Top Tagging Using Spatial Distribution of Subjets

Danielle Wilson
Supervisor: Professor Deepak Kar

University of the Witwatersrand

High Energy Particle Physics (HEPP) Workshop 2020

January 29, 2020

University of the Witwatersrand, Johannesburg

Outline

- Introduction
  - Jets
  - Boosted Objects
  - Top Jets
  - QCD Jets
- Classifying Events Spatially
  - Notation
  - Identifying Configurations of Subjets
  - Probabilities of Configurations
- Creating the Tagger
  - Combining Probabilities
  - How did it perform?
  - Shortcomings of the Tagger

Introduction

Jets

- Quarks and gluons do not occur freely in nature
  - Immediately after production, they fragment and hadronise $\implies$ collimated shower of energetic hadrons which is referred to as a jet
- Can identify original "parton" by measuring jet energy and direction
  - Concept of "parton" is ambiguous
    $\implies$ Jets must be well defined
    - Jets defined by algorithm used to assemble them and a radius parameter
    - No single universal definition
- Majority of events at the LHC contain jets

Boosted Objects

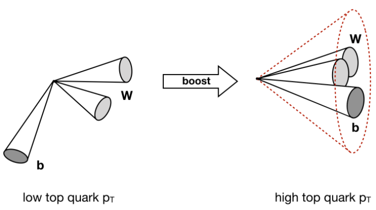
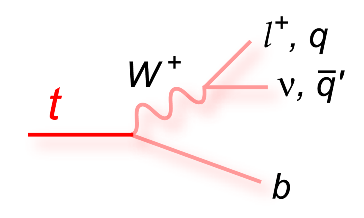
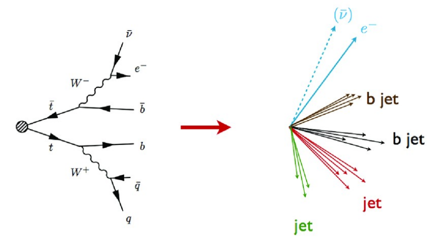


Figure: Illustration of the decay of a boosted top quark. *Ref:* ***arXiv:1712.01391***

- If heavy particles are produced with large transverse momentum $\Longrightarrow$ decay products collimated into a single large-radius jet
- Large-radius jets contain intricate substructure
  - Observables are constructed to characterise this substructure
- Using one or more of these observables to identify boosted objects is referred to as tagging

Top Jets





- Top quark has a very short lifetime $\implies$ decays before it hadronises $\implies$ unique opportunity to study bare quarks
- Top quarks decay mainly via $t \rightarrow Wb$
  - $W \rightarrow q\hat{q}$ occurs 67% of the time
  - $W \rightarrow l\nu$ has a branching ratio of 11% for each lepton flavor
  - Pairs of top quarks : 45% hadronic, 35% semileptonic, rest are dileptonic and hadronic tau decays

Top Jets

- Standard methods to identify Top Quarks:
  - B-tagging
  - Identifying W boson
  - Invariant mass of 3 jets is comparable to the top mass
- Highly boosted top quarks

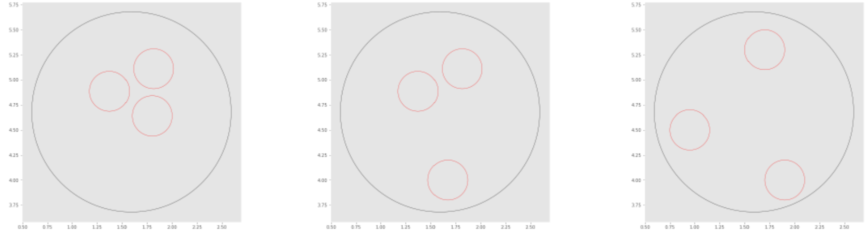
  $\Longrightarrow$ Standard methods are hindered

  $\Longrightarrow$ Jet substructure analysis is the natural next step

QCD Jets

- Efficient top taggers discriminate features unique to top quarks from those of the background
- QCD jets describe the background
  - They orginate from high p_T light quarks or gluons that shower into many soft and collinear particles
    $\implies$ Not easily resolved

Classifying Events Spatially



(a) Subjets arranged in a '123' Configuration

(b) Subjets arranged in a '12 3' Configuration

(c) Subjets in arranged in a '1 2 3' Configuration

Figure: Illustration of the notation used in this investigation

- Plots in the eta-phi plane
- Large radius jets have a radius of 1, whilst the subjets had radius 0.2.
- Only events with more than 2 subjets were considered.

Identifying Configurations of Subjets

- In order to consistently classify configurations of subjets, a clustering algorithm was implemented
- K-means clustering algorithm was chosen for this analysis
  - Separates data into K pre-defined clusters
  - Clusters do not overlap
  - Aims to maximize similarity between cluster points and the distance between clusters.
  - It is easy to implement

- Problem : pre-defining number of clusters
  $\Longrightarrow$ Silhouette Analysis applied
  - It determines optimal number of clusters $\Longrightarrow$ greatest separation between clusters
  - Silhouette score $\in [-1, 1]$ assigned to measure degree of separability
  - 1 $\Longrightarrow$ very good clustering
    -1 $\Longrightarrow$ very bad clustering
- Problem : Not possible to define 1 cluster or for clusters to have single data points
  $\Longrightarrow$ distance cuts applied

Probabilities of Configurations

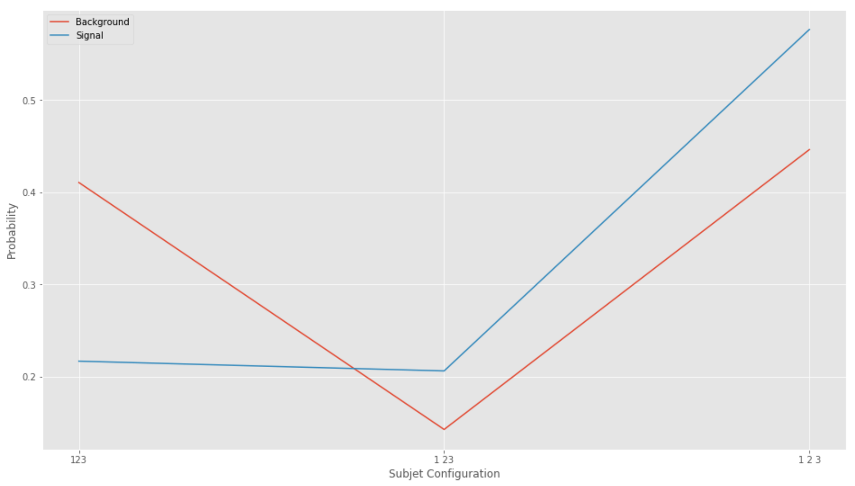


Figure: Plot comparing the probabilities of different spatial configurations for an event containing 3 subjets

Probabilities of Configurations

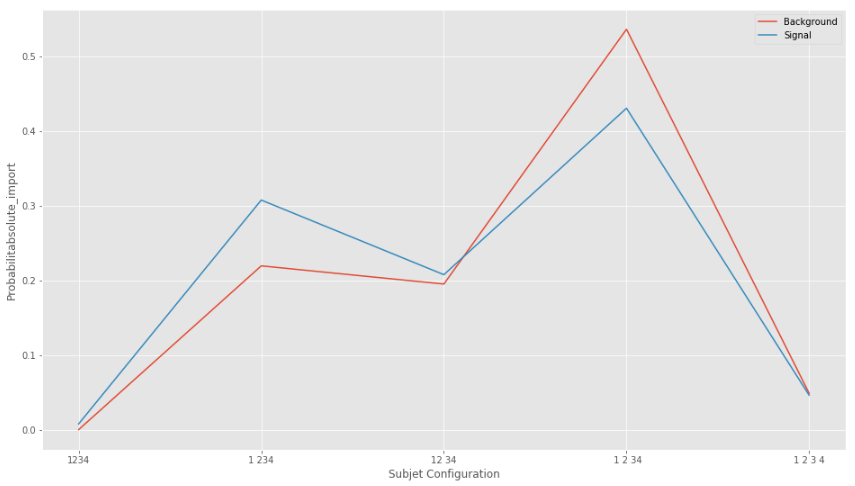


Figure: Plot comparing the probabilities of different spatial configurations for an event containing 4 subjets

Creating the Tagger

Combining Probabilities

- Two variables from events used to create final tagger:
  - Number of Subjets
  - Spatial configuration of Subjets
- Is the event more likely to be signal or background?

  $\Longrightarrow$ Implementation of Naive Bayes Classifier
  - It combines probabilities of the two variables from event data and previously determined probabilities from “training” data

How did it perform?

- Not well..
  - **Signal Efficiency** : $\epsilon_S = 47.1\%$
  - **Background Rejection** : $1 - \epsilon_B = 50.6\%$
- BUT creating a super efficient tagger was not the purpose of the project
- Analysis was too simple to obtain viable results
- Important qualitative results
  - **QCD Jets** : subjets tended to be closer together
    - 3 Subjets : '123'
  - **Top Jets** : subjets tended to be more distinct
    - 3 Subjets: ' 1 2 3 '

Shortcomings of the Tagger

- K-means algorithm works best clustering large amounts of data
  - This investigation dealt mainly with only 3 - 6 subjets
- K -means initially assigns clusters at random
  $\implies$ clustering is not unique
  - Number of events was small
    $\implies$ significantly different results obtained each run of the program
- The sample of events analysed had unrealistic proportions of Signal to Background events.
- The “training” and “testing” data had different proportions of Signal to Background events
  $\implies$ Naive Bayes classifier was compromised

Thank you for **Listening**!