An overview on the b-tagging landscape

algorithms and efficiencies

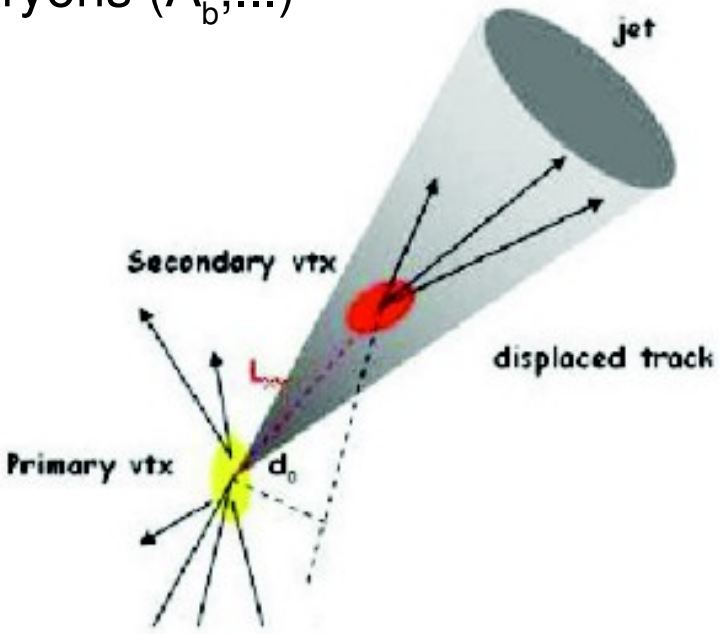
Joris Maes

IIHE top quark meeting

24th November 2008

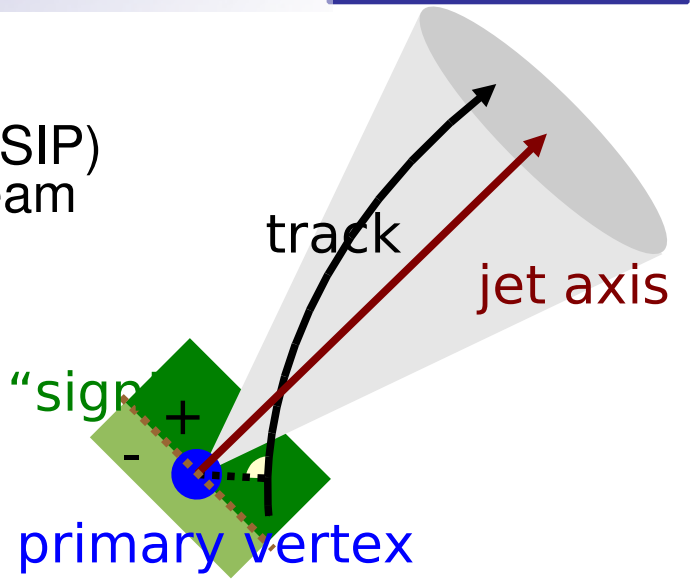
Introduction

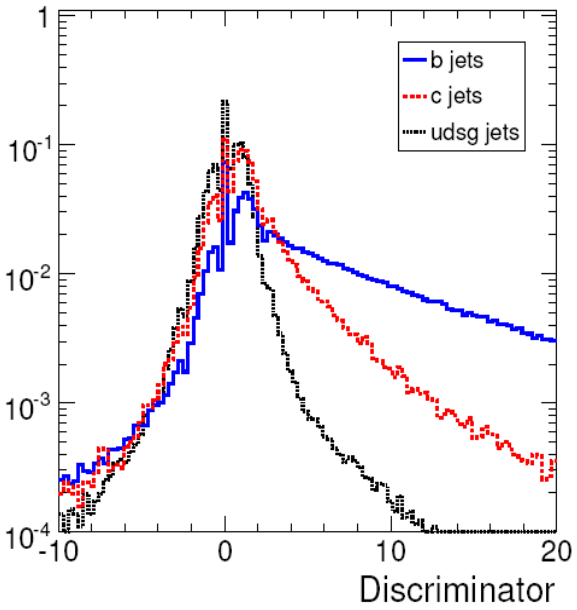
- In different interesting processes at the LHC b quarks play an important role
  - H → bbbar,
  - t → Wb for about 100%
- Identifying jets coming from b quarks (so called b jets) can help to reduce background in complex final state topologies.
- The difference of a jet from a b/c quark w.r.t uds- or gluon-jets can be used to identify them:
  - During fragmentation of heavy flavour quarks (b/c) jets are formed containing heavy mesons (B,D,...) and baryons (Λ_b,...)
  - the b hadron has a long lifetime τ of 1.6 ps corresponding to $c\tau$ 490 μm
  - this long lifetime is reflected in the presence of tracks not compatible with the primary vertex (displaced secondary vertex)
  - b jets contain (19%) a non isolated $e^{\mp}/\mu^{\mp}$ which has different properties w.r.t. other leptons formed inside jets



Impact parameter taggers

- Impact parameter IP = smallest distance of a track and the considered primary vertex
- One can assign a sign to the impact parameter (SIP) whether the track was produced up- or downstream
- To take into account the experimental resolutions the significance of the impact parameter is used = IP/σ(IP)
- This algorithm needs tracks and primary vertex with some additional quality cuts on the objects

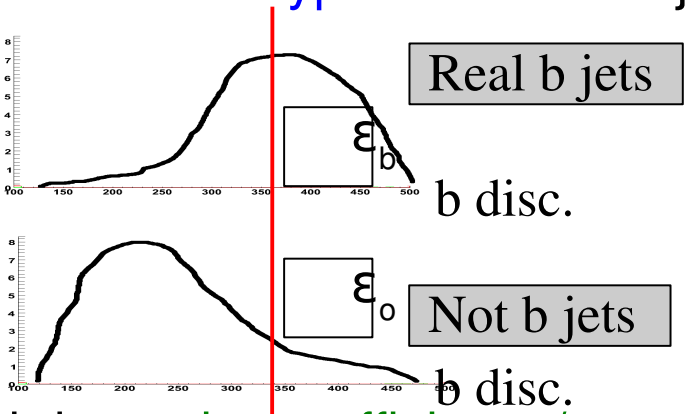


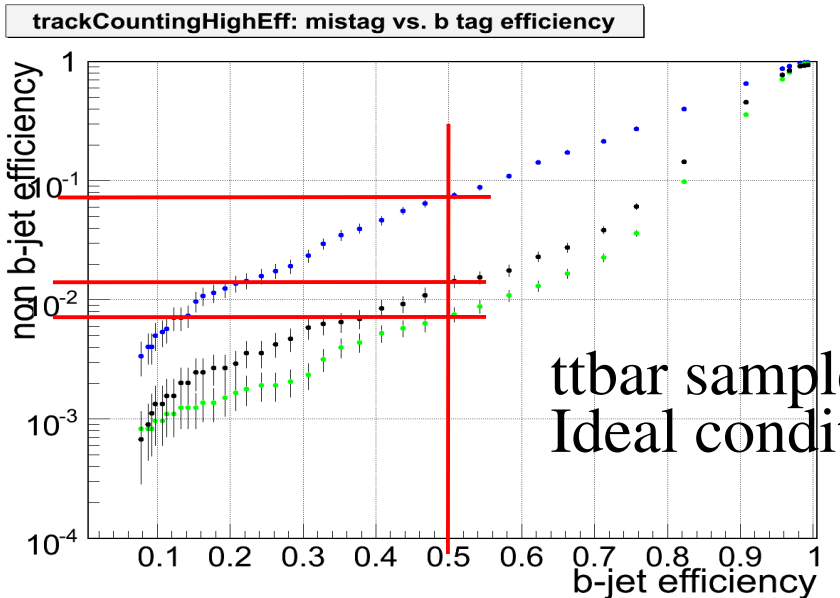


- Algorithms using this info
  - `TrackCountingHighEff`: SIP^{3d} of 2nd track (HLT)
  - `TrackCountingHighPur`: SIP^{3d} of 3rd track
  - `JetProbability`: Likelihood that each track comes from a b or light quark, combined for each track in the jet
  - `JetBProbability`: Only use first 4 tracks
  - `ImpactParameterMVA`: Combine all track Ips via MVA

Impact parameter taggers - performance

- The algorithms give a b discriminant variable to test the hypothesis that a jet is originating of a b quark.
- Two mistakes can be made:
  - accept the hypothesis when it is wrong
  - reject it while it was correct.
- This is reflected in two efficiencies
- ε_b: efficiency to tag a b jet as a b jet
- ε_o: efficiency to tag an 'other' (udscg) jet as a b jet =mistag efficiency/rate
- Interpretation: typically 50% b tag efficiency gives about a 10% c-mistag rate and a 1% udsg mistag rate

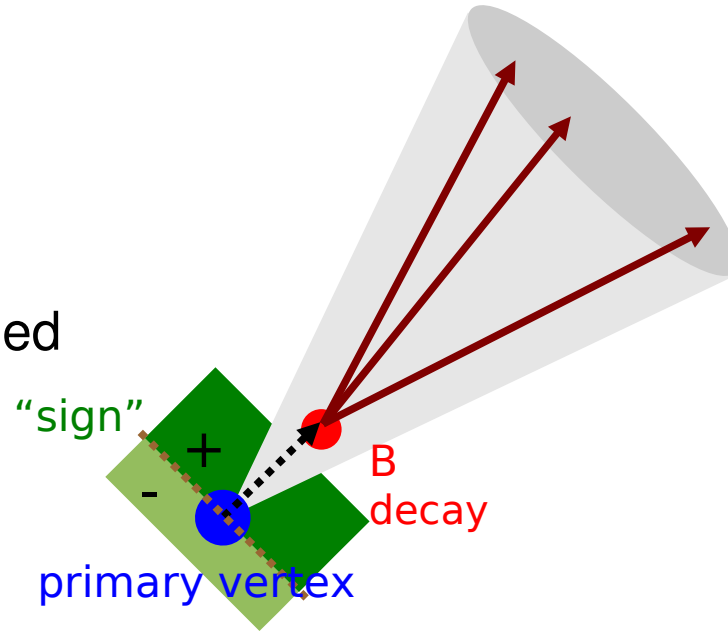




Secondary vertex taggers

- Besides the primary vertex the secondary vertex from the B decay is reconstruced. This one is used as extra input w.r.t impact parameter taggers

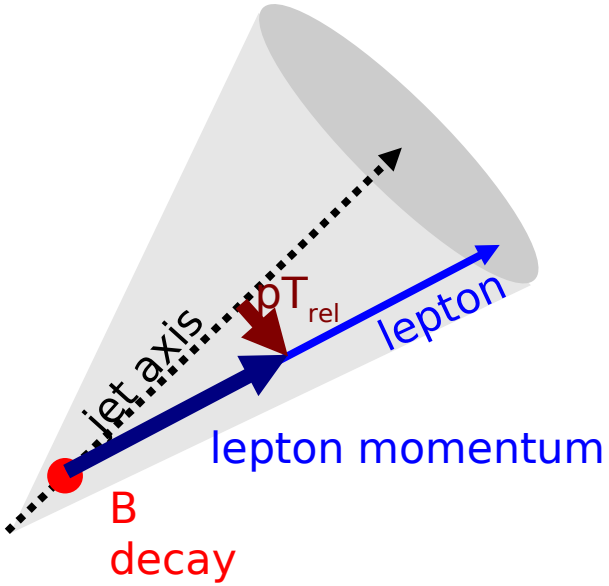
- Advantage: in general a lower mistag rate but more input is needed and can only be applied with enough detector knowledge



- algorithms
  - `SimpleSecondaryVertex`: Discriminate using the (signed) significance 3D flight distance
  - `CombinedSecondaryVertex`: likelihood to come from a b, c or light quark. Ratio of combined likelihoods as b-tag discriminator
  - `CombinedSecondaryVertexMVA`: neural network instead of Likelihood

Soft lepton taggers

- The branching fraction of direct and cascade decay of b hadrons in electrons and muons is about 19% due to $b \rightarrow W^{-*}X$ or $b \rightarrow W^{-*}c \rightarrow W^{-*}W^{+*}X$ where $W^{-*} \rightarrow l\nu$
- This is exploited by looking for a non isolated lepton in the jet



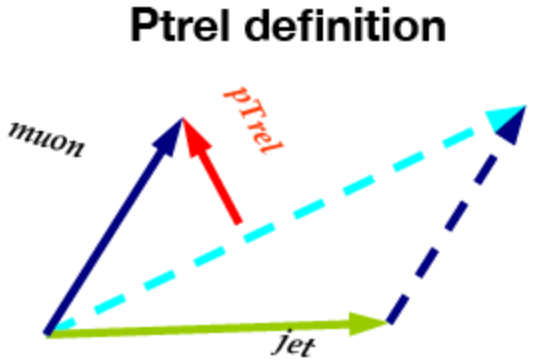
- These algorithms are less correlated w.r.t track impact and secondary vertex algorithms
- Input used are muons or electrons, primary vertex and tracks to improve jet direction
- Algorithms:
  - (simple) `SoftMuonByPt`: requires a muon with pTrel > cut (HLT)
  - `SoftMuonByIP3d`: requires a muon with SIP3d > cut
  - `SoftMuon`, `SoftElectron`: Combine lepton kinematic informations via a NN
  - `SoftMuonNoIP`: NN without SIP^{3d}

b tag performance measurement

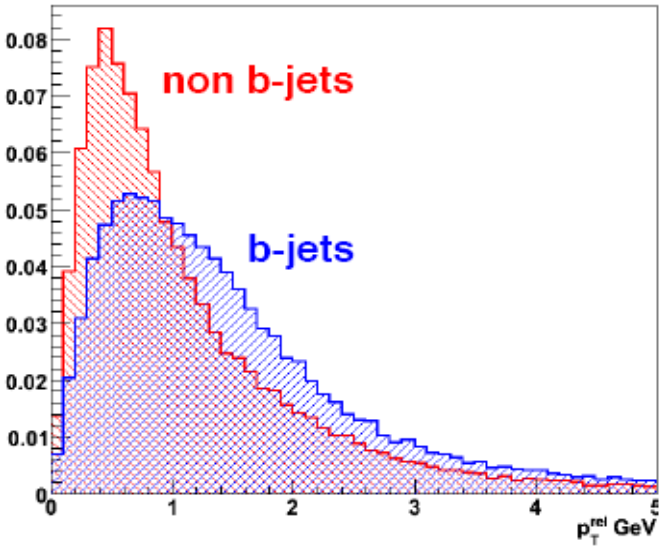
- In CMS different methods are used to measure the b-tagging efficiency from data:

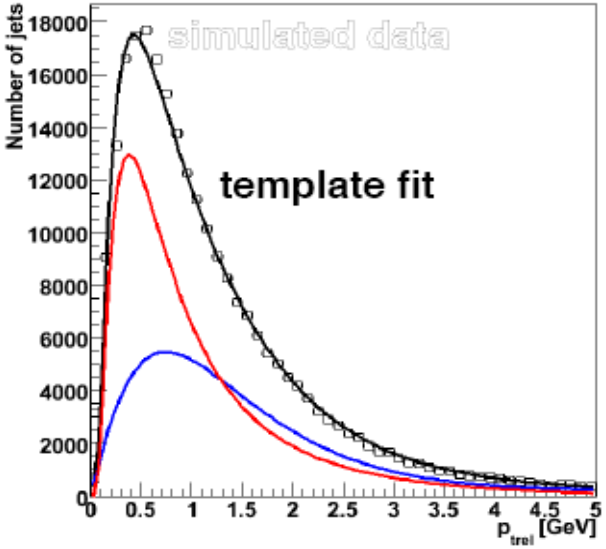
Ptrel (CMS Note 2007/046)

- Based on events with at least 2 jets and a non-isolated muon ($\Delta R(j,\mu)<0.4$)
- Method relies directly on fit of pTrel distr.
- Ptrel distr. fitted with linear combination of b and c+light jet templates
- Process repeated after tagging
- Btag efficiency calculated as the ratio between b jets after and before tagging



- An extention uses additional info from data: counting method





b tag performance measurement

System 8 (CMS Note 2008/081)

- Based on events with at least 2 jets and a non-isolated muon ($\Delta R(j,\mu)<0.4$)
- Two different data samples
  - Muon-in-jet + away-jet sample (n)
  - Muon-in-jet + tagged-away-jet sample (p)
- Two independent taggers
  - Life Time Tag (Track Counting/ Track Prob.)
  - Soft Muon Tag (pTrel > 0.8 GeV)

- A system of 8 equations with 8 unknowns.
- Depends minimally on MC: MC is only used to evaluate correlation factors between taggers.

$$
\begin{cases}
\text{muon-in-jet+away-jet} & n = n_b + n_{cl} \\
\text{muon-in-jet+tagged-away-jet} & p = p_b + p_{cl}
\end{cases}
$$

$$
\begin{cases}
\text{"probe" tagger} & n^{tag} = \varepsilon_b^{tag} n_b + \varepsilon_{cl}^{tag} n_{cl} \\
 & p^{tag} = \beta\, \varepsilon_b^{tag} p_b + \alpha\, \varepsilon_{cl}^{tag} p_{cl}
\end{cases}
$$

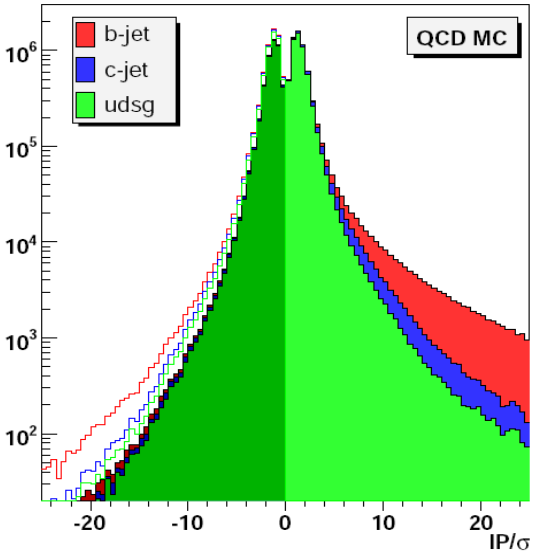
$$
\begin{cases}
\text{"tag" tagger} & n^{p_{Trel}} = \varepsilon_b^{p_{Trel}} n_b + \varepsilon_{cl}^{p_{Trel}} n_{cl} \\
 & p^{p_{Trel}} = \delta\, \varepsilon_b^{p_{Trel}} p_b + \gamma\, \varepsilon_{cl}^{p_{Trel}} p_{cl}
\end{cases}
$$

$$
\begin{cases}
\text{"tag"+"probe" tagger} & n^{tag,p_{Trel}} = \kappa_b\, \varepsilon_b^{tag} \varepsilon_b^{p_{Trel}} n_b + \kappa_{cl}\, \varepsilon_{cl}^{tag} \varepsilon_{cl}^{p_{Trel}} n_{cl} \\
 & p^{tag,p_{Trel}} = \kappa_b\, \beta\, \delta\, \varepsilon_b^{tag} \varepsilon_b^{p_{Trel}} p_b + \kappa_{cl}\, \alpha\, \gamma\, \varepsilon_{cl}^{tag} \varepsilon_{cl}^{p_{Trel}} p_{cl}
\end{cases}
$$

b tag performance measurement

Mistag rate using negative tags
CMS Note 2007/048

- Measures the tagging efficiency for uds-g jets (= mistag rates)
- Method uses jets with pT > 20 GeV from QCD samples, this data contains only small fraction of b and c jets



- Then tag the jets by using tracks with negative IP significance in order to compute

$$\varepsilon_{neg}^{Data}$$

One can then introduce a scale factor R_l

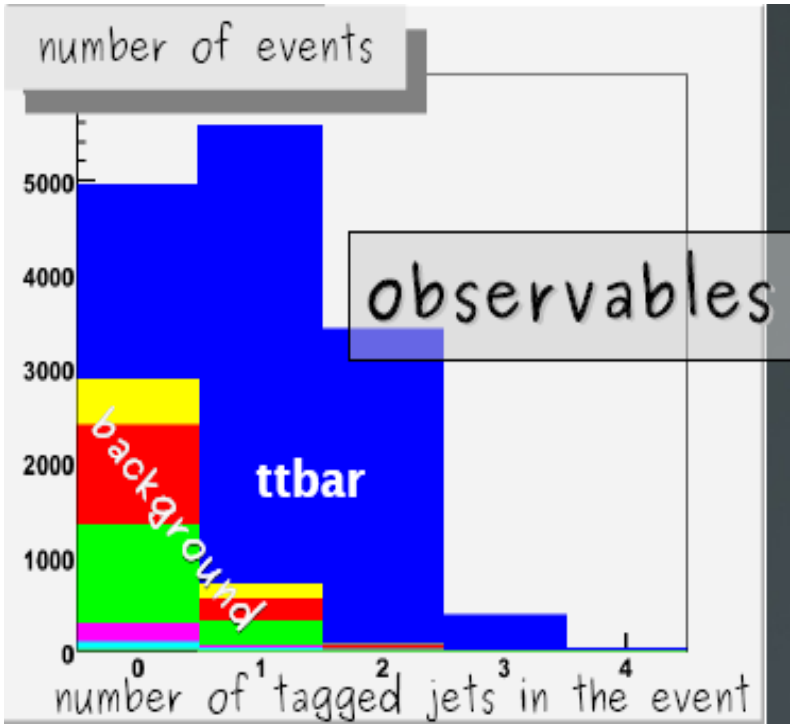
$$\varepsilon_{tag}^{Data}(udsg) = R_l \times \varepsilon_{neg}^{Data}(all)$$

$$R_l = \frac{\varepsilon_{tag}^{MCqcd}(udsg)}{\varepsilon_{neg}^{MCqcd}(all)}$$

b tag performance measurement

B-tag efficiency using ttbar samples
tag consistency method

- Currently implemented in the semi-muonic decay channel, dilepton channel is coming
- The number of events (F_{ijk}) with i,j,k, of b,c and light jets is extraced from MC (= jet flavor composition of sample)
- These fractions are used together with the ttbar cross section, the acceptance and the ε_b ε_c ε_l to compare the expected and observed number of tagged jets in an event
- The comparison is done using a maximum likelihood fit to extract the b an c tagging efficiency



Likelihood ratio method

Implemented in CMSSW_1_6_12 with TQAF running on CSA07 samples, for the semi-muonic and the dileptonic decay channels

Event Selection

Use cuts on lepton, jets and MET to select the ttbar semi-muon decay channel

Jet combination selection

Use likelihood ratio based selection to correclty assign the jets

Jet sample

This leads to a jet sample with enriched b jet content

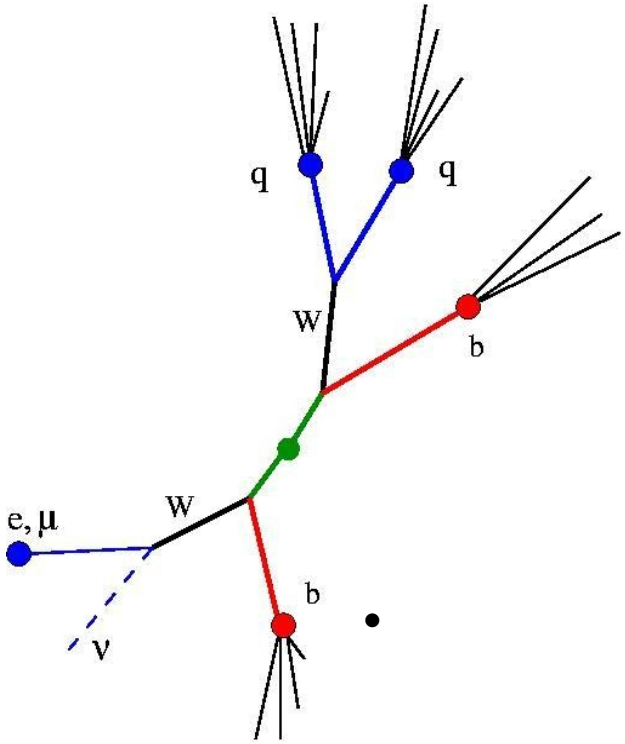
Data driven background estimation

Current work

εb measurement

$$\epsilon_b = \frac{1}{x_b}[x_{tag} - \epsilon_o(1 - x_b)]$$

Selection



Event Selection

- Muon with $|\eta| < 2.1$, $p_T > 30$ GeV and isolation criteria (trackIso < 3.0 Gev, caloIso < 6.0 GeV)
- At least 4 jets with $|\eta| < 2.4$ & $p_T > 40$ GeV
- Missing $E_T > 30$ GeV

Jet combination selection

Construct a combined likelihood ratio to choose the jet combination among the 12 possible solutions

- For chosen jet combination
  - Require the b coming from the hadronic decaying top to be loosely btagged: trackCountingHighEffJetTags > 2 (reduce W+jets background)
  - Require only solutions with converged kinematic fit $\chi^2 > 0$ for m_{thad}, m_{tlep}, m_{Whad}, m_{Wlep} constraints.
- The jet from the leptonic decaying top is used to form

Jet sample

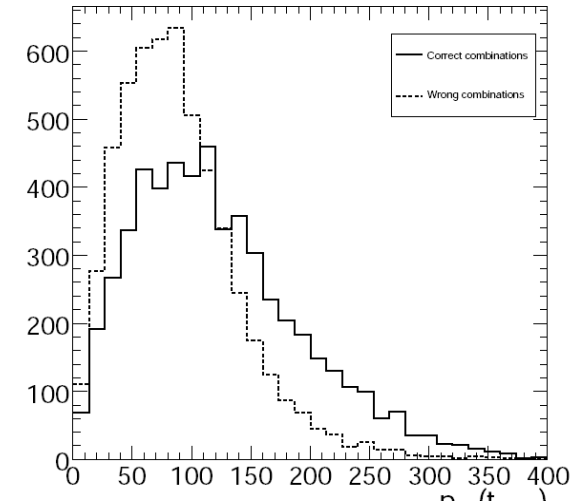
Likelihood ratio

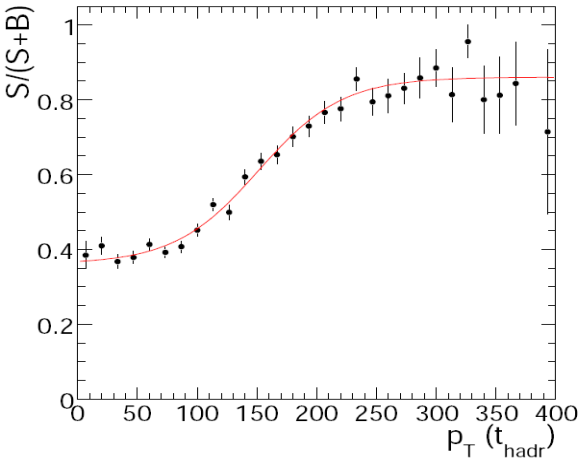
- After the basic event selection to select the semi-muonic ttbar events the correct assignment of jets should be found
- I use a Likelihood ratio based method to find the most likely solution among the 12 possible jet combination. (Based on the Neyman-Pearson hypothesis)
- As input 11 observables are used which have a different shape for good and bad assignements (solutions)

- Full line shows p_T of hadronic top for the good combination
- Dashed line for the bad combinations (norm.)

- From these distributions in a bin-by-bin way the ratio S/(S+B) is calculated
- Via a fit you obtain a function which gives for each value of the had. top p_T the probability to have signal (correct jet combination) at this value

- These probabilities are multiplied for several observables to obtain a total probability





Likelihood ratio

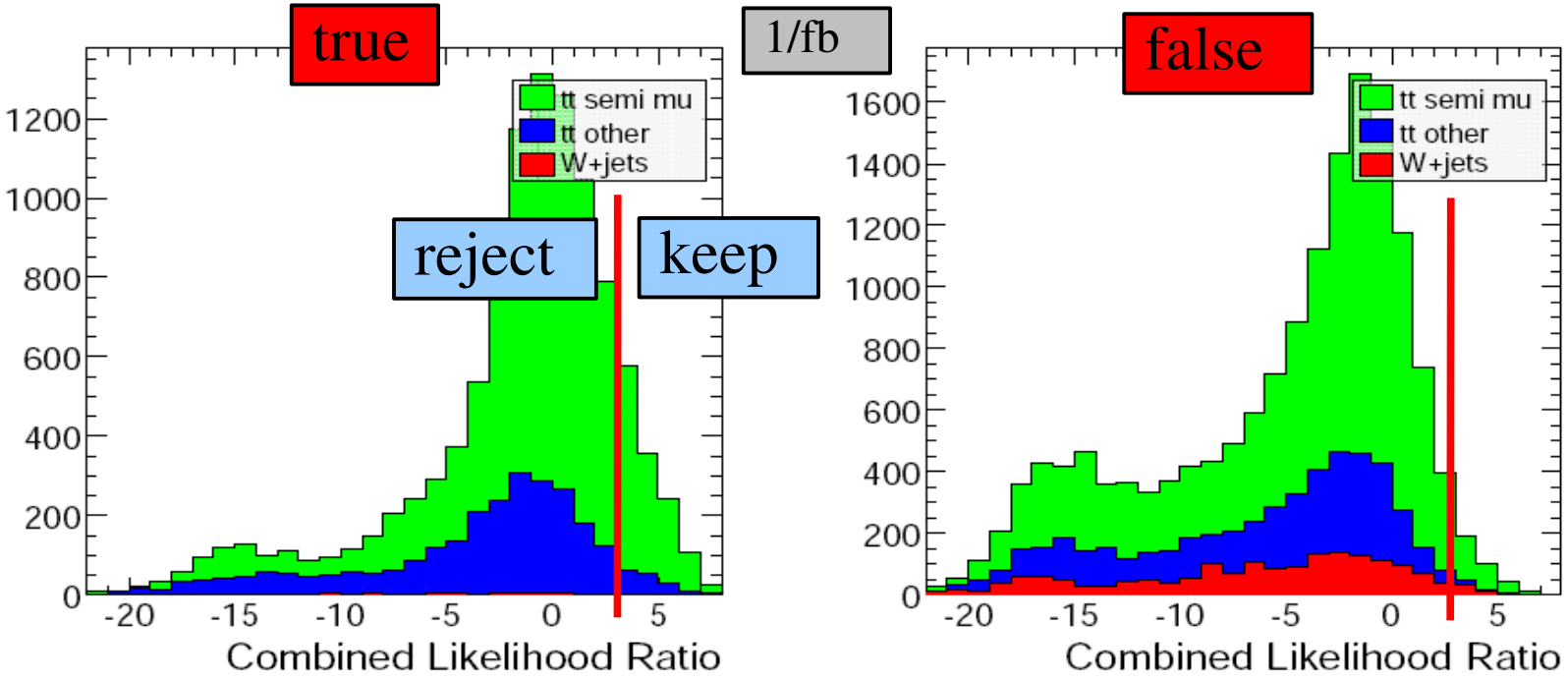
- For each observable the S/(S+B) are calculated and parameterized with $f_i(x_i)$, from this the Likelihood Ratio is calculated

$$\mathcal{L}_i(x_i) = \frac{f_i(x_i)}{1 - f_i(x_i)}$$

- The Likelihood Ratio functions are combined into one observable:

$$\mathcal{L} = \prod_i \mathcal{L}_i(x_i)$$

- Solution with highest value is most probably the correct one
- Leptonic b jet has b flavour ($\Delta(b_{quark}, b_{jet})<0.3$) in most likely jet combination?



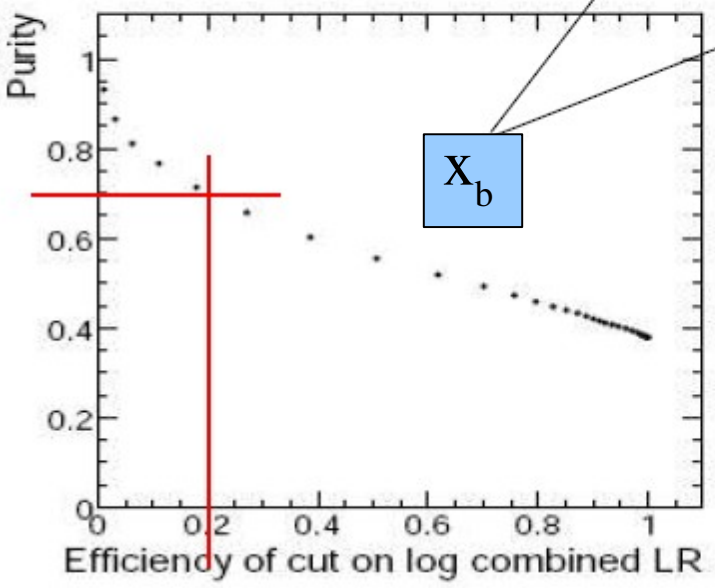
measurement

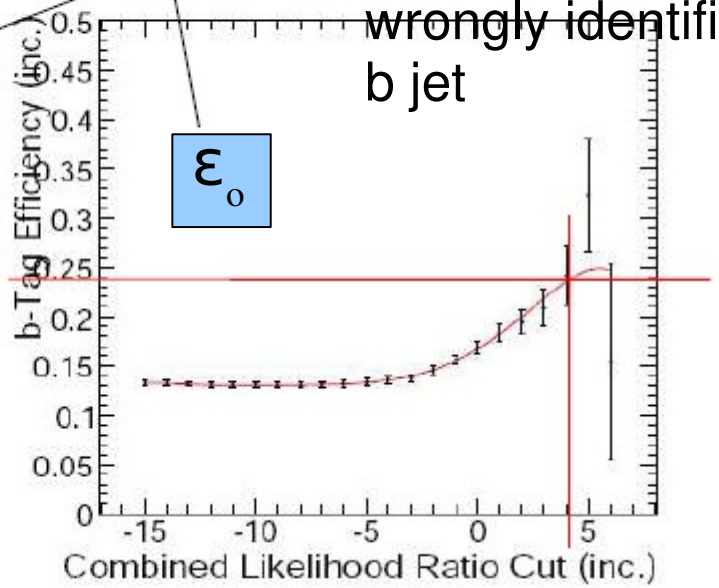
- By cutting harder on the combined likelihood ratio value we tend to have more correct event solutions so the purity of the jet sample, formed from the leptonic b jets increases → b-enriched jet sample
- a cut at 4 results in a purity x_b of 75% resulting in a jet sample with about 1100 jets for 100/pb
- For these jets we test if they are tagged by a b-tagging algorithm
  x_{tag} = #positive identified / #total number of jets in sample

$$\epsilon_b = \frac{1}{x_b}[x_{tag} - \epsilon_o(1 - x_b)]$$

from data

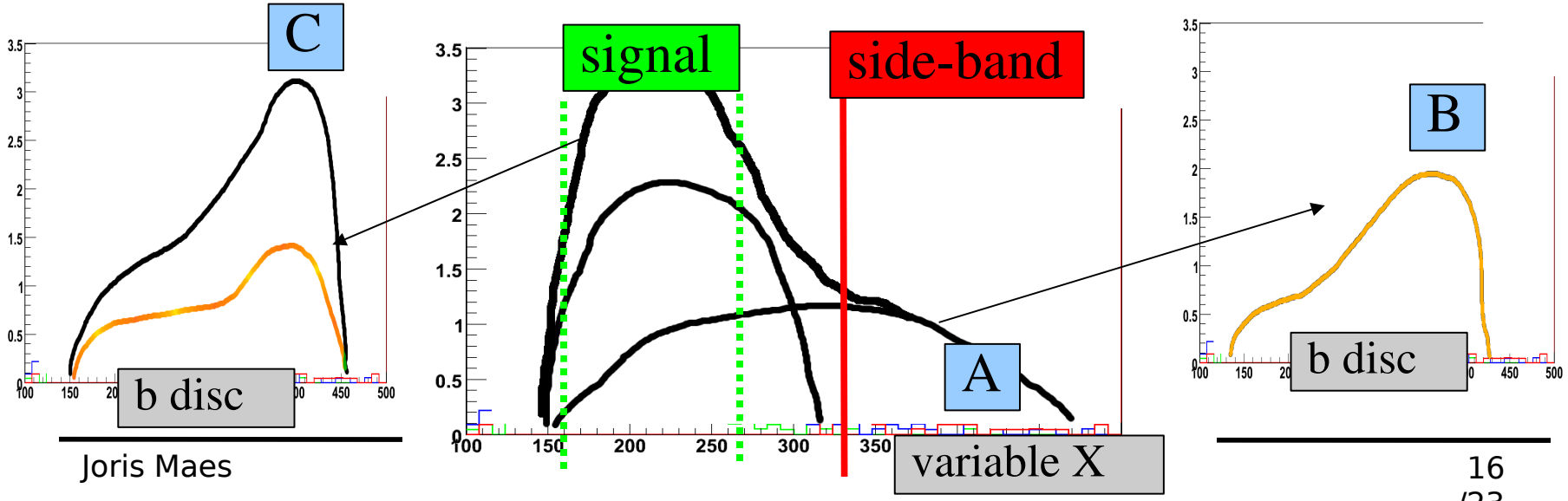
Faction of non b jets wrongly identified as b jet





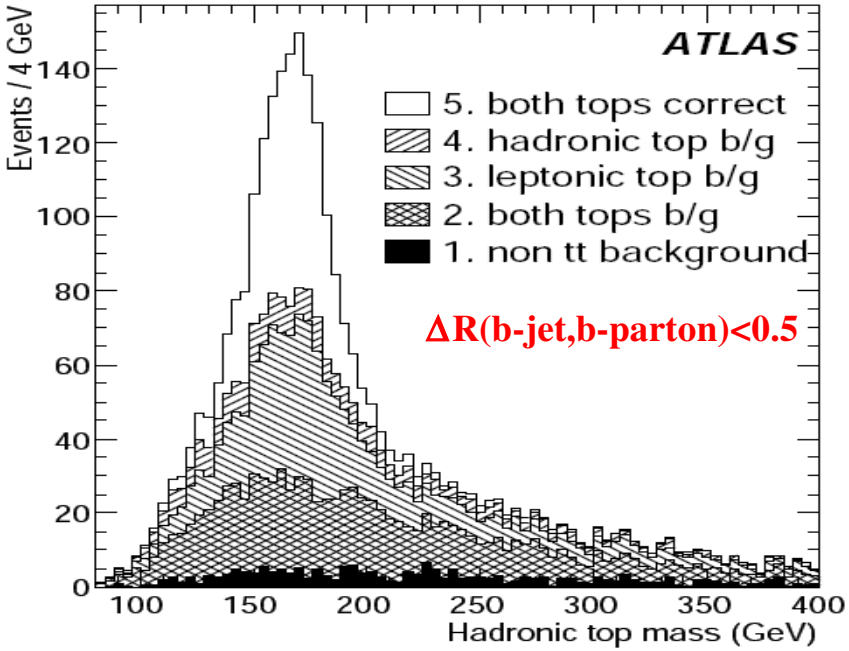
Data driven background control

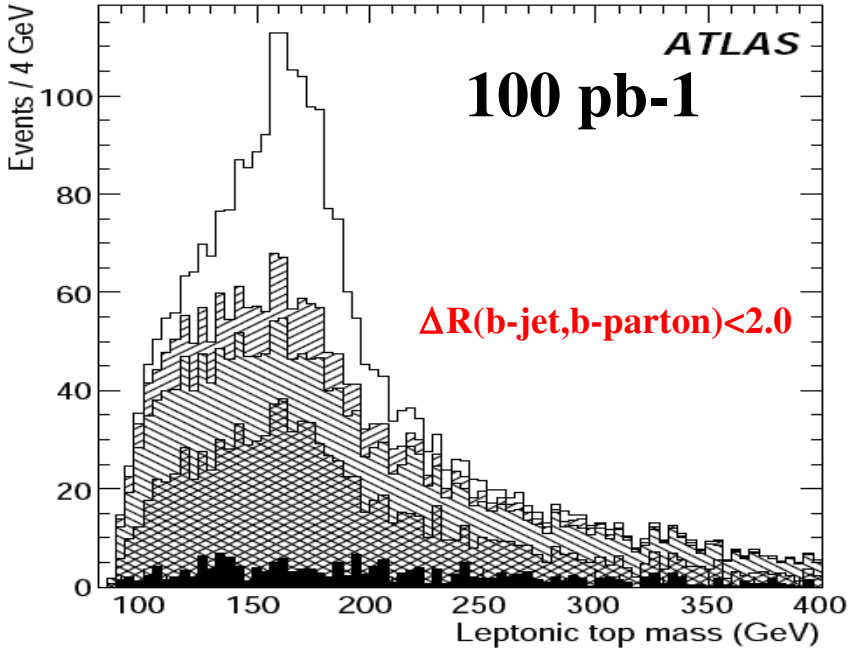
- The method uses the flavor composition from Monte Carlo and is therefor strongly dependent on the model we use.
- To overcome this we would like to estimate the b-tagging efficiency in a slightly different way using data
  - A. Find a variable with a clear background dominated region (side-band)
  - B. Obtain the b-tag discriminator distribution in this region (data)
  - C. Correct the b-tag discriminator value in the signal region (data) with the sideband distribution normalized with a factor from data
  - The side-band region distribution of the b-tag discriminator should be the same as in the signal region (to be checked)



ATLAS 'topological' method

- **Hadronic top:**
  - Di-jet combinations with $60 < m_{jj} < 100$ GeV/c^2 (j_1,j_2 are anti-b-tagged)
  - Combine this di-jet system with a b-tagged jet
  - $E_T(j_1 \text{ or } j_2)$ & $E_T(b) > 40$ GeV
- **Leptonic top:**
  - W formed with lepton and neutrino (p_z from W mass constraint)
  - One of the remaining jets is used to reconstruct the leptonic top
- **Combination with largest scalar sum of p_T of the 2 top quarks is chosen**

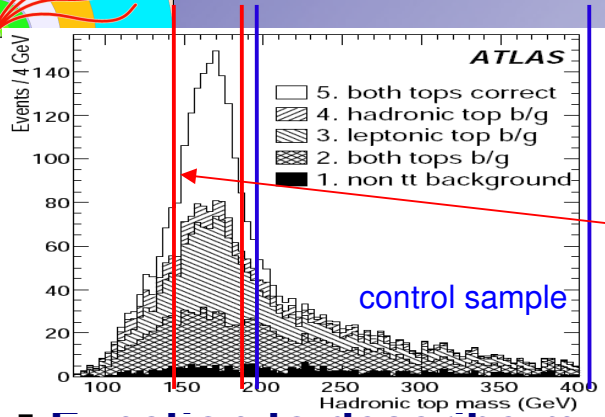




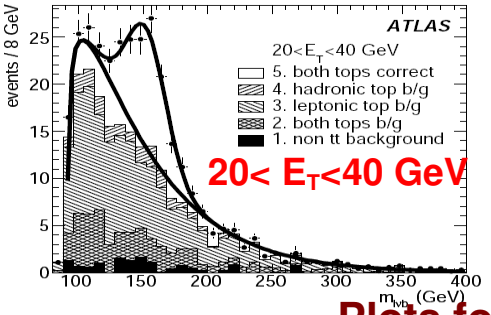
- **Goal: calculate b-tag weight for jets in the b-enriched sample, measure ε_b → background has to be well-estimated (next slide) and subtracted (next-to-next slide)**

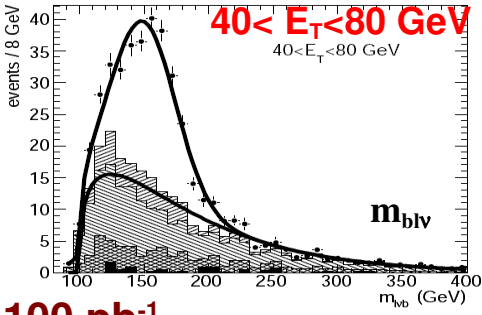
ATLAS

Background estimation

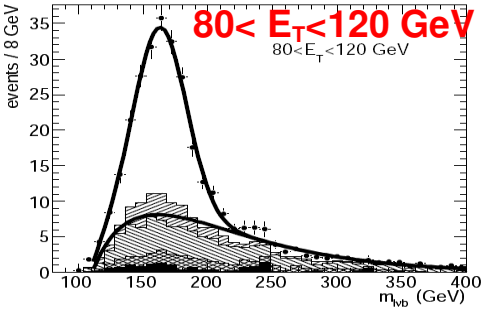


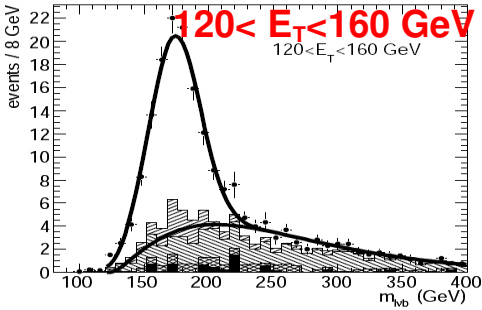
- **Function to describe m_{bln}**
  - **Signal sample: 140 <m_{bjj}<190 GeV/c^2 5 subsamples for leptonic b-jet ET**
  - **Control sample: 200 <m_{bjj}<400 GeV, leptonic b-jet anti-b-tagged → describes background-shape in m_{bln}**
  - **Fit simultaneously background shape and normalization to m_{bln} distributions of signal and control sample (fully data-driven)**

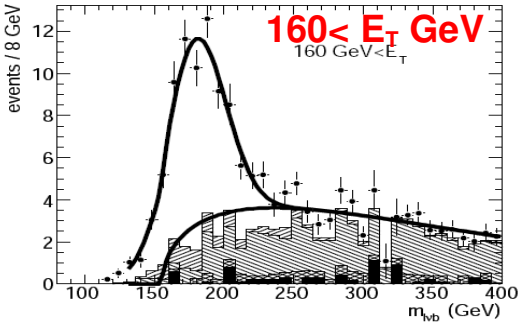




Plots for 100 pb^{-1}





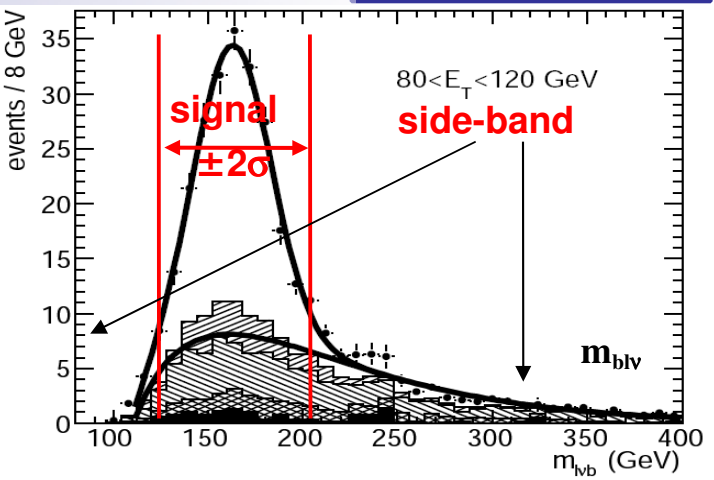


$$F_s(m_{b\ell\nu}) = c_4 F_b(m_{b\ell\nu}) + c_5 G((m_{b\ell\nu} - c_6)/c_7)$$

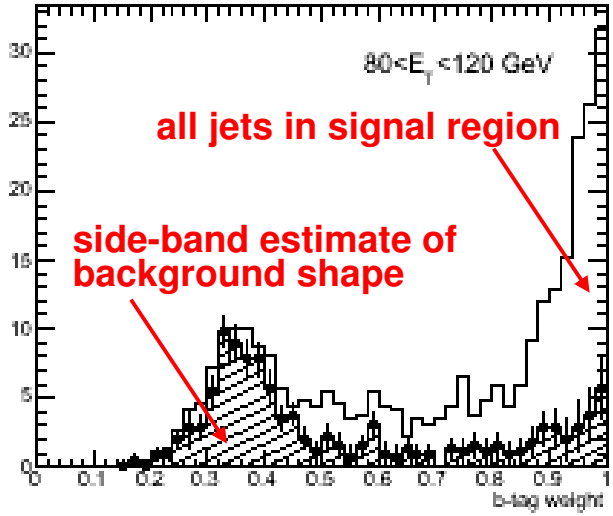
$$F_b(m_{b\ell\nu}) = E((m_{b\ell\nu} - c_0)/400),$$

$$E(x) = \begin{cases} c_1 x^{c_2} \exp(-c_3 x) & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

ATLAS
Background subtraction



- **Obtain b weight distribution in signal region**
  - **Jet sample from signal region: b-jets + other jets (from comb. background)**
  - **Other jets have same b-tag weight distribution as jets from events in the side-band**
- **Subtract scaled side-band b-tag weight distribution from signal b-tag weight distribution**

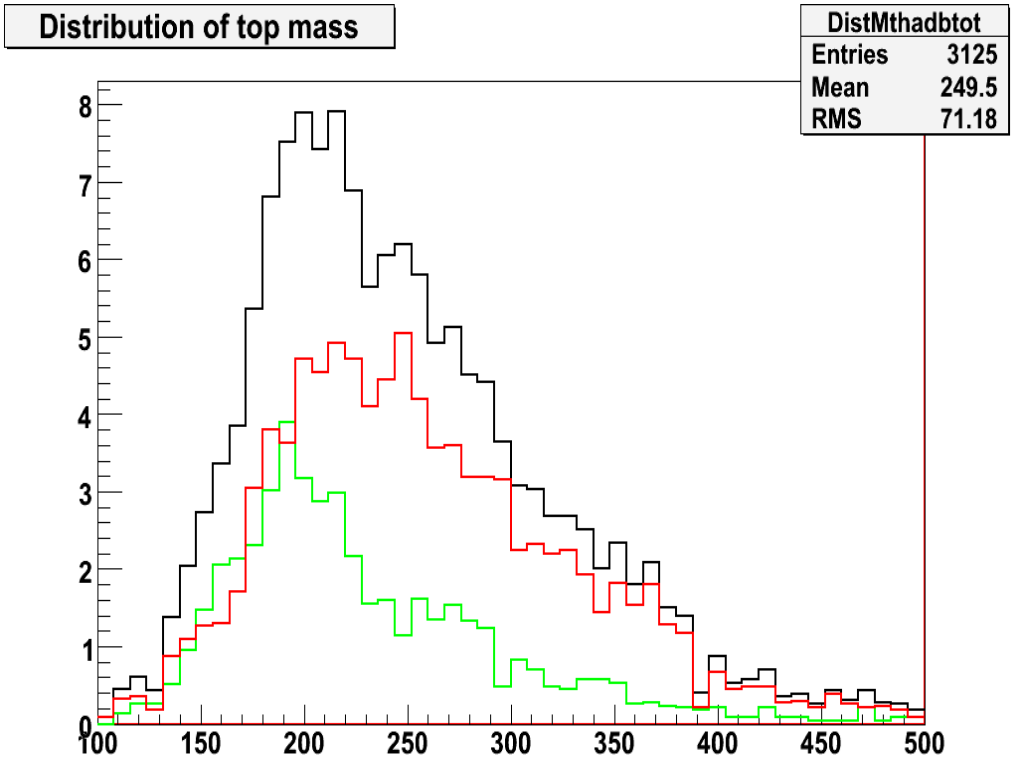


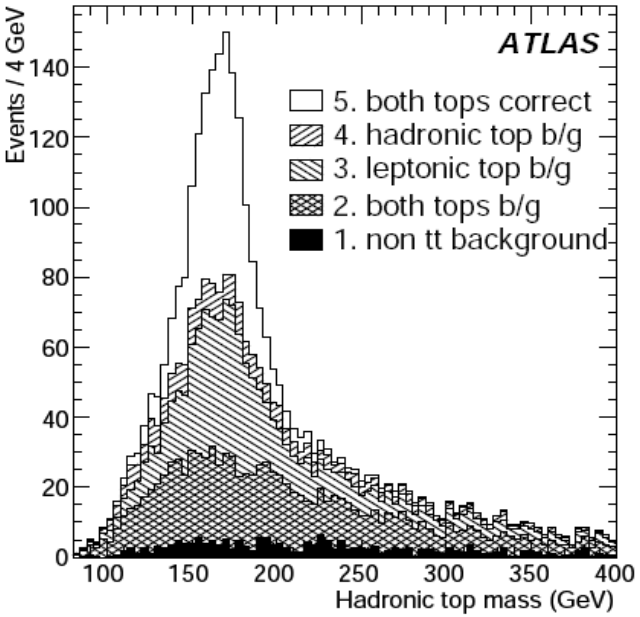
$$S = \frac{\int_{s_1}^{s_2} F_b(m_{b\ell\nu})\, dm_{b\ell\nu}}{\int_{b_1}^{s_1} F_b(m_{b\ell\nu})\, dm_{b\ell\nu} + \int_{s_2}^{b_2} F_b(m_{b\ell\nu})\, dm_{b\ell\nu}}$$

- **Do this for each E_T sub-sample**
- **The side-band estimate does subtract the background in the signal region properly**
- **Apply b-tagging criterion (cut on b-tag weight)**
- **Calculate ε_b from the remaining jets above this cut value**

Sideband region

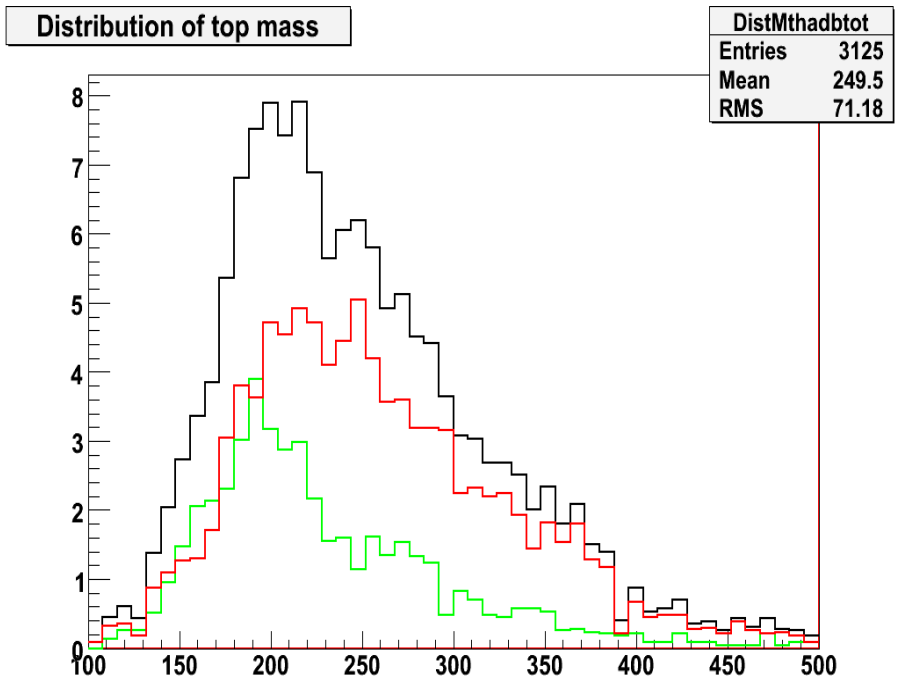
- We look at the distribution for calibrated hadronic top mass, 2 categories
  - Signal: leptonic top matches with generated parton ΔR<0.3 (green)
  - Background: leptonic doesn't match with generated parton (red)
- Sideband is contaminated with signal events
- Mass is not as peaked as in ATLAS case

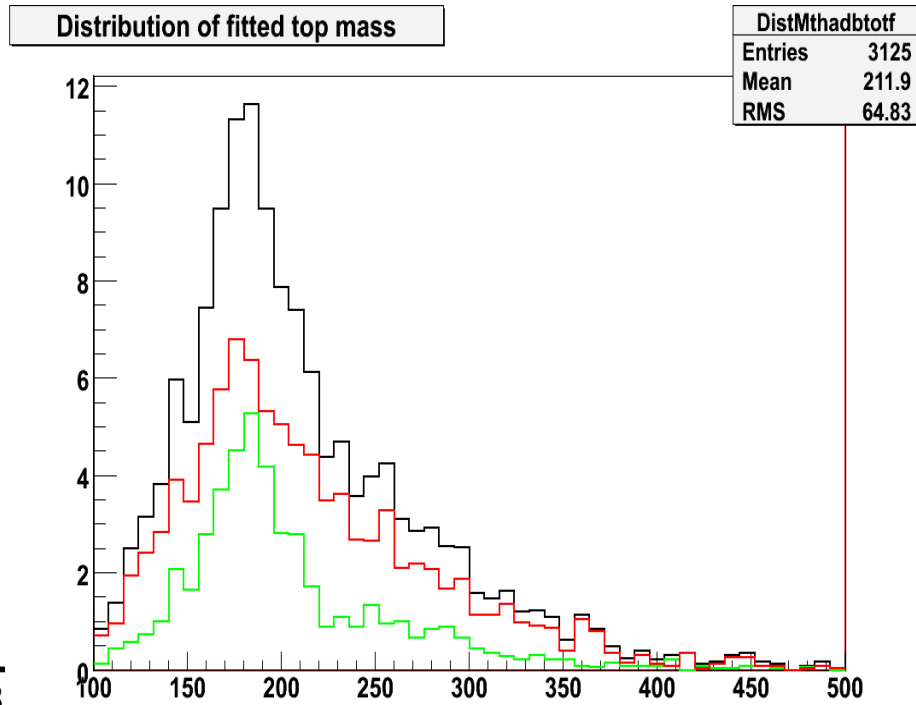




Sideband region

- In stead of the calibrated hadronic top mass we look at the fitted hadronic top mass. The 3 jets serving as input for the top mass calculated are adjusted by using the kinematic fit with m_{Whad} and m_{Wlep} contraint.
- The mass is more peaked
- Background region still contains too much b jets
- Other observables (e.g. ΔΦ(top, antitop)) show similar problem





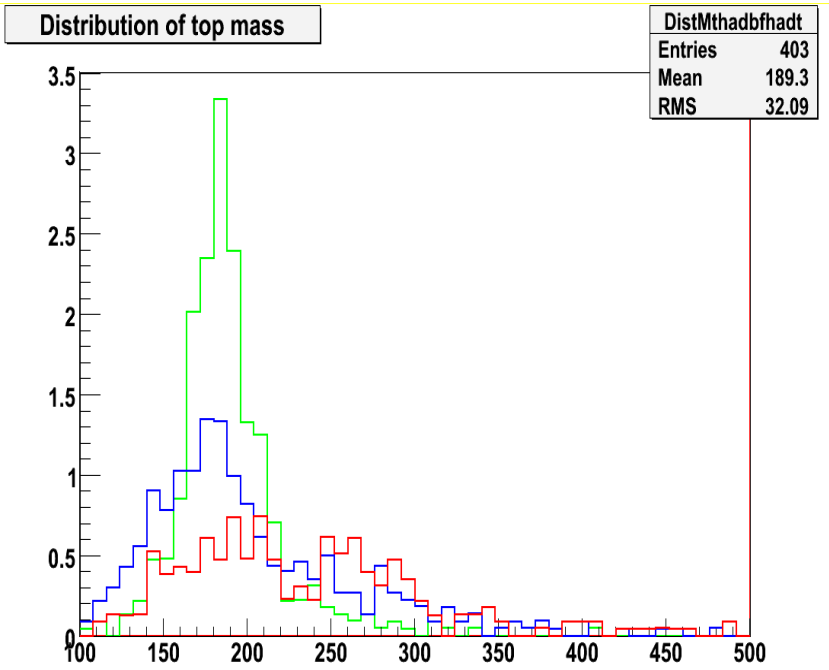
2008

Background region

- A breakdown of the top mass distribution (green) where the leptonic b jet has really b flavor, this shows that there is an irreducible part.
  - Green: top mass reconstructed with leptonic b and hadronic b jet matched with generated b jets
  - Red: top mass reconstructed with two b jets interchanged
  - Blue: hadronic b jet has not b flavour
    - can be that a light jet is interchanged with hadronic b jet
    - A wrong jet is picked up in the event
    - ...

- This can be understood by the fact that the likelihood ratio is biasing our sample.

  The likelihood ratio tends to select events which are compatible with the ttbar topology, so distinguishing inside this sample might be impossible



Future plans/ideas

- Since it is very hard to find an observable to isolate background after applying the likelihood ratio we should look for a different way to select the correct jet combination
- Study for which event selection which observables suits best for isolating a background sample
  - Should have significant amount of events

$$\mathrm{Sign} = \frac{S}{\sqrt{S+B}}$$

  - Should have a good separation of the signal vs. background

$$\mathrm{Sep} = \sum_{\mathrm{bin}} \left(S(i) * B(i) \right)$$

  - See if the b discriminant distribution can be used to substract the light/c jets in the signal region

- Port code to CMSSW_2_1_X, use PAT
- Run on the Summer08 samples