Jet Substructure and Tagging with Tracks

Northwest Terascale Workshop: Using Jet Substructure

Maximilian Swiatlowski

1SLAC National Accelerator Laboratory

23 April, 2012
Why Tracks For Substructure?

• We have already seen from experiments and theorists the motivation for jet substructure variables—can dramatically improve signal-to-background for selecting jets from boosted objects, etc.

• Sophisticated pileup correction schemes—jet grooming, jet areas correction for $p_T$ and shapes, etc.—reduce the significant impact that pileup can have on these observables

• But there is another scheme to avoid the affects of pileup: use objects (i.e. tracks) that are inherently immune to pileup!
  • Excellent $z$ resolution of tracks allows efficient matching of tracks to vertices, and rejection of pileup tracks

• And sometimes—$b$-tagging, $q/g$-tagging, jet charge, occasionally substructure—tracks are actually optimal!
Not Just Substructure, Of Course...

- Of course, tracks used with jets for more than substructure in ATLAS
- JVF, for example, is one of our main pileup cleaning tools
  - Fraction of $p_T$ of tracks coming from hard scatter, compared to all PV
- Cuts here significantly improve data/MC agreement: useful for all analyses
- $b$-tagging is similar: combine tracks with jets to improve analyses
Outline

1. Introduction
   Overview

2. Quark-Gluon Tagging
   Monte-Carlo Templates
   Conclusions

3. $g \rightarrow \bar{b}b$ Tagging
   Overview
   Discriminating Variables

4. Tagger Performance

5. Conclusions
   Closing Thoughts

6. Quark/Gluon Templates

---

M. Swiatlowski (SLAC) Jet Substructure and Tagging with Tracks 23 April, 2012
Quark-Gluon Tagging
History and Motivation

- Quark-initiated and gluon initiated jets have long been known to have different properties
  - Well measured at JADE, LEP, others
  - Discrimination will never be as robust as $b$-tagging, but...
- Two papers from Schwartz and Gallicchio in 2011, along with previous efforts in ATLAS, led to a push for creating and commissioning a quark-gluon tagger at ATLAS
  - Theory paper (here) investigated the best variables to use to train a tagger, in parallel to our own efforts
- Many potential applications in searches for new physics and standard model measurements
  - Separate hadronically decaying bosons from gluon dominated backgrounds (diboson searches, Higgs, etc.), improve discrimination in dijet searches, monojet characterization, many more
A Monte-Carlo Based Tagger at ATLAS

- $S+G$ showed $n_{trk}$ and Width as most powerful values
  - Switch width to track width for pileup reasons (and nearly as performant...)
- Pythia MC shows good discrimination at moderate jet $p_T$!
MC Tagger Performance

- Construct a tagger using 2D combination of those variables: define $L = Q/(Q + G)$ and cut on various values of $L$
- Only about a factor of 2 worse performance in reco (at high $p_T$)
  - Promising initial results
First Signs of Trouble?

- But when we add different MC’s, the story becomes more complicated.
- Pythia alone is not a good place to look!
  - Neither Herwig or Pythia are perfect: use data driven templates
Template Methods

- Take percentages from MC, measure $\gamma + \text{jet}$ and dijet in data: solve for quark and gluon distributions in data
- More information on method in backup

![Diagram showing quark and gluon distributions](attachment:template_methods_diagram.png)
MC-labeled distributions in $\gamma$+jet and dijets agree very well with templates derived in MC

- Gives us confidence that the algorithm is doing something sensible!
Finally, Some Data (Comparing to Pythia)

- But data disagrees with Pythia in $n_{\text{trk}}$, and in a way that leads to worse performance!
- Track Width is a little better, though bad again at high $p_T$
Finally, Some Data (Comparing to Herwig)

- Better agreement with Herwig overall and in some regions, but still some significant disagreement!
Validating with Pure Data Samples

- Can also compare templates to results of G+S’s first paper: kinematic cuts to form 90-95% pure samples of q/g in $\gamma+2j$ and trijet events
Comparing Data to Data (Pure Samples)

- Generally speaking, see good agreement—better than in MC (limited by statistics, though)
  - Templates agree with independent q/g sample
Lessons from Quark/Gluon

- Analysis is finishing up systematics and tagger now, paper very much in progress. What have we learned?
  1. Exciting results from theorists should be closely tested, with multiple MC’s
  2. Important to test both signal and background—dijet and $\gamma+$jet agreement is much closer than for pure templates, for example
  3. Data-driven techniques for testing tagging and substructure techniques are difficult, but can help us understand data/MC differences
  4. And finally: tracking information is useful from a pileup perspective, but also on its own! $n_{\text{trk}}$, even this weak, is still the strongest variable!

- Conf note available: **ATLAS-CONF-2012-138**
$g \rightarrow \bar{b}b$ Tagging
Why Tag $g \to \bar{b}b$?

- Moderate and high $p_T$ gluons easily fragment to $\bar{b}b$: can produce heavy flavor from jet fragmentation instead of the hard scatter or particle decays!
  - Contaminates measurements of $b$ cross-section, introducing large theoretical uncertainties
- Sometimes the $b$ quarks will be separated widely and reconstructed separately: little hope of tagging these as originating from gluon splitting
- But often, gluon will have had enough boost to be reconstructed as single anti-$k_t$ $R = 0.4$ jet: a b-jet with two b-hadrons inside!
  - This jet will have different properties... can discriminate!
Variables in MC

- Performed search across many variables to find the best discrimination
- $n_{trk}$ and Track Width make an (unsurprising?) return
- A new variable: $\Delta R(k_t^1, k_t^2)$
  - Cluster tracks into 2 $k_t$ subjets: take $\Delta R$ between them
  - Related to n-Subjettiness, but turns out stronger
  - Note: another track variable, so automatically pileup resistant!
Variable Correlations

- Did not have an approved version of this plot for $q/g$, but qualitatively very similar
- Two distinct regions, with gluon jets having broader distribution and more tracks
  - Variables are clearly correlated, but information can be gained from using them together
Another Interesting Variable: $\tau_2$

- Another (suprisingly) powerful variable is $\tau_2$: n-Subjettiness variable
  - Suprising, because the individual variable, and not the ratios, were most powerful
  - And: merged has **higher** value than single
• Surprisingly, given our experience with q/g, all variables are very well modelled between data and MC!

• $b$-hadron fragmentation and showering is different from light quarks: perhaps expected that data/MC agreement should be different for this

• Strong level of agreement in all variables, across all $p_T$ bins, allows for training MVA in MC for use on data
The Tagger

- Train simplest likelihood output from TMVA
- Observe very strong $g \rightarrow \bar{b}b$ rejection! Third substructure related variable increases discrimination significantly compared to q/g
Conclusions and Open Points

- $g \to \bar{b}b$ tagger trained in Pythia MC performs strongly in discriminating merged jets against signal jets
  - Jet substructure **using tracks** plays a crucial role in the strong performance
    - With $b$-jets, might as well use tracks: you have a SV, so you know you have good tracks, etc.
- Opposite problem exists to q/g: good data/MC agreement means the tagger is easy to train, but it is very difficult to find a calibration sample to verify performance on
  - Whereas q/g was difficult to train, but with trijet + $\gamma+2$jet samples for calibration
- Conf note available: ATLAS-CONF-2012-100
Subject Energy Scales
What Else Can We Do With Tracks For Substructure?

- Often use track-jets, for example, to measure systematics on mass, other variables, using double ratio in data/MC for LCTopo jet vs track-jet
- This is useful for “full” jets— easy to identify matches between track-jet and calo jet
- But what about calibrating/measuring subjets? Often have wildly erratic shapes, and are very small— no guarantees of matches!
  - And even, the subset of jets with matches is probably biased compared to the inclusive sample
- How can we measure independently the energy of subjets?
A Simple Event Display...

- Dots from green to red are clusters (and their energy), purple-blue are jets (and their energy: blue is highest). Tracks would appear similarly to clusters.
- These are anti-$k_t$ $R = 1.0$ jets in a very crowded, busy environment: many jets are not cones!
- Subjets are going to have the same problem: any assumption on the shape of the subjet is not going to be valid in a busy (albeit smaller) environment
- Even worse for subjets: non-circular algorithms!
Associating Tracks to (Sub)Jets

- Traditionally, associate tracks to jets using $\Delta R$ cone, but often this shape assumption is not justified
  - Solution: use **ghost association** of tracks to jets
    - Place tracks on the grid: ask where they go? Practically: place a 0-energy copy of the track at the same $\eta/\phi$, and add to list of constituents for clustering, and see which jet it gets clustered to
    - Identical to method of calculating jet areas: pretend there is a tiny particle at some point, and see which jet it would join

- On the right, a jet with $k_t$ subjets: the active area, not cones, describes the region tracks (black points) should join
**ATLAS Simulation**

**MC11 PYTHIA (AUET2b), √s=7 TeV**

- Leading $k_t$ subjets ($R_{sub}=0.3$)
  - $100 \leq p_T^{jet} < 150$ GeV, $\Delta R$ association
  - $300 \leq p_T^{jet} < 400$ GeV, $\Delta R$ association

- Sub-leading $k_t$ subjets ($R_{sub}=0.3$)
  - $100 \leq p_T^{jet} < 150$ GeV, Ghost association
  - $300 \leq p_T^{jet} < 400$ GeV, Ghost association

**Figure 1:** Distributions of track multiplicity ($n_{trk}$) for subjets in the di-jet sample, for Monte Carlo simulation (PYTHIA AUET2b tune). Subjets are built with $k_t$ (figures 1(a) and 1(b)) or C/A (figures 1(c) and 1(d)) algorithm using $R_{sub}=0.3$ from large-$R$ anti-$k_t$ ($R=1.0$) jets calibrated at the Local Cluster Weighting scale (no jet energy scale applied). Tracks are associated to subjets using the ghost-association (tracks are given negligible momentum and cluster to the subjet using the jet finder algorithms) and geometrical-association (tracks are matched to the subjet if they are less than $R_{sub}$ apart from the subjet axis in the pseudorapidity-azimuthal plane, hence also named $R$-association). Tracks associated by the $R$-matching and by GA show similar overall performance. For the leading subjets (figures 1(a) and 1(c)) the mean value of these distributions is an increasing function of transverse momentum the parent jet. Many subleading subjets (figures 1(b) and 1(d)) have no tracks associated to them.

- Leading subjet on the left: ghost method (open) agrees with $\Delta R$ (closed)
- But for subleading, significant disagreement!
$r_{trk}$ For Subjets

- As subjets get closer to other subjets, $\Delta R$ on bottom starts getting confused: very unstable energy scale measurement

- Ghost association, on top, remains constant across distance: tracks associated this way correspond well to the subjet
**Validation of Subjet JES**

- Can use this measurement to compare data/MC agreement, and measure subjet JES uncertainty for, for example, HEP Top Tagger
- See generally good agreement

\[ r_{trk}^{\text{subjet}} \]

*ATLAS Preliminary*

C/A LCW jets with R=0.4

\[ \int L dt = 3.6 \, fb^{-1}, \sqrt{s} = 7 \, TeV \]

- Data 2011
- \( \ll \eta_{\text{jet}} < 0.8 \)

\[ p_T^{\text{jet}} \]

\[ \int L dt = 55 \, pb^{-1}, \sqrt{s} = 7 \, TeV \]

- Data 2011
- \( \ll \eta_{\text{jet}} < 0.8 \)

\[ <7 \mu \]

Average number of interactions: \( 4 < \mu < 7 \)

\[ <7 \mu \]

Average number of interactions: \( 4 < \mu < 7 \)
The same strategy— independent measurements of mass using tracks—is used to contrain the JMS uncertainty on ATLAS as well.
Conclusions for Subjets

- A nominally difficult problem—subjet energy scale measurement—seems to be possible, and robust, with a new application of the jet areas method for associating tracks to jets
  - Using an independent measure of the subjet energy, we can constrain the uncertainty on the calorimeter energy measurement
- Tracks can be accurately associated to the particular shapes of jets in dense event topologies—what else can we use this information for?
  - Subjet JVF?
  - Subjet moments? Width, etc.?
Conclusions
What Have We Learned?

- Tracks provide (at least?) three key capabilities to jet substructure and tagging measurements:
  1. Sometimes, tracking information is the most powerful tagging tool—\(b\)-tagging, \(n_{trk}\) in \(q/g\)
  2. Tracking can provide pileup-resistant versions of substructure variables—Track Width in \(q/g\), \(\Delta R\) between \(k_t\) subjets in \(g \rightarrow \bar{b}b\)
  3. Tracking also provides independent measurements of energy, useful for verifying energy scales and systematics—subjet JES with ghost-tracking, mass systematics using calo-track jet double ratios
Where Is the Future?

- In the future, how best do we combine information between tracking and calorimeter variables? Are they fully correlated? Or can we use variables from both?
- Will we be able to continue to rely on tracks as pileup resistant?
- Can we use other features of calibrations with tracks—Global Sequential Calibration—to improve properties of fat jets and substructure?
- Will tracking substructure variables be useful, or is the performance degradation too large?
- Other measurements using tracks, like Jet Charge, are incoming
  - Once validated and measured, what searches can use these tools?
Thank You For Your Attention!
Backup
Extracting Templates

• Goal: to better understand quark/gluon shapes in data, extrapolate data to 100% purity with fractions from MC

• Ideally, solve for \( q/g \) on bin-per-bin basis from:

\[
\begin{align*}
 h^{\gamma+j} &= P_Q^{\gamma+j} q + P_G^{\gamma+j} g \\
 h^{\text{dijet}} &= P_Q^{\text{dijet}} q + P_G^{\text{dijet}} g
\end{align*}
\]

\( P_Q \) = percentage quark  
\( h \) = histogram value  
\( q/g \) = templates  
\( (\gamma + \text{jet})/(\text{dijet}) \) = different sample

• But, need to account for \( b \) and \( c \) fractions (for now, taken from MC):

\[
\begin{align*}
 h^{\gamma+j} &= P_Q^{\gamma+j} q + P_G^{\gamma+j} g + P_B^{\gamma+j} b + P_C^{\gamma+j} c \\
 h^{\text{dijet}} &= P_Q^{\text{dijet}} q + P_G^{\text{dijet}} g + P_B^{\text{dijet}} b + P_C^{\text{dijet}} c
\end{align*}
\]

From Data  
From MC  
Solving for This

• Then, compare pure data shapes to pure MC shapes (used for training tagger)