# Identifying Boosted Objects with $N$-subjettiness and Linear $k$-means Clustering

by

Ken Van Tilburg

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degrees of

Bachelor of Science in Physics

and

Bachelor of Science in Mathematics

at the



MASSACHUSETTS INSTITUTE OF TECHNOLOGY 

June 2011



Author .. .....
Department of Physics
May 6, 2011

Certified by.. .....
Jesse Thaler
Assistant Professor of Physics
Thesis Supervisor

Accepted by ... .....
Nergis Mavalvala
Professor, Senior Thesis Coordinator, Department of Physics

# Identifying Boosted Objects with $N$-subjettiness and Linear $k$-means Clustering

by

Ken Van Tilburg

Submitted to the Department of Physics
on May 6, 2011, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science in Physics
and
Bachelor of Science in Mathematics

## Abstract

In this thesis, I explore aspects of a new jet shape — $N$-subjettiness — designed to identify boosted hadronically-decaying objects (with a particular focus on tagging top quarks) at particle accelerators such as the Large Hadron Collider. Combined with an invariant mass cut on jets, $N$-subjettiness is a powerful discriminating variable for tagging boosted objects such as top quarks and rejecting the fake background of QCD jets with large invariant mass. In a crossover analysis, the $N$-subjettiness method is found to outperform the common top tagging methods of the BOOST2010 conference, with top tagging efficiencies of 50% and 20% against mistag rates of 4.0% and 0.19%, respectively. The $N$-subjettiness values are calculated using a new infrared- and collinear-safe minimization procedure which I call the linear $k$-means clustering algorithm. As a true jet shape with highly effective tagging performances, $N$-subjettiness has many advantages on the experimental as well as on the theoretical side.

Thesis Supervisor: Jesse Thaler
Title: Assistant Professor of Physics

# Acknowledgments

I would like to take this opportunity to first and foremost thank my thesis supervisor, Prof. Jesse Thaler. He not only contributed tremendously to this work by enlightening me with his ample expertise in this specific field, but he also provided me with academic and non-academic mentoring in the later stages of my undergraduate career. At certain times, he selflessly set aside what would be best to fulfill our direct research aims, to make room for my personal development in physics. His broad, extensive knowledge in various fields of physics is something I will strive to achieve in years to come. It has been a pleasure working with him for now exactly a year, and I hope I will get more chances to work with him in the future.

My sincere gratitude goes out to Prof. Pablo Jarillo-Herrero and his graduate students Javier Sanchez-Yamagishi and Leonardo Campos, who showed me that the early stages of an academic career can actually be quite fun and exciting. When I joined his research group, Pablo's young lab at MIT was in full expansion, but nevertheless he took the time to invest in an initially quite useless sophomore student. Over the course of several semesters, I saw his handful of students and the physical lab space grow into the successful research group it is now. The feeling of having been a small part of that success has been extremely rewarding to me. In addition, I thank Pablo for guiding me even after my research interests veered away from condensed matter expèriments.

I am also grateful for the many opportunities MIT has presented me. Without the many semesters of funding from the Undergraduate Research Opportunities Program (UROP), I would have been more likely to venture into a career outside of academia, a decision I am sure I would have regretted later. Many of my professors and recitation instructors in physics, mathematics and other fields at MIT and Harvard University have inspired me on many occasions, and have set the bar high when it comes to teaching. These people are Michael Artin, Gabriel Carroll, Hung Cheng, Isaac Chuang, Jan Egedal, Daniel Freedman, Robert Jaffe, Wolfgang Ketterle, Hong Liu, Nergis Mavalvala, John Negele, Kaiy Quek, Iain Stewart and Andrew Strominger.

This work and in fact almost everything else I did over the past four years would not have been possible without the unconditional support of my family and friends both in Belgium and the United States. Specifically in relation to this thesis, I would like to thank my Chi Phi fraternity brothers (and Course VI majors) Colin Hom, Eric Lubell, Christian Therkelsen and David Witmer for helpful conversations about programming and algorithms.

I owe most of my gratitude to my direct family — my sister Britt, parents Anne & Eric and four grandparents — who let me and helped me pursue my ambitions far away from home, and who I know will always be there for me when I fly back across the Atlantic.

# Contents


**1 Introduction** **15**

**2 Boosted Top Quarks and $N$-subjettiness** **19**

2.1 Defining $N$-subjettiness . . . . . 22

2.2 Finding Candidate Subjets . . . . . 25

2.3 Infrared- and Collinear Safety . . . . . 29

2.4 Summary . . . . . 29

**3 Efficiency Study** **31**

3.1 Analysis Overview . . . . . 32

3.2 Boosted Top results . . . . . 33

3.3 Further Optimization . . . . . 35

**4 Conclusions** **39**

**A The Unique Minimum of a Sum of Cones** **41**

A.1 Relevance to $N$-subjettiness and Linear $k$-means Clustering . . . . . 41

A.2 Statement and Proof of Theorem . . . . . 42

**B Generalized Formulations of $N$-subjettiness and Minimization** **47**

B.1 Generalized Definition of $N$-subjettiness . . . . . 47

B.2 Generalized Weighted $k$-means Clustering Algorithm . . . . . 49

B.3 Comparison of Subjet Finding Methods . . . . . 51

**C Additional Event Displays** **55**

# List of Figures

2-1 Schematic of fully hadronic decay sequences in (a) $t\bar{t}$ and (b) dijet QCD events. Whereas a top jet is typically composed of three distinct lobes of energy, a QCD jet acquires invariant mass through multiple splittings. 20

2-2 Left: Event displays for (a) top jets and (c) QCD jets with invariant mass near $m_{\text{top}}$. For displaying purposes, the particles were clustered into virtual calorimeter cells of size 0.1 by 0.1; the marker size for each cell is proportional to the logarithm of its scalar transverse momentum in units of GeV (cells with $p_T < 1$ GeV were not displayed). The blue open diamond indicates the total jet direction as calculated by the anti-$k_T$ algorithm. The red open square indicates the direction which yields $\tau_1$, while the red circles indicate the two subjet directions and the crosses the three subjet directions corresponding to $\tau_2$ and $\tau_3$, respectively. The discriminating variable $\tau_3/\tau_2$ measures the relative alignment of the jet energy along the crosses compared to the circles. Right: $\tilde{\tau}_1$ distributions of the same (b) top and (d) QCD jets as a function of position. Note that the global minimum of $\tilde{\tau}_1$ does not necessarily agree with the jet direction, e.g. the red square, which corresponds to the darkest spot in (b), does not line up with the blue diamond. . . . . 21

2-3 Distributions of (a) $\tau_1$, (b) $\tau_2$ and (c) $\tau_3$ for boosted top and QCD jets. For these plots, I impose an invariant mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV on jets with $R = 1.0$ and 500 GeV $< p_T <$ 600 GeV. . 23

2-4 Distributions of (a) $\tau_2/\tau_1$ and (b) $\tau_3/\tau_2$ for boosted top and QCD jets. The selection criteria are the same as in Fig. 2-3. One can see that $\tau_3/\tau_2$ is a good discriminating variable between top jets and QCD jets. In this thesis, I do not explore $\tau_2/\tau_1$ for top jets, though it appears to contain additional information. . . . . 24

2-5 Density plots in the (a) $\tau_1$–$\tau_2$ plane and (b) $\tau_2$–$\tau_3$ plane for boosted top and QCD jets. The selection criteria are the same as in Fig. 2-3. These plots suggest further improvement in boosted top identification is possible with a multivariate method (see Sec. 3.3). . . . . 24

3-1 Kinematic information of anti-$k_T$ jets. (a) Transverse momentum distribution of jets in the combined sample of parton-level $p_T$ between 200 and 800 GeV. Note that this is an unphysical $p_T$ distribution, and merely serves as a fair "testing ground" for the various tagging methods. (b),(c) Mass distributions of jets in the combined sample as well as the 500 GeV $< p_T <$ 600 GeV sample. Note that the $N$-subjettiness cut $\tau_3/\tau_2 < 0.5$ eliminates nearly all QCD jets as well as top jets with a mass much smaller than $m_{\text{top}}$, but leaves most of the top resonance peak intact. . . . . 34

3-2 Top jet signal efficiency/background rejection plots. The orange curve corresponds to tagging rates of the method based on a sliding $N$-subjettiness cut of $\tau_3/\tau_2$ combined with a fixed mass window. The rightmost points in each plot correspond to just applying a jet invariant mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV, and points to the left of these are obtained from progressive cutting on the $\tau_3/\tau_2$ ratio only. The other curves were taken from the BOOST2010 study [34]. . . . . 36

3-3 Optimized top jet signal efficiency/background rejection plot for the combined sample of jets generated by partons of transverse momenta between 200 GeV and 800 GeV. The orange curve corresponds to tagging rates of the method based on a sliding $N$-subjettiness cut of $\tau_3/\tau_2$ combined with a fixed mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV (the same curve as in Figs. 3-2(a) and 3-2(b)). The gray points correspond to tag/mistag rates for the multivariate selection method within part of the parameter space $(s, \tau_{2;0}, m_{\min}, m_{\max})$, as described in Sec. 3.3. Recalling that lower curves are better than higher ones, we see that modest improvements are possible with this multivariate analysis. . . 38

B-1 Top jet signal efficiency/background rejection plots for the combined sample of jets generated by partons of transverse momenta between 200 GeV and 800 GeV, for the four selection methods of Sec. B.3. All methods first impose a fixed mass window 160 GeV $< m_{\text{jet}} <$ 240 GeV and impose progressive sliding cuts on the ratio of their 3-subjettiness and 2-subjettiness measures. One can see that the classical "quadratic" $k$-means algorithm and its discriminant ($\beta = 2$) do worst, while their linear variants ($\beta = 1$) discussed in the body of the thesis do best. However, lower values of $\beta$ do not further improve top identification, and even do worse (see e.g. the curve for $\beta = 1/2$). We also observe that using the linear $k$-means minimization procedure over exclusive $k_T$ subjet finding yields significant improvements across the entire efficiency space. . . . . 53

C-1 Top row: boosted top jets. Bottom row: QCD jets with invariant mass close to $m_{\text{top}}$. The coloring and labeling is the same as in Fig. 2-2. . . 56

# List of Tables

3.1 Summary of mistag rates at different working points for different top-taggers. Recall that smaller mistag rates (for a fixed tagging efficiency) are desired over higher mistag rates. The univariate $N$-subjettiness method, described in Sec. 3.2, requires $\tau_3/\tau_2 < s$ and 160 GeV $< m_{\text{jet}} <$ 240 GeV, with $s$ variable. The multivariate $N$-subjettiness method, described in Sec. 3.3, requires $\tau_3/(\tau_2 - \tau_{2;0}) < s$ and $m_{\text{jet}} \in [m_{\text{min}}, m_{\text{max}}]$, with $s, \tau_{2;0}, m_{\text{min}}, m_{\text{max}}$ all variable. . . . . 35

# Chapter 1

# Introduction

The Large Hadron Collider (LHC) will search for new physics by probing a previously unexplored kinematic regime. Most new physics scenarios that provide a solution to the hierarchy problem predict that the LHC will produce new heavy particles with decay channels involving top quarks, $W/Z$ bosons, and Higgs bosons. In addition, many extensions of the standard model including technicolor and Higgs compositeness invoke new heavy resonances within the LHC reach with large branching fractions to pairs of gauge bosons and top quarks. Therefore, a key task in the search for physics beyond the standard model is to efficiently identify final state electroweak gauge bosons and top quarks in a variety of kinematic configurations.

With its current center-of-mass energy of $\sqrt{s} = 7$ TeV, the LHC is already able to produce new TeV-scale resonances which can decay to highly boosted electroweak bosons and/or top quarks. For a large enough boost factor, the decay and fragmentation of such a boosted object yields a collimated spray of hadrons which a standard jet algorithm would reconstruct as a single jet. Thus, standard reconstruction methods for electroweak bosons and top quarks become ineffective due to the immensely large background of ordinary QCD jets. One possibility is to focus on channels where the boosted object decays leptonically, though such methods discard much of the original signal and may therefore not be optimal for detecting new heavy resonances.

Recently, there has been considerable progress in identifying boosted hadronically-decaying objects using jet substructure techniques. Algorithmic methods use infor-

mation from the jet clustering procedure to extract the internal structure of jets [1, 2, 3, 4, 5, 6, 7], and are able to successfully distinguish between jets originating from boosted electroweak boson and top quarks (denoted here as "$W$ jets", "top jets", etc.) and those originating from light quarks or gluons ("QCD jets"). Jet shape methods efficiently tag boosted objects with jet-based observables that take advantage of the different energy flow in the decay pattern of signal jets and background jets [8]. In addition, there are "jet grooming" techniques such as filtering [9, 10], pruning [11, 12], trimming [13], and their combinations [14] which aid in the identification of boosted objects by reducing the smearing effects of jet contamination from initial state radiation, underlying event activity, and pileup. Taken together, these jet substructure methods show much promise for enhancing searches for new physics in all-hadronic decay channels.[1]

In this thesis, I discuss aspects of the jet shape "$N$-subjettiness" (denoted by $\tau_N$), and a new tagging method for hadronically-decaying boosted objects based on this variable. It was recently introduced in Ref. [29], and originally adapted from the event shape "$N$-jettiness" advocated in Ref. [30] as a way to veto additional jet emissions and define an exclusive jet cross section.[2] Here, I take advantage of the multi-body kinematics in the decay pattern of boosted hadronic objects, and use $N$-subjettiness to effectively "count" the number of subjets in a given jet. I find that $\tau_3/\tau_2$ is an effective discriminating variable for three-prong objects like boosted top quarks, and though not explored in this thesis, $\tau_2/\tau_1$ can be used to efficiently tag boosted electroweak bosons (see Ref. [29]).

Compared to previous jet substructure techniques, $N$-subjettiness has a number of advantages. First, to identify boosted objects, one wants to find jets that contain two or more lobes of energy. While previous jet shape measures do capture the deviation of a jet from a one-lobe QCD-like configuration, $N$-subjettiness is a more direct measure of the desired energy flow properties. Second, it is convenient to be able

[1] Additional recent related work appears in Refs. [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28].

[2] A few days before Ref. [29], an independently discovered tagging method based on $N$-subjettiness (though defined slightly differently) was presented in Ref. [31]. This thesis is an extension of the work in the former and will thus mostly use the same definitions and concepts as those in Ref. [29].

to adjust the relative degree of signal efficiency and background rejection without having to perform computationally intensive algorithmic adjustments. Like for other jet shape methods, $\tau_N$ can be calculated for every jet, and a flexible one-dimensional cut on a function $f(\tau_1, \ldots, \tau_N)$ can determine the efficiency/rejection curve. Similarly, the set of $\tau_N$ values can be used as input to a multivariate discriminant method for further optimization. Third, $N$-subjettiness is an inclusive jet shape and is defined and calculated without any reference to a jet substructure algorithm.[3] This will likely make $N$-subjettiness more amenable to higher-order perturbative calculations and resummation techniques (see, e.g. recent work in Ref. [32, 33]) compared to algorithmic methods for studying substructure. Finally, $N$-subjettiness gives favorable efficiency/rejection curves compared to other jet substructure methods (see Chapter 3 for an apples-to-apples comparison against other top-tagging methods).

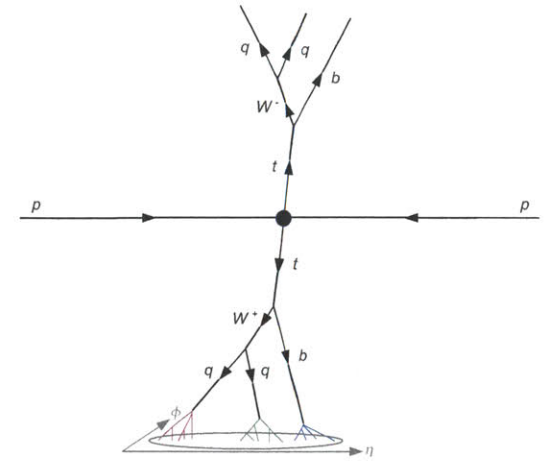
The remainder of this paper is organized as follows. In Chapter 2, we define $N$-subjettiness and discuss our general methods. We present a tagging efficiency study in Sec. 3, where we use $N$-subjettiness to identify individual hadronic top quarks, and compare our method against those of the BOOST2010 study [34]. Our conclusions follow in Chapter 4. The appendices will contain further information, such as a proof of an important claim made in Chapter 2 (App. A), a discussion about a generalized definition of $N$-subjettiness and associated minimization procedures (App. B), and additional event displays (App. C).

[3] In the previous work of Ref. [29], there was still some residual algorithmic dependence in the definition of $N$-subjettiness, namely in finding the candidate subjet directions, but this has now been completely removed. See Sec. 2.2 for further discussion.

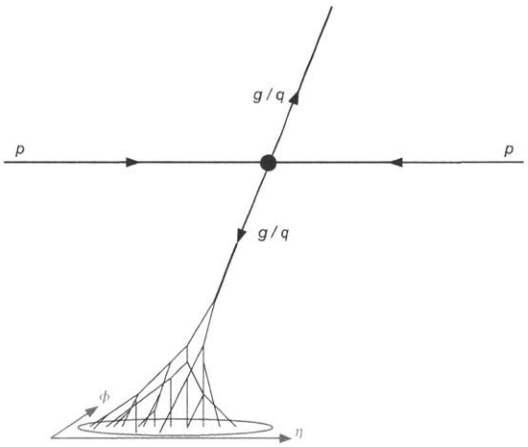
# Chapter 2

# Boosted Top Quarks and $N$-subjettiness

Boosted hadronic objects have a fundamentally different energy pattern than QCD jets of comparable invariant mass. For concreteness, we will consider the case of a boosted top quark as shown in Fig. 2-2(a), though a similar discussion holds for boosted electroweak bosons, Higgs bosons or new physics objects (see for example Ref. [29]). In a particle accelerator, the top quark will decay almost instantly into a $b$-quark and a $W$ boson, after which the latter decays into two quarks about 2/3 of the time. Therefore, a jet originating from a sufficiently boosted top quark should be composed of three distinct—but not necessarily easily resolved—hard subjets with a combined invariant mass of around 171 GeV, as seen in Fig. 2-1(a). A boosted QCD jet with an invariant mass of around 171 GeV usually originates from a single hard parton and acquires mass through large angle soft splittings, as is depicted in Fig. 2-1(b). We want to exploit this difference in expected energy flow to differentiate between these two types of jets by "counting" the number of hard lobes of energy within a jet.




(a)




(b)

Figure 2-1: Schematic of fully hadronic decay sequences in (a) $t\bar{t}$ and (b) dijet QCD events. Whereas a top jet is typically composed of three distinct lobes of energy, a QCD jet acquires invariant mass through multiple splittings.

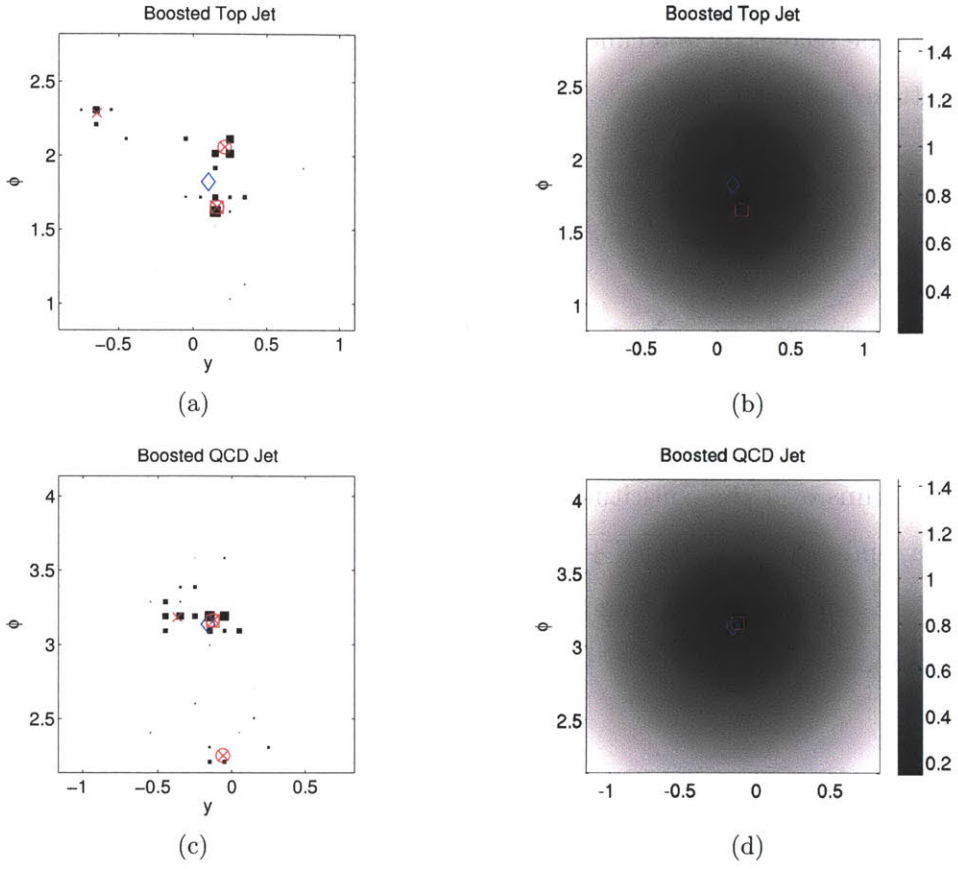



Figure 2-2: Left: Event displays for (a) top jets and (c) QCD jets with invariant mass near $m_{\text{top}}$. For displaying purposes, the particles were clustered into virtual calorimeter cells of size 0.1 by 0.1; the marker size for each cell is proportional to the logarithm of its scalar transverse momentum in units of GeV (cells with $p_T < 1$ GeV were not displayed). The blue open diamond indicates the total jet direction as calculated by the anti-$k_T$ algorithm. The red open square indicates the direction which yields $\tau_1$, while the red circles indicate the two subjet directions and the crosses the three subjet directions corresponding to $\tau_2$ and $\tau_3$, respectively. The discriminating variable $\tau_3/\tau_2$ measures the relative alignment of the jet energy along the crosses compared to the circles. Right: $\tilde{\tau}_1$ distributions of the same (b) top and (d) QCD jets as a function of position. Note that the global minimum of $\tilde{\tau}_1$ does not necessarily agree with the jet direction, e.g. the red square, which corresponds to the darkest spot in (b), does not line up with the blue diamond.

## 2.1 Defining $N$-subjettiness

I start by defining the inclusive jet shape called "$N$-subjettiness" and denoted by $\tau_N$. First, one reconstructs a candidate top jet using some jet algorithm. I define the $N$-subjettiness of a jet to be

$$\tau_N = \min_{N \text{ subjet directions}} \tilde{\tau}_N, \tag{2.1}$$

$$\tilde{\tau}_N = \frac{1}{d_0} \sum_i p_{T,i} \min \{\Delta R_{1,i}, \Delta R_{2,i}, \ldots, \Delta R_{N,i}\}. \tag{2.2}$$

Here, $i$ runs over the constituent particles in a given jet, $p_{T,i}$ are their transverse momenta, and $\Delta R_{J,i} = \sqrt{(\Delta y)^2 + (\Delta \phi)^2}$ is the distance in the rapidity-azimuth plane between a candidate subjet $J$ and a constituent particle $i$. The normalization factor $d_0$ is taken as

$$d_0 = \sum_i p_{T,i} R_0, \tag{2.3}$$

where $R_0$ is the characteristic jet radius used in the original jet clustering algorithm. The minimization over the candidate subjet directions in Eq. (2.1) is not a trivial step and may at first look like a computationally (too) daunting procedure, but in Sec. 2.2 I present a fast algorithm which can perform this task.

It is straightforward to see that $\tau_N$ quantifies how $N$-subjetty a particular jet is, or in other words, to what degree it can be regarded as a jet composed of $N$ subjets. Jets with $\tau_N \approx 0$ have all their radiation aligned with the candidate subjet directions and therefore have $N$ (or fewer) subjets. Jets with $\tau_N \gg 0$ have a large fraction of their energy distributed away from the candidate subjet directions and therefore have at least $N + 1$ subjets. Figs. 2-2(b) and 2-2(d) show typical $\tilde{\tau}_1$ distributions of a top jet and a QCD jet. Plots of $\tau_1$, $\tau_2$ and $\tau_3$ comparing top jets and QCD jets are shown in Fig. 2-3.

Less obvious is how best to use $\tau_N$ for identifying boosted top quarks. While one might naively expect that an event with small $\tau_3$ would be more likely to be a top

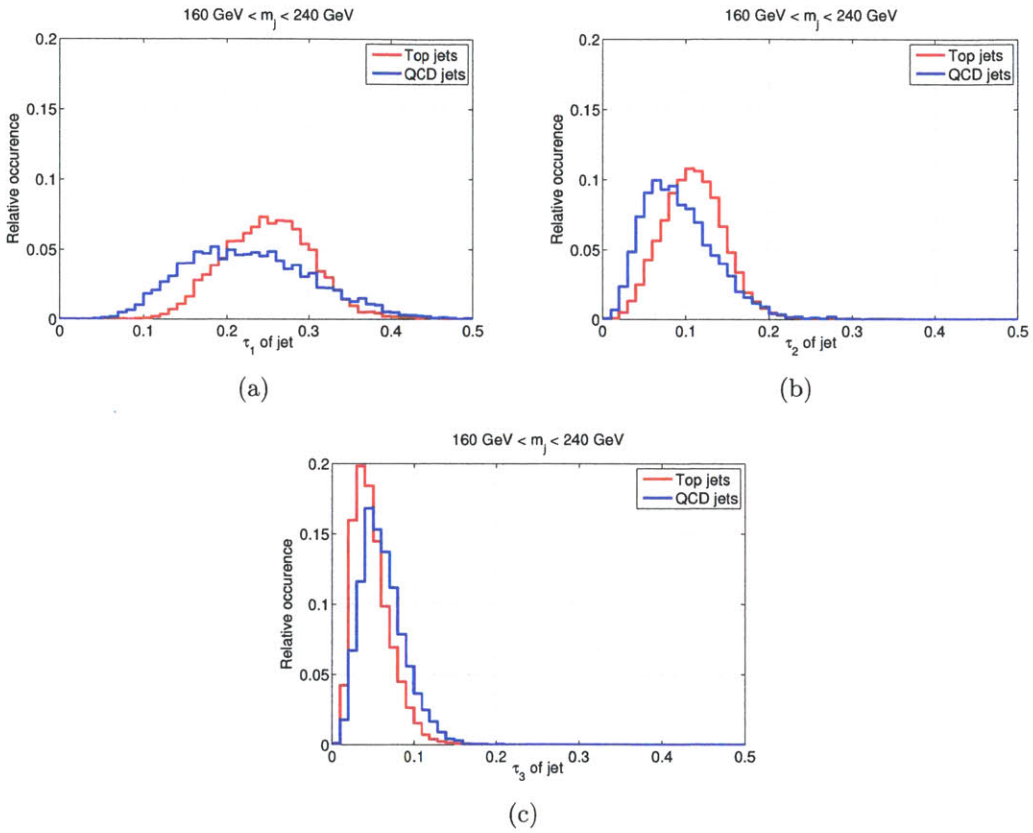
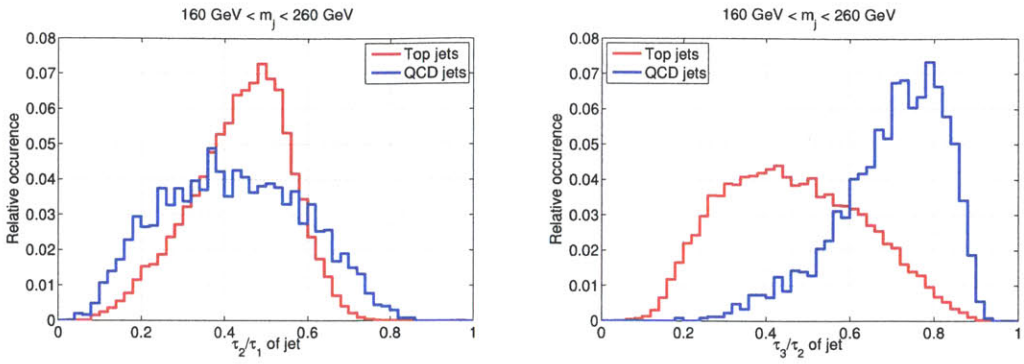
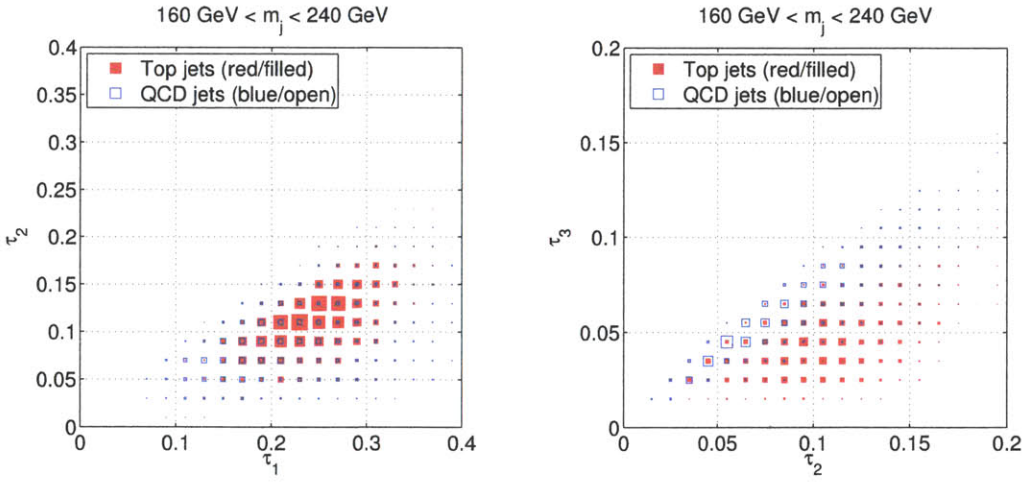



Figure 2-3: Distributions of (a) $\tau_1$, (b) $\tau_2$ and (c) $\tau_3$ for boosted top and QCD jets. For these plots, I impose an invariant mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV on jets with $R = 1.0$ and 500 GeV $< p_T <$ 600 GeV.




(a) (b)

Figure 2-4: Distributions of (a) $\tau_2/\tau_1$ and (b) $\tau_3/\tau_2$ for boosted top and QCD jets. The selection criteria are the same as in Fig. 2-3. One can see that $\tau_3/\tau_2$ is a good discriminating variable between top jets and QCD jets. In this thesis, I do not explore $\tau_2/\tau_1$ for top jets, though it appears to contain additional information.




(a) (b)

Figure 2-5: Density plots in the (a) $\tau_1$–$\tau_2$ plane and (b) $\tau_2$–$\tau_3$ plane for boosted top and QCD jets. The selection criteria are the same as in Fig. 2-3. These plots suggest further improvement in boosted top identification is possible with a multivariate method (see Sec. 3.3).

jet, observe that a QCD jet can also have small $\tau_3$, as shown in Fig. 2-3(c). Similarly, though top jets are likely to have large $\tau_1$ and $\tau_2$, QCD jets with a diffuse spray of large angle radiation can also have large $\tau_1$ and $\tau_2$, as shown in Figs. 2-3(a) and 2-3(b). However, those QCD jets with large $\tau_2$ typically have large values of $\tau_3$ as well, so it is in fact the *ratio* $\tau_3/\tau_2$ which is the preferred discriminating variable. As seen in Fig. 2-4(b), top jets have smaller $\tau_3/\tau_2$ values than QCD jets. Of course, one can also use the full set of $\tau_N$ values in a multivariate analysis, as suggested by Figs. 2-5(a) and 2-5(b), and we will briefly explore this possibility in Sec. 3.3.

As mentioned in the introduction, $N$-subjettiness is adapted from the similar quantity $N$-jettiness introduced in Ref. [30]. There are three important differences: the sum over $i$ only runs over the hadrons in a particular jet and not over the entire event, I do not have candidate (sub)jets corresponding to the beam directions, and the distance measure is only longitudinally boost invariant and not fully Lorentz invariant. The definition of $\tau_N$ is by no means unique, and some variations are discussed in App. B, though Eqs. (2.1) and (2.2) appear to be well-suited for boosted object identification.

## 2.2 Finding Candidate Subjets

A key step for defining $N$-subjettiness is to appropriately choose the candidate subjet directions. As also mentioned in Ref. [30], ideally one would determine $\tau_N$ by minimizing over all possible candidate subjet directions, analogously to how the event shape thrust is defined [35]. In that case, $\tau_N$ is a strictly decreasing function of $N$, and $0 < \tau_N/\tau_{N-1} < 1$. In Ref. [29], it was thought that a search for the global minimum of $\tilde{\tau}_N$ for each jet would be too computationally intensive, so the candidate subjet directions were determined by using the exclusive $k_T$ clustering algorithm [36, 37], forcing it to return exactly $N$ jets. This was found to work well in terms of efficiency, but it introduced an algorithmic dependence and a certain sense of arbitrariness in the jet shape. Below, however, I present a fast infrared- and collinear safe minimization procedure which can determine the candidate subjet directions of the global minimum

(which is inherently independent of any algorithm) of $\tilde{\tau}_N$ to arbitrary precision. A comparison between different subjet finding procedures can be found in App. B.

Minimizing the function $\tau_N$ of Eq. (2.1) is similar to the classical computer science problem of finding $k$ number of clusters in a data set, and is called the $k$-means clustering problem, which is to find the $k$ cluster centers (or "means") that minimize the in-cluster variance (i.e. the weighted sum of the distances *squared* between the data points and the closest cluster center).[1] A solution to this problem is commonly known as Lloyd's algorithm [38], which terminates in polynomial time and produces $k$ means which form a (local) minimum of the cluster variance. Combined with sufficiently many reseedings of the initial $k$ cluster centers, Lloyd's algorithm can find the global minimum of the cluster variance.

I present here an adaptation of Lloyd's algorithm, which aims to execute the minimization hinted at in Eq. (2.1). Note that $\tilde{\tau}_N$ is the weighted sum of the *linear* rapidity-azimuth distances between particles and the closest subjet direction. Suppose for a moment that we want to minimize $\tilde{\tau}_1$ of a particular cluster $C$. Taking the first-order partial derivatives with respect to the coordinates $(y_0, \phi_0)$ of the only subjet direction and setting them to zero gives:

$$\frac{\partial \tilde{\tau}_1}{\partial y_0} = 0 = \frac{1}{d_0} \sum_{i \in C} \frac{p_{T,i}(y_i - y_0)}{\sqrt{(y_i - y_0)^2 + (\phi_i - \phi_0)^2}}$$
$$\frac{\partial \tilde{\tau}_1}{\partial \phi_0} = 0 = \frac{1}{d_0} \sum_{i \in C} \frac{p_{T,i}(\phi_i - \phi_0)}{\sqrt{(y_i - y_0)^2 + (\phi_i - \phi_0)^2}}.$$

Any pair $(y_0, \phi_0)$ which solves these two equations for a given distribution of particles $(p_{T,i}, y_i, \phi_i)_{i \in C}$ corresponds to a local minimum of $\tilde{\tau}_1$. Finding such a pair, let alone finding the one that corresponds to the global minimum, does indeed not seem straightforward. However, it turns out that one can find such a pair to arbitrary precision by a fast iterative algorithm. Furthermore, I prove in App. A that there is only one local minimum (which is thus also the global minimum) for a typical

[1] The "$k$" is standard notation in the computer science literature, though strictly speaking we are trying to find "$N$" number of clusters in the thesis. Henceforth, I will use "$k$" when referring to the algorithms, and $N$ to the jet shape $N$-subjettiness, but know that technically $N = k$ throughout.

distribution of particles.

In the spirit of Lloyd's algorithm, let us suppose we already have a "guess" or initial seeding of the candidate subjet direction; call it $\left(y_0^{(0)}, \phi_0^{(0)}\right)$. Now let us define the following recursive procedure:

$$y_0^{(n+1)} = \frac{\sum_{i \in C} \frac{p_{T,i} y_i}{\sqrt{\left(y_i - y_0^{(n)}\right)^2 + \left(\phi_i - \phi_0^{(n)}\right)^2}}}{\sum_{j \in C} \frac{p_{T,j}}{\sqrt{\left(y_j - y_0^{(n)}\right)^2 + \left(\phi_j - \phi_0^{(n)}\right)^2}}}, \qquad \phi_0^{(n+1)} = \frac{\sum_{i \in C} \frac{p_{T,i} \phi_i}{\sqrt{\left(y_i - y_0^{(n)}\right)^2 + \left(\phi_i - \phi_0^{(n)}\right)^2}}}{\sum_{j \in C} \frac{p_{T,j}}{\sqrt{\left(y_j - y_0^{(n)}\right)^2 + \left(\phi_j - \phi_0^{(n)}\right)^2}}}.$$

It is straightforward to see that if $\left(y_0^{(n+1)}, \phi_0^{(n+1)}\right) = \left(y_0^{(n)}, \phi_0^{(n)}\right)$, we have found a local minimum. In general, the sequence $\left(y_0^{(n)}, \phi_0^{(n)}\right)$ does not converge in finite time, but instead quickly tends to an asymptotic value. Hence we can find a local minimum of $\tilde{\tau}_1$ to arbitrary precision. Furthermore, the analytic proofs of App. A show that, with probability one, any given cluster of particles has only one local minimum of $\tilde{\tau}_1$, which is thus the global minimum (see for example Figs. 2-2(b) and 2-2(d)).

Another way of looking at the calculation of the global minimum of $\tilde{\tau}_N$, is that such a procedure simply amounts to dividing up the jet into $N$ distinct subclusters, calculating the sum of the $\tau_1$ of the subclusters, and then finding the partition into subclusters which gives rise to the lowest such sum. From Eq. (2.2), we see that particles should be assigned to the cluster associated with the nearest subjet direction, as the particle's contribution to $\tilde{\tau}_N$ enters through the $\tilde{\tau}_1$ value of that specific subjet. In the above paragraph, I argued that it is possible to construct a sequence which in practice always converges to the global minimum of $\tilde{\tau}_1$ of any one cluster. With this motivation in hand, we are ready to formulate a linearized variant of Lloyd's algorithm. Again, when this algorithm converges (to within arbitrary precision of $N$ asymptotic directions), then it converges to a local minimum of $\tilde{\tau}_N$.

**Linear weighted $k$-means algorithm.** ($N = k$)

1. *Initialization step.* Pick initial subjet directions $\left(y_{0,J}^{(0)}, \phi_{0,J}^{(0)}\right)_{J\in\{1,\ldots,N\}}$. Set $n = 0$.

2. *Assignment step.* Divide the jet into clusters $\left(C_J^{(n)}\right)_{J\in\{1,\ldots,N\}}$ by assigning the particles to the closest subjet direction. In other words, $i \in C_J^{(n)}$ if and only if

$$\sqrt{\left(y_i - y_{0,J}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,J}^{(n)}\right)^2} < \sqrt{\left(y_i - y_{0,M}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,M}^{(n)}\right)^2} \quad \forall M \neq J. \tag{2.4}$$

3. *Update step.* Based on the previous subjet directions $\left(y_{0,J}^{(n)}, \phi_{0,J}^{(n)}\right)_{J\in\{1,\ldots,N\}}$, calculate new subjet directions $\left(y_{0,J}^{(n+1)}, \phi_{0,J}^{(n+1)}\right)_{J\in\{1,\ldots,N\}}$ as follows:

$$y_{0,J}^{(n+1)} = \frac{\sum\limits_{i\in C} \frac{p_{T,i} y_i}{\sqrt{\left(y_i - y_{0,J}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,J}^{(n)}\right)^2}}}{\sum\limits_{j\in C} \frac{p_{T,j}}{\sqrt{\left(y_j - y_{0,J}^{(n)}\right)^2 + \left(\phi_j - \phi_{0,J}^{(n)}\right)^2}}}, \qquad \phi_{0,J}^{(n+1)} = \frac{\sum\limits_{i\in C} \frac{p_{T,i} \phi_i}{\sqrt{\left(y_i - y_{0,J}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,J}^{(n)}\right)^2}}}{\sum\limits_{j\in C} \frac{p_{T,j}}{\sqrt{\left(y_j - y_{0,J}^{(n)}\right)^2 + \left(\phi_j - \phi_{0,J}^{(n)}\right)^2}}}. \tag{2.5}$$

4. *Iteration.* Repeat Steps (b) and (c) until the average directional change of the subjets $\overline{\Delta}^{(n+1)}$ is smaller than some threshold $\delta_p$ (which encodes the desired precision), i.e. until

$$\overline{\Delta}^{(n+1)} \equiv \frac{1}{N} \sum_{J=1}^{N} \sqrt{\left(y_{0,J}^{(n+1)} - y_{0,J}^{(n)}\right)^2 + \left(\phi_{0,J}^{(n+1)} - \phi_{0,J}^{(n)}\right)^2} < \delta_p. \quad (n \geq 0) \tag{2.6}$$

For all practical purposes, the linear $k$-means algorithm is very fast. For example, for $\delta = 10^{-4}$, a non-optimized `C++` implementation of the algorithm running on a standard 2.54 GHz processor can calculate a local minimum of $\tilde{\tau}_3$ of a typical top jet in less than $10^{-4}$ seconds on average (given an initial seed). For every *two orders of*

*magnitude* in $\delta$ improvement, about a factor of *two* in computing time is needed; this empirical law roughly holds at least up to $\delta < 10^{-10}$. In none of millions of analyzed jets did the algorithm fail to converge to the desired precision of at least $\delta < 10^{-4}$.

## 2.3 Infrared- and Collinear Safety

The speed of the algorithm allows one to run the algorithm many times with different seeds to ensure that the global minimum is found, as the definition of Eq. (2.2) is such that a jet typically has relatively few local minima. For more discussion, see App. A.

Once the candidate subjets are identified through the algorithm of the previous section, $N$-subjettiness is a proper inclusive jet shape. Since Eq. (2.1) is linear in each of the constituent particle momenta, $\tau_N$ is an infrared- and collinear-safe observable. That is, the addition of infinitesimally soft particles does not change $N$-subjettiness (infrared safety), and the linear dependence on the particle momenta combined with the smooth angular dependence ensures that the same $\tau_N$ is obtained for collinear splittings (collinear safety). Crucially, the candidate subjets used in $N$-subjettiness must be determined via a method that is also infrared- and collinear-safe. The arguments presented in App. A show that the addition of infinitesimally soft particles does not affect the minimization procedure, as they cannot create extra local minima. I conclude that infrared- and collinear safety is thus automatic with the global minimization procedure of the previous section or by using exclusive $k_T$ declustering to find the subjet directions.

## 2.4 Summary

To summarize, $N$-subjettiness is an inclusive jet shape that offers a direct measure of how well jet energy is aligned into subjets, and is therefore an excellent starting point for boosted object identification. The ratio $\tau_N/\tau_{N-1}$ is an easily adjustable offline parameter which can be varied to adjust signal efficiency/background rejection

without having to redo the clustering of the particles in an event. It is suspected that $N$-subjettiness will lend itself better to theoretical studies than algorithmic boosted object tagging methods, either in fixed-order or resummed QCD calculations, though I will not attempt to perform such studies in this thesis. As we will see in the next section, $N$-subjettiness compares favorably to other boosted object tagging methods in terms of discriminating power.

# Chapter 3

# Efficiency Study

In this chapter, we investigate the tagging efficiencies for top jets and the mistagging rates for QCD jets. Here, we select candidate boosted objects using an invariant mass cut augmented with an $N$-subjettiness criterion and compare our results to the common top tagging methods described in the BOOST2010 study [34]. For a detailed account on how the other top tagging methods are used to produce the results of Fig. 3-2 and Table 3.1, I refer the reader to Ref. [34]. For detailed descriptions of the methods themselves, see their respective original sources, Refs. [2, 3, 4, 5, 6, 7, 12, 23, 24, 39, 40].

In Ref. [29], $N$-subjettiness was already compared to simple implementations of the Johns Hopkins Top Tagger [7] and the ATLAS YSplitter method [4] on event samples produced with the Pythia8 Monte Carlo generator [41, 42]. That study had different underlying event (UE) as well as different initial- and final state radiation (ISR/FSR). Both the Hopkins and ATLAS method were likely also not applied optimally. The BOOST2010 study was aimed to even out the playing field and make apples-to-apples comparison between top taggers possible by letting research groups apply their tagging methods themselves on a set of benchmark samples.

## 3.1 Analysis Overview

The basic criterion for tagging a boosted top quark is that the jet invariant mass should fall near $m_{\text{top}} \approx 171$ GeV. For concreteness, we consider the mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV for top jets in Sec. 3.2. The upper limit of this mass range is relatively high (about 70 GeV above $m_{\text{top}}$, while the lower limit is chosen to be only about 10 GeV below $m_{\text{top}}$), because boosted top jets often acquire additional mass from additional soft splittings. We then apply a cut on the $\tau_3/\tau_2$ ratio, where the cut is adjusted to change the relative signal tagging efficiency and background mistagging rate.

For all of the results and figures in this paper, I used the benchmark samples of the BOOST2010 study [34]. The event samples used in the study are publicly available at:

- http://www.lpthe.jussieu.fr/~salam/projects/boost2010-events/
- http://tev4.phys.washington.edu/TeraScale/boost2010/

These events were produced with HERWIG 6.510 [43] to simulate proton-proton collisions at a center-of-mass energy of 7 TeV. The generated events include a description of the underlying event. HERWIG is used in conjunction with JIMMY [44] that takes care of the underlying event generation. The samples relied on a tune from ATLAS [45] [1].

For the efficiency studies in this chapter (and for the figures in the previous chapter), I used the samples for QCD dijet production and Standard Model $t\bar{t}$ production. They are divided in sub-samples of equal size, with parton $p_T$ ranges from 200–300 GeV, 300–400 GeV, ..., 700–800 GeV. Together they yield an approximately flat parton transverse momentum distribution in a kinematic regime that is interesting for new physics searches at the LHC. It must be clarified that such a flat transverse momentum of partons is artificial (as physical cross-sections fall off roughly exponen-

[1] The parameters of this tune are set as such: PDF = MRST2007LOMOD, PTJIM = 3.6 $\times$ $(\sqrt{s}/1800\text{ GeV})^{0.274}$, JMRAD(73) = 2.2, JMRAD(91) = 2.2 and PRSOF = 0.0 (i.e. HERWIG's internal soft UE turned off).

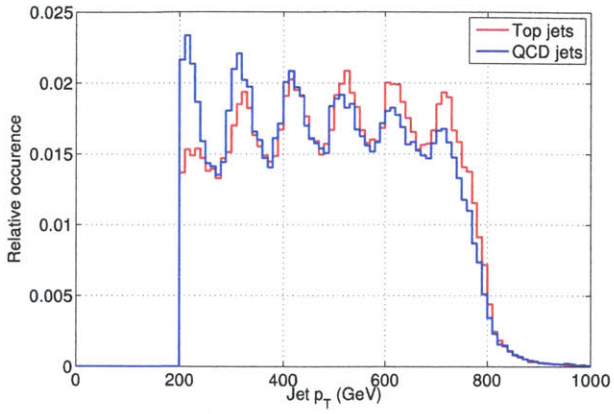
tially with $p_T$), and is only of interest as a benchmark sample to test the performance of tagging methods.

Also in accordance with the BOOST2010 proceedings, I cluster particles into jets with the anti-$k_T$ algorithm [46] with a jet radius parameter of $R = 1.0$ in `FastJet 2.4.3` [47, 48]. No simulation of detector effects is performed, but only stable particles with pseudorapidity $|\eta| < 5.0$ (except neutrinos and muons) are used. Per event, I consider a maximum of two jets with $p_T > 200$ GeV.

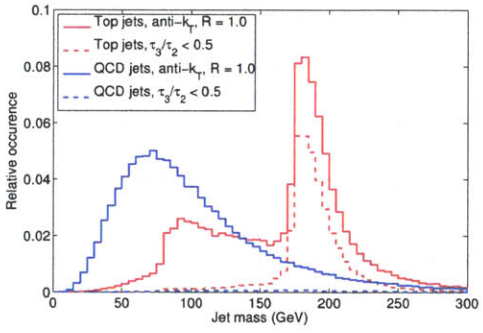
To compute $\tau_N$, the jets are first reclustered with the exclusive-$k_T$ algorithm [36, 37] into exactly $N$ candidate subjets. Subsequently, random noise is added to the rapidity-azimuth coordinates of these subjets. The resulting $N$ distorted positions then serve as input to the linear weighted $k$-means algorithm of Sec. 2.2. The above procedure is repeated 50 times (with different random seeds), and the "best" $N$ subjet direction outcomes of the algorithm (i.e. those with the lowest value of $\tilde{\tau}_N$) are used in the calculation of $\tau_N$ of the jet.

## 3.2 Boosted Top results

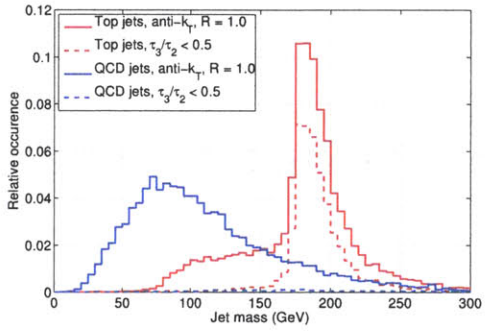
Fig. 3-1(a) shows the approximately flat jet $p_T$ distribution of the roughly $10^5$ jets analyzed jets of the combined Herwig samples with parton $p_T$ between 200–800 GeV. In Figs. 3-1(b) and 3-1(c), which show the invariant mass distribution for the combined $p_T$ sample as well as the sample with parton $p_T$ between 500–600 GeV, one can see that a $\tau_3/\tau_2$ cut decreases the background rate more than the signal rate. In addition to the large concentration of jets near or just above $m_{\text{top}}$, we see that a significant number of jets fall between roughly $m_W \approx 80$ GeV and $m_{\text{top}}$. One intuitive explanation for the occurrence of jets with such low masses is that the jet algorithm sometimes fails to capture all decay products of the top quark within the jet radius (an effect which becomes more dramatic at low transverse momenta). Indeed, we see in Figs. 3-1(b) and 3-1(c) that jets with a mass below 160 GeV generally have $\tau_3/\tau_2 > 0.5$, while a majority of jets with mass between 160 GeV and 240 GeV have $\tau_3/\tau_2 < 0.5$, a fact which strongly indicates that at least one of three decay products




(a) all $p_T$ samples




(b) all $p_T$ samples




(c) 500 GeV $< p_T <$ 600 GeV

Figure 3-1: Kinematic information of anti-$k_T$ jets. (a) Transverse momentum distribution of jets in the combined sample of parton-level $p_T$ between 200 and 800 GeV. Note that this is an unphysical $p_T$ distribution, and merely serves as a fair "testing ground" for the various tagging methods. (b),(c) Mass distributions of jets in the combined sample as well as the 500 GeV $< p_T <$ 600 GeV sample. Note that the $N$-subjettiness cut $\tau_3/\tau_2 < 0.5$ eliminates nearly all QCD jets as well as top jets with a mass much smaller than $m_{\text{top}}$, but leaves most of the top resonance peak intact.

of the top quark is missing in the low-mass jets.

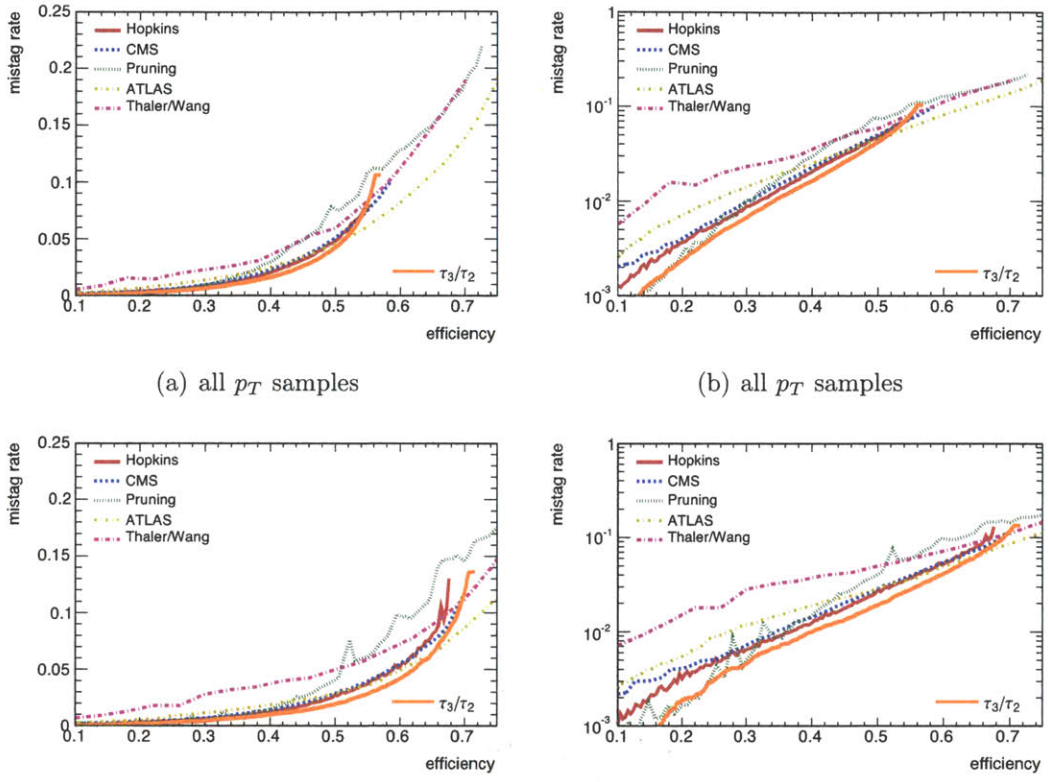
In Fig. 3-2, I show the results of a basic tagging application based on $N$-subjettiness values, namely a method only utilizing a fixed mass window of [160 GeV, 240 GeV] and a flexible $\tau_3/\tau_2$ cut. It outperforms the top tagging methods considered in the BOOST2010 study [34] for signal efficiencies below 50%. At 20% signal efficiency, the $\tau_3/\tau_2$ cut yields signal versus background improvements by a factor of 15.7 in $S/B$ and 2.37 in $S/\sqrt{B}$. Table 3.1 contains the mistag rates for the other top taggers with optimized parameters for the specific working points of 20% and 50% tagging efficiency, as well as the mistag rates at the same working points for the $N$-subjettiness method discussed above. Note that the mistag rates for the method presented in this thesis are lower than for the other top tagging methods; in Fig. 3-2(b), one can see that this statement remains true for any signal efficiencies between these working points.

| Tagger working point | Mistag rate | |
|---|---|---|
| | $\epsilon = 20$ % | $\epsilon = 50$ % |
| Hopkins | 0.4% | 4.9% |
| CMS | 0.4% | 5.2% |
| Pruning | 0.3% | 7.6% |
| ATLAS | 0.7% | 4.6% |
| Thaler/Wang | 1.5% | 6.0% |
| **$N$-subjettiness (univariate)** | **0.24%** | **4.3%** |
| **$N$-subjettiness (multivariate)** | **0.19%** | **4.0%** |

Table 3.1: Summary of mistag rates at different working points for different top-taggers. Recall that smaller mistag rates (for a fixed tagging efficiency) are desired over higher mistag rates. The univariate $N$-subjettiness method, described in Sec. 3.2, requires $\tau_3/\tau_2 < s$ and 160 GeV $< m_{\text{jet}} <$ 240 GeV, with $s$ variable. The multivariate $N$-subjettiness method, described in Sec. 3.3, requires $\tau_3/(\tau_2 - \tau_{2;0}) < s$ and $m_{\text{jet}} \in [m_{\text{min}}, m_{\text{max}}]$, with $s, \tau_{2;0}, m_{\text{min}}, m_{\text{max}}$ all variable.

## 3.3 Further Optimization

In the above efficiency study, I only considered cuts on the $\tau_3/\tau_2$ ratio combined with a fixed mass window to tag boosted tops. One could certainly generalize this approach




(a) all $p_T$ samples

(b) all $p_T$ samples

(c) 500 GeV $< p_T <$ 600 GeV

(d) 500 GeV $< p_T <$ 600 GeV

Figure 3-2: Top jet signal efficiency/background rejection plots. The orange curve corresponds to tagging rates of the method based on a sliding $N$-subjettiness cut of $\tau_3/\tau_2$ combined with a fixed mass window. The rightmost points in each plot correspond to just applying a jet invariant mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV, and points to the left of these are obtained from progressive cutting on the $\tau_3/\tau_2$ ratio only. The other curves were taken from the BOOST2010 study [34].

to include floating mass windows or more complicated cuts in the $\tau_2$–$\tau_3$ plane, or even to perform a multivariate analysis on the full set of $\tau_N$ values. Such studies are beyond the scope of the present work, but as a small step towards optimization, I consider a simple generalization of $\tau_3/\tau_2$ with a fixed mass window: a general linear cut in the $\tau_2$–$\tau_3$ plane with a floating mass window.

For $N$-subjettiness, I utilize a cut of $\tau_3/(\tau_2 - \tau_{2;0}) < s$ for a slope parameter $s \in [0, 1]$ with steps of 0.02 and an intercept parameter $\tau_{2;0} \in [0, 0.1]$ with steps of 0.01. The floating mass window $[m_{\text{min}}, m_{\text{max}}]$ was varied with $m_{\text{min}} \in [150\text{ GeV},\ 170\text{ GeV}]$ and $m_{\text{max}} \in [220\text{ GeV},\ 300\text{ GeV}]$, with steps of 2 GeV and 10 GeV, respectively. In Fig. 3-3, the gray data points are the attainable points in the efficiency space corresponding to the region of the four-dimensional parameter space $(s, \tau_{2;0}, m_{\text{min}}, m_{\text{max}})$ specified by the above parameter ranges. The orange curve is the same as the one in Fig. 3-2, i.e. it corresponds to the parameter subspace $(s, 0, 160\text{ GeV}, 240\text{ GeV})$ with $s \in [0, 1]$. Fig. 3-3 shows that for a given signal efficiency, further rejection of QCD jets by a factor of roughly 1.1 to 1.25 is possible, warranting a further exploration of multivariate methods. In particular, for the working points of 20% and 50% top signal efficiency, the mistag rates are now improved to 0.19% and 4.0%, respectively.

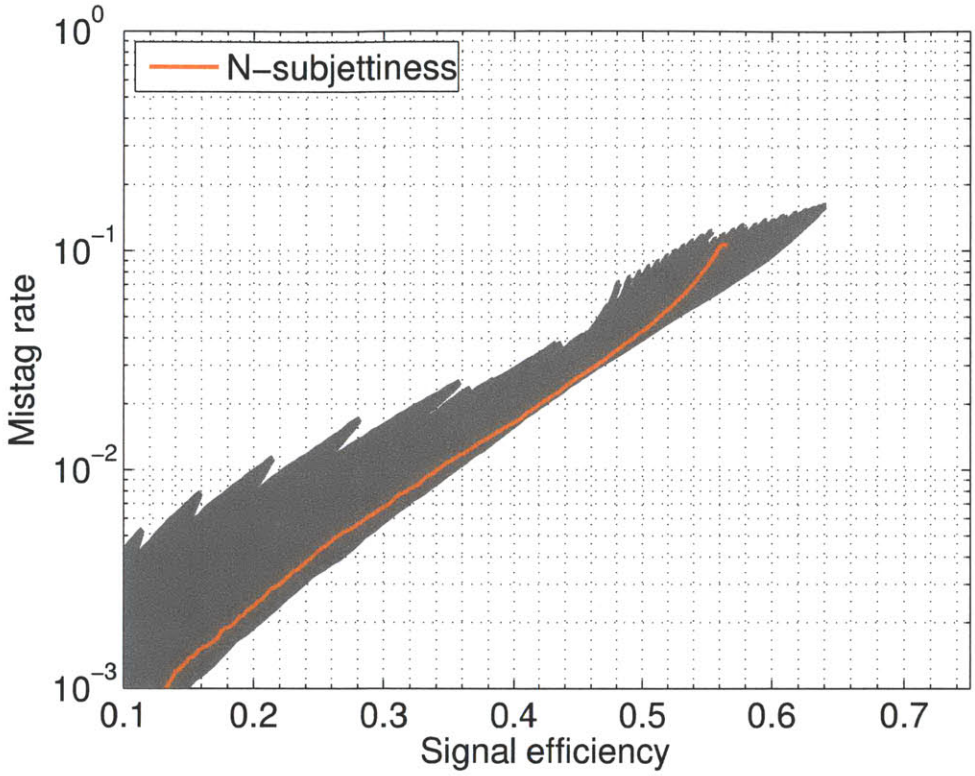



Figure 3-3: Optimized top jet signal efficiency/background rejection plot for the combined sample of jets generated by partons of transverse momenta between 200 GeV and 800 GeV. The orange curve corresponds to tagging rates of the method based on a sliding $N$-subjettiness cut of $\tau_3/\tau_2$ combined with a fixed mass window of 160 GeV $< m_{\text{jet}} <$ 240 GeV (the same curve as in Figs. 3-2(a) and 3-2(b)). The gray points correspond to tag/mistag rates for the multivariate selection method within part of the parameter space $(s, \tau_{2;0}, m_{\text{min}}, m_{\text{max}})$, as described in Sec. 3.3. Recalling that lower curves are better than higher ones, we see that modest improvements are possible with this multivariate analysis.

# Chapter 4

# Conclusions

In this thesis, I explored aspects of the inclusive jet shape $N$-subjettiness, which is designed to tag boosted hadronic objects. I found how the ratio $\tau_3/\tau_2$ is an effective variable to isolated boosted top quarks from the background of QCD jets with large invariant mass. This result can be extended to efficiently tag any $N$-prong boosted object by the ratio $\tau_N/\tau_{N-1}$ (see for example Ref. [29], where an analogous tagging technique is used for tagging boosted $W$ bosons, for which $N = 2$). Selection methods based on $N$-subjettiness methods were shown to offer better tagging efficiencies than other commonly used discriminating methods for identification of boosted objects. The discriminating power of $N$-subjettiness thus has important beneficial implications for phenomenological studies at the LHC, in particular for searches of physics beyond the standard model.

Other possible avenues for further research could be to use minimization of $N$-jettiness as a jet finding algorithm, where now the generalized $k$-means clustering algorithm of App. B is the jet algorithm. The speed and the behavior of the $k$-means algorithm certainly do not rule out an application of this kind, and the fact that the 1-subjettiness minimum does not generally line up with the anti-$k_T$ jet direction means that an $N$-subjettiness minimization method for jet finding would definitely have different (and perhaps better) behavior than the anti-$k_T$ jet algorithm in certain situations. The results of App. B also motivate going beyond "thrust-like" $\beta = 2$ measures [35]. Perhaps the linear $\beta = 1$ measure, which is more akin to jet broadening

[49] and less studied in the literature, deserves a more prominent place in jet physics.

$N$-subjettiness exhibits several desirable properties which warrant further experimental and theoretical investigations. On the experimental side, $\tau_N$ can be calculated on a jet-by-jet basis and thereby offers considerable flexibility of application. While we focused just on ratios of $\tau_N$ as discriminating variables, multivariate optimization along the lines of Sec. 3.3 could improve signal efficiency and background rejection even further. In addition, some of the $N$-subjettiness variations mentioned in App. B might also be effective discriminating variables by themselves or in combination. On the theoretical side, $\tau_N$ is an infrared- and collinear-safe inclusive jet shape which can be defined without the need for an algorithmic subjet finding method. Thus, the prospects for theoretical calculations involving $N$-subjettiness look promising both using fixed-order perturbative calculations and using resummation techniques.

With the first LHC data on the books, the search for new physics is already underway. New phenomena may be revealed in the production of highly boosted top quarks and/or electroweak bosons, and I expect that $N$-subjettiness will prove to be a useful variable for exploring such extreme kinematic regimes.

# Appendix A

# The Unique Minimum of a Sum of Cones

## A.1 Relevance to $N$-subjettiness and Linear $k$-means Clustering

In the motivation for the linear, weighted $k$-means algorithm at the beginning of Sec. 2.2, I claimed that the distribution of

$$\tilde{\tau}_1(y,\phi) = \frac{1}{d_0}\sum_{i=1}^{|C|} p_{T,i}\sqrt{(y-y_i')^2+(\phi-\phi_i')^2}$$

for a cluster $C$ of finitely many particles (i.e. $|C| < \infty$) has a unique minimum with probability one. This statement is equivalent to saying that for any local minimum $(y^*, \phi^*)$ of $\tilde{\tau}_1(y,\phi)$ for a cluster of finitely many particles, we must have that

$$\tilde{\tau}_1(y^*,\phi^*) = \tau_1.$$

This statement has profound consequences for the clustering method presented in Sec. 2.2. It means that whenever the linear, weighted $k$-means converges for $N = k = 1$, it converges to the global minimum of $\tilde{\tau}_1(y,\phi)$, so we can trivially read off $\tau_1$!

Similarly, it implies that the calculation of $\tau_N$ for $N = k \geq 2$ with weighted, linear $k$-means is essentially reduced to finding the partition of the cluster $C$ of particles into $N$ number of non-empty subclusters $C^{(j)}$ which yield the lowest $\tilde{\tau_N}$ value.

More precisely, Theorem A.2.1 and the fact that in practice the linear, weighted $k$-means algorithm always converges to arbitrary precision, together imply that computing $\tau_N$ of a jet amounts to finding the partition of the cluster

$$C = \{C^{(1)}, C^{(2)}, \ldots, C^{(N)}\}$$

into non-empty subclusters $C^{(j)}$ which gives rise to the lowest value of $\tau_1^{(1)} + \tau_1^{(2)} + \cdots + \tau_1^{(N)}$, where the $\tau_1^{(j)}$ are the values at the (now easily calculable) global minima of the $\tilde{\tau_1}(y, \phi)$ distributions associated with their respective subclusters $C^{(j)}$.

In practice, the number of local minima of $\tau_N$ is relatively small. Crucially, iterating the linear, weighted $k$-means algorithm only 10 times with different random seeds for each jet seemed to be enough to find the global minimum in virtually all cases, as repeating it more did not generally lower the found $N$-subjettiness values. As mentioned in Sec. 3.1, I rerun the algorithm 50 times per jet in all of my studies (to err on the conservative side).

## A.2 Statement and Proof of Theorem

Let me first introduce a definition which will make the language (and hopefully the intuition) in this section easier.

**Definition** Any function $z(\mathbf{x}) = z(x^1, x^2, \ldots, x^n) = s\sqrt{\sum_{d=1}^{n}(x^d - m^d)^2}$ is defined to be a cone on $\mathbb{R}^n$ with slope $s > 0$ and center $\mathbf{m} = (m^1, m^2, \ldots, m^n)$. A sum of cones $z_i$ with slopes $s_i$ and centers $\mathbf{m}_i$ on $\mathbb{R}^n$ is defined as the function $z(\mathbf{x}) = \sum_i z_i(\mathbf{x}) = \sum_i s_i\sqrt{\sum_{d=1}^{n}(x_i^d - m_i^d)^2}$.

Comparing the above definition for $n = 2$ with Eq. (2.2) for $N = 1$, in the language of jets, cones can be identified with particles, slopes with transverse momenta, $\mathbb{R}^2$ as

the rapidity-azimuth plane[1] and the sum of cones as $\tilde{\tau}_1$.

To justify the claim made in Sec. 2.2 and reiterated in the section above, we are to prove the following theorem.

**Theorem A.2.1.** *A finite sum of arbitrary cones on $\mathbb{R}^2$ has, with probability one, a unique local minimum. Furthermore, this local minimum is the global minimum.*

As a warm-up case, let us consider the analogous case for $n = 1$ first.

**Lemma A.2.2.** *A finite sum of arbitrary cones on $\mathbb{R}^1$ has, with probability one, a unique local minimum. Furthermore, this local minimum is the global minimum.*

*Proof.* Observe that the sum of cones can be represented by the function

$$z(x) = \sum_i s_i |x - m_i|. \tag{A.1}$$

Note that $z(x) \to +\infty$ as $x \to \pm\infty$, so it must have a global minimum. Its derivative is

$$z'(x) = \begin{cases} -\sum\limits_{i|x<m_i} s_i + \sum\limits_{i|x>m_i} s_i & : x \notin M \\ \text{undefined} & : x \in M \end{cases}$$

where $M$ is the set of minima $(m_i)$. We see that $z'(x)$ behaves as an increasing step function. Now take $x = x^*$ to be the global minimum. By definition, we must have $z'(x^* + \varepsilon) > 0$ and $z'(x^* - \varepsilon) < 0$ for infinitesimal $\varepsilon > 0$ with probability one (a flat sum is a zero probability event for a finite number of arbitrary cones). But because the derivative of the sum of cones behaves as an increasing step function, we must then have $z'(x) > 0$ for any $x > x^*$ and $z'(x) < 0$ for any $x < x^*$, ruling out local minima other than the global minimum $x = x^*$. □

We are now ready to prove Theorem A.2.1.

[1]Note that technically the restricted rapidity-azimuth plane is, unlike $\mathbb{R}^2$, topologically equivalent to an annulus. This technicality does not affect the equivalence with the claim of Sec. 2.2, as a jet only occupies a small patch of the whole $(y, \phi)$ space, and can thus be thought of as living on a patch which is topologically equivalent to a disk.

*Proof.* Again, because a sum of cones on $\mathbb{R}^2$ tends to infinity far away from the cluster centers, the sum must have a global minimum $(x^{*1}, x^{*2})$. Now consider any line $L \subset \mathbb{R}^2$ which contains $(x^{*1}, x^{*2})$, and let this line $L$ be parametrized by the coordinate

$$x = \begin{cases} \sqrt{(x^1 - x^{*1})^2 - (x^2 - x^{*2})^2} & : x \in L^+ \\ -\sqrt{(x^1 - x^{*1})^2 - (x^2 - x^{*2})^2} & : x \in L^- \end{cases} \tag{A.2}$$

where $L^+$ and $L^-$ are both halflines, with $L^+ \cup L^- = L$, $L^+ \cap L^- = \emptyset$ and $(x^{*1}, x^{*2}) \in L^+$. Let $z_L(x)$ be the restriction of $z(\mathbf{x})$ to the line $L$. In this construction, it is obvious that for $(x^{*1}, x^{*2})$ to be a global minimum, we must have $z'_L(\varepsilon) > 0$ and $z'_L(-\varepsilon) < 0$ for infinitesimal $\varepsilon > 0$.

We now calculate each cone's contribution to $z_L(x)$. If the cone's center $\mathbf{m}_i \in L$, then its contribution is just $z_{L,i}(x) = s_i|x - m_i|$, where $m_i$ is the distance between $\mathbf{m}$ and $\mathbf{x}^*$ if $\mathbf{m} \in L^+$ (and negative the distance if $\mathbf{m} \in L^-$). If the cone's center is not on $L$, then its contribution to $z_L$ will be a conic section, more specifically the positive leaf of a hyperbola, so it will be of the form

$$z_{L,i}(x) = a_i \sqrt{1 + \frac{(x - m_i)^2}{b_i^2}} \tag{A.3}$$

with derivative

$$z'_{L,i}(x) = \frac{a_i}{b_i^2} \frac{x - m_i}{\sqrt{1 + \frac{(x-m_i)^2}{b_i^2}}}, \tag{A.4}$$

where $m_i$ is the perpendicular projection of $\mathbf{m}_i$ on $L$, $a_i$ is the distance between $\mathbf{m}_i$ and its perpendicular projection on $L$ times $s_i$, and $b_i = s_i^{-1}$. What matters is that $z'_{L,i}(x)$ is a strictly increasing function.

Hence the derivative of the sum of cones restricted to the line $L$ is the sum of an increasing step function and a strictly increasing function, and consequently $z'_L(x)$ is strictly increasing on the whole line $L$ (apart from points which coincide with cone centers, where it is undefined). As in the proof of Lemma A.2.2, we conclude that nowhere on $L$ can we find another local minimum. Since $L$ was arbitrary, this must hold for any line through the global minimum. But the set of lines through $(x^{*1}, x^{*2})$

covers all of $\mathbb{R}^2$, so no other local minimum other than the global minimum can exist. This proves the theorem. □

**Remark** Note that the crucial element of the proof of the 2D case above is that the first derivative of any "vertical slice" of the potential function for each particle (in the case of conic potential functions, these slices were either hyperbolas or 1D cones) is a weakly increasing function. Hence the proof naturally extends to potential functions for which the second derivative on any vertical slice is positive everywhere (or zero, as long as the first derivative does not make decreasing steps). Therefore, similar arguments as above would also prove that the *sum of any finite sum of paraboloids on* $\mathbb{R}^2$ *has only one minimum.* (In fact, in this case the second derivative of any slice is strictly positive, so there cannot even be a degenerate line of minima.)

The proof no longer holds for potential functions for which the second derivative is negative on some vertical slices, or for which the first derivative makes a decreasing step. This category of potential functions includes those where the potential grows slower than linearly in the separation from the particle's positions (if it grows as the square root of the separation, for example), and also the Snowmass cone potential used in many iterative cone algorithms, which is shaped as a paraboloid for small separations but is flat for separations large than a characteristic cone radius. The fact that in the latter local minima can be created by infinitesimally soft particles is indeed the Achilles' heel of many cone algorithms (such as the CDF Midpoint Cone Algorithm [50]), as infrared safety is compromised. Other (seedless) cone algorithms, most notably SISCone [51], have since found a way around this problem.

# Appendix B

# Generalized Formulations of $N$-subjettiness and Minimization

## B.1 Generalized Definition of $N$-subjettiness

The definition of $N$-subjettiness in Eqs. (2.1) and (2.2) is not unique, and different choices for $\tilde{\tau}_N$ can be used to give different weights to the emissions within a jet. These generalizations of $N$-subjettiness are similar to different "angularities" [52] used in $e^+e^- \to$ hadrons measurements.

Analogously to Refs. [29, 30], a general $N$-subjettiness measure is

$$\tau_N^{\text{gen}} = \min_{N \text{ subjet directions}} \tilde{\tau}_N^{\text{gen}} \tag{B.1}$$

$$\tilde{\tau}_N^{\text{gen}} = \frac{1}{d_0} \sum_i \min_J \left\{ d(p_J, p_i) \right\}, \tag{B.2}$$

where $d_0$ is a normalization factor, $J$ runs over the $N$ candidate subjets, and $d(p_J, p_i)$ is a distance measure between a candidate subjet $p_J$ and a jet constituent $p_i$. Like in Sec. 2.2, one needs a method to figure out the value of $\tau_N$ by minimizing $\tilde{\tau}_N$ over possible candidate subjets $p_J$.

There are many choices for $d(p_J, p_i)$, but a nice two-parameter, longitudinally

boost-invariant choice for the distance measure is

$$d^{\alpha,\beta}(p_J, p_i) = p_{T,i} \left(p_{T,J}\right)^{\alpha} \left(\Delta R_{J,i}\right)^{\beta}. \tag{B.3}$$

In the next section, I present a minimization-based subjet-finding method for $\alpha = 0$ and $0 < \beta \leq 2$. If desired, one could replace $p_{T,J}$ with $E_{T,J} = \sqrt{p_{T,J}^2 + m_J^2}$ to include information about the subjet mass; similarly, one could also use $E_{T,i}$ instead of $p_{T,i}$ to include the mass of the jet constituents. For $e^+e^-$ applications, one would replace the transverse momentum $p_T$ with the total momentum $|\mathbf{p}|$ (or the energy $E$) and $\Delta R$ with the opening angle $\Delta\Omega$. A natural choice for the normalization factor to keep $0 < \tau_N < 1$ is

$$d_0 = \max_J \left\{\left(p_{T,J}\right)^{\alpha}\right\} \left(R_0\right)^{\beta} \sum_i p_{T,i}, \tag{B.4}$$

where $R_0$ is the characteristic jet radius.

By making $d(p_J, p_i)$ linear in $p_{T,i}$, $\tau_N$ is automatically an infrared-safe observable. Collinear-safety requires linearity in $p_{T,i}$ as well, but imposes the addition requirement that $\beta \geq 0$. The value of $\alpha$ is unconstrained. Of course, I am assuming that the candidate subjet finding method is also infrared- and collinear-safe.

In the body of the paper, we used $\alpha = 0$, $\beta = 1$. This choice corresponds to treating each subjet democratically, and using a $k_T$-like distance measure. This distance measure makes $\tau_N$ similar to jet broadening [49],[1] and I found that this was an effective choice for boosted object identification. By varying $\beta$, we can change the angular weighting. A thrust-like [35] weighting corresponds to $\beta = 2$, while other angularities [52] with $-\infty < a < 2$ are given by $\beta = 2 - a$. By varying $\alpha$, we can weight the distance measure by the hardness of the subjet directions. Large positive (negative) $\alpha$ means that the minimum in Eq. (B.2) is given by the distance to the hardest (softest) candidate subjet. Further studies of boosted object identification using different values of $\alpha$ and $\beta$ would be interesting, since studies of jet angularities have shown that additional information about jet substructure can be gleaned by

[1] By similar, we mean the distance measure has the same $\Delta R_{J,i} \to 0$ limit. Because thrust-like observables are defined in a preferred rest frame and we are working with a longitudinally boost-invariant measure, the correspondence is inexact.

combining different angular information [32].

## B.2 Generalized Weighted $k$-means Clustering Algorithm

For one class of generalizations of $N$-subjettiness as in Eqs. (B.1) and (B.2) for which $\alpha = 0$ in the distance measure of Eq. (B.3),

$$\tau_N^{(\beta)} = \min_{N \text{ subjet directions}} \tilde{\tau}_N{}^{(\beta)}, \tag{B.5}$$

$$\tilde{\tau}_N{}^{(\beta)} = \frac{1}{d_0} \sum_i p_{T,i} \min_J \left( (y_i - y_{0,J})^2 + (\phi_i - \phi_{0,J})^2 \right)^{\frac{\beta}{2}}, \quad (0 < \beta \leq 2) \tag{B.6}$$

one can define a generalized weighted $k$-means clustering algorithm. The linear weighted $k$-means algorithm of Sec. 2.2 was for the specific case $\beta = 1$. For general $\beta$, similar (heuristic) arguments can be given for the formulation of the algorithm. In particular, the assignment step shall be the same, but now the angular weighting parametrized by $\beta$ will require slightly different update steps. See Sec. 2.2 for how I motivated the update step for $\beta = 1$ by setting the partial derivative of $\tilde{\tau}_N$ equal to zero and requiring a convergence condition at the minimum. A completely analogous derivation produces the following generalized method for performing the minimization in Eq. (B.5).

**General weighted $k$-means algorithm.** ($N = k$)

1. *Initialization step.* Pick initial subjet directions $\left(y_{0,J}^{(0)}, \phi_{0,J}^{(0)}\right)_{J\in\{1,\ldots,N\}}$.

2. *Assignment step.* Divide the jet into clusters $\left(C_J^{(n)}\right)_{J\in\{1,\ldots,N\}}$ by assigning the particles to the closest subjet direction. In other words, $i \in C_J^{(n)}$ if and only if

$$\sqrt{\left(y_i - y_{0,J}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,J}^{(n)}\right)^2} < \sqrt{\left(y_i - y_{0,M}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,M}^{(n)}\right)^2} \quad \forall M \neq J. \tag{B.7}$$

3. *Update step.* Based on the previous subjet directions $\left(y_{0,J}^{(n)}, \phi_{0,J}^{(n)}\right)_{J\in\{1,\ldots,N\}}$, calculate new subjet directions $\left(y_{0,J}^{(n+1)}, \phi_{0,J}^{(n+1)}\right)_{J\in\{1,\ldots,N\}}$ as follows:

$$y_{0,J}^{(n+1)} = \frac{\sum_{i\in C} \frac{p_{T,i} y_i}{\left(\left(y_i - y_{0,J}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,J}^{(n)}\right)^2\right)^{\frac{2-\beta}{2}}}}{\sum_{j\in C} \frac{p_{T,j}}{\left(\left(y_j - y_{0,J}^{(n)}\right)^2 + \left(\phi_j - \phi_{0,J}^{(n)}\right)^2\right)^{\frac{2-\beta}{2}}}}, \phi_{0,J}^{(n+1)} = \frac{\sum_{i\in C} \frac{p_{T,i} \phi_i}{\left(\left(y_i - y_{0,J}^{(n)}\right)^2 + \left(\phi_i - \phi_{0,J}^{(n)}\right)^2\right)^{\frac{2-\beta}{2}}}}{\sum_{j\in C} \frac{p_{T,j}}{\left(\left(y_j - y_{0,J}^{(n)}\right)^2 + \left(\phi_j - \phi_{0,J}^{(n)}\right)^2\right)^{\frac{2-\beta}{2}}}} \tag{B.8}$$

4. *Iteration.* Repeat Steps (b) and (c) until the average directional change of the subjets $\overline{\Delta}^{(n+1)}$ is smaller than some threshold $\delta_p$ (which encodes the desired precision), i.e. until

$$\overline{\Delta}^{(n+1)} \equiv \frac{1}{N} \sum_{J=1}^{N} \sqrt{\left(y_{0,J}^{(n+1)} - y_{0,J}^{(n)}\right)^2 + \left(\phi_{0,J}^{(n+1)} - \phi_{0,J}^{(n)}\right)^2} < \delta_p. \quad (n \geq 0) \tag{B.9}$$

It is instructive to think about the different powers $\beta$ as "distance weights". Note that for $\beta = 2$, the update step simplifies dramatically, as all particles are assigned equal distance weights.[2] For $\beta < 2$, more weight is put on particles closer to the can-

[2]In the computer science literature, the term "(weighted) $k$-means algorithm" always refers to this specific $\beta = 2$ case. To my best knowledge, the generalization to different powers $\beta$ has never

didate subjet directions. If $\beta > 2$, more weight is put on particles that are far away, as the power $\frac{2-\beta}{2}$ now becomes negative. Besides providing no useful physical interpretation, it also has the consequence that the algorithm no longer converges. The $\beta = 0$ case is simply not interesting because $\tau_N^{(0)}$ would no longer include positional information, and for $\beta < 0$ we would have to give up collinear safety. Hence the only physically relevant measures are those in which the angular dependence scales with powers $\boxed{0 < \beta \leq 2}$. I should comment that the cases in which $0 < \beta < 1$ may not be very practical, as the remark at the end of Sec. A.2 shows that for these angular dependencies, $\tilde{\tau}_1$ can in principle have many local minima. Therefore, I suspect that only the cases for which $1 \leq \beta \leq 2$ are both practical and feasible.

One could also interpret the separation dependence of the distance weight as a quantification of the probability that a particle actually belongs to a jet. In this interpretation, one could justify giving less weight to particles that are relatively far away. After all, these particles are more likely to originate from initial state radiation or from radiation of other nearby jets; one could thus make the argument that they should have less influence on the position of the subjets of a jet.

## B.3 Comparison of Subjet Finding Methods

The generalizations of the previous two sections widen the range of possible selection methods. For example, one could attempt to redo the analysis of Chapter 3 with $\tau_3^\beta/\tau_2^\beta$ as a top-tagger, where $\beta$ can now take on any value between 0 and 2. One could also imagine using a combined selection method where jets are selected on their "quadratic" ($\beta = 2$) and "linear" ($\beta = 1$) $N$-subjettiness values.

In this section, I succinctly present three alternatives to using linear $k$-means clustering and $N$-subjettiness. I consider the two cases where $k$-means is used to find candidate subjets, but now with $\beta = 1/2$ and $\beta = 2$. The third alternative is to use the subjet directions found by reclustering the jets with the exclusive $k_T$ algorithm in Eq. (2.2), which was the method utilized in [29]. I consider three simple discriminants

been explored before, though $k$-medians clustering [53] is somewhat similar to the linearized version of $k$-means but unfortunately not rotationally invariant.

corresponding to these three subjet-finding methods (in addition to a mass window of [160 GeV, 240 GeV]): $\tau_3^{(1/2)}/\tau_2^{(1/2)}$, $\tau_3^{(2)}/\tau_2^{(2)}$ and $\tilde{\tau}_3^{(k_T)}/\tilde{\tau}_2^{(k_T)}$, respectively. Fig. B-1, which is otherwise entirely analogous to Fig. 3-2, displays the results of these methods. The three alternative discriminants and associated subjet finding methods are seen to be less powerful than the linear $N$-subjettiness and $k$-means clustering method considered in Chapter 3.

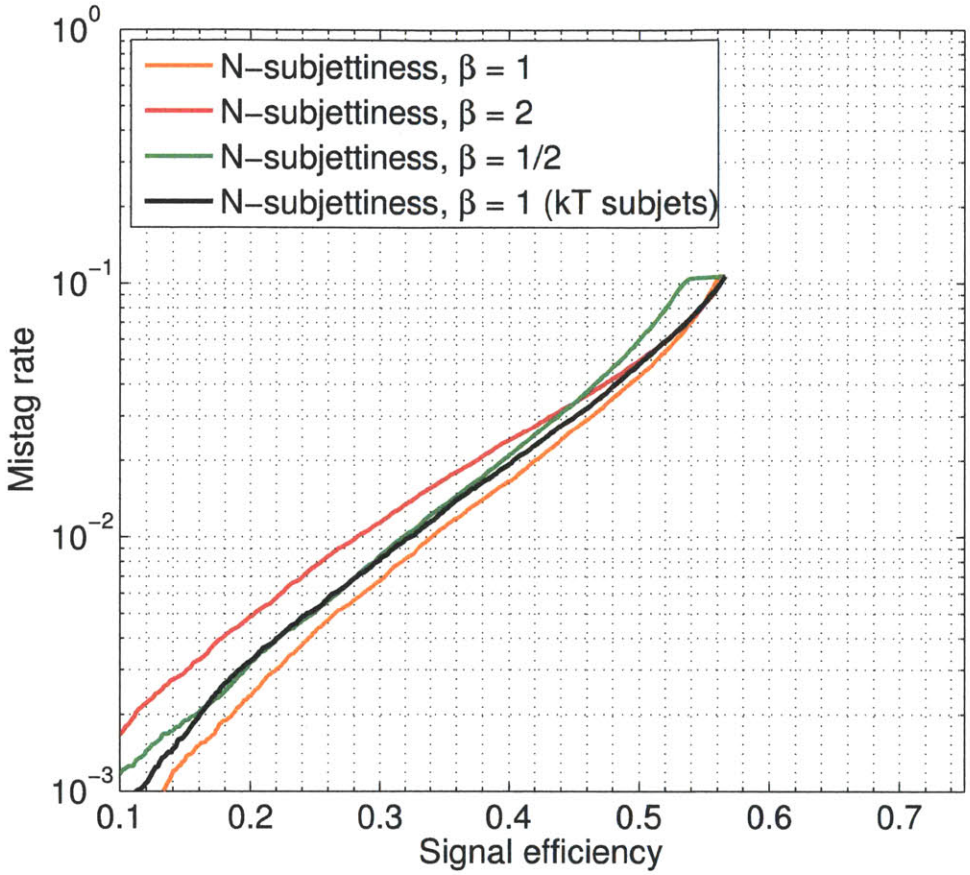



Figure B-1: Top jet signal efficiency/background rejection plots for the combined sample of jets generated by partons of transverse momenta between 200 GeV and 800 GeV, for the four selection methods of Sec. B.3. All methods first impose a fixed mass window 160 GeV $< m_{\text{jet}} <$ 240 GeV and impose progressive sliding cuts on the ratio of their 3-subjettiness and 2-subjettiness measures. One can see that the classical "quadratic" $k$-means algorithm and its discriminant ($\beta = 2$) do worst, while their linear variants ($\beta = 1$) discussed in the body of the thesis do best. However, lower values of $\beta$ do not further improve top identification, and even do worse (see e.g. the curve for $\beta = 1/2$). We also observe that using the linear $k$-means minimization procedure over exclusive $k_T$ subjet finding yields significant improvements across the entire efficiency space.

# Appendix C

# Additional Event Displays

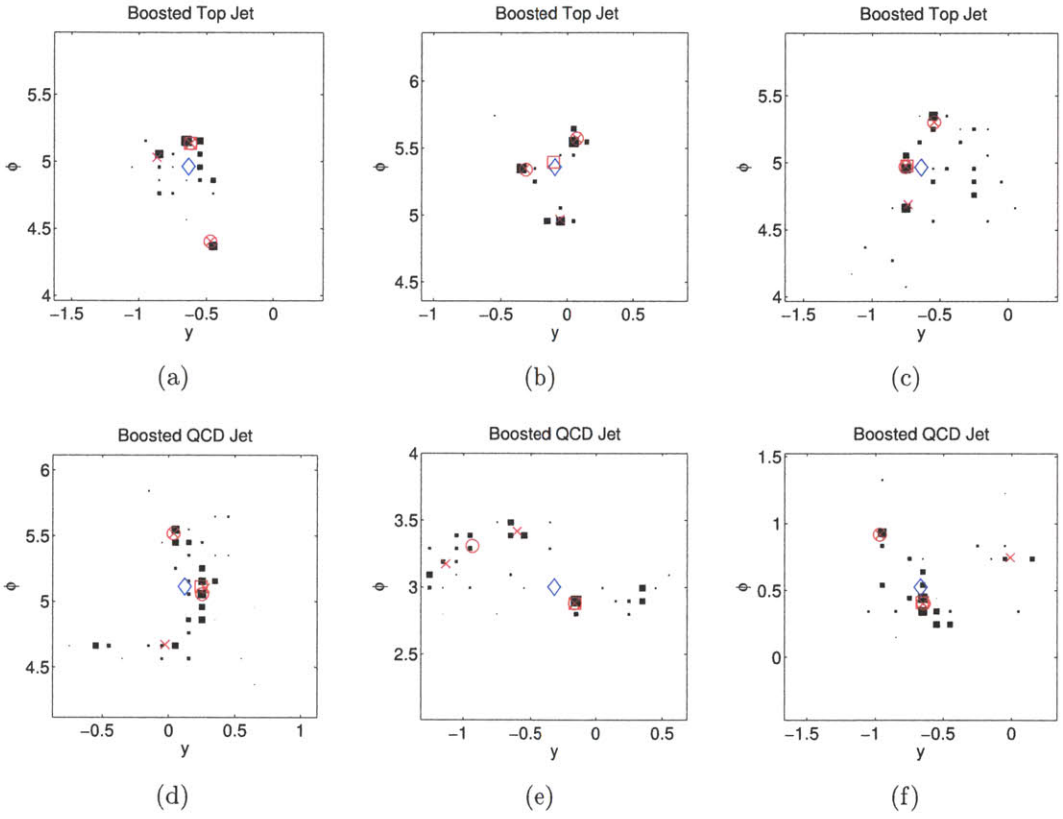



Figure C-1: Top row: boosted top jets. Bottom row: QCD jets with invariant mass close to $m_{\text{top}}$. The coloring and labeling is the same as in Fig. 2-2.

# Bibliography


[1] M. H. Seymour, Z. Phys. C **62**, 127 (1994).

[2] J. M. Butterworth, B. E. Cox and J. R. Forshaw, Phys. Rev. D **65**, 096014 (2002) [arXiv:hep-ph/0201098].

[3] J. M. Butterworth, A. Davison, E. Ozcan, and P. Sherwood, ATL-PHYS-INT-2007-015.

[4] G. Brooijmans, ATL-COM-PHYS-2008-001 (2008).

[5] **ATLAS** Collaboration, **ATL-PHYS-PUB-2009-081**.

[6] J. Thaler and L. T. Wang, JHEP **0807**, 092 (2008) [arXiv:0806.0023 [hep-ph]].

[7] D. E. Kaplan, K. Rehermann, M. D. Schwartz and B. Tweedie, Phys. Rev. Lett. **101**, 142001 (2008) [arXiv:0806.0848 [hep-ph]].

[8] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung and J. Virzi, Phys. Rev. D **79**, 074017 (2009) [arXiv:0807.0234 [hep-ph]].

[9] J. M. Butterworth, A. R. Davison, M. Rubin and G. P. Salam, Phys. Rev. Lett. **100**, 242001 (2008) [arXiv:0802.2470 [hep-ph]].

[10] J. M. Butterworth, A. R. Davison, M. Rubin and G. P. Salam, arXiv:0810.0409 [hep-ph].

[11] S. D. Ellis, C. K. Vermilion and J. R. Walsh, Phys. Rev. D **80**, 051501 (2009) [arXiv:0903.5081 [hep-ph]].

[12] S. D. Ellis, C. K. Vermilion and J. R. Walsh, Phys. Rev. D **81**, 094023 (2010) [arXiv:0912.0033 [hep-ph]].

[13] D. Krohn, J. Thaler and L. T. Wang, JHEP **1002**, 084 (2010) [arXiv:0912.1342 [hep-ph]].

[14] D. E. Soper and M. Spannowsky, JHEP **1008**, 029 (2010) [arXiv:1005.0417 [hep-ph]].

[15] J. M. Butterworth, J. R. Ellis, A. R. Raklev and G. P. Salam, Phys. Rev. Lett. **103**, 241803 (2009) [arXiv:0906.0728 [hep-ph]].

[16] T. Plehn, G. P. Salam and M. Spannowsky, Phys. Rev. Lett. **104**, 111801 (2010) [arXiv:0910.5472 [hep-ph]].

[17] G. D. Kribs, A. Martin, T. S. Roy and M. Spannowsky, Phys. Rev. D **81**, 111501 (2010) [arXiv:0912.4731 [hep-ph]].

[18] J. Gallicchio and M. D. Schwartz, Phys. Rev. Lett. **105**, 022001 (2010) [arXiv:1001.5027 [hep-ph]].

[19] C. R. Chen, M. M. Nojiri and W. Sreethawong, JHEP **1011**, 012 (2010) [arXiv:1006.1151 [hep-ph]].

[20] A. Falkowski, D. Krohn, L. T. Wang, J. Shelton and A. Thalapillil, arXiv:1006.1650 [hep-ph].

[21] G. D. Kribs, A. Martin, T. S. Roy and M. Spannowsky, arXiv:1006.1656 [hep-ph].

[22] L. G. Almeida, S. J. Lee, G. Perez, G. Sterman and I. Sung, Phys. Rev. D **82**, 054034 (2010) [arXiv:1006.2035 [hep-ph]].

[23] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, *JHEP* **10** (2010) 078, [arXiv:1006.2833 [hep-ph]]

[24] B. Bhattacherjee, M. Guchait, S. Raychaudhuri and K. Sridhar, Phys. Rev. D **82**, 055006 (2010) [arXiv:1006.3213 [hep-ph]].

[25] K. Rehermann and B. Tweedie, arXiv:1007.2221 [hep-ph].

[26] C. Hackstein and M. Spannowsky, arXiv:1008.2202 [hep-ph].

[27] C. Englert, C. Hackstein and M. Spannowsky, arXiv:1010.0676 [hep-ph].

[28] A. Katz, M. Son and B. Tweedie, arXiv:1010.5253 [hep-ph].

[29] J. Thaler and K. Van Tilburg, JHEP **1103**, 015 (2011) [arXiv:1011.2268 [hep-ph]].

[30] I. W. Stewart, F. J. Tackmann and W. J. Waalewijn, Phys. Rev. Lett. **105**, 092002 (2010) [arXiv:1004.2489 [hep-ph]].

[31] J. H. Kim, Phys. Rev. D **83**, 011502 (2011) [arXiv:1011.1493 [hep-ph]].

[32] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig and C. Lee, arXiv:1001.0014 [hep-ph].

[33] A. Banfi, M. Dasgupta, K. Khelifa-Kerfa and S. Marzani, JHEP **1008**, 064 (2010) [arXiv:1004.3483 [hep-ph]].

[34] A. Abdesselam *et al.*, arXiv:1012.5412 [hep-ph].

[35] E. Farhi, Phys. Rev. Lett. **39**, 1587 (1977).

[36] S. Catani, Y. L. Dokshitzer, M. H. Seymour and B. R. Webber, Nucl. Phys. B **406**, 187 (1993).

[37] S. D. Ellis and D. E. Soper, Phys. Rev. D **48**, 3160 (1993) [arXiv:hep-ph/9305266].

[38] S. P. Lloyd, *Least Squares Quantization in PCM, IEEE Transactions on Information Theory* **IT-28**, 2 (1982).

[39] **CMS** Collaboration, R. Adolphi *et al.*, *CMS Physics Analysis Summary* **CMS-PAS-EXO-09-002** (2009).

[40] **CMS** Collaboration, R. Adolphi *et al.*, *CMS Physics Analysis Summary* **CMS-PAS-JME-09-001** (2009).

[41] T. Sjostrand, S. Mrenna and P. Z. Skands, JHEP **0605**, 026 (2006) [arXiv:hep-ph/0603175].

[42] T. Sjostrand, S. Mrenna and P. Z. Skands, Comput. Phys. Commun. **178**, 852 (2008) [arXiv:0710.3820 [hep-ph]].

[43] G. Corcella *et al.*, *HERWIG 6.5 release note*, [arxiv:0210213 [hep-ph]].

[44] J. M. Butterworth, J. R. Forshaw, and M. H. Seymour, *Z. Phys.* **C72** (1996) 637–646, [arxiv:9601371 [hep-ph]].

[45] **ATLAS** Collaboration, Tech. Rep. **ATL-PHYS-PUB-2010-002**, CERN, Geneva, March, 2010.

[46] M. Cacciari, G. P. Salam and G. Soyez, JHEP **0804**, 063 (2008) [arXiv:0802.1189 [hep-ph]].

[47] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet." http://fastjet.fr/.

[48] M. Cacciari and G. P. Salam, Phys. Lett. B **641**, 57 (2006) [arXiv:hep-ph/0512210].

[49] S. Catani, G. Turnock and B. R. Webber, Phys. Lett. B **295**, 269 (1992).

[50] G. C. Blazey *et al.*, arXiv:hep-ex/0005012.

[51] G. P. Salam and G. Soyez, JHEP **0705**, 086 (2007) [arXiv:0704.0292 [hep-ph]].

[52] C. F. Berger, T. Kucs and G. F. Sterman, Phys. Rev. D **68**, 014012 (2003) [arXiv:hep-ph/0303051].

[53] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*: Prentice-Hall, 1981.