University of Heidelberg

Department of Physics and Astronomy

Institute for Theoretical Physics

Bachelor Thesis in Physics

Investigation of invisible Higgs decays through Weak Boson Fusion

submitted by

Fabian Sascha Keilbach

born in Heidelberg (Germany)

This Bachelor Thesis has been carried out by Fabian Keilbach

at the Institute for Theoretical Physics Heidelberg

under the supervision of

Prof. Tilman Plehn

2017

Abstract

Invisible decays of a Higgs boson created through Weak Boson Fusion and corresponding background processes are simulated. The power of Boosted Decision Trees in separating these events is investigated as well as the suitability of quark gluon discrimination as input variables. We see that Boosted Decision Trees perform very well while also being stable to variations in the tree parameters. Quark gluon variables are found to be well suited in separating signal from background and my be valuable as a substitute for variables that are harder to obtain.

Zusammenfassung

Unsichtbare Zerfälle eines durch schwache Boson Fusion produzierten Higgs Teilchens und zugehörige Hintergrundprozese werden simuliert. Wir untersuchen die Eignung von Boosted Decision Trees und Quark Gluon Variablen bezüglich der Separation von Signal und Hintergrund Ereignissen. Es zeigt sich das Boosted Decision Trees dafür gut geeignet sind und gleichzeitig stabil bezüglich Variation der Parameter sind. Quark Gluon Variablen stellen sich ebenfalls als geeignet heraus und können insbesondere als Ersatz für schwieriger zu messende Variablen hilfreich sein.

Contents

1 Introduction

Particles and their fundamental interactions are currently best described by the *Standard model of particle physics*. It consits of an electroweak sector which is a unified description of quantum electrodynamics (QED) and the theory of weak interactions, a quantum chromodynamics (QCD) sector descriping the strong interaction and the Higgs sector. With the discovery of the Higgs Boson in 2012 [1] all particles predicted by the standard model have been found and so far no prediction derived from the standard model undoubtedly disagrees with experimental results [2].
Despite the huge success of the SM many open questions remain. On the one hand theoretical principles strongly suggest a more fundamental theory which may include extra dimensions or supersymmetry [3]. On the other hand various physical processes are known which are not described by the standard model in the first place: Most notably gravity but also the existence of neutrino masses or the nature of dark matter.
Some extensions of the standard model allow invisible Higgs decays (that is, the final state products can not be detected) in addition to the invisible Higgs decay already incorporated in the standard model $H \to ZZ \to 4\nu$ which in this thesis will be neglected due to its small branching fraction compared to the reach of current detectors. Such invisible decays may open a link from the standard model to dark matter through a Higgs portal [4]. Constraining the branching fraction of such decays is therefore of great importance.

Higgs production occurs through several channels of which Weak Boson Fusion (WBF) is one of the most promising for invisible decays [5]. As a standard technique kinematic cuts are applied to distinguish signal from background but in recent times methods using multivariate analysis have been introduced in high energy physics with remarkable success.
In this thesis new ideas and methods of probing invisible Higgs decays are reviewed and tested with an emphasis set on boosted decision trees (BDT). The structure is as follows: In chapter 2 the properties of Weak Boson Fusion are briefly reviewed and the process of event generation is shortly described. Chapter 3 gives an introduction to boosted decision trees. The ROOT toolkit TMVA is used to test and evaluate various properties of BDT's. Chapter 4 describes the notion of quark-gluon discrimination and the applicability of some quark gluon discriminating variables is checked. Chapter 5 gives a summary and some ideas for further analysis.

2 Weak Boson Fusion

2.1 Basic collider observables

In order to evaluate collider experiments one must be able to identify particles and measure their properties and kinematics. Some of the most used observables are listed below:

- **particle 4-momentum** $p = (E, \vec{p}) = (E, p_x, p_y, p_z)$
- **transverse momentum** p_T
  If the z-axis is defined as pointing along the beam line the transverse momentum is perpendicular to it: $p_T = \sqrt{p_x^2 + p_y^2}$
- $\eta - \phi$ **plane**
  Since measurements can not be performed along the line of the beam axis the interesting spatial information can be described by two angles. Conventionally this is done by a polar angle ϕ and an azimuthal angle θ. Instead of θ the rapidity $y = \frac{1}{2}\ln(\frac{E+p_z}{E-p_z})$ is often used. In the massless limit, which is an excellent approximation for the high energies we consider, it becomes the *pseudorapidity* $\eta = \lim_{m\to 0} y = -\ln(\tan\frac{\theta}{2})$
- **missing transverse momentum** $\not{p}_T$ (missing energy)
  From the definition of the z-axis it is clear that the total incoming transverse momentum is zero. Momentum conservation then tells us that this must also be true for the final state. Therefore a nonzero total transverse momentum indicates incomplete measurement which may arise from invisible particles like the standard model neutrinos or pure detector effects like momentum mismeasurement or energy deposited outside calorimeters
- **invariant mass** $m_{1,2}$
  $m_{1,2}^2 = (p_1 + p_2)^2$ which is a Lorentz invariant quantity of two particle

2.2 Weak Boson Fusion

Higgs production through Weak Boson Fusion is a process in which two incoming quarks each radiate a vector boson (W or Z) which then fuse to a Higgs. Since we require the Higgs to decay into invisible particles we expect, at lowest order, two final state jets and large missing transverse energy (MET). There are several backgrounds that exhibit the

same behavior:

- Electroweak W or Z production: qq $\rightarrow$ qqVV $\rightarrow$ qqV (V = W or Z)
- QCD W or Z production : qq $\rightarrow$ qq gg $\rightarrow$ ggV / gg $\rightarrow q\bar{q}\ q\bar{q} \rightarrow q\bar{q}V$
- any QCD multijet event with large missing transverse energy

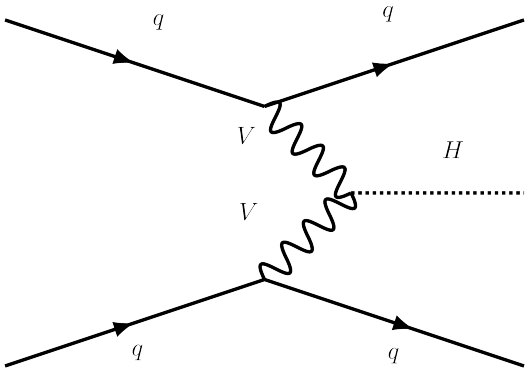
The electroweak or QCD produced vector bosons mimic the signal if they subsequently decay invisibly, that is Z $\rightarrow \nu\nu$ or W $\rightarrow l\nu$ where ν is a neutrino and l any misidentified lepton. Example Feynman diagrams of these processes are shown in Figure 2.1 [18] . Missing energy from general QCD multijet events can arise from mismeasured or undetected particles.

The standard approach to separate signal from background events is to use a cut flow. One applies a number of selection cuts on the whole sample and vetoes events that fail a cut. The idea is to make use of the kinematic properties of the signal to filter out more and more background, thus increasing the signal to background ratio. A possible cut flow is briefly described below, it follows the suggestions made in [5].

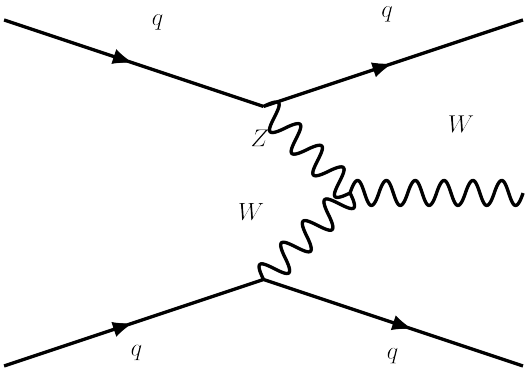
- The main properties of WBF are two jets with substantial missing energy. Therefore we must ensure to tag such events.

$$\begin{gathered}\text{Number of Jets} >= 2 \\ \text{Number of Leptons} = 0 \\ p_{T,j} > 40 \text{ GeV} \\ \not{p}_T > 100 \text{ GeV} \\ |\eta_j| > 5,\ |\Delta\eta_j| > 4.4,\ \eta_1 \cdot \eta_2 < 0 \\ m_{1,2} > 1200 \text{ GeV} \\ |\Delta\phi_{1,2}| < 1 \\ (\text{j} = 1, 2)\end{gathered}$$

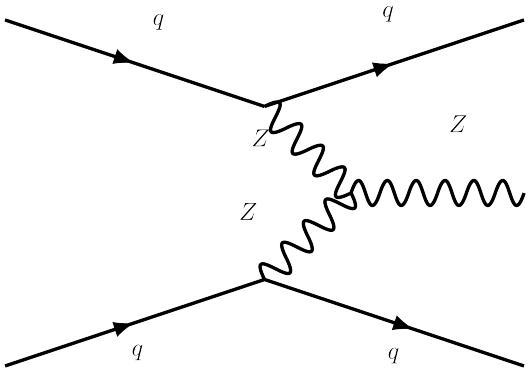
  The first two conditions make sure we have at least two jets with no identified leptons as we want to consider invisible decays. A minimal transverse momentum requirement for the two tagging jets is necessary to make sure they can be detected. In addition to that they also reduce backgrounds, especially those arising from



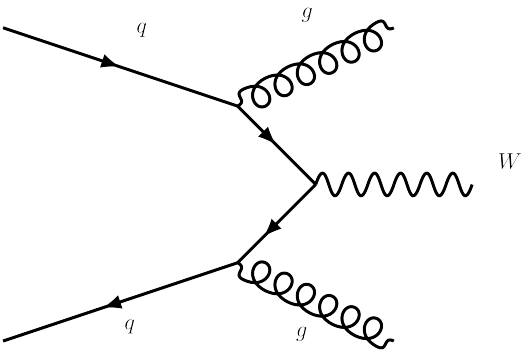
(a) WBF signal.



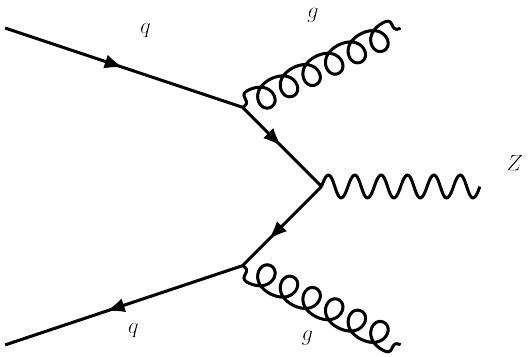
(b) Electroweak W production.



(c) Electroweak Z production.

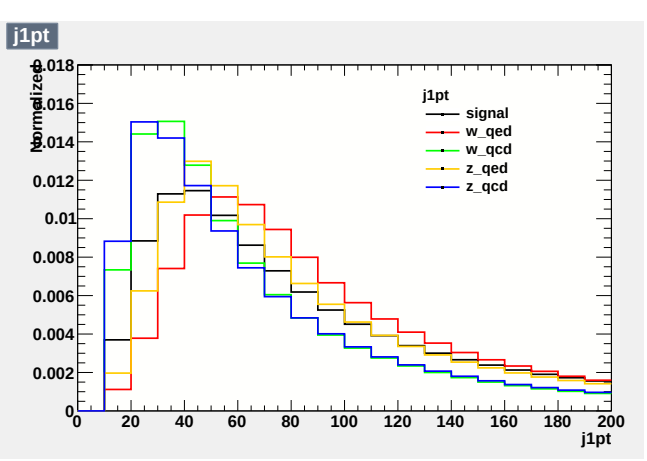


(d) QCD W production.

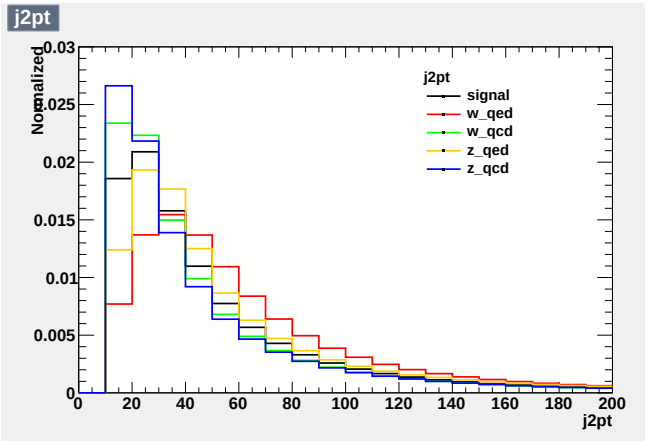


(e) QCD Z production.

Figure 2.1: Example Feynman diagrams for signal and backgrounds.



(a) p_T distribution of the hardest jets.



(b) p_T distribution of the second hardest jets.

Figure 2.2: p_T distribution of signal and background. Cutting at 40 GeV increases the signal to background ratio.

QCD processes as can be seen in Figure 2.2: There the p_T distribution of signal and background is plotted for the hardest and second hardest jet. Since measurements can not be performed close to the beam axis a minimal pseudorapidity is required. Finally the signal is expected to have two well separated jets going into opposite hemispheres.

- Finally a central jet veto (CJV) may be applied: Any event with one or more jets that have a pseudorapidity between those of the tagging jets and at least a transverse momentum of 20 GeV are vetoed.
  This is especially useful to veto QCD background events which are expected to display a different color structure than the signal.

2.3 Event generation

Events were generated using SHERPA [13] which is a Monte Carlo event generator. It calculates cross sections from matrix elements, adds soft radiation, applies parton showering and hadronization and simulates decays to stable particles. In order to avoid divergences some cuts must be applied upon event generation. They are basically less strict versions of some of the cuts described above and therefore do not interfere with the actual cut flow approach.

Since real data will be influenced by detector effects like smearing, false reconstruction or mismeasurements these effects must be taken into account before evaluating simulated data. This may be done using a detector simulation. In this thesis DELPHES [14] is used. DELPHES is a fast detector simulation including a tracking system, calorimeters and a muon system. It simulates effects arising from magnetic fields, calorimeter granularity or limited detector resolution. It does however not account for such things as dead clusters. The output generated by DELPHES includes all relevant observables like reconstructed particles, transverse momentum, missing energy, etc. It may subsequently be used as input data for boosted decision trees.

Boosted decision trees were generated and evaluated using the TMVA package of the ROOT Framework ([15], [16]). TMVA offers a variety of machine learning algorithms including several implementations for boosted decision trees.

3 Boosted Decision Trees

(Boosted) Decision Trees are multivariate classifiers that aim to seperate signal from background events from a given sample by extending simple cutflow approaches. They (generally) do not rule out events that fail a single criterion but analyze them further. The simplest decision trees are rooted binary trees: A sample containing weighted events that is described by a set of variables is split into two subsamples: A training sample which the tree has full information of (i.e whether or not a given event is signal or background) and a test sample which will be used to evaluate the performance of the tree. The input data sample makes up the root. Then a cut on one of the variables is applied and the root is split into two branches: One contains all events that pass the cut criterion and the other one those which do not. These nodes are further split until some stopping condition is fulfilled and the node is turned into a leaf. Depending on the events that land in them each leaf is classified as "signal-like" or "background-like". After the tree has been grown by this algorithm the test sample is passed through it. Every event moves down the tree, following the cuts at each branch, until it reaches a leaf. There it is categorized as signal or background event, depending on the value assigned to the leaf. A generic representation of this algorithm is shown in Figure 3.1.

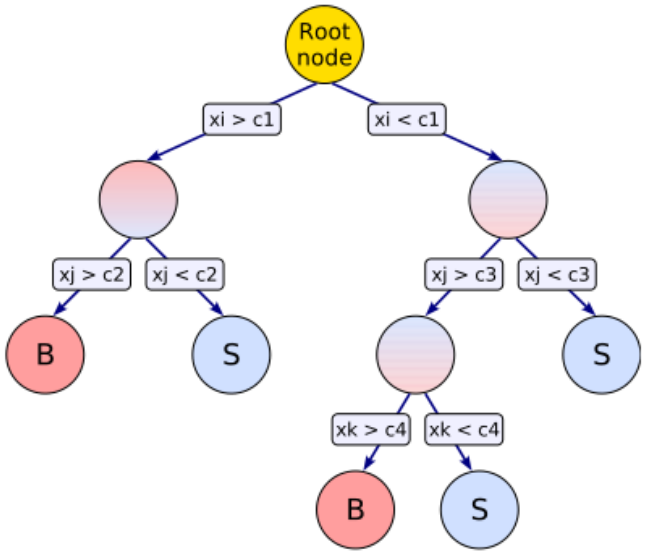


Figure 3.1: Generic decision tree. Each node is split into two, controlled by the cut variables x_i, x_j, x_k and the corresponding cut values c_1, c_2, c_3, c_4. Leafs are either classified as Signal (S) or Background (B). Figure taken from [6], p.109.

Before applying this algorithm to data several steps have to be specified:

- How shall the optimal split at each node be determined? (i.e. what measure to use)
- What is the output of the decision tree? When is a leaf classified as signal-like or background-like?
- What are the conditions to stop further splitting of the decision tree?

In the next subsections these questions are answered thereby illustrating the growing of a decision tree. This is followed by a description of boosted decision trees and their advantages and disadvantages as compared to other classifying methods.

3.1 Decision Tree algorithm

Since the goal of the decision tree is to separate signal from background events one needs a measure stating how well this separation is achieved. A common choice is the purity p

$$p = \frac{s}{s+b}$$

where s (b) is the sum of weights of signal (background) events. One can then proceed by defining a so called impurity function which tells how much signal and background is still mixed. The best split is then determined as the one maximizing the *decrease* of impurity. Some standard choices for the impurity function are:

- *Gini index*: $p \cdot (1-p)$
- *cross entropy*: $-p \cdot ln(p) - (1-p) \cdot ln(1-p)$
- *misclassification error*: $1 - max(p, 1-p)$
- *statistical significance*: $\frac{s}{\sqrt{s+b}}$

Note that the first three are maximal for $p = 0.5$ and tend to zero as $p \to 0$ or $p \to 1$ as it should be since a purity of 0.5 indicates a fully mixed sample and isolating only signal (p = 1) is as good as isolating only background (p = 0).

Once the tree is grown and all terminal leafs are determined an output value has to be assigned to each leaf. Usually this is the purity of the training events that constitute the leaf. If this exceeds a certain threshold events in it will be classified as signal and background otherwise. Now a cut value p_{cut} can be defined. If a leaf fulfills $p_{leaf} > p_{cut}$ events that end up in it during testing phase while be classified as signal, otherwise background. Likewise in the case of discrete output a leaf is determined as signal-like (+1) if $p_{leaf} > p_{cut}$ and background-like (-1) else.

It has yet to be answered under which circumstances a node should be turned into a leaf. Some of the most important stopping conditions are:

- perfect separation in a node, i.e. p = 0 or p = 1
- minimum node size: in order to ensure proper statistics one usually requires each node to contain a minimum number of events and a node will turn into a leaf once further splitting would undercut this value
- no further split can increase the purity in the sense that for no split the decrease of impurity surpasses a fixed value
- maximal depth: even if none of the above applies it is necessary to define a maximal amount of splits (counted from the root) after which the flow will end in a leaf. On the one hand this is required by limited computing resources on the other hand it is an important way to prevent overtraining, an issue treated in a moment

3.2 Boosting

From the previous section it is clear that decision trees have many favorable properties:

- Since for each split each variable is evaluated individually using many variables is not as resource intensive as most other multivariate classifiers. In fact the required CPU resources scale as $\propto n$ where n is the number of variables used
- All variables and events are sorted as they are evaluated, therefore one does not need to pay attention to how variables and events are presented to the decision tree. Especially there is no problem in having the same information encoded in multiple variables (e.g. when two variables are correlated)

- Furthermore variables that have poor to no discriminating power will simply be ignored and do not decrease the performance of the tree. This is again due to the sorting algorithm

All in all it seems like decision trees can be operated very safely without the need for too much manual adjustment. However, there is one huge shortcoming of the aforementioned procedure. While a single tree may be trained to perform very well on the initially given sample this does not necessarily carry over to further data sets on which the tree may be used. A single tree can enhance its performance only by growing larger branches. As the number of events in each node decrease, any further splitting will make use of very particular properties of the remaining events. The decrease of impurity will therefore not be the result of general discrepancies of signal and background but of splits that are tailored to the very events in the node. Once the node size undergoes a certain limit these discrepancies, while being existent, will not be statistically significant anymore and therefore of no use when applied to other samples. This is generally referred to as *overtraining*: The decision tree is very efficient at separating the sample at hand but has been trained too specifically on this sample. In practice one can check for overtraining by comparing the tree's performance on the training sample to its performance on the test sample.
There exist several strategies to reduce overtraining. One is the already mentioned maximal depth of each branch. If the tree is not allowed to split indefinitely it can be prevented from using too specialized splits. The same holds true for requiring a minimal nodesize.

Boosted decision trees try to minimize overtraining by considering not one but many decision trees in the classification process. Instead of one large tree many smaller trees are used which will suffer less from overtraining. These small trees will generally perform not as well as a deeper tree on their own but can outperform it when taken together. Furthermore the many trees that make up the boosted decision tree are not grown nor evaluated separately but take in account the performance of their predecessors.
The main idea of boosting is that each subsequent tree is not trained with the initial sample but a reweighted one: From the events misclassified by the k-th tree a boosting weight α_k is derived. The weight of all misclassified events is multiplied by some function of α_k and then make up, together with the correct classified events, the sample used for the k+1-th tree. One very common boosting method is AdaBoost [7]. The algorithm proceeds as follows:

- Train the k-th decision tree with the sample derived from reweighting the (k-1)-th sample
- evaluate which events were misclassified (during training phase)
- calculate the misclassification rate $\epsilon_k = \frac{\sum_{i,misclassified} w_i^k}{\sum_i w_i^k}$ as a weighted sum over the misclassified events of the k-th sample
- from this the weight of the of the k-th tree, $\alpha_k = \ln \frac{1-\epsilon_k}{\epsilon_k}$ is calculated
- reweight the misclassified events by multiplying each event with some function $f(\alpha_k)$

The misclassification rate is constructed such that it takes values ≤ 0.5 where 0.5 corresponds to random guessing. A common choice for the boosting function is $f(\alpha_k) = f(\alpha_k, \beta) = \exp(\alpha_k * \beta)$ where β is a user defined boosting parameter.
Thus events that were classified correctly are left unchanged while misclassified ones get a higher (boosted) weight and will therefore be more carefully considered in the growing process of the next tree. Even more so, since α_k depends on the misclassification rate ϵ_k the boosting function $f(\alpha, \beta)$ will become bigger as ϵ_k gets smaller so that events that were misclassified by a high performing tree will receive a bigger boost.

The power of boosted decision trees can be best understood when one looks at the overall evolution of the growing process: The first tree is grown using the original sample. It is likely that this tree will also give the best single performance. The subsequent trees will focus more and more on the events misclassified by their predecessors as these events get higher weights each time they are misclassified and may at some point dominate the sample. However this means in return that the splitting will get more specialized towards single events, neglecting the features of the overall sample. It is thus expected that such trees perform relatively badly, corresponding to a high misclassification rate ϵ and a low tree weight α. In summary, later trees are not better versions of the first one but classifiers specialized to give corrections for events hard to classify. They may perform not so well on most other events but only add a small weight to the overall output which is the weighted sum of all single tree outputs. In fact one observes in actual applications a tendency towards $\epsilon \to 0.5$ and $\alpha \to 0$ for high tree numbers, as also shown in the next subsection.

3.3 TMVA implementation and BDT results

TMVA offers various decision tree algorithms but in this thesis only AdaBoost will be used. In order to operate a boosted decision tree its parameters must be specified. A full list of available options can be found in [6], page 110ff. In addition to select the tree parameters it is possible to specify preselection cuts which effectively act as a cut flow before decision trees are generated. They may be used to increase the signal to background ratio or to ensure events entering the BDT build a more realistic scenario. In the rest of this section some of these options will be tested and the stability of the boosted decision trees under variation of its parameters is evaluated. This will give valuable hints which settings may be considered optimal in a further analysis and how much they realistically can be tuned in order to optimize the performance of the BDT. As a measure of performance we will consider and compare the Receiver-Operating-Characteristic-Curve (ROC-curve). If not explicitly stated otherwise BDT parameters are set to the TMVA default values. The evaluations in this section were made with a sample consisting of 1 million events for the signal channel, 1 million events for the background channels Z-QCD, Z-QED, W-QCD and 10.000 events for the W-QED channel. No preselection cuts were applied at this stage.

The variables entering the BDT are the following:

$$p_{T,1},\ p_{T,2},\ \eta_1,\ \eta_2,\ \not{p}_T,\ \text{number of jets},\ |\Delta\phi_{1,2}|,\ m_{1,2},\ \eta_1 \cdot \eta_2$$

These are basic variables that will almost always be used in any BDT analysis. To check if a BDT trained with them is sufficient we compare its performance to a cut flow. This is done in Figure 3.2 where the signal efficiency vs Z-QED background efficiency is displayed. The BDT consisted of 800 individual trees and each node was required to contain at least 1000 events. The cut flow used here as a reference was the following:

$$p_{T,2} > 40\ \text{Gev},\ |\Delta\eta_{1,2}| > 4.4,\ \eta_1 \cdot \eta_2 < 0,\ \not{p}_T > 100\ \text{GeV},\ m_{1,2} > 1200\ \text{GeV},\ \text{Central Jet Veto},\ |\Delta\phi_{1,2}| < 1$$

The BDT clearly outperforms the cut flow and we conclude that the basic variables described above are sufficient to generate a realistic BDT performance.
Figure 3.3 shows the corresponding tree weights and error fraction. They support our earlier expectation that the error fraction ϵ rapidly tends towards 0.5 while the boost weight α tends towards zero.

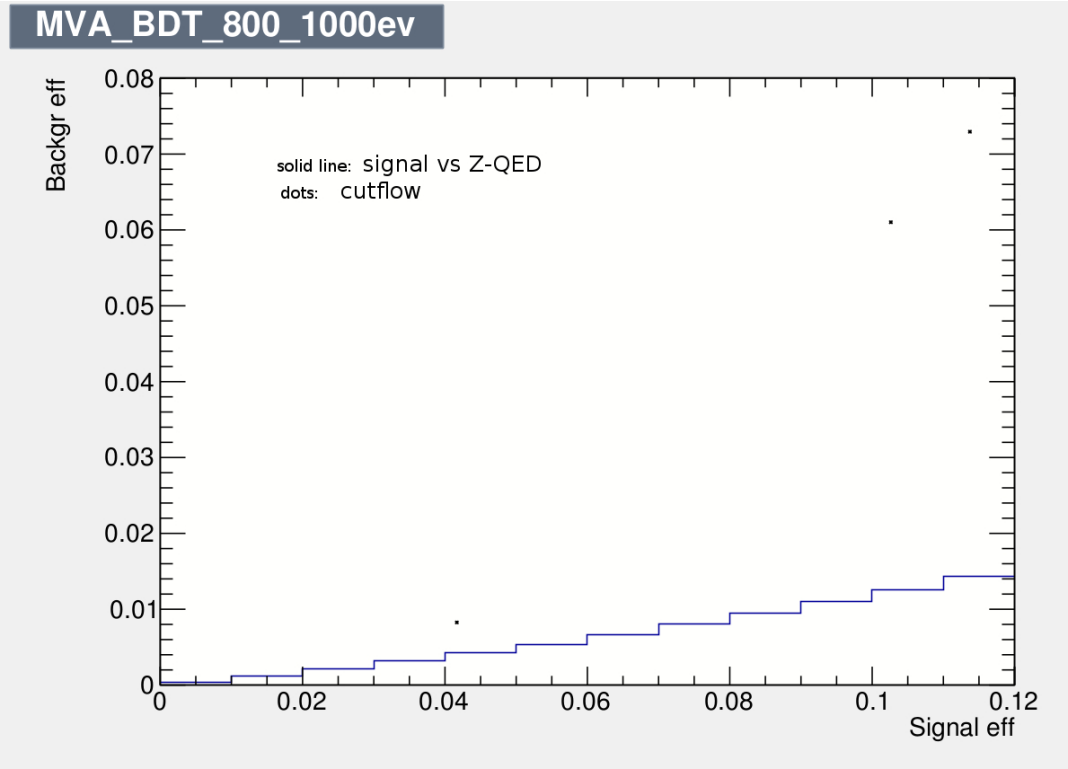


Figure 3.2: ROC curve signal vs Z-QED. The three dots represent the cutflow at $m_{1,2} > 1200$, central jet veto and $|\Delta\phi_{1,2}| < 1$ (from higher to lower signal efficiency).

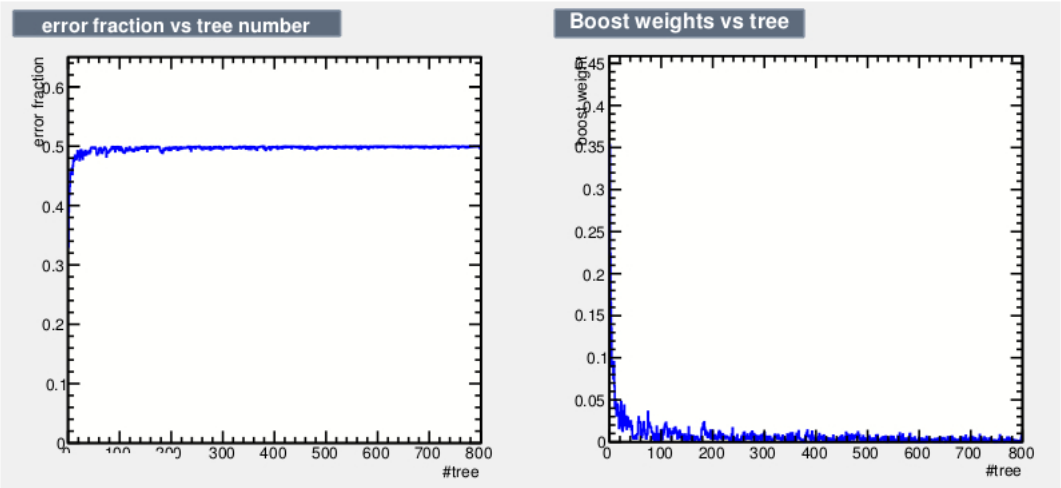


Figure 3.3: Error fraction and boost weight as function of tree number.

- The most straightforward BDT parameter is the number of individual trees that make it up. From a pure logical viewpoint more trees should never decrease the overall performance however this is only true if the sample entering the BDT is a perfect model in the sense that there occur no numerical limitations whatsoever. Aside from that more trees require more computing resources which may limit the amount of trees one wants to use in practice, especially when the sample contains millions of events. Figure 3.4 (a) depicts four BDTs built from 10, 50, 100 and 400 individual trees. As expected the BDT with only 10 trees performs slightly worse than the others but basically no difference can be seen with at least 50 trees.This can be understood as evidence for the earlier mentioned notion of the BDT growing process: Only a rather small amount of trees perform very well on the complete sample while the following merely give corrections to difficult events. It should be noted that the addition of more variables may allow more sophisticated splits which in return could make a larger amount of trees beneficial. Despite that it seems like there is no benefit in exceeding a few hundred trees, an amount still suitable for numerous calculations

- Another central parameter is the maximal depth of the trees i.e how many splits are allowed before a node is mandatory. Figure 3.4 (b) shows the ROC curves of four different BDTs, two with a hundred individual trees and a maximal depths of 3 or 6 and two with four hundred individual ones and 3 or 6 maximal splits respectively. Virtually no performance difference occurs indicating that the variables at hand can make no use of deep splittings. As previously stated deeper splittings increase the likelihood of overtraining. To check this Figure **??** shows the classifier output of signal and background during testing and training phase for the BDTs with 400 individual trees. Difference between testing and training results indicates overtraining. As can be seen in both cases no difference occurs for the signal sample whereas for the background some distinction is visible. As expected the BDT allowing a maximal depth of 6 splits shows greater difference than the one with only 3 maximal splits. Apparently the slight overtraining is not big enough to decrease performance in the current case but as it clearly increases with the maximal depth there seems to be no reason to use a maximal depth larger 4

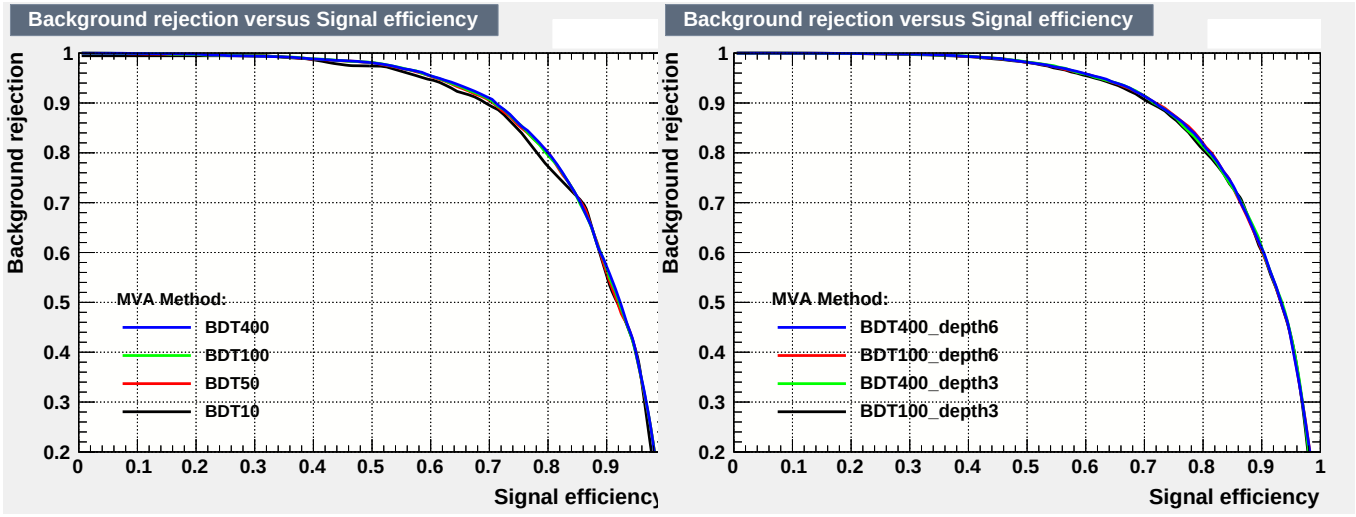
- Figure 3.6 displays ROC curves testing a variation of the minimal node size (a), the boosting parameter β (b), the number of evaluated cuts per split (c) and the impurity function determining the best split (d)

– Only very little discrepancy is observable for the different minimal node sizes. One could have expected that a minimum as little as 50 events leads to strong overtraining resulting in a decrease in performance but since we do not allow very deep splitting it is reasonable to assume that the separating splits are not strong enough to reach that minimum and therefore no stronger overtraining occurs. Thus we conclude it is unproblematic to chose the minimal node size large enough to ensure proper statistics

– Varying the boosting parameter β leads to no significant variation in performance. This parameter determines how strong subsequent samples are reweighted so a larger value puts more emphasis on falsely classified events. However, as we use at least a hundred individual trees it appears likely that there are enough tree variations to deal with difficult events regardless of boosting strength. Since we have no interest in going far below hundred trees adjusting the boosting parameter in a further analysis seems to be unnecessary

– Most variables entering the BDT are in fact continuous. Although measurements will only yield discretized results a determination of the best cut at each split would require the algorithm to consider a very large amount of variations. In practice only a limited number of cuts is checked. Figure 3.6 (c) shows no difference between 20, 30 and 50 considered cuts per split. As a bigger number results in increased computation time the TMVA default value of 20 requires no modification.

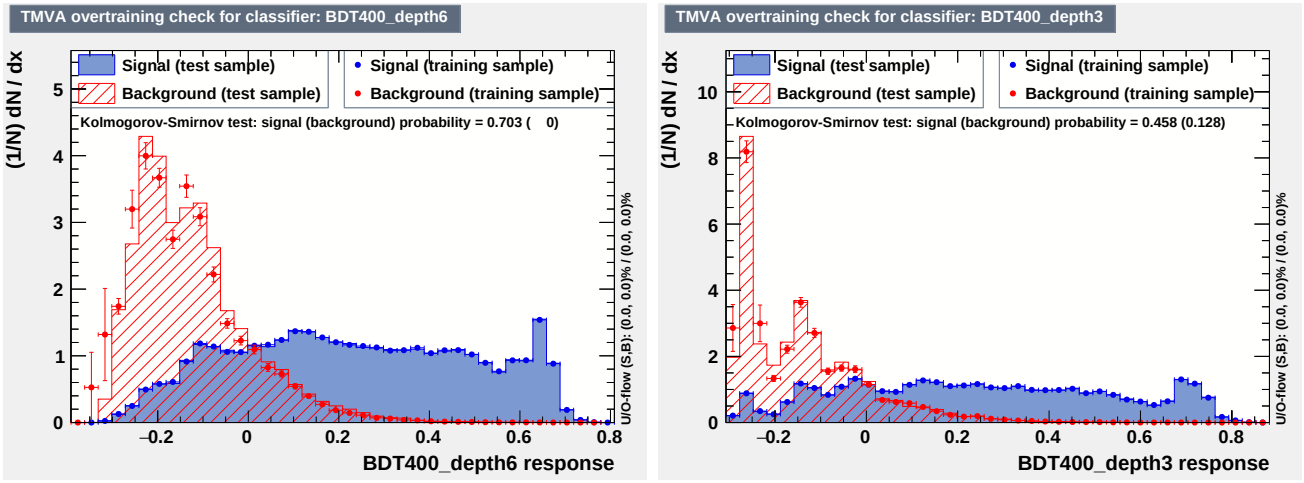
– At the beginning of Section 3.1 four different impurity measures were presented: The *Gini index*, the *cross entropy*, the *misclassification error* and the *statistical significance*. Figure 3.6 (d) depicts four BDTs consisting of a hundred individual trees and using one of the four impurity functions to determine splitting. One sees that there is no difference in performance between Gini index, cross entropy and misclassification error whereas the BDT using statistical significance performs clearly worse than the others. Again we conclude that the default value (Gini index) needs no modification



(a) ROC curves for BDTs consisting of 10, 50, 100 and 400 individual trees.

(b) ROC curves for BDTs with different maximal depth.

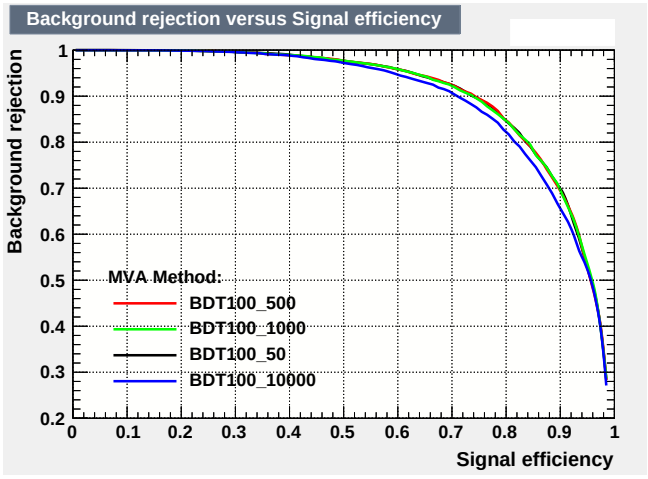
Figure 3.4: Variation of BDT parameters.



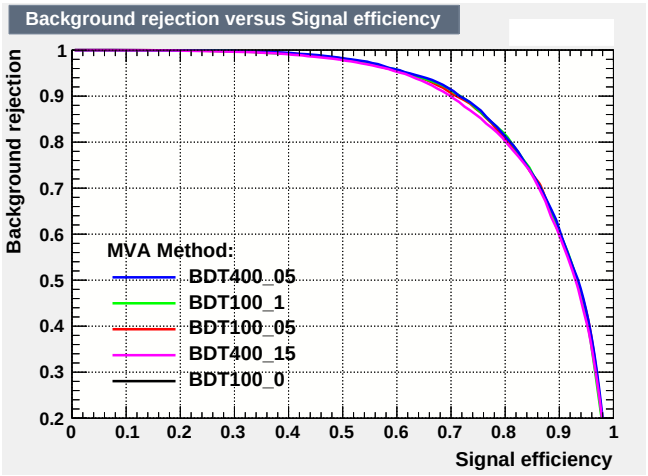
(a) number of trees = 400, maxdepth = 6.

(b) number of trees = 400, maxdepth = 3.

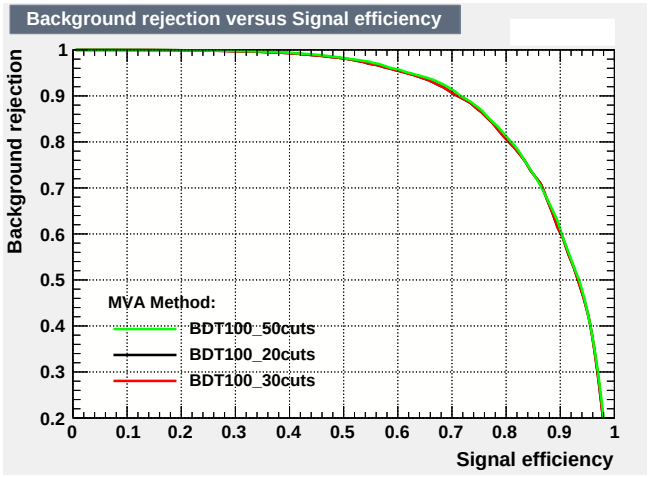
Figure 3.5: Overtrain check for a BDT with a maximal depth of 6 (left) and one with maximal depth of 3 (right). Signall and background are superimposed. The solid line gives the classifier output during testing phase, the dots at training phase. A difference between the two indicates overtraining.



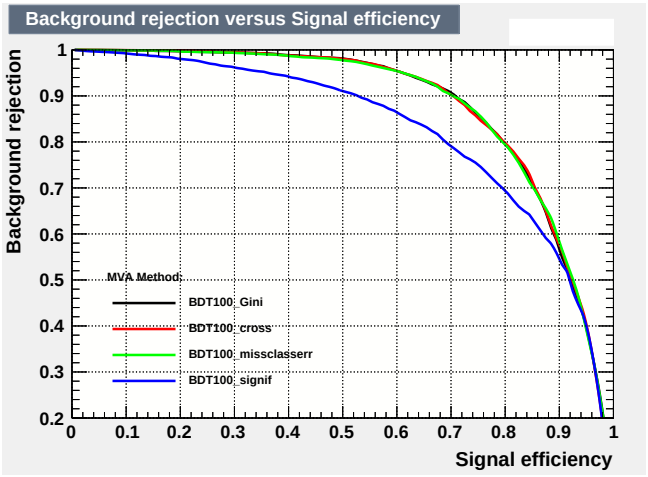
(a) ROC curvees for BDTs requiring a minimal node siize of 50, 500, 1000 and 10000 events. The total sample size was 4010000 events.



(b) ROC curves for different values of the boosting parameter β. Tested were β = 0, 0.5, 1.0 and 1.5.



(c) ROC curves for different numbers of cuts considered per split.



(d) ROC curves for different impurity functions.

Figure 3.6: ROC curves for variation of more BDT parameters.

The results can be summarized as follows: First, the boosted decision tree algorithm seems to be very robust when it comes to variation of parameters. Most of the changes from default values tested showed no significant change in performance. This means an overall BDT setup can be expected to function very well under a wide range of circumstances but a meaningful increase in performance, if possible at all, may only result after tweaking the settings according to specific properties like sample size, number of input variables or the kind of input variables.
We expect a few hundred hundred individual trees to be the best compromise between performance and computing time and do not want to have a bigger maximal depth than 4 to not counteract the notion of weak classifiers. Furthermore it was seen that there is no problem in choosing the minimal node size large enough to ensure statistical significance. However it must be kept in mind that some of these findings may require modification if the properties of an analysis strongly differ from those used for this tests as could be induced by much bigger samples or vastly more available information through more variables.

4 Quark Gluon Discrimination

Being able to distinguish quark-initiated jets from gluon-initiated ones can be valuable in different topics in. Consequently quite some research is devoted to Quark-Gluon discrimination with improving success (see for example [8], [9]). The task is however not an easy one: First of all it is not clear what is meant by the terms "quark jet" or "gluon jet". Only fully hadronized final state jets can be measured and the quantum nature of this process does not allow to conclude the initiating parton. Therefore a sensible definition of quark- or gluon jet can only be given after one has established a showering algorithm. A thorough treatment of this issue can be found in [10]. Since in this thesis all processes are evaluated at lowest order parton level the most straightforward definition is sufficient: Upon event generation any final state quark will result in a quark jet and the equivalent holds true for gluons. From the aforementioned it is clear that one must consider statistical kinematic distributions of large samples. For weak boson fusion the tagging jets of the signal are quark initiated so one can hope that quark gluon discrimination improves the separation between signal and qcd backgrounds considerably.

4.1 Quark gluon discriminating variables

Before discriminating variables can be defined it must first be understood why any difference between quark and gluon jets is expected at all. At a fundamental level quarks enter the QCD Lagrangian as spinors in the fundamental representation while the gluon, being the particle associated with the SU(3) gauge field, is given in the adjoint representation. This leads to different values in the constant associated with the Casimir operator which is often useful in perturbative calculations: For SU(N) one finds for the fundamental representation

$$C_2(r) = \frac{N^2 - 1}{2N}$$

while for the adjoint representation it is $C_2(G) = N$.
In the case of QCD (SU(3)) this means $C_2(r) = 4/3$ and $C_2(G) = 3$. This factors often enter perturbative calculations of Feynman diagrams so qaurks and gluons will display different splitting behavior. For this and the fact that gluons have no electric charge it might be possible to distinguish quark jets from gluon initiated ones: Without kinematic differences it is expected that gluon jets have more soft wide angle emission and split more in general, leading to bigger numbers of particles in a jet.

On this basis the following variables will be used and tested, for a broader explanation see [11] and the references there.

$$
\begin{aligned}
n_{\mathrm{PF}} &= \sum_{\mathrm{PF}\in\mathrm{jet}} \\
w_{\mathrm{PF}} &= \frac{\sum_{\mathrm{PF}\in\mathrm{jet}} p_{T,\mathrm{PF}} \Delta R_{\mathrm{PF,jet}}}{\sum_{\mathrm{PF}\in\mathrm{jet}} p_{T,\mathrm{PF}}} \\
Q^{\kappa} &= \frac{1}{(p_T)^{\kappa}} \sum_{\mathrm{trk}\in\mathrm{jet}} q_{\mathrm{trk}} (p_{T,\mathrm{trk}})^{\kappa} \\
C_{\beta} &= \frac{\sum_{i,j\in\mathrm{jet}} E_{T,i} E_{T,j} \left(\Delta R_{ij}\right)^{\beta}}{\left(\sum_{i\in\mathrm{jet}} E_{T,i}\right)^2}, \quad i,\, j \in \mathrm{PF} \\
f_{\mathrm{largest}} &= \frac{E_{\mathrm{largest}}}{E_{\mathrm{jet}}}
\end{aligned}
$$

PF refers to particle flow objects as identified by Delphes while a lowercase T indicates transverse momentum. For Q^{κ} the sum runs over all reconstructed objects with nonzero charge. ΔR is defined as $\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$. $E_{largest}$ is the total energy of the reconstructed object which carries the biggest energy fraction of the whole jet. The parameters β and κ are freely adjustable (as long as they are kept bigger than zero) with default values of $\beta = 0.2$, $\kappa = 1$.[12]

4.2 Implementation and testing

The samples used for the following evaluation consisted of 10000 Z qcd background events. One sample contained only quark jets the other only gluon jets in the final state. In actual data both will be mixed of course but this idealized situation allows to easily check the maximal possible discrimination power.

As a first check one should compare the histogram of each variable for the pure samples. They are shown in Figure 4.1. In the histogram for n_{PF} entries with only one object were omitted, in the w_{PF} histogram those with $w_{PF} = 0$ since they dominate the histogram as shown in Figure 4.2. Every histogram shows some distinction between the quark and gluon samples, however the variables w_{PF} and $C_{0.2}$ clearly give the best separation. This is not completely surprising: w_{PF} extends the information encoded in the number of objects n_{PF} by also taking angular information and energy deposit into account.
The variables $Q^{1.0}$ and $f_{largest}$ show no strong difference between the two samples. Since

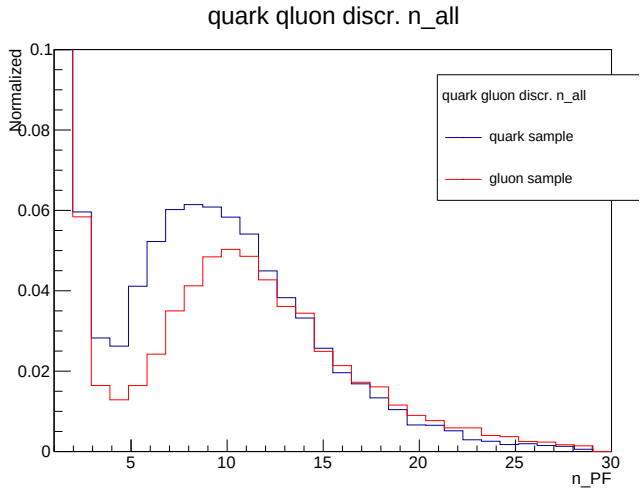
this means their possible separation power will be even further reduced in a less idealized case they might not be suited to greatly contribute to event classification.

In order to check the variables separation power more clearly it is interesting to look at their ROC-Curves, in this case a plot of quark acceptance vs gluon rejection. They were manually created by applying varying cuts on the respective variable and counting how many events fall below it. This fraction was then divided by the total number of events. For the quark sample this gives the quark acceptance while for the gluon sample the rejection (= 1 - acceptance) was calculated. The amount of different cuts used vary between 8 to 18 and they were chosen such that the whole acceptance region, the interval $[0, 1]$, was filled as broad as possible. The individual points thus gathered have been connected to a smooth curve. Note that only events with $n_{PF} > 1$ and $w_{PF} > 0$ were considered. These curves are plotted in Figure 4.3. Again one sees that the variables $C_{0.2}$ and w_{PF} outperform the others by far. w_{PF} shows the best separation in the regime of low quark acceptance while there is a cross-over point at a quark acceptance around 0.7 where $C_{0.2}$ becomes the best variable. The variables $f_{largest}$ and n_{PF} show decent separation power while $Q^{1.0}$ performs the worst.

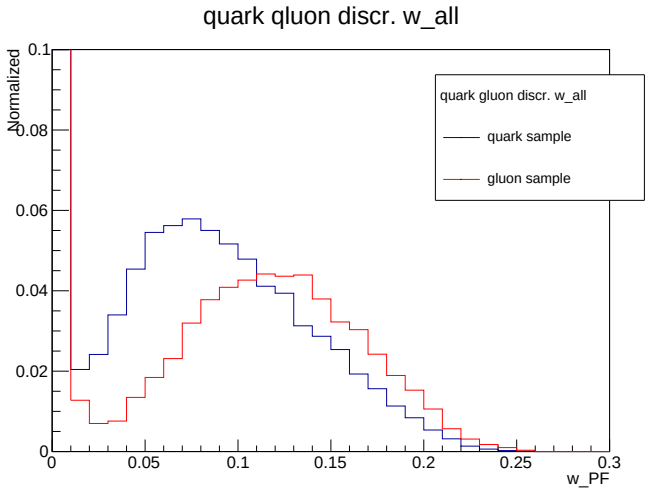
One may hope to increase the performance of Q^κ and C_β by varying the parameters. For C_β the values $\beta = 0.1, 0.2, 0.5, 1, 2$ were probed. The resulting histograms are depicted in Figure 4.4. All seem to give a separation power comparable to the standard value $\beta = 0.2$, in order to determine the best values these histograms were also used to create ROC-curves as before. They are also shown in Figure 4.4 and clearly indicate that the value of β does not significantly change the performance. Only in the quark-acceptance region of 0.2 to 0.6 does the curve corresponding to $\beta = 0.1$ fall off slightly. Overall $\beta = 0.5$ performs marginally better than the others.

A similar analysis was done with Q^κ. Here the probed values were $\kappa = 0.1, 0.2, 0.5, 1.0, 2.0$. The resulting histograms and ROC-curves are shown in Figure 4.5. Since the parameter κ only affects the transverse momentum of jet constituents high values of κ stress this difference strongly and only the highest p_T objects contribute significantly to the sum. This in return leads to very poor separation power between quark- and gluon jets as can be seen in Figure 4.5 (e) for $\kappa = 2$. The problem might be that only so few objects contribute that statistical uncertainties get too big.

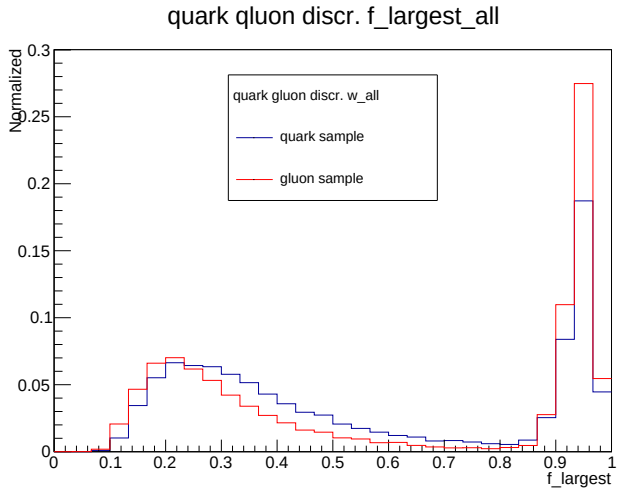
For small κ the in principle discrete nature of Q^κ becomes apparent: The reconstructed objects posses after clustering charges of integer value. These discrete values get somewhat smeared by the p_T weighting but small values of κ mitigate the influence of



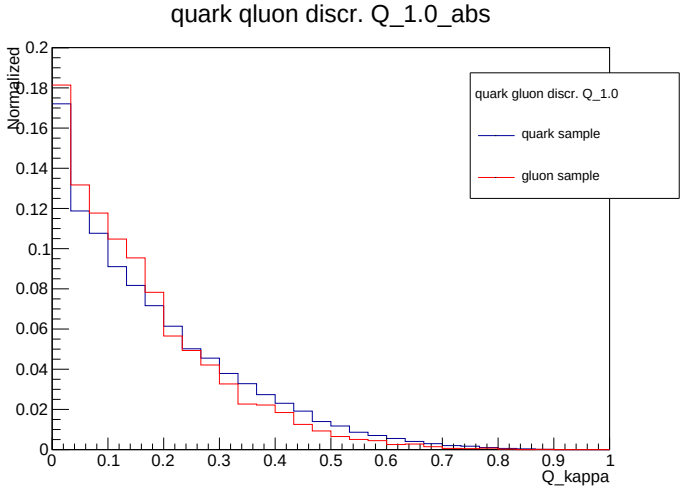
(a) n_{PF} with zoom on entries larger 1.



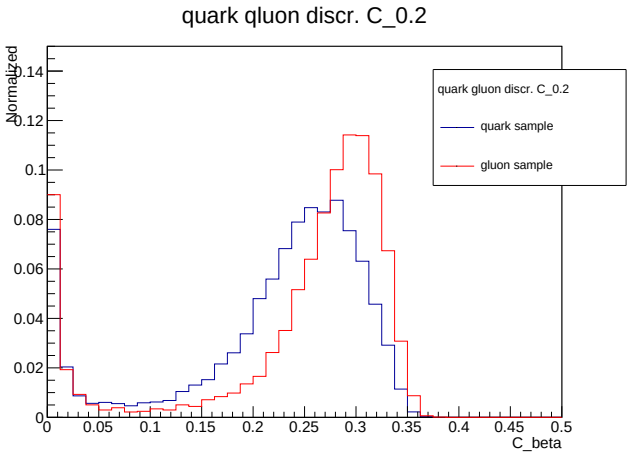
(b) w_{PF} with zoom on entries larger 0.



(c) $f_{largest}$

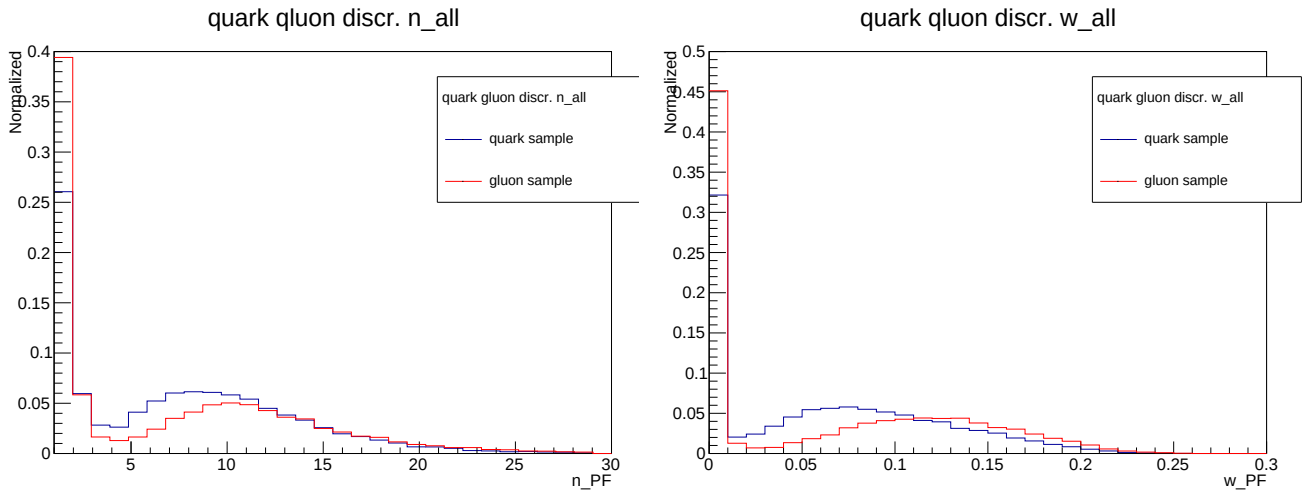


(d) absolute value of Q^{κ} for the default value $\kappa = 1$.



(e) C_{β} with default value $\beta = 0.2$

Figure 4.1: Normalized histogram for each quark-gluon discriminating variable for the pure quark- and pure gluon sample.



(a) n_{PF} with full y-range.

(b) w_{PF} with full y-range.

Figure 4.2: n_{PF} and w_{PF}. The samples used are dominated by jets consisting of only one reconstructed particle.

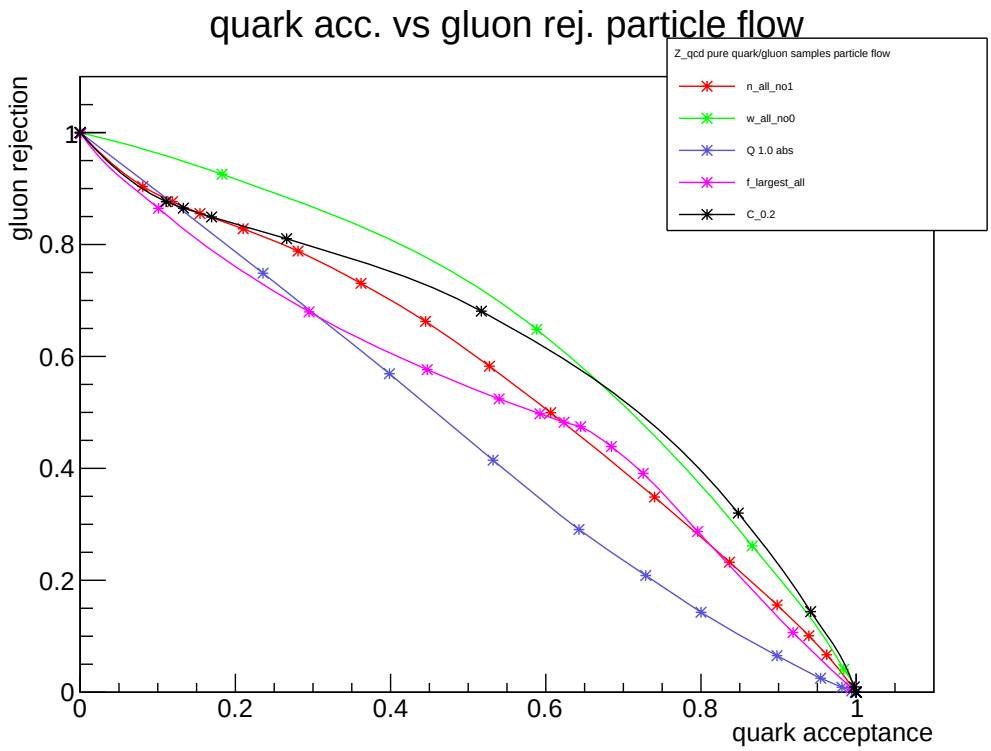
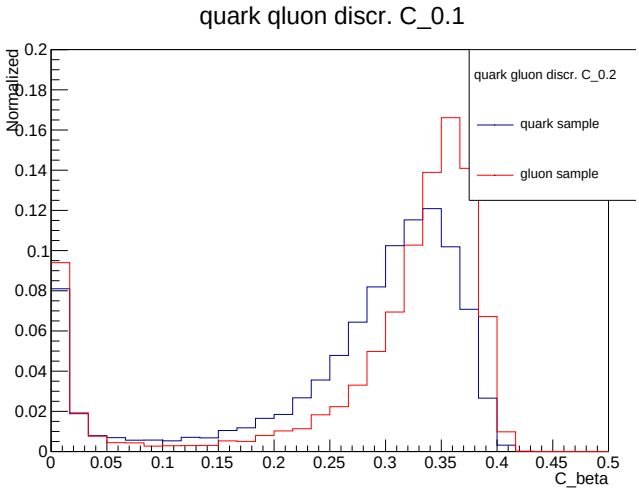
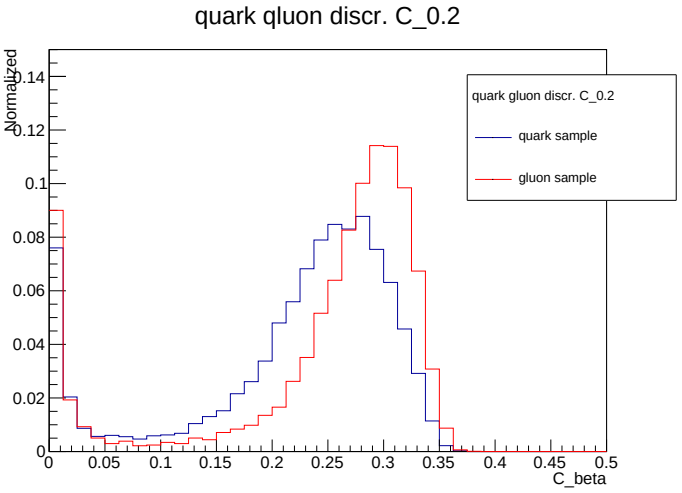


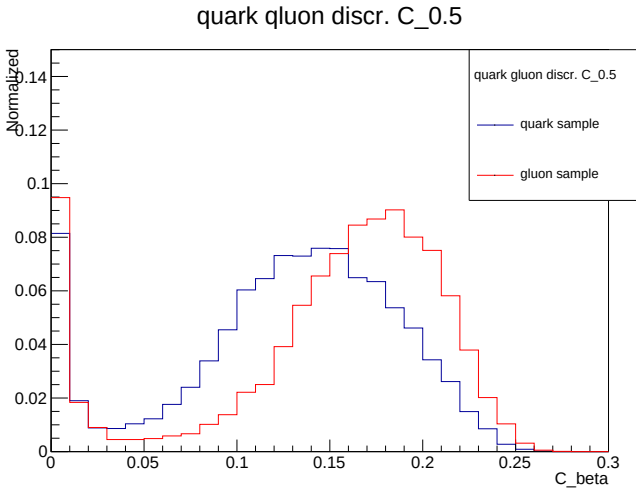
Figure 4.3: Manually created ROC-curves for the quark-gluon discriminating variables.



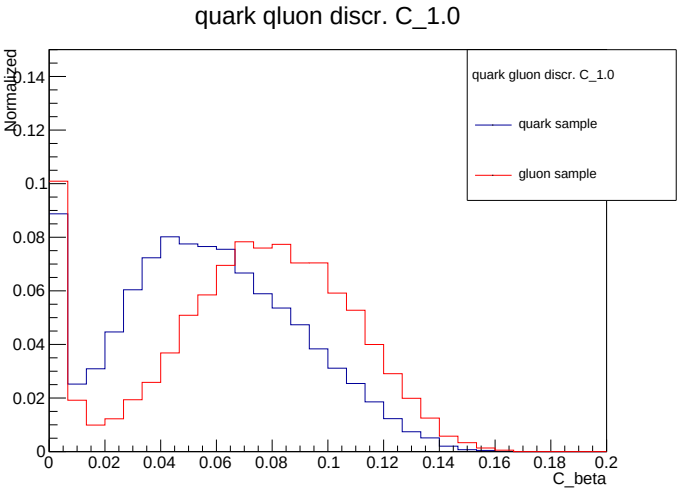
(a) $\beta = 0.1$



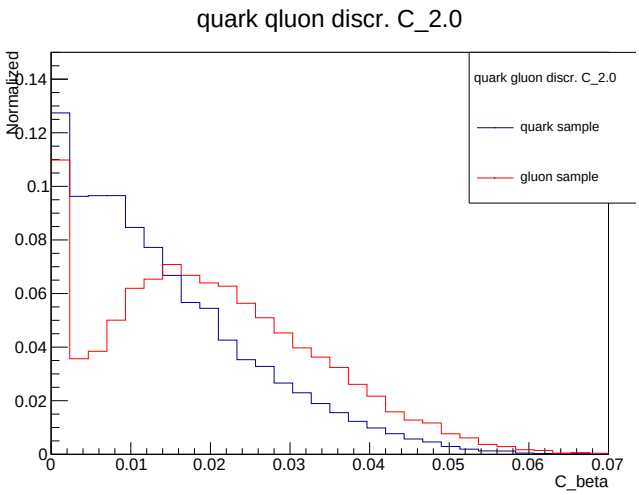
(b) $\beta = 0.2$



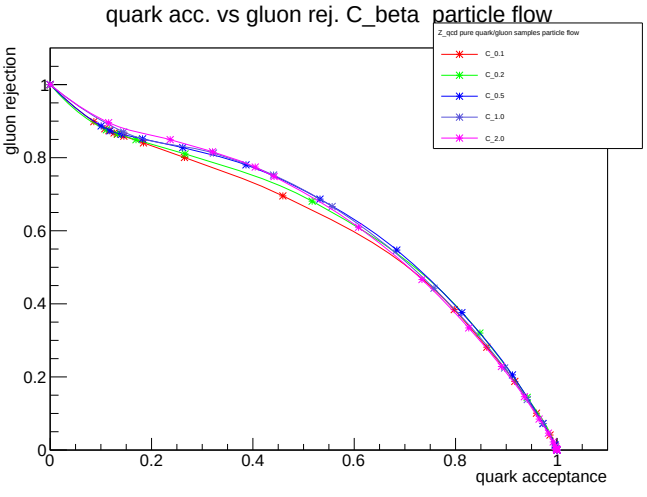
(c) $\beta = 0.5$



(d) $\beta = 1.0$

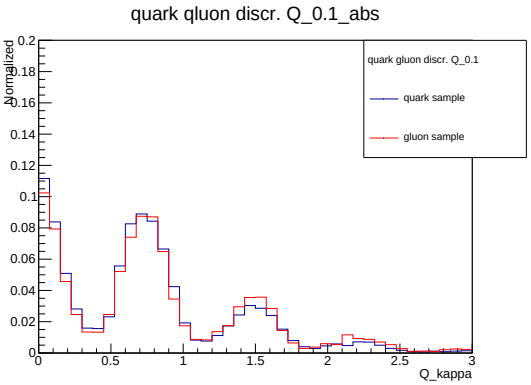


(e) $\beta = 2.0$

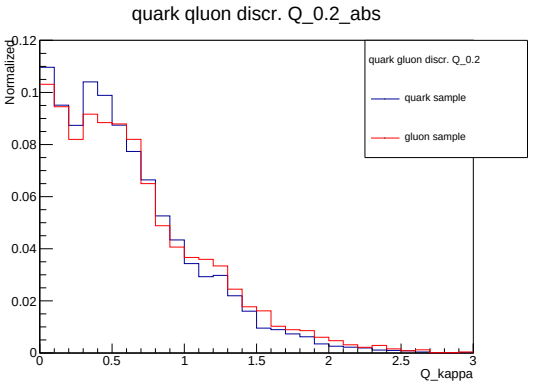


(f) Manually created ROC-curves for the different values of β.

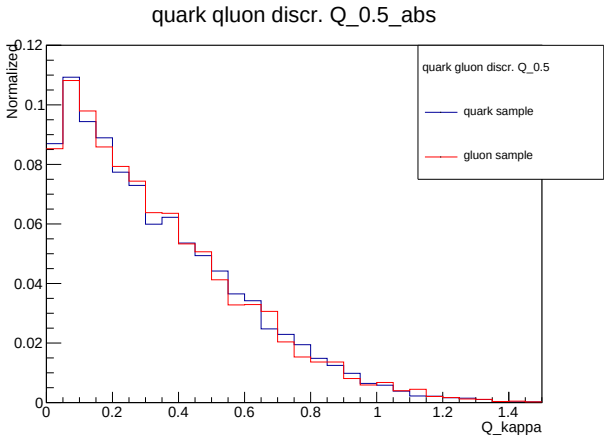
Figure 4.4: Normalized histograms for C_β with variable β (a-e) and ROC-curve (f).



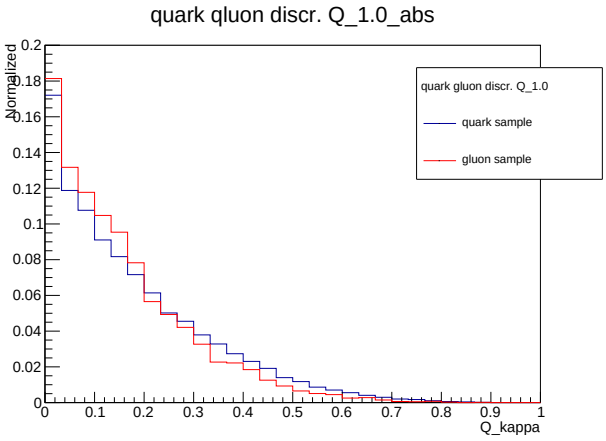
(a) $\kappa = 0.1$



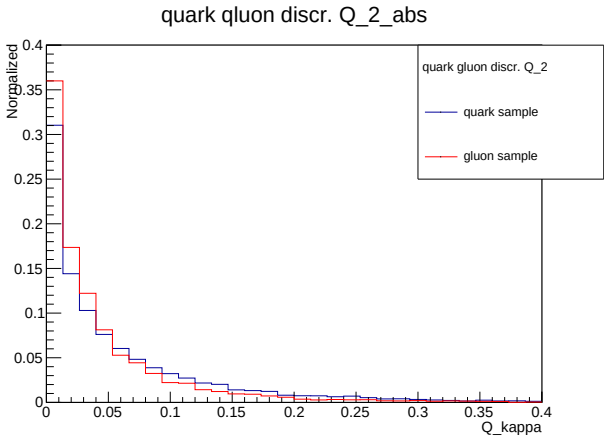
(b) $\kappa = 0.2$



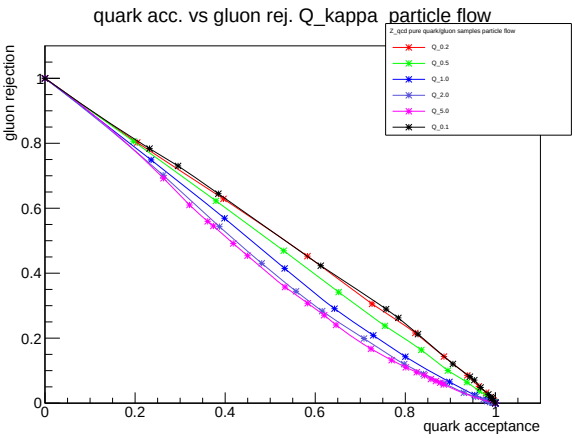
(c) $\kappa = 0.5$



(d) $\kappa = 1.0$



(e) $\kappa = 2.0$



(f) Manually created ROC-curves for the different values of κ.

Figure 4.5: Normalized histograms for Q^κ with variable κ (a-e) and ROC-curve (f).

transverse momentum differences. Since the discriminating power of Q^κ stems from gluons being electrically neutral it is reasonable to assume that small values of κ, which emphasize charge differences, lead to a better separation. This notion is supported by the ROC-curve (g) in Figure 4.5 in which the small values outperform high values by far.

Despite this relative improvement the variable Q^κ still performance very poorly when compared to the other quark gluon discriminating variables as can be seen in Figure 4.6 which shows the same as Figure 4.3 except that now the values $\kappa = 0.1$ and $\beta = 0.5$ were used.

4.3 BDT implementation of quark gluon discriminating variables

In the following the performance of quark gluon discrimination as part of a BDT is tested. The samples used for this analysis consisted of 6 million events for signal and each of the four major backgrounds. In all BDTs the variables from Section 3.3 will be used. In addition to that some of the following may be included:

- 3rd Jet p_T and angular information: $p_{T,j3}$, ϕ_{j3}, η_{j3}
- 4th jet transverse momentum $p_{T,j4}$
- n_{PF}, w_{PF}, $Q^{1.0}$ and $C_{0.2}$ for the hardest and second hardest jet ($j_{1,2}$)

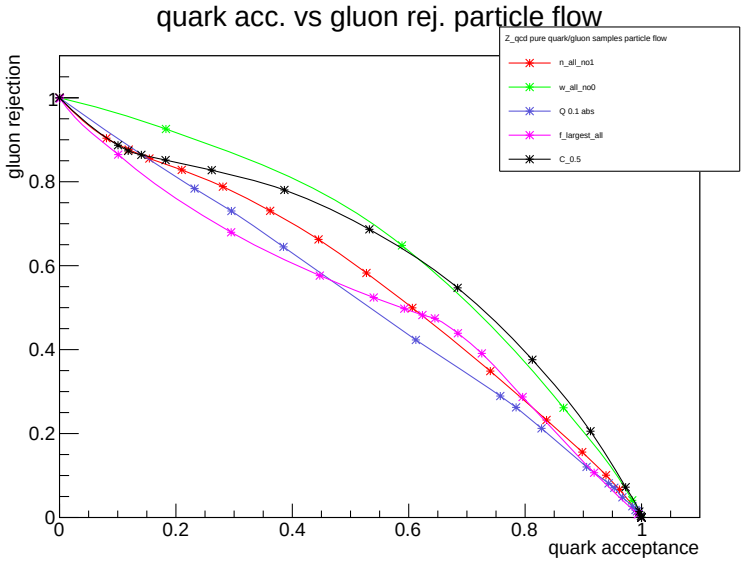


Figure 4.6: Manually created ROC-curves for the quark-gluon discriminating variables with optimal values for κ and β.

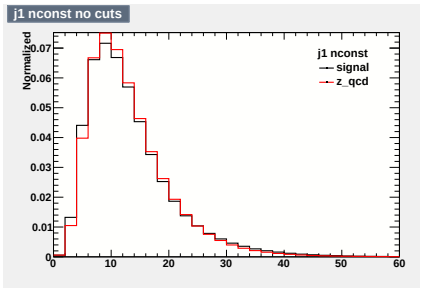
As a consistency check we again look at the histograms for the quark gluon variables. They are depicted for signal and Z-QCD in Figure 4.7. Again we see that the variables w_{PF} and $C_{0.2}$ show the biggest difference between signal and background whereas $Q^{1.0}$ shows very little difference. Furthermore the separation is smaller for all variables as compared to their counterparts in Figure 4.1. This is due to the fact that we are now not dealing with pure but mixed samples.

In the analysis that follows preselection cuts will be applied to all BDT's since they heavily increase the signal to background ratio prior to decision tree building. The preselection used cuts are those suggested by CMS [17]

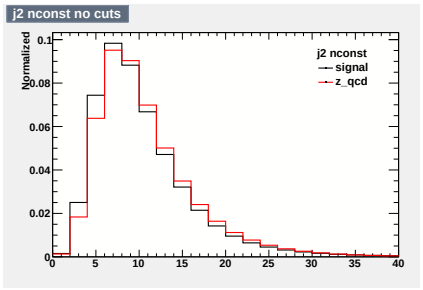
$$\eta_1 \cdot \eta_2 < 0,\ |\Delta\eta_{jj}| > 3.5$$
$$p_T(j_1) > 40 \text{ GeV},\ p_T(j_2) > 40 \text{ GeV}$$
$$m_{jj} > 600 \text{ GeV},\ \not{p}_T > 140 \text{ GeV}$$

In order to check if we still have decent statistics after these cuts and to get a feeling for possible discrimination power of the quark gluon variables their histograms are once again plotted. Figure 4.8 shows the histograms of signal and Z-QCD background after these preselection cuts.
Several things are remarkable:

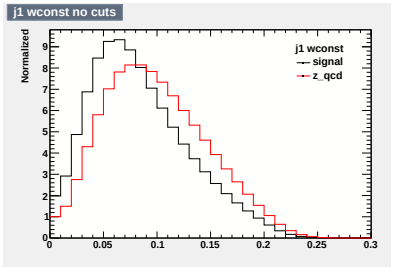
- the difference between signal and background for the variable $Q^{1.0}$ is left unchanged (at its low separation power)
- both C_β variables exhibit slightly less separation power than without cuts but still show the best or second best distinction of all quark gluon discriminating variables
- the variables w_{PF} and n_{PF} basically reverse their roles: After cuts there is now very little difference between signal and background for w_{PF} while the separation for n_{PF} has increased in contrast to all other variables



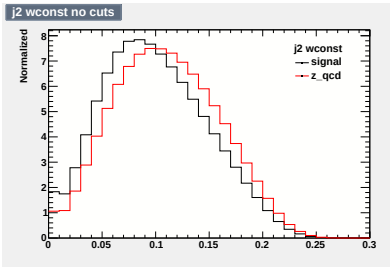
(a) j1nconst



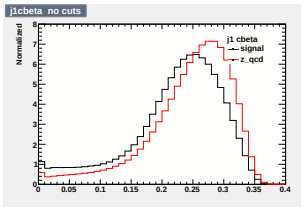
(b) j2nconst



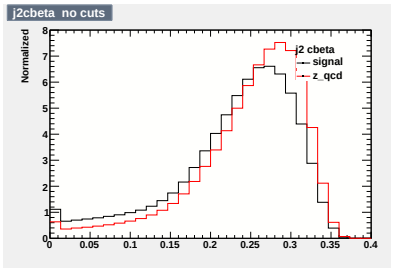
(c) j1wconst



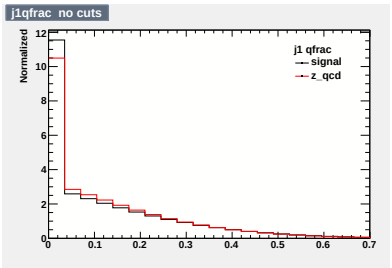
(d) j2wconst



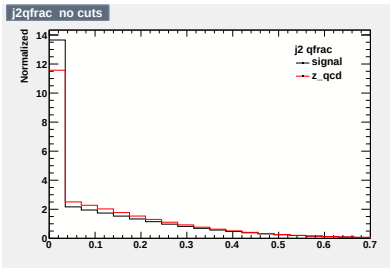
(e) j1 C0.2



(f) j2 C0.2

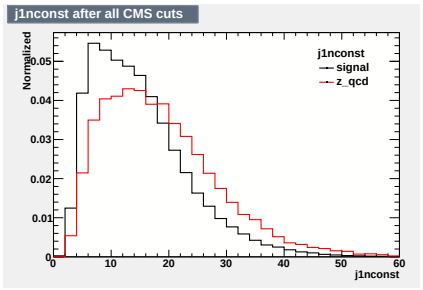


(g) j1 Q1.0

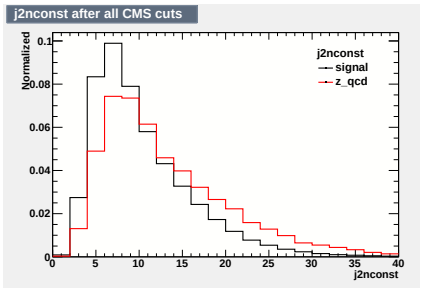


(h) j2 Q1.0

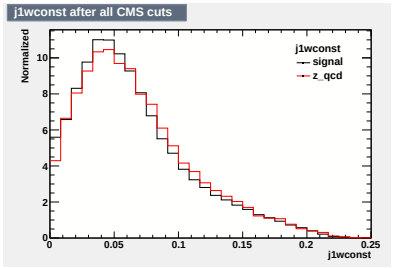
Figure 4.7: Normalized histograms for the quark gluon variables implemented for a BDT analysis without preselection cuts.



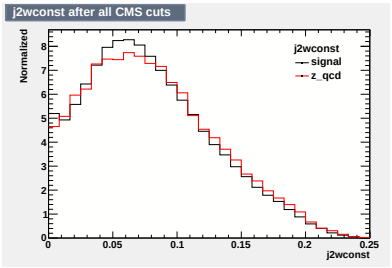
(a) j1nconst



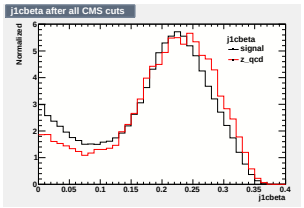
(b) j2nconst



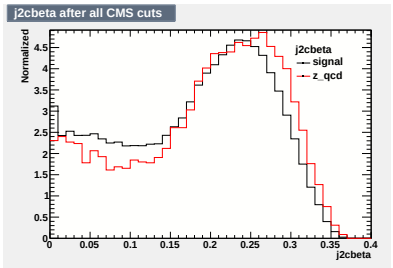
(c) j1wconst



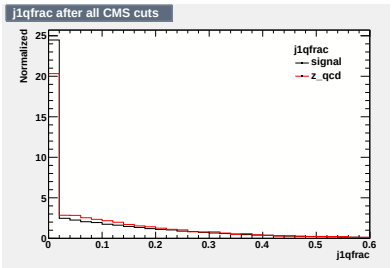
(d) j2wconst



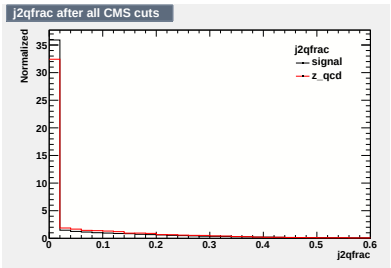
(e) j1 C0.2



(f) j2 C0.2



(g) j1 Q1.0



(h) j2 Q1.0

Figure 4.8: Normalized histograms for the quark gluon variables implemented for a BDT analysis after the CMS preselection cuts.

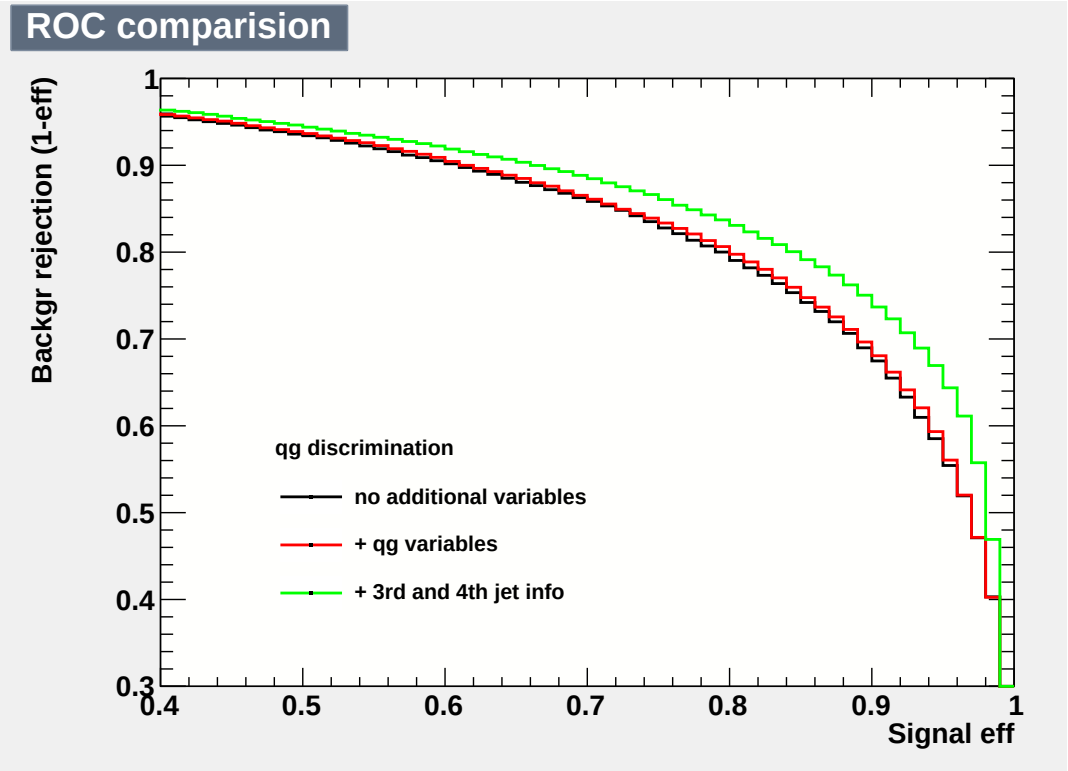
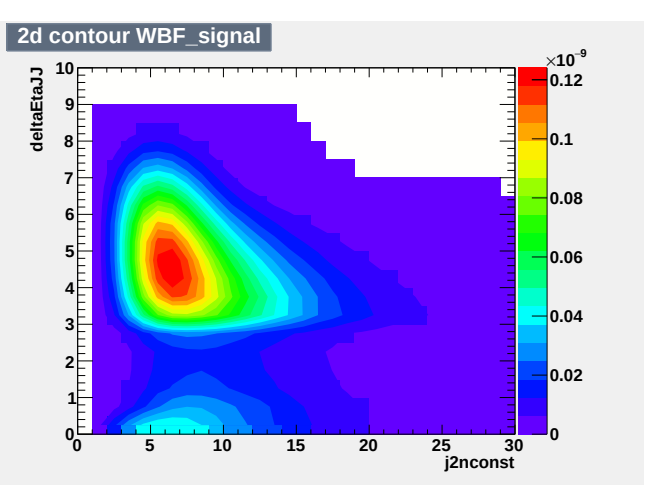
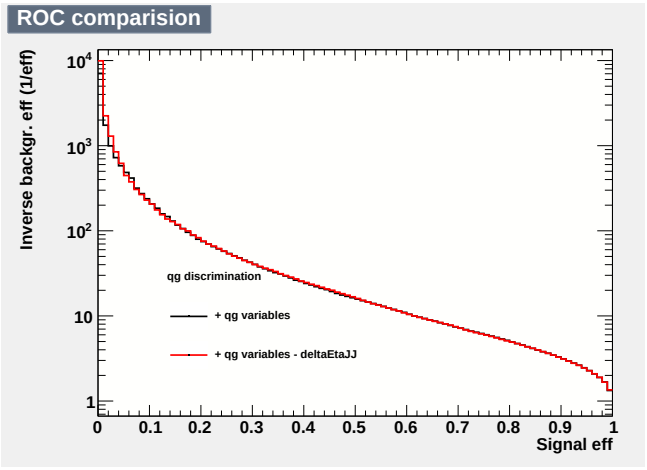


Figure 4.9: ROC curves of BDT with basic variables, with addition quark gluon variables and with additional 3rd and 4th jet information.

We want to test the performance of quark gluon discriminating variables as part of a BDT. Figure 4.9 shows the ROC curves for a BDT with only the standard variables, one with additional information of the quark gluon variables and one with full information of the 3rd and 4th jet. While additional jet information clearly increases the BDT performance the addition of quark gluon variables has merely a noticeable effect. An explanation for this is that the classification problem is over constrained: All information encoded in the quark gluon variables seem to be already provided by the basic variables. To check this Figure 4.10 (a) shows a 2d contour plot of $n_{PF,2}$ vs $|\Delta\eta_{1,2}|$ for the signal file. Clearly some correlation is present although it seems to be not too strong but it is anyways expected that the information is provided by the sum of all basic variables. Some evidence of this can be seen in Figure 4.10 (b) which compares the ROC curve of a BDT with quark gluon variables to one where the second best performing variable $|\Delta\eta_{1,2}|$ does not enter the BDT. Despite that the two curves are almost identical, this indicates that quark gluon variables can in fact be very useful when substituting for other variables.



(a) Contour plot $n_{PF,2}$ vs $|\Delta\eta_{1,2}|$.



(b) Comparison of performance with and without $|\Delta\eta_{1,2}|$.

Figure 4.10: Overconstrain check.

We continue the investigation of quark gluon variables by also considering them for the 3rd hardest jet. This may be especially useful when it is hard to obtain full additional jet information. In Figure 4.11 three settings are compared: A BDT that only has information of the basic variables with central jet veto information, one that has in addition to that information of the 3rd jet quark gluon variables and one that has full 3rd and 4th jet information. The central jet veto ensures that any difference is not due to information on the existence of additional jets.

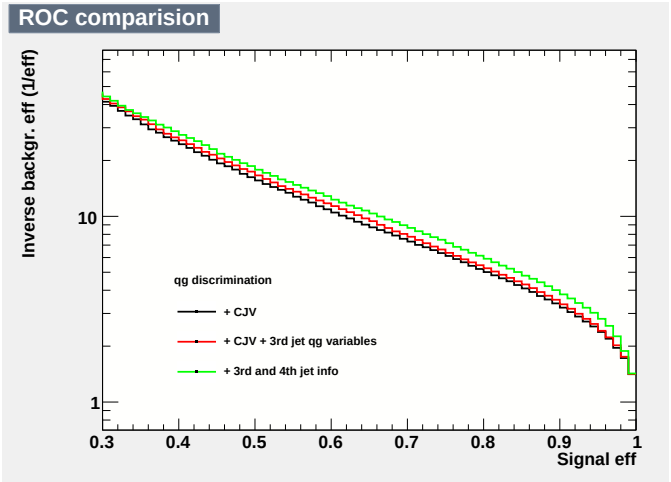
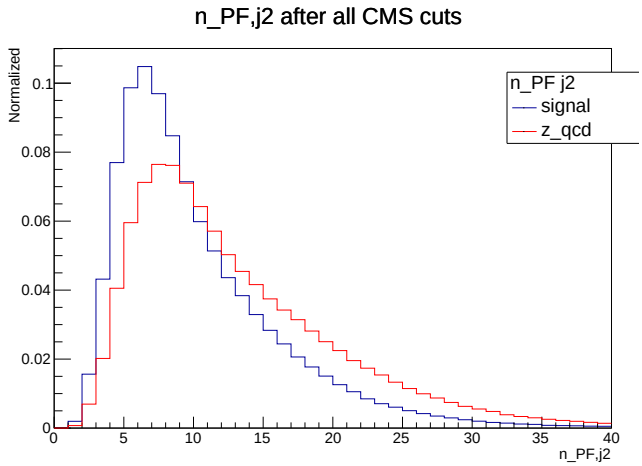
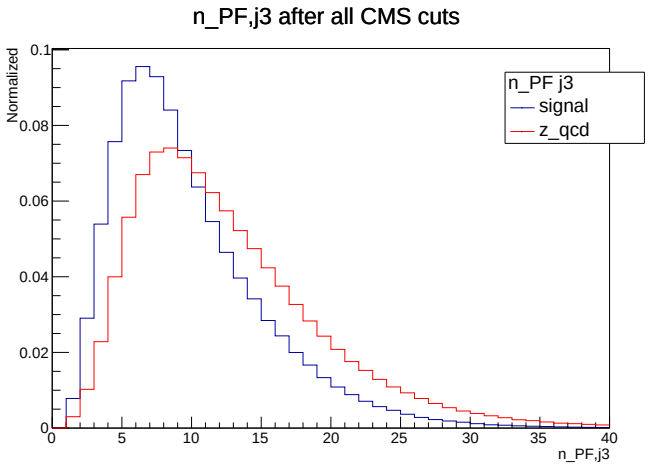


Figure 4.11: ROC curves of BDT with additional CJV, additional CJV and 3rd jet quark gluon variables and full 3rd and 4th jet info.



(a) Distribution of $n_{PF,2}$ for signal and Z QCD background after CMS preselection cuts.



(b) Distribution of $n_{PF,3}$ for signal and Z QCD background after CMS preselection cuts.

Figure 4.12: Distributions of n_{PF} for the second and third jet. The third jet distribution of the signal sample looks quark like.

The addition of quark gluon variables increase the BDT performance, out of them $n_{PF,3}$ is the highest ranked. This is in so far nice as it is the easiest to measure.

Something very interesting can be seen in the distribution of $n_{PF,3}$ for signal and background shown in Figure 4.12 using the Z-QCD background. As we expect the signal to have two outgoing quarks an additional third jet would result from a radiated gluon. However the distributions look very alike, that is the third jet appears quark like. A possible explanation is that in contrast to the naive expectation it is not unlikely for the radiated gluon to have a higher p_T than one of the tagging quarks and therefore induce a tagging jet itself. Alternatively it may be that the radiated gluon splits less than expected and therefore shows quark like distributions. As this seems to contradict our basic understanding of quark gluon discrimination in may be worthwhile to investigate this issue in a further analysis.

To summarize our findings we can say that quark gluon discrimination is an effective way of separating signal and background. When the quark gluon variables make use of the first and second jet they do not provide much additional information to the basic variables already used in the BDT but can replace some of them when they are easier to obtain. With the current preselection cuts of CMS n_{PF} appears to be the best quark gluon variable. We also saw that quark gluon variables for a third jet may increase the performance of the BDT and can be a good substitute in case full further jet information is not available.

5 Conclusion and Outlook

We investigated Higgs production through Weak Boson Fusion in an attempt to constrain invisible Higgs decays. We saw that the task of separating signal from background events is very well suited for Boosted decision trees which can improve cut flow approaches. To understand this machine learning algorithm and make best use of it various settings of the TMVA implementation were tested. It was seen that there is no need to use more than a few hundred trees in one BDT and that they are overall very stable. Furthermore the weak classifier approach of BDTs allows us to ensure proper statistics when a large enough sample is provided.

As an attempt to increase the separation power of Boosted decision trees quark gluon discrimination was investigated. We saw that the proposed quark gluon variables show different distributions for signal and background and therefore can contribute to the classifier problem but we also realized that the current BDT variables may already over constrain this problem so an increase in performance requires truly additional information. This is given by quark gluon variables for higher jets although they can not substitute the full information on 3rd and 4th jet.

It was noted that the n_{PF} distribution for a 3rd jet in the signal sample looks quark like but a gluon like distribution was expected. This hints that the process of quark gluon discrimination is not fully understood and additional research may help in improving the usage of quark gluon variables.

An alternative attempt to increase the BDT performance would be to test various trigger settings. As noted above it seems difficult to add useful variables with the information already provided so an approach to optimize the usage of given information seems natural.

References

[1] Chatrchyan, Serguei, et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC." Physics Letters B 716.1 (2012): 30-61.

[2] Carlson, Carl E. "The proton radius puzzle." Progress in Particle and Nuclear Physics 82 (2015): 59-77.

[3] Wilczek, Frank. "Beyond the standard model: An Answer and twenty questions." arXiv preprint hep-ph/9802400 (1998).

[4] He, Xiao-Gang, and Jusak Tandean. "Low-mass dark-matter hint from CDMS II, Higgs boson at the LHC, and darkon models." Physical Review D 88.1 (2013): 013020.

[5] Eboli, Oscar JP, and Dieter Zeppenfeld. "Observing an invisible Higgs boson." Physics Letters B 495.1 (2000): 147-154.

[6] Hoecker, A., Speckmayer, P., Stelzer, J., Therhaag, J., von Toerne, E., Voss, H. (2007). TMVA 4. arXiv preprint physics/0703039.

[7] Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." Icml. Vol. 96. 1996.

[8] Gallicchio, Jason, and Matthew D. Schwartz. "Quark and Gluon Tagging at the LHC." Physical review letters 107.17 (2011): 172001.

[9] CMS collaboration. "Performance of quark/gluon discrimination in 8 TeV pp data." CMS Physics Analysis Summary CMS-PAS-JME-13-002, CERN (2013).

[10] Banfi, Andrea, Gavin P. Salam, and Giulia Zanderighi. "Infrared-safe definition of jet flavour." The European Physical Journal C-Particles and Fields 47.1 (2006): 113-124.

[11] The ATLAS collaboration, "Discrimination of Light Quark and Gluon Jets in pp collisions at s=8 TeV with the ATLAS Detector", ATLAS-CONF-2016-034

[12] A. J. Larkoski, G. P. Salam and J. Thaler, Energy Correlation Functions for Jet Substructure, JHEP 1306 (2013) 108, arXiv: 1305.0007 [hep-ph].

[13] https://sherpa.hepforge.org/doc/SHERPA-MC-2.2.4.html

[14] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, M. Selvaggi
"DELPHES 3, A modular framework for fast simulation of a generic collider experiment" arXiv:1307.6346 [hep-ex]

[15] Rene Brun and Fons Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. Meth. in Phys. Res. A 389 (1997) 81-86. See also http://root.cern.ch/.

[16] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," PoS A CAT 040 (2007) [physics/0703039].

[17] The CMS collaboration, Khachatryan, V., Sirunyan, A.M. et al. J. High Energ. Phys. (2017) 2017: 135. https://doi.org/10.1007/JHEP02(2017)135

[18] Thorsten Ohl
"Drawing Feynman Diagrams with LaTeX and Metafont"
arXiv:hep-ph/9505351

Acknowledgements

I would like to thank Prof. Tilman Plehn for immediately offering me a wide range of interesting topics as my Bachelor project. Special thanks must go to Anke Biekötter who never failed to find a way when things seemed to be stuck. I also owe thanks to Jennifer Thompsons and Rhea Moutafis for introducing me to the topic of WBF with great enthusiasm as well as the whole group for offering a nice place to work.

Erklärung

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

Heidelberg, den 13.10.2017

Fabian Keilbach