CMS B-Tag Algorithms

Aaron Dominguez
SUSY 2009, Northeastern University

Partially supported by NSF grant PHY-0547780

Saturday, June 6, 2009
B-Tagging: Basic Idea

- Some interesting physics involves b production & decays of heavier particles into b quarks: top, Z, Higgs, SUSY. Identifying b’s helps you identify the other interesting particles that decayed into them.

- B hadrons have lifetimes and decay lengths distinct from other species

- Decay length is measurable by finding a “secondary” vertex and measuring distance to the “primary” vertex

- Can also find B decays from semileptonic decays into electrons, muons (or taus)

- B jets also have slightly different kinematics and charge/multiplicity
B-Tagging at CMS

• CMS is building on many years of experience from previous experiments’ success at b-tagging: LEP, SLD, CDF, D0 etc

• We try to take advantage of all the differences listed on the previous slide to separate b jets from lighter flavored jets

• Our b-tagging algorithms range from simple and robust which are suitable for early data taking and online triggering, to sophisticated multivariate techniques that require a very good understanding of the detector
Look at all Si tracker

One Quadrant of Tracker in r-z

Blue=Double-Sided  Red= Single-Sided

- 3 Barrel Pixel Layers, 2 Forward Pixel Disks
- 4 Inner Barrel Layers (TIB), 6 Outer Layers (TOB)
- 3 Forward Inner Disks (TID), 9 Outer Disks (TEC)

Saturday, June 6, 2009
Pixel Tracker
- 3 space points to $|\eta|=2.5$
- \(~10\mu m\) resolution in $r\phi$
- \(~20\mu m\) resolution in $rz$
  (depends on $\eta$, cluster size)
- >95% efficiency for seeds
- Use in HLT standalone

3 barrel pixel layers
- 4.4 cm
- 7.3 cm
- 10.2 cm

4 forward/backward disks
- ±35.5 cm
- ±46.5 cm

Saturday, June 6, 2009
Alignment of Tracker

- As you can see, our all silicon tracker consists of many different detector elements (thousands), both strip & pixels
- Lifetime information is the most discriminating piece of information to ID b-jets
- The tracker must be very well aligned before this information is sensible
- We have surveyed the tracker during construction & installation
- We will align it finally with tracks from collisions. The more tracks we have, the better the alignment.
- We are also recently finalizing an alignment using cosmic rays we’ve taken over the past several months. Results not yet public, but could be quite helpful and provide a very good start for b-tagging before we’ve collected lots of collision data
The track’s impact parameter is one of the standard helix parameters for charged tracks. Directly determined in tracking. It’s the perpendicular distance of the track to a reference point: primary vertex. Tracks are selected after a set of basic quality cuts. Impact par is signed based on angle w.r.t. jet axis. 2D or 3D calculation. Tracks from b decays have larger impact pars than those from prompt decays. Basis for robust, simple taggers.
Track Impact Parameter

QCD Monte Carlo (80-120 GeV jets) Perfect Alignment

CMS Preliminary

Saturday, June 6, 2009
Secondary Vertex

- An adaptive fitter takes clusters of tracks and down-weights incompatible tracks until a vertex converges with no more rejection of tracks.

- The invariant mass of all the tracks in a vertex is used to reject $K_s$ and junk.

- Calculate 2D or 3D flight distance with same signing convention as impact parameter: distance from primary vertex, $L$.

- Secondary vertex significance is used as a discriminator: $L/\sigma_L$.

- Tracks from b decays have larger secondary vertex significance than those from prompt decays.

- Basis for robust, but slightly more sophisticated taggers.
Secondary Vertex

ttbar Monte Carlo with different tracker alignment scenarios
“Soft Leptons”

- Inclusive semileptonic branching ratio of b hadrons is about 11% for each lepton
- Leptons give an orthogonal piece of information from the lifetime variables
- Identify b-jets with a muon or electron inside a jet with other signatures indicating heavy flavor
  - 3D signed impact parameter
  - Lepton momentum w.r.t. jet axis (a few GeV)
  - Opening angle of lepton w.r.t. jet
- Different lepton ID algorithms, muons are somewhat easier than electrons. (Also can use leptonic decays of taus)
Soft Muons

ttbar and QCD Monte Carlo with perfect tracker alignment
Simplest Tagger: Track Counting

ttbar, QCD Monte Carlo with different tracker alignment scenarios

- Sort tracks by decreasing 3D IP significance
- Use significance of 2nd or 3rd track as discriminator: HighEff, HighPur

Saturday, June 6, 2009
Simple Tagger: Jet Probability

- Calculate $CL$ that tracks in jet are consistent with primary vertex
- Small $CL$'s indicate the presence of lifetime tracks in the jet
• Find secondary vertex with at least 3 tracks, or 2 for pseudo-vtx
• Combine vertex distance, mass & impact parameters into a likelihood
• Combine muon’s momentum w.r.t. jet, 3D IP significance, opening angle with jet and other variables into a Bayesian neural network
Measuring Eff. from Data

- Tevatron developed techniques to measure tagging efficiencies from data instead of solely from MC.
- Muon-in-jet dijet events are used, compare events with a loose lifetime (TrackCount) tag on the muon jet to events without tag.
- This is done with lifetime taggers (like Track Counting & Jet Probability) and requiring $P_{T\text{rel}}>0.8$ GeV for the muon in the jet.
- This leads to 8 simultaneous equations, which can be solved for tagging efficiencies as a function of eta and $P_T$.
- The method (“System8”) assumes that the efficiency to tag a jet with lifetime and $P_{T\text{rel}}$ can be determined as a product of two efficiencies.

Uncertainty on efficiency expected to be:
- 15% for 10 pb$^{-1}$
- 10% for 100 pb$^{-1}$
- 5% for 1 fb$^{-1}$
Summary

• Our tracker is working well and we await collision data later this year.

• We hope to use an alignment from cosmic ray data as a starting point on day 1.

• We have several b-tag algorithms implemented and ready to go. Data driven methods will be used to measure fake rate and efficiency ("$P_T^{rel}$," "System8").

• We plan on using the simpler algorithms in the first data and then as we understand the detector better and better, take advantage of more sophisticated analyses.
Backup Slides
MVA framework

- Multi Variate Analysis in CMS reconstruction software
  - modular approach
    - combine simple modules to build more complex algorithms
  - provide common modules
    - linear discriminants
    - Likelihood ratios
  - abstraction layer for different
    - Neural Networks
    - Boosted Decision Trees
  - common interface over different back-ends
    - ROOT's MultiLayer Perceptron
    - Toolkit for Multi Variate Analysis (TMVA)
  - Seamless integration in the CMS reconstruction software
    - calibration and learning
    - analysis, reconstruction and identification
    - store / retrieve calibrations to / from the experiment conditions database

26/03/2009    A.Bocci - An overview of the b-tagging algorithms in the CMS offline software
Concrete Algorithms

- **Impact parameter:**
  - TrackCountingHighEff, TrackCountingHighPur
  - JetProbability, JetBProbability
  - ImpactParameterMVA

- **Secondary Vertex:**
  - SimpleSecondaryVertex
  - CombinedSecondaryVertex, CombinedSecondaryVertexMVA

- **Soft Lepton**
  - (simple) SoftMuonByPt, SoftMuonByIP3d
  - SoftMuon, SoftMuonNoIP
  - SoftElectron
\( p_{\text{TrEl}} \) Method

- Consider a dataset
  - two reconstructed jets where one jet has a muon
- Subset, where the muon jet is tagged
- Use \( p_{\text{TrEl}} \) template fit to find the amount of
  - b jets with muons \( (n_b) \) - from the main set
  - Tagged b jets with muons \( (n_b^{\text{tag}}) \) - from the subset
- b-tagging efficiency will be
  \[
  \epsilon_b = \frac{n_b^{\text{tag}}}{n_b}
  \]
- This is measured as a function of jet pseudorapidity and transverse momentum
- The main systematical uncertainty comes from the dependence on the Monte Carlo for the \( p_{\text{TrEl}} \) templates

**CMS PAS BTV-07-001**
Mistag rate measurement using negative tags

\[ \epsilon_{\text{mistag}} = \epsilon_{\text{data}} \cdot R_{\text{light}} \]

\[ \epsilon_{\text{data}}^{-} \] is the negative tag rate in multi-jet data

\[ R_{\text{light}} = \epsilon_{\text{mistag}}^{\text{MC}} / \epsilon_{\text{MC}}^{-} \]

- \( R_{\text{light}} \) is the ratio between the mistag efficiency of \( udsg \)-jets and the negative tag rate for all jets in simulation

- The evaluation of the mistag efficiency is sensitive to the fractions of \( c \) and \( b \) quarks in the negative tag jet sample. The \( c \) and \( b \) fractions can be significantly reduced by applying a positive tag veto: the negative tag jet is rejected if it has any track with \( IP/\sigma_{IP} > 4 \)
“System 8”

\[
\begin{align*}
    n &= n_b + n_{cl} \\
    p &= p_b + p_{cl} \\
    n_{tag} &= \varepsilon_{b}^{tag} n_b + \varepsilon_{cl}^{tag} n_{cl} \\
    p_{tag} &= \beta \varepsilon_{b}^{tag} p_b + \alpha \varepsilon_{cl}^{tag} p_{cl} \\
    n_{mu} &= \varepsilon_{b}^{mu} n_b + \varepsilon_{cl}^{mu} n_{cl} \\
    p_{mu} &= \varepsilon_{b}^{mu} p_b + \varepsilon_{cl}^{mu} p_{cl} \\
    n_{tag,mu} &= \kappa_b \varepsilon_{b}^{tag} \varepsilon_{b}^{mu} n_b + \kappa_{cl} \varepsilon_{cl}^{tag} \varepsilon_{cl}^{mu} n_{cl} \\
    p_{tag,mu} &= \kappa_b \beta \varepsilon_{b}^{tag} \varepsilon_{b}^{mu} p_b + \kappa_{cl} \alpha \varepsilon_{cl}^{tag} \varepsilon_{cl}^{mu} p_{cl}.
\end{align*}
\]

\( n \) = number of muon-jets
muon-jet+away-jet sample;

\( p \) = number of muon-jets
muon-jet+tagged-away-jet

\( n_{tag} p_{tag} \) = tagged with a lifetime tag

“mu” = requirement of a muon \( p_{T_{rel}} > 0.8 \text{ GeV} \)
Figure 2: Distribution of the transverse impact parameter measurement error of the second track (ordered by I.P. significance) for the various scenarios and jet flavors, presented for light flavor jets (top left), charm jets (top right) and b-jets (bottom).