Università degli studi di Trieste

Exploration of advanced statistical-learning methods for the optimal classification of flavor-physics collider data

Tesi di laurea in Fisica

Author:
Marco Sciorilli

Supervisor:
dr. Diego Tonelli

Co-Supervisor:
Eldar Ganiev

Ottobre 2020

Università degli studi di Trieste

Esplorazione di metodi avanzati di auto-apprendimento statistico per la classificazione ottimale di collisioni finalizzata alla fisica del sapore

Tesi di laurea in Fisica

Autore:
Marco Sciorilli

Relatore:
dr. Diego Tonelli

Correlatore:
Eldar Ganiev

Ottobre 2020

Riassunto

Lo scopo di questa tesi di fisica sperimentale delle particelle è lo sviluppo e l'applicazione di un classificatore ad auto-apprendimento guidato alla selezione del canale di decadimento $B^0 \to D^-[\to K^+\pi^-\pi^-]\pi^+$ nei primi dati raccolti dall'esperimento Belle II. Belle II è un rivelatore dedicato alla ricostruzione dei prodotti di collisioni elettrone-positrone ad alta intensità ad energia di 10 GeV, al fine di sondare i limiti del Modello Standard, la teoria attualmente accettata.

Il lavoro consiste di due parti. Nella prima parte studio e seleziono le variabili discriminati utili per l'isolamento del canale d'interesse usando una simulazione. Nella seconda, applico il metodo di classificazione scelto, un *boosted decision-tree*, al primo campione di dati raccolto da Belle II, e ne valuto le prestazioni osservate.

Il classificatore sviluppato in questo lavoro apporta un miglioramento di un fattore 4 sul rapporto segnale su fondo a fronte di un'inefficienza sul segnale del 16 %.

Abstract

This is an experimental particle-physics thesis devoted to the development and implementation of a supervised-learning classifier for the selection of the decay channel $B^0 \to D^-[\to K^+\pi^-\pi^-]\pi^+$ in the first data collected by the Belle II experiment. Belle II is a detector dedicated to the study of the products of high-intensity 10 GeV electron-positron collisions, to explore extensions of the Standard Model, the currently accepted theory.
My work is organized in two parts. In the first part I explore and select the discriminating variables useful for the identification of the decay channel of interest using a simulation. Then I apply the chosen classification method, a boosted decision-tree, to the first data collected by Belle II, in spring 2019, and assess performance. The classifier developed in this work improves the signal-to-background ratio by a factor of 4 with a signal inefficiency of 16 %.

Contents

Chapter 1

Introduction

The Standard Model is the currently accepted theory of particle physics. Even though it reproduces thousands of experimental results at energies ranging from eV to TeV, theoretical considerations and experimental inconsistencies support the common belief that the Standard Model needs to be completed by a more general theory valid over a broader range of energies in the higher end of the spectrum. Completing the Standard Model is the principal objective of particle physics today.

Belle II is an experiment located in Tsukuba, Japan, designed, built, and operated by over 1000 physicists, with the main purpose of exploiting high-intensity 10 GeV electron-positron collision produced by the superKEKB accelerator to explore extensions of the Standard Model. It started taking data in March 2019, and has now collected samples corresponding to 80 million pairs of B mesons (bound states of a b quark and a lighter u or d quark).

Not all events observed are interesting for the Belle II scientific program. The decay channels used to test the Standard Model are a small subset of the total processes that occur in the collisions. It is therefore necessary to isolate such decays, the signal of interest, from the large amount of background. This kind of analysis of data is well suited for automation, exploiting the tools developed in recent years dedicated to classifying clusters of data in fields other than particle physics.

This thesis focuses on devising an optimized selection for a specific decay channel of the B^0 meson, the $B^0 \to D^-[\to K^+\pi^-\pi^-]\,\pi^+$ decay, reconstructed in the very first Belle II data, which correspond to about 5 million pairs of B mesons. I explore the use of a statistical learning method, the boosted decisions tree, as a way to automatize in an optimal way the identification of this decay channel from the prevailing background. My work mainly focuses on the development of a suitable middle ground between the need for an accurate classifier, and the constrains imposed by a reasonable calculation time. The results are compared with traditional classification approaches.

In the first chapter, I give a brief overview of Belle II physics and on the experiment. In the second chapter I discuss the general features of the decay channels targeted in this study. In the third chapter I introduce decision trees as binary classifiers and give an

overview on the variables used for the classification, their discriminating power, and their similarity with simulation. In the fourth chapter, I apply such classification to an early Belle II data set, use common metrics to determine the performance of the classification, and compare the results obtained with the initial performance. Finally, I summarize the project.

Chapter 2

Flavor physics with the Belle II detector

Since its theoretical formulation was completed in the 70's, the Standard Model (SM) has been successfully describing three of the four known fundamental forces (electromagnetic, weak, and strong interactions). Even though the SM stands on solid experimental bases tested over decades with increasing precision, it leaves open various questions that suggest the need for a more general full theory valid over a broader range of energies. Such questions include the large discrepancy in strength between gravity and other forces, the lack of an explanation for a dynamical origin for the observed asymmetry between matter and antimatter in the universe, and the postulated large amounts of non interacting matter (so-called dark matter) introduced to justify cosmological observations. Identifying the theory that completes the Standard Model is one of the principal goals in fundamental physics. Two approaches are used. The energy-frontier direct approach consists in colliding beams of particles to reach collision energies high enough to produce non-SM particles of large mass, detect then their decay products, and thus gain a direct evidence of their existence and study their phenomenology. The approach has proved effective in the past, but it is limited in reach, since increasingly large energies are required to detect particles with masses of 10 TeV and beyond, leading to major technological and budgetary challenges.

The indirect, intensity-frontier approach relies on a peculiar aspect of the quantum-mechanical nature of the dynamics of particles. It consists in seeking small but significant differences between measurements and SM predictions of equal precision in lower-energy processes sensitive to non-SM contributions. Such approach relies on the temporary non-conservation of energy allowed by the Heisenberg's uncertainty principle, which enables observation of the contributions of (virtual) particles of arbitrary high mass through the alterations they induce in amplitudes of predictable processes. This approach is frequently used for processes associated with the weak interaction of quarks (so called 'flavor physics'). Among those, heavy charm and bottom quarks are specially promising due to their property of decaying in hundreds of different processes, many of which can be precisely predicted and measured. The violation of charge-parity (CP) symmetry (the symmetry that consist in exchanging all particles with their antiparticles and inverting

all spatial coordinates) has a special role in the study of flavor dynamics as it allows for probing the existence of new particles or interactions through measurements of the phases they introduce in the amplitudes, thus enhancing the probing sensitivity.

The focus of this thesis is the reconstruction of the decay $B^0 \rightarrow D^-[\rightarrow K^+\pi^-\pi^-]\,\pi^+$. B mesons are ground states composed by a valence bottom antiquark and one among u, d, s, c quarks. B^0 mesons are made by a $d\bar{b}$ pair, have zero electric charge, isospin $\frac{1}{2}$, spin 0, bottomness -1, a rest mass of about 5280 MeV/c^2 and a mean lifetime of about 1.5 ps. The decay $B^0 \rightarrow D^-[\rightarrow K^+\pi^-\pi^-]\,\pi^+$ is important as it is very similar to the much rarer decay $B^0 \rightarrow D^-[\rightarrow K^+\pi^-\pi^-]\,K^+$, which offers privileged access to crucial measurements associated with CP violation. Optimizing the reconstruction and identification of the $B^0 \rightarrow D\pi$ decay in early data allows preparing for performing analyses of $B \rightarrow DK$ decays, once the data set size will be sufficient.

2.1 The Belle II experiment at the SuperKEKB accelerator

Belle II is a particle-physics experiment located at the SuperKEKB electron-positron collider in Tsukuba, Japan. It succeeds the Belle experiment, and its main purpose is to study billions of $B^0\overline{B^0}$ and B^+B^- pairs produced with low background in decays of the $\Upsilon(4S)$ resonance. The collisions occur at an energy of 10.58 GeV, which corresponds to the mass of the $\Upsilon(4S)$ meson. The $\Upsilon(4S)$ meson is a bound state composed by a b quark and a $\bar{b}$ quark.

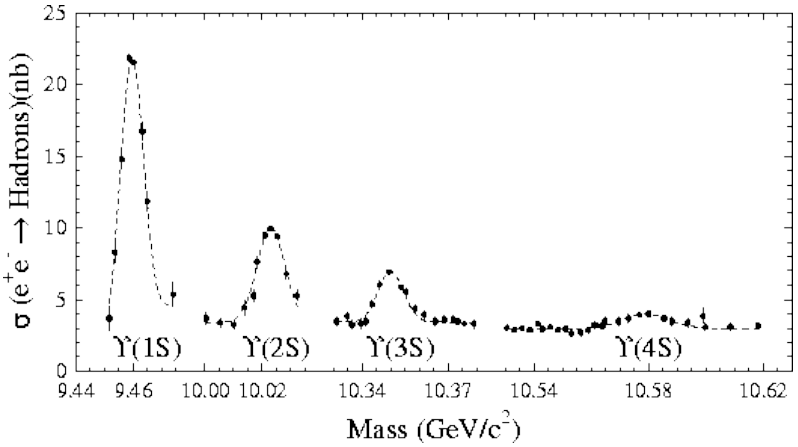

Figure 2.1: Cross section for the production of hadrons in electron-positrons collisions, as a function of collision energy.

Its fourth radially excited state has sufficient mass to decay in a $B^0\overline{B^0}$ or B^+B^- pair and nothing else 96% of the time (Fig.2.1), which offers an abundant low-background source of bottom mesons to be studied in the Belle II experiment in pristine experimental conditions.

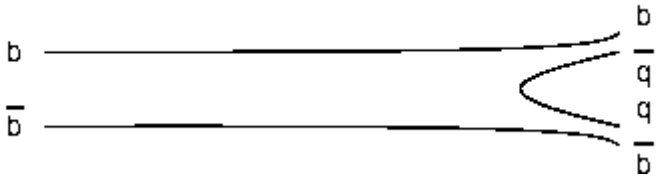

Figure 2.2: Scheme of the B-meson pair formation for a $\Upsilon(4S)$ decay.

The SuperKEKB accelerator is expected to reach a luminosity of 8×10^{35} cm^{-2}s^{-1} at its peak performance, corresponding to the production of about 1000 $B\overline{B}$ meson pairs per second. The SuperKEKB complex (shown in Fig.2.2) consist of a chain of accelerators that culminates in two storage rings of asymmetric energy; a 4 GeV ring for positron and and a 7 GeV ring for electron, with circumference per ring of 3016 m.

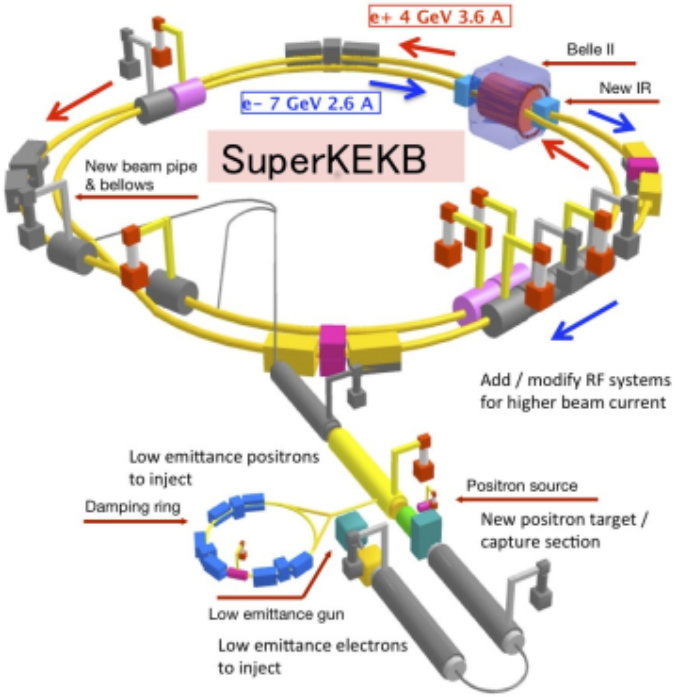

Figure 2.3: Scheme of the superKEKB collider.

2.2 The Belle II detector

With a width, height, and depth of 8 meters each, the Belle II detector (Fig. 2.4) is an assembly of sub-detectors built for of recording information on decays of B and D mesons, and τ leptons produced by SuperKEKB. Its main components are as follows.

- **The vertex detector**. This is an approximately cylindrical structure made of 6 radially concentric layers of two types of silicon sensors, microstrip sensors and pixel

sensors. They allow determining the position of the decay vertices with a precision of 30 μm and sample the trajectories of the charged particle close to the interaction region. This is important to distinguish B mesons, which are long-lived and travel a measurable distance before decay, from backgrounds.

- **The central drift chamber**. This is a large-radius drift chamber installed radially around the collision point. It is an hollow cylinder made of 56 layers of wires immersed in a gaseous mixture of He and C_2H_6, with a radius that span from 160 mm to 1130 mm. It reconstructs charged tracks, measures their momenta, identifies particles using measurements of specific ionization energy loss, and it provides trigger signals for charged particles.

- **The particle identification detectors**. These are two detectors based on the Cherenkov light technology: an aero-gel and a quartz bar ring-imaging Cherenkov counter. The former discriminate charged kaons and pions over most of their momenta, and pions, muons and electrons with momenta smaller than 1 GeV/c. The latter measures the time of propagation of the Cherenkov photons internally reflected inside the quartz radiators.

- **The electromagnetic calorimeter**. This is made of an highly-segmented array of thallium-doped, cesium iodide crystals assembled in a projective geometry pointing to the interaction region. It is organized as a 3-m-long barrel section with an inner radius of 1.25 m that covers the polar region $12.4^\circ < 0 < 155.1^\circ$. Its detection principle is based on scintillation, and its purpose is to measure the energy of photons, electrons, and kaons, for trigger tasks and luminosity measurements.

- **The K_L^0 and μ system**. This is a detector made of alternating patterns of 4.7-cm-thick iron plates and active detector elements. Its main purpose is to detect muons and K_L^0 mesons that escape from the internal region and its iron structure is used as return yoke for the magnetic field.

- **The superconducting NbTi/Cu coil**. Powered with 44 A, is used to provide an homogeneous magnetic field of 1.5 T parallel to the beam direction in the inner region to curve the charged-particle trajectories.

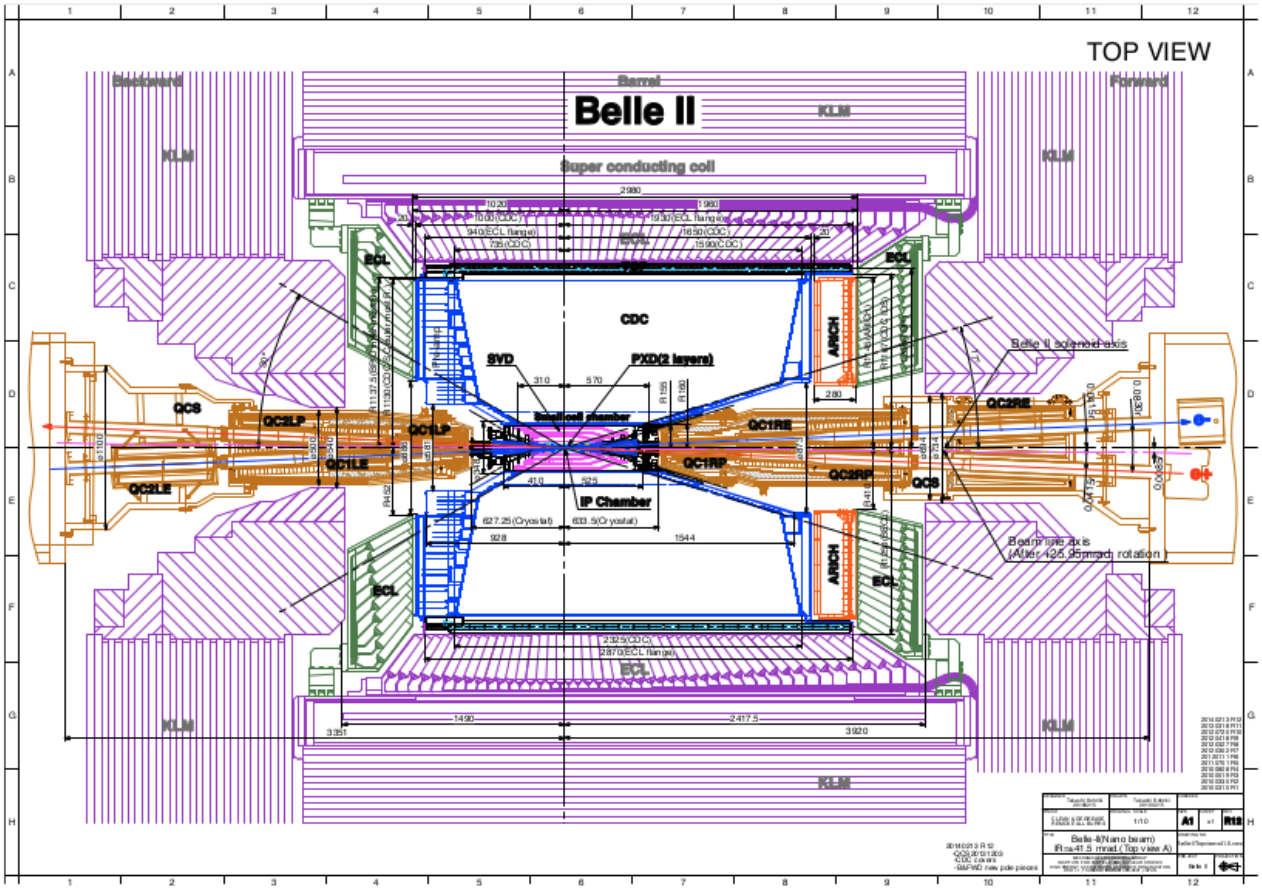

Figure 2.4: Top view of the Belle II detector.

2.3 Overview

The main goal of this thesis is the optimal selection of the decay channel $B^0 \to D^-[\to K^+\pi^-\pi^-]\pi^+$ in the very first data collected by Belle II. In order to achieve that, I developed and trained a boosted decision tree as a statistical learning tool. The decision tree is a supervised learning algorithm that requires an already classified data set. For this purpose I used Monte Carlo simulation to generate a data set in which events are known to belong to signal or background. Once the classifier is trained on simulation, I use it to classify the data collected by Belle II.

Both the Monte Carlo simulation and the data collected are formatted as ntuples, which are data structures organized into the basic unit of "events", which contain all information associated with the particles reconstructed in a collision, real or simulated.

Prior to introduce the reader to the details of my work, it may be useful to provide a

general overview of the reconstruction of a physics event in Belle II. In the SuperKEKB accelerator millions of collision occur every second. Of those only 0.4% create a $\Upsilon(4S)$ meson, which decay in a $B\overline{B}$ 96% of the times. Of those, only two out of thousand decay as $D^-[K^+\pi^-\pi^-]\pi^+$, the final state of interest, hence the necessity of an optimal classification. In signal events the B^0 meson is produced with ≈ 1 Gev/c momentum in the laboratory and, owing of its 1.5 ps lifetime, travels typically 150 μm before decaying into a D^- meson and a pion. The D^- meson travels another 50 μm before decaying into a kaon and three pions. The tracking volume is pervaded by an axial magnetic field, so the trajectories of all the charged particles are curved by the Lorentz force. While charged particles go through the position-sensitive layers of tracking detectors, they excite small electrical signals, which are clustered into hits. A geometric fit of the spatial pattern showed of hits allows the reconstruction of the trajectories. Then, the knowledge of magnetic field, detector material and trajectories space point of origin of the final-state particles, combined with the constrains imposed by the momentum and energy conservation, is used to reconstruct a kinematic fit of the decay, producing the informations stored in the ntuple.

Chapter 3

Data and principal observables

3.1 First look at data

In my first approach to data, I focused only on three baseline quantities that are straightforward to understand and already capture important features of the physics of the decay: the transverse (d_0) and longitudinal impact parameters (z_0), which indicate the transverse and longitudinal distance of the particle trajectory from the collision point, and the transverse momentum (p_t), which is the component of the momentum transverse to the direction of the beam measured in the center of mass.

I compared distributions of these kinematic quantities in simulated signal and background events to familiarize with the Belle II basic physics.

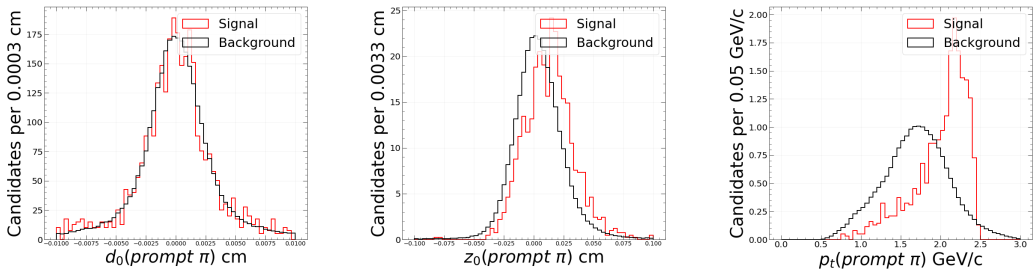

Figure 3.1: Comparison between simulated signal and background distributions of (left) transverse impact parameter, (center) longitudinal impact parameter, and (right) transverse momentum recorded in the center of mass, associated to the pion originated directly from the B decay.

Fig. 3.1 shows kinematic distributions of the quantities of interest for pions originated from the B decay. The d_0 distributions are similar between signal and background, while the distributions for z_0 and p_T show differences. In particular, signal in z_0 has higher mean valued than background. The transverse momentum also shows a clear difference with a harder spectrum for signal. The signal B meson has approximately 5 GeV/c^2 mass

and decays into a smaller number of final-state particles than those contributing to the background, which explains why signal transverse momenta are typically higher.

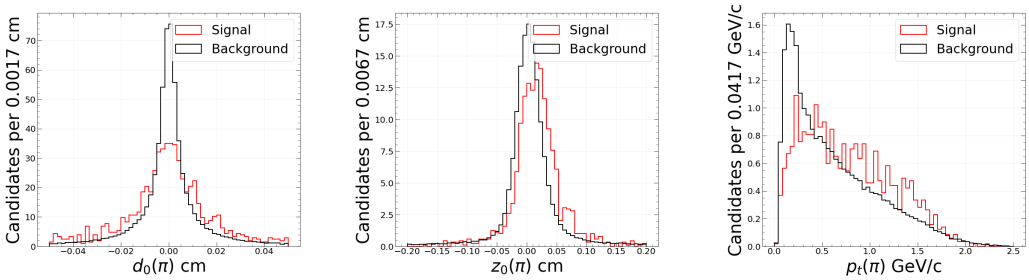

Figure 3.2: Comparison between simulated signal and background distributions of (left) transverse impact parameter, (center) longitudinal impact parameter, and (right) transverse momentum recorded in the center of mass, associated with the pion from the D decay.

Fig. 3.2 shows the kinematic distributions associated with the pions from the D decay. The d_0 distribution for signal is broader than for background. As the B meson has a mean lifetime of about 1.5 ps and is produced with momentum of ≈ 1 GeV, it decays after having traveled a measurable distance of about 150 μm. In addiction, the produced D meson travels about 50 μm further. The reconstructed trajectories of the D decay products are therefore unlikely to project back to the collision point. This leads to the signal d_0 distribution to be wider compared to the background, whose particles are generated in the center of the collision. This is reflected in z_0 too.

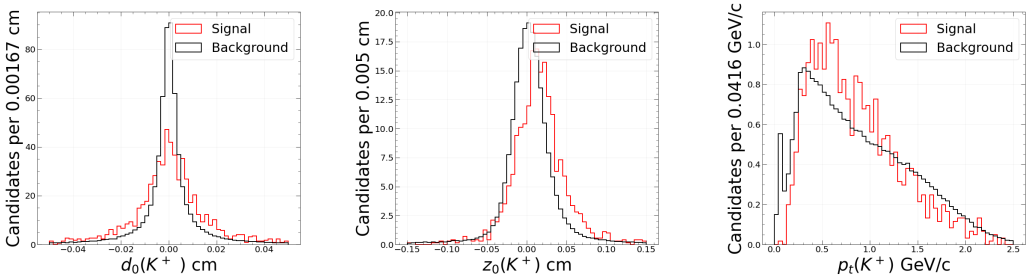

Figure 3.3: Comparison between simulated signal and background distribution of (left) transverse impact parameter, (center) longitudinal impact parameter, and (right) transverse momentum recorded in the center of mass, associated with K.

The features of the kaon distributions resemble those discussed for pions.

After a first exploration of the kinematic variables, it is useful to introduce the distribution of the difference between the expected and observed B candidate energy ΔE, which is the principal variable we use to identify the presence of signal and measure its size. If a B meson is produced in a collision and correctly reconstructed, the energy

of the decay products equals approximately half of the collision energy. Therefore, signals peak at zero in the ΔE distribution, while continuum background from light-quark pairs produced in non resonant collisions follows a smooth distribution, offering a striking discriminating information to identify signal.

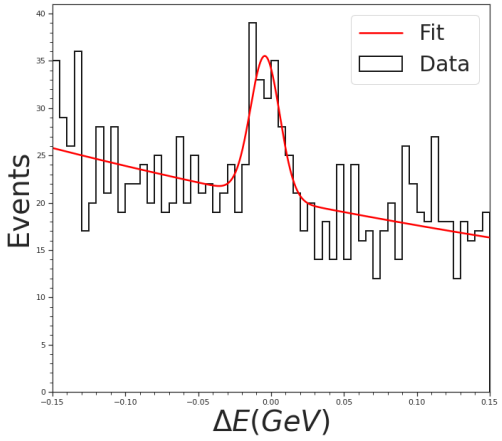

Figure 3.4: Distribution of difference between expected and observed B candidate energy (ΔE) in the first Belle II data.

Fig. 3.4 shows the ΔE of the data collected by Belle II in the spring of 2019. The data shown are subjected to preliminary criteria that suppress the most obvious background. Signal candidates not meeting the following restrictions were discarded:

- Transverse impact parameter of the D decay products: $|d_0| < 0.5$ cm
- Longitudinal impact parameter of the D decay products: $|z_0| < 3$ cm
- A lower threshold on a measure of kaon identification: $kaonID > 0.6$
- Difference between expected and observed B candidate energy: 0.15 GeV $< \Delta E <$ 0.15 GeV

The first two restrictions aim to remove tracks produced outside the interaction region, which are typically associated with particles from beam-background events that interact with the material surrounding the interaction region.

In the data distributions of ΔE I identify two features: a narrow structure centered on 0, which can be interpreted as signal, and an overlapping smoothly decreasing background. I fitted the distribution using a χ^2 fit with an empiric model made of the sum of a negative exponential, to account for background distribution, and a Gaussian function, to capture the shape of the signal distribution. The fit reports 75 $\pm$ 16 signal events, and 1241 $\pm$ 366 background events, with a signal-to-background fraction at peak of approximately 0.6.

Chapter 4

Classifier design

The main goal of the algorithm that is the object of this thesis is to identify differences between signal and background events, and exploit them for a statistical discrimination. The classifier combines variables chosen to exploit maximally the discriminating differences between the decay of interest, and everything else. However, increasing the number of variables employed comes with a cost in computation time. The choice of such variables plays then a major role in terms of classification performance and realistic applicability.

Here I illustrate the design of the classification algorithm I used, the study of variables I fed to it, and a straightforward criterion to sort them in a calculation-constrained situation. Among popular statistical learning methods suited for binary classification (such as neural networks, support vector machines, etc.), I choose a boosted decision tree solution. This is suited for its versatility on non linear classification problems, as those typically associated with collider data classification, and its algorithmic simplicity, which makes it low duty on calculation and easy to parallelize in a cluster-kind environment.

To implement the boosted decision tree, I used the dedicated software tool (FBDT) that is part of the Belle II Analysis Software Framework (BASF2), a larger framework that includes all the software necessary to analyse the data collected by the Belle II detector [10].

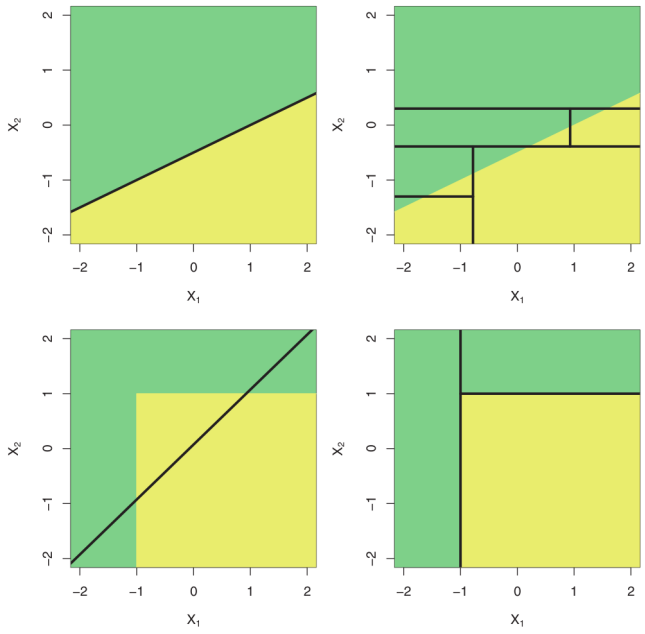

Figure 4.1: top row: A two-dimensional classification example in which the true decision boundary is linear. The two discriminating variables are labeled as x_1 and x_2; the two class of events are depicted in different colors. The use of a linear boundary (left) outperforms a decision tree that implements splits parallel to the axes (right).
Bottom row: a two-dimensional classification example in which the true decision boundary is non-linear. A linear model is unable to fully capture the true decision boundary (left), whereas a decision tree is optimal (right).

4.1 Decision trees

Decision trees are algorithm commonly used as multivariate classifiers. They operate in two sequential steps:

1. Given a predictor space (the set of all the possible values assumed by the variable considered), a decision tree subdivides it into distinct non overlapping regions, labelling them based on their discriminating-information contents. That is called "training".

2. A new set of observations is fed to the algorithm, which categorize them based on the region they populate. That is called "classification" or "prediction".

In the particular case of classification trees, as used in this work, any new observation is labelled as the most commonly occurring class present in the region it belongs to. A classification tree grow (that is, it increases the number of partition in the prediction space) searching recursively for the best binary selections over a set of requirements applied to a set of discriminating input variables. Given a training data set, the classifier iterates recursively on all variables, and all requirements on each, to search for the combination of binary decisions that minimize a loss function (and through that, maximise

the separation). The most basic loss function is the classification error rate,

$$E = 1 - \underset{k}{max}(\hat{p}_{mk}), \tag{4.1}$$

defined as the fraction of the training observations that does not belong to the most common class, where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

Boosted decision trees

Single large decision trees often struggle in delivering good performance when dealing with difficult-to-cluster data sets, and are prone to overfitting. Overfitting occurs when a classifier develop needlessly complex structures in order to mirror as closely as possible a particular set of training data, following the random fluctuations associated to finite sampling rather than the genuine features of the population, and is therefore prone to fail to classify future observations reliably. In the case of decision trees, this translates into too many partitions in the prediction space. Many strategies have been developed over the years to address this issue including use of random forest, bagging, and boosting. In our case a boosting approach is adopted. This implies the construction of several low depth trees called weak-learners (depth is the maximum number of subsequent cuts in the tree). All trees are grown sequentially, using information from previously grown trees: given the current model, we fit a decision tree to the residuals from the model. In order to achieve that, a first weak-learner is grown, and a weight is attributed to his candidate. Afterwards, the weights of the misclassified candidates are increased and another weak-learner is constructed. The weight of the new candidate is the gradient descent,

$$w_i = -\frac{\partial L}{\partial \sigma(\vec{x}_i)}. \tag{4.2}$$

The final output is obtained through a weighted sum over the candidates of each weak-learner. This way, the algorithm singles out the candidates that are hard to classify and increases their weight. To further increase the ability of the boosted decision tree to generalize, only a random subset of the candidates is used for the construction of the individual trees, hence the name stochastic gradient boosted decision trees.

4.2 Choice of discriminating variables

The decision tree learns how to classify a data set exploring the discriminating variables of the decay of interest previously "learned" form a data set of known classification, usually derived from simulation. For an optimal classification, the discriminating variables have to be chosen according to two principal criteria. Firstly, they need to offer discriminating power, that is the distributions of signal events have to differ as much as possible from the one of the background events. Secondly, the distributions of the variables in simulation have to reproduce as closely as possible those from data observed in the detector. Otherwise the results of the classification will be sub-optimal, as the discriminant threshold calculated for each variable will not mirror the reality outcomes of data, leading to a considerable fraction of false positive and negative classifications.

I compare spring 2019 data and simulation to identify possible options for discriminating variables. I explore especially variables constructed to exploit the differences in shape and energy-momentum flow between the generic $q\overline{q}$ production which make up the continuum background and the signal of interest. These variables have been developed in the past B-factory experiments to leverage on a fundamental kinematic difference between signal and background from e^+e^- collision at the $\Upsilon(4S)$ threshold. The energy of the collision is 10 GeV, much higher that the masses of the pairs of light quarks populating the non resonance background, which therefore are produced back-to-back in the collision with high momenta. $B\overline{B}$ pairs are instead produced nearly at rest in the center-of-mass frame. Because the mass of a $B\overline{B}$ pair is nearly equal to the collision energy, the $B\overline{B}$ daughters spread isotropically in space forming spherically-shaped events (Fig. 4.2). These topological and energy-momentum flow differences are captured by a number of variables discussed in the following.

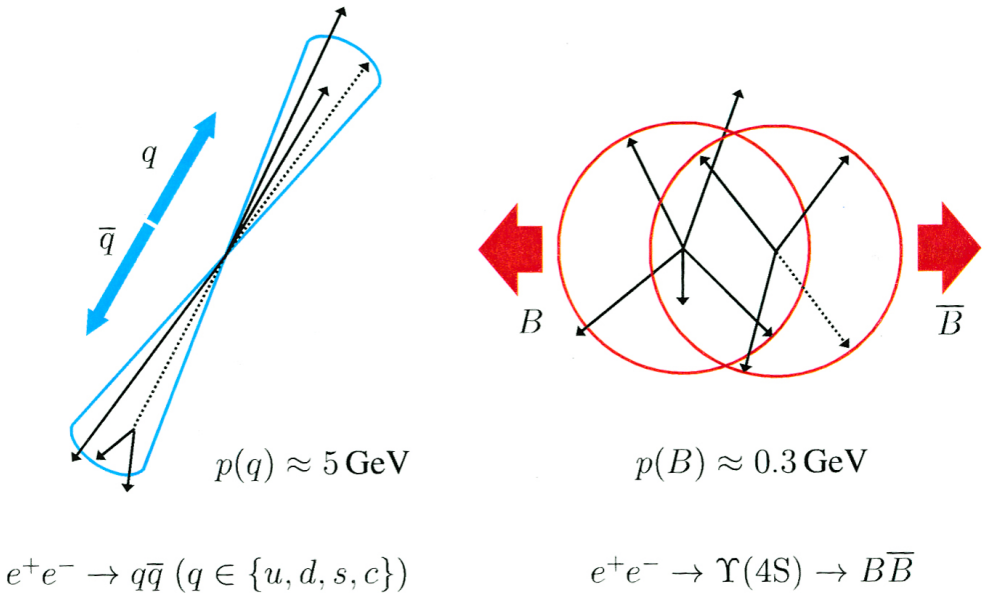

Figure 4.2: Illustration of event shapes for (left) jet-like continuum events (left) and (right) spherical $B\overline{B}$ events.

Thrust

The thrust is defined as

$$T = \frac{\sum_{i=1}^{N} |\mathbf{T} \cdot \mathbf{p}_i|}{\sum_{i=1}^{N} |\mathbf{p}_i|} \tag{4.3}$$

Where $\mathbf{T}$ is the vector that maximizes T. The value of T and of its axis are calculated both for the candidate decay particles (T_{sig} and $\mathbf{T}_{sig}$) and for the momenta of the rest of the particles reconstructed in the event (T_{tag} and $\mathbf{T}_{tag}$).

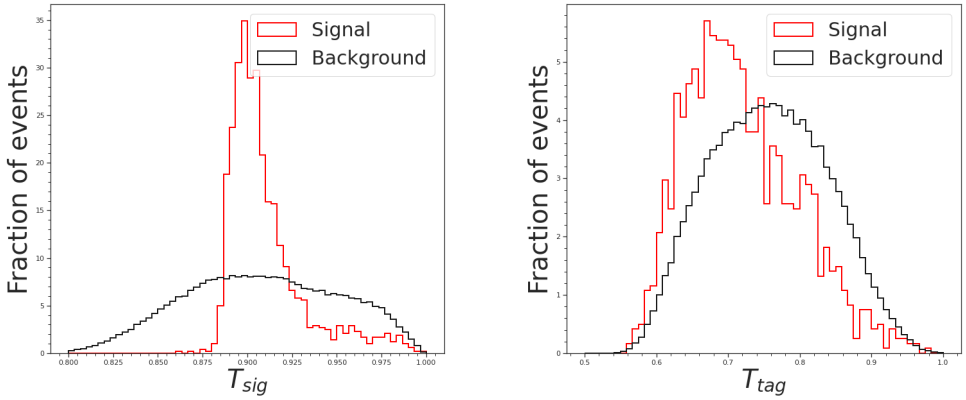

Figure 4.3: Comparison between simulated signal and background distributions of T_{sig} and T_{tag}.

Figure 3.5 shows a comparison between signal and background distribution of thrust. T_{sig} for signal shows a narrow structure while background is wider. T_{tag} display a similar but less pronounced behaviour. This support thrust as a powerful discriminating variable.

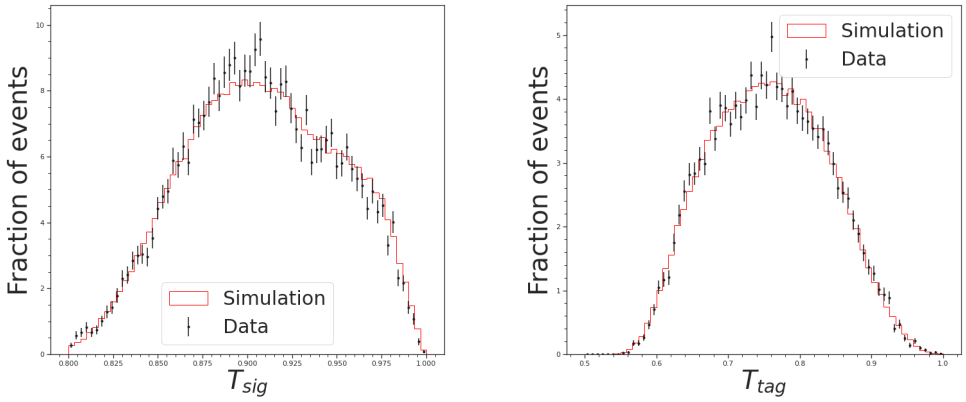

Figure 4.4: Comparison between the distributions of T_{sig} and T_{tag} in data and simulation.

The distributions of T_{sig} and T_{tag} shows no notable differences between simulation and data, which indicates that thrust is certainly a good candidate variable to be included in the BDT.

Thrust angles

Given their low momenta, the decay of particles B^0_{sig} (the candidate signal B meson) and B^0_{tag} (the candidate signal B meson) are isotropically distributed, so $\mathbf{T}_{sig}$ and $\mathbf{T}_{tag}$ are randomly distributed. Lighter background particles form instead jet-like distribution. Hence $\mathbf{T}_{sig}$ and $\mathbf{T}_{tag}$ for background are strongly directional and collimated. The

variable $|cos\theta_T^{sig,tag}|$, where $\theta_T^{sig,tag}$ is the angle between $\mathbf{T}_{sig}$ and $\mathbf{T}_{tag}$, has therefore considerable suppression power. Based on similar considerations, another useful variabile is $|cos\theta_T^{sig,beam}|$, where $\theta_T^{sig,tag}$ is the angle between $\mathbf{T}_{sig}$ and the beam axis.

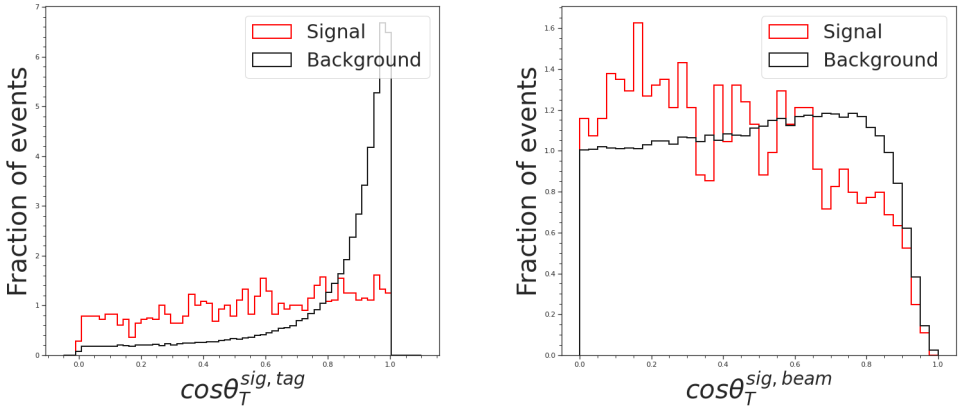

Figure 4.5: Comparison between thrust-angle distributions for simulated signal and background.

As the products of the decay of interest spread isotropically, we expect every angular configuration of the decay to be roughly equiprobable, resulting in a uniform distribution for signal in $cos\theta_T^{sig,tag}$. This is observed in the distribution of Fig. 4.5, where differences due to jet-like back-to-back decay of background are evident.

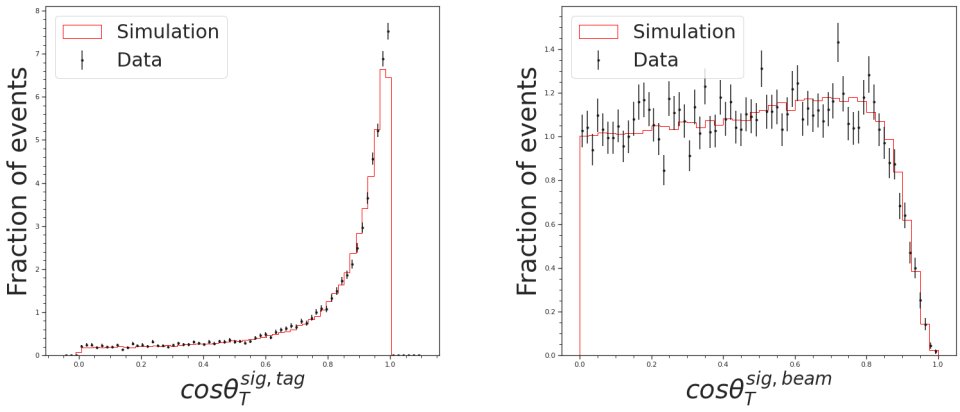

Figure 4.6: Comparison between the distributions of thrust angles in data and simulation.

The distributions of thrust angles in data (Fig. 4.6) are consistent with the distribution observed in simulations, supporting thrust angles as good candidates for BDT inputs.

CLEO cones

CLEO cones are refinements of the concept of thrust. They are based on the sum of the absolute values of the momenta of all particles within angular sectors around the thrust axis, in intervals of 10°, for a total of 9 concentric cones.

]

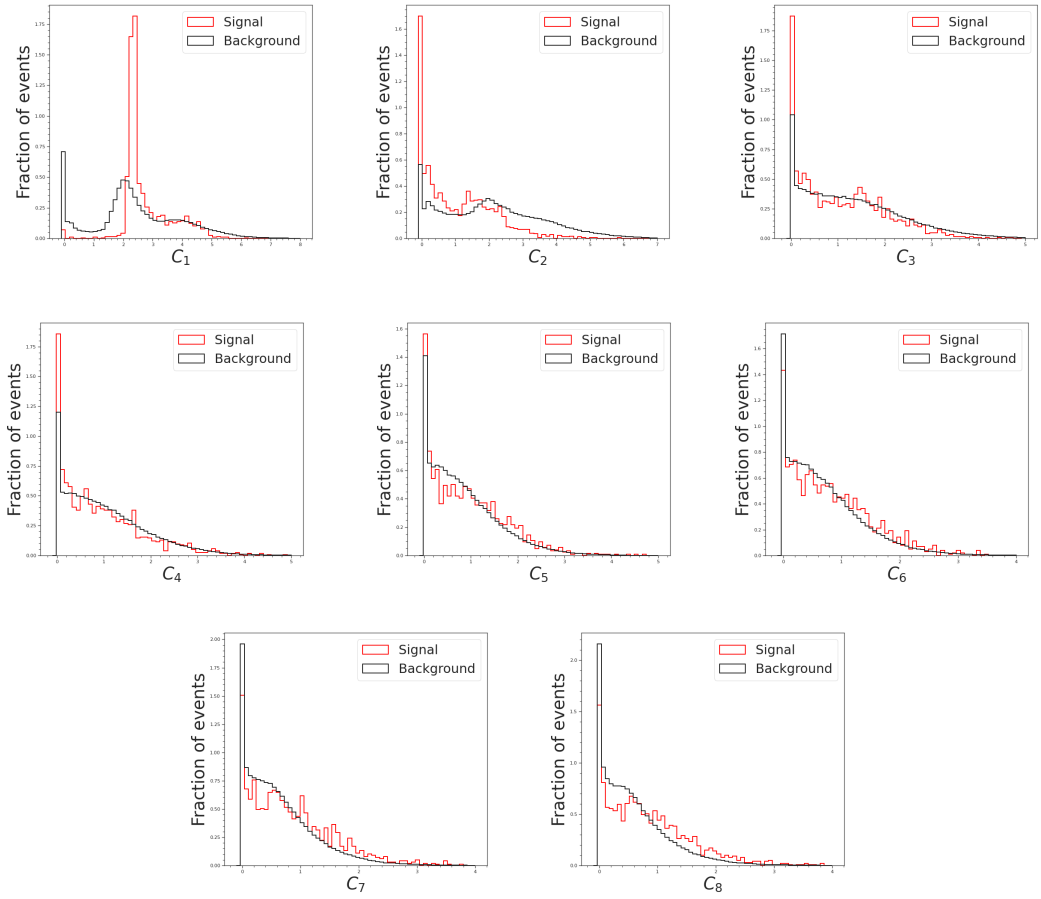

Figure 4.7: Comparison between signal and background in simulation of CLEO cones.

Fig. 4.7 shows significant differences between the distributions of signal and background events in C_1 (mainly) and C_2. Distributions of C_3 to C_8 show little to no discriminating power.

]

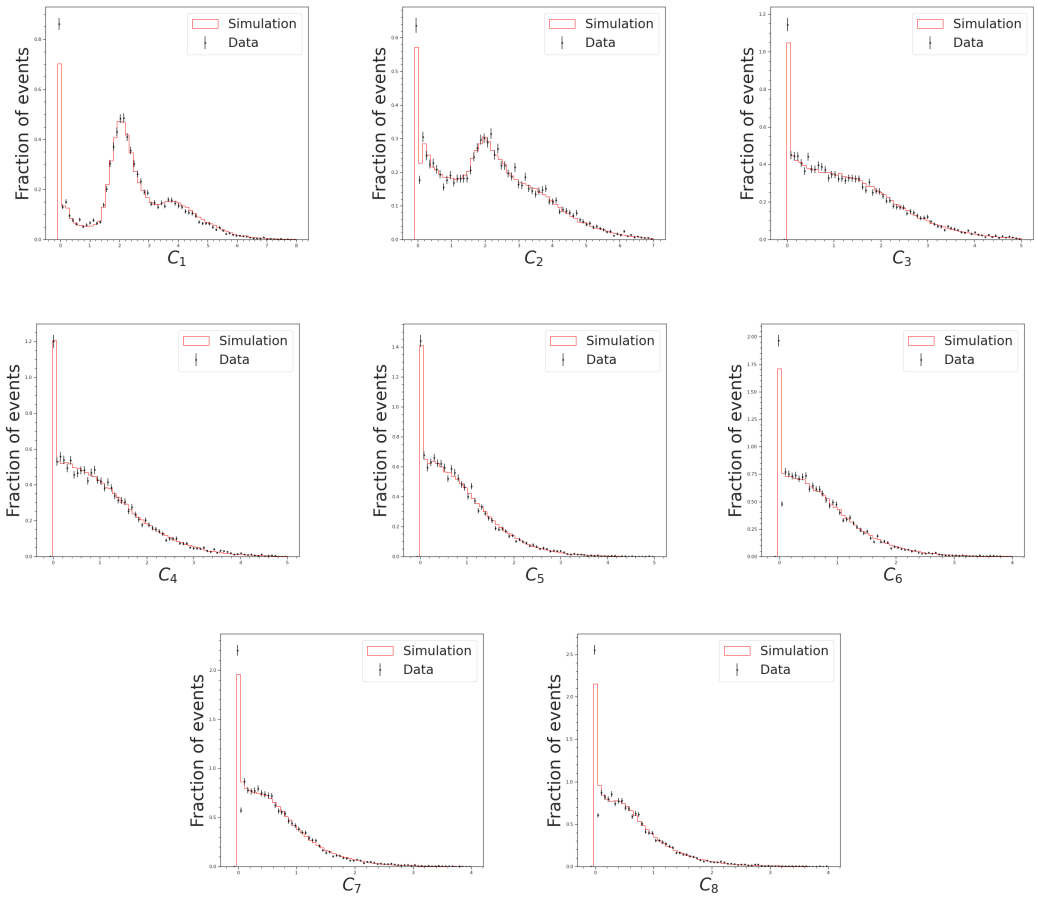

Figure 4.8: Comparison between the distributions of CLEO cones in data and simulation.

The data distributions of CLEO ones are well reproduced by the simulation (Fig. 4.8, supporting their use as good discriminating variables.

Fox-Wolfram moments

The l-th order Fox-Wolfram moment H_l is defined as

$$H_l = \sum_{i,j}^{N} \frac{|\mathbf{p}_i^*| \cdot |\mathbf{p}_j^*|}{s} \cdot P_l(cos\theta_{i,j}^*), \tag{4.4}$$

where $\theta_{i,j}^*$ is the angle between the momenta $\mathbf{p}_i^*$ and $\mathbf{p}_j^*$, $\sqrt{s}$ is the total energy in the $\Upsilon(4S)$ frame, and P_l is the l-th order Legendre polynomial.

At Belle II, the normalized ratio $R_2 = \frac{H_2}{H_0}$ is used as a discriminating variable as it capture the "shape" of the final-state particle trajectories in the detector volume.

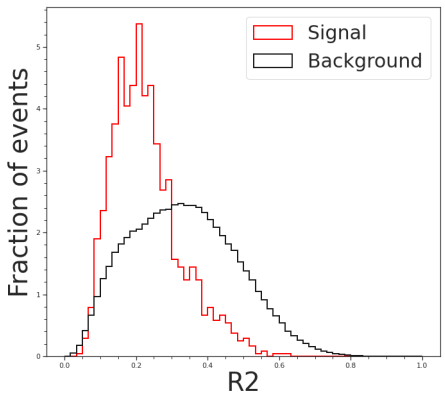

Figure 4.9: Comparison between R_2 distributions for simulated signal and background.

Dependence on the angle of emission between daughters particles through the Legendre polynomial, makes the distribution of R_2 distinctive between signal and background (Fig. 4.9). The small angles of the jet-like background translate into a flatter, wider distribution, while signal shows a narrow distribution with a maximum at a value of R_2 of about 0.2, showing the expected significant discriminating power of this variable.

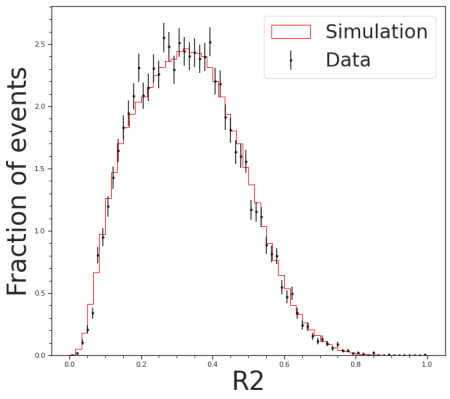

Figure 4.10: Comparison between R_2 distributions of data and simulation.

No clear differences between data and simulation distributions for R_2 is seen, supporting the usage of R_2 as input in our BDT.

Kakuno-Super-Fox-Wolfram moments

The discriminating power provided by the Fox-Wolfram moments is dependent by the successful detection of all the particles produced in an event. Kakuno-Super-Fox-Wolfram moments H_{xl}^{so} and H_{l}^{oo} were developed in an attempt to account for particles that could

go undetected due to acceptance of inefficiencies. The superscript is used to label the recostructed particles: "s" denotes the B^0_{sig} candidate daughters, and "o" denotes the tag-side particles. The subscript is instead used to sort whether the particle is charged (c), neutral (n), or missing (m).

For l even,

$$H^{so}_{xl} = \sum_i \sum_{j_x} |\mathbf{p}^*_{j_x}| \cdot P_l(cos\theta^*_{i,j_x}), \tag{4.5}$$

where i extends over primary B^0_{sig} daughters, and j_x over all the tag-side particles. For odd l $H^{so}_{nl} = H^{so}_{ml} = 0$,

$$H^{so}_{cl} = \sum_i \sum_{j_x} q_i \cdot q_{jx} \cdot |\mathbf{p}^*_{j_x}| \cdot P_l(cos\theta^*_{i,j_x}), \tag{4.6}$$

where q_i and q_{jx} are the charges of the particles i and j_x, and

$$H^{oo}_l = \begin{cases} \sum_j \sum_k |\mathbf{p}^*_j| \cdot |\mathbf{p}^*_k| \cdot P_l(cos\theta^*_{j,k}), & \text{if } l \text{ is even} \\ \sum_j \sum_k q_i \cdot q_{jx} \cdot |\mathbf{p}^*_{j_x}| \cdot |\mathbf{p}^*_k| \cdot P_l(cos\theta^*_{i,j_x}), & \text{if } l \text{ is odd} \end{cases} \tag{4.7}$$

where j and k extend over all the tad side particles.

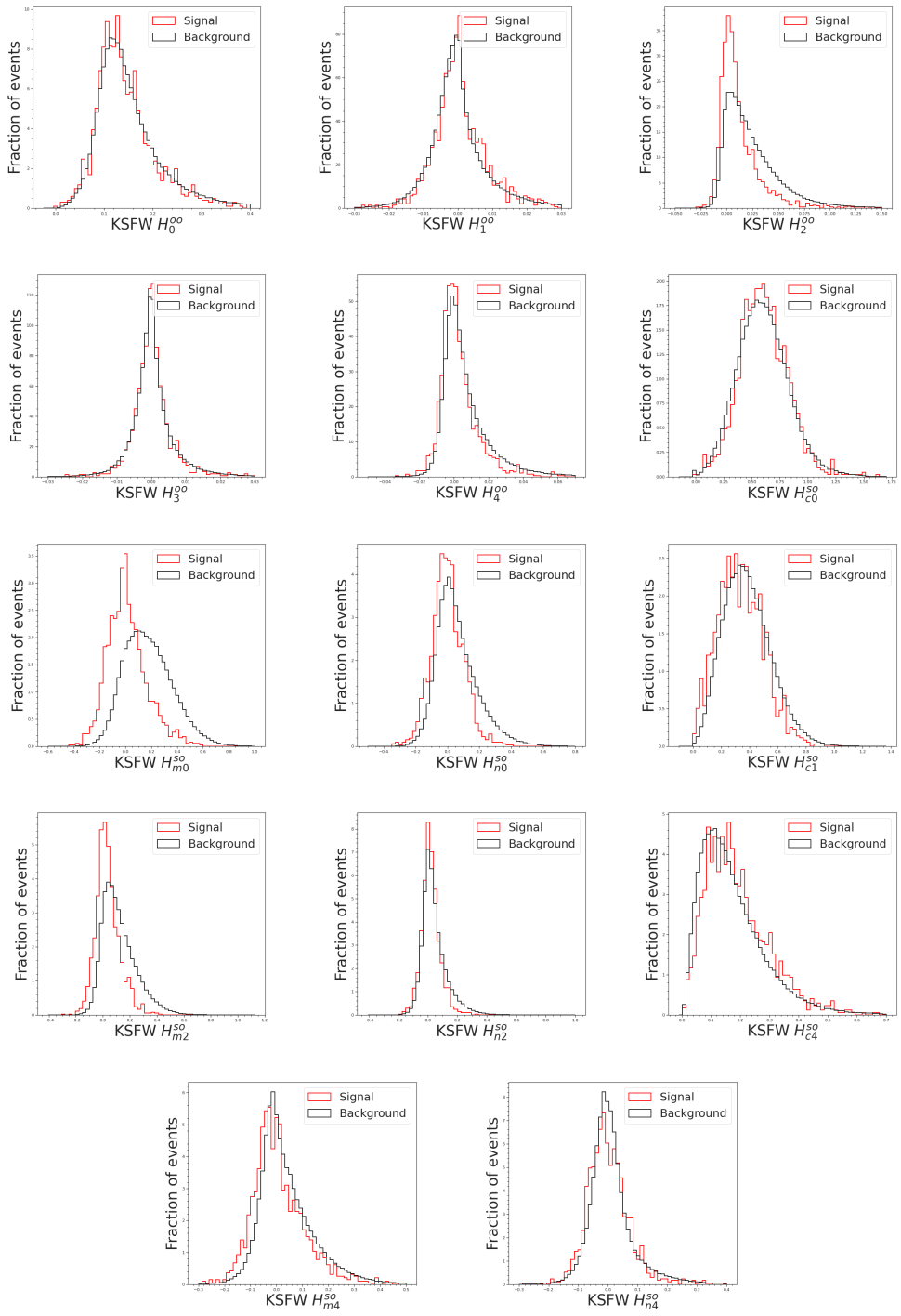

Figure 4.11: Comparison between Kakuno-Super-Fox-Wolfram moments distributions for simulated signal and background.

Most Kakuno-Super-Fox-Wolfram moments show limited difference between signal and background. This result in little discriminating power. Exceptions are the distribu-

tions of H_2^{oo}, H_{m0}^{so}, and H_{m2}^{so}, in which the signal distributions is significantly narrower than the one of the background.

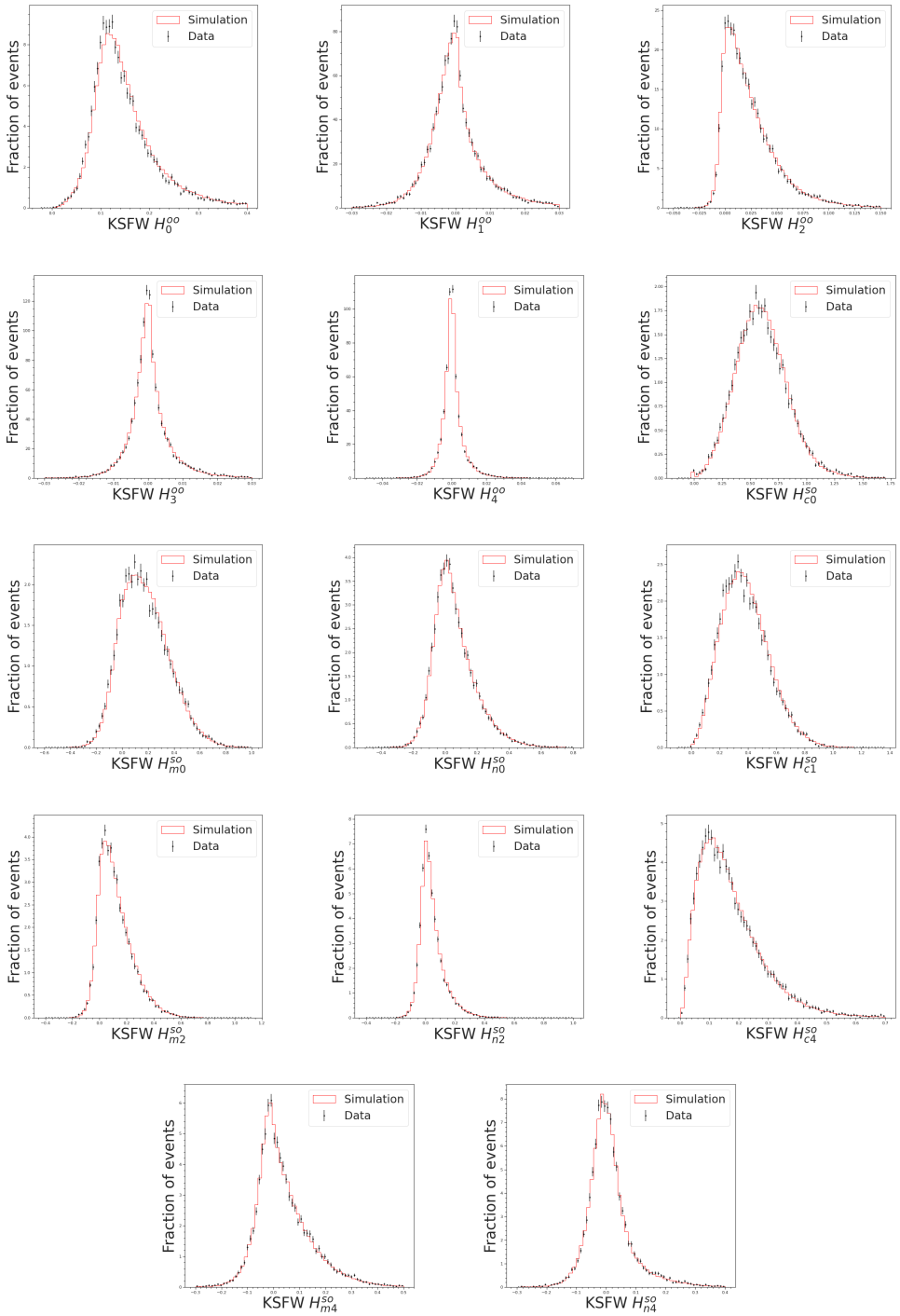

Figure 4.12: Comparison between the distributions of Kakuno-Super-Fox-Wolfram moments in data and simulation.

The distributions in data are compatible with those in simulation, supporting the use of these variables as discriminators.

4.3 Discriminating variable pruning

When it comes to decide which variables should be used in the training of a decision tree, the more the better is a good rule of thumb. If a variable as little to no discriminating power, it will be simply weighted less and less during the building of new branches, or even eventually be discarded. However, if limitations due to constrains associated with training time, or to the computational power of the machine used arise, data cleaning beforehand is useful. If the only reference used to choose the set of input variables are the one-dimensional distributions, it is not always straightforward to understand which ones are the most effective, as frequently the discriminating power is shared among many variables and hidden in their correlations. A common and straightforward way to perform a screening is using a scatterplot matrix. Given a set of discriminant variables and a training data sample where every event is already tagged as signal or background, a scatterplot matrix is a matrix in which every entry is a scatter plot of any pair of discriminating variables. This makes evident which variables (or which combination of them) lead to a clearer separation between the categories of interest. An example of such matrix is reported in Fig. 4.13, where only five variables where considered.

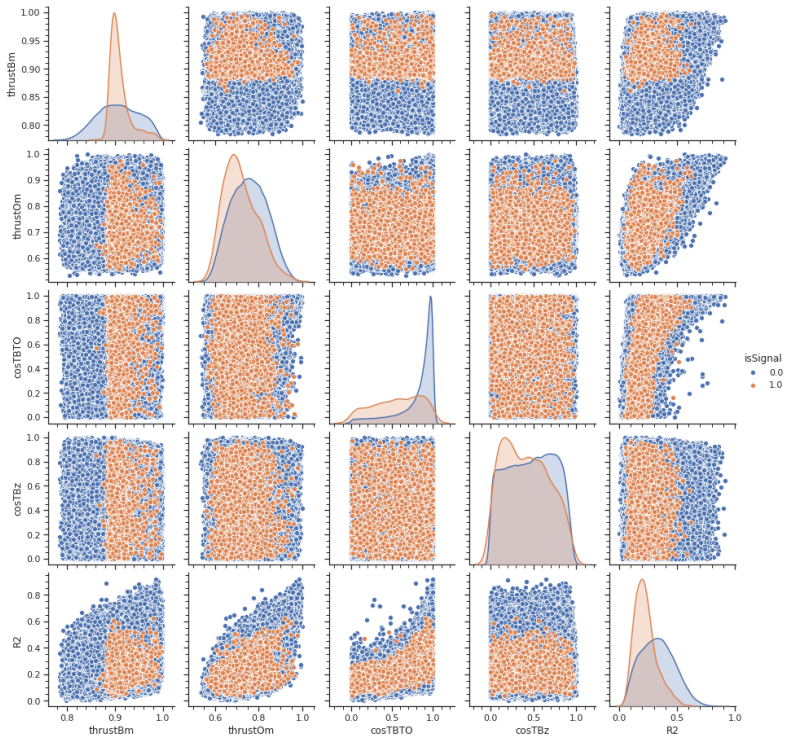

Figure 4.13: Scatterplot matrix of T_{sig}, T_{tag}, $cos\theta_T^{sig,tag}$, $cos\theta_T^{sig,beam}$, R2.

The general case in which all variables were considered is not reported due to his size.

Chapter 5

Results

Starting from first Belle II data recorded in early 2019, the ideal goal is to obtain the best possible separation between signal and background. The starting point is the distribution of the difference between expected and observed B candidate energy (ΔE) after the baseline selection discussed in sec. 3.2. The signal peak is recognizable overlapping a smooth background. An ideal output of the binary classification should resemble the plot of the distribution of energies of signal simulation. Fig. 5.1 shows the starting point of the distribution for data (left), and the ideal distribution after the classification (right), represented by the distribution of signal events from simulation.

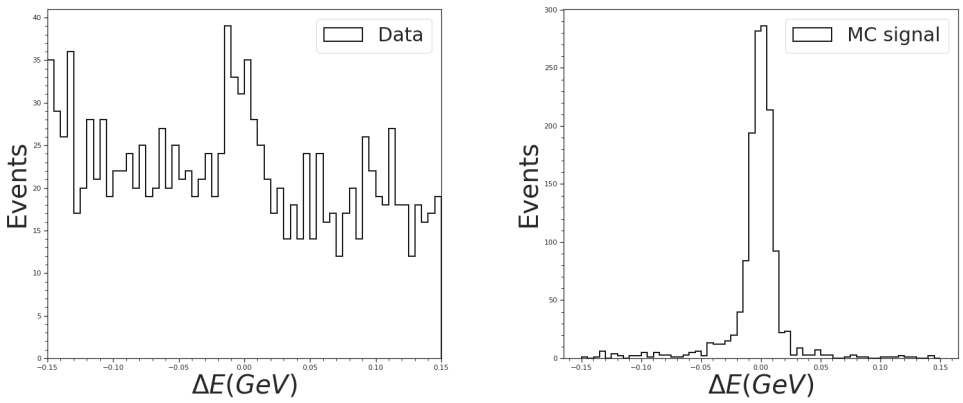

Figure 5.1: Comparison between the distributions of difference between expected and observed B candidate energy (ΔE) in (left) data and (right) simulation.

5.1 Classification performance

To evaluate the performance of a binary classifier, we use the following nomenclature:

		Predicted class		
		Null	Non-Null	Total
True class	Null	True negative (TN)	False positive (FP)	N
	Non-Null	False Negative (FN)	True positive (TP)	P
	Total	N*	P*	

Where null and non-null are the two classes of events that have to be divided. In the case studied, null is background, and non-null is signal. A number of commonly used metrics exists that can help to give a quantifiable evaluation of the classifier performance:

- False positive rate (FP/N).
- True positive rate (TP/P).
- Positive prediction value (TP/P*).
- Negative prediction valure (TN/N*).

The ROC curve (receiver operating characteristic curve) displays the relation between the background efficiency (the false positive rate), and the signal efficiency (1 - true positive rate). The overall performance of a classifier is given by the area under the (ROC) curve (AUC), as it represents the probability that the classifier will label correctly a random event (see Fig. 5.2). An ideal ROC curve approaches the top right corner, so the larger the AUC, the better the classifier.

ROC curves on different trainings

It is interesting to compare ROC curves obtained training the boosted decision trees with different choices of input variables. It is common practice in the use of supervised learning algorithms to test the final results of the training on a data set of known classification different from the one used for training. This tests the versatility of the statistical learning algorithm on new data, and checks if outfitting occurred. I splitted the simulation in a larger batch dedicated to the training itself, and a smaller test batch. I trained the classifier algorithm on the same data set using different collections of discriminating variables, and both the results of the BDT on the training and test data are reported. Fig.5.2 shows the ROC curves of a training in which all the discriminating variables are used. Fig. 5.3 shows the ROC curves for of a training in which only the variables with the highest discriminating power was used and CLEO cones from C_3 to C_8, and all the Kakuno-Super-Fox-Wolfram moments but H_2^{oo}, H_{m0}^{so}, H_{n0}^{so}, H_{m2}^{so}, H_{m4}^{so} are discarded.

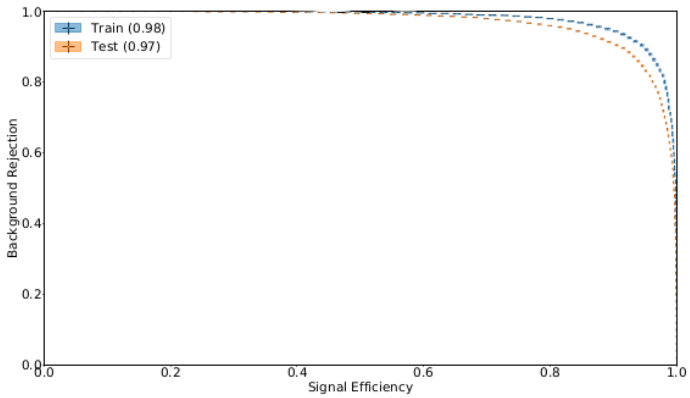

Figure 5.2: ROC curve using 27 variables. The upper curve indicates performance on the training sample, the lower curve on the test sample.

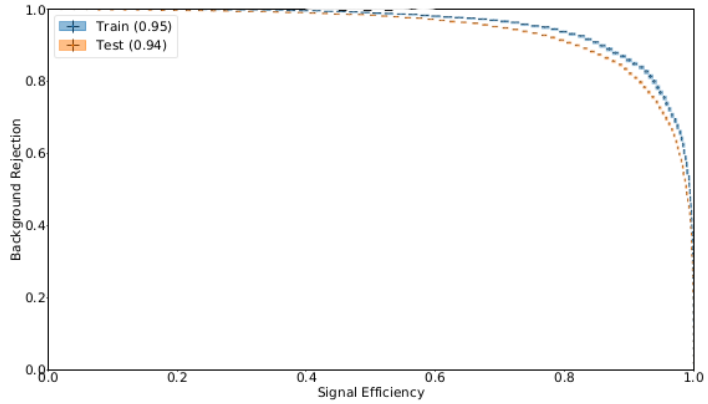

Figure 5.3: ROC curve using 11 variables.The upper curve indicates performance on the training sample, the lower curve on the test sample.

As expected, the use of an higher number of variables led to better results as supported by the ROC curves. However, the second training used only about a third of the variables, the ones with the higher discriminating power. The optimal tradeoff should be evaluated depending on the computation power available. Also, in both plot, no major discrepancy is displayed between ROC curves of training and test data set, suggesting the negligible overfitting.

5.2 Output study

Feeding a data set to the classifier, it gives back as an output a probability value raging from 0 to 1 for every event. This represents the classification operated by the decision

tree. If an event is associated with the number 1 (0), it means that it was tagged as signal (background) (Fig. 5.4).

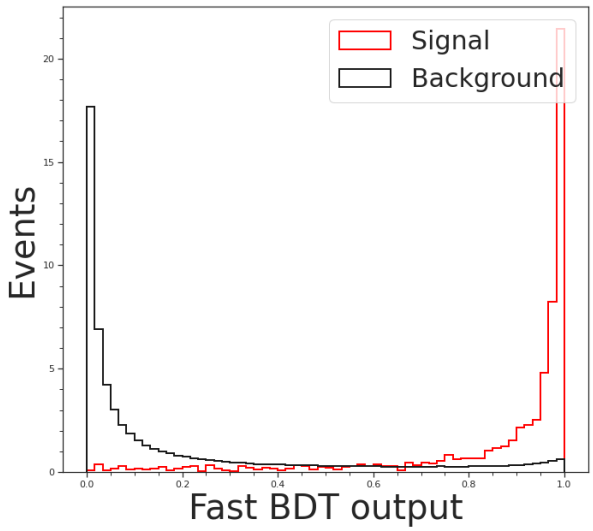

Figure 5.4: Distribution of the FBDT output variable for simulated signal and background events of known classification.

As an event final output value get further away from one of the two extremes of the interval, its classification becomes less and less clear, and cases of false positive (or false negative) increase.

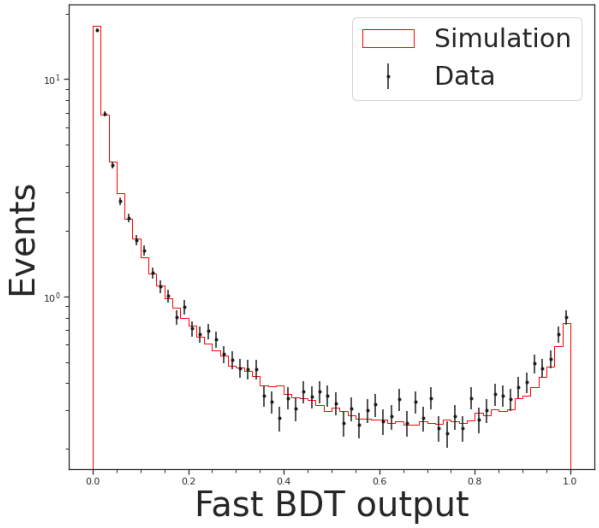

Figure 5.5: Comparison between the distributions of the FBDT output variable of data and simulation.

Fig. 5.5 shows no clear difference between the distribution of the BDT output for data and simulation. This support the quality of our classifier which is an accurate model of the features in data.

5.3 Selection optimization

Given the results of the classification, one has to decide the threshold on which an event would be considered signal. The optimal decision should minimize the false positive (as the purity of the signal is the main goal) without discarding a too large portion of the data set. T maximize the figure of merit (FOM) $\frac{N_S}{\sqrt{N_S+N_B}}$, in the region of energy difference $-0.15 < \Delta E < 0.15$. Fig. 5.6 shows the value of FOM as a function of the threshold imposed on the BDT output. The final criterion is $FBDT > 0.90$, which retains 84% of the signal, while rejecting 90% of the background.

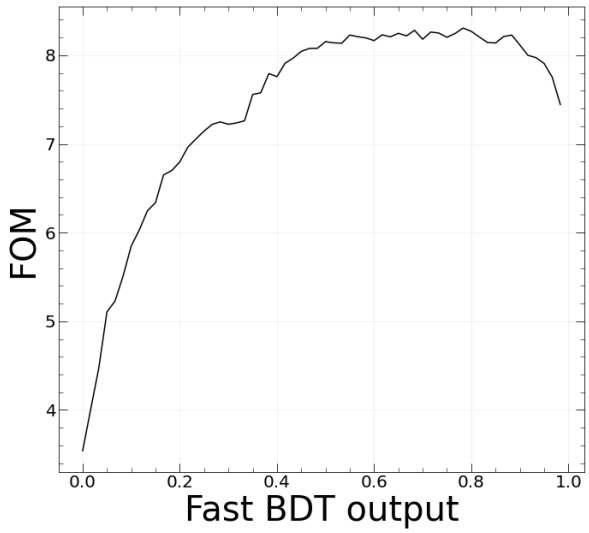

Figure 5.6: Value of FOM as a function of the threshold imposed on the fast BDT output.

5.4 Classification results

Having classified the events recorded in the first Belle II data, I display the results in Fig.5.7.

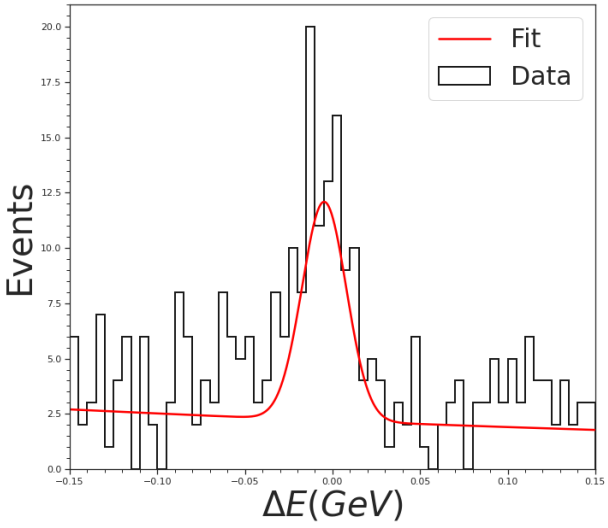

Figure 5.7: Distribution of difference between expected and observed B candidate energy (ΔE) in data after BDT classification.

A fit of this distribution determines 63 ± 7 signal events with a 2.5 signal-to-background fraction at the peak. This corresponds to an improvement of a factor of 4 in signal purity at the modest price of 16% of signal reduction. These numbers encapsulate the power of the classifier developed in this thesis.

Chapter 6

Summary

Completing the Standard Model is the principal objective of particle-physics today. Belle II is an experiment located in Tsukuba, Japan, designed, built, and operated by over 1000 physicists, with the main purpose of exploiting high-intensity 10 GeV electron-positron collision produced by the superKEKB accelerator to explore extensions of the Standard Model. In this experimental particle-physics thesis I developed and implemented a supervised-learning classifier for the selection of the decay channel $B^0 \to D^-[\to K^+\pi^-\pi^-]\pi^+$ in the data first collected by the Belle II experiment. In the first part I explored and selected the discriminating variables useful for the identification of the decay channel of interest using a simulation. Then I applied the chosen classification method, a boosted decision-tree, to the first data set collected by Belle II, in spring 2019, and assess performance. The classifier developed in this work improves the signal-to-background ratio by a factor of 4 with a signal inefficiency of 16 %, offering promising chances of impact when the Belle II data set will be sufficient to allow world-leading results.

Bibliography

[1] F. Abudinén, E. Ganiev, R. Manfredi, S. Raiz, and D. Tonelli. *Charmless B decay reconstruction in 2019 data.* (Belle II Collaboration), 2020. BELLE2-NOTE-PH-2020-007.

[2] Belle II Collaboration and B2TiP theory Community. *The Belle II Physics Book.* 2018. Prog. Theor. Exp. Phys.

[3] F. Abudinén et al. *Charmless B decay reconstruction and first measurements in 2019 and 2020 data.* (Belle II Collaboration), 2020. BELLE2-NOTE-PH-2020-041.

[4] F. Abudinén et al. *Charmless B decay reconstruction in 2019 Belle II data.* (Belle II Collaboration), 2020. BELLE2-CONF-PH-2020-001.

[5] Geoffrey C. Fox and Stephen Wolfram. Observables for the analysis of event shapes in e^+e^- annihilation and other processes. *Phys. Rev. Lett.*, 41:1581–1585, Dec 1978.

[6] E. Ganiev, J. Libby, N. Rout, D. Tonelli, and K. Trabelsi. *Hadronic B decay reconstruction in Early Phase III data.* BELLE2-NOTE-PH-2019-039. 2019.

[7] Sheldon L. Glashow. Partial-symmetries of weak interactions. *Nuclear Physics*, 22, 1961.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer Texts in Statistics. Springer Verlag, 2013.

[10] Thomas Keck. *The Full Event Interpretation for Belle II.* BELLE2-THESIS-2015-001. 2014.

[11] Sebastiano Raiz. *First charmless B decay reconstruction in Belle II data.* 2019.

[12] A. Salam. *Weak and Electromagnetic Interactions.* 1968. Conf. Proc. C 680519.