Estimation of btagging efficiency

Igor Marfin

August 17, 2010

Outline

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 $\bar{N}_n = L \times \sigma_{ttbar} \times \epsilon_{sel}^{ttbar} \times$ $\times \sum_{i,j} f_{ij}^{ttbar} \times \sum_{\acute{i}+\acute{j}=n}^{\acute{i}\leq i,\acute{j}\leq j} [C_i^{\acute{i}} \epsilon_b^{\acute{i}} \times (1-\epsilon_b)^{(i-\acute{i})} \times C_j^{\acute{j}} \epsilon_{nonb}^{\acute{j}} \times (1-\epsilon_{nonb})^{(j-\acute{j})}] +$
- $+ L \times \sigma_{bkg} \times \epsilon_{sel}^{bkg} \times \sum_{i,j} f_{ij}^{bkg} \times \sum_{\acute{i}+\acute{j}=n}^{\acute{i}\leq i,\acute{j}\leq j} [\cdots],$
- 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 $f_{ij}^{ttbar(bkg)}, C_i^{\acute{i}}$ 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}

Kinematical Fit

This method is based on the *TKinFitter package* . More details are available in **CMS AN 2005/025**.

- 4 constrainsts are used: $mW_{lep}, mW_{had}, mTop_{lep}, mTop_{had}$.
- The constraints are Gaussian smeared: $mW = 80.4GeV \pm \Gamma_W(2.1GeV)$, $mTop = 173.GeV \pm \Gamma_{Top}(12.7GeV)$
- The parametrization is : $\vec{p} = (E_T cos\phi, E_T sin\phi, E_T sinh\eta), E = E_T cosh\eta$
- Up to 7 jets are used to construct the χ^2.
- Only such combination of 4 fitted jets with fit. muon and MET is used which has minimal χ^2 and has been converged.

Fig: Mass distribution of the fitted W boson

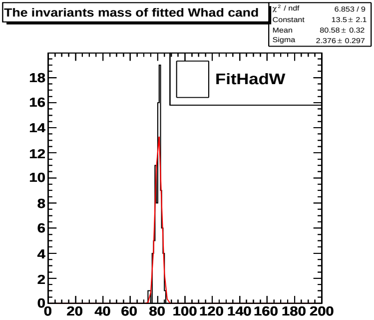
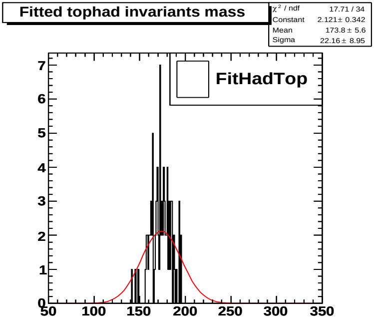


Fig: Mass distribution of the fitted top quark



Events preselection

The selection of semimuonic ttbar events among overwhelming background is done using the following steps.

- SisCone algorithm with $\Delta R = 0.5$ is used to construct jets.
- JES Corrections L2L3 are used
- The lepton impact parameter d0 is calculated 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, $\eta < 2.5$, $Pt > 30GeV$, $Reliso < 0.05$)
Step4	Reconstructed MET. No cuts on MET are applied

Kinematical variables for MVA training

Three different sets of the kinematical variables were used to train MVA. Two of them consider properties of fitted objects.

- Kinematical varibles from CMS NOTE 2006/013. KinFitter is used.
- Kinematical variables of TopEventSelection package with Kinematical Fit.
- Kinematical variables of TopEventSelection package without Kinematical Fit.

The first set of the variables was chosen at the same way as in CMS NOTE 2006/013

Kinematical variables from CMS NOTE 2006/013

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

The second set of the kinematical variables was build on fitted objects as well as non-fitted ones

Kinematical variables of TopEventSelection package with/without Kinematical Fit

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

Procedure of the kinematical fit

- Perform non-linear least-square kinemtical fit.
- The fit produces combinations of four fitted jets corresponding to b-,light-quarks fitted fitted muon and neutrino.
- Choose the combination with minimal χ^2.
- Construct $W_{lep}, W_{had}, t_{lep}, t_{had}$ candidates from fitted objects.
- Use kinematical variables to train MVA.

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
- Wbb :/Wbb0Jets-alpgen/Summer09-MC_31X_V3_7TeV-v2/GEN-SIM-REC
- Zbb: /Zbb0Jets-alpgen/Summer09-MC_31X_V3_7TeV-v1/GEN-SIM-RECO

All events from datasets were asked to process.
$'total\_number\_of\_events = -1'$

Preselecton. Cut Flow Table

Kinematical variables of TopEventSelection package with KinFit

Table for TTbarPresel

Xsection in pb 94.3

Evnt tot	Evnt jets rej	Evnt muon rej	Evnt ele rej	Evnt rej MET	GenEvt rej
209630	191503	9278	8963	8963	8963

Table for WjetsPresel

Xsection in pb 7899

Evnt tot	Evnt jets rej	Evnt muon rej	Evnt ele rej	Evnt rej MET	GenEvt rej
2022023	790086	283	283	283	283

Table for WbbPresel

Xsection in pb 5.0724

Evnt tot	Evnt jets rej	Evnt muon rej	Evnt ele rej	Evnt rej MET	GenEvt rej
377004	306843	22	21	21	21

Kinematical variables of TopEventSelection package with KinFit

Table for ZbbPresel

Xsection in pb 1.8046

Evnt tot	Evnt jets rej	Evnt muon rej	Evnt ele rej	Evnt rej MET	GenEvt rej
36618	30506	6	6	6	6

Table for QCDPresel

Xsection in pb 6.116e+07

Evnt tot	Evnt jets rej	Evnt muon rej	Evnt ele rej	Evnt rej MET	GenEvt rej
6411818	6409883	3	3	3	3

MVA Training. Cut Flow Table

Kinematical variables of TopEventSelection package with KinFit

Table for TTbarTrainMVA

Xsection in pb 94.3

Evnt tot	Accept to train
4696	4414

Table for WjetsTrainMVA

Xsection in pb 7899

Evnt tot	Accept to train
283	265

Table for WbbTrainMVA

Xsection in pb 5.1528

Evnt tot	Accept to train
21	21

Kinematical variables of TopEventSelection package with KinFit

Table for ZbbTrainMVA

Xsection in pb 1.8046

Evnt tot	Accept to train
6	6

Table for QCDTrainMVA

Xsection in pb 6.116e+07

Evnt tot	Accept to train
3	3

MVA Efficiencies

Fig:TopEventSelection kinematical variables with kinfit

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

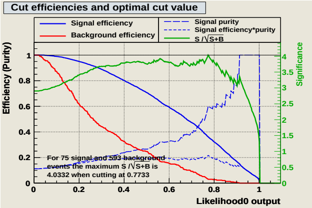


Fig:TopEventSelection kinematical variables w/o kinfit

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

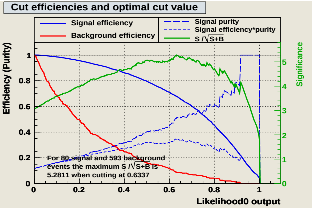


Fig:CMS NOTE 2006/013 kinematical variables with kinfit

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

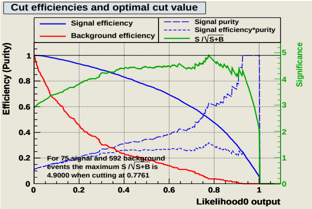


Table: Sig and Bkg events accepted to train @ $20pb^{-1}$

Kin. set	Sig	Bkg
TQAF +kinfit	75	593
TQAF -kinfit	80	594
CMS NOTE	75	592

Normalized distributions of the variables. I

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

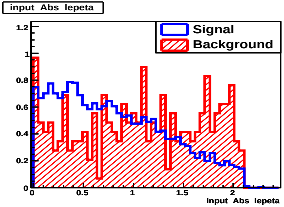


Fig:CMS NOTE 2006/013 kinematical variable η_{hadTop}

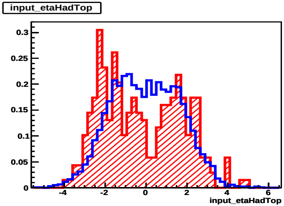


Fig:TopEventSelection -kinfit kinematical variable $|\eta_{lep}|$

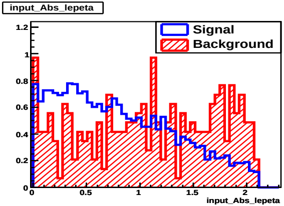
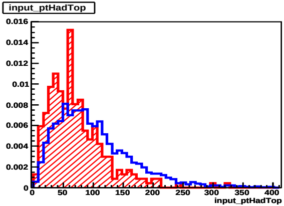


Fig:CMS NOTE 2006/013 kinematical variable pT_{hadTop}



Normalized distributions of the variables. II

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

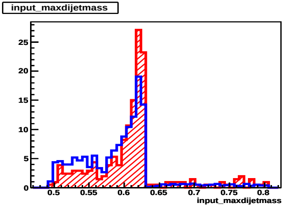


Fig:CMS NOTE 2006/013 kinematical variable $\Delta(\theta(q), \theta(\bar{q}))$

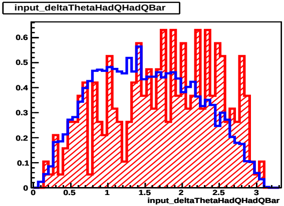


Fig:TopEventSelection -kinfit kinematical variable $max(M_{j1,j2})$

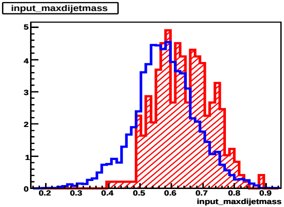
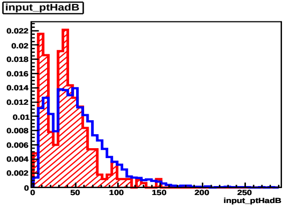


Fig:CMS NOTE 2006/013 kinematical variable pT_{hadB}



Analysis @ $20pb^{-1}$. Cut Flow Table

One needs to calculate x_b fraction and f_{ij} coefficients from MC events. They will be used for the estimation of btagging efficiency from real data later.

Kinematical variables of TopEventSelection package with KinFit
Table for TTbarAnalEstimMVA
Xsection in pb 9.430000e+01
Lumi: 2.000000e+01

Evnt tot	Evnt pass pres	Presel Eff	Evnt pass mva	MVA Eff
1886	80.6384	0.04275	27.4061	0.33986

<x_b>	btag-j/b-j	nonbtag/nonb
0.3087	0.79551	0.10332

Kinematical variables of TopEventSelection package with KinFit
Table for WjetsAnalEstimMVA
Xsection in pb 7.899000e+03
Lumi: 2.000000e+01

Evnt tot	Evnt pass pres	Presel Eff	Evnt pass mva	MVA Eff
157980	22.1107	0.00013	2.34389	0.106

<x_b>	btag-j/b-j	nonbtag/nonb
0.0124	0	0.06918

Kinematical variables of CMS NOTE
Table for WjetsAnalEstimMVA
Xsection in pb 7.899000e+03
Lumi: 2.000000e+01

Evnt tot	Evnt pass pres	Presel Eff	Evnt pass mva	MVA Eff
157980	22.1107	0.00013	0.46877	0.0212

<x_b>	btag-j/b-j	nonbtag/nonb
0	0	0.03333

Kinematical variables of CMS NOTE
Table for TTbarAnalEstimMVA
Xsection in pb 9.430000e+01
Lumi: 2.000000e+01

Evnt tot	Evnt pass pres	Presel Eff	Evnt pass mva	MVA Eff
1886	80.6384	0.04275	38.3788	0.4759

<x_b>	btag-j/b-j	nonbtag/nonb
0.3034	0.79574	0.09876

Kinematical variables of TopEventSelection package w/o KinFit
Table for TTbarAnalEstimMVA
Xsection in pb 9.430000e+01
Lumi: 2.000000e+01

Evnt tot	Evnt pass pres	Presel Eff	Evnt pass mva	MVA Eff
1886	80.6384	0.04275	54.4858	0.67568

<x_b>	btag-j/b-j	nonbtag/nonb
0.3067	0.79548	0.1014

Table: Comparison of different kinematical sets.

Kin. set	**Sig. presel.**	$\frac{S}{S+B}$ **presel.**	**Sig. MVA**	$\frac{S}{S+B}$ **MVA**
TQAF + kinfit	80.6384	0.11	27.4061	0.9209
CMS NOTE + kinfit	80.6384	0.11	38.3788	0.9876
TQAF -kinfit	80.6384	0.11	54.4858	0.9679

btagging efficiency and x_b distribution plots

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

Fig:TopEventSelection +kinfit **btagging eff.**

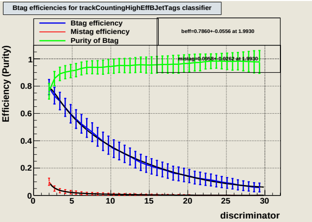


Fig:CMS NOTE 2006/013 **btagging eff.**

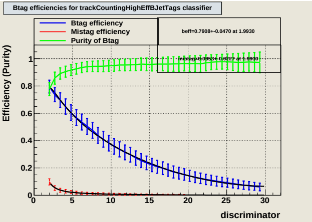


Fig:TopEventSelection +kinfit x_b

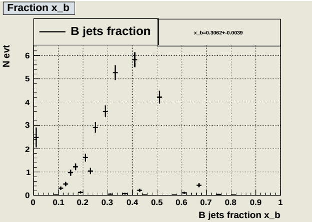
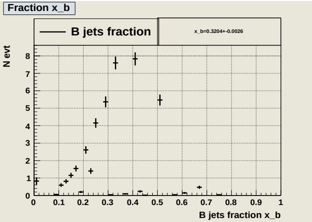


Fig:CMS NOTE 2006/013 x_b



Estimation of btagging efficiency from pseudo-experiments

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

Fig:Likelihood technique

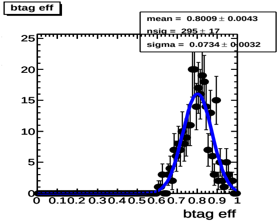


Fig:Flavor-tag consistency method

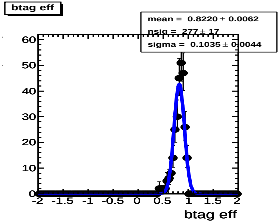


Fig:Likelihood technique. Pull distribution

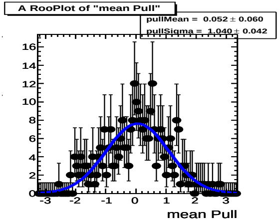
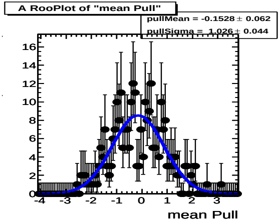


Fig:Flavor-tag consistency method. Pull distribution



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

Fig:Likelihood technique

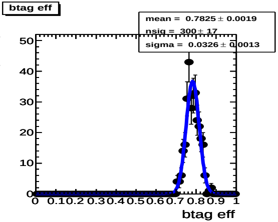


Fig:Flavor-tag consistency method

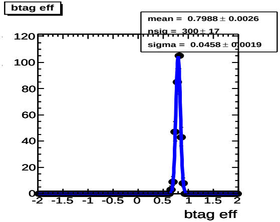


Fig:Likelihood technique. Pull distribution

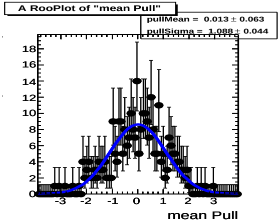


Fig:Flavor-tag consistency method. Pull distribution

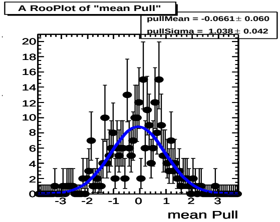


Table: Estimation of the btagging efficiency. TopEventSelection & kinfit kinematical variables

Method	**Lumi,** pb^{-1}	**estim. eff.**	**MC eff.**
Likelihood technique	20	0.80 ± 0.07	0.78 ± 0.05
Flavor-tag consistency method	20	0.82 ± 0.10	0.78 ± 0.05
Likelihood technique	100	0.78 ± 0.03	0.78 ± 0.02
Flavor-tag consistency method	100	0.79 ± 0.04	0.78 ± 0.02

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

Fig:Likelihood technique

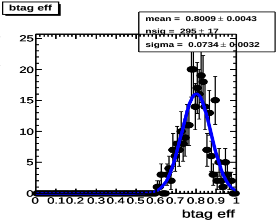


Fig:Flavor-tag consistency method

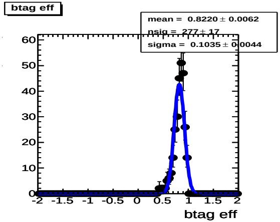


Fig:Likelihood technique. Pull distribution

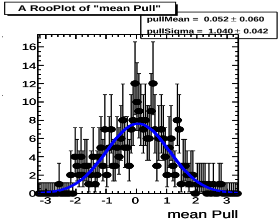
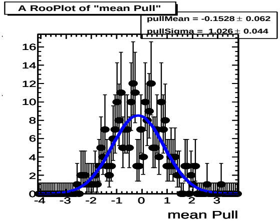


Fig:Flavor-tag consistency method. Pull distribution



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

Fig:Likelihood technique

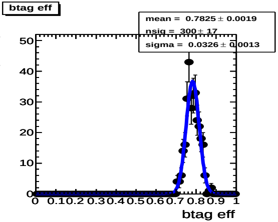


Fig:Flavor-tag consistency method

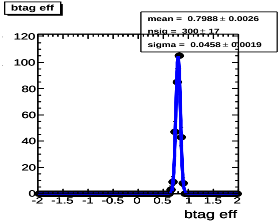


Fig:Likelihood technique. Pull distribution

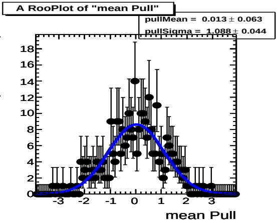


Fig:Flavor-tag consistency method. Pull distribution

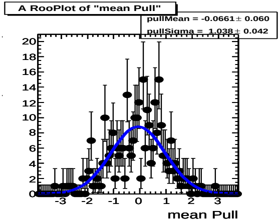


Table: Estimation of the btagging efficiency. CMS NOTE kinematical variables

Method	**Lumi,** pb^{-1}	**estim. eff.**	**MC eff.**
Likelihood technique	20	0.80 ± 0.07	0.78 ± 0.05
Flavor-tag consistency method	20	0.82 ± 0.10	0.78 ± 0.05
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. TopEventSelection w/o kinfit kinematical variables

Method	**Lumi,** pb^{-1}	**estim. eff.**	**MC eff.**
Likelihood technique	20	0.77 ± 0.06	0.78 ± 0.05
Flavor-tag consistency method	20	0.83 ± 0.08	0.78 ± 0.05
Likelihood technique	100	0.78 ± 0.02	0.78 ± 0.02
Flavor-tag consistency method	100	0.80 ± 0.03	0.78 ± 0.02

Summary

- The study has been finished. The following was done:
- MVA selection of the process have been developed and implemented in CMSSW.
- Two methods of btagging calibration were studied.
- Three sets of kinematic variables were used.
- The kinematic variables determined as in CMS NOTE 2006/013 are the best choice to make the estimation of btagging efficiency.
- Thanks for your attention.