

# Extracting B-tag distribution from MC to data-driven!

A. Jafari

# Outline

- **Event Selection**
- **A review on  $\chi^2$** 
  - **Efficiency and purity study**
- **The analysis approach**
- **Studying the method from fully based on MC to fully data driven**

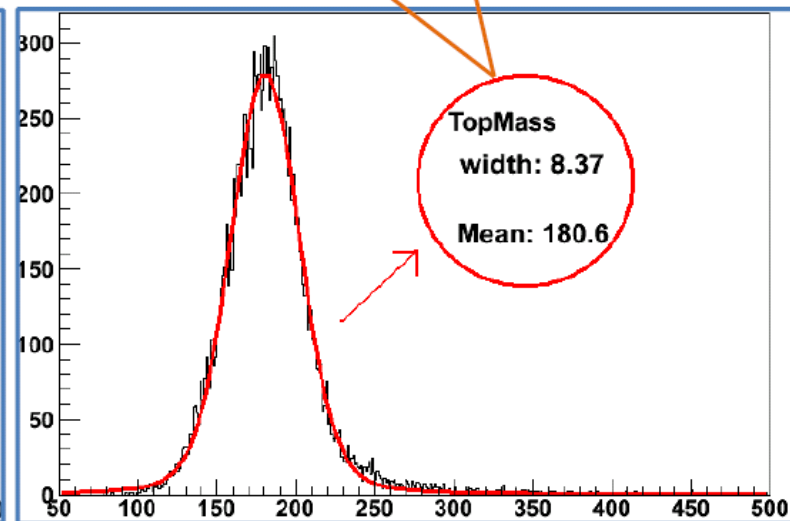
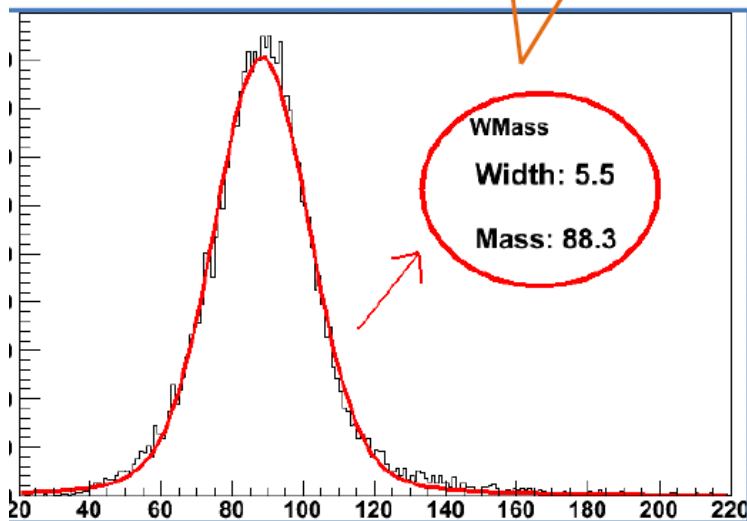
# Event selection

- Events are selected according to what we had discussed in the meeting except here there is no request for “*at least on b-jet*”.
- Passing the HLT of having a non-Isolated electron.
- At least one electron,  $|\eta| < 2.4$  (gap is excluded),  $P_t > 20$  GeV, isolated, identified, with  $d_0 < 200$   $\mu\text{m}$ .
- Exactly one electron with the criteria above.
- No second electron with  $|\eta| < 2.4$  (gap is excluded),  $P_t > 20$  GeV, and identified as robustLoose.
- No isolated muon with with  $|\eta| < 2.1$ ,  $P_t > 20$  GeV,  $d_0 < 200$   $\mu\text{m}$ ,  $\chi^2 < 10$ ,  $n\text{ValidHits} > 11$ .
- Jets are cleaned from electrons (Maryam’s Method)
- At least 4 jets with  $|\eta| < 2.4$ ,  $P_t > 25$  GeV,  $n\text{CaloTowers} > 5$ 
  - ❖ *These four leading jets are matched with partons*
  - ❖ *Software and release: CMSSW314*
  - ❖ *DataSet: summer09 ttbar Pythia, skimmed to semielectronic decay channel*

# Review on $\chi^2$ method

- Thanks to Maryam, I finalize the matching algorithm as precise as possible to measure the efficiency and the purity correctly
- I changed the definition of the purity. Now it only looks at the leptonic b-candidate instead of the hadronic top combination. It accepts the hadronic b-jets as matched b-jets, too.
- To select the leptonic b-candidate, we go through this procedure:
  - In each event, the 4 leading jets are taken.
  - All possible hadronic top combinations (12) are made.
  - The one with the lowest  $\chi^2$  value is nominated as the hadronic top.
  - The remaining jet is the leptonic b-candidate.

$$\chi^2 = \left( \frac{\text{rec}W.\text{Mass}() - W\text{Mass}}{\sigma_W} \right)^2 + \left( \frac{\text{recTop.Mass}() - \text{TopMass}}{\sigma_{\text{Top}}} \right)^2$$



# Signal and background sets

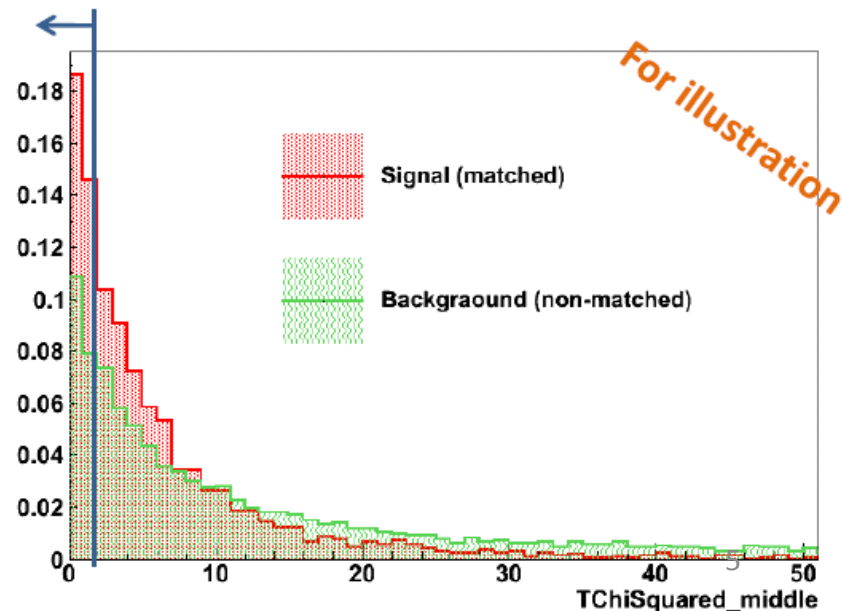
- To calculate the efficiency and the purity, I did as follows:
  - After selecting the top combination using the minimum  $\chi^2$  value, we have a set of jets, 3 for the hadronic top and one for the leptonic b.
  - The  $\chi^2$  value of the top candidate is attached to this set.
  - I define my **signal set** as the set whose leptonic b-candidate is matched with a b-jet. Therefore the **background sets** are those whose leptonic –candidate does not match.
  - I can put a cut on this minimum  $\chi^2$  value and see if my **signal set** pass the cut (**signal efficiency**):

$$\mathcal{E}_{signal} = \frac{\# SignalsPassedTheCut (leftOrange)}{\# AllSignals (totalOrange)}$$

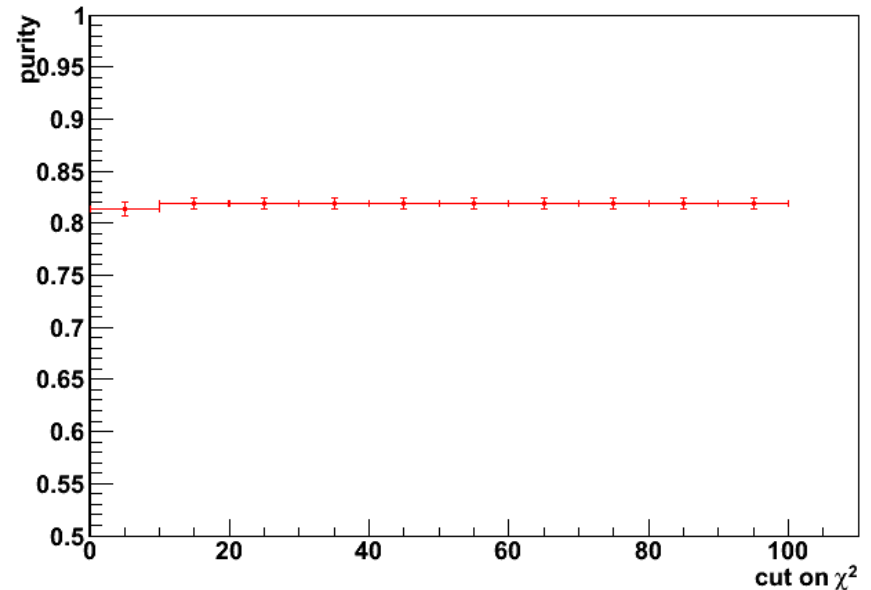
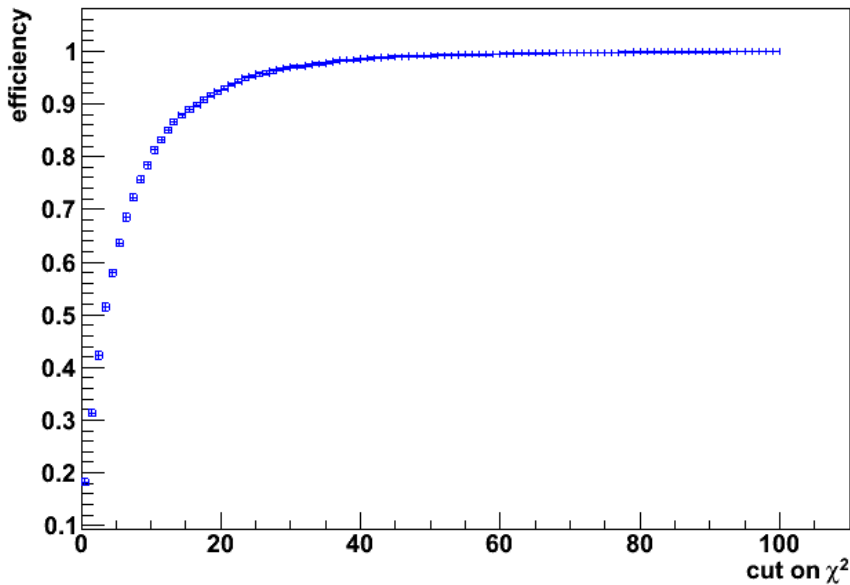
- The **purity of the selected sets** then is the number of **signal sets** in the left side, divided by the whole number of sets there:

$$P_{selectedSets} = \frac{\# SignalsPassedTheCut (leftOrange)}{\# AllSetsPassedTheCut (allLeft)}$$

- Also, I multiplied the signal efficiency and the purity and plotted this value vs. the minimum  $\chi^2$  value to find an optimal point for the cut on  $\chi^2$ .



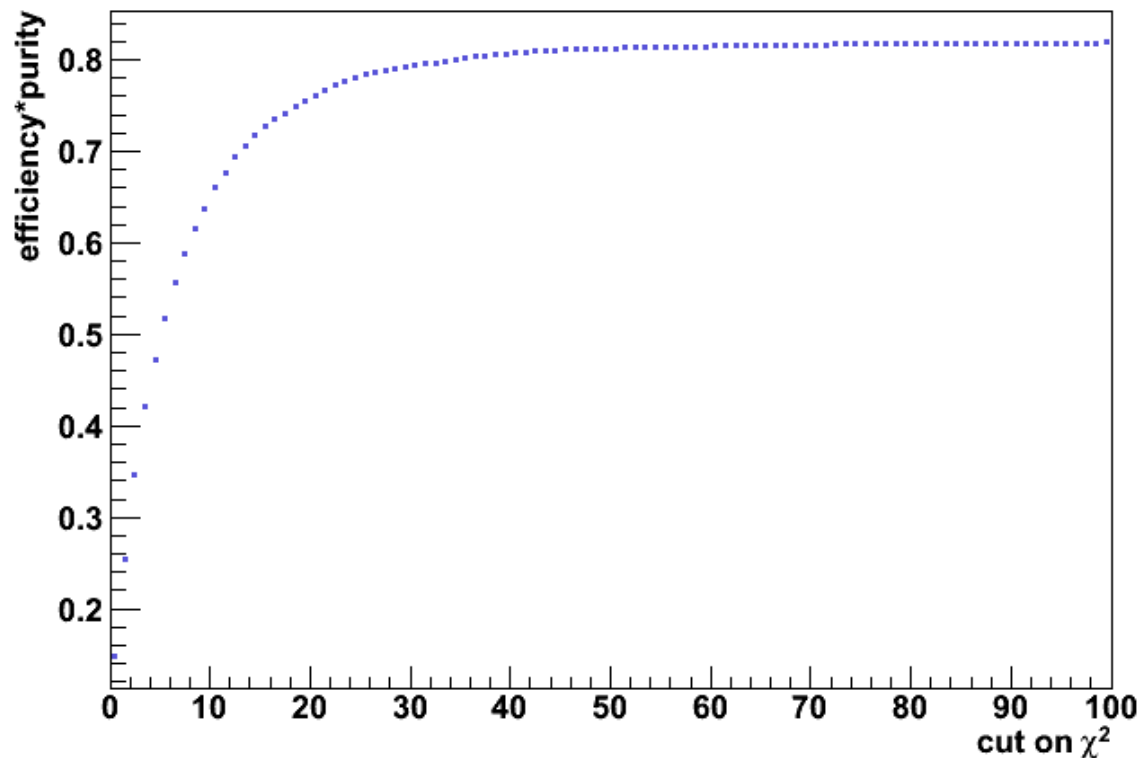
# Efficiency and purity



- As we worked with jets which are matched, by a random choice, we'll have a purity of about 50%. (both b-jets are accepted)
- By selecting sets with the minimum  $\chi^2$  value, it reaches to 80%
- The purity is kind of stable vs. the cut on  $\chi^2$ , because both signal and background sets have a very similar  $\chi^2$  distribution.
- The similarity is coming from the fact that we already chose the minimum value of  $\chi^2$  in each event. So,  $\chi^2$ 's are tending to smaller values in both signal and background sets.

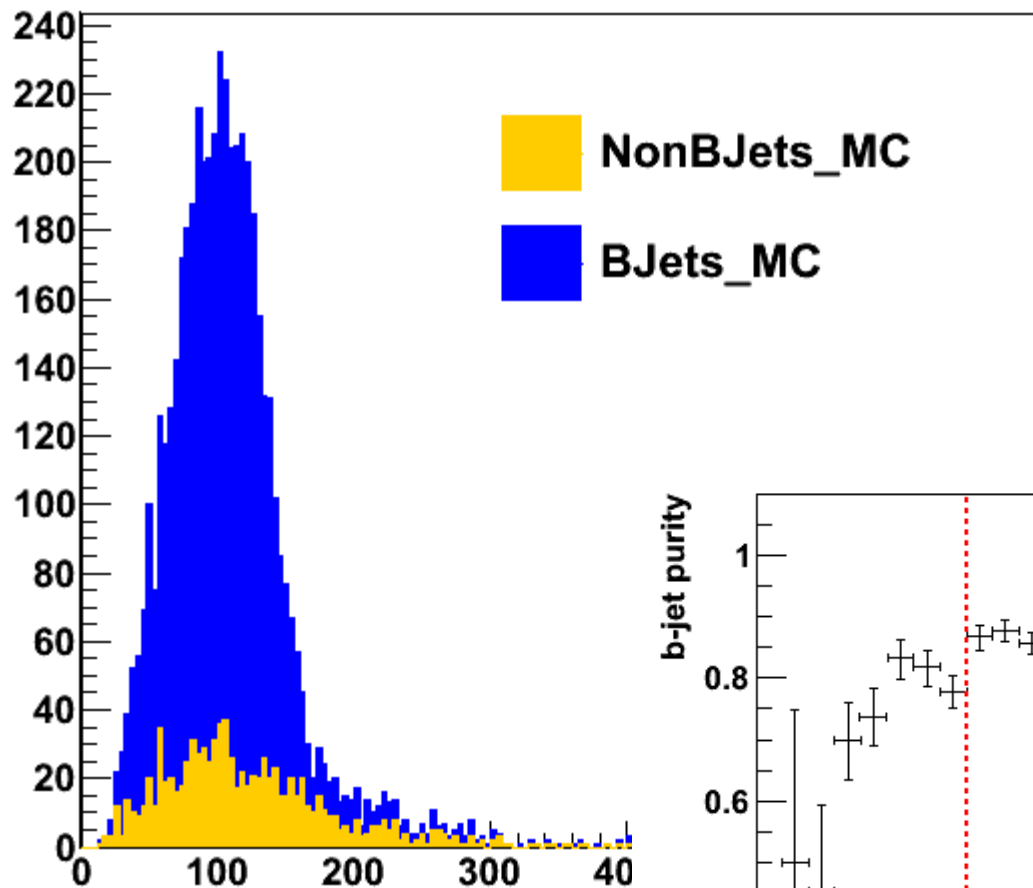
# The optimum cut on $\chi^2$

- Efficiency\*Purity is usually helping us to find an optimal cut for  $\chi^2$ .
- As the purity is stable, here Efficiency\*Purity, has a very similar shape to the efficiency itself.
- To avoid the systematic uncertainties, in the plot below I look for a region in which the Efficiency\*Purity is stable. This region is in high values of  $\chi^2$ .
- On the other hand, if we put a cut on smaller values of  $\chi^2$ , we'll gain no more purity in the price of the efficiency loss.
- Therefore, in the following study, no cut will apply on  $\chi^2$ . Note that we have already take the sets with the minimum value of  $\chi^2$ .



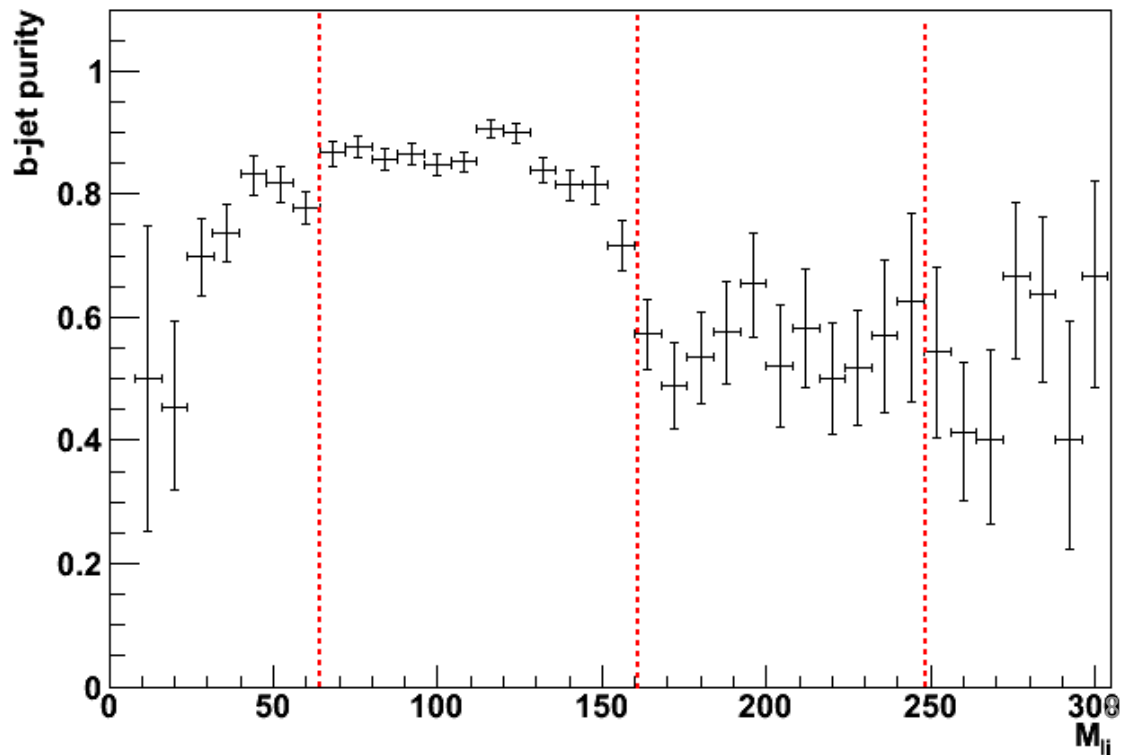
# Analysis approach, start with $M_{ij}$

Leptonic b-candidate from the  $\chi^2$  method:



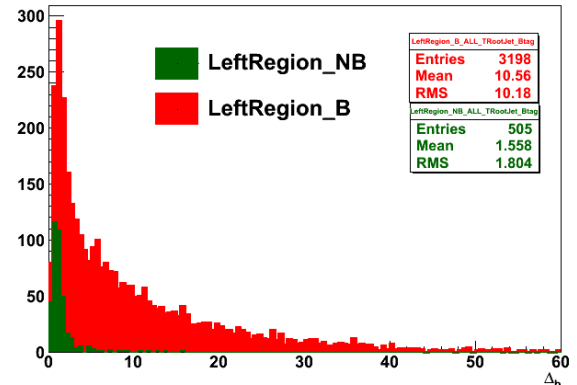
- In the peak area, the b-purity is about 80%.
- With no cut on the minimum  $\chi^2$  value, we expected a purity about 50%
- The increment around the peak shows that the right region is really dominated by the signal sets.

*B-tagging algorithm:*  
*TrackCountingHighEfficiency*



# Analysis approach: How to estimate the b-tag efficiency

$$\Delta_L^{lep bCand} = \Delta_L^{total} - \Delta_L^{lightJets}$$



How to estimate  $\Delta_L$  in the left region without using the MC information of the left region?  
Assumption\*:

No correlation between  $\Delta$  and  $M_{lj}$

→ the shape of  $\Delta^{lightJets}$  is similar in right and left region

→ for light jets,  $\Delta_L \approx F \cdot \Delta_R$  where  $F = N_L / N_R$

$$\Delta_L^{lep bCand} = \Delta_L^{total} - F \cdot \Delta_R^{lightJets}$$

F can be obtained

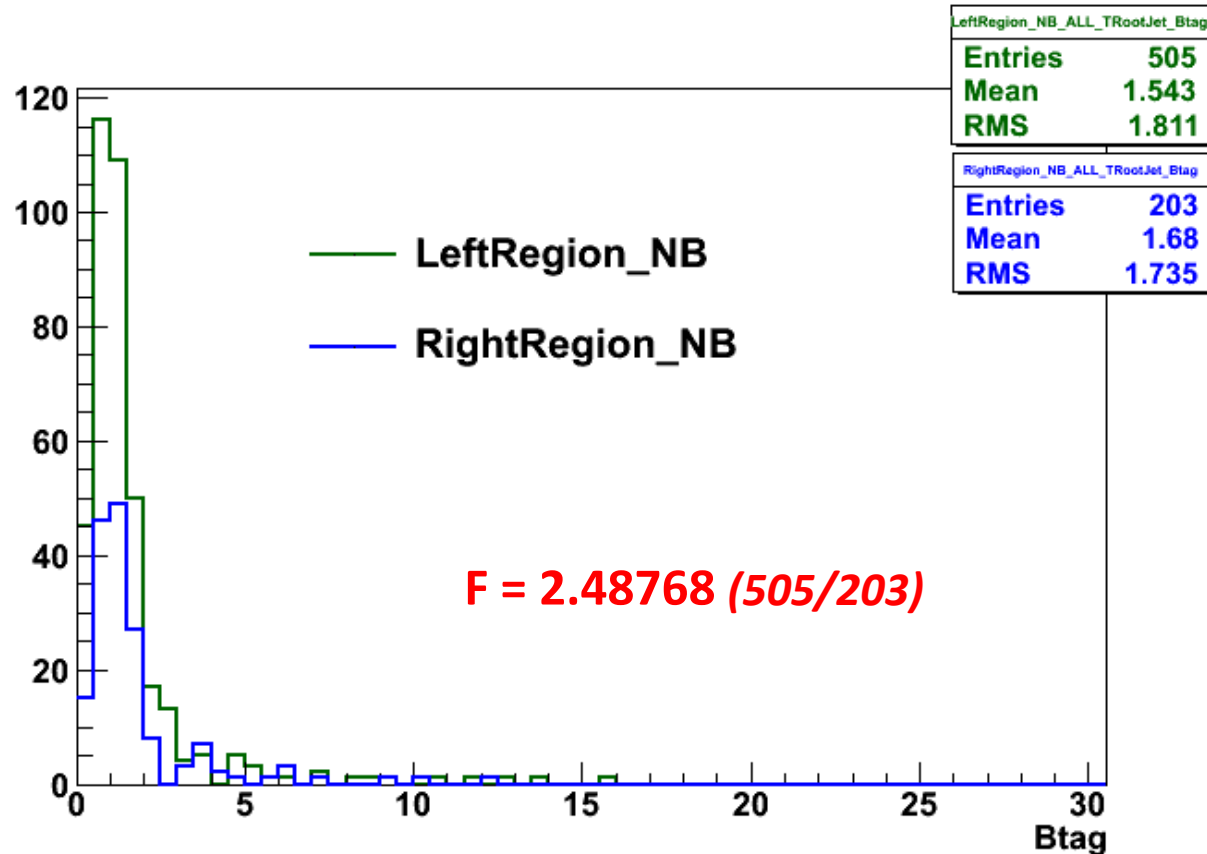
- From MC i.e. counting light jets in both side, using MC truth
- From data, using the control sample

For  $\Delta_R$  of light jets,

- It is trivial from MC
- It can be estimated as the  $\Delta$  of all jets in right region because this region is dominated by light jets

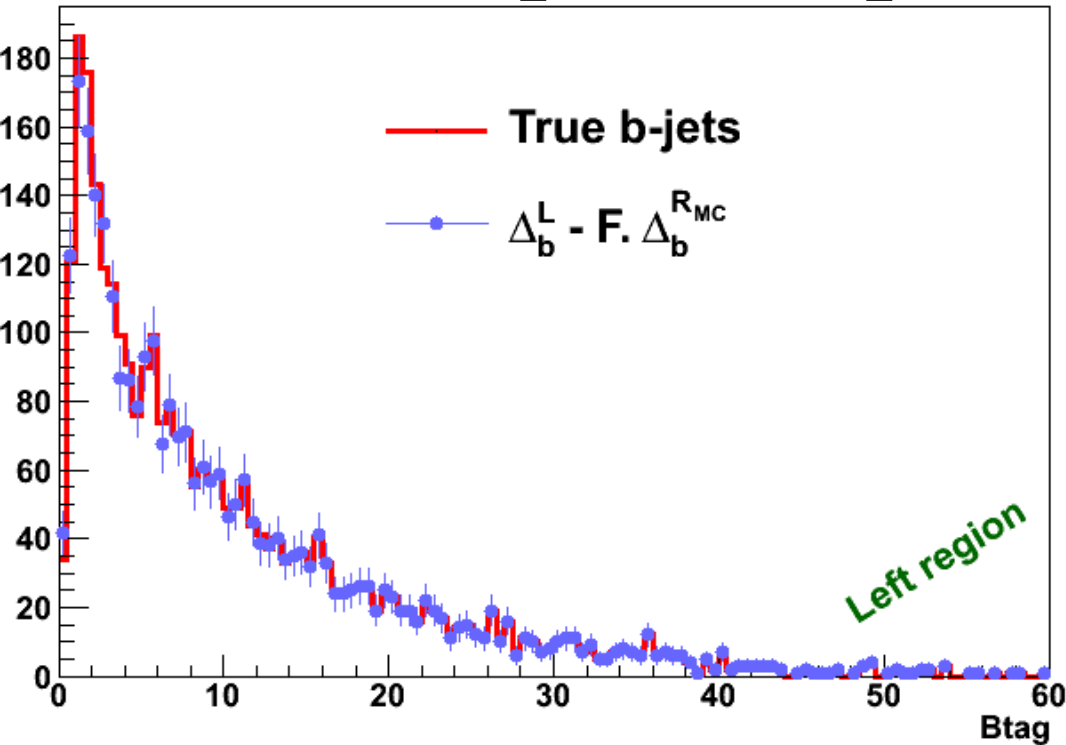
\* We will come back to this assumption later.

# First Try: MC based study, F from MC truth, Light jets in right region are matched with light quarks

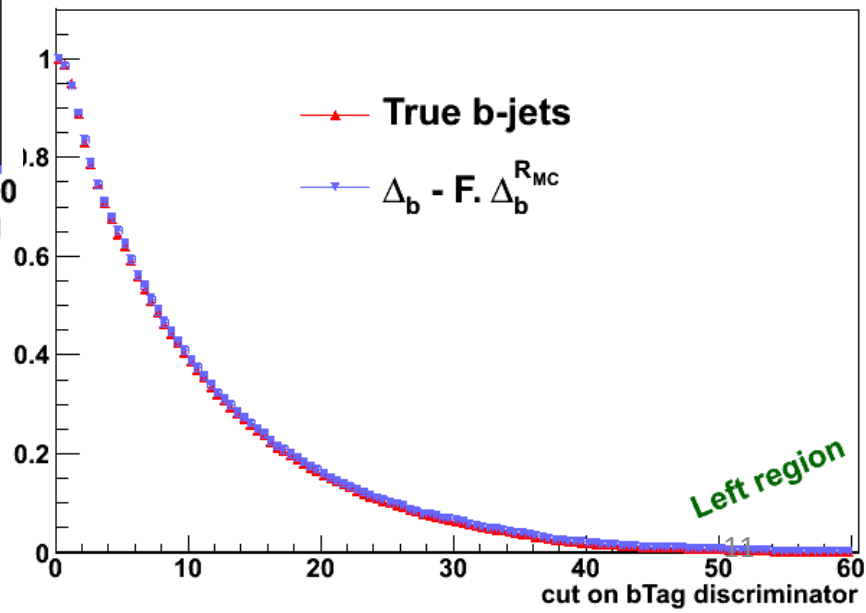


# First Try: MC based study, F from MC truth, Light jets in right region are matched with light quarks

$$\Delta_L^{lep b Cand} = \Delta_L^{total} - F \cdot \Delta_R^{light Jets MC}$$

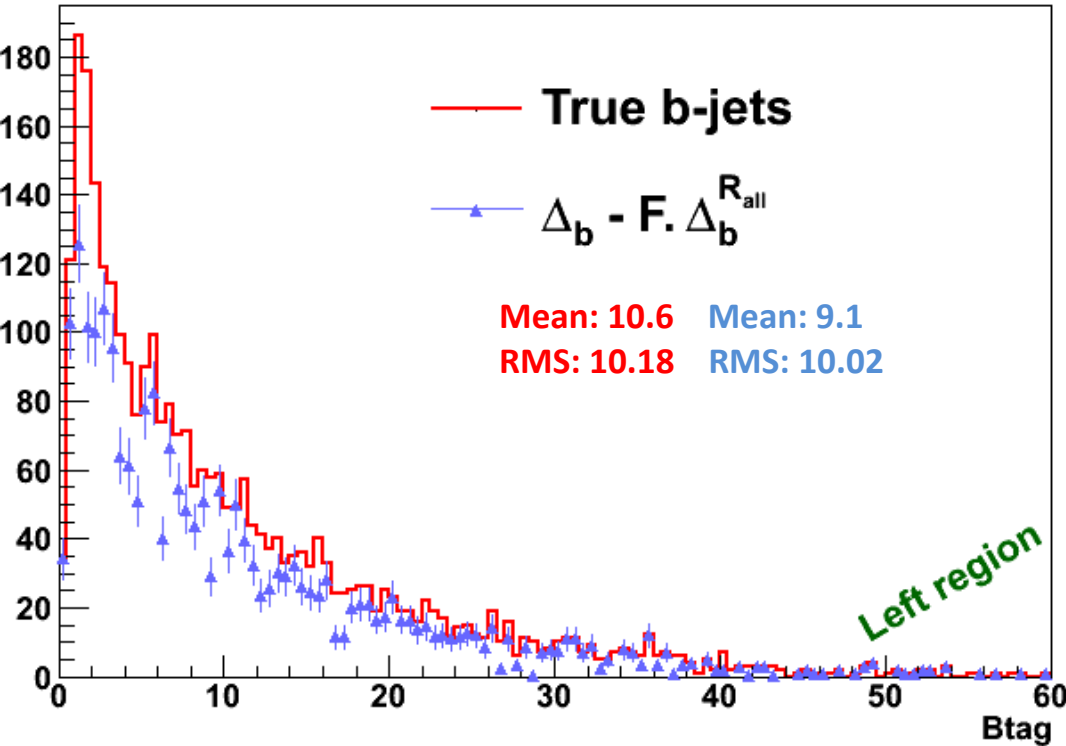


At MC level, the method works fine

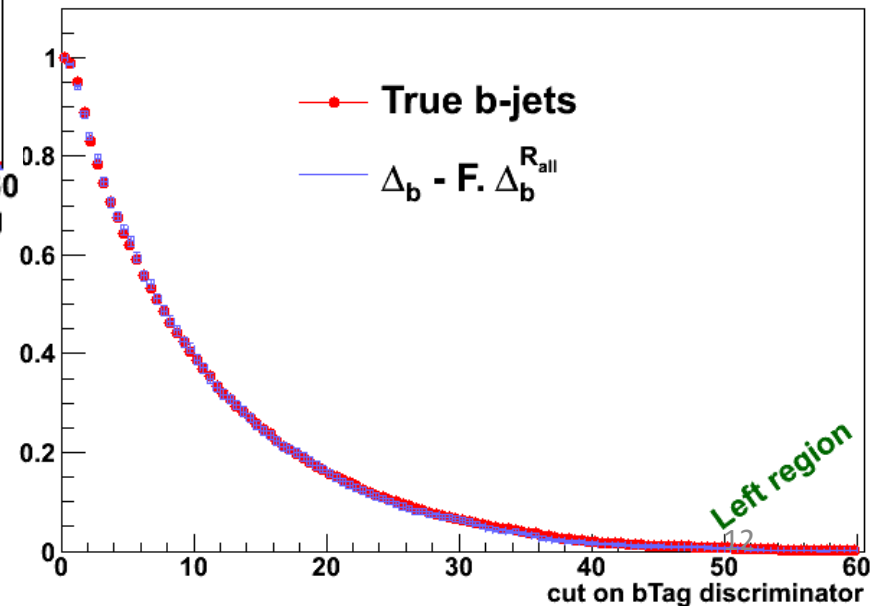


# Second Try: MC based study, F from MC truth, All jets in right region are taken as light

$$\Delta_L^{lep b Cand} = \Delta_L^{total} - F \cdot \Delta_R^{all Jets}$$



- The bTag distribution is different
- The difference is mainly in the number of entries, not the shape. Comparing with the previous case which we had a very good match, now we take all jets in right region as non-b, so we overestimate non-bs and underestimate b-jets in left side (\*).

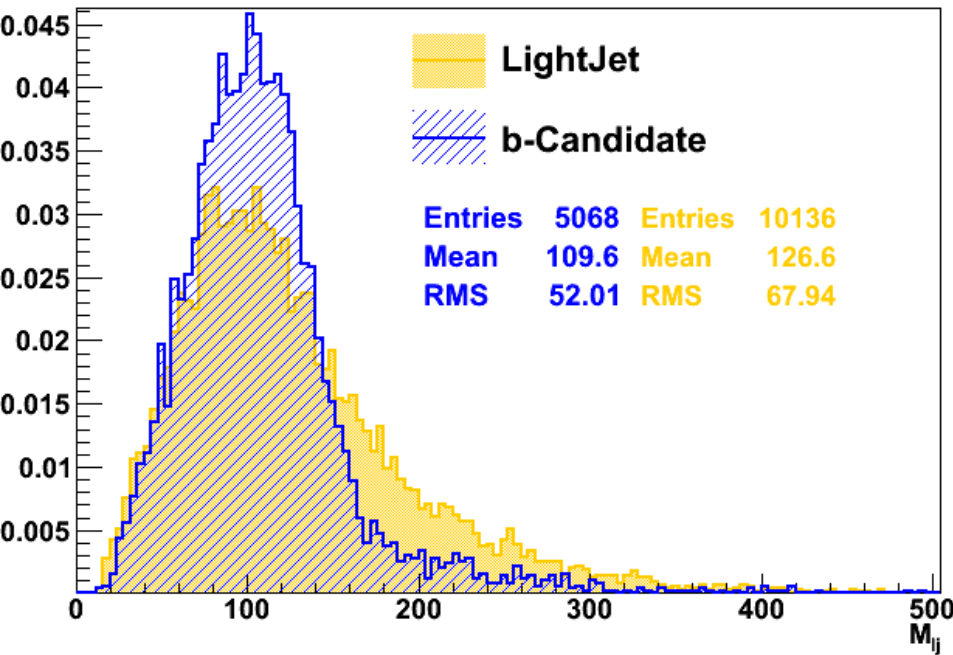


- As the shapes are not too different, the b-tag efficiencies are in kind of agreement.

(\*): b-purity in right region is about 50%

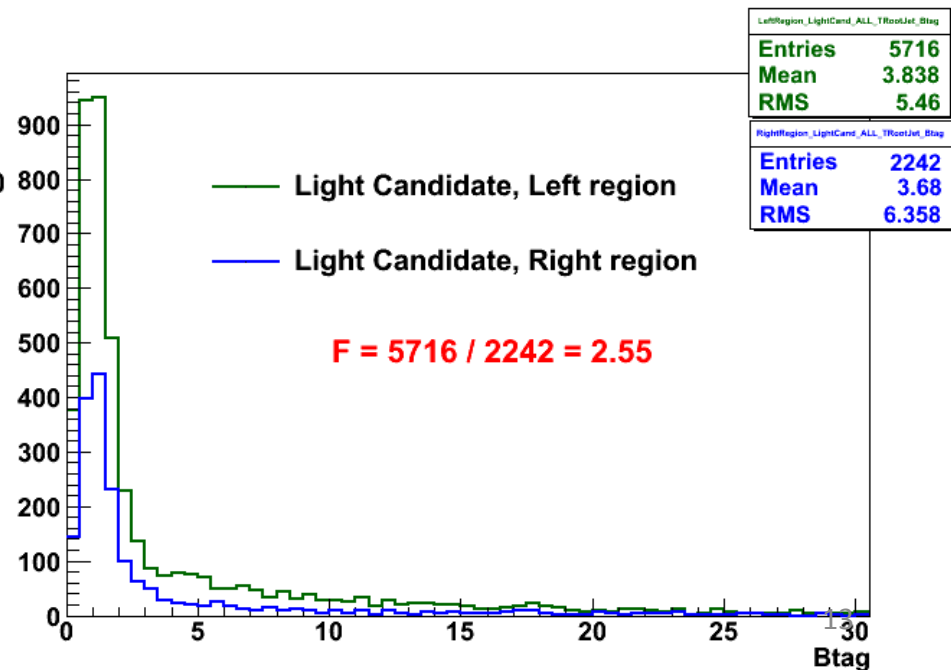
# Third Try: MC based study, F from control sample, Light jets in right region are matched with light quarks

$$\Delta_L^{lep bCand} = \Delta_L^{total} - F_{ControlSample} \cdot \Delta_R^{lightJets}$$



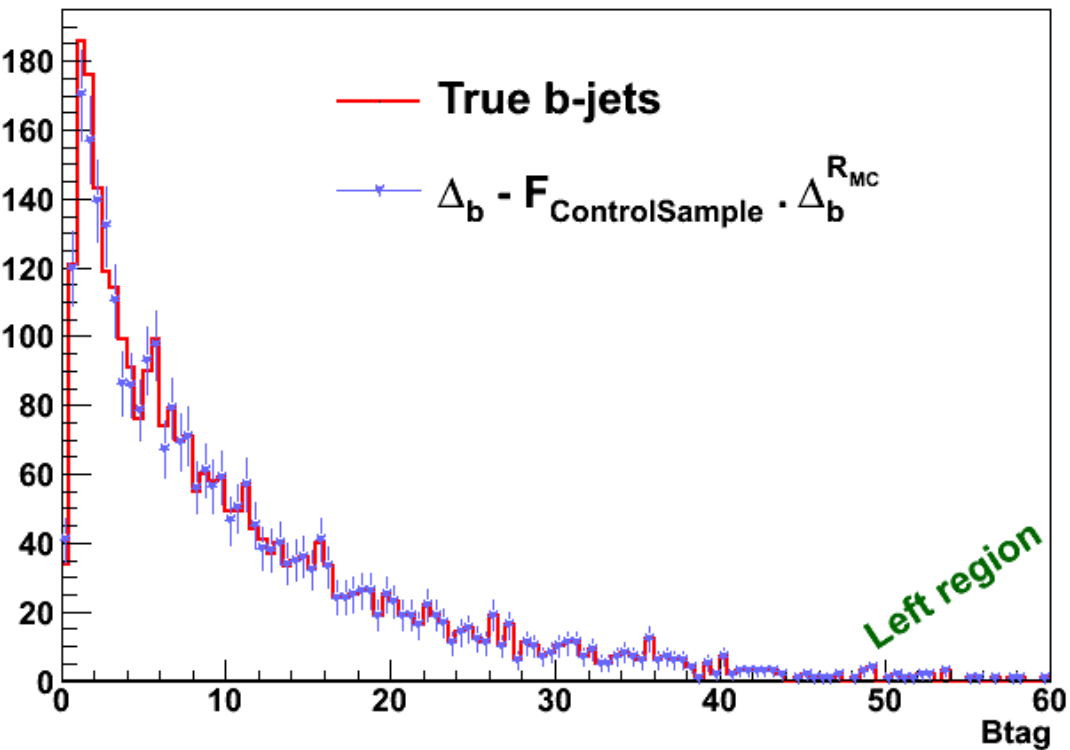
- Light candidates are those jets make W in the  $\chi^2$  formula.
- $F_{ControlSample}$  is the ratio between the number of light candidates in left and right region.
- This F can be obtained from data.
- It's value here is very close to  $F_{MC}$  (2.48)

- The peak in b-candidate distribution is more recognizable.
- The distribution for light jets, has a larger tail.



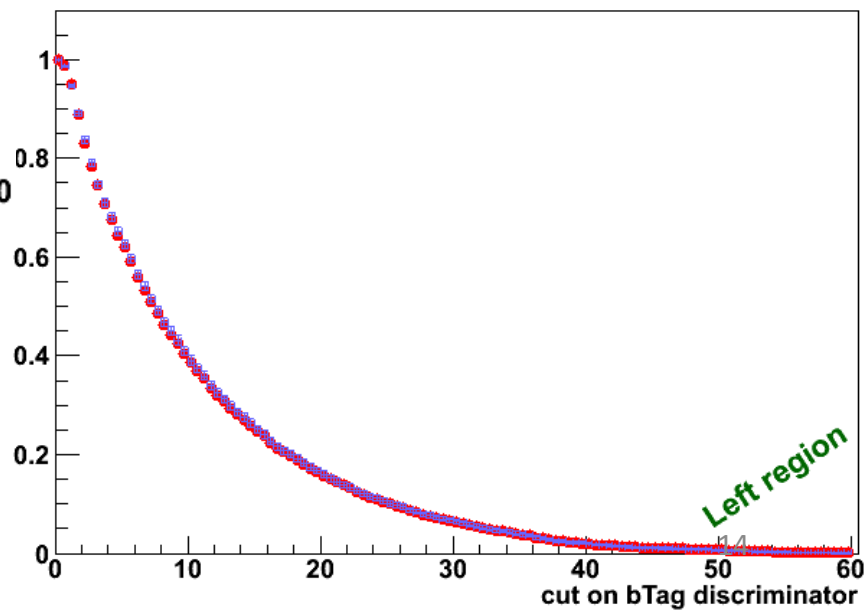
# Third Try: MC based study, F from control sample, Light jets in right region are matched with light quarks

$$\Delta_L^{lep b Cand} = \Delta_L^{total} - F_{ControlSample} \cdot \Delta_R^{lightJetsMC}$$



- Replacement of  $F_{MC}$  with the  $F_{ControlSample}$  and take the  $\Delta_R$  (bTag distribution in right side) from MC truth.

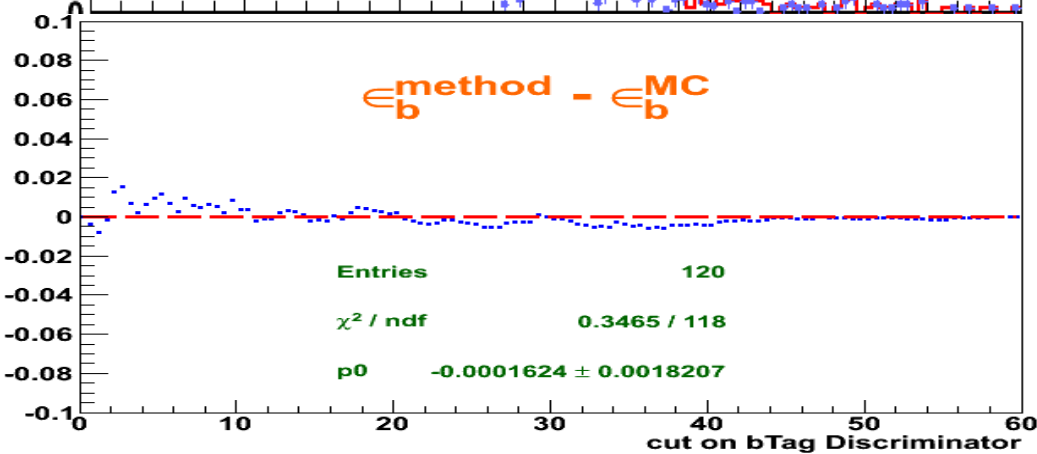
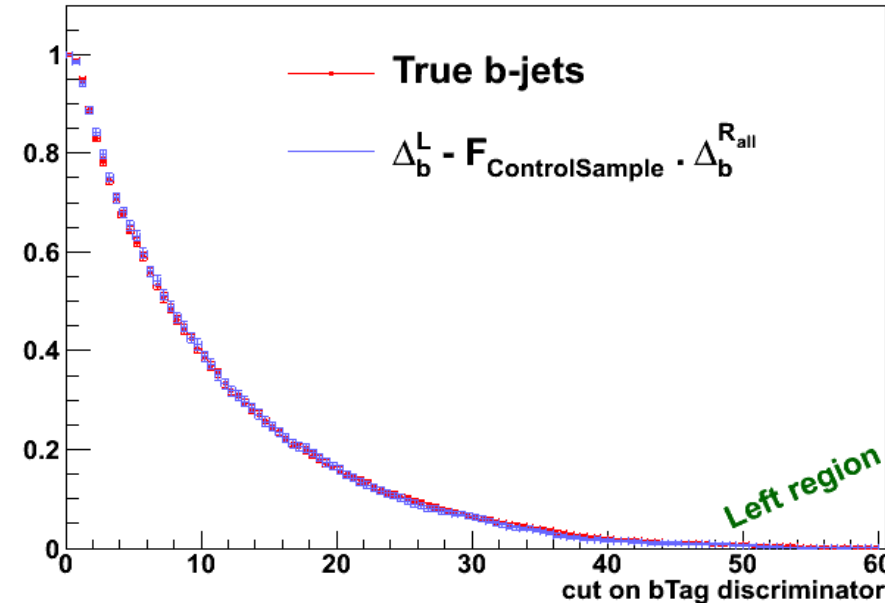
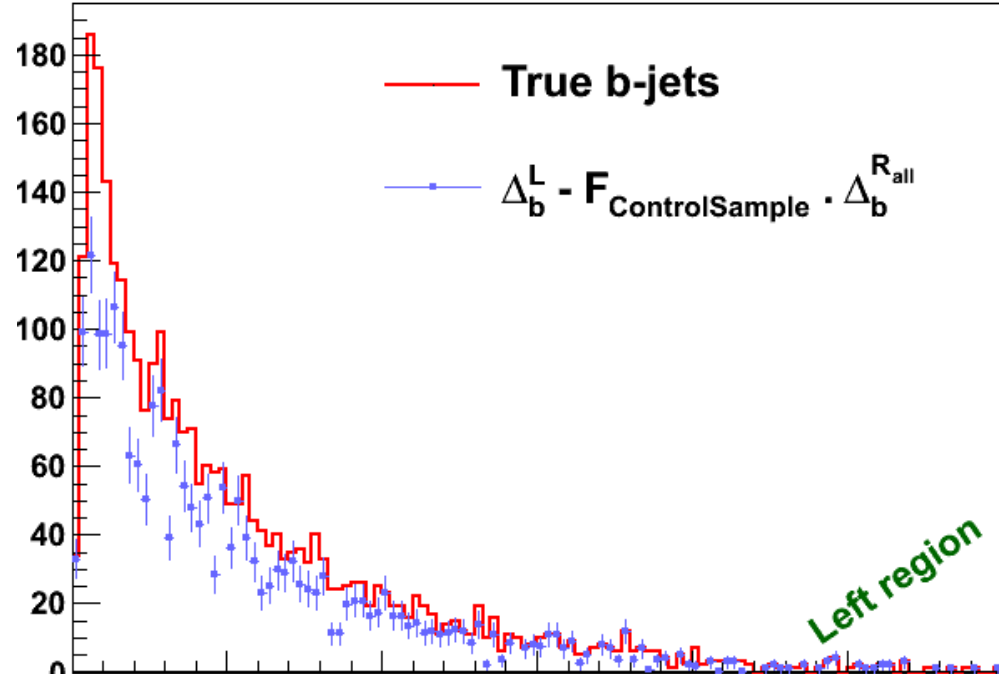
- As  $F_{ControlSample} \approx F_{MC}$ , the results are similar



# Final Try: MC based study, F from control sample, All jets in right region are taken as light

Fully data-driven

$$\Delta_L^{lep bCand} = \Delta_L^{total} - F_{ControlSample} \cdot \Delta_R^{allJets}$$



- The ratio is  $F_{ControlSample}$
- All b-candidates in right region are considered
- Method still works fine

# Conclusion (I)

- The ability of  $\chi^2$  method is studied using only signal events.
  - No cut is applied on  $\chi^2$  value
  - The minimum value is already chosen.
- The b-tag distribution of b-jets in left side is extracted using different ways, from fully MC based to fully data driven.
- The method is stable and working fine.

# Adding Backgrounds

7TeV samples are not completely ready.

# Event selection, 100 pb<sup>-1</sup>

	Ttbar Pythia		TtBar MadGraph		W + jets	Z +jets	SingleTop		
	signal	Others	signal	Others			S-channel	T-channel	TW-channel
Initial	1140.7	6559.3	1140.7	6559.3	856000	84000	100	2600	580
Trigger	1027.3	3726.0	1032.7	3726.0	203898.3	29300.7	48.7	1218.8	318.1
>= 1GE	610.1	339.4	608.8	345.0	107909.3	17884.1	16.3	469.1	76.5
==1 GE	609.8	305.2	608.6	310.8	107906.0	11649.5	16.3	469.0	73.5
!2 <sup>nd</sup> Ele	586.5	277.7	583.7	282.5	107776.4	8804.5	15.9	459.6	70.2
!Muon	586.4	207.6	583.7	209.5	107771.7	8771.7	15.9	459.3	64.2
>= 4 GJ	288.6	48.3	305.6	51.2	200.1	33.1	0.8	21.8	12.1

	QCD						
	BEToE 20-30	BEToE 30-80	BEToE 80-170	EM 20-30	EM 30-80	EM 80-170	
Initial	3.84 E+06	4.8 E+06	4.56 E+05	6.4 E+07	9.4 E+07	5.7 E+06	
Trigger	6.0 E+06	1.8 E+06	2.8 E+05	1.2 E+07	2.6 E+07	2.5 E+06	
>= 1GE	3.7 E+03	5.6 E+03	1.9 E+02	5.1 E+04	5.5 E+04	5.0 E+03	
==1 GE	3.7 E+03	5.6 E+03	1.9 E+02	5.1 E+04	5.5 E+04	5.0 E+03	
!2 <sup>nd</sup> Ele	3.7 E+03	5.6 E+03	1.8 E+02	5.1 E+04	5.5 E+04	4.9 E+03	
!Muon	3.7 E+03	5.6 E+03	1.8 E+02	5.1 E+04	5.5 E+04	4.9 E+03	
>= 4 GJ	0	14.2	15.4	12.8	244.8	350.4	

# Event selection, 100 pb<sup>-1</sup>

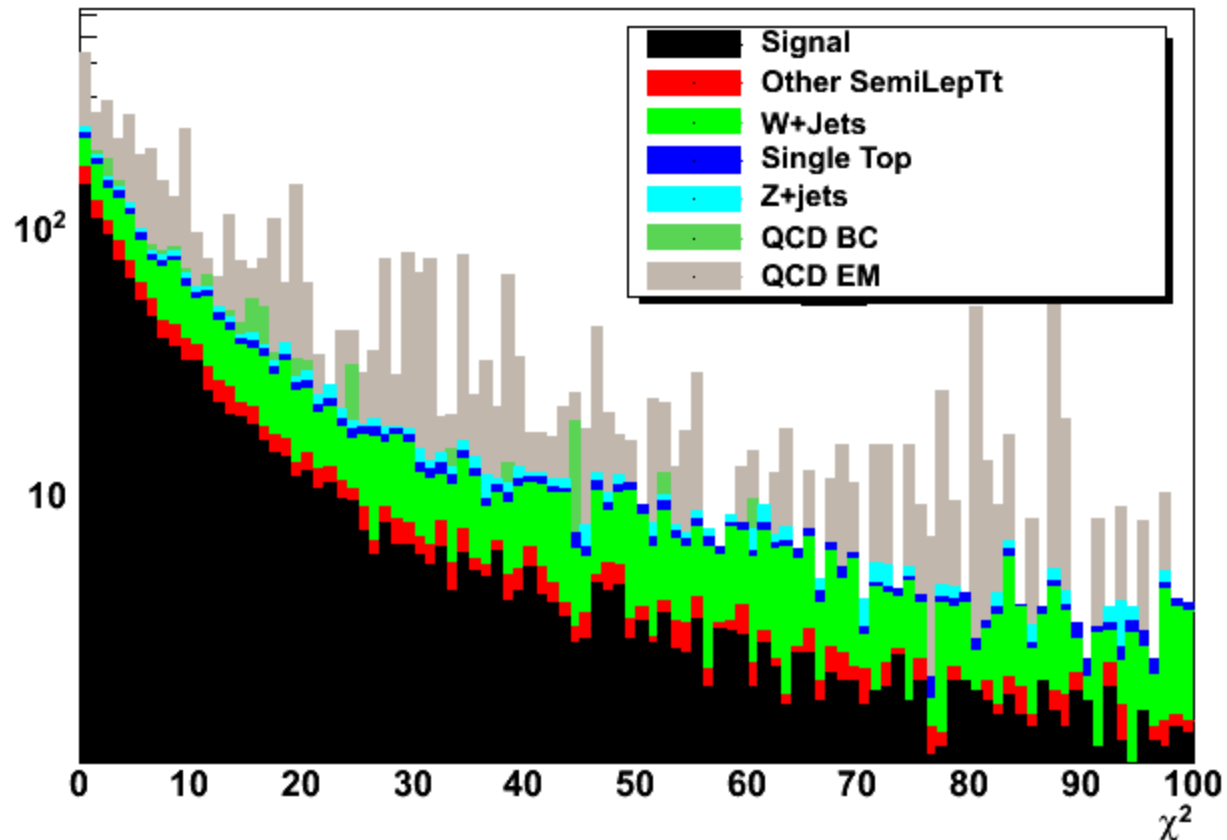
- QCD EmEnriched, specially in high p<sub>T</sub> range, is really a problem
- W+Jet has a very large contribution, too.
- Possibilities are
  - B-tagging:

	Ttbar Pythia		TtBar MadGraph		W + jets	Z +jets	SingleTop		
	signal	Others	signal	Others			S-channel	T-channel	TW-channel
>=1 bjet	268.4	45.1	284.9	47.7	61.8	12.3	0.8	20.0	10.5

	QCD					
	BEToE 20-30	BEToE 30-80	BEToE 80-170	EM 20-30	EM 30-80	EM 80-170
>=1 bjet	0.0	7.1	10.1	0	75.3	111.5

- Although the effect is promising, there is still ~200 QCD events
- The b-tag efficiency measurement method, should change accordingly e. g. , the tagged jet should only participate in hadronic top reconstruction
  - Trying other variables like HT
  - Asking for only Ecal-Driven electrons
- More studies are postponed to be done with 7TeV samples.

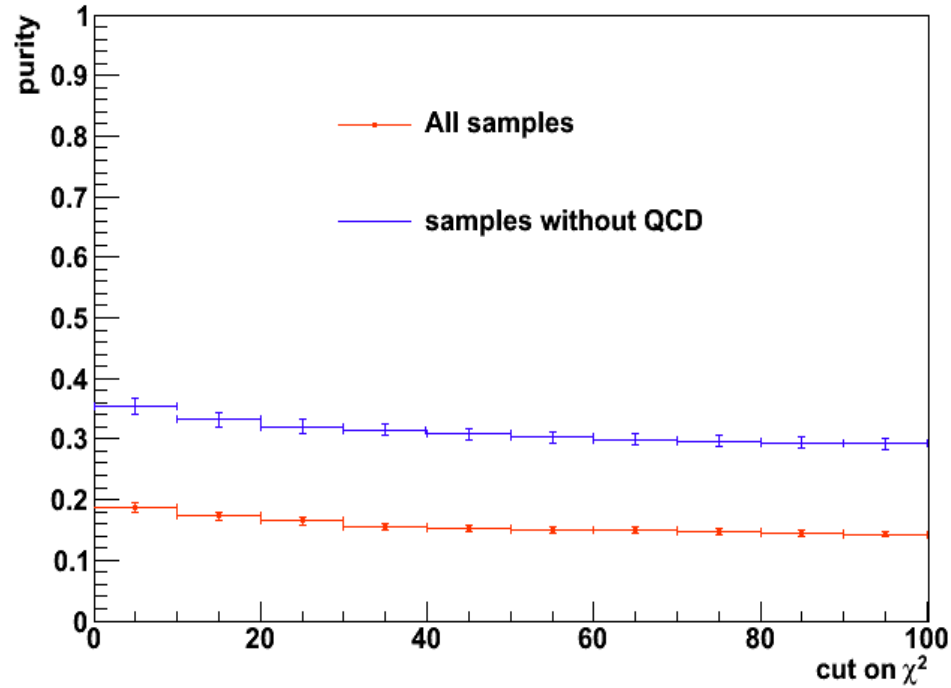
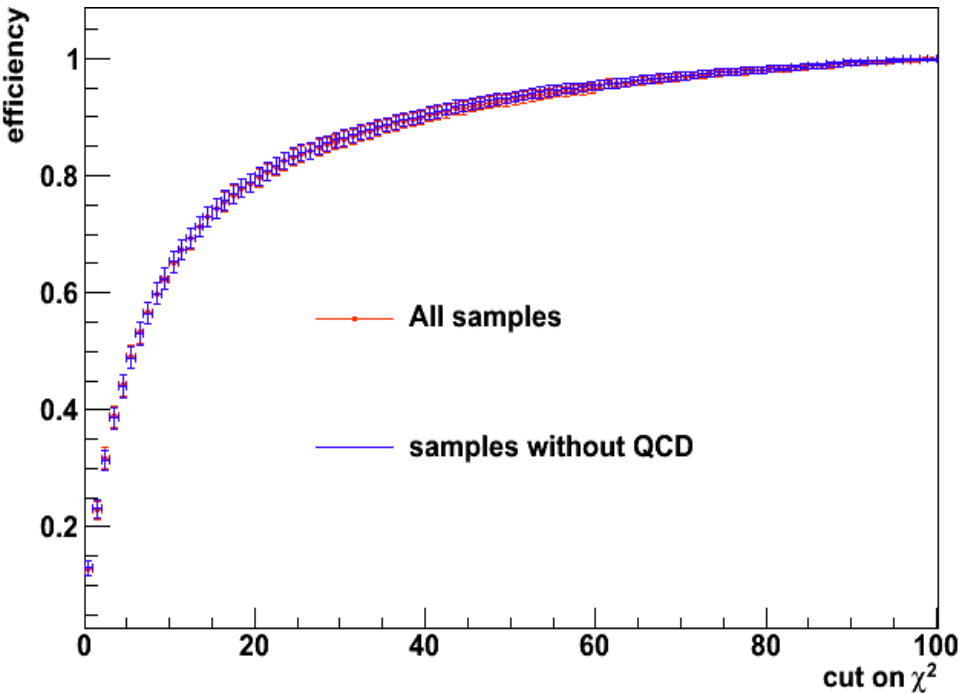
# $\chi^2$ Study



- Consistent with the event selection, most of the contamination is from QCD EM Enriched and W+jets.
- As the genInfo for the diLep and fullHad ttbar is not available in TopTrees, from hereafter only semiLep ttbar events are participated in the efficiency-purity study. Because in this part I need the generator level information.

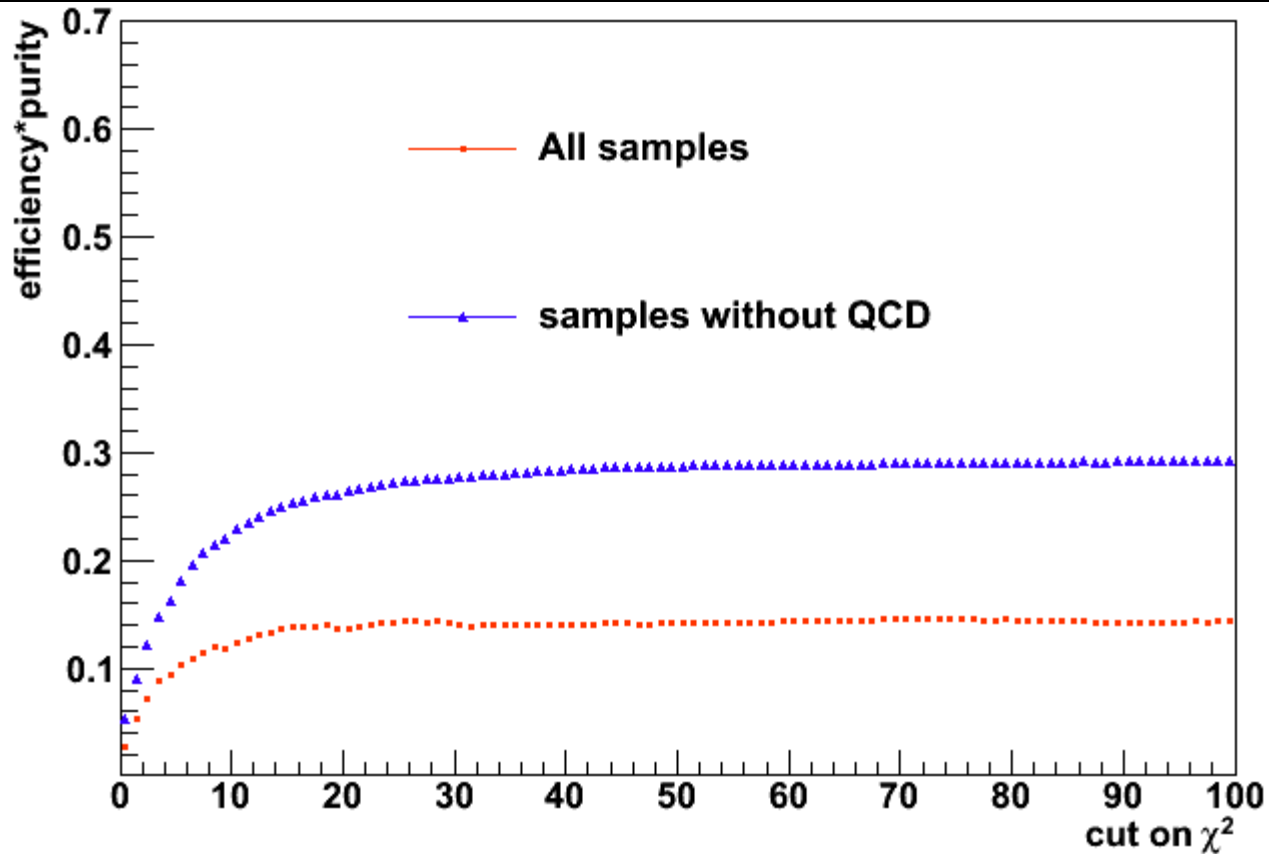
# Efficiency and purity

- Using the same procedure as before, the efficiency and purity are defined.
- Due to the uncertainty on the QCD events (large weights), each property is studied w/o QCD.



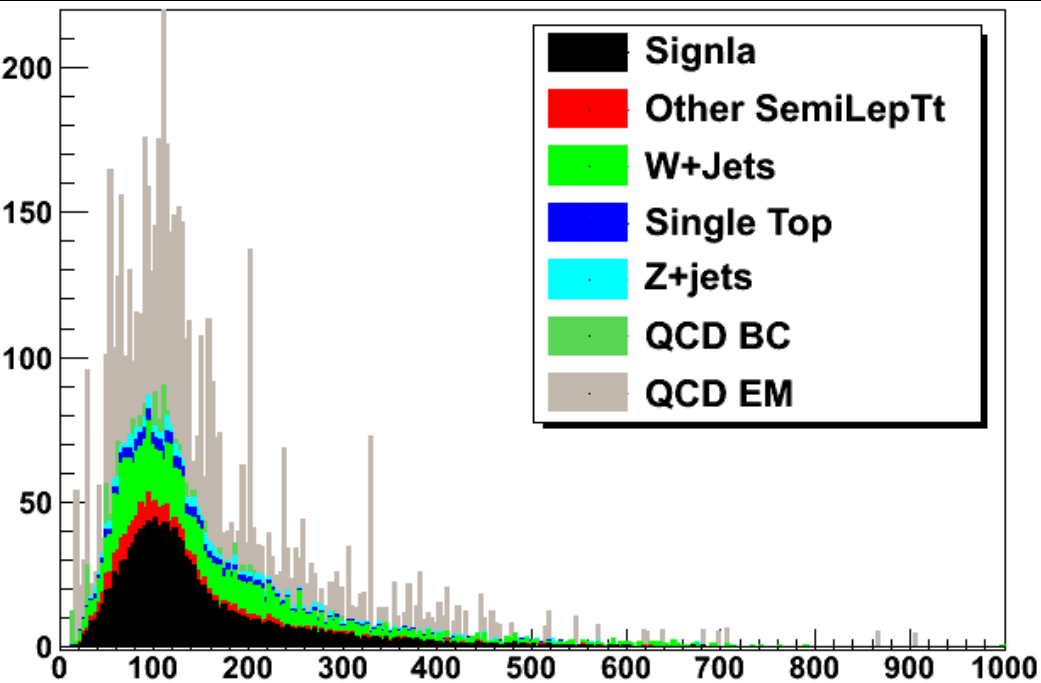
- The efficiency is not so different. The reason is that in both qcd and non-qcd events, the  $\chi^2$  distribution is tending to small values.
- The purity is of course different because the fraction of b's in qcd multi-jet samples are really small.

# Efficiency and purity

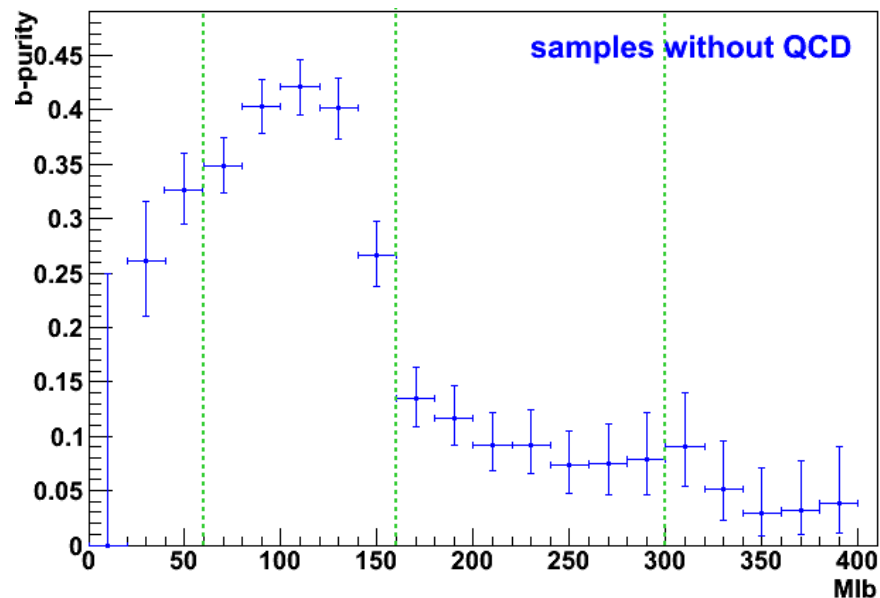
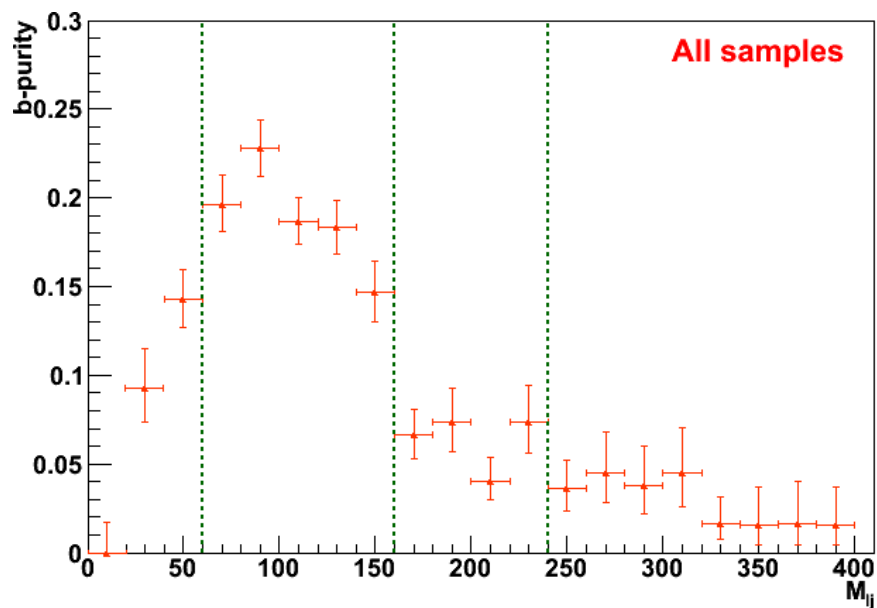


- A good point to cut is somewhere between 20 and 30 (should be studied more)
- The point is that cutting on  $\chi^2$ , does not change the purity while it costs about 20% of efficiency
- I go for the study without cutting until I find a convincing reason for it.

# $M_{lj}$ and Left/Right regions



- Many QCD events have entries in our signal (left) region. I should really find a way to get rid of them



## Conclusion (II)

- Many QCD EMEnriched events can survive the event selection
  - More variables should be tried to reduce this contamination
- These QCD events, can pass the cut on  $\chi^2$  and don't have any special effect on the efficiency.
- They reduce the purity by about 16% because they don't have real b-jets.
- Also, these QCD events are mostly in the Left region (peak area of  $M_{lj}$ ). So they have a considerable effect on the b-tag efficiency measurement.
- Working w/o QCD, may change the boundaries of the left and right regions.
- More studies on the event selection and QCD rejection is to be done with 7TeV samples.

# Backup

