Substructure and tagging in CMS

James Dolen
Purdue University Northwest

Large Hadron Collider Physics (LHCP) conference - Bologna Italy - June 6th, 2018
Jet Tagging at CMS

• Jets are ubiquitous at hadron colliders.
  - Collimated spray of hadrons resulting from an initial state quark or gluon

• LHC jet tagging revolution
  - Enabled by new reconstruction techniques (ex. particle flow, sequential recombination jet algorithms)
  - Discriminate quark and gluon jets, noise and pileup jets
  - Boosted heavy object jet tagging (ex. top, W, Z, Higgs)
    ‣ all decay products of reconstructed within one jet
  - Advanced b-tagging methods
    ‣ Subjet b-tagging, double b-tagging

• Critical tools for discovery!
• Critical tools for measurements!
Jet tagging

- $u,d$ or $s$ jet
- Gluon jet
- $c$ or $b$ jet
- Pileup jet

- $W$ or $Z$ jet
- Higgs jet
- Top jet

Diagram by Nhan Tran
Identifying quark and gluon jets

- Gluon more likely to radiate a gluon ($C_A = 3$ vs $C_F = 4/3$)
- Gluon fragmentation function softer
- Gluon jets
  - Wider than quark jets ($\eta$-Φ plane)
  - Larger multiplicities
  - Fewer hard particles
- Quark jets
  - Narrow
  - Smaller multiplicities
  - Asymmetrical energy shared between constituents
Quark/gluon discrimination at CMS

- Three tools:
  - Likelihood discriminator (3 variable)
  - BDT discriminator (5 variable)
  - Deep neural net (jet constituents)

- Quark/gluon discriminator variables:
  - Jet particle multiplicity
    - Total or charged
  - Jet shape
    - major axis width
    - minor axis width
    - \( \Delta R \) weighted \( p_T \)
  - Energy sharing (\( p_T D \))

\[
p_T D = \sqrt{\frac{\sum_i p_{T,i}^2}{\sum_i p_{T,i}}} \rightarrow 1 \text{ if all momentum carried by one particle} \\
\rightarrow 0 \text{ if jet has infinite number of particles}
\]
Pileup jet tagging

- Pileup jets are formed from overlapping low p_T pileup particles
  - Each pileup vertex contributes ~0.7 GeV of energy per unit area (η,Φ) of the detector

- Pileup jet tagging - tool used to identify and reject jets originating from pileup
  - Utilize vertex, track, and jet shape information
  - Charged particles inside pileup jets are not associated with the primary vertex
  - Pileup jets are more defuse (overlapping soft particles from multiple vertices)
**Boosted heavy object tagging**

- **Example: 2 body decay**
  - decay product angular separation $\Delta R \sim 2M/p_T$

- **Boosted heavy object jets are identifiable based on their mass and via jet substructure**
  - The jet, its constituents, and its clustering history, contain useful information which can be used to identify these objects
  - examples: jet mass, “subjettiness”, $W$ mass within a top jet
Jet mass as a tagging variable

- Merged top/W/Z/H jet → LO jet mass is the heavy object mass

- For background jets (quark/gluon) the LO jet mass is ~ 0, but perturbative effects lead to measured mass

\[ \langle M^2 \rangle \simeq C' \cdot \frac{\alpha_s}{\pi} p_t R^2 \]

Jet grooming

Algorithmic jet substructure techniques designed to remove isolated soft radiation in jets (contamination from ISR, UE, pileup)

**Trimming, Filtering** - Recluster jet with smaller distance parameter. Condition based subjet removal.

**Pruning** - Recluster jet. Remove soft large angle particles.

**BDRS, MMDT, Soft Drop, JHU top tagger, CMSTT** - Recursively decluster jet. Remove sub-clusters not satisfying algorithm condition. Stop declustering when both subjets satisfy condition.

Reduces jet mass dependence on pileup. Reduces measured QCD jet mass (improves discrimination). Improved jet mass resolution for boosted heavy object.
Jet shape tagging variables

Particle energy pattern within a jet used to identify “multi-prong” jets

**N-subjettiness**

\[ \tau_N = \frac{\sum_{i=1}^{n_{\text{constituents}}} p_{T,i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, \ldots, \Delta R_{N,i}\}}{\sum_{i=1}^{n_{\text{constituents}}} p_{T,i} R} \]

Determines how consistent a jet is with having N or fewer subjets

Better discrimination by using ratios (ex. \( \tau_3/\tau_2 \))

J. Thaler, K. Van Tilburg, JHEP 2011:15

**Energy Correlation Functions (ECF)**

\[ \text{ECF}(N, \beta) = \sum_{i_1 < i_2 < \ldots < i_N \in J} \left( \prod_{a=1}^{N} p_{T,i_a} \right) \left( \prod_{b=1}^{N-1} \prod_{c=b+1}^{N} R_{i_b i_c} \right)^\beta \]

Jet constituent based observables sensitive to N subjet substructure

ECF ratios (C2,D2, etc.) for boosted object discrimination (Small C2 \( \Rightarrow \) 2 subjets)

\[ C_N^{(\beta)} = \frac{\text{ECF}(N+1, \beta) \text{ECF}(N-1, \beta)}{\text{ECF}(N, \beta)^2} \quad D_2(\beta) = \frac{\text{ECF}(3, \beta) \text{ECF}(1, \beta)^3}{\text{ECF}(2, \beta)^3} \]

\( \beta \) parameter allows access to different angular scales

M. Jankowiak and A. J. Larkoski JHEP 1106 (2011) 057
A. Larkoski, G. Salam, J. Thaler JHEP 2013:108
2-prong tagging “V-tagging”

- Boosted W, Z, H tagging
- General technique: select V jets based on groomed jet mass and jet shape variable
- CMS default:
  - Anti-kt R=0.8 PUPPI jet
  - MMDT/Soft drop jet mass
  - N-subjettiness ratio \( \frac{\tau_2}{\tau_1} \)

[Diagrams of boosted W/Z jet and boosted H → bb]

James Dolen (Purdue University Northwest)
Top jet tagging

Jet mass = top mass

Substructure (3 subjets)

One subjet b-tagged

Two subjets with pairwise mass = W mass
Top tagging algorithms

- Taggers utilizing general algorithms
  - Groomed mass + N-subjettiness ($\tau_3/\tau_2$)
  - Shower deconstruction [1]
  - Splitting scale
  - Subjet b-tagging
- Dedicated algorithms
  - CMS Top Tagger (JHU Top Tagger) [2]
    - Decluster jet twice to find 1-4 subjets
    - Select tops based on top mass, W mass, Nsubjets
  - HEP Top Tagger v1 [3]
    - Very large jets (R=1.2-1.5)
    - Multistep decluster + filter procedure
    - Select tops based on top mass, W mass
  - HEP Top Tagger v2 [4]
    - Multiple algorithm improvements + multi R approach
    - Select tops based on top mass, W mass, optimal jet size

Top tagging performance

CMS default:  
- Anti-kt R=0.8 PUPPI jet  
- MMDT/Soft drop jet mass  
- N-subjettiness ratio $\tau_3/\tau_2$
Multi-category taggers

- BEST (Boosted Event Shape Tagger)
- Deep AK8
Top and W jet validation in data

- Semileptonic ttbar selection $\rightarrow$ very pure sample of boosted Ws

- Data-MC scale factors measured
Heavy flavour tagging at CMS

- Identify jets originating from a b or c quark
- Tagging algorithms rely on track and secondary vertex information
Heavy flavour taggers

- Optimized algorithms for Run II
- CSVv2
  - 19 variable shallow neural network (1 hidden layer)
- cMVAv2
  - BDT combines CSVv2 with taggers based on soft leptons and track probability information
- DeepCSV
  - Similar to CSVv2 but with a deep neural network (4 hidden layers with 100 notes each) and with more tracks considered
- Deep Flavour (*in development*)
  - Deep neural network using particle flow particles and secondary vertices as input
b-tagging in boosted topologies

• **Boosted top**
  - **Fat jet method** - require the large AK8 jet be b-tagged
  - **Subjet method** - Use grooming algorithms to identify subjets and require 1 b-tagged subjet

• **Boosted Higgs → bb, boosted Z → bb, etc.**
  - **Subjet method** - identify two subjets which can be b-tagged
  - **Dedicated double b-tagger** - trained specifically for this topology

James Dolen (Purdue University Northwest)
Double $b$-tagged H-jet

CMS DP 2017/032

CMS Experiment at LHC, CERN
Data recorded: Wed Dec 31 19:00:00 1969 EDT
Run/Event: 1 / 363
Lumi section: 2
Orbit/Crossing: -1 / -1
**b-tagging in boosted topologies**

**CMS PAS BTV-16-002**  
**ARXIV 1712.07158**

### $t \rightarrow bW$

#### CMS Simulation

**Misid. probability (multijet)**

- **AK8 jet**

<table>
<thead>
<tr>
<th>$300 &lt; p_T &lt; 500$ GeV, $135 &lt; m &lt; 200$ GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^{-2}</td>
</tr>
</tbody>
</table>

**Tagging efficiency ($t \rightarrow b$)**

- **Subjet method = best performance**

#### $H \rightarrow bb$

**CMS Simulation**

**Misid. probability (multijet)**

- **AK8 jet**

<table>
<thead>
<tr>
<th>$1200 &lt; p_T &lt; 1800$ GeV, $135 &lt; m &lt; 200$ GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^{-3}</td>
</tr>
</tbody>
</table>

**Tagging efficiency ($H \rightarrow b \bar{b}$)**

- **Subjet method = best performance**

James Dolen (Purdue University Northwest)

21
Conclusions

• New jet tagging techniques have proven to be indispensable to LHC analyses
  - Impressive progress in the last 7 years developing and commissioning jet tools
  - Boosted objects are now mainstream
    ‣ Thoroughly calibrated and commonly used
    ‣ example: groomed jet mass is now in the CMS trigger
    ‣ Strong community of theorists and experimenters
      ‣ Short turnaround time between new ideas and results
  - Modern techniques have led to advances in heavy flavour tagging
**b-tagging validation in data**

- **b-jet rich events**
  - Events with muons from b-decays
  - tt events
- **QCD, DY enriched + c-jet enriched topologies**
- **Double b-tagger**
  - $g \rightarrow bb$ jet with two tagged muons

---

**Graphs and Figures**

1. **CMS Muon-enriched multijet sample**
   - Double-muon-tagged AK8 jets
   - $p_T$ (AK8 jets) $> 250$ GeV
   - Data vs. MC comparison

2. **CMS Single-lepton t$\bar{t}$ sample**
   - $p_T$ (AK8 jets) $> 250$ GeV
   - Data vs. MC comparison
**Pileup removal before jet clustering**

- Charged Hadron Subtraction (CHS)
  - enabled by particle flow
  - remove charged particles originating from pileup vertices before clustering jets
  - does not remove neutral pileup

- New tools which utilize additional information for pileup removal
  - **Constituent subtraction** [1] - per particle area subtraction
  - **Jet cleansing** [2] - charged particle vertex information used to correct jets at the subjet level
  - **Soft killer** [3] - progressively remove soft particles until the average pileup density in the event is 0
  - **Pileup Per Particle Identification (PUPPI)** [4] - jet shape and charged particle vertex information used to suppress pileup

---

PileUp Per Particle Identification (PUPPI)

• Pileup handles:
  - Tracking/vertexing → we know which charged particles come from pileup and which ones come from the hard scatter
  - Pileup is randomly distributed, while collinear radiation from a particle from the hard scatter is preferentially radiated at small angles
  - $p_T$ spectrum of pileup falls quickly

• PUPPI assigns each particle a weight based on the likelihood it originated from pileup
  - Handle: Calculate $\alpha = \sum p_T^j / \Delta R_{ij}$ for each particle $j$ within an annulus around particle $i$
    - $\alpha$ small for particles from pileup (nearby particles are soft and at large angle)
    - $\alpha$ large for particles from the LV (radiation around the particle harder and at small angle)
PUPPI performance

- PUPPI jet mass stable with pileup
- PUPPI corrects jet shapes for pileup → improves tagging discrimination
- Improved MET resolution
- Improved lepton isolation
- With pileup the top jet N-subjettiness value is shifted to be more QCD-like
Boosted jet menu

- **Boosted W/Z jet**
  - $W$ jet
- **Boosted $H \rightarrow bb$**
  - $H$ jet
- **Boosted $H \rightarrow \tau \tau$**
  - $H$ jet
- **Unmerged hadronic top**
  - $t \rightarrow W q q b$
- **Partially merged hadronic top (W jet + b jet)**
  - $t \rightarrow W q q b$
- **Fully merged hadronic top jet**
  - $t \rightarrow W q q b$
- **Leptonic top with non-isolated lepton**
  - $t \rightarrow W q q b$
- **RPV gluino $\rightarrow 3$ quarks**
  - $\tilde{g}$
- **SUSY Cascade**
- **SUSY Accidental Substructure**

James Dolen (Purdue University Northwest)
Pileup jets

- Many overlapping soft particles from PU can add jet of substantial pT
- Identifying pileup jets relies on charged PU vertex information and jet shape

\[
\beta = \frac{\sum_{i \in \text{PV}} p_{Ti}}{\sum_i p_{Ti}} \quad \beta^* = \frac{\sum_{i \in \text{other PV}} p_{Ti}}{\sum_i p_{Ti}}
\]
Soft drop

- Recursively decluster

\[
\text{Soft Drop Condition: } \frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{\text{cut}} \left( \frac{\Delta R_{12}}{R_0} \right)^\beta
\]

- Soft wide angle radiation fails the condition
  - As \( z_{\text{cut}} \uparrow \Rightarrow \) more aggressive grooming
  - As \( \beta \downarrow \Rightarrow \) more aggressive grooming

- Example (\( z_{\text{cut}} = 0.1 \)):
  - If \( \beta = 0 \), remove softer subjet if \( p_T \) fraction < 0.1
    (~equivalent to MMDT)
  - If \( \beta > 0 \), remove softer subjet if \( p_T \) fraction < \( x \), where \( x \) increases with \( \Delta R \) and has maximum value 0.1
  - \( \beta \rightarrow \infty \) no grooming
  - \( \beta < 0 \) soft drop becomes a tagger instead of a groomer (finds jets with hard, large angle subjets)