CERN SUMMER STUDENT PROJECT

Boosting Impact Parameter based B-Tagging at High Transverse Momentum for the ATLAS Experiment

Ville Pyykkönen

Supervised by
Dr. Andrea COCCARO

September 18, 2016

Contents

1 Foreword

This report describes the studies that I carried out as a CERN Summer Student project working for the ATLAS experiment in the Summer 2016. I worked under the supervision of Andrea Coccaro who works at the University of Geneva.

I would like to thank the CERN Summer Student Program for giving this delightful opportunity to spend the summer working at the CERN, joining the leading research in the particle physics, one of the greatest adventures of the mankind to delve to the deepest layers of the universe.

I give special thanks to my supervisor, Andrea Coccaro for giving me chance to work with b-tagging algorithms and having time for me while his schedule was booked quite full. Andrea showed me that it is all right to ask questions even if they do not show some prolific understanding of the matter at hand but just the opposite.

Working at the CERN is about international collaboration and relationships. Therefore, a summer at CERN wouldn't be anything without meeting new friends. I would like to thank the group of people that arrived to CERN at the same time as I did, collectively know as the Cool Summies. Especially, I would like to thank Håkon and Kalle for nice conversations.

A man has to have a roof above his head even when it comes to working long hours at CERN. Housing is especially comfortable with such a group of Finns I lived and shared the daily experiences with. I would like to thank Mikael, Jyri, Toni and Leevi for many thorough conversations and support during the summer. Especially, I thank Toni for sharing room, nice guitar jams (bearing my shredding) and many thoughts during the summer, at home and at work.

I guess one has to see world to appreciate the nature's diversity and its deep imagination (far deeper than that of human beings.) I would like to thank Jan for driving me around the central Europe and openness for ideas, Toni for photographing, Jussi for the jokes and Panu for keeping me down to earth. I would specially like to thank Laura for organizing many nice trips and asking me to come along. Otherwise, I would have spent many more inefficient hours at the office or playing video games. Her appetite for adventure and seeing the world is something I wish I could have as well.

Finally, life without background support would be on very unstable basis. I would like to thank my family: Eeva, Ari and Kalle for constant support for my doings even though they do not understand what I'm doing, giving wise advice in life and sharing couple of adventurous days around the lake Geneva and Chamonix during the Summer. I would like to thank my aunt Raija for giving me the book that gave me the enthusiasm all those years ago to find out more about this world and science; it is probable that I would have never been at CERN without her. Last, I thank my flatmate in Finland, Tuomo for relaxing moments in the Middle Earth during the Summer.

Ville Pyykkönen

2 Introduction

The bottom (b) quark has prominent role in the physics program of the ATLAS Collaboration. Importantly, b quarks are present in many significant decay channels of the Standard Model [1, 2]. For instance, b quarks are part of the decay chain of the Higgs boson with the largest branching ratio. Also, mesons including a b quark, the B mesons provide an important way of studying the CP violation. Furthermore, b quarks also have significance in the searches of physics Beyond the Standard Model (BSM): many BSM models predict particles that couple to b quarks [3, 4].

To use b quarks in an analysis, the experimentalist must identify the b-jets, the collimated sprays of particles coming from the processes initiated by a b-quark. The route from a b quark to a b-jet contains fragmentation of quarks to hadrons and the subsequent decays of b-hadrons.

The identification of b-jets is called b-tagging. The b-tagging is made possible by unique features of b quarks that are inherited by the b-hadrons, composite particles that b quarks form with other quarks [1, 2]. The b-hadrons have sufficiently long lifetime so that they travel a measurable distance in the detector before decaying. Thus the secondary and tertiary vertices, that is the points where the subsequent decays of the inclusive chain $b \rightarrow c \rightarrow X$ happen, can be reconstructed inside the tracker, the instrument measuring the trajectories of charged particles. On the other hand, the lifetime is longer than the charm hadron lifetime providing clear separation in this end as well. Another useful feature is the higher mass of b-hadrons compared to the other hadrons composed of light quarks. The high mass difference is seen as high transverse momentum of the resulting jet.

As a multipurpose detector at the LHC, the ATLAS Experiment has great interest in identifying the b-jets. To do b-tagging, the ATLAS Collaboration has developed algorithms for b-tagging using the presented special features of b-hadrons. In this report, we are focusing in the lifetime based algorithms. In this category, ATLAS Collaboration presently uses three different algorithms: IP3D, SV, and JetFitter [2]. Of these, IP3D uses the large average impact parameter of tracks in a b jet. SV makes use of the reconstruction of secondary vertex. JetFitter uses the reconstruction the decay chain in a jet. The outputs of these algorithms are combined by multivariate analysis based MV2 algorithm that uses a boosted decision tree [2].

The present b-tagging algorithms used in ATLAS physics program are suboptimal at high transverse momentum, the component of momentum transverse to the beam direction in the ATLAS experiment[2]. There are many possible reasons causing this. In this paper, we give a hypothesis for a possible cause and present two ideas to improve b-tagging algorithms based on this hypothesis. The paper is constructed in the following way. In the theory section we introduce the bottom flavoured jets with their peculiar properties, the lifetime based b-tagging method: IP3D that is used as framework for the presented analysis. We introduce the concepts and tools of b-tagging performance evaluation. Further, to motivate this study, in the theory section we describe the sub-optimality of present algorithms more carefully and give a hypothesis about one source of the sub-optimality. In the methods part we present the tools that are used for studying b-tagging performance: simulated data samples and the default jet and track selection schemes. We then proceed to give two possible ways to improve IP3D In the Results and the Discussion section we present the results and discuss them and the future.

3 Theory and Motivation

In this section, we first present some general notions and the choice of units. Secondly, we present the essential properties of b quark and its hadrons. Based on the long lifetime of b quarks, we present the IP3D algorithm for b-tagging. Further, we describe the performance studying concepts and tools, namely the visualisations that present the performance and can be used to compare different methods. Finally, using the introduced methods, we discuss the sub-optimality of present b-tagging algorithms at high transverse momentum and propose a possible source, setting the

More background on the strong interaction, jets and the CKM matrix referenced in this section can be found in the appendix. We strongly recommend the reader to read it if the material is not familiar.

3.1 Units and definitions

In this report, we use the **natural units** in which the speed of light c and the reduced Planck's constant $\hbar$ are $c = \hbar = 1$. These units are especially comfortable since these constants would otherwise appear in most of the equations.

We use the **ATLAS coordinate system** [5]. The interaction point in the detector is defined as the origin. The z-axis runs along the beam line, positive direction being anti-clockwise direction when watching from above the LHC ring. The $x - y$ plane is perpendicular to the beam line and is called **transverse plane**. We define positive x-axis pointing towards the center of the LHC ring and the y-axis pointing upwards. We also adapt polar coordinates in the transverse plane (r, ϕ) where r is the **transverse distance** from the origin and ϕ is the right-handed **azimuthal angle** from x-axis. We also use the **polar angle** θ which is the angle with respect to the positive z-axis.

We adapt a typical choice for kinematic parameters: the **transverse momentum** p_T which is the component of the momentum vector in the transverse plane, and the **pseudorapidity** η defined using the polar angle θ as $\eta = -\ln \tan(\theta/2)$. The pseudorapidity η is an alternative measure for the polar angle. When considering solid angle we usually refer to **distance in** $\eta - \phi$ **space** $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

We often talk about **Monte Carlo (MC)** methods and **event** generation based on them. It is sufficient here to think MC methods as synonymous to simulations. In practice, the MC methods utilize repeating some process to calculate integrals, get distributions of quantitative effects in various situations and so forth. Simulation tasks in particle physics usually have quite many degrees of freedom so the MC methods, as simpleminded they seem in low dimensions, are the only choice since they do not suffer from " the curse of dimensions" , that is the performance is relatively good also for many dimensions.

Here by an **event** we mean the result of a bunch crossing of proton beams in LHC in the detector.

3.2 Bottom flavoured jets

Bottom flavoured jets, or b-jets for short, are jets which have been initiated by a b-quark fragmentation They have many unique properties that make possible their identification.

First of all, b quarks have relatively high mass so a good fraction of the initial momentum of the quark is transfered to the B-hadron. Secondly, the CKM matrix elements related to b are [6] $|V_{tb}| = 0.9991$, $|V_{cb}| = 0.041$, and $|V_{ub}| = 0.0035$ so we see that almost all t quarks decay into b quarks but b quarks have relatively low amplitudes to decay further. Importantly, even though b-quarks, and thus its hadrons, have relatively high mass, their lifetime is comparatively long. Their proper lifetime is ~ 1.5 ps [6]. The proper mean distance of travel of approximately $c\tau = 0.45$ mm is rather long and can be detected in a detector. The collision point, that is the primary vertex and the first decay point, that is secondary vertex and furthermore the tertiary vertex can be measured by the ATLAS tracker [1, 2].

3.3 IP3D Algorithm for b-tagging

Impact parameter is by definition the shortest distance from the line of a track to the primary vertex. As events including b-hadrons have a secondary vertex clearly separated from the primary vertex and the b jets are arising from the secondary vertex, the impact parameters of tracks in these b jets are relatively large. This is illustrated in the left figure 1. This fact is used in the Impact Parameter 3D (IP3D) algorithm [1, 2]. The IP3D algorithm uses signed impact parameter significances. First of all, a sign is given to impact parameter according to the right figure in 1. If the point of a track of the closest approach to the primary vertex is in the front (back) of the plane perpendicular to the line representing the jet direction at the primary vertex, the track is given a positive (negative) sign as for track 1 (track 2) in the figure.

Furthermore, we want to add information about the measurement accuracy. We use the signed impact parameter significances which are defined as

$$d_{0,\mathrm{sig}} = \frac{d_0}{\sigma_{d_0}} \tag{1}$$

where d_0 is the signed impact parameter and σ_{d_0} is the standard deviation of the signed impact parameter.

To take the geometry of the detector into account, we consider the signed impact parameter significance in two perpendicular components, the transverse $d_{0,\mathrm{sig}}$ and longitudinal $z_{0,\mathrm{sig}}$.

Now that we have defined the input parameters of IP3D algorithm, the signed impact parameter significances of the tracks in a jet, we describe the algorithm itself. First, a training set of jets with truth information about their flavours is given to the

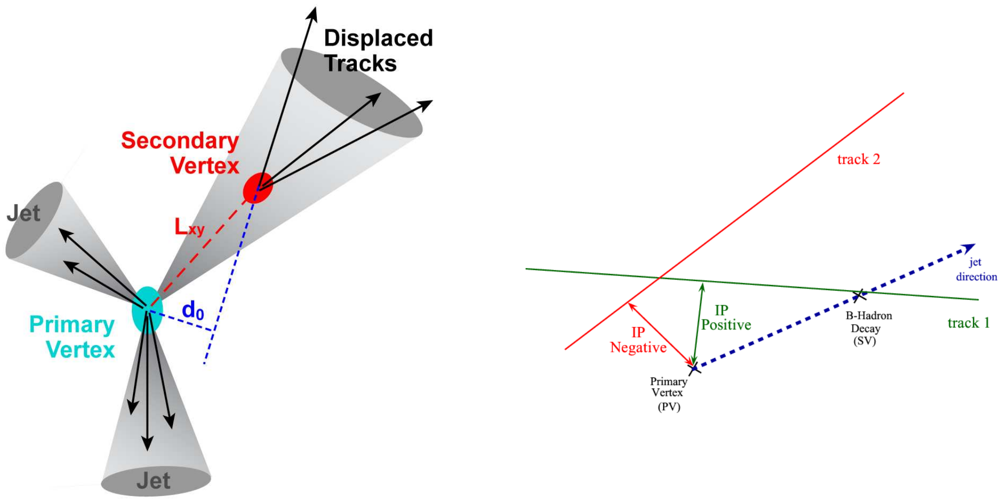

Figure 1: (Left figure) shows an event of three jets in which the secondary vertex is clearly separated from the primary vertex. Tracks coming from the secondary vertex are from the decay there and are clustered as one jet. Here the impact parameter is defined as the shortest distance from the line of a track to the primary vertex. In the figure, d_0 is the impact parameter and L_{xy} is the distance form the primary vertex to the secondary vertex. Figure is taken from [7]. (Right figure) Shows an event with primary vertex, secondary vertex and two tracks is shown. The figure defines the sign of impact parameter. Let's consider the jet that is indicated as blue dashed line arrow. If the point of a track of the closest approach to the primary vertex is in the front of the line perpendicular to the jet at primary vertex, the track is given a positive sign (as for track 1) and vice versa if it is in the back (track 2). Figure is taken from [8]

.

algorithm to get the probability distribution functions for tracks from different kind of jets. Further, tracks are divided into different categories based on their quality. Secondly, when evaluating the algorithm it gives a discriminating number to each jet in the following manner. Using the probability distribution functions, the algorithm calculates likelihood for a track to be in a b-jet P_b and from lighter jets P_l. The IP3D parameter is calculated using log-likelihood ratio method

$$\text{IP3D} = \sum_{\text{tracks}} \log\left(\frac{P_b}{P_l}\right) . \tag{2}$$

The probability distribution functions of $d_{0,sig}$ for different types of jets are shown in the left figure of 2 and the resulting IP3D parameter distribution is shown in the right figure of 2. Some discrimination power is obtained by the track distribution solely but for the IP3D parameter clear discrimination is obtained even moderately for moderately large value. This motivates the algorithm.

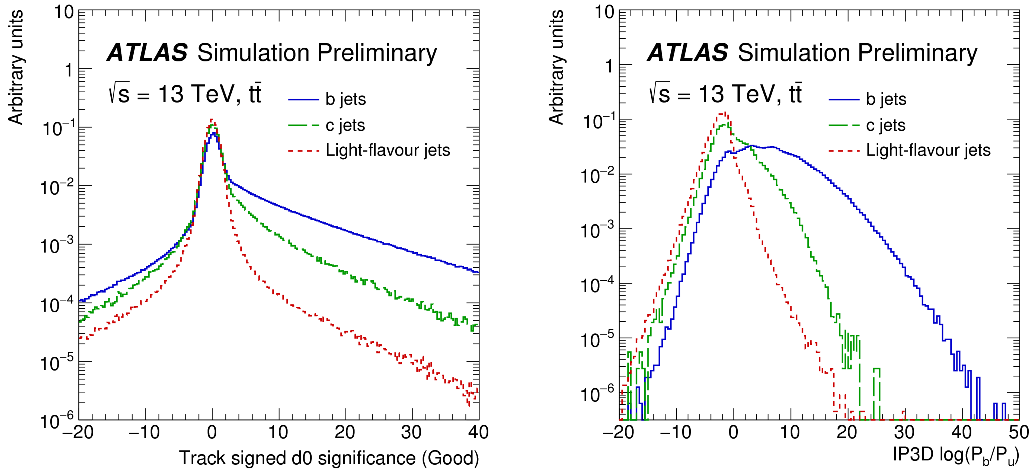

Figure 2: The signed transverse impact parameter significance distribution for tracks (left) and the IP3D output distribution for jets (right) for different types of jets. It is seen that the signed impact parameter significance gives some discrimination and the combined algorithm IP3D gives even more which tells that the algorithm works properly. The figures are taken from [1].

3.4 Concepts of performance studies

A b-tagging algorithm tries to separate b-jets from other jets. In terms of multivariate analysis, b-jets are thus the signal and other jets are background. The task is to separate signal from the background. The basic concepts to study the performance of an algorithms are signal efficiency and background rejection. Signal efficiency is defined as the portion of the signal that the algorithm tags correctly. Background rejection is defined as the inverse of the portion of background that algorithm tags as signal. Inverse is taken so that the number is larger when the number of rejected background is larger. These may be calculated if the distribution of algorithm output is known for the signal and background separately. The output distribution for MV2 algorithm is shown in the figure 3. A cut value in the distribution must be given to algorithm: it accepts all the jets that have value above this point and discards all below. The b-jet efficiency is calculated as the ratio of the area above the given cut value and the total area in the b jet distribution.

$$\text{b-jet efficiency} = \frac{P_{b,MV2>c}}{P_{b,MV2,tot}} \tag{3}$$

where c is the cut value and P indicates the probability for a jet being in the indicated region. Likewise, the light jet (up,down, and strange) rejection is calculated by similar manner from the distribution for these jets and taking inverse.

$$\text{light-flavour jet rejection} = \frac{P_{l,MV2,tot}}{P_{l,MV2>c}} \,. \tag{4}$$

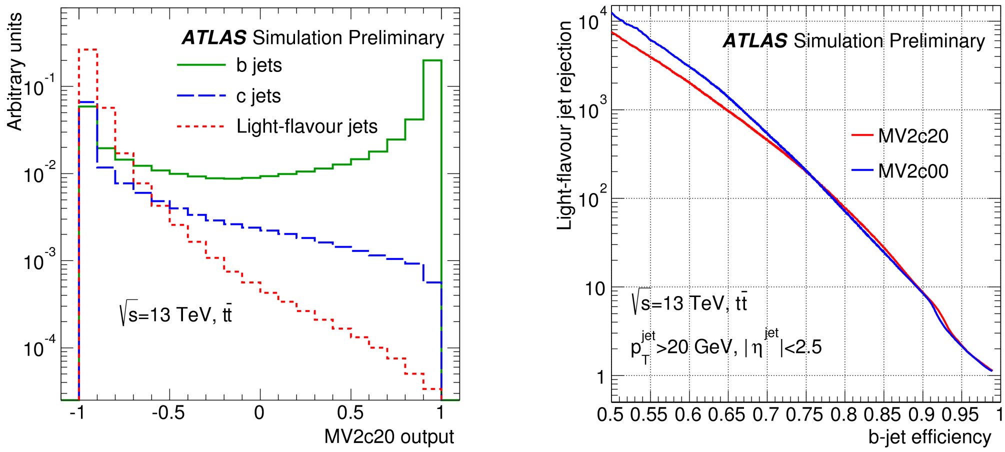

Figure 3: (Left figure)The MV2c20 output distribution for the b-jets (signal), the c-jets (background) and the light jets (background). Clear discrimination is shown. (Right figure) The ROC curve for MV2c20 and MV2c00 The difference between these algorithms is that the training sample for MV2c20 included 20 % of c-jets while for MV2c00 all the jets were light jets. Here it can be seen that MV2c00 handles light-flavour jets a bit better than MV2c20. The figures are taken from [1].

Furthermore, since charm jets have different features compared to light jets, they are considered as separate background and charm jet rejection is defined likewise. However, in this study we consider charm jets as the same background as other light jets.

Since the optimization of performance is about having as large background rejection while maintaining large signal efficiency, the ROC curve is of interest. The ROC curve is defined as background rejection as function of signal efficiency. A good algorithm offers a nice trade off: steeper the ROC curve, the better. The curves for the current MV2 algorithms MV2c20 and MV2c00 are presented in the right figure in 3. The ROC curve is calculated using the distribution of the algorithm output for different kind of jets, and calculating b-jet efficiencies and light-flavour jet rejections for various cut values and drawing the results. Thus, the curve is parametrized by the cut value: greater the b-jet efficiency, smaller the cut value.

As we are interested in the performance of algorithms in high transverse momentum (p_T) range, we are additionally interested in b-jet efficiency and light jet rejection vs. p_T. The b-jet efficiency vs. p_T curve is calculated by taking a cut value and evaluation algorithm output values for different p_T ranges. Light jet rejection vs. p_T is, on the other hand, evaluated with a constant b-jet efficiency.

3.5 Suboptimal performance of present algorithms at high p_T

We have introduced the necessary concepts to understand the basic of b-tagging studies in the past section. In this section, we present a problem in the present b-tagging

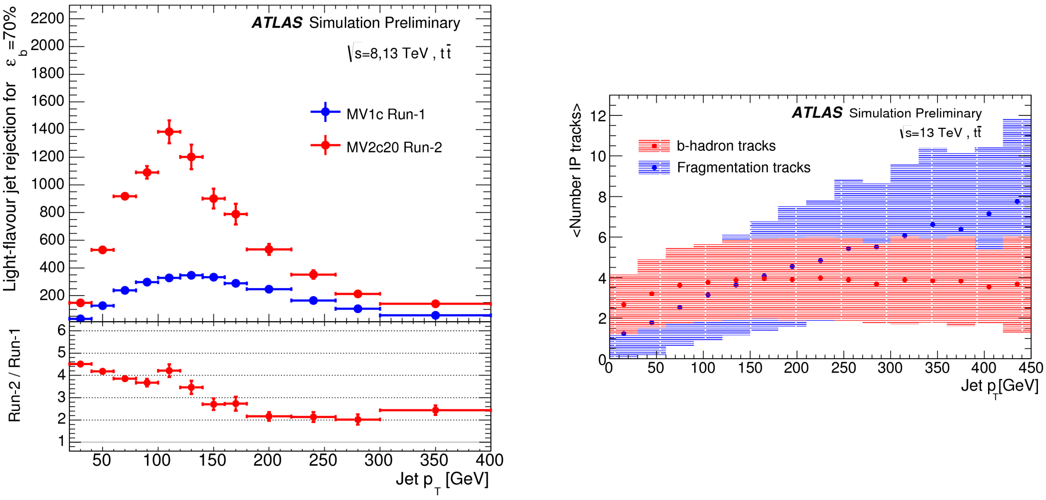

Figure 4: The performance of ATLAS IP3D algorithm as function of p_T is shown in left figure. It is seen that the light-flavour jet rejection is substantially lower in p_T range above 150 GeV compared to region around 100 GeV. This hinders the overall performance of the algorithm. The average number of tracks in a b-jet coming from b-hadrons and fragmentation is shown in the right figure. Fragmentation tracks mean the tracks of particles coming from the initial fragmentation process. The portion of fragmentation tracks in a b-jet is significantly larger than the portion of tracks from b-hadron. This may have connection with the low performance of IP3D at high p_T as the performance is low in the region where the portion of fragmentation tracks is substantially greater. Figures are taken from [2].

algorithms and a hypothesis for a possible cause. At the present state the performance of b-tagging algorithms is suboptimal at high p_T range, above 150 GeV [2] This can be seen in the figure 4. The light-flavour jet rejection is significantly lower above 150 GeV than below it. Therefore, the overall performance of the algorithm is worse than would be optimal. Moreover, interesting physics may lie in the high p_T region so having good performance there is of essence.

The ATLAS Note [2] indicates a possible source of the sub-optimality. The source is the large quantity of fragmentation particles clustered to a b-jet at high p_T, partly covering the b-hadron tracks and confusing the algorithms. This is shown in the right figure in 4. Thus b-jets of high p_T look more like a background jet to the algorithm meaning that light flavour rejection is worse at same b-jet efficiency. In this paper, we investigate ideas of making the jets more pure by discarding fragmentation tracks by more sophisticated track selection schemes.

4 Methods

In the frame set by the Theory and Motivation chapter in this chapter we present the methods of studying performance and trying to solve the presented challenge in impact parameter based b-tagging at high p_T.

4.1 Monte Carlo samples, default jet and track selection

In the study presented in this paper, we used Monte Carlo simulated $t\bar{t}$ events corresponding to $\sqrt{s} = 13$ TeV proton-proton collisions at LHC. We use the same generation as the paper [2]. View it for more details; here we present the generic characteristics. The events are simulated with Powheg interfaced with Pythia6 for the parton shower and hadronization using the CT10 parton distribution function set. The decays of b- and c-hadrons are simulated by EvtGen. Additional interactions and pile-up of 20 extra hard scattering collisions are added using Pythia8. The detector simulation of ATLAS is made using GEANT4. The ATLAS experiment is reviewed in [5].

Jets are clustered from calorimeter clusters using anti-k_t algorithm [9] with distance parameter $R = 0.4$. The calorimeter clusters are calibrated at the EM-scale and hadronic scale. Only jets with $p_T > 20$ GeV and $|\eta| < 2.5$ are used. In addition, Jet Vertex Tagger (JVT) algorithm is used [10]. The JVT algorithm combines track parameter of a jet in a multivariate analysis in order to determine whether the jet is pile-up (signal) or not (background). Demand that JVT score is less than 0.59 is used.

In training and performance studies jets are given the flavour label by the following procedure using the truth information. Jets are matched to b- and c-hadrons and τ leptons within a cone of radius $\Delta R = 0.3$ in solid angle around the jet axis. It is required that b- and c-hadron have p_T larger than 5 GeV. The procedure first searches for b-hadron within the cone. In the case a b-hadron is found, the jet is labeled as b-jet and if none is found, algorithm is repeated for c-hadrons. When a c-hadron is found, the jet is labeled as c-jet. If no c-hadrons are found, the search is repeated for τ leptons. In the case that a τ is found, the jet is labeled as τ-jet. If not, the jet is labeled as a light flavour jet. Further, to discard possible overlap between jets and leptons a jet is removed if it overlaps with a truth electron or muon from W/Z boson decays inside the considered cone of $\Delta R = 0.3$.

For IP3D a track selection scheme is applied [2]. Tracks are associated to jets using p_T dependent criteria on the ΔR between track and jet axis. The ΔR is demanded to be wider for low p_T jets and narrower for high p_T jets since high p_T jets are more collimated physically. At 20 GeV it is demanded that the tracks are within 0.45 and at 150 GeV the ΔR cut is 0.26. It is also demanded that track p_T is above 1 GeV, $|d_0| < 1$ mm and $|z_0 \sin\theta| < 1.5$ mm, and that there are seven or more silicon hits in the tracker.

4.2 Track selection schemes to improve IP3D at high p_T

Our hypothesis is that the suboptimal performance of IP3D at p_T higher than 150 GeV is caused by additional fragmentation tracks confusing the algorithm. Fragmentation tracks means the trajectories of particles coming from fragmentation of quarks at primary

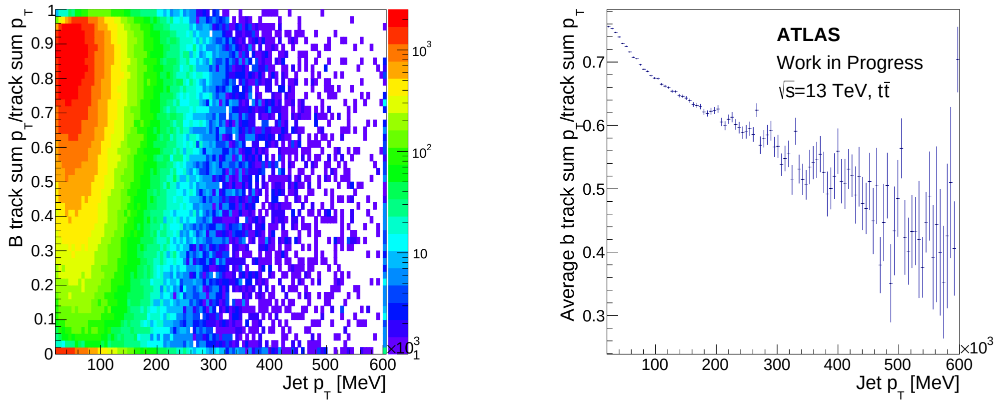

Figure 5: The b-track sum p_T/track sum p_T distribution (left) and average (right) vs. jet p_T of b-jets. It is seen that the portion of p_T sum carried by b-tracks is smaller at high p_T than at low p_T as is expected. Also, the amount of jets at high p_T is relatively low. It was suggested in [2] that a Z' MC sample would provide more jets in this region as well.

vertex. As shown by right figure in 4, the fragmentation tracks are significant source of tracks: after about 150 GeV their multiplicity is above that of b-hadron decay product tracks. Based on this hypothesis, we present two possible approaches to try to boost the performance of the IP3D at high p_T. Both approaches try to do tracks selection within tracks clustered to jets to discard the fragmentation tracks to make the b-jets more pure.

4.2.1 Track selection based on fraction of b-hadron track p_T of a b-jet

It is know that the portion of p_T of a b-jet carried by b-hadron decay products, namely b-tracks, varies as function of p_T because the amount of tracks from fragmentation varies. Also we know that b-hadron decay product tracks have higher signed impact parameter significances in average than the fragmentation tracks. We can use these fact to select tracks that are from b-hadrons decays.

The sum of p_T of the tracks from b-hadrons versus the sum of p_T of all tracks in b-jets is determined using truth information provided by the MC samples. The b-track sum p_T/track sum p_T distribution and the average versus jet p_T are shown in the figures 5. It is seen, as expected at the basis of the right figure of the 4, that the average jet p_T fraction of the b-tracks is less at higher p_T. Also, the amount of jets at high p_T is seen to be insufficient. It was suggested in [2] that choosing Z' MC sample instead of $t\bar{t}$ sample would provide more jets in high p_T region.

Using the presented features, we use the following track selection scheme for jets:

1. Apply the default track selection
2. Order remaining tracks in a jet in descending impact parameter significance order
3. From the ordered tracks, choose n tracks so that the their sum p_T corresponds to the fraction of p_T carried by b-tracks in average.

To motivate the second step, as summing track p_T is a discrete operation, we have to decide how near we want the sum p_T to be to the average. When the p_T fraction of the tracks chosen by the scheme goes over the average portion, we must decide whether the total fraction is more or less than the average. Ideally we do not want to lose any b hadron tracks and as we see from left figure 5, the p_T portion of b-tracks varies rather much at a chosen jet p_T level. Judging from the relative multiplicities of different types of tracks, losing b-tracks has more severe consequences to the performance than having the same amount of excess fragmentation tracks. Thus, we used the following two schemes for choosing n:

1. Choose all tracks until the track sum p_T goes over the average and choose the track going over if the resulting track sum p_T is closer to the average than without or otherwise discard it
2. Choose all tracks until the track sum p_T goes over the average and also the one going over.

4.2.2 Track selection based on MVA on track parameters

Another approach to select only the b-tracks is to use Multivariate analysis techniques to select the tracks. We used a Boosted Decision Tree (BDT) as implemented in TMVA package [11] . We used the Gradient boosting with bagging fraction 0.50 and shrinking 1, with 1000 trees. In addition, maximum depth of a tree was set to 2 and the number of steps during node cut optimisation was set to 20.

In the analysis we considered the b-tracks to be the signal. B-tracks include also the tracks clustered to jet which come from inclusive $b \to c \to X$ decay. The track parameters inserted to the BDT are

1. Track p_T
2. Track p_T/jet p_T
3. Track $d_{0,sig}$
4. Track $z_{0,sig}$
5. Track η/jet η
6. Track ϕ/jet ϕ .

5 Results

We implemented the presented track selection schemes using MC generated events described in the Method section. The results are gathered in this section. We discuss the results in the Discussions and Outlook section.

5.1 B-track p_T fraction based track selection

We implemented the scheme and carried out the analysis for 200000 events. The ROC curve for b-track p_T fraction based track selection scheme 1 (in the numbering presented in the Methods section) is shown in the figure 6. In the figure it is compared to the standard track selection output. The b-jet efficiency vs. p_T for the cut value IP3D = 0 and the light-flavour jet rejection vs. p_T for constant b-jet efficiency of 0.77 for the scheme 1 is presented in the figure 7.

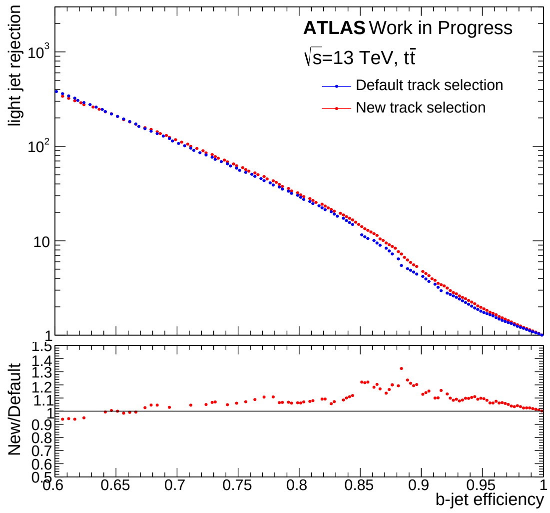

Figure 6: The ROC curve for p_T fraction based track selection scheme 1 and for the default track selection.

Similarly, the ROC curve for b-track p_T fraction based track selection scheme 2 (in the numbering presented in the Methods section) is shown in the figure 8. In the figure it is compared to the standard track selection output. The b-jet efficiency vs. p_T for the cut value IP3D = 0 and the light-flavour jet rejection vs. p_T for constant b-jet efficiency of 0.77 for the scheme 2 is presented in the figure 9.

5.2 The MVA based track selection for IP3D

We implemented the training of the Boosted Decision Tree with parameters mentioned in the Methods section. The resulting output distribution is shown in the figure 10 .

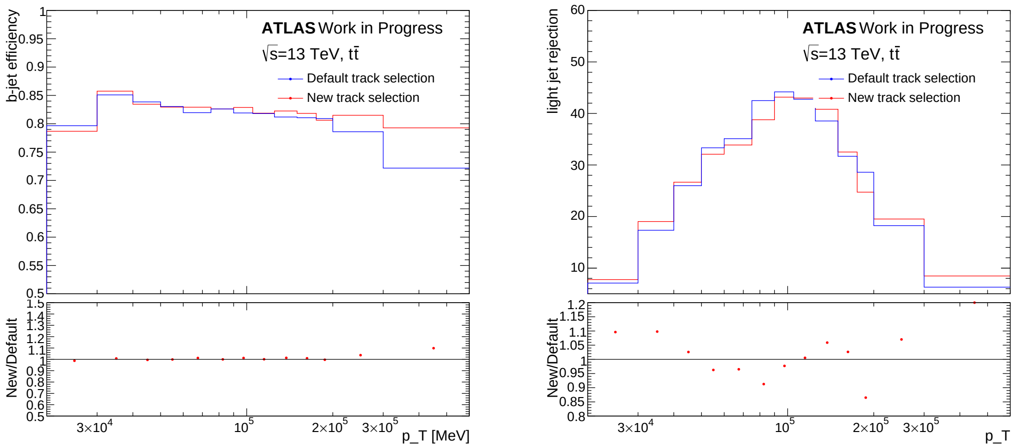

Figure 7: (Left figure) b-jet efficiency vs p_T for jets cut below value IP3D = 0 for p_T fraction based track selection scheme 1 and for the default selection. (Right figure) light-flavour jet rejection vs p_T for constant b-jet efficiency of 0.77 for p_T fraction based track selection scheme 1 and for the default selection

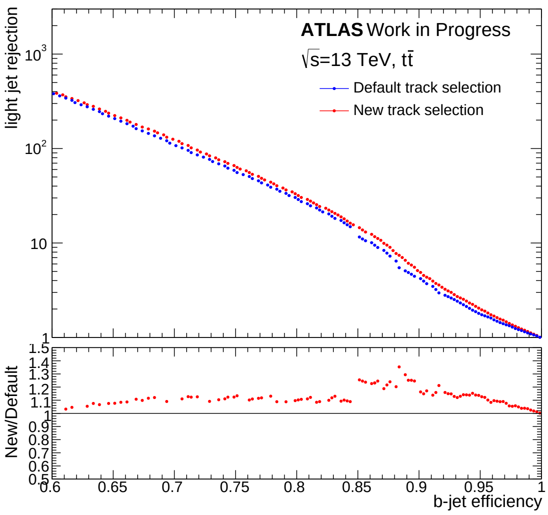

Figure 8: The ROC curve for p_T fraction based track selection scheme 2 and for the default track selection.

We applied the Boosted Decision Tree decision for track selection with criteria for tracks that the output must be over 0.5. The resulting ROC curve is presented in figure 11.

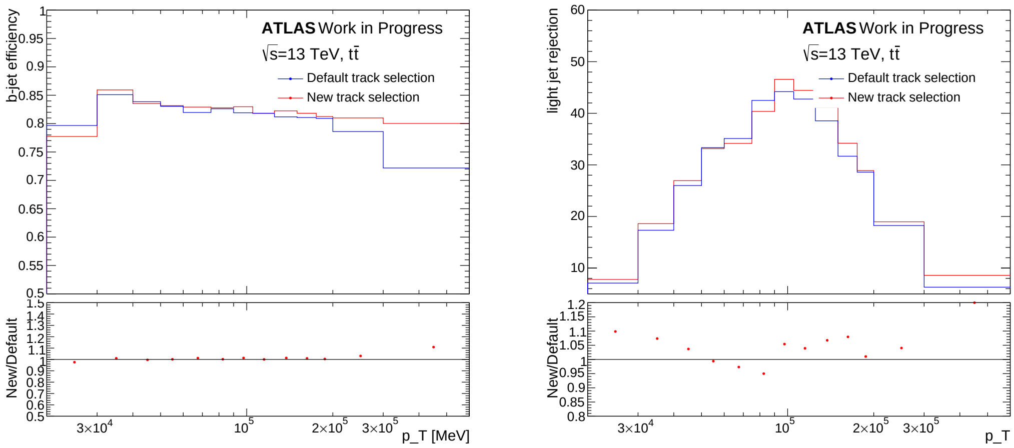

Figure 9: (Left figure) b-jet efficiency vs p_T for jets cut below value IP3D = 0 for p_T fraction based track selection scheme 2 and for the default selection. (Right figure) light-flavour jet rejection vs p_T for constant b-jet efficiency of 0.77 for p_T fraction based track selection scheme 2 and for the default selection

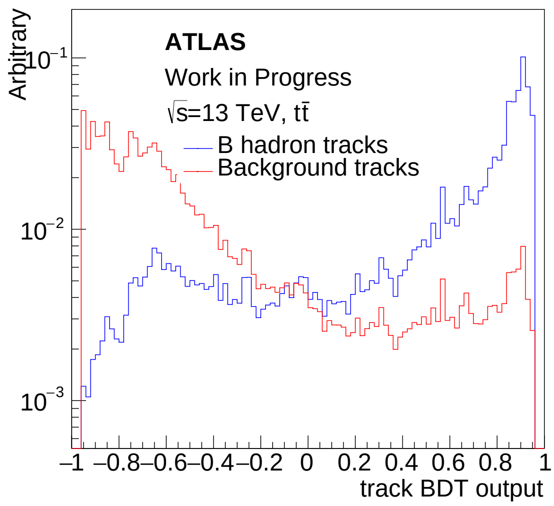

Figure 10: The output value distribution of the BDT used for track selection in MVA based track selection scheme. Clear discriminative power is seen.

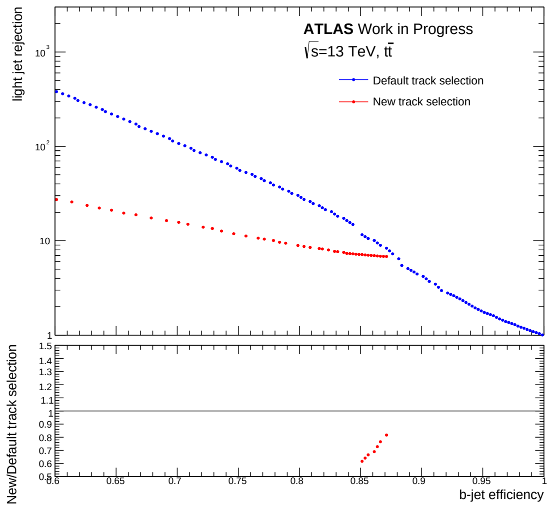

Figure 11: The ROC Curve for IP3D with track selection based on the Boosted Decision Tree demanding that tracks must have the output parameter over 0.5.

6 Discussions and Outlook

We saw in the Theory section that default b-tagging algorithms are sub-optimal in the region of $p_T > 150$ GeV. We presented a hypothesis that this is caused by the increase of the fraction of fragmentation tracks clustered to b-jets. Working based on this hypothesis, we developed two track selection schemes to discard the fragmentation tracks and thus make b-jets more pure. The first is based on two facts: the tracks from b-hadron decays have large impact parameter significance and the fraction of p_T of these tracks can be determined and varies with jet p_T. The second in based on Boosted Decision Tree algorithm using kinematic properties of tracks to separate b-tracks from fragmentation tracks.

The results are shown in the Results section. For the first scheme, with both track selection variates we see improvement as expected compared to the default track selection. The performance is increased ~ 10 % in terms of ROC curve, 5 %-unit in b-jet efficiency and 20 % in light-flavour jet rejection. The improvement for the more loose track selection scheme, 2, is slightly better as was expected: the amount of lost b-tracks is less in average.

For the MVA scheme, track selection BDT output distribution shows good discrimination. For the ROC curve, the overall efficiency worse than before. There are couple of possible reasons for this. The choosen cut value for tracks based on BDT output is too large. There might be also some problems with the implementation still so the results are not definite; just the opposite in fact.

As outlook, we are planning to develop the MVA scheme further. We will try also to give a BDT output as a free parameter to IP3D instead of performing track selection with it. Further, we will evaluate the best performance boost obtainable with track selection schemes discarding all other tracks than b hadron tracks by using truth information to select the tracks. We take only b-tracks and discard others, including all the other jets and see what is the outcome of this. Thus, we will determine the fate of this approach to the b-tagging performance boosting. We are also planning to rerun the analysis using a Z' sample that should give more high p_T jets as discussed.

References

[1] ATLAS Collaboration. Expected performance of the ATLAS b-tagging algorithms in Run-2. ATL-PHYS-PUB-2015-022. URL: https://cds.cern.ch/record/2037697.

[2] ATLAS Collaboration. Optimisation of the ATLAS b-tagging algorithm performance for the 2016 LHC Run. ATL-PHYS-PUB-2016-012. URL: https://cds.cern.ch/record/2160731.

[3] ATLAS Collaboration. Search for resonances in the mass distribution of jet pairs with one or two jets identified as b-jets in proton-proton collision at $\sqrt{s}$=13 TeV with the ATLAS detector. *Phys. Let. B*, 759, 2016.

[4] ATLAS Collaboration. Search for resonances below 1.2 TeV from the mass distribution of b-jet pairs in proton-proton collisions at $\sqrt{s}$ = 13 TeV with ATLAS detector. ATLAS-CONF-2016-031. URL: https://cds.cern.ch/record/2161136.

[5] ATLAS Collaboration. The ATLAS experiment at the CERN Large Hadron Collider. *JINST*, 3(S08003), 2008.

[6] Particle Data Group. Review of Particle Physics. *Chin. Phys. C*, 38(9), 2014.

[7] DØ Collaboration. Observation of Single Top Quark Production. https://www-d0.fnal.gov/Run2Physics/top/singletop_observation. Accesory figures of the article published in March 2009.

[8] Saptaparna Bhattacharya. Efficiency measurement of b-tagging algorithms developed by the CMS experiment. In *Particles and fields. Proceedings, Meeting of the Division of the American Physical Society, DPF 2011, Providence, USA, August 9-13, 2011*, 2011.

[9] M. Cacciari, G.P. Salam, and G. Soyez. The Anti-k_t jet clustering algorithm. *JHEP*, 04(063), 2008.

[10] ATLAS Collaboration. Tagging and suppression of pileup jets with the ATLAS detector. ATLAS-CONF-2014-018. URL: https://cds.cern.ch/record/1700870.

[11] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[12] David J. Griffiths. *Introduction to Elementary Particles*. Wiley-VCH, Weinheim, 2008.

[13] Ahmed Ali and Gustav Kramer. Jets and QCD: A Historical Review of the Discovery of the Quark and Gluon Jets and its Impact on QCD. *Eur. Phys. J. H*, 36:245–326, 2011.

A Appendix

In this appendix, we give brief introduction to important concepts behind this work, the strong interaction, the jets and the CKM matrix of weak interaction. A reader not familiar with these concepts can get a general picture of things from here. This section is based on materials from [12], [13] and [6].

A.1 Strong interaction

For a long time physicists thought that all the fundamental interactions are such that they get weaker when the distance between interacting particles increases. However, nature seems to surprise and amaze physicists with phenomena that break this kind of prejudices. Indeed, when more information about strong interaction, the force that keeps nuclear particles such as protons and neutrons together, was obtained it became increasingly more evident that in its case the increasing distance will increase the force. On the other hand, when the particles have **asymptotic freedom**: near each other or with high energy they barely interact.

The matter, that is fermionic particles that feel strong interaction are the quarks and the force is mediated by gluons. There are six types of quarks called flavours and each flavour has corresponding antiparticle. The flavours are up (u), down (d), strange (s), charm (s), bottom (b), and top (t). The quarks come in three generations, the pairs u and d, c and s , and t and b each forming a generation. First of each pair in this order has electric charge $+2/3$ and the second has charge $-1/3$. All the quarks are symmetric in the sense of strong interaction meaning that one can switch any two quarks or replace a quark by another and it would not make any difference in the equations or parameter of this interaction. This means that the each quark is similarly charged in strong interaction sense, namely they have property of color charge that is not dependent on the flavour. However, different quarks behave differently with respect to weak interaction and electromagnetic interaction.

The mediators i.e. bosonic particles feeling strong interaction are the gluons. There are in total eight different gluons. The gluons interact neither via electromagnetic nor via weak interaction. However, each gluon carry two color charges that means that they have self-interaction and interact with each other. This causes the fact that the force becomes stronger when the distance between interacting particles increases.

As the fact that the strong interaction becomes stronger as distance increases hints, there cannot be color charge roaming free, that is quarks and gluons always come in lumps of no color. This phenomenon is called **color confinement**. Because of this, quarks and gluons form hadrons: mesons with two valence quarks being bosons and baryons with three valence quarks being fermions.

A.2 Jets

Let's now consider what happens when quarks or gluons have large amount of energy when they are formed or have being kicked by other particles. These processes happen

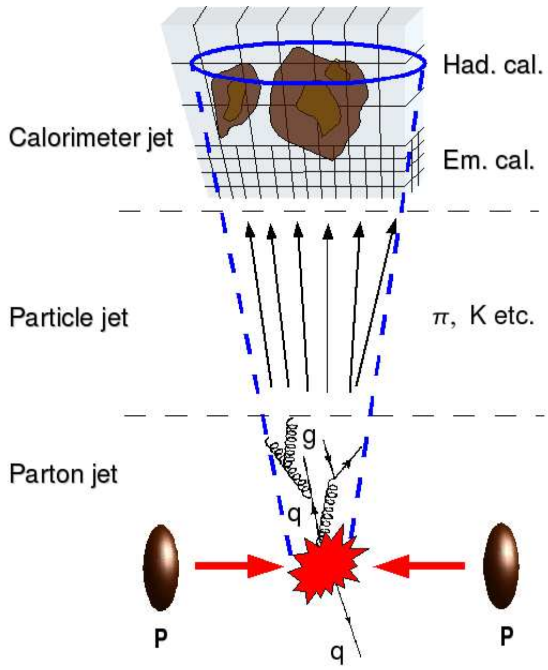

Figure 12: Formation of a jet at LHC. Quarks and gluons of high energy are formed in a pp collision. The quarks and gluons hadronize due to color confinement. The hadrons are shown in the detector calorimeter as clusters that can be clustered as jets.

and can be probed in particle collisions. When the highly energetic colored particles fly away from each other, the interaction energy becomes so large at about 1 fm separation that new pairs of quarks are created spontaneously so that finally the result is a new bunch of quarks and gluons. All the resulting colored particles will form hadrons of neutral color that fly out of the detector. This process is called fragmentation or hadronization and it leads to collimated jets of the resulting particles flying out. There are one of such jet for each initial quark or gluon that got kicked in the first place. Because of the conservation of energy and momentum (four-momentum) in the process, the jets resemble the initial quark or gluon. Collimation of a jet is guaranteed by the high momentum in the lab frame. The formation and detection of a jet is sketched in figure 12. Jets are prominent objects in the LHC physics. They are the method of obtaining information about the interactions between two protons because they show the properties of gluons and quarks inside. Different kind of distributions of jet properties can be calculated using the Standard Model and information about the inner structure of protons. Jet analyses can thus be used to test the Standard Model provide also more information about the structure of proton. Through flavour tagging, jets can also be used to Higgs studies, top quark studies and searching for physics beyond the standard model. In short, if the LHC is considered to be a stage, jets are the main actors in the drama.

A jet is given flavour by its initial source: it is said to be a gluon jet if it is from a gluon and it is given the flavour of the quark if the initial particle was a quark.

The definition of a jet as the result of fragmentation of a energetic quark is quite clear. However, a detector can only detect the out-coming particles so it sees the underlying fragmentation only in a indirect manner. Furthermore, the quantum field theory of strong interaction, Quantum Chromodynamics cannot predict the soft hadronization process and phenomenological models like Lund string model are used instead. Thus in practice, the jet seen in a detector is a phenomenological concept and must be given a phenomenological definition.

In this paper we use the anti-k_t jet clustering algorithm [9] which provides us also the definition of a jet. It is based on a metric giving distance between reconstructed particles i, j d_{ij} and distance from a particle i to the beam B d_{iB}

$$\begin{aligned} d_{ij} &= \min(k_{ti}^{-2}, k_{tj}^{-2})\frac{\Delta_{ij}^2}{R^2} \\ d_{iB} &= k_{ti}^{-2} \end{aligned} \quad (5)$$

where $\delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and k_{ti}, y_i and ϕ_i are the transverse momentum rapidity and azimuthal angle in detector frame of particle i. R is a parameter between 0 and 1 which determines the size of a jet in solid angle and which must be specified. In this study we use $R = 0.4$ to get well collimated jets without too much excess tracks.

The anti-k_t algorithm proceeds as finding the smallest distances between particles and the beam. If the smallest distance in this metric is between two particles, the particles are clustered together into a pseudo jet is repeated. If the smallest distance is between a particle or a pseudo jet and the beam, the particle or a pseudo jet is considered as a proper jet and is removed from the list of particles. The algorithm continues as long as there are tracks left. It is called anti-k_t as the metric has inverse exponent. It thus first clusters the particles with large k_t and proceeds to more soft particles. By construction the algorithm is collinear and infrared safe.

There are also two frequently used algorithms that are similar to anti-k_t: Cambridge-Aachen with the exponent of k_t 0 and k_t with the exponent 2. Cambridge-Aachen is used, if all jets are particles are wanted to cluster in equal footing and k_t if soft particles have privilege [9].

A.3 The CKM matrix

The CKM matrix encodes the information about the strength by which different quarks decay to other flavours i.e. how the weak interaction affects quarks. The CKM matrix is a unitary matrix of the form

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}, \quad (6)$$

The elements give weight of each quark of down-type (d,s,b) in a down-type superposition state to couple with a up-type quark state (u,c,t) For instance, the combination that couples to a u quark is $d' = V_{ud}d + V_{us}s + V_{ub}b$ where d', d, s, and b are states in which there is a quark of the flavour. The elements thus give probability amplitude for a down-like quark (d,s,b) to decay into a up-like quakr (u,c,t) or vice versa.

The matrix is a piece in Standard Model encoding essential information about CP violation.