Status update on $H \to b\bar{b}$ analysis on semileptonic channel

Antonio Branca, Ugo Gasparini, Andrea Gozzelino, Tommaso Dorigo, Kostya Kanishev, **Stefano Lacaprara**, Paolo Ronchese, Mia Tosi, Alberto Zucchetta

INFN Padova

Hbb meeting,
CERN, 18/11/2011

Strategy

Get QCD background from data

- Major background source is QCD
- Build B-tagging matrixes for $bb + j$ sample in control region;
- Estimate *bbb* background in signal region starting from *bbj*:

$$F(bbb) = F(bbj) \times P_b^{3rd-j}(j)$$

where

$$P_b^{3rd-j}(j) = \epsilon_b \cdot f_b + \epsilon_c \cdot f_c + \epsilon_l \cdot f_l$$

- Get $\epsilon's$ from MC and check on Data;
- and $f_{b,c,l}$ from Data (not for today's talk)
- Compare single distribution (M_{bb}) or MVA variable;

Baseline Selection

- HLT trigger fired;
- At least one μ, $p_t > 15\ GeV$, no isolation requirement;
- At least 3 jets
  - AK5PF, `PFNoPU` and `L1FastJet` correction, JetID loose;
  - inside $|eta| < 2.6$
  - $E_t^{(1,2)} > 25\ GeV$ $E_t^{(3)} > 20\ GeV$;
- B-tagging: CSV (Combined Secondary Vertex);
  - thresholds: $CSV > 0.8$ for first two jets, $CSV > 0.7$ for third;
- for final analysis
  - The first 3 jets, sorted in E_t, has to pass CSV threshold as above;
  - μ must be inside one of the first two jets;
- Still to be optimized

Data Analized

Dataset	$\int \mathcal{L}dt\ [pb^{-1}]$		num. triggers	
	Delivered	Selected	All P.D.	Just Hbb HLTs
May10ReReco-v1	248.859	215.733	9'267'331	3'370'074
PromptReco-v4	1037.	930.211	28'130'919	754'129
Run2011A-05Aug2011-v1	439.810	336.590	6'295'087	806'901
PromptReco-v6	510.046	543.693	8'008'055	1'082'103
Run2011B-PromptReco-v1	2833.	2478.	30'527'184	still running
Total	5069.72	4414.173	82'228'576	6'013'207

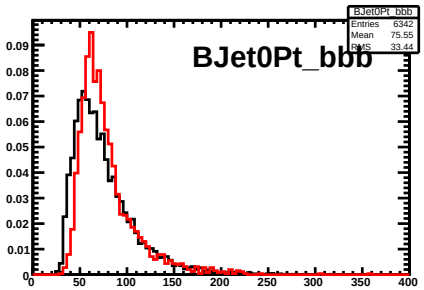
All numbers to be checked

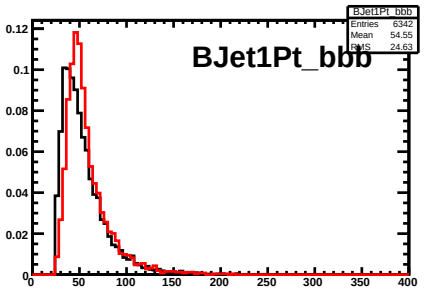
Build a Discriminator

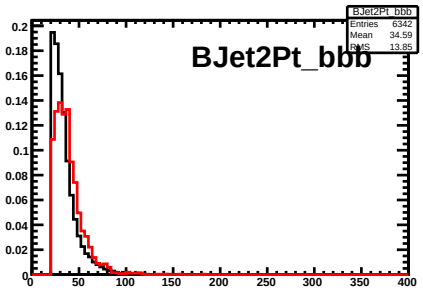
Compare

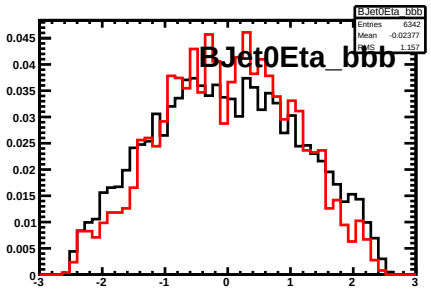
- Compare several distributions for signal $H(M = 120\ GeV)$ and background QCD MC
- Cannot rely on QCD well enough to use directly the distributions or the discriminator to compare with data
- But we can find a control region signal-free.
- Can use data in control region to train B-tagging matrices and check the prediction from *bbj* to *bbb*
- Can be used also as MVA analysis once the background distrubtion are obtained from data ($bbj \rightarrow bbb$).

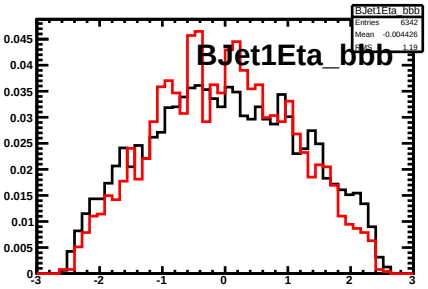
Discriminator variables I

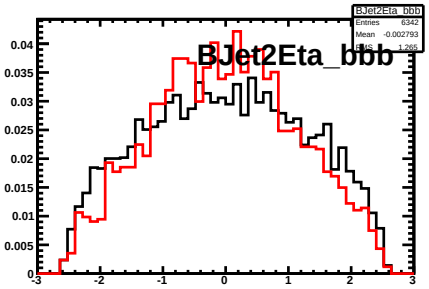


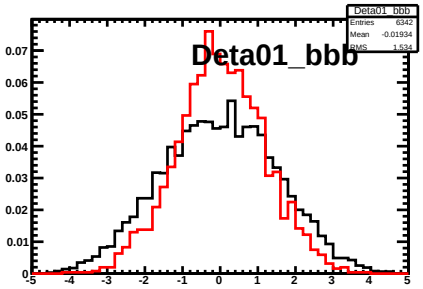


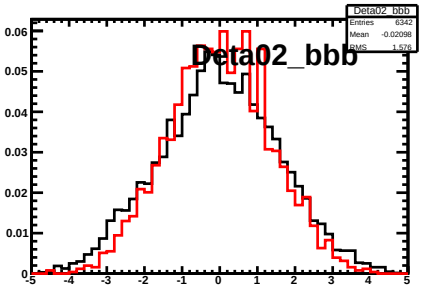


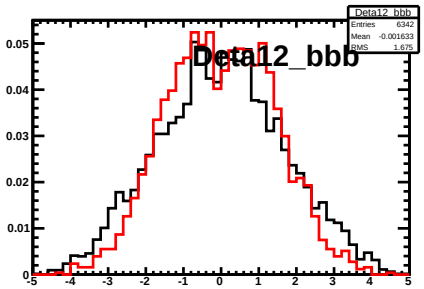


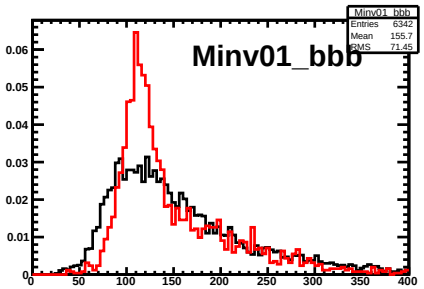


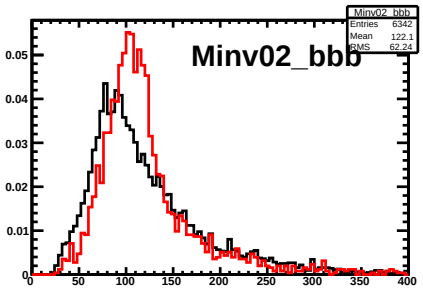


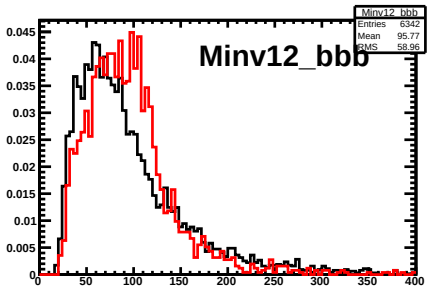




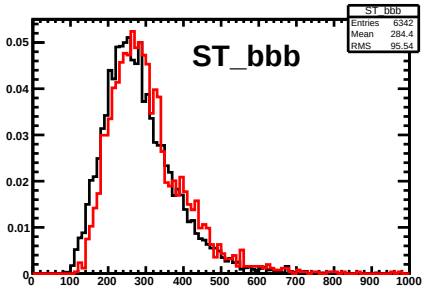


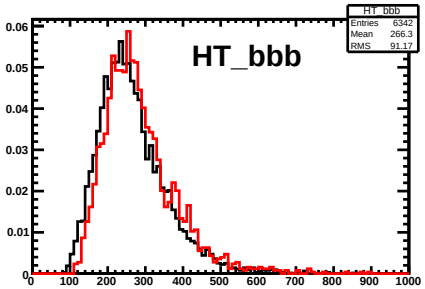


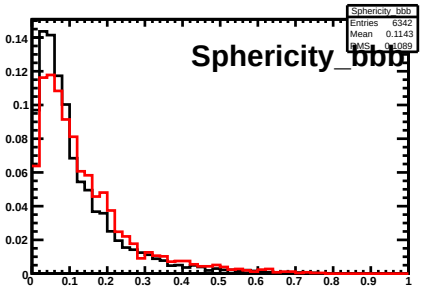


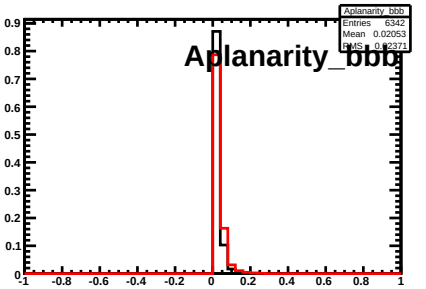


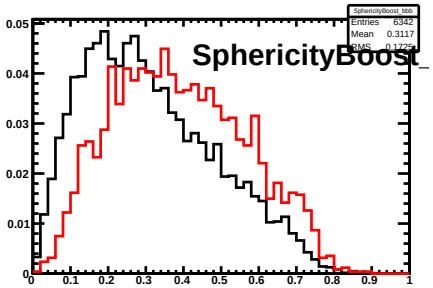
Discriminator variables II

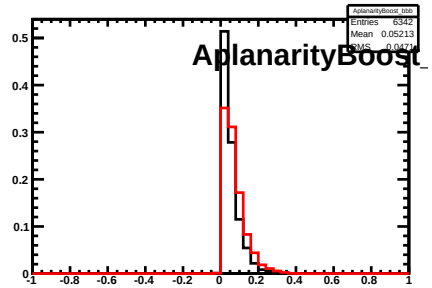


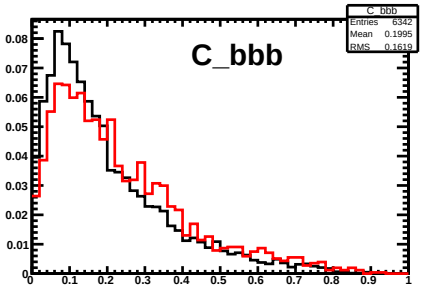


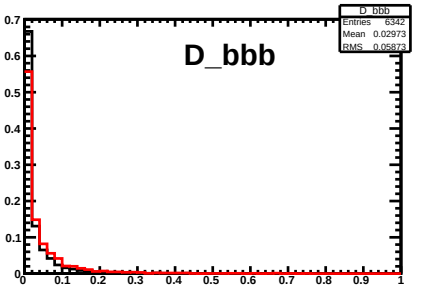


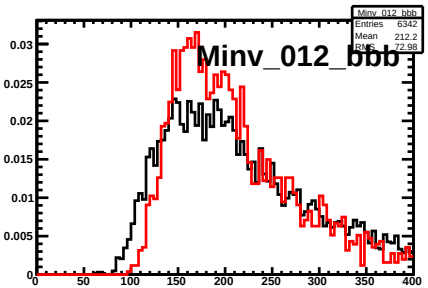


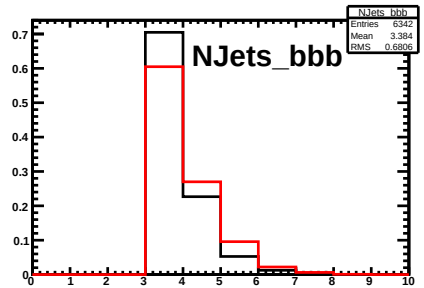


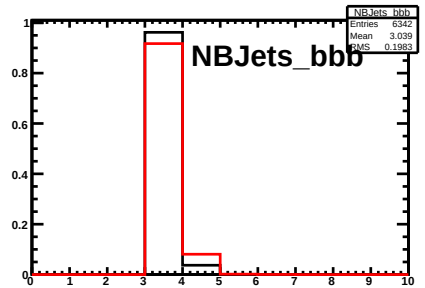




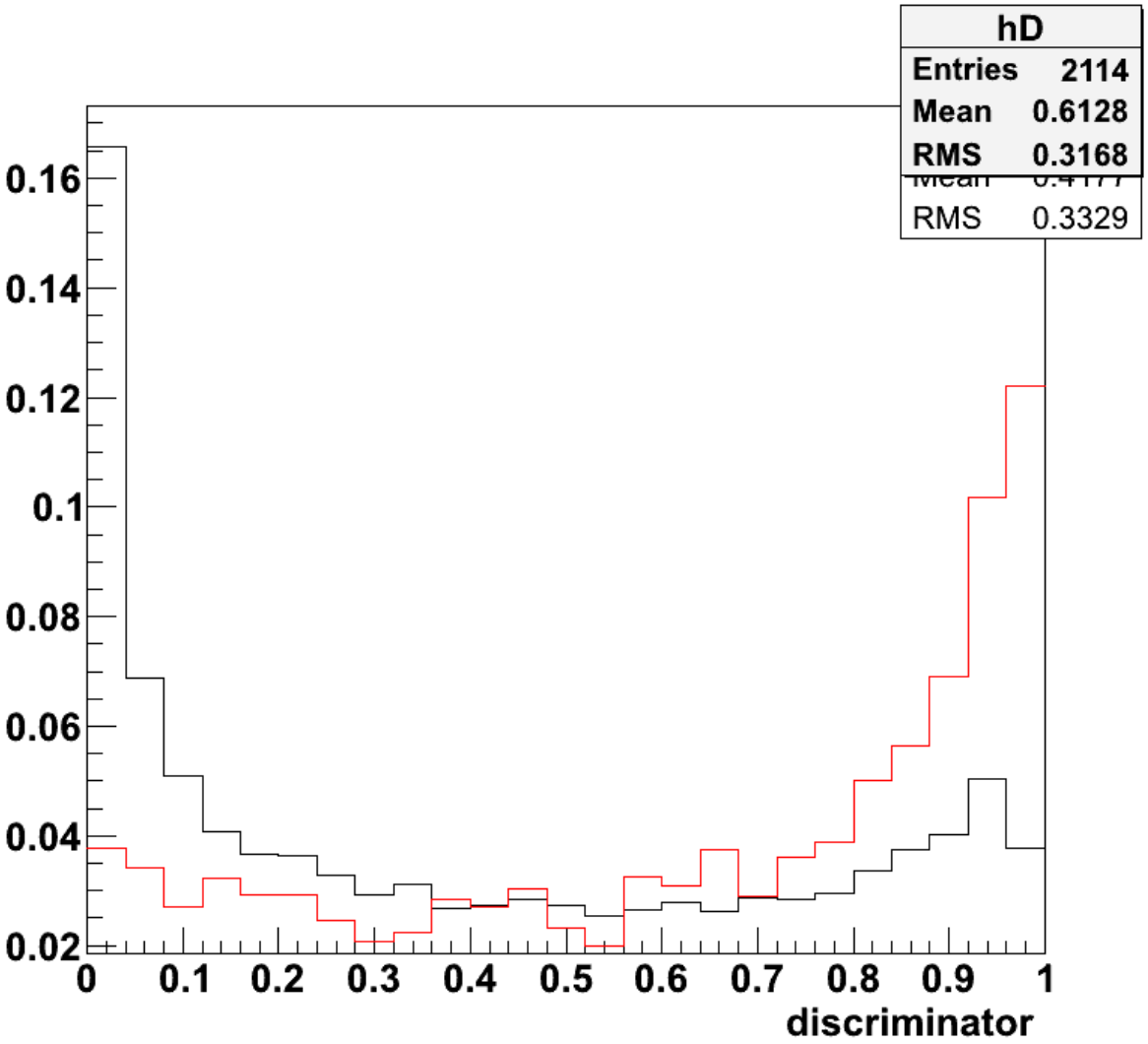








Discriminator



$$Discr = \frac{\Pi_i p_i^{(signal)(x_i)}}{\Pi_i p_i^{(signal)}(x_i) + \Pi_i p_i^{(QCD)}(x_i)}$$

Likelihood ratio build using the most discriminating variables from MC.

Still to be optimized

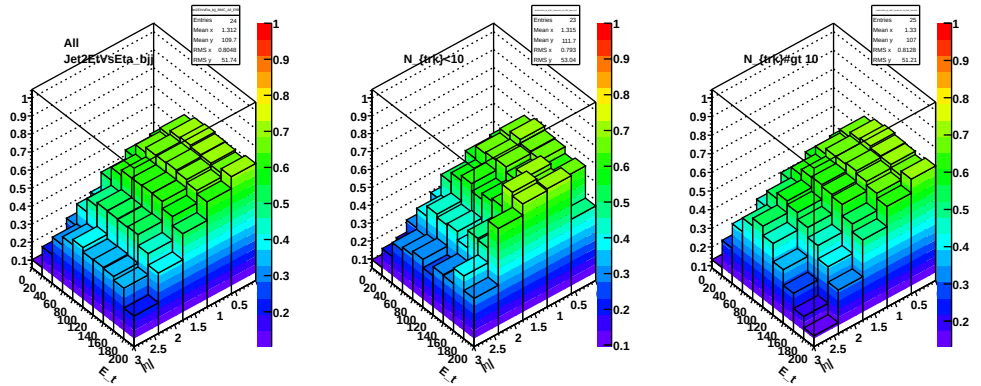
Can be used also as MVA if we can predict the variables from data *bbj* to *bbb*

B-matrices and co.

B-matrices and co.

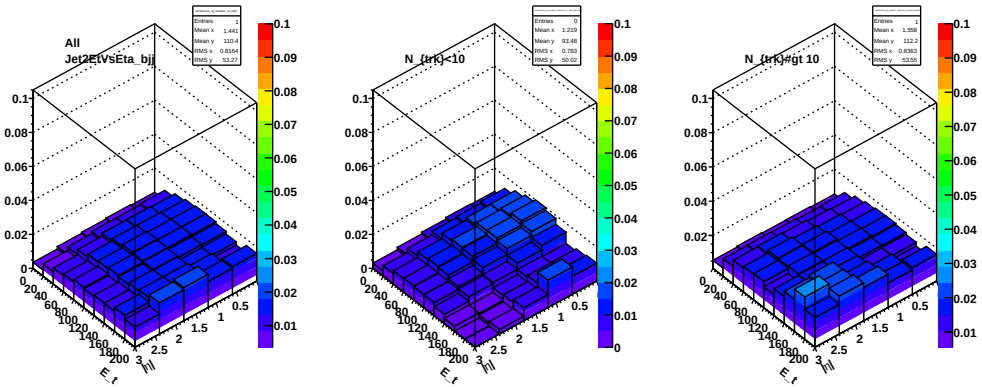
- Get f_b, f_c, ϵ_b, ϵ_c and ϵ_l from MC;
- Separately for:
  - *bjj* vs *bbj* (first jet b-tag, look at second one);
  - *jbj* vs *bbj* (second jet b-tag, look at first one);
  - *bbj* vs *bbb* (first two b-tags, look at third);
- Many different parametrization has been tried
  - vs E_t, $|\eta|$;
  - vs E_t, ΔR;
  - vs E_t, ΔR separating B, C and light;
  - vs E_t, ΔR and N_{trk};
  - vs E_t, $\Delta\eta$;
  - ...
- Will show only some example

ϵ_b, ϵ_q, f_b for bjj vs $|\eta|$, E_t vs N_{trk}



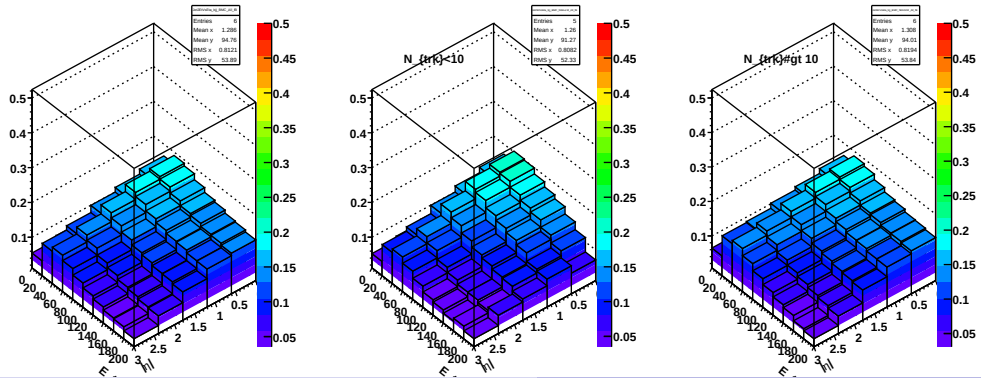
Left to Right:

ϵ_B All, $N_{trk} < 10$, $N_{Trk} \geq 10$



Left to Right:

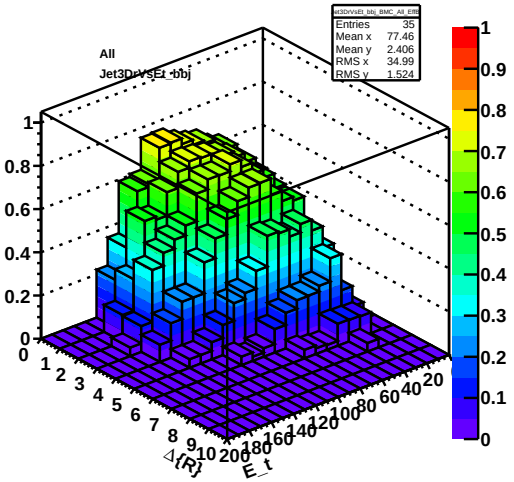
ϵ_Q All, $N_{trk} < 10$, $N_{Trk} \geq 10$

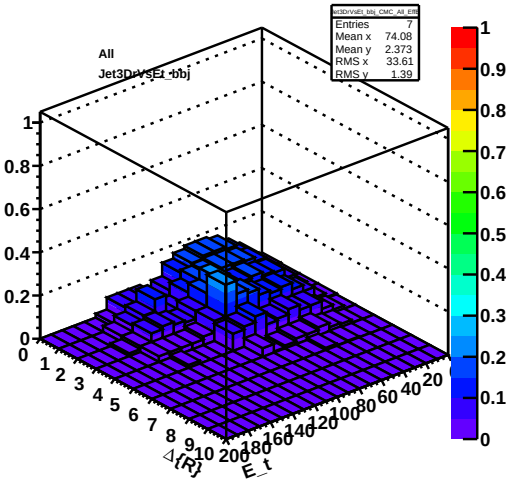


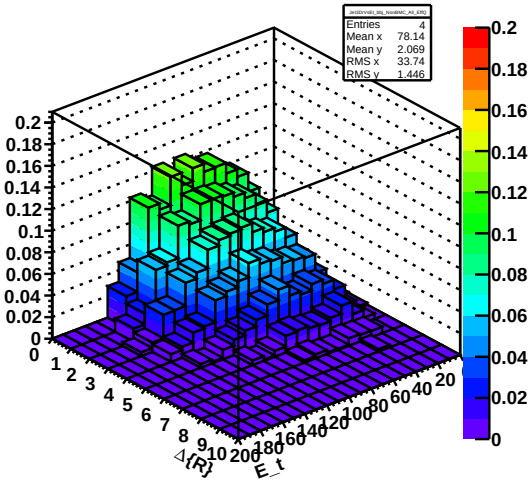
Left to Right:

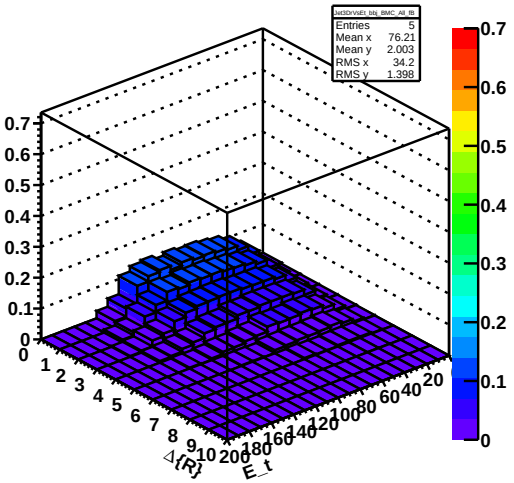
F_B All, $N_{trk} < 10$, $N_{Trk} \geq 10$

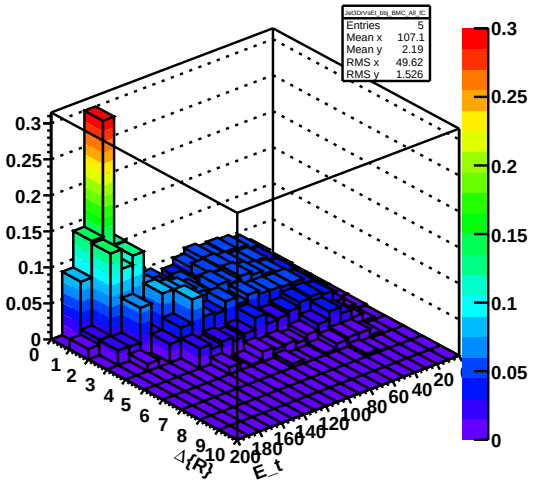
ϵ_b, ϵ_c, ϵ_l, $f_{b,c,l}$ for bbj vs $|\eta|, E_t$

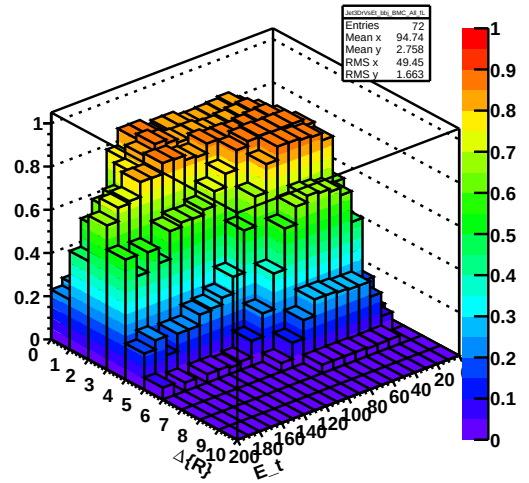




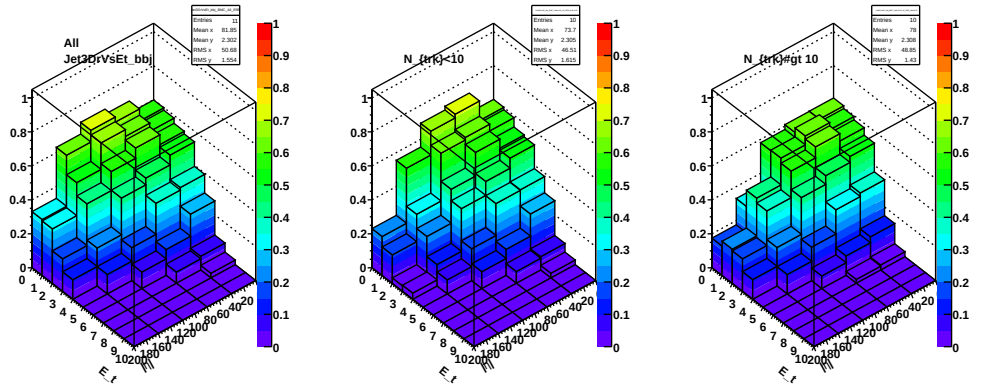






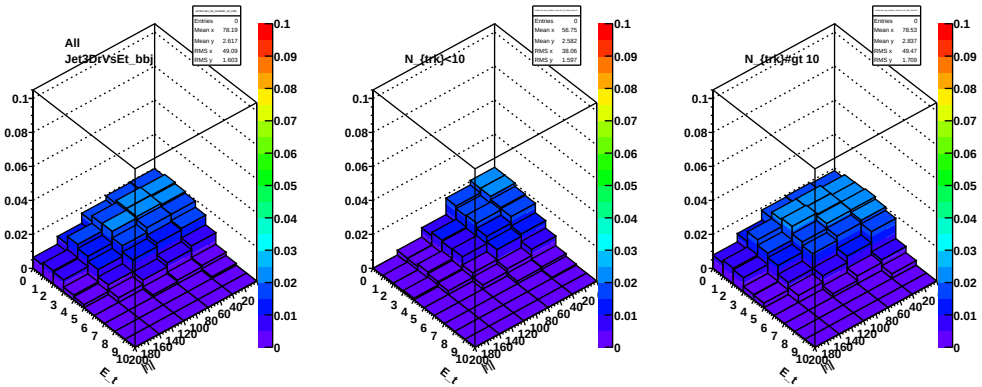


ϵ_b, ϵ_q, f_b for bbj vs $E_t, \Delta R$



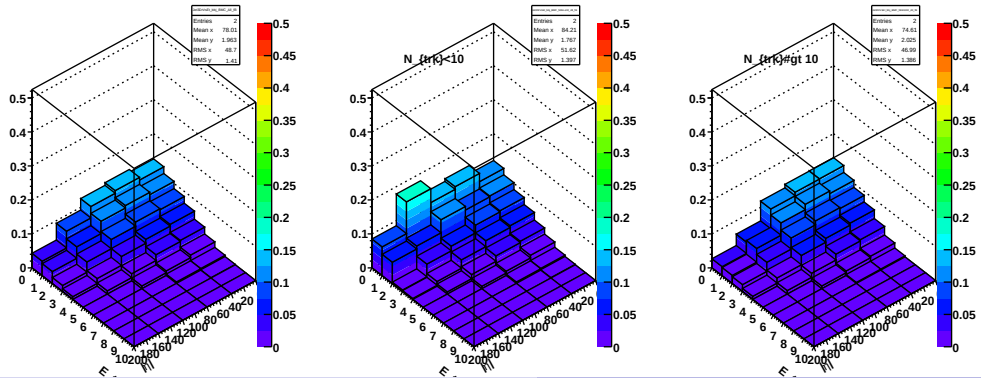
Left to Right:

ϵ_B All, $N_{trk} < 10$, $N_{Trk} \geq 10$



Left to Right:

ϵ_Q All, $N_{trk} < 10$, $N_{Trk} \geq 10$



Left to Right:

F_B All, $N_{trk} < 10$, $N_{Trk} \geq 10$

Check $\epsilon's$ on Top sample

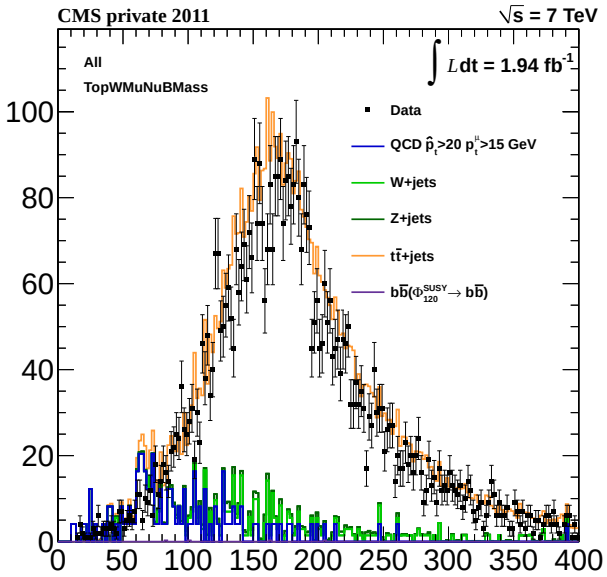
Check on Top

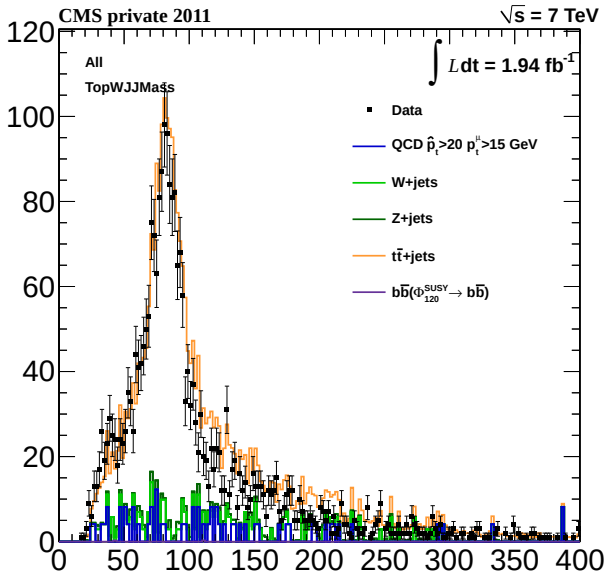
- On $t\bar{t}$ sample (MC), check ϵ_q and ϵ_b on $t \to bW$ candidates:
- Select a clean top sample, semileptnic decay $(tt \to (b\mu\nu)(bjj))$
  - Use $W \to jj$ to check mistag probability on light and charm quark
  - Use b for b-tagging efficiency:
  - Already done: `CMS PAS BTV-11-003` and corresponding `AN`
  - For CSV, Data and MC b-tagging efficiency agrees within $\sim 5\%$ or better;

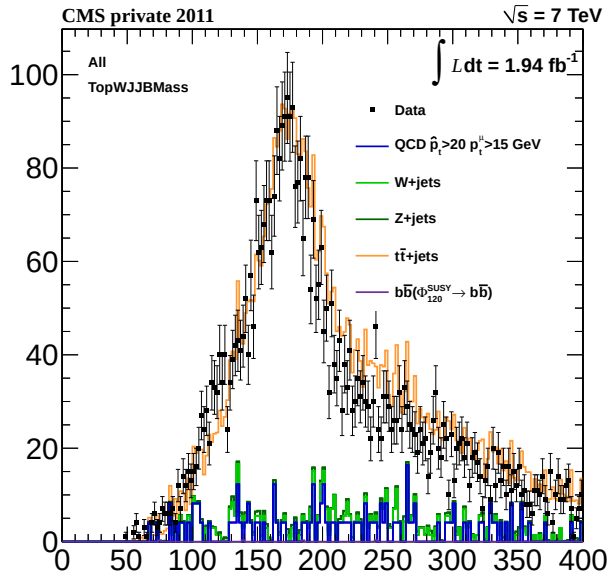
Top Selection

- Good, isolated muon $p_t > 25\ GeV$ $(\Sigma p_t^{trk} + \Sigma E_t^{em} + \Sigma E_t^{had})/p_t^{\mu} < 0.1$
- exactly 4 jets $E_t > 25$ and $|\eta| < 2.6$ JetID loose, of which exactly 2 B-jets $(CSV > 0.8)$.
  - $W \to jj$ from the two non-b jets
  - $t \to bW \to bjj$ from $W \to jj + b$ (combinatorics not resolved, $\times 2$ candidates)
  - $t \to bW \to b\mu\nu$ from μ and MET.

Mass of: $t \to b\mu\nu$, $W \to jj$, $t \to bjj$ (Run2011A)



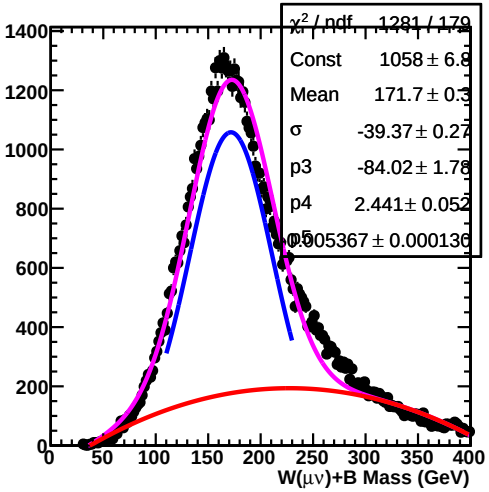


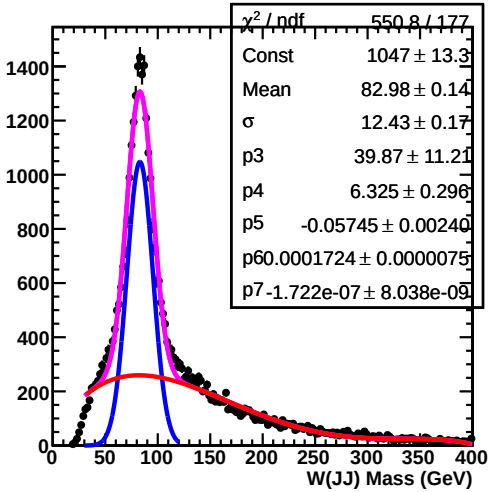


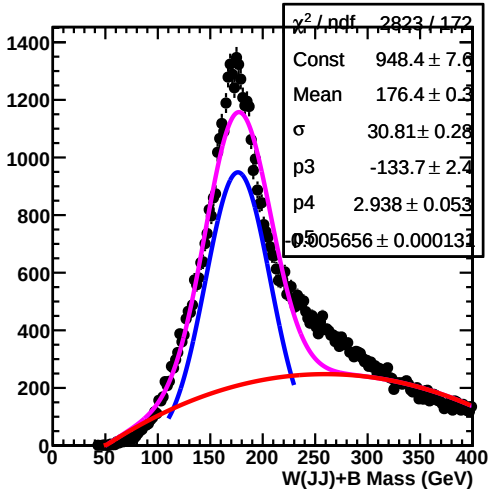
Looking at $t\bar{t}$ yield in data: only RunA

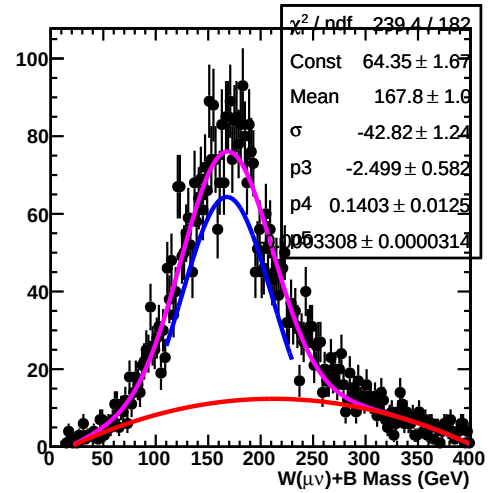
Need to select jets w/o any b-tag requirement to check b-tagging efficiency

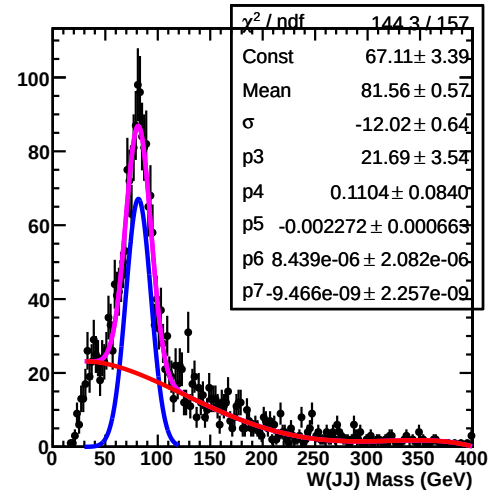
Mass of: $t \to b\mu\nu$, $W \to jj$, $t \to bjj$ (MC/RunA)

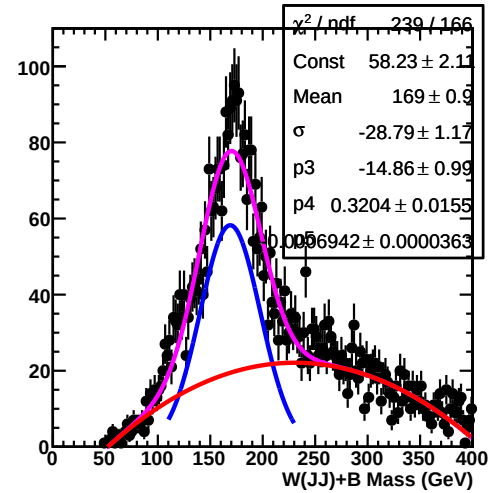




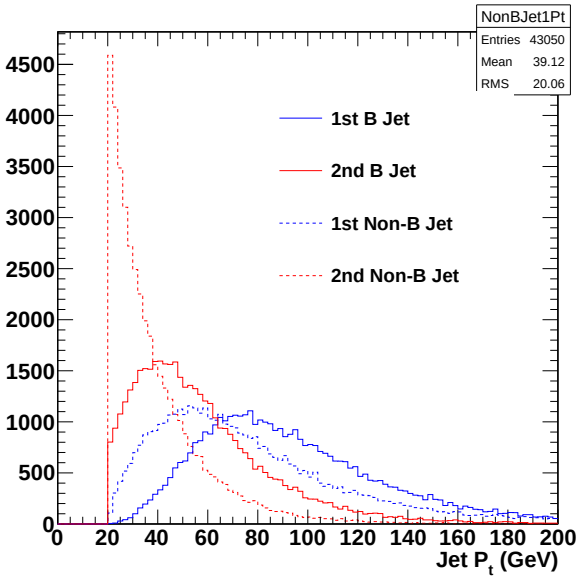




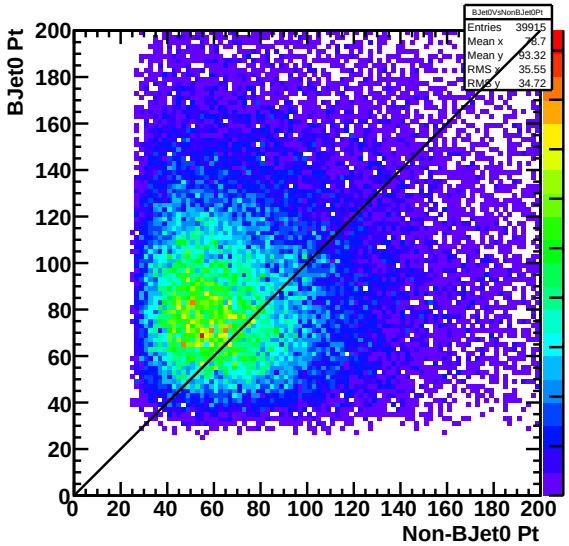


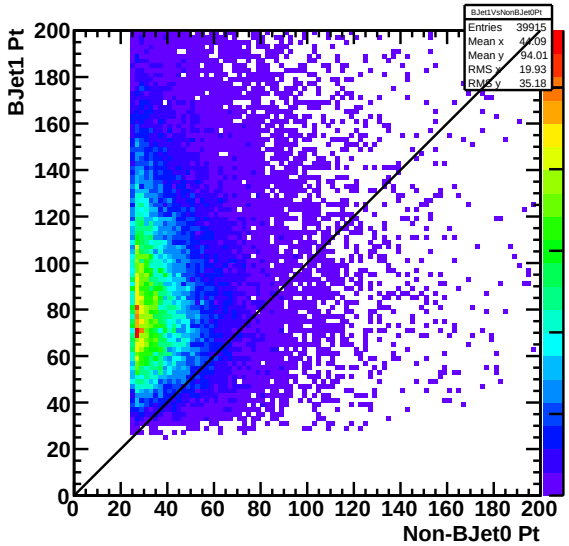


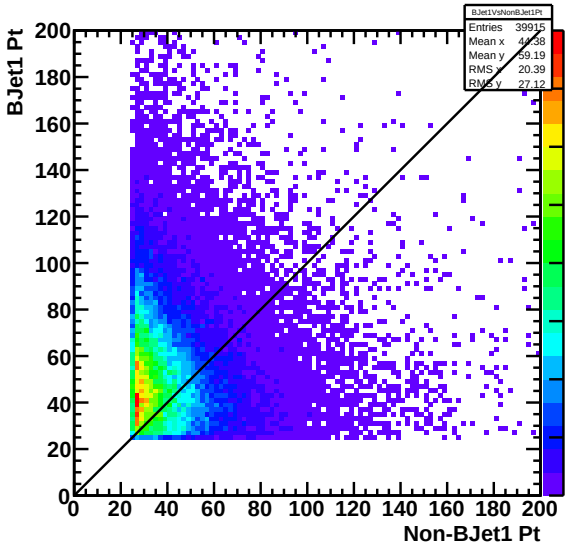
P_t of first/second B-jet/Non-B-jet



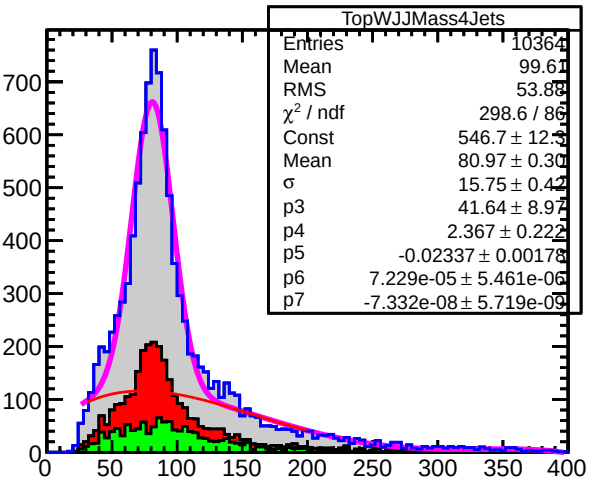
- Typically p_t ordering is the following: $Bjet^{(1)} > Jet^{(1)} > Bjet^{(2)}$
- Look at second and forth jet (sorted in E_t) as coming from $W \to jj$
- Can do better: will try simultaneoud t and W mass fit (as in PAS sec.9).

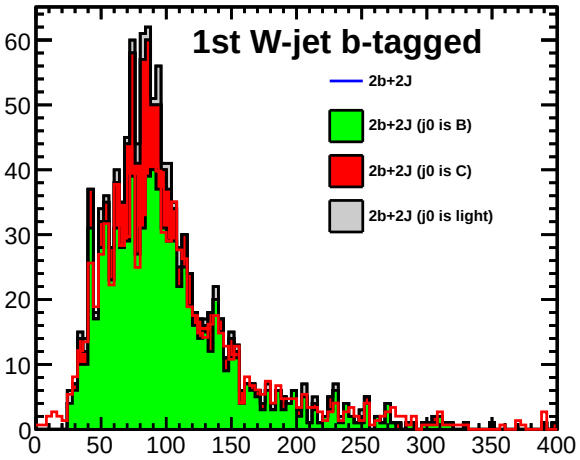


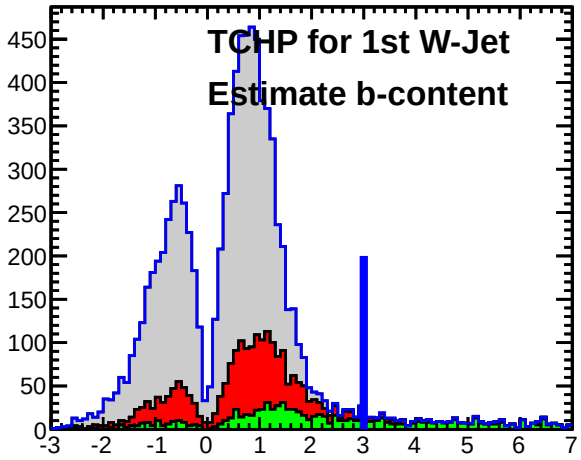


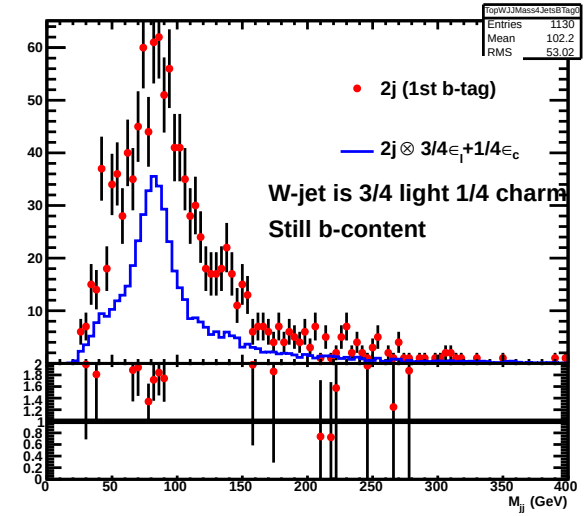


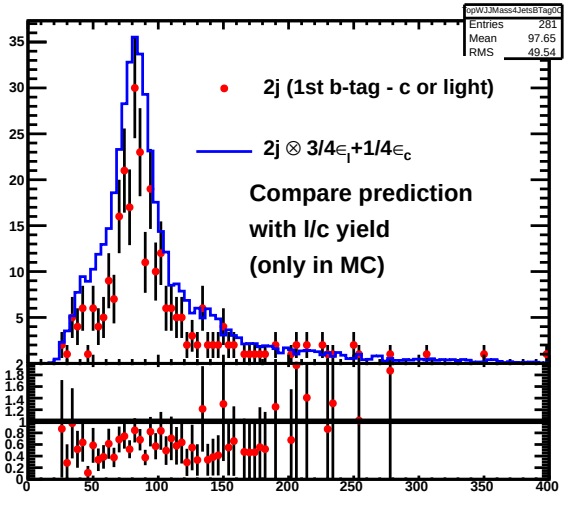
Test of ϵ_l/ϵ_c on W first jet (MC)

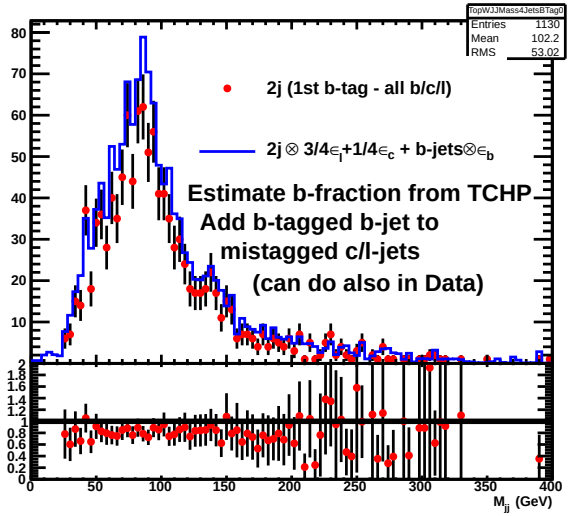












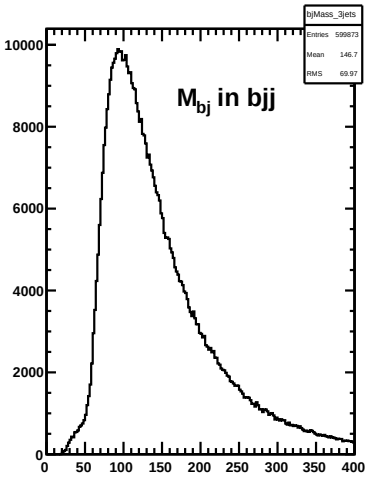
found a last minute bug in analogous data distributions

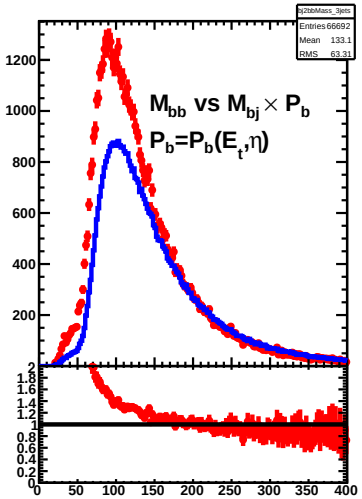
Check on QCD (MC and Data)

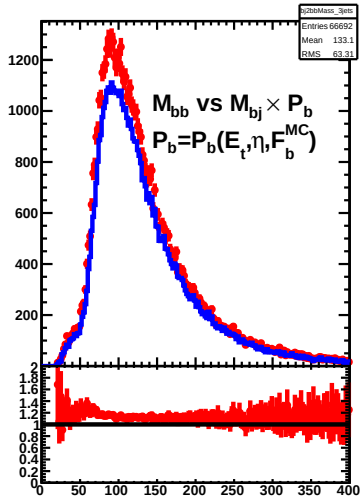
Check on multi-b sample (QCD MC)

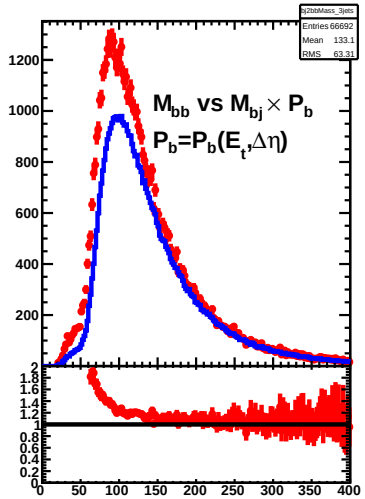
- On *QCD* sample (MC), check $P_b(j) = f_b\epsilon_b + (1. - f_b) \cdot \epsilon_q$ on *bjj* samples;
- compare M_{jj} distribution (two highest j) in sample *bbj* vs $bjj \otimes P_b(j^{2nd})$;
- likewise for *bbj* vs $jbj \otimes P_b(j^{1st})$;
- *bbb* vs $bbj \otimes P_b(j^{3rd})$;
- *bbb* vs $bbj \otimes P_b(j^{3rd})$;
  - Also in Data, only Control Region;

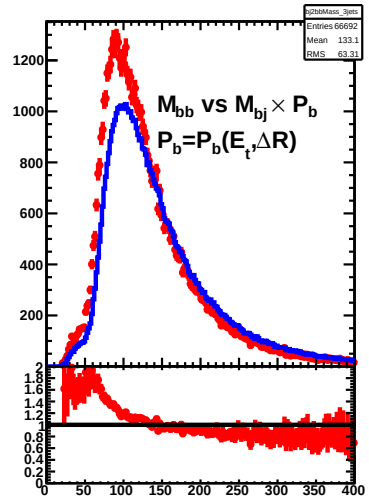
Test of $P_q(j)$ on *bjj* to *bbj* (MC)

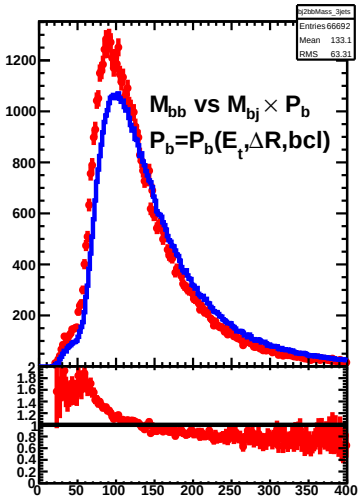


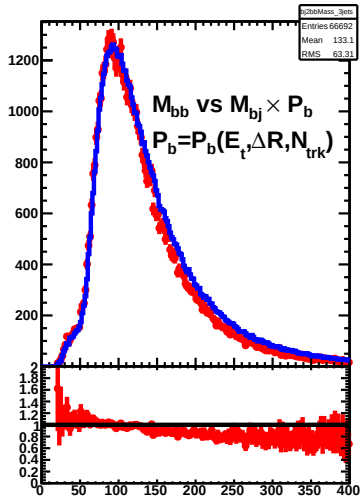




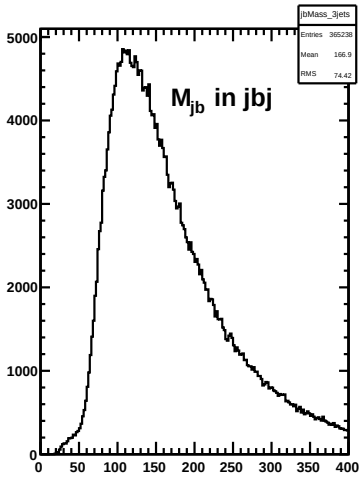


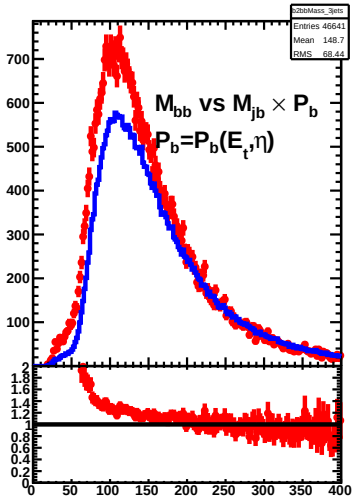


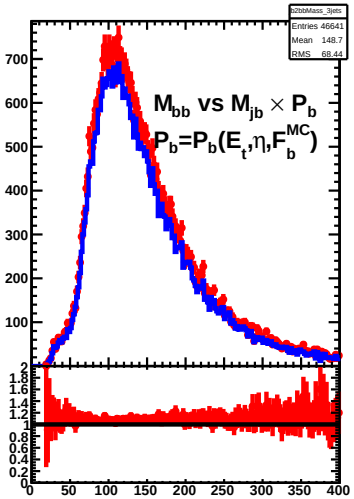


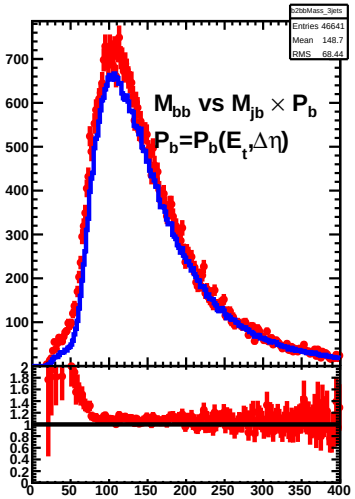


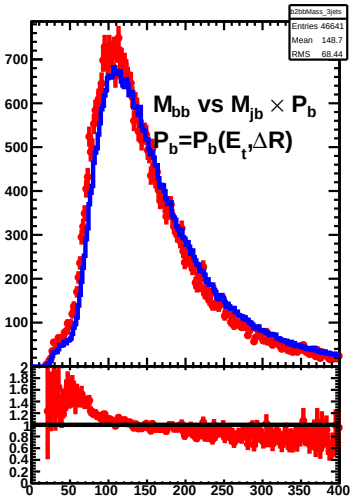
Test of $P_q(j)$ on *jbj* to *bbj* (MC)

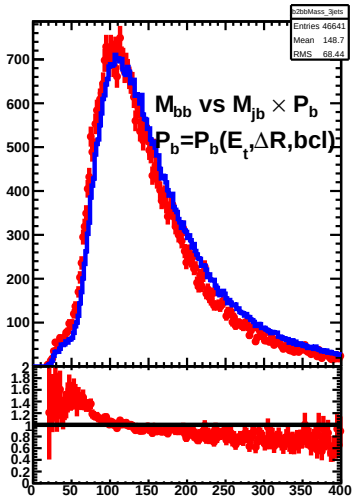


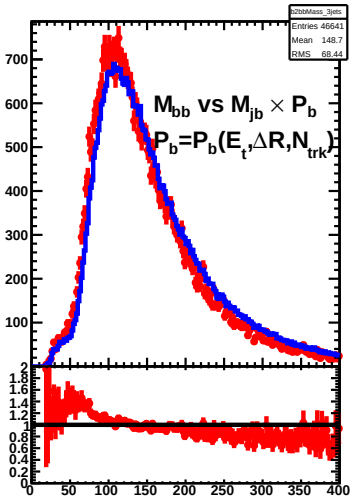




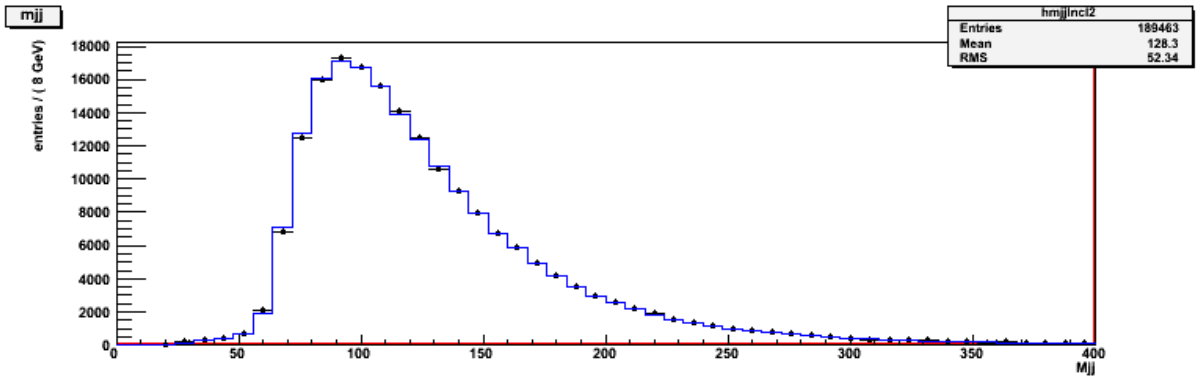


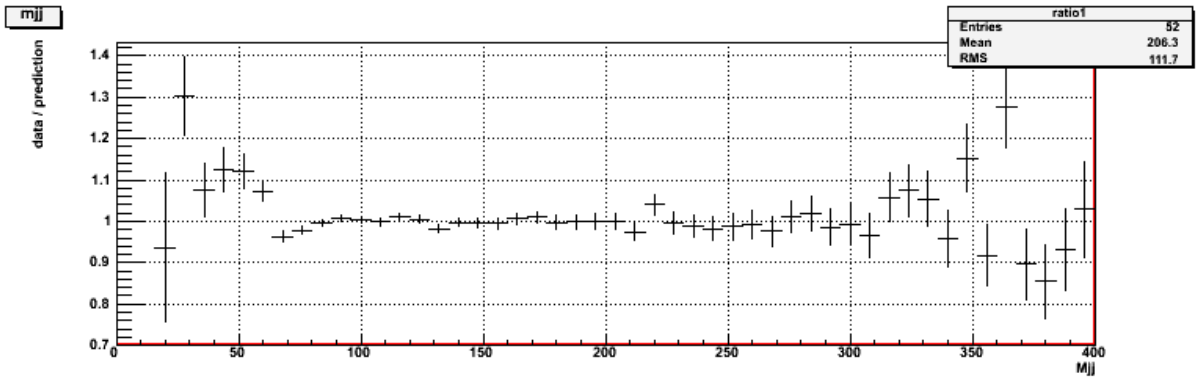


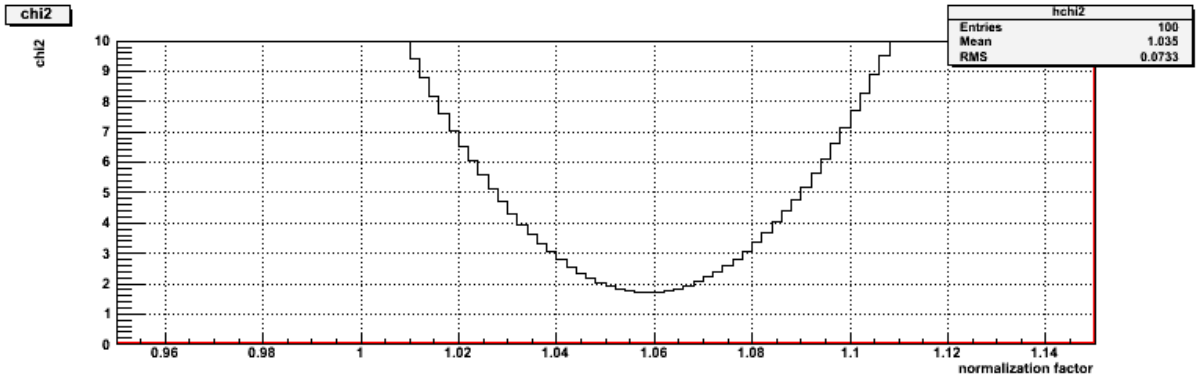




Prediction *bjj* to *bbj* (Data)



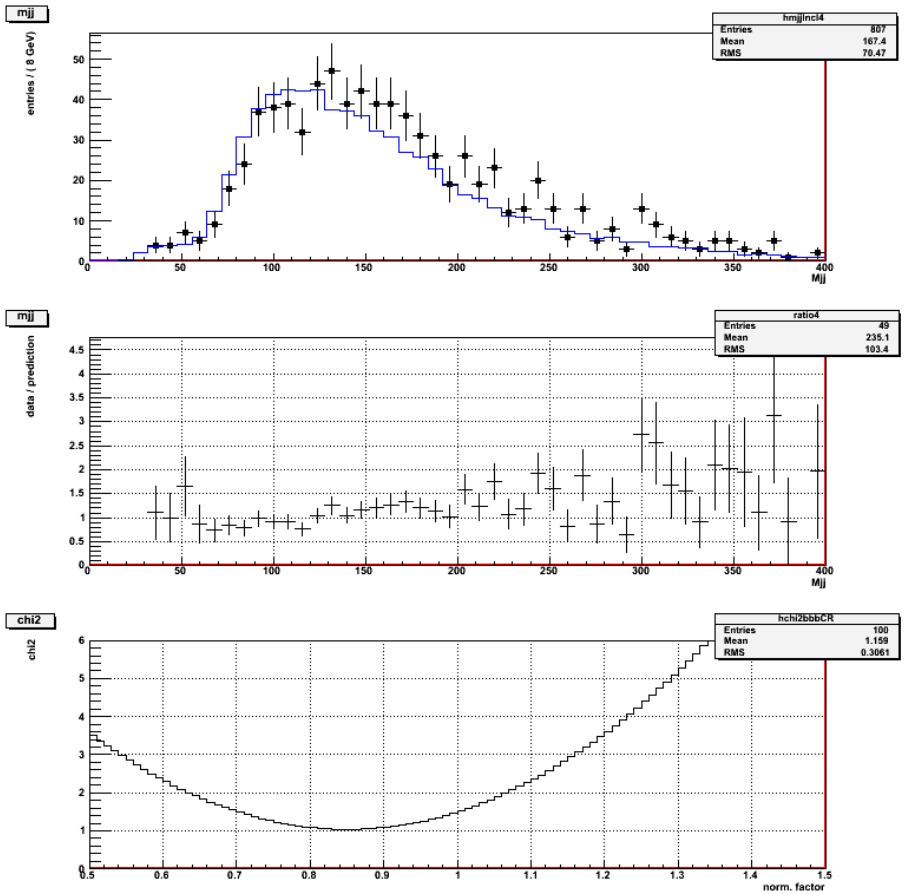




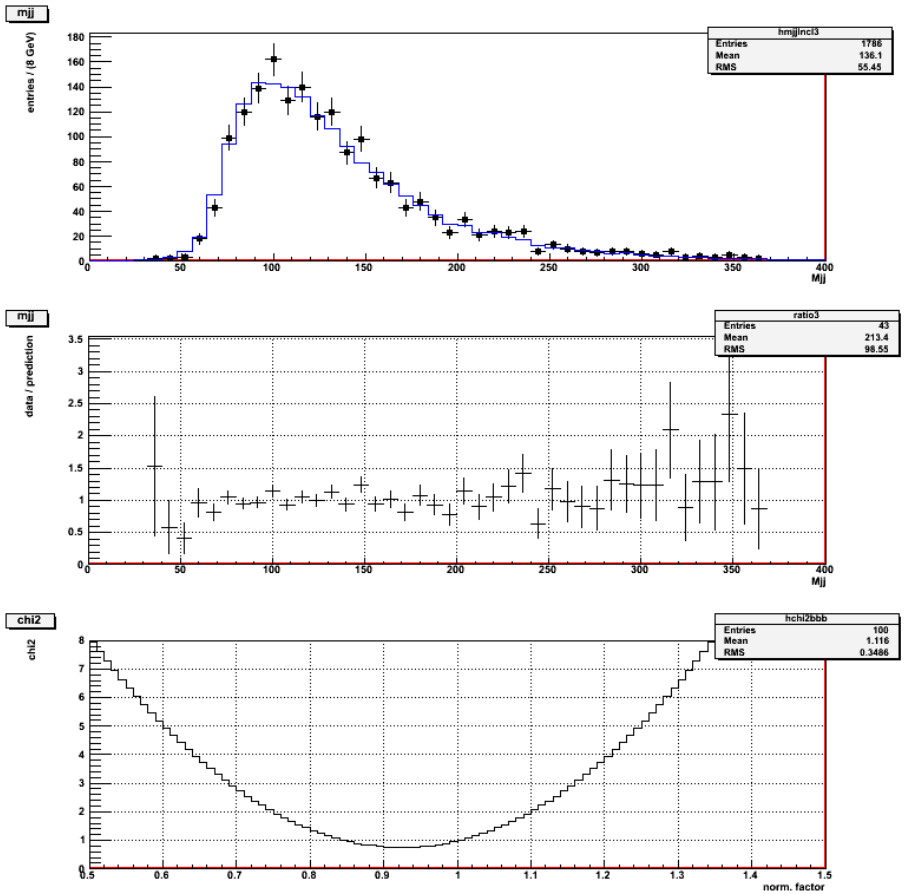
Also the normalization is good.
Best χ^2 for *ScaleFactor* = 1.06

Prediction *bbj* to *bbb* (MC)

Control Region



Signal Region

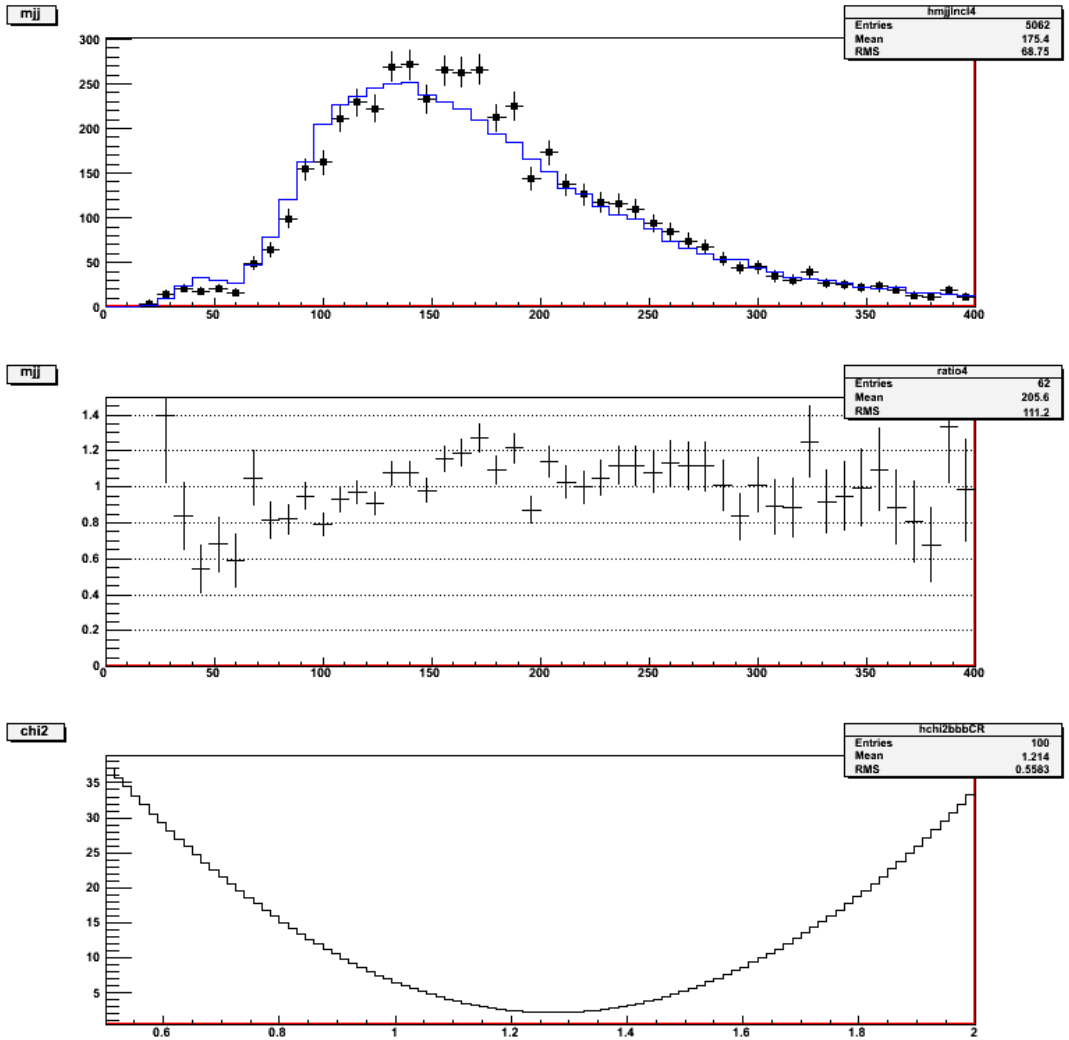


Still VERY preliminary Optimization of parametrization is still underway

Signal region looks better than control one ...

Prediction *bbj* to *bbb* (Data)

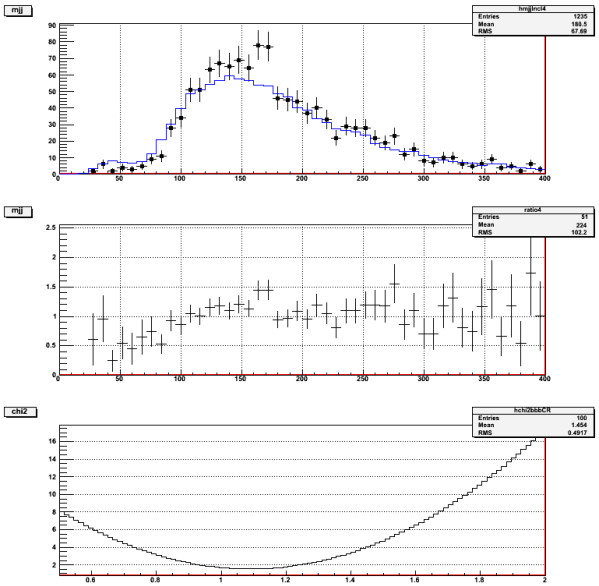
- Control Region Only
- Reminder f_B still from MC
- ScaleFactor ~ 1.3
- VERY PRELIMINARY!



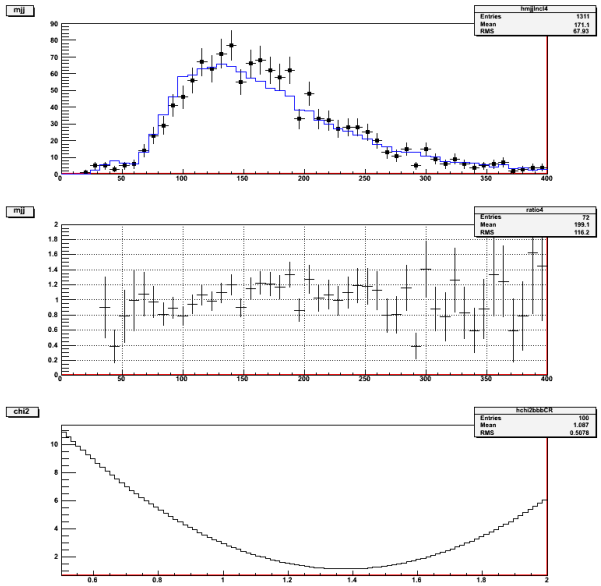
Prediction *bbj* to *bbb* (Data)

Control Region for three different HLT Path 1b, 2b, 2b_eta2.1

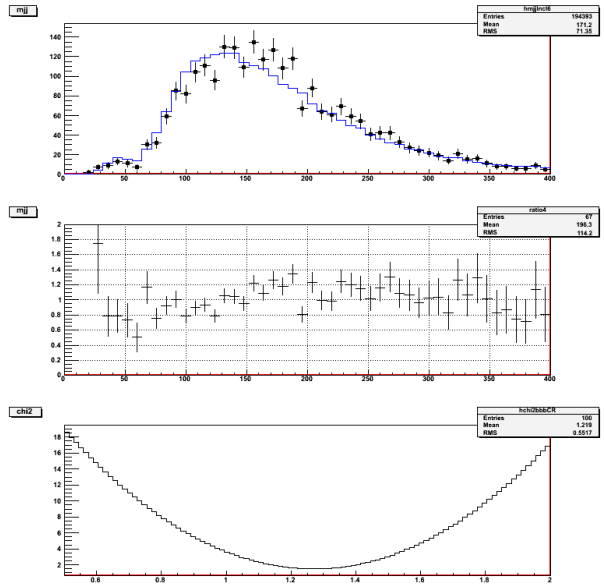
Mu12
DiCentralJet30
BTagIP3D



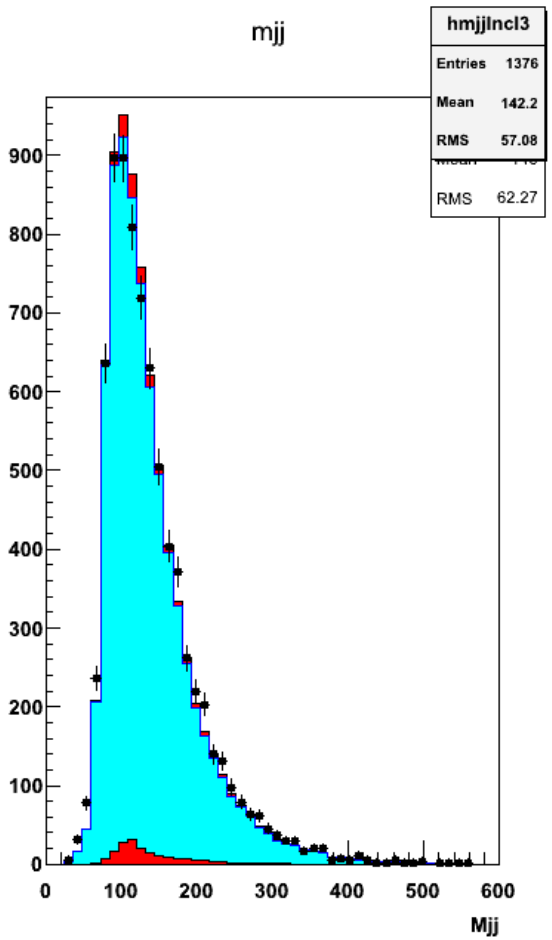
Mu12
DiCentralJet20
DiBTagIP3D1stTrack

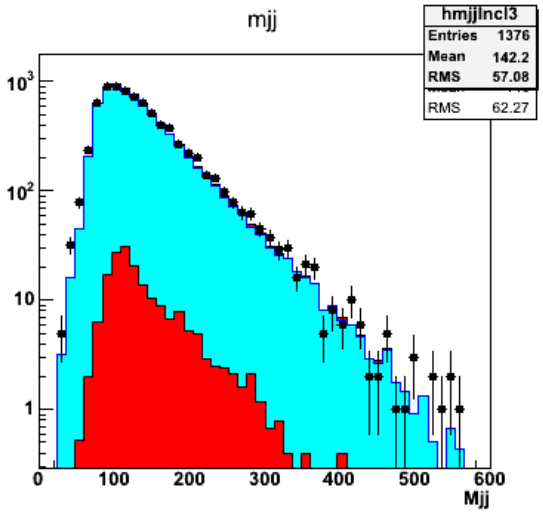


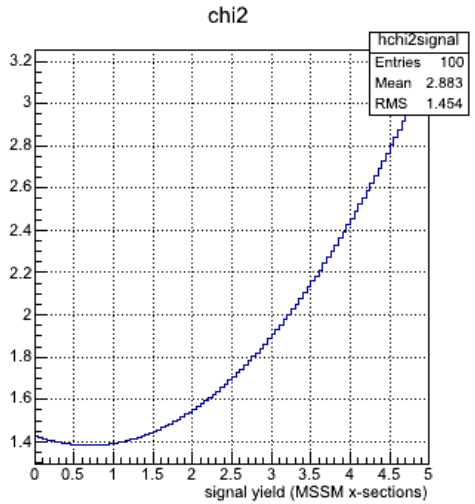
Mu12eta2p1
DiCentralJet20
DiBTagIP3D1stTrack



Expected sensitivity $M_H = 120\ GeV$

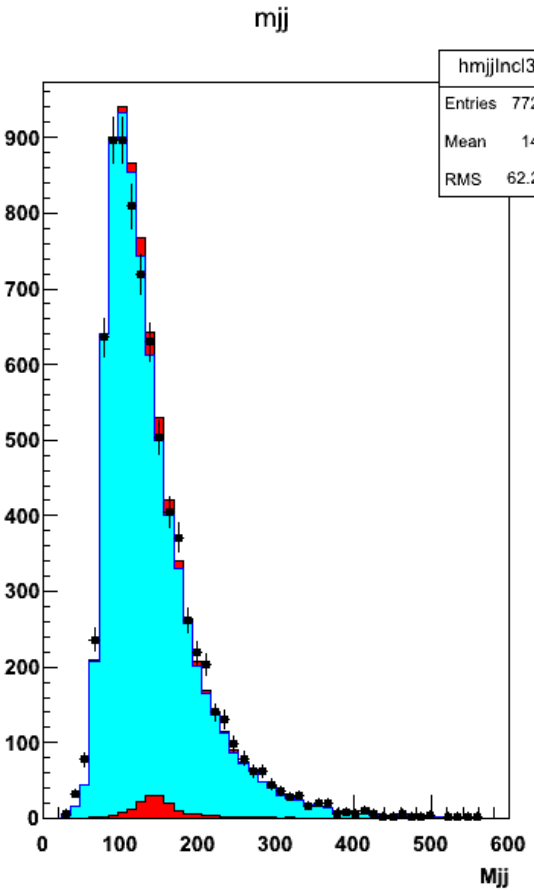


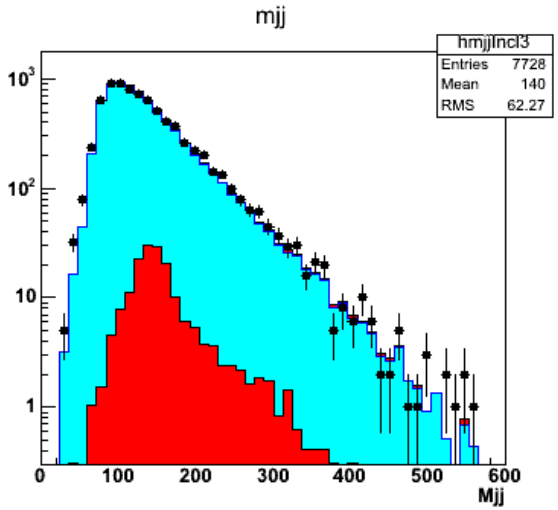


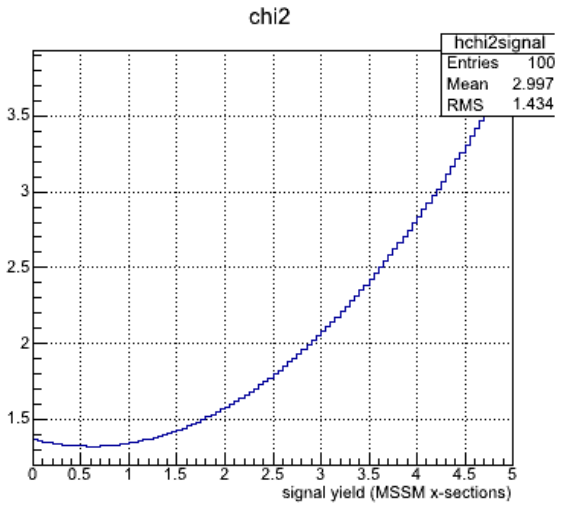


- MSSM: $M_H = 120\ GeV$, $\tan\beta = 30$;
- Sensitivity for $\sim 0.5\ fb^{-1}$ (available MC);
- QCD yield normalized to Data (control region);
- Only statistical errors;
- Can exclude up to $\lesssim 4 \times \sigma_{MSSM,\tan\beta=30}$
- Actual $\int \mathcal{L} dt$ collected in 2011 $\sim 4.4\ fb^{-1}$;
- Only M_{bb}, no MVA analysis (yet);

Expected sensitivity $M_H = 160\ GeV$







Conclusion

- To Be Done
  - Turn-on curves for Trigger;
  - Get f_b from data;
  - Check $\epsilon_{c,l}$ from $W \to jj$ on Data and improve methodology;
  - Optimize control/signal region;
  - Detailed comparison (predicted vs measured) of kinematical variables in control region
  - Try MVA analysis with predicted spectra;
  - Sensitivity including systematics;
  - ...
- Analysis Note in progress
- Will try to have updates for CMS week