Estimation of btagging efficiency

Igor Marfin

DESY, Zeuthen,Germany NC PHEP, Minsk,Belarus

September 7, 2010

Outline

Methods of performance measuring

- Top-quark based methods:
  - Likelihood technique. Use likelihood cut to obtain highly enriched b-jet content of ttbar.
  - Flavor-tag consistency method. Minimize the log-likelihood function $L = 2\log\prod_n P(N_n, \bar{N}_n)$.
    N_n - measured ($\bar{N}_n$ - expected) number of events with $n = 0, 1, 2$ tagged jets
- The Ptrel method.
- The System8 method.

Top-quark based methods

Likelihood technique. Estimation of the bjets/tagged fraction in MC/data. Use likelihood cut.

- Preselection of events.
- Likelihood function $L = \prod_i f_i(x_i)$ from MC for x_i observables, $f_i = \frac{S}{S+B}$, $S(B)$ - x_i bin-by-bin distributions for Signal (Background). Use MVA.
- Impose a cut on Likelihood.
- The fraction of bjets $x_b = \frac{bjets}{alljets}$ from MC
- The mistag rate ϵ_0 from MC.
- The fraction of tagged jets $x_{tag} = \frac{tagjets}{alljets}$ **from data**.
- Btagging efficiency $\epsilon_b = (x_{tag} - \epsilon_0 * (1 - x_b))/x_b$.

Flavor-tag consistency method. Enforce a consistency between the predicted number of events with 0,1,2 tagged jets to the actual number of observed events.

- Log-likelihood $L = -2 \log \prod_n P(N_n, \bar{N}_n)$ to minimize.
- $\chi^2 = \sum_n \frac{(N_n - \bar{N}_n)^2}{\bar{N}_n}$ to minimize instead of the log-likelihood function.

Kinematical Fit

- 4 constraints on: $mW_{lep}, mW_{had}, mTop_{lep}, mTop_{had}$.
- The parametrization: $\vec{p} = (E_T cos\phi, E_T sin\phi, E_T sinh\eta), E = E_T cosh\eta$.
- Up to 7 jets descending ordered to construct the χ^2.
- Use only the combination of 4 fitted jets + fit. muon + MET for the minimal converged χ^2.
- Look at **CMS AN 2005/025**.

Events preselection

The selection of semimuonic ttbar events among overwhelming background.

- SisCone algorithm with $\Delta R = 0.5$ to construct jets.
- Use kT/antikt jet-clustering in newer versions of CMSSW.
- JES corrections L2L3.
- The lepton impact parameter d0 with respect to the offline Beamspot.
- $Reliso = (E_{calo}(Iso) + P_T(tracker, Iso))/P_T(\mu)$

Table: The Selection derived from TOP-09-003

Step	Description
Step1	$\geq$ 4 jets with $Pt > 30GeV$ (corrected), $\eta < 2.4$
Step2	One GM muon with : $Pt > 30GeV$, $\eta < 2.1$, $N(hits) >= 11$, $d0 < 200\mu$, $\chi^2/ndf < 10$, $Reliso < 0.05$
Step3	veto on electrons (no electrons which are GsfElectron, $\eta < 2.5$, $Pt > 30GeV$, $Reliso < 0.05$)

Kinematical variables for MVA training

Three different sets of the kinematical variables to train MVA. Two sets use fitted objects.

- Kinematical variables from CMS NOTE 2006/013 (CMSNOTE) with KinFitter.
- Kinematical variables of TQAF/TopEventSelection subpackage(TESKinFit) with Kinematical Fit.
- Kinematical variables of TQAF/TopEventSelection subpackage (TES) without Kinematical Fit.

The set CMSNOTE of the variables from CMS NOTE 2006/013

- p_T, η_T of: $hadtop, leptop, hadB, lepB$.
- $\Delta\phi(hadB, hadtop), \quad \Delta\theta(hadQ, hadQBar)$
- $\Delta\phi(hadB, hadW), \quad \Delta\phi(lepB, lepW)$
- pT_{3jet}/pT_{4jet}
- $\Delta M(leptop, lepW), \quad \Delta R(leptop, lepW)$
- $\Delta M(hadtop, hadW), \quad \Delta R(hadtop, hadW)$
- $Prob(\chi^2)$

The set TES of the kinematical variables with (non-)fitted objects

- $sum_{E_T} = \sum_{i=1}^{4} E_T(j_i)$
- $relEt1 = E_T(j_1)/sum_{E_T}$
- $MET.Et()$
- $mindijetmass = Min(Mass(j_i, j_k))/\sum_k M(j_k)$
- $maxdijetmass = Max(Mass(j_i, j_k))/\sum_k M(j_k)$
- $mindRjetlepton = Min(\Delta R(muon, j_k))$
- $lepeta = abs(\eta(muon))$
- $dphiMETlepton = \Delta\phi(MET, muon)$

Preselecton. MVA Trainig

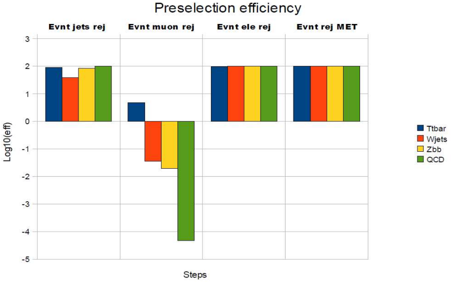


Fig:Preselection of events

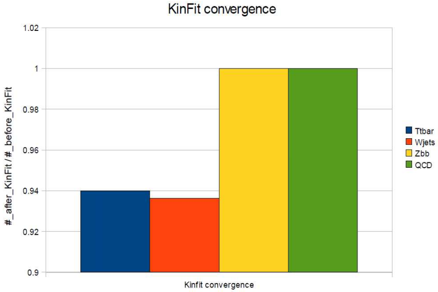


Fig:Convergence of kinematical fit

MVA Efficiencies

Fig:TESKinFit kinematical variables

MVA Cut. $= 0.773$ @ $20pb^{-1}$

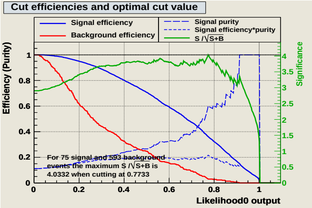


Fig:TES kinematical variables

MVA Cut. $= 0.633$ @ $20pb^{-1}$

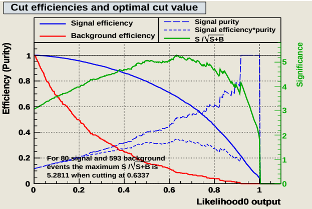


Fig:CMSNOTE kinematical variables

MVA Cut. $= 0.776$ @ $20pb^{-1}$

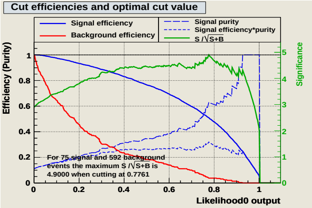


Table: Sig and Bkg events as input to train @ $20pb^{-1}$

Kin. set	Sig	Bkg
TES +kinfit	75	593
TES -kinfit	80	594
CMSNOTE	75	592

Analysis @ $20pb^{-1}$

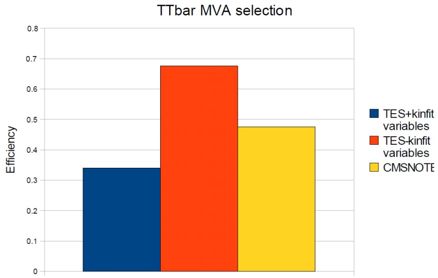


Fig:MVA filtering of signal events

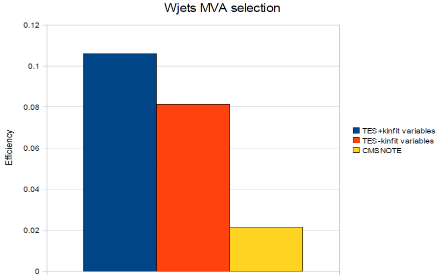


Fig:MVA filtering of wjets events

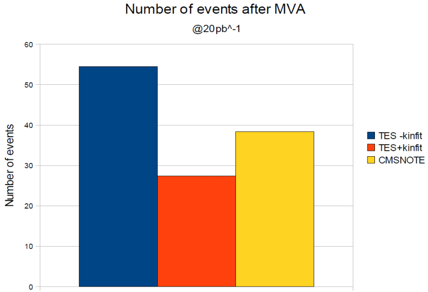


Fig:TTbar events after MVA @20pb^{-1}

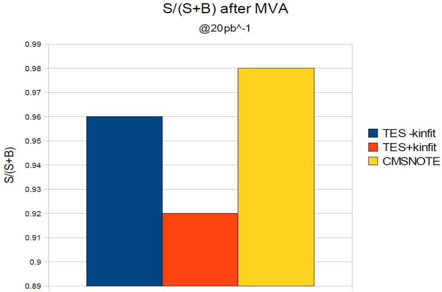


Fig:$S/(S+B)$ after MVA @20pb^{-1}

btagging efficiency plots

The plots corresponded to 'trackCountingHighEffBJetTags' btagging at the 'loose' operation point.

Fig:TES +kinfit variables **btagging eff.**

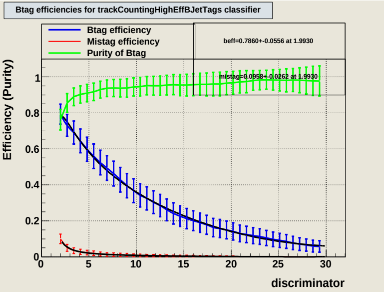
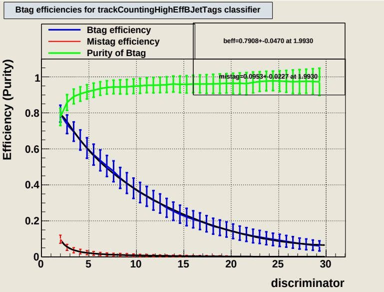


Fig:CMSNOTE variable**btagging eff.**



Estimation of btagging efficiency from pseudo-experiments

- 'trackCountingHighEffBJetTags' btagging algo at the 'loose' operation
- 300 pseudo-experiments on $20pb^{-1}$ data.
- CMSNOTE kinematical variables

Fig:Likelihood technique

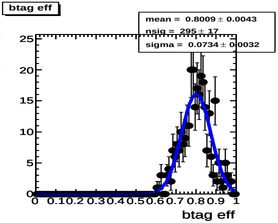


Fig:Flavor-tag consistency method

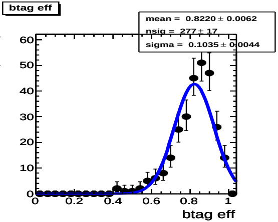


Fig:Likelihood technique. Pull distribution

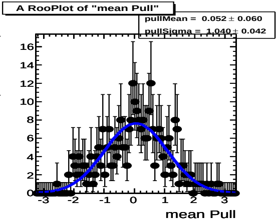
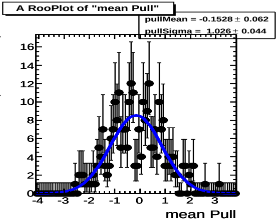


Fig:Flavor-tag consistency method. Pull distribution



- 300 pseudo-experiments on $100pb^{-1}$ data.

Fig:Likelihood technique

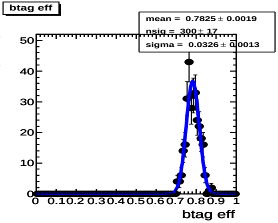


Fig:Flavor-tag consistency method

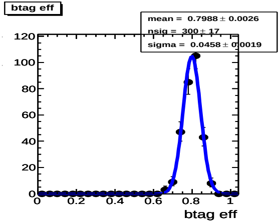


Table: Estimation of the btagging efficiency. TES +kinfit kinematical variables

Method	Lumi,pb^{-1}	estim. eff.	MC eff.
Likelihood technique	20	0.80 ± 0.07	0.78 ± 0.02
Flavor-tag consistency method	20	0.86 ± 0.09	0.78 ± 0.02
Likelihood technique	100	0.77 ± 0.03	0.78 ± 0.02
Flavor-tag consistency method	100	0.78 ± 0.05	0.78 ± 0.02

Table: Estimation of the btagging efficiency. CMSNOTE kinematical variables

Method	Lumi,pb^{-1}	estim. eff.	MC eff.
Likelihood technique	20	0.80 ± 0.07	0.78 ± 0.02
Flavor-tag consistency method	20	0.82 ± 0.10	0.78 ± 0.02
Likelihood technique	100	0.78 ± 0.03	0.78 ± 0.02
Flavor-tag consistency method	100	0.79 ± 0.04	0.78 ± 0.02

Table: Estimation of the btagging efficiency. TES -kinfit kinematical variables

Method	Lumi,pb^{-1}	estim. eff.	MC eff.
Likelihood technique	20	0.77 ± 0.06	0.78 ± 0.02
Flavor-tag consistency method	20	0.83 ± 0.08	0.78 ± 0.02
Likelihood technique	100	0.78 ± 0.02	0.78 ± 0.02
Flavor-tag consistency method	100	0.80 ± 0.03	0.78 ± 0.02

- 'trackCountingHighEffBJetTags' btagging algo at the 'medium' operation point
- 300 pseudo-experiments on $20pb^{-1}$ data.
- CMSNOTE kinematical variables

Fig:Likelihood technique

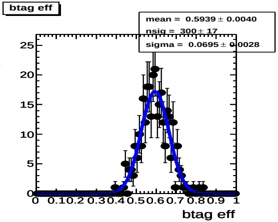


Fig:Flavor-tag consistency method

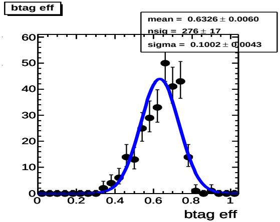


Table: Estimation of the btagging efficiency at 'medium' operation point. CMSNOTE variables

Method	Lumi, pb^{-1}	estim. eff.	MC eff.
Likelihood technique	20	0.59 ± 0.06	0.62 ± 0.02
Flavor-tag consistency method	20	0.63 ± 0.10	0.62 ± 0.02
Likelihood technique	100	0.60 ± 0.03	0.62 ± 0.02
Flavor-tag consistency method	100	0.61 ± 0.06	0.62 ± 0.02

Summary

- Two methods of b-jets identification efficiency on data were studied
- MVA selection of the process to have b-jets enriched samples/data have been developed and implemented in CMSSW.
- Three sets of kinematical variables were used for MVA.
- The kinematical set proposed in CMS NOTE 2006/013 gives the best estimation of btagging efficiency.
- Thanks for your attention.

BACKUP SLIDES

Methods of performance measuring

- The Ptrel method. This method is based on measuring $Ptrel = p_{\mu} \times p_{\mu+jet}/|p_{\mu+jet}|$ from events with two reco jets and one non-isolated muon before and after btagging. Then number of bjets before and after btagging can be fitted from Ptrel distribution with MC templates.
- The System8 method. Based on the same events as before but taking into the account cut on Ptrel and number of jets before and after btagging. Solving 8 equations on numbers of jets the performace is estimated.
- Top-quark based method: Likelihood technique. Using likelihood cut one can obtain semimuonic ttbar events with highly enriched b-jet content and suppressed background. Then the fraction of b-jets $x_b = \frac{bjets}{alljets}$ is calculated before and after btagging. Using x_b and mistag rates (estiamted from MC) one can get btagging efficiency.
- Top-quark based method:Flavor-tag consistency method. The btag efficiency and mistag rates can be obtained from minimazing log-likelihood function $L = 2 \log \prod_n P(N_n, \bar{N}_n)$, where N_n and $\bar{N}_n$ are the measured and expected number of events with $n = 0, 1, 2$ tagged jets. P is the Poisson distribution.

Description of the likelihood ratio technique

This method is based on estimating the fraction of bjets in MC and measuring tagged jets in experiments after preselection and likelihood cut.
Likelihood function is used to suppress remaining background (mainly from W+jets events) and get events with the highly enriched b-jet content.
The **MVA technique** is used to construct the likelihood function.

- The preselection of events is done.
- The likelihood function $L = \prod_i f_i(x_i)$ is constructed from MC, where x_i is some observable, $f_i = \frac{S}{S+B}$ with $S(B)$ - x_i distribution derived bin by bin way for Signal (Background).
- The fraction of bjets $x_b = \frac{bjets}{alljets}$ is estimated from MC events survived the selection and the likelihood cut.
- The mistag rate ϵ_0 is estimated from MC.
- The fraction of tagged jets $x_{tag} = \frac{tagjets}{alljets}$ is **measured from data** passed through the selection and the likelihood cut.
- One can calculate btagging efficiency as $\epsilon_b = (x_{tag} - \epsilon_0 * (1 - x_b))/x_b$.

As it was mentioned before, likelihood is being built using MVA. The cut is chosen at the value when the significance $\frac{S}{\sqrt{S+B}}$ reaches a maximum.

Description of the flavor-tag consistency method

Within the SM, top quarks are expected to decay almost to W boson accompanied by a b-quark.
In the semimounic ttbar events, given b efficiency and non-b mistag rate, the number of events with n_b tagged b-jets and n_{nonb} tagged nonb-jets can be predicted from MC.
By enforcing a consistency between the predicted number of events with no,one,two and more tagged jets to the actual number of observed events with that particular combination, the b-tag and non-btag efficiencies can be measured.

- The preselection of events is done.
- The MVA selection is performed to suppress the remain background (see the previous method).
- The following log-likelihood $L = -2 \log \prod_n P(N_n, \bar{N}_n)$ must be minimized. Here $N_n, \bar{N}_n, P$ are the measured number of events with $n = 0, 1, 2$ tagged jets, the expected number of events, the Poisson distribution.
- The function $\chi^2 = \sum_n \frac{(N_n - \bar{N}_n)^2}{\bar{N}_n}$ is minimized instead of the log-likelihood function.

- The expected number of events $\bar{N}_n$ is calculated as
  $\bar{N}_n = L \times \sigma_{ttbar} \times \epsilon_{sel}^{ttbar} \times$
  $\times \sum_{i,j} f_{ij}^{ttbar} \times \sum_{i'+j'=n}^{i' \leq i, j' \leq j} [C_i^{i'} \epsilon_b^{i'} \times (1 - \epsilon_b)^{(i-i')} \times C_j^{j'} \epsilon_{nonb}^{j'} \times (1 - \epsilon_{nonb})^{(j-j')}] +$
- $+L \times \sigma_{bkg} \times \epsilon_{sel}^{bkg} \times \sum_{i,j} f_{ij}^{bkg} \times \sum_{i'+j'=n}^{i' \leq i, j' \leq j} [\cdots],$
- where $L, \sigma_{ttbar(bkg)}, \epsilon_{sel}^{ttbar(bkg)}$ are the luminosity, cross section of signal (background), the preselection and MVA combined efficiency.
- The coefficients $f_{ij}^{ttbar(bkg)}, C_i^{i'}$ are the fraction of events with i, j of $b-$ and $nonb-$jets respectively, and the binomial coefficients.
- The method gives ϵ_b and ϵ_{nonb}

Samples used for study

- PYTHIA6 from SUMMER09@7TeV samples.
- https://twiki.cern.ch/twiki/bin/view/CMS/ProductionSummer2009at7TeV
- ttbar events: /TTbar/Summer09-MC_31X_V3_7TeV-v5/GEN-SIM-RECO/
- W+jets events: /Wmunu/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
- QCD events: /InclusiveMu15_Pt30/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO
- Zbb: /Zbb0Jets-alpgen/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO

Normalized distributions of the variables. I

Fig:TES +kinfit kinematical variable $|\eta_{lep}|$

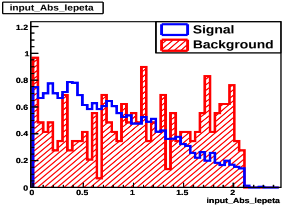


Fig:CMSNOTE kinematical variable η_{hadTop}

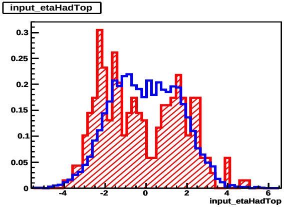


Fig:TES -kinfit kinematical variable $|\eta_{lep}|$

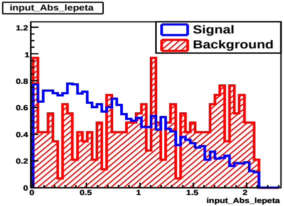
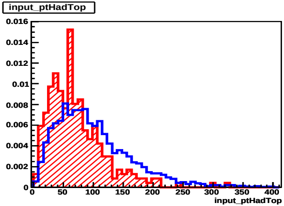


Fig:CMSNOTE kinematical variable pT_{hadTop}



Normalized distributions of the variables. II

Fig:TES +kinfit kinematical variable $max(M_{j1,j2})$

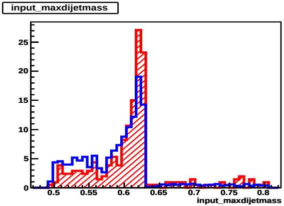


Fig:CMSNOTE kinematical variable $\Delta(\theta(q), \theta(\bar{q}))$

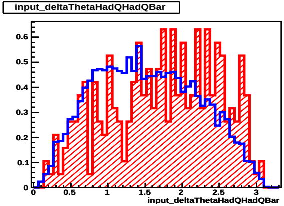


Fig:TES -kinfit kinematical variable $max(M_{j1,j2})$

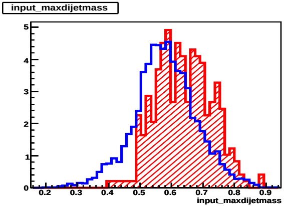
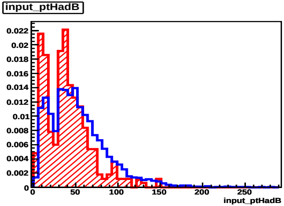


Fig:CMSNOTE kinematical variable pT_{hadB}



x_b distribution plots

There are several plots corresponded to 'trackCountingHighEffBJetTags' btagging algo at the 'loose' operation point.

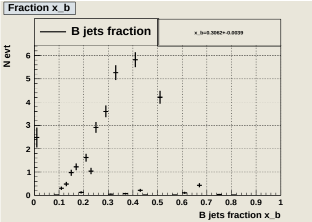


Fig:TES +kinfit variables x_b

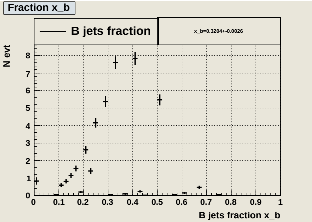


Fig:CMSNOTE variables x_b