. The walk distribution of a given parameter as a histogramm.

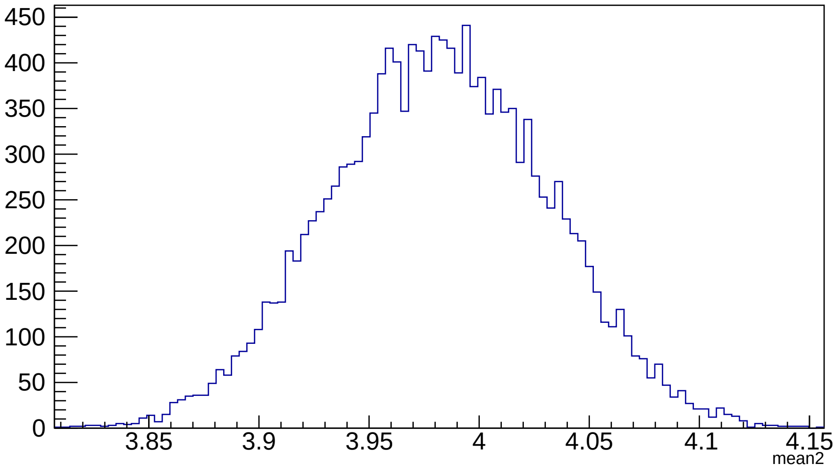


Figure 6: Walk distribution of μ_2 in equation 4 as a histogramm.

This plot can be used to check, which error strategy should be used. If the walk distribution histogramm is gaussian, gaussian errors can be assumed.

4. The scatterplot between two given parameters.

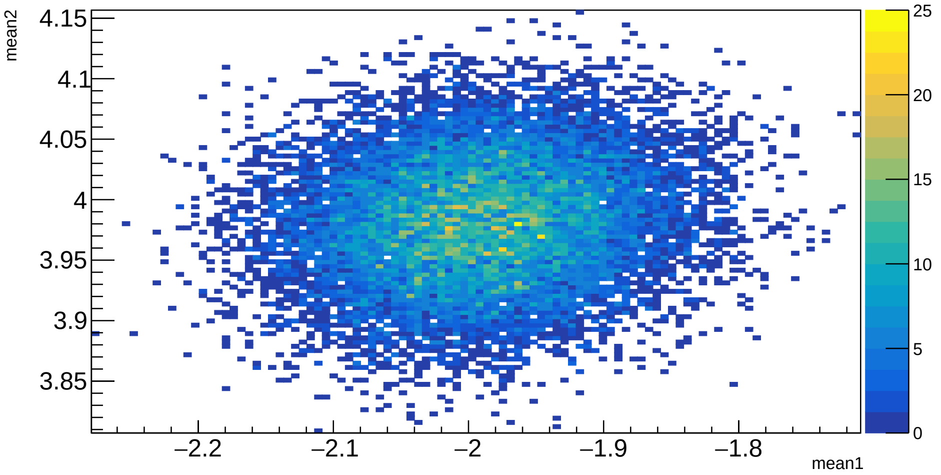


Figure 7: Scatterplot between μ_1 and μ_2 in equation 4.

This plot shows the correlation between two parameters graphically.

3 The LHCb Experiment

In this section the experimental setup of the detector is presented, which recorded the data used in this thesis. The Large Hadron Collider beauty experiment (LHCb) is one of four large experiments based at the CERN laboratory near Geneva in Switzerland. It is part of the Large Hadron Collider (LHC), a proton-proton accelerator and collider located in a vast unterground tunnel with 26.7 km circumference beneath the Swiss-French countryside. The other three experiments are CMS and ATLAS which are dedicated to a wide range of physics and have therefor very large detectors. ALICE investigates quark-gluon plasma and therefor needs geavy ion collisions, instead of proton collisions.
The protons in the LHC have a kinetic energy of 7 TeV, which allows a collision energy, in the LHCb detector, of 13 TeV. In the year 2016 the LHCb had a recorded luminosity of 1906 pb^{-1}. For this thesis 2280 pb^{-1} of data, collected at LHCb during the years 2011 to 2016 are used. LHCb is dedicated to falvour physics. It therefor investigates rare decays and CP violation in beauty and charm hadrons.

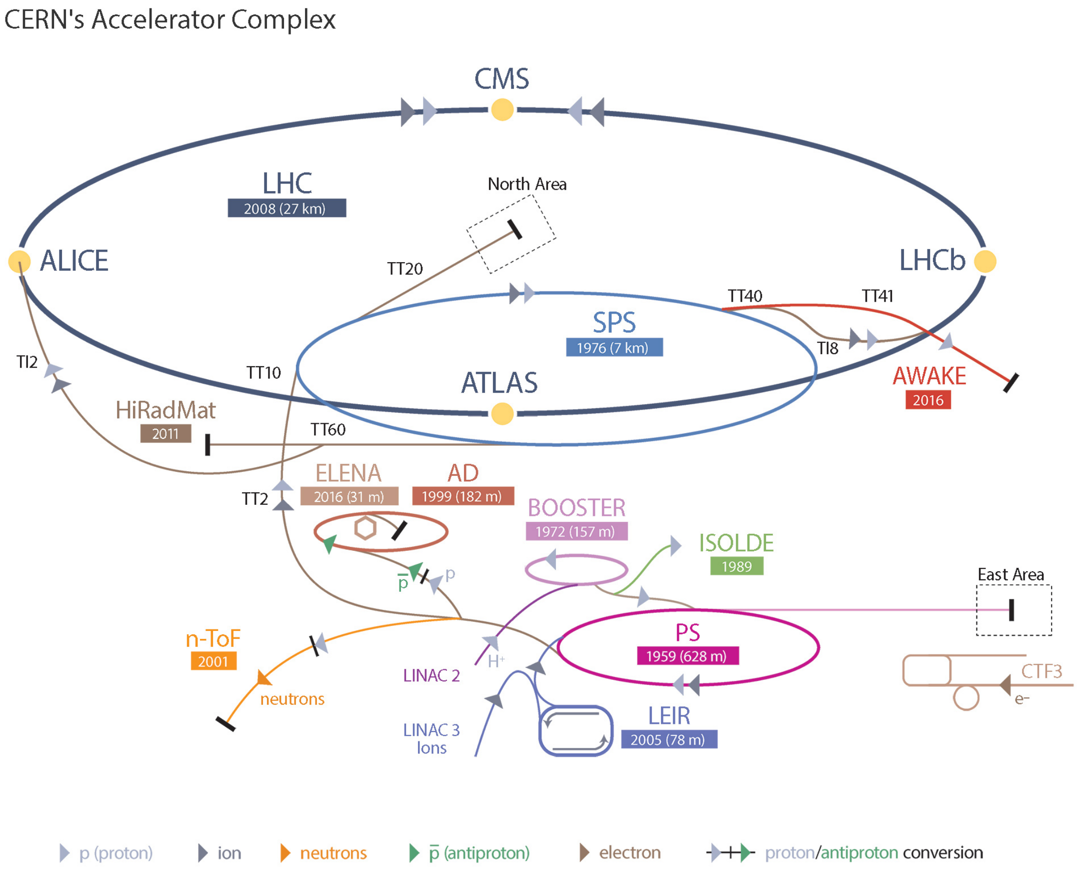


Figure 8: CERN's Accelerator Complex.[6] The protons get injected in the lineare accelerator LINAC2. Then they get pre-accelerated in 3 synchrotons (BOOSTER,PS,SPS) where the protons reach a kinetic energy of 450 GeV. That is the entering energy of the LHC which accelerates them futher up to 7 TeV, before they collide at the four detectors: CMS, ATLAS, LHCb and ALICE.

3.1 The LHCb Detector

The LHCb Detector has a fix target geometry, because beauty hadrons are manily produced at small angeles with respect to the beam pipe.

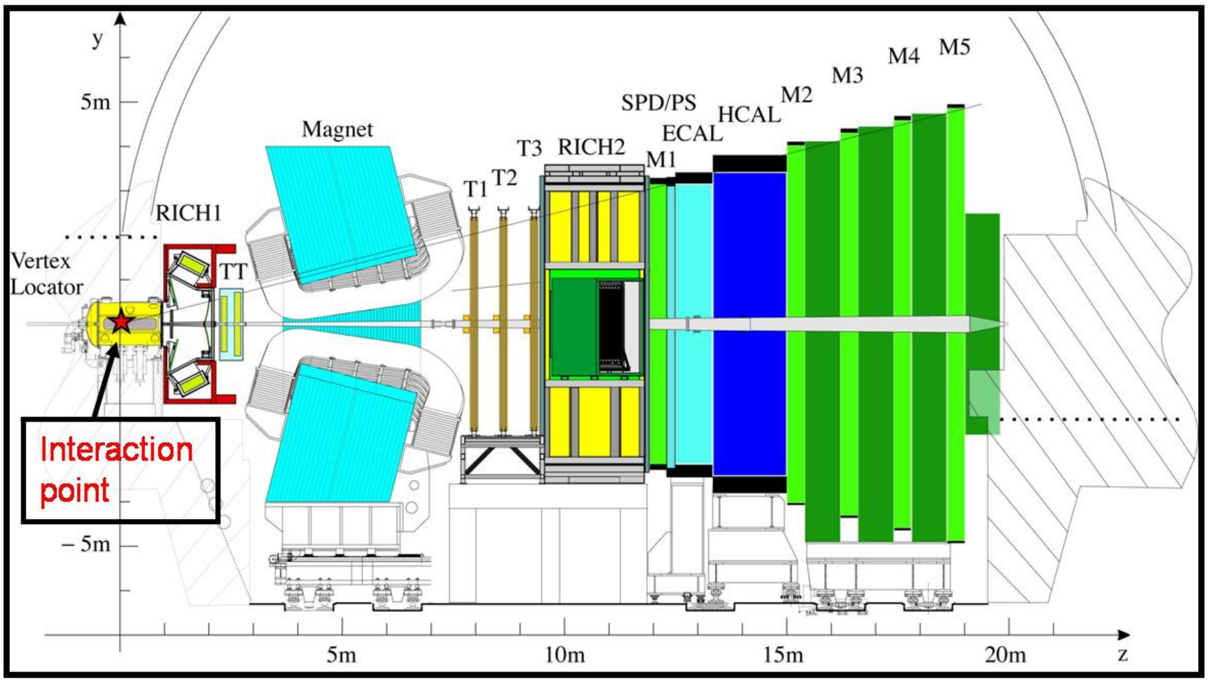


Figure 9: Basic layout of the LHCb detector [7]. The interaction point is inside the vertrex detector and the beam pipe passes through the center. The different subdetecors are the two Ring Imaging Cherenkov Detecors (RICH1 and RICH2), the tracking stations (TT and T1 to T3), the scintillator pad detector (SPD), the preshower electromagnetic calorimeter (ECAL), the hadronic calorimeter (HCAL) and the muon stations (M1-M5).

VErtex LOcator (VELO) [8] : Velo picks out B mesons from the multitude of other particles produced. This is a complex task since B mesons have very short livetimes spent colse to the beam. The VELO's silicon detecor elements must be placed at a distance of just five milimetres to the interaction point. To prevent damage to the detector during beam injection and stabilization it is mechanically moved to a safe distance. Velo measures B mesons indirectly be detecting its decay particles, nevertheless it has a resolution of 10 microns. .

Ring Imaging Cherenkov (RICH) detectors [9] : The RICH detectors meeasure the emission of Cherenkov radiation, which happens when a charged particle passes through a medium faster than light does. It is a similar effect like the sonic boom a aircraft produces by breaking the sound barrier. The shape of the light cone depends on the particle's velocity, enabling the detector to determine its speed.

Magnet [10] : The big magnet of the LHCb experiment weights 27 tonnes ands is mounted inside a 1,450 tonne steel frame. This powerful magnet forces all charged particels to change there trajectory. By examining the curvature of the path one can calculate its momentum.

Trackers [11] : The LHCb's tracking system consists of a series of four large rectangular stations, each covering an area of 40 m^2. While flying through this area charged particles will leave a trace, therefor one can estimate the trajectory of a particle. The trajectory is used to link the signals left in other detecor elements to the corresponding particle. In LHCb two different tracker technologies are used: The silicon tracker placed close to the beam pipe, uses silicon microstripes. If a charged particle passes such a stripe it collides with the silicon atoms, liverating electrons and creating an electric current, which is then recorded. The outer tracker situated further from the beam pipe consists of gas-filled tubes. The gas ionizes when a charged particle hits a gas molecules, producing electrons. These reach an anode wire situated in the centre of each tube. The position of the track is found by timing how long it takes

electrons to reach it.
Calorimeters [12] : Calorimeters stop particles as they pass through, measuring the amount of energy lost. In LHCb there are two different types: The elctromagnetic calorimeter responsible for light particles like electrons and photons and the hadronic calorimeter responsible for heavier particles containing quarks. Both have a sandwich-like structure, with alternating layers of metal and plastic plates. If a particles hits a metal plate it produces a shower of secondary particles. These will excite polystyrenne molecules in the plastic plates, which then emit ultraviolet light. The energy lost by the particle in the metal plate is propoertional to the amount of UV light produced in the platic plates.
Muon System [13] : The muon system consistes of 5 rectangular stations, which cover an area of 435 m^2. Each station has chambers filled with three gases: carbon dioxide, argon and tetrafluoromethane. Passing muons react with the mixture and electrodes detect the result.

3.2 The LHCb trigger system

The rate of events at the LHCb interaction point is 40 MHz. But the rate to have a B meson contained in the detector is 15 kHz. But the offline computing power just allows 2 kHz to be recorded. The LHCb trigger system aims to 'fill' this 2 kHz with intresting B decays and important control decays like J/ψ decays. The trigger has two levels:
The **Level Zero (L0)** trigger reduces the beginning 40 MHz to 1 MHz. To get this high rate it can only rely on fast sub-detectors as the calorimeters and the muon system. The L0 trigger looks for events with high transverse momentum with respect to the patrticle beam axis (pT), because particles from a B decay have this attribute, since B Mesons are always produced almost parallel to the beam axis. In addition the L0 trigger performs a simplified vertex reconstruction with the signal of two silicon layers of the VELO to identifie events with multple proton-proton collisions. They are rejected because for this kind of events its much more difiicult to reconstruct B meson decays, since it is harder to distinguish primary and secondary vertex of the B decay.
The **High Level Trigger (HTL)** is an algorithm that runs on a farm of 1000 16-core computers. It has two stages: HLT1 which reduces the event rate to a few tens of kHz and HLT2 which reduces the rate to the 2 kHz which are recorded. HLT1 gets all the candidates of the L0 trigger and uses the full detector information on them to search for particles with a high impact parameter with respect the proton-proton collision. These particles are most likly decay products from B mesons, because of its relatively long life-time. They typically fly 1 cm away from the collision before decaying resulting in a high impact parameter for the decay products. HLT2 does a complete reconstruction of the events. It starts with the track of the VELO and connects them to the tracks in the other sub-detectors. Most important are displaced vertices, since they are strong indicator for B decays. The selection is devided into two parts. The inclusive selection searches for resonance decays like D^* or J/ψ. The exclusive selection is desigened to provide the highest possible efficiency to fully reconstruct B decays of interest. It therfor uses all information available such as mass and vertex quality and intermediate resonances.

4 The $B \to K^*$ μ μ decay

4.1 Motivation to analyse the $B_0 \to K^* \mu^+ \mu^-$ decay

According to the Standard Model the three leptons electron (e), muon (μ), and tau (τ), differ only in their masses. Therefore, in high energy regions (> 1 TeV) where masses become negligible, these leptons should behave the same. This phenomena is called lepton universality.
Recent experimental results of the LHCb collaboration [2] suggest a violation of the lepton universality: They analysed the decay of the B meson with a Kaon and 2 muons in the final state, and the decay of the B with a Kaon and 2 electrons in the final state.
The B_0 is a pseudoscalar meson made of an anti-b- and a d-quark. K^* (892) is a vector meson made of an anti-s and a d-quark which promptly decays to a charged kaon and a pion.
The LHCb collaboration measured the fractions of four B0 decays with a b to s quark transition and two leptons in the final state:

$$\begin{aligned} &B_0 \to K_0^* \mu^+ \mu^- \\ &B_0 \to J/\psi(\to \mu^+ \mu^-) \\ &B_0 \to e^+ e^- \\ &B_0 \to J/\psi(\to e^+ e^-) \end{aligned} \tag{5}$$

The following double ratio of these branching fractions are considered as well defined test of lepton universaility and reduces systematic uncertainties.

$$\begin{aligned} R_{K*0} &= \frac{\mathcal{B}(B^0 \to K^{*0} \mu^+ \mu^-)}{\mathcal{B}(B^0 \to K^{*0} J/\psi(\to \mu^+ \mu^-))} \Bigg/ \frac{\mathcal{B}(B^0 \to K^{*0} e^+ e^-)}{\mathcal{B}(B^0 \to K^{*0} J/\psi(\to e^+ e^-))} , \\ R_{K*0} &= \begin{cases} 0.66^{+0.11}_{-0.07}(\text{stat}) \pm 0.03(\text{syst})\text{for}0.045 < q^2 < 1.1\text{GeV}^2/c^4, \\ 0.69^{+0.11}_{-0.07}(\text{stat}) \pm 0.05(\text{syst})\text{for}1.1 < q^2 < 6.0\text{GeV}^2/c^4. \end{cases} \end{aligned} \tag{6}$$

Where q is the invariant mass of the di-muon system. The mesurement shows a 2.1-2.3 and 2.4-2.5 σ deviation from the Standart Modell in the two q^2 regions, respectively. To investigate further the same measurment is performed with new data from the LHCb detector. The goal of this thesis is to contribute to this new measuremnt, by testing a new method and by reweighting the monte carlo simulation to match the data.

4.2 Kinematics

In this section the decay itself and its kinematics are explained: The Decay is a flavor changing neutral current (FCNC) with four charged particles in the final state. The FCNC is a current, which changes the falvor of a fermion without changing its electric charge. The four particles in the final stage are:
The K^+ and π^- from the K^* decay and two leptons from the loop or box diagrams:

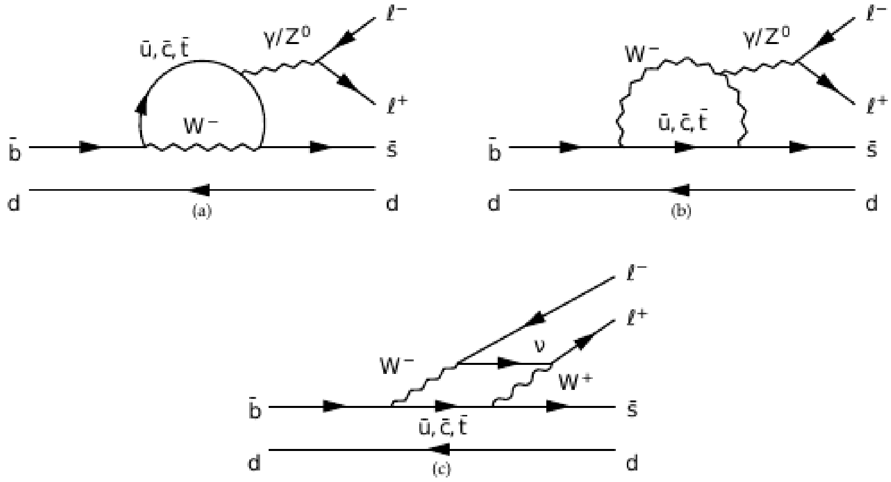


Figure 10: Feynman diagrams for decay $B_0(\bar{b}d) \to K(\bar{s}d)\ l^+\ l^-$ at lowest order, where l denotes leptons.

The kinematics of the decay is defined by the three angels θ_K, θ_L and ϕ, shown in figure 11 and the invariant di-muon mass square q^2.

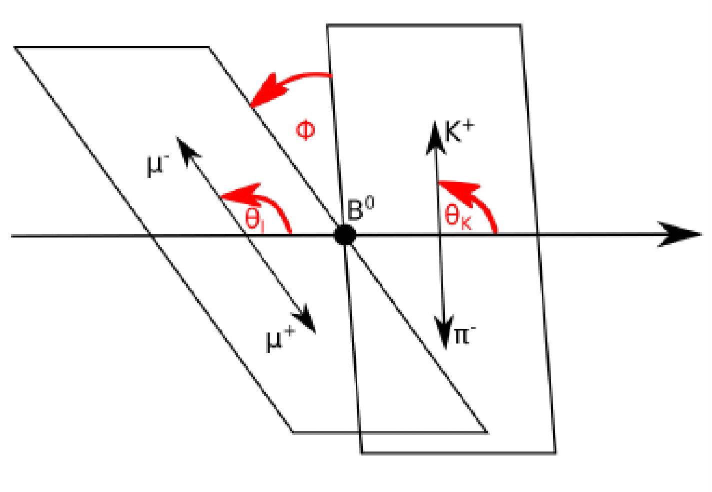


Figure 11: kinematic variables of the decay $B^0 \to K^{*0}\ \mu\ \mu$

The decay rate of the B_0 depends only on the kinematic variables θ_K, θ_L and ϕ and the invariant di-muon mass square q^2. The mathematical paramatrisation of the differential decay rate can be written as:

$$
\begin{aligned}
\frac{d^4\Gamma}{d\cos\theta_L d\cos\theta_K d\phi dq^2} &= \frac{9}{32\pi} I(q^2, \theta_L, \theta_K, \phi) \\
\text{with: } I(q^2, \theta_L, \theta_K, \phi) &= I_1^S \sin^2(\theta_K) + I_1^C \cos^2\theta_K + \left(I_2^S \sin^2\theta_K + I_2^C \cos^2\theta_K\right) \cos^2\theta_L \\
&+ I_3 \sin^2\theta_K \sin^2\theta_L \cos 2\phi + I_4 \sin 2\theta_K \sin 2\theta_L \cos\phi \\
&+ I_5 \sin 2\theta_K \sin\theta_L \cos\phi \\
&+ \left(I_6^S \sin^2\theta_K + I_6^C \cos^2\theta_K\right) \cos\theta_L + I_7 \sin 2\theta_K \sin\theta_L \sin\phi \\
&+ I_8 \sin 2\theta_K \sin 2\theta_L \sin\phi + I_9 \sin^2\theta_K \sin^2\theta_L \sin 2\phi
\end{aligned}
\tag{7}
$$

Where the angular coefficents I_i are functions of q^2. This decay rate is defined as the probability per unit time that the particle will decay. In section 2 the fitting algorithm is described which will be able to fit the distribution of this decay rate and determ the coeffients $I_{1,..,9}$.

4.3 Operators for $B \to X_s l^+ l^-$ decays

In this section all the operators of the B decay into a stange Meson (X_s) and two leptons are derived from the effective Lagrangian. The operators $\mathcal{O}_i$ in Quantum field theory are used to describe the creation or destruction of particles, by applying them to a quantum field. This section should just give an overview. If you need more detailed information consider reading [3].
The effective Lagrangian for $B \to X_s l^+ l^-$ decays has the form:

$$
\begin{aligned}
\mathcal{L}_{eff} &= \mathcal{L}_{QCD,QED}(u, d, s, c, b, e, \mu, \tau) \\
&+ \frac{4G_F}{\sqrt(2)} \left[V_{us}^* V_{ub} (C_1^c \mathcal{O}_1^u + C_2^c \mathcal{O}_2^u) + V_{cs}^* V_{cb} (C_1^c \mathcal{O}_1^c + C_2^c \mathcal{O}_2^c)\right] \\
&+ \frac{4G_F}{\sqrt(2)} \sum_{i=3}^{10} \left[(V_{us}^* V_{ub} + V_{cs}^* V_{cb}) C_i^c + V_{ts}^* V_{tb} C_i^t\right] \mathcal{O}_i.
\end{aligned}
\tag{8}
$$

Where V_{ij} is an element of the CKM Matrix, for the tranistion from the quark i to the quark j. C_i are the Wilson coeffecients described in section 4.4. The first term in equation 8 contains the kinetic terms of the light SM particles as well as their QCD and QED interactions. The remaining two terms consist of $\Delta B = -\Delta S = 1$ local operators of dimension ($d \leq 6$), which contain those light fields. The mass of the s quark can be negleted in comparisson with the b mass. One gets the following operators [3]:

$$
\begin{aligned}
\mathcal{O}_1^u &= (\bar{s}_L \gamma_\mu T^a u_L)(\bar{u}_L \gamma^\mu T^a b_L), & \mathcal{O}_6 &= (\bar{s}_L \gamma_{\mu_1} \gamma_{\mu_2} \gamma_{\mu_3} T^a b_L) \sum_q (\bar{q} \gamma^{\mu_1} \gamma^{\mu_1} \gamma^{\mu_1} T^a q), \\
\mathcal{O}_2^u &= (\bar{s}_L \gamma_\mu u_L)(\bar{u}_L \gamma^\mu b_L), & & \\
\mathcal{O}_1^c &= (\bar{s}_L \gamma_\mu T^a c_L)(\bar{c}_L \gamma^\mu T^a b_L), & \mathcal{O}_7 &= \frac{e}{g^2} m_b (\bar{s}_L \sigma^{\mu\nu} b_R) F_{\mu\nu}, \\
\mathcal{O}_2^c &= (\bar{s}_L \gamma_\mu c_L)(\bar{c}_L \gamma^\mu b_L), & \mathcal{O}_8 &= \frac{1}{g} m_b (\bar{s}_L \sigma^{\mu\nu} T^a b_R) G_{\mu\nu}^a, \\
\mathcal{O}_3 &= (\bar{s}_L \gamma_\mu b_L) \sum_q (\bar{q} \gamma^\mu q), & & \\
\mathcal{O}_4 &= (\bar{s}_L \gamma_\mu T^a b_L) \sum_q (\bar{q} \gamma^\mu T^a q), & \mathcal{O}_9 &= \frac{e^2}{g^2} (\bar{s}_L \gamma_\mu b_L) \sum_l (\bar{l} \gamma^\mu l), \\
\mathcal{O}_5 &= (\bar{s}_L \gamma_{\mu_1} \gamma_{\mu_2} \gamma_{\mu_3} b_L) \sum_q (\bar{q} \gamma^{\mu_1} \gamma^{\mu_1} \gamma^{\mu_1} q), & \mathcal{O}_{10} &= \frac{e^2}{g^2} (\bar{s}_L \gamma_\mu b_L) \sum_l (\bar{l} \gamma^\mu \gamma_5 l),
\end{aligned}
\tag{9}
$$

Where u,s,c,b are the Up, Strange, Charme and Beauty quark respectivly. The γ_μ denotes the Dirac Gamma-Matrices and T^a the QCD color matrices. The $\sum_q$ and $\sum_l$ denote the sums over light quarks and all leptons, respectivly. F_μ^ν In the following the effect of each operator on a quatnum field is decribed:

- $\mathcal{O}_1^u$: The right side destroys a beauty quark (b) and creates an up quark (u), while the left side destroys an up quark (u) and creates an anti-strange quark ($\bar{s}$). Both transistion happen over QCD gluons (T^a).
- $\mathcal{O}_2^u$ describes the same transistion as $\mathcal{O}_1^u$, just via QED photons (γ_u).
- $\mathcal{O}_1^c$ and $\mathcal{O}_2^c$ describe the same QCD and QED transistion, just with a charme (c) instead of an up quark.
- $\mathcal{O}_3$ and $\mathcal{O}_4$ are first creating a quark to destroy it again in one loop, since the quark is arbitrary there is a sum over all quarks. $\mathcal{O}_3$ does it with gluons and $\mathcal{O}_4$ with photons.
- $\mathcal{O}_5$ ...

4.4 Wilson coefficients

In this section the Wilson coefficents are derived from the effective Hamiltonion of the deacy. The wilson coefficents determ how strong a certain coupling is, while the operators in section 4.3 define how the particles can couple together. The effective Hamiltonian for $b \to s\mu^+\mu^-$ transitions can be written as [5]:

$$H_{eff} = -\frac{4G_F}{\sqrt{2}}\left(\lambda_t H_e f f^{(t)} + \lambda_u H^{(}_{eff} u)\right) \tag{10}$$

The λ_i can be expressed with CKM combinations $\lambda_i = V_{ib}V_{is}^*$.

$$H_{eff}^{(t)} = C_1\mathcal{O}_1^c + C_2\mathcal{O}_2^C + \sum_{i=3}^{6} C_i\mathcal{O}_i + \sum_{i=7,8,9,10,P,S}(C_i\mathcal{O}_i + C_i'\mathcal{O}_i') \tag{11}$$

$$H_{eff}^{(u)} = C_1(\mathcal{O}_1^C - \mathcal{O}_1^u) + C_2(\mathcal{O}_2^C - \mathcal{O}_2^u). \tag{12}$$

The contribution of $H_{eff}^{(u)}$ has a double Cabibbo supression and is therfore usually dropped. It is kept here since it is sensitive to complex phases of decay amplitudes. The operators $P_{i\leq 6}$ are the same as for general $B \to X_s l^+ l^-$ decays, see equation 9. The remaining ones are given by:

$$\begin{aligned}
\mathcal{O}_7 &= \frac{e}{g^2}m_b(\bar{s}\sigma_{\mu\nu}P_R b)F^{\mu\nu}, & \mathcal{O}_7' &= \frac{e}{g^2}m_b(\bar{s}\sigma_{\mu\nu}P_L b)F^{\mu\nu},\\
\mathcal{O}_8 &= \frac{1}{g}m_b(\bar{s}\sigma_{\mu\nu}T^a P_R b)G^{\mu\nu a}, & \mathcal{O}_8' &= \frac{1}{g}m_b(\bar{s}\sigma_{\mu\nu}T^a P_L b)G^{\mu\nu a},\\
\mathcal{O}_9 &= \frac{e^2}{g^2}(\bar{s}\sigma_\mu P_L b)(\bar{\mu}\gamma^\mu\mu), & \mathcal{O}_9' &= \frac{e^2}{g^2}(\bar{s}\gamma_\mu P_R b)(\bar{\mu}\gamma^\mu\mu),\\
\mathcal{O}_{10} &= \frac{e^2}{g^2}(\bar{s}\gamma_m u P_L b)(\bar{\mu}\gamma^\mu\gamma_5\mu), & \mathcal{O}_{10}' &= \frac{e^2}{g^2}(\bar{s}\gamma_\mu P_R b)(\bar{\mu}\gamma^\mu\gamma_5\mu),\\
\mathcal{O}_S &= \frac{e^2}{16\pi^2}m_b(\bar{s}P_R b)(\bar{\mu}\mu), & \mathcal{O}_S' &= \frac{e^2}{16\pi^2}m_b(\bar{s}P_L b)(\bar{\mu}\mu),\\
\mathcal{O}_P &= \frac{e^2}{16\pi^2}m_b(\bar{s}P_R b)(\bar{\mu}\gamma_5\mu), & \mathcal{O}_P' &= \frac{e^2}{16\pi^2}m_b(\bar{s}P_L b)(\bar{\mu}\gamma_5\mu),
\end{aligned} \tag{13}$$

where m_b denotes the running b mass in the $\overline{MS}$ scheme and g is the strong coupling constant and $P_{L,R} = (1 \pm \gamma_5)/2$. In the Standart Modell the primed Operators with opposite chirality to the unprimed operators vanish or are highly suppresd as are the $\mathcal{O}_S$ and $\mathcal{O}_P$. The contributions of $\mathcal{O}_{1,2,3,4,5,6}$ are neglected, since they are either heavily constrained or their impact turns out to be generically very small. For example in the left-right symmetric models or throughout gluino contributions in a general Minimal Supersymmetric Standard Model.

The C_i coefficients in the equations 11 and 12 are called Wilson coefficients. They encode short-distance physics and New Physics effects. For the calculation a matching scale $\mu = m_W$ is chosen, in a pertubative expansion in powers of $\alpha_s(m_W)$. Then the Wilson coefficents are evolved down to scales $\mu = m_b$ according to the solutions of the renomalization group equations. Contributions by New Physics enter through $C_i(m_W)$, while the low scales are determined by the Standart Modell. To allow a more organized expansion of the Wilson coefficients in pertubation theory the factors $16\pi^2/g^2 = 4\pi/\alpha_S$ are included into the definitions of the operators $\mathcal{O}_{i\geq 7}$. All the C_i expand as:

$$C_i = C_i^{(0)} + \frac{\alpha_s}{4\pi} C_i^{(1)} + \left(\frac{\alpha_s}{4\pi}\right)^2 C_i^{(2)} + O(\alpha_s^3) \tag{14}$$

where $C_i^{(0)}$ is the tree-level contribution, which is quale to zero for all operators except $\mathcal{O}_2$ and $C_i^{(n)}$ denotes the n-loop contributions. Before discussing the Wilson coefficents in details, lets look at the Operators again; the operators $\mathcal{O}'_S$ and $\mathcal{O}'_P$ are given in terms of conserved currents. They carry no scale-dependence. They do not mix with other operators and their Wilson coefficents are at the matching scale. $\mathcal{O}_9$ is also given by conserved curents. It mixes with $\mathcal{O}_{1,2,3,4,5,6}$ via a virtual photon decaying into $\mu^+\mu^-$. In addition their is a scale depedence from the factor $1/g^2$. This dependence is also present in C_{10} which otherwise would be scale independent.

In equation **??** one can see that C_7 and C_9 always appear in a particular combination with oder Wilson coefficents in matrix elements. Therfore effective coefficients are defined:

$$\begin{aligned}
C_7^{eff} &= \frac{4\pi}{\alpha_s} C_7 - \frac{1}{3} C_3 - \frac{4}{9} C_4 - \frac{20}{3} C_5 - \frac{80}{9} C_6, \\
C_8^{eff} &= \frac{4\pi}{\alpha_s} C_8 + C_3 - \frac{1}{6} + 20 C_5 - \frac{10}{3} C_6, \\
C_9^{eff} &= \frac{4\pi}{\alpha_s} C_9 + \mathcal{Y}(q^2), \\
C_{10}^{eff} &= \frac{4\pi}{\alpha_s} C_{10}, \\
C_{7,8,9,10}^{\prime eff} &= \frac{4\pi}{\alpha_s} C'_{7,8,9,10},
\end{aligned} \tag{15}$$

$$\begin{aligned}
\text{where } \mathcal{Y}(q^2) = h(q^2, m_c) &\left(\frac{4}{3} C_1 + C_2 + 6 C_3 + 60 C_5\right) \\
&- \frac{1}{2} h(q^2, m_b) \left(7 C_3 + \frac{4}{3} C_4 + 76 C_5 + \frac{64}{3} C_6\right)_{env} \\
&- \frac{1}{2} h(q^2, 0) \left(C_3 + \frac{4}{3} C_4 + 16 C_5 + \frac{64}{3} C_6\right) \\
&+ \frac{4}{3} C_3 + \frac{64}{9} C_5 + \frac{64}{27} C_6.
\end{aligned} \tag{16}$$

The function $h(q^2, m_q)$ comes from the fermion loop and for completness is presented in equation 17 below. If you need more details consider reading [5].

$$h(q^2, m_q) = -\frac{4}{9}\left(\ln\frac{m_q^2}{\mu^2} - \frac{2}{3} - z\right) - \frac{4}{9}(2+z)\sqrt{|z-1|}\cdot\begin{cases}\arctan\frac{1}{\sqrt{z-1}} & z > 1\\ \ln\frac{1+\sqrt{1-z}}{\sqrt{z}} - \frac{i\pi}{2} & z \leq 1\end{cases} \tag{17}$$

$$z = \frac{4m_q^2}{q^2}$$

5 Classifiers test

In this section it is explained how to use the Machine Learning method Classification to separate combinatorial background from signal:

To do so Monte Carlo simulated data which contains only $B \to K^*\ \mu\ \mu$ decays is labeled with probability 1 to be signal. Then it is merged with real data and used to train the classifiers. But since the classification becomes naturally biased if the data to classify is the same as the training data, a technique called K-folding is used. K-folding seperates the data Monte Carlo mix into several parts called folds. To classify one fold all the other folds are used for training. After iterating over all the folds one has a complety classified data set without any bias.
The conribution of this thesis is that several different classifiers are tested to find the one best suited:

First the following list of classifers where tested and compared in terms of perfomance:

- Ada Boost [15]
- uGB [16] + knnAda (k-nearest-neighbor AdaBoost)
- uBoost [17]
- uGB [16] + Fl (flatness loss)
- xgb [**?**]
- sk_bdtg
- sk_bdt

The test was performed with 30000 events from the 2016 LHCb $B \to K^*\ \mu\ \mu$ data and 10000 events from the Monte Carlo simulation. The following list of parameters are used in brackets are the names of the parameters in the root files.

- Decay vertex location for reconstructed particles (ENDVERTEX)
- Primary vertex location (OWNPV)
- Impact parameter (IP_OWNPV)
- Flight distance (FD_OWNPV)
- The cosine of the angle between primary vertex and decay vertex and recorded momentum (DIRA_OWNPV)

To compare the different classifiers the ROC curves and the correlations to the kinematic variabels of the decay (see section 4.2) are used. The receiver operating characterisitc (ROC) curve is a graphical plot that illustrates the ability of a classifier to correctly classify the data as its dicrimination threshold varies. Since the curves for the different folds all look alike only one is presented in figure 12. One can find the other nine curves in the appendix **??**. It turns out that all the classifiers classify the data correctly with just some minor variances. The ROC curve is in that case not a good tool to compare the different classifiers.

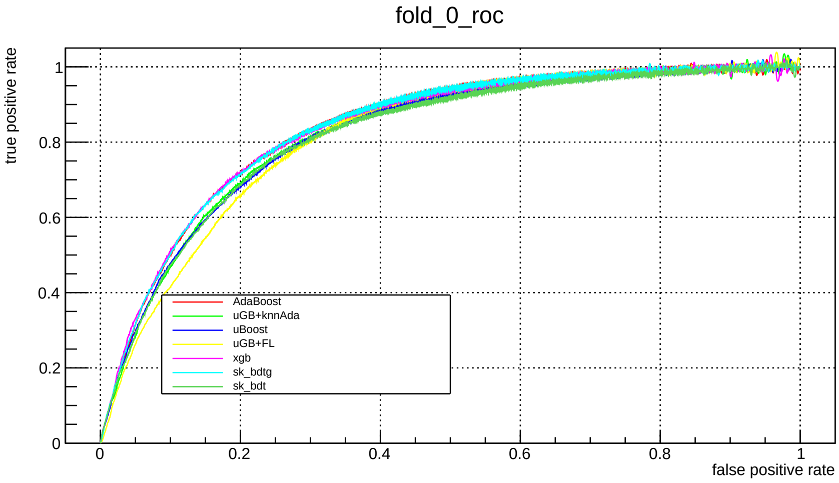


Figure 12: ROC curve of the first fold. One can see that all the classifiers are competetive in terms of classifing correctly

The next step is to check for correlations between the classifier response and the kinematic variables described in section 4.2. There should be no correlation between them, since the classifier uses other parameters to classify the data. But the classifier might pick up a correlation since the paramerters used for classification are not independent of the kinematic variables. That behavior should be avoided since it will bias the

only the correlations with the mass are shown here. The others can be found in appendix **??**.

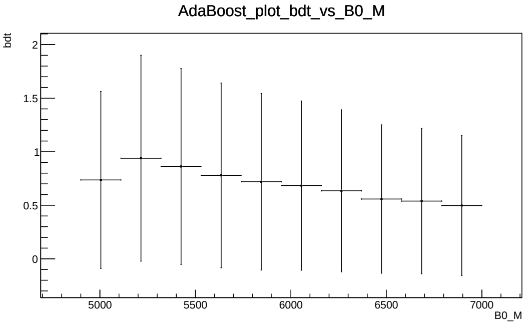


Figure 13: This plot shows the correltation between the bdt decision of the AdaBoost classifier and the B_0 mass. There is clearly a correlation starting from the second bin from the left.

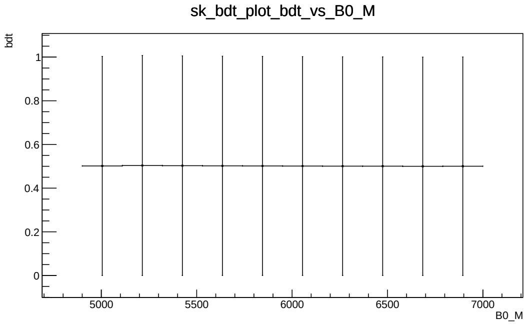


Figure 14: This plot shows the correltation between the bdt decision of the sk_bdt classifier and the B_0 mass. There is no correlation.

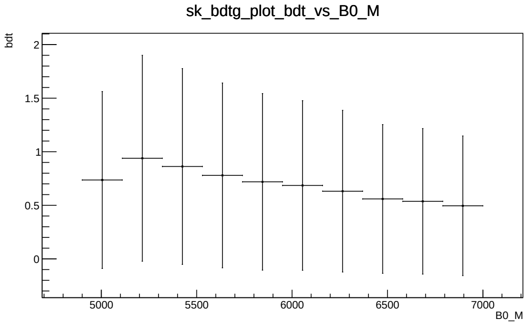


Figure 15: This plot shows the correltation between the bdt decision of the sk_bdtg classifier and the B_0 mass. There is clearly a correlation starting from the second bin from the left.

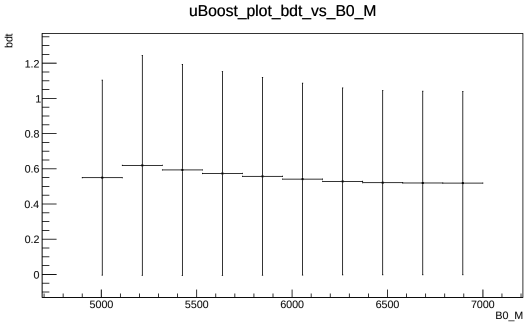


Figure 16: This plot shows the correltation between the bdt decision of the uBoost classifier and the B_0 mass. There is just a very small correlation compared to other classifiers.

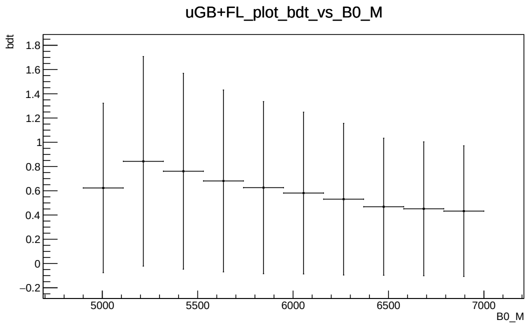


Figure 17: This plot shows the correltation between the bdt decision of the uGB+FL classifier and the B_0 mass. There is clearly a correlation starting from the second bin from the left.

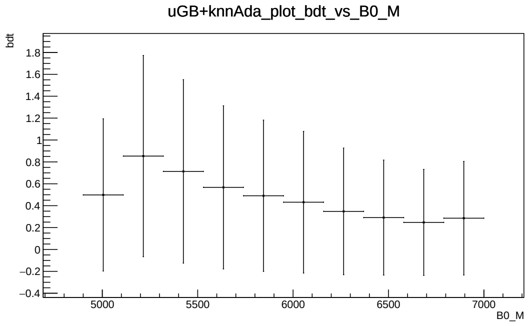


Figure 18: This plot shows the correltation between the bdt decision of the uGB+knnAda classifier and the B_0 mass. There is clearly a correlation starting from the second bin from the left.

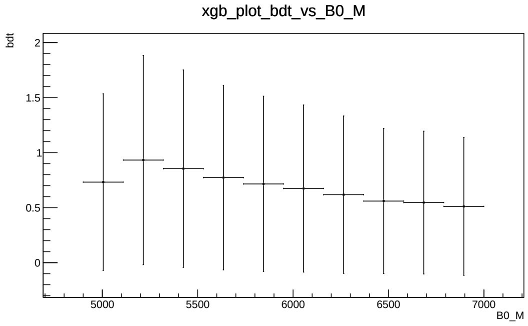


Figure 19: This plot shows the correltation between the bdt decision of the xgb classifier and the B_0 mass. There is clearly a correlation starting from the second bin from the left.

The two classifers with the least correlation are the sk_bdt and the uBoost classifers. Therfore they were used to do the seperation between signal and background in the $B_0 \rightarrow K^* \ \mu \ \mu$ data.

6 Reweighting

In this section the results from the reweighting of the $B_0 \rightarrow K^* \ \mu \ \mu$ Monte Carlo simulation is presented:

6.1 SPlot

In this section the SPlot Technique to sperate two or more merged distributions is explained. It has been used to get the inital weights for the Reweighting in chapter 6.

6.1.1 Likelihood method

Consider an analysis of a data sample, which consists of several types of events. These types represent signal components and background components, for example from different experiments. The log-liklihood of such a data sample is expressed as:

$$L = \sum_{i=1}^{N} \ln \left[\sum_{j=1}^{N_S} N_j f_j(y_i) \right] - \sum_{i=j}^{N_S} N_j \tag{18}$$

– N = total number of events

– N_S = number of types

– N_i = expected average number of events for type i

– y = set of diciminating variables

– f_j = PDF of the ith type

– $f_j(y_i)$ = value of PDF for event y_i

– x = control variable, not a part of L by construction

The yields N_i and the free parameters of the PDF are obtained by maximizing the above log-likelihood (eq 18).

6.1.2 $_{in}Plot$ technique

Consider a varibale x which can be expressed as a function of the dicriminating variables y used in the fit. Furthermore a fit has been performed to determine the yields N_i for all types. From the knowledge of the PDF and the values of N_i a naive weight can be defined as:

$$P_n(y_i) = \frac{N_n f_n(y_i)}{\sum_{k=1}^{N_s} N_k f_k(y_i)} \tag{19}$$

which will leed to the x-distribution $\tilde{M}_n$ defined by:

$$N_n \tilde{M}_n(\bar{x}) = \sum_{i \subset \delta x} P_n(y_i) \tag{20}$$

where sum $\sum_{i \subset \delta x}$ contains alle events for which x_i lies in the interval centered on $\bar{x}$ and of total width δx. Therefor $N_n \tilde{M}_n(\bar{x}) \delta x$ is the x-distribution of the histogrammed events, using the weights of eq 19.

With this procedure one can on average reproduce the true distribution $\mathbf{M}_n(x)$. One can even replace the sum in eq 20 by an integral:

$$\left\langle \sum_{i\subset\delta x} \right\rangle \to \int dy \sum_{j=1}^{N_s} N_j f_j(y)\delta(x(y)-\bar{x})\delta x \tag{21}$$

Furthermore through identifying the number of events N_i from the fit one gets:

$$\langle N_n\rangle \tilde{M}_n(\bar{x}) = \int dy \sum_{j=1}^{N_s} N_j f_j(y)\delta(x(y)-\bar{x})P_n(y) \tag{22}$$

$$= \int dy \sum_{j=1}^{N_s} N_j f_j(y)\delta(x(y)-\bar{x}) \cdot \frac{N_n f_n(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \tag{23}$$

$$= N_n \int dy\delta(x(y)-\bar{x})f_n(y) \tag{24}$$

$$= N_n \mathbf{M}_n(\bar{x}) \tag{25}$$

One can see that the sum over enents of the daive weight P_n provides a direct estimate of the x-distribution for the nth type. But this procedure has a major drawback, since x is correlated to y, the PDFs of x enter implicity in the definition of the naive weight. Therfor the $\tilde{M}_n$ distributions are a bad estimate for the quality of the fit, since the these distributions are biased in a difficult way, when the PDFs $f_i(y)$ are not accurate.

Consider for example a data sample where one of the types has events on the tail of the x-distribution. Such events require the true ditribution to account for the tail. But since the events are averaged the weights on the tail are going to be very small missing those events in the extimated true distribution. Only the core of the x-distribution can be examined with ${}_{in}Plots$.

6.1.3 ${}_{S}Plot$ technique

In the previous section it was shown that if a variable x belongs to a set y of discriminating varibales, one can reconstruct the expected x distribution. Consider now two sets of variables x and y, where x does not belong to y and which are uncorrelated , hence the total PDFs $f_i(x,y)$ all factorize into products $\mathbf{M}_i(x)f_i(y)$. The equation 25 does not hold anymore because, when summing over the events the x-PDFs $\mathbf{M}_j(x)$ appear:

$$\langle N_n\rangle \tilde{M}_n(\bar{x}) = \int\int dydx \sum_{j=1}^{N_s} N_j \mathbf{M}_j(x) f_j(y)\delta(x-\bar{x})P_n \tag{26}$$

$$= \int dy \sum_{j=1}^{N_s} N_j \mathbf{M}_j(\bar{x}) f_j(y) \frac{N_n f_n(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \tag{27}$$

$$= N_n \sum_{j=1}^{N_s} \mathbf{M}_j(\bar{x}) \left(N_j \int dy \frac{f_n(y) f_j(y)}{\sum_{k=1}^{Ns} N_k f_k(y)} \right) \tag{28}$$

$$\neq N_n \mathbf{M}_n(\bar{x}). \tag{29}$$

The correction term

$$N_j \int dy \frac{f_n(y) f_j(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \tag{30}$$

is not identical to the kroenecker delta δ_{jn}. In fact the $N_n \tilde{M}_n$ distribution obtained by the naive weight is a linear combination of the true distribution $\mathbf{M}_j$.

To go forward one has to realize that the correction term is related to the inverse of the covariance matrix, given by the second derivatives of $-L$, after the minimization.

$$\mathbf{V}_{nj}^{-1} = \frac{\partial^2(-L)}{\partial N_n \partial N_j} = \sum_{i=1}^{N} \frac{f_n(y_i) f_j(y_i)}{\left(\sum_{k=1}^{N_s} N_k f_k(y_i)\right)^2} \tag{31}$$

If one averages and is replacing the sum over events by intergals (eq 21) the varaince matrix reads:

$$\langle \mathbf{V}_{nj}^{-1} \rangle = \int \int dydx \sum_{e=1}^{N_s} N_e \mathbf{M}_e(x) f_e(y) \frac{f_n(y) f_j(y)}{\left(\sum_{k=1}^{N_s} N_k f_k(y)\right)^2} \tag{32}$$

$$= \int dy \sum_{e=1}^{N_s} N_e f_e(y) \frac{f_n(y) f_j(y)}{\left(\sum_{k=1}^{N_s} N_k f_k(y)\right)^2} \cdot \int dx \mathbf{M}_l(x) \tag{33}$$

$$= \int dy \frac{f_n(y) f_j(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \tag{34}$$

Therefor equation 26 can be rewritten as:

$$\langle \tilde{M}_n(\bar{x}) \rangle = \sum_{j=1}^{N_s} \mathbf{M}_j(\bar{x}) N_j \langle \mathbf{V}_{nj}^{-1} \rangle. \tag{35}$$

To get the distribution of intrest one has to invert this matrix equation:

$$N_n \mathbf{M}_n(\bar{x}) = \sum_{j=1}^{N_s} \langle \mathbf{V}_n j \rangle \langle \tilde{M}_j(\bar{x}) \rangle \tag{36}$$

The true distribution of x can still be reconstructed using the naive weight (eq 19), through a linear combination of ${}_{in}Plots$. In other words: When x does not belong to the set y, the weights are not given by equation 19, they are given by a covariance-weighted quantity called sWeight defined by:

$${}_sP_n(y_i) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_n j f_j(y_i)}{\sum_{k=1}^{N_s} N_k f_k(y_i)} \tag{37}$$

With the sWeights on can obtain the distribution of the x variable by histogramming the ${}_sPlot$:

$$N_{ns} \tilde{M}_n(\bar{x}) \delta x = \sum_{i \subset \delta x} {}_sP_n(y_i) \tag{38}$$

On average it reproduced the true distribution:

$$\langle N_{ns} \tilde{M}_n(x) \rangle = N_n \mathbf{M}_n(x) \tag{39}$$

In the case were x is significantly correlated with y, the ${}_sPlots$ from equation 38 connt be compared with the pure distributions of the various types. To solve that problem one can perform a Monte Carlo simulation of the procedure and obtain the expected distributions to which the ${}_sPlots$ should be compared with.

For more information on ${}_sPlots$ consider reading [14].

References

[1] MINUIT Home page, *https://seal.web.cern.ch/seal/snapshot/work-packages/mathlibs/minuit/*

[2] JHEP08 (2017) 055

[3] C. Bobeth, M. Misiak and J. Urban, Nucl. Phys. B 574 (2000) 291 [arXiv:hep-ph/9910220].

[4] arXiv:1709.01051v1 [hep-ph] 4 Sep 2017

[5] arXiv:0811.1214 [hep-ph]

[6] CERN Accelerator Complex, *http://www.stfc.ac.uk/research/particle-physics-and-particle-astrophysics/large-hadron-collider/cern-accelerator-complex/*

[7] Science and Technology Facilities Council article about LHCb , *https://www.ppd.stfc.ac.uk/Pages/LHCb.aspx*

[8] VELO description, *http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/VELO2-en.html*

[9] RICH description, *http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/RICH2-en.html*

[10] Magnet description, *http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/Magnet2-en.html*

[11] Tracker description, *http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/Trackers2-en.html*

[12] Calorimeters description, *http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/Calorimeters2-en.html*

[13] Muon system description, *http://lhcb-public.web.cern.ch/lhcb-public/en/Detector/Muon2-en.html*

[14] sPlot: a statistical tool to unfold data distributions, arXiv:physics/0402083 [physics.data-an]

[15] A Short Introduction to Boosting, by Yoav Freund and Rovert E. Schapire *http://www.site.uottawa.ca/~stan/csi5387/boost-tut-ppr.pdf*

[16] J.H. Friedman Ǵreedy function approximation: A gradient boosting machine, 2001.

[17] J. Stevens and M. Williams, uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers, JINST 8, P12013 (2013). [arXiv:1305.7248]

[18] arXiv:1603.02754 [cs.LG]