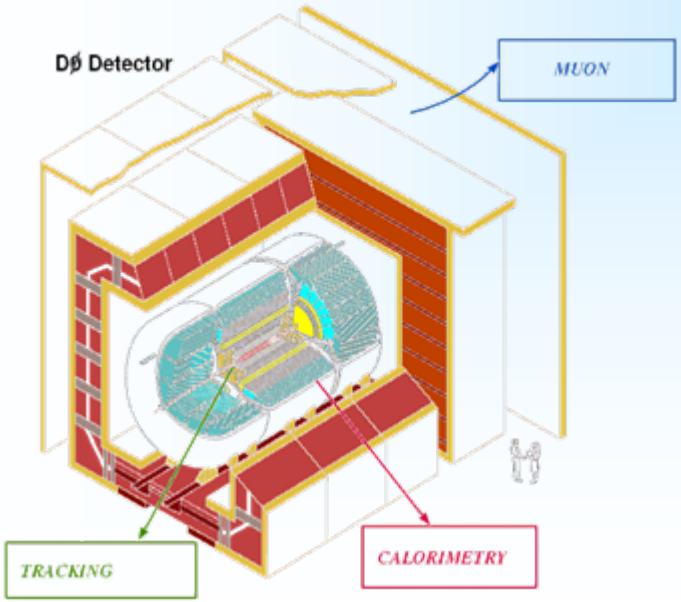
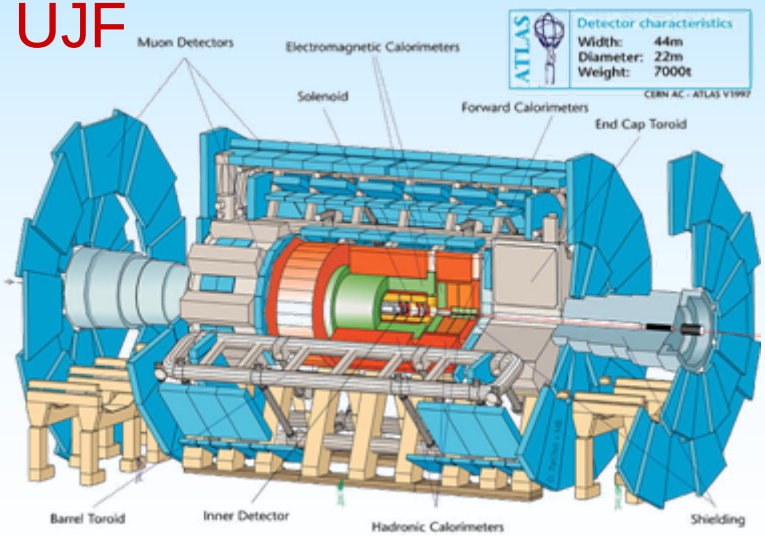
Using real data in physics analyses

Tips and tricks from DØ

Benoit Clément

LPSC / UJF





Correcting MC to model data

MC is never perfect : Several level of discrepencies

- Generator level : LO/NLO, showering model, multiple interaction…
- Detector level : missing material, bad description of physical properties, error in geometry…
- Digitization level : detector response, electronic noise…

Discrepencies affects :

- Distribution normalizations
- Distribution shapes

The right solution : **FIX EVERYTHING**

- requires a long time : months ? years ? decades ?
- rarely doable on a short timescale

The practical solution

Correcting distribution **shapes** (usually p_T spectrum) :

Degrades resolution in MC

Smearing : $E_T^{smeared} = E_T^{raw} \times S + O$

S : shift, random number following a gaussian distribution : $\sigma^2 = \sigma_{data}^2 - \sigma_{mc}^2$

O : constant offset

Can be more complex for non gaussian errors…

Applied to jets, EM objects, muons

-> propagate correction to E_T^{miss}

The practical solution continued

Corrections to cut efficiencies -> 3 possible ways

❶ **Apply the cut** in MC samples and **weight the event** by a corrective scaling factor :

$$SF = \varepsilon_{data} / \varepsilon_{MC}$$

->SF close to 1, almost constant

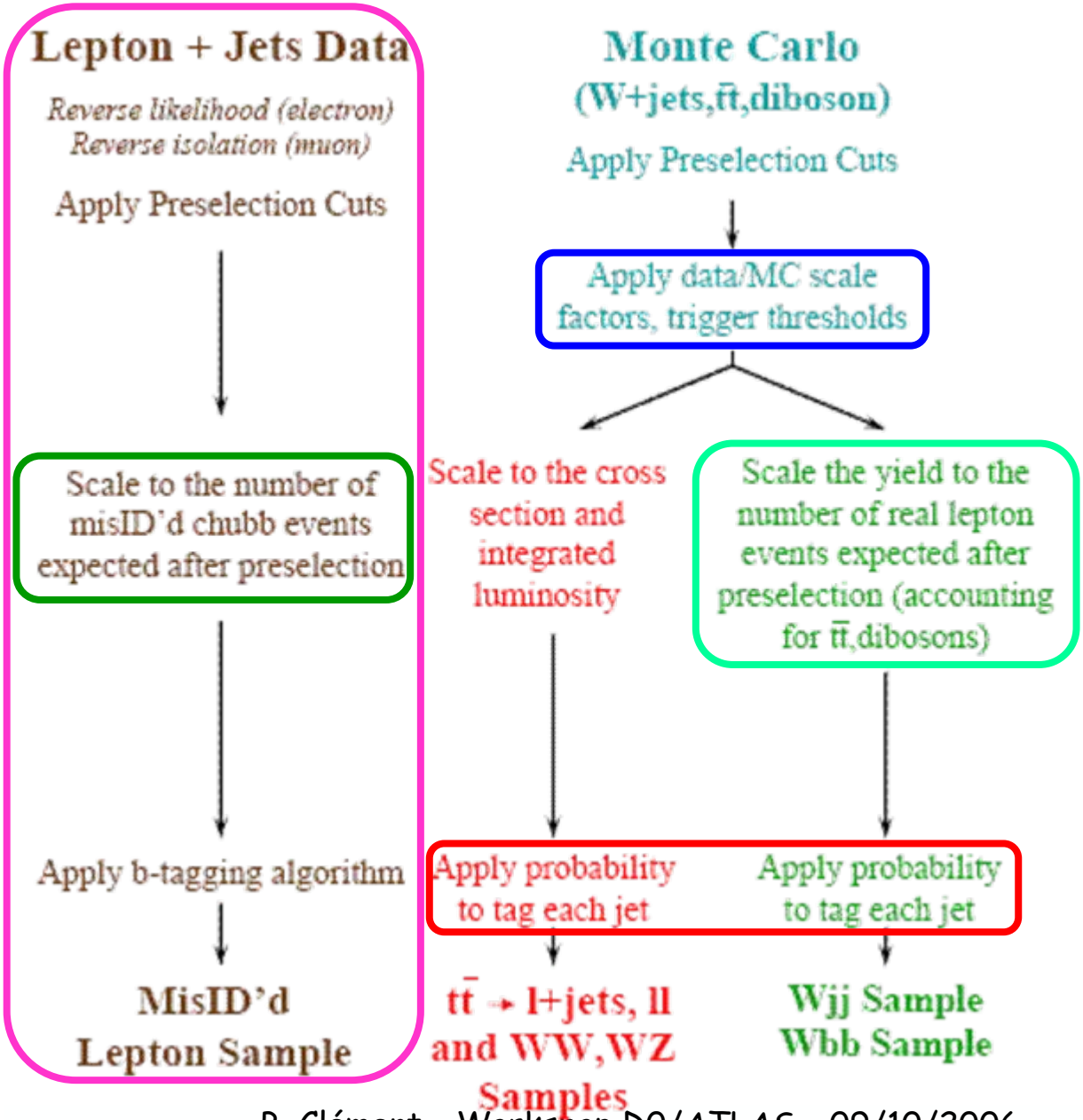
❷ **Do not apply the cut**, **weight the event** by a probability to pass the cut measured in data

-> Statistical method, needs a correct parametrization of probability densities(=cut efficiency)

❸ **Do not apply the cut**, use the probability to pass the cut measured in data to **reject events/objects**

-> Can introduce large statistical fluctuation if small probability

Analysis flow in l+jet (tt and single top)



Multijet background initial sample

Multijet background normalization

Correction for selection cuts efficiencies (including triggers)

W+jets background normalization

b-tagging efficiency

Jets and Lepton

Jets and leptons

How to measure selection efficiency on DATA ?
Need an unbiased sample w/ respect to studied cut

Reconstructed jets : Easy to get jet samples in an hadronic environment !
-> Use jet triggers
-> veto on EM triggers

Electrons/Muons :
Soft p_T (<15 GeV) : Use $J/\psi \to l^+l^-$ Hard p_T (<100 GeV) : $Z \to l^+l^-$
so-called **« tag and probe » method**

Count lepton pairs within mass peak (fit or window cut)
- 1 well identified « tag » lepton
- use second « probe » lepton to mesure efficiency by counting the number of Z or J/ψ ;

$$\varepsilon_{data} = \frac{\#\ Z\ \textit{with probe lepton satifying the tested cut}}{\#\ Z\ \textit{Bosons}}$$

Use similar method in MC, to remove pentential bias.

Triggers

Triggers : event weights

Turn on curves are computed :

- For **each kind of object** (mu, jets, electrons)
- At **each level** (L1, L2, L3)

A trigger requires L1, L2 and L3 terms for each kind of object

eg. EM15_2JT15 :

L1 : 1 EM cal tower (E_T>10 GeV)	2 Cal towers (E_T>5 GeV)
L2 : 1 L2 EM object (E_T>10 GeV)	2 L2 jets (E_T>10 GeV)
L3 : 1 L3 EM object (E_T>15 GeV)	2 L3 jets (E_T>15 GeV)

For each reco object : $P_{object}(L1L2L3) = P(L1) x P(L2|L1)\ x P(L3|L1L2)$

For n reconstructed object in the event:

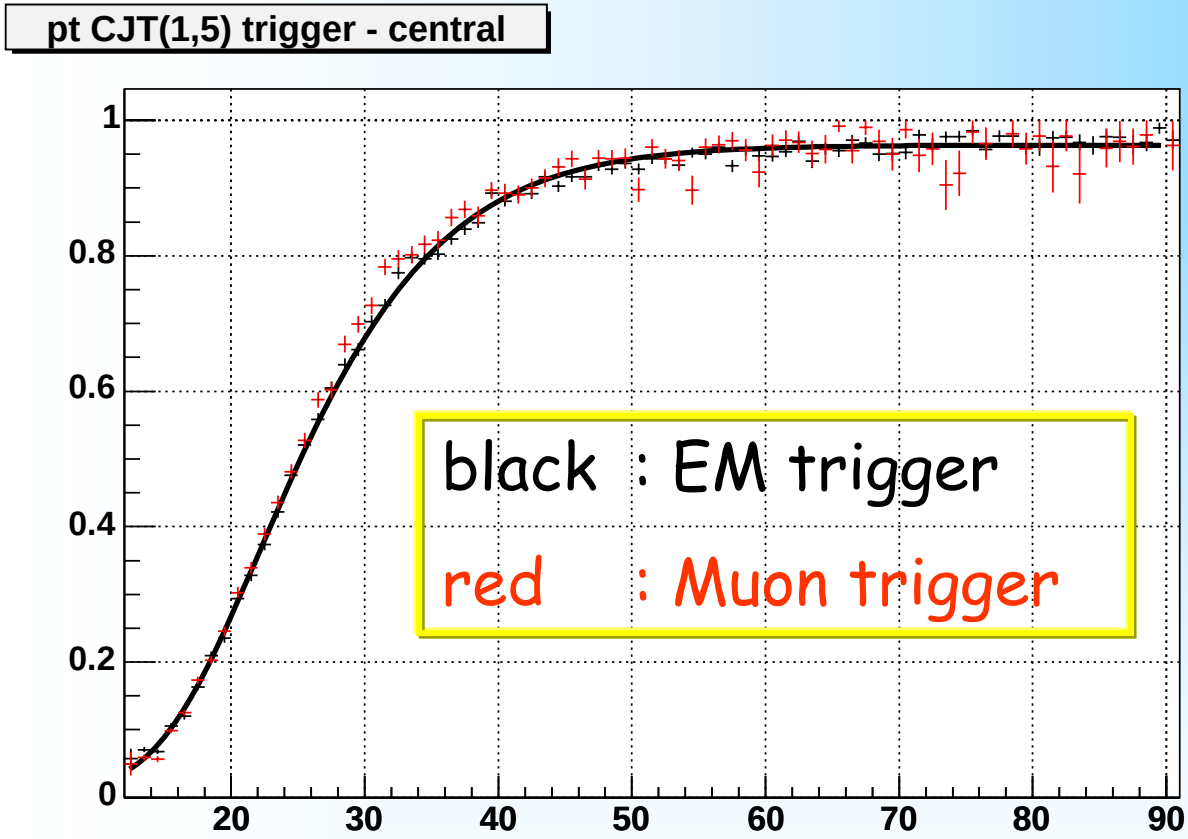
$$P_{\geq 1obj} = 1 - \prod_{i=1}^{n} (1 - P_{object\ i}) \qquad P_{\geq 2obj} = P_{\geq 1obj} - \sum_{i=1}^{n} \left(P_{object\ i} \prod_{j=1, j\neq 1}^{n} 1 - P_{object\ j} \right)$$

Event trigger weight : $P_{trigger} = P_{em} \times P_{mu} \times P_{jet} \times ...$

more complicated if correlations…

Example : Jet trigger turn-ons

- Use EM trigger (with reco EM matched to the triggering object)
- Use muon trigger : bias from larger HF fraction ?



Turn On :

L1 : 1 tower E_T>5GeV

match reco jets to L1 jets (dR<0.5)

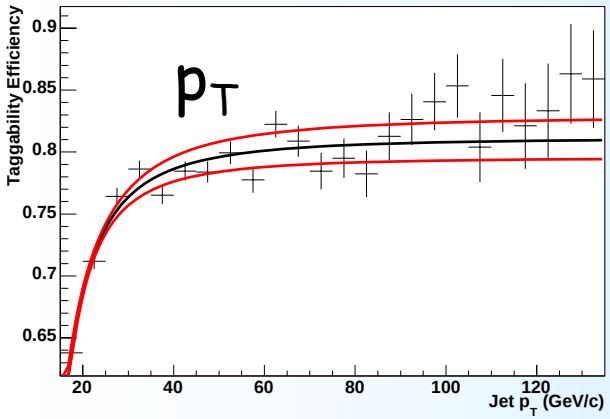
$$\varepsilon_{L1} = \frac{\#\ Matched\ reco\ jets\ (E_T^{L1} > 5)}{\#\ Matched\ reco\ jets}$$

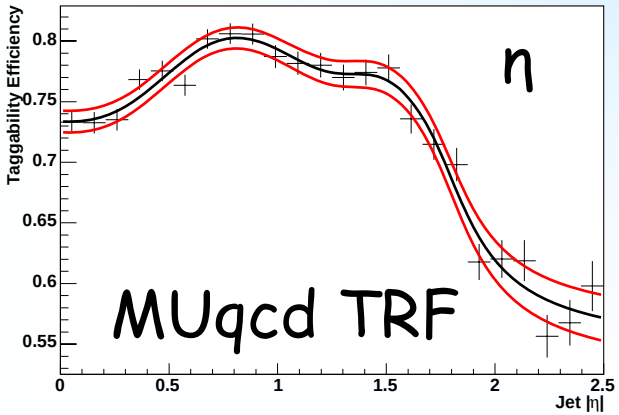
B-tagging

B-Tagging : taggability

Jets -> calorimeter only **Lifetime tagging -> tracks**

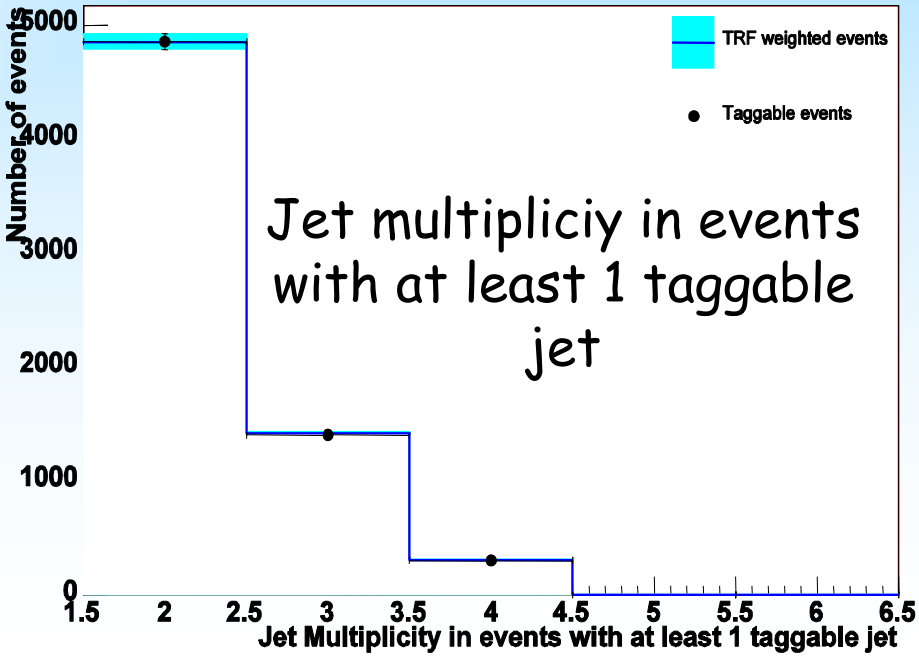
Taggability : matching Jets / tracks -> minimal requirement to run btagging algorithm





Efficiency estimated on data with similar topology :

1 loose lepton, >=2 jets, Met



B-tagging efficiencies

Lifetime tagging algorithms -> based on tracking

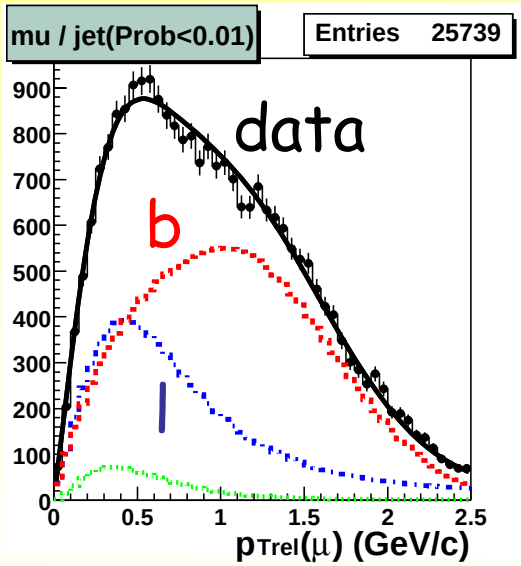
- signed IP
- signed 2^{ndary} vertex decay length

Efficiency measurement:

Heavy flavor enriched sample from **semi-leptonic decays** of b quarks (muon within the jet cone)

Several methods :

- b-fraction estimated from p_{Trel} fits
- System8



All methods needs muonic jets : correction from muonic to inclusive decay from MC (efficiency ratios)

Mistag rate measurment:

Use negative tags in multijet data

B-tag efficiency : System8

Combine 3 (almost) **uncorrelated** identification criteria :

- **The tagging algorithm**.
- Cut on muon momentum w.r.t jet axis (p_{Trel}).
- Look for a second tag jet in the event (b are produced by pairs).

Write and solve a system of 8 non linear equations with 8 unknowns.

There is no MC input

Small correlation between criteria can be estimated on MC.
Several hypothesis/approximation -> systematic uncertainties

Has been tested end verified at DØ.
Standard B-tagging efficiency measument method.
Also used for photon Id efficiency.

B-tag efficiency : System8

fractions passing a cut (from data)

signal / background fractions

$$
\begin{cases}
1 = n_b + n_{cl} \\
q^{(T)} = \varepsilon_b^{(T)} n_b + \varepsilon_{cl}^{(T)} n_{cl} \\
q^{(M)} = \varepsilon_b^{(M)} n_b + \varepsilon_{cl}^{(M)} n_{cl} \\
q^{(B)} = \varepsilon_b^{(B)} n_b + \varepsilon_{cl}^{(B)} n_{cl} \\
q^{(T,M)} = \varepsilon_b^{(T)} \varepsilon_b^{(M)} n_b + \varepsilon_{cl}^{(T)} \varepsilon_{cl}^{(M)} n_{cl} \\
q^{(M,B)} = \varepsilon_b^{(M)} \varepsilon_b^{(B)} n_b + \varepsilon_{cl}^{(M)} \varepsilon_{cl}^{(B)} n_{cl} \\
q^{(B,T)} = \varepsilon_b^{(B)} \varepsilon_b^{(T)} n_b + \varepsilon_{cl}^{(B)} \varepsilon_{cl}^{(T)} n_{cl} \\
q^{(T,M,B)} = \varepsilon_b^{(T)} \varepsilon_b^{(M)} \varepsilon_b^{(B)} n_b + \varepsilon_{cl}^{(T)} \varepsilon_{cl}^{(M)} \varepsilon_{cl}^{(B)} n_{cl}
\end{cases}
$$

(T) : Tagger

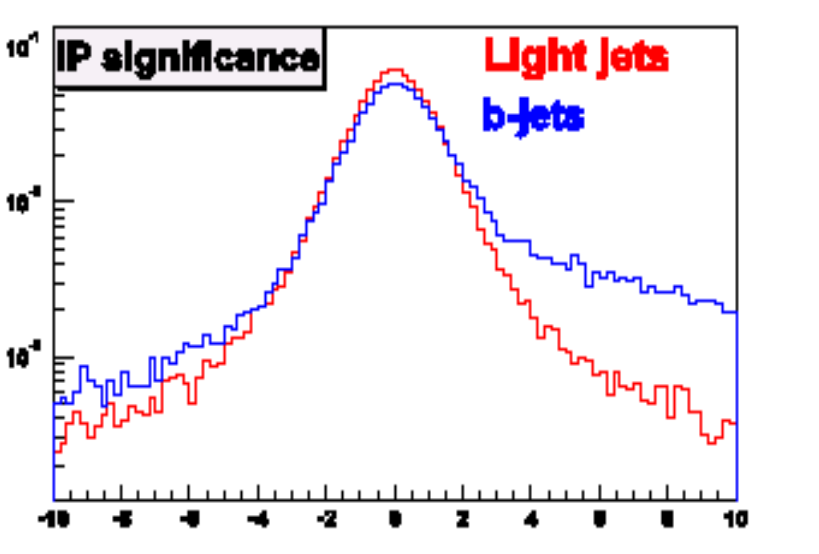
(M) : Muon tag

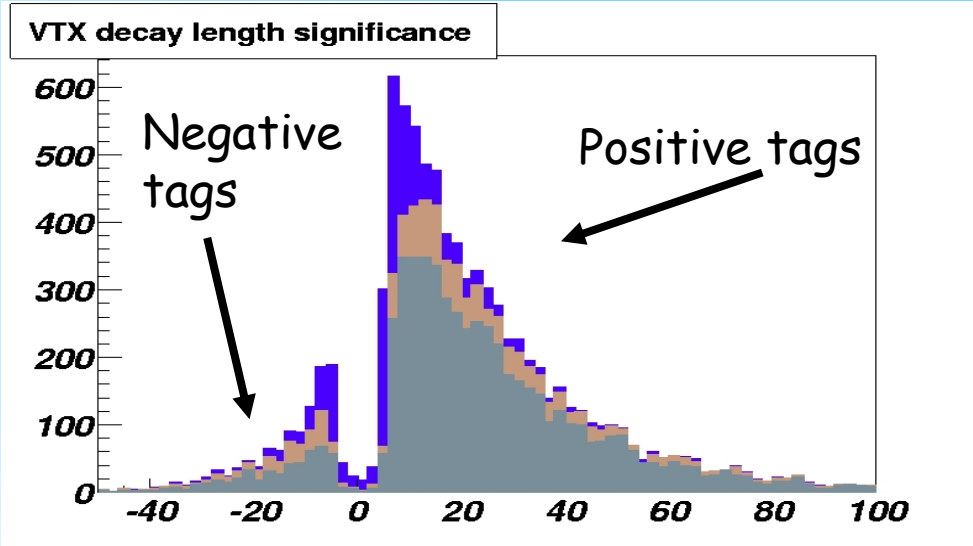
(B) : 2nd jet

selection efficiencies

Mistag Rate

Lifetime tag
- tracks with **positive** IP
- 2ndary vertex with **positive** decay length





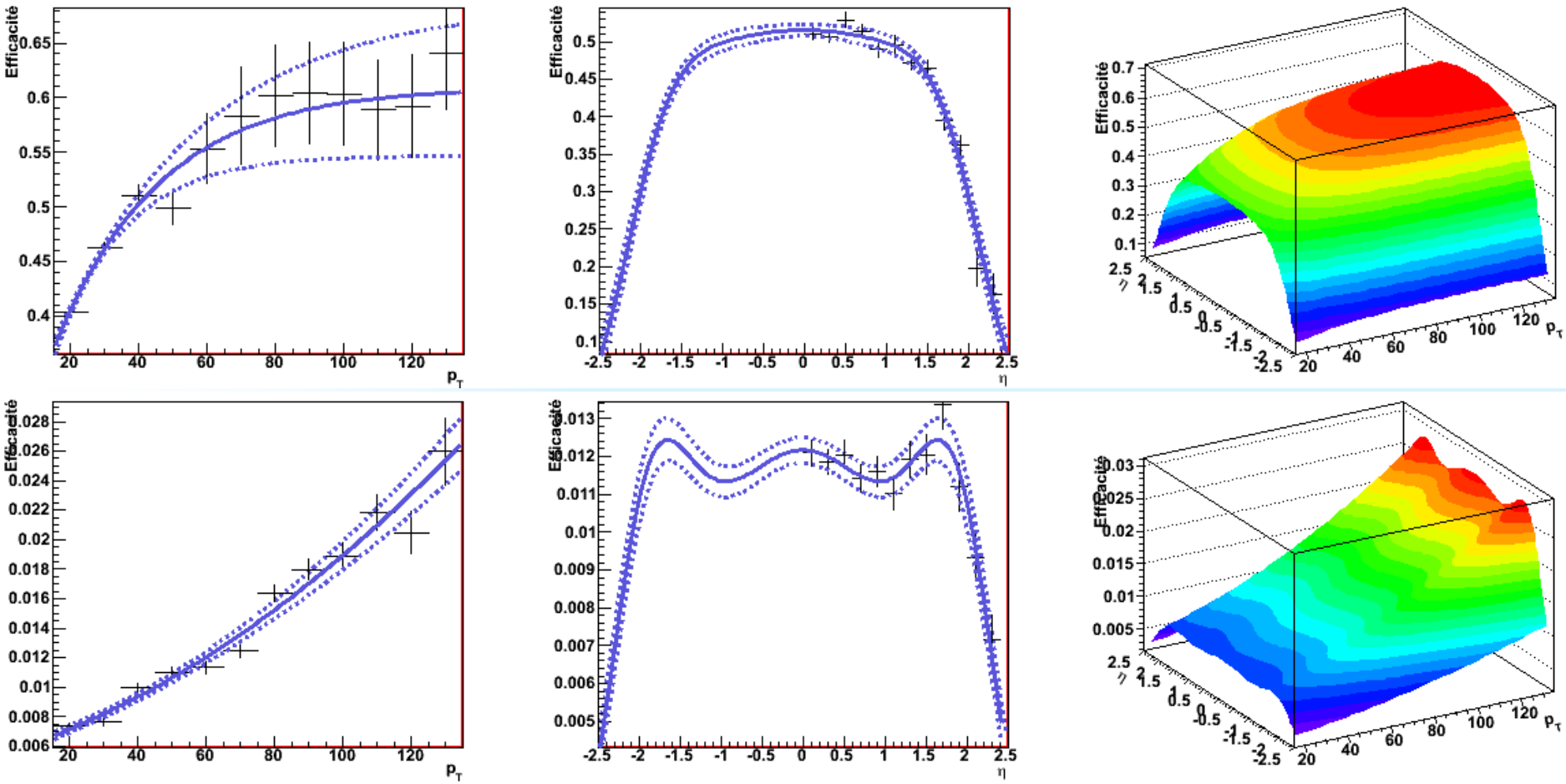
Negative tag : run tagging algorithm with negative IP/decay length
- gives a rough estimation of mistag
- needs to be corrected for :
  - **HF content in multijet data**
  - **Long lived particles (K^0_S, Λ) in light jets**

These scaling factors are derived from MC.

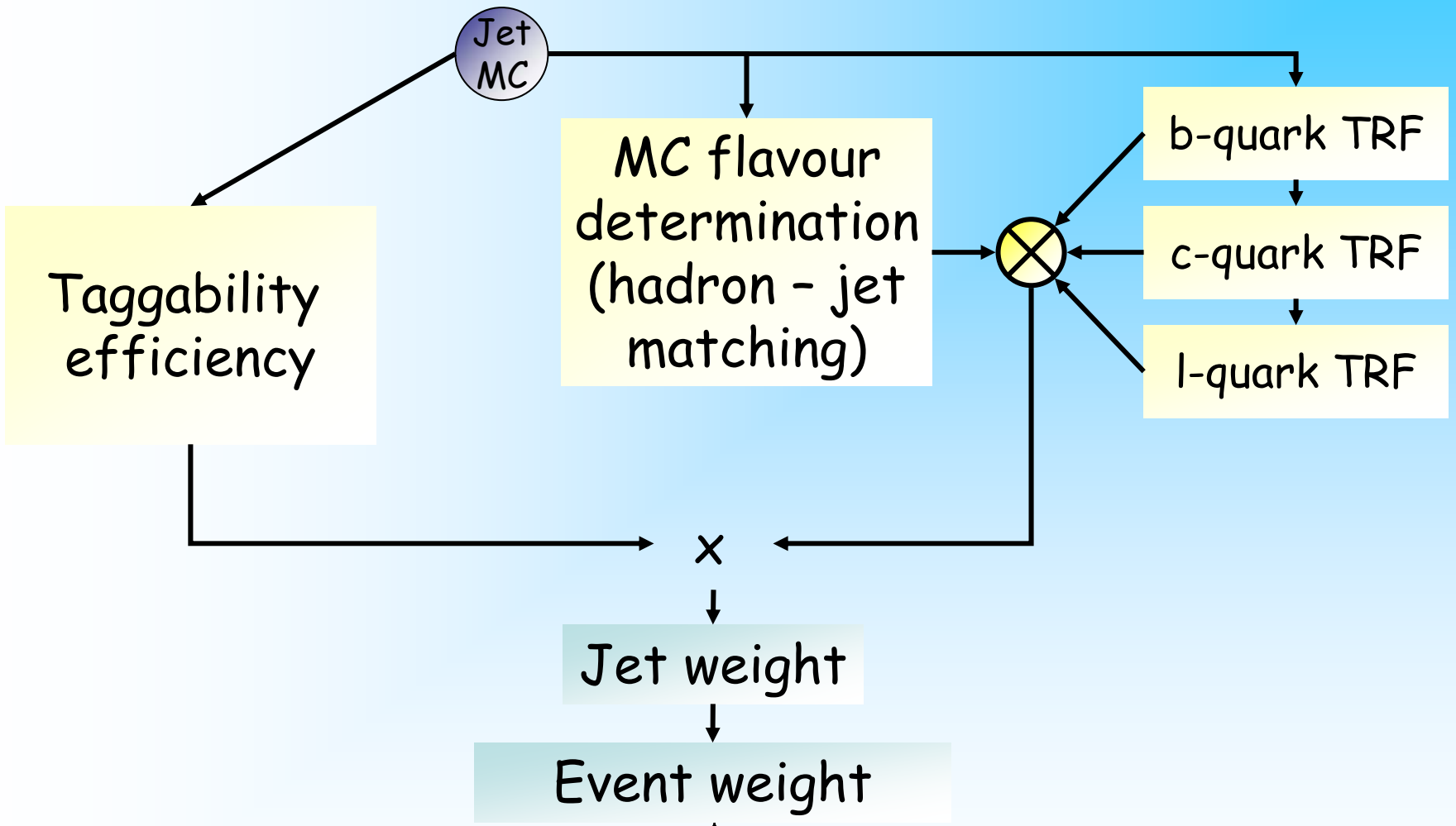
Tag Rate Function (TRF)

Efficiency parametrisation as a function of:

- jet tranverse momentum : p_T
- jet pseudorapidity : η
- jet flavour : b, c, light



Using TRFs



$$P_{\geq 1tag} = 1 - \prod_{i=1}^{N_{jets}} (1 - P_i) \qquad P_{\geq 2tag} = 1 - \prod_{i=1}^{N_{jets}} (1 - P_i) - \sum_{i=1}^{N_{jets}} \left(P_i \prod_{j=1, j \neq 1}^{N_{jets}} 1 - P_j \right)$$

Background normalization

Background normalisation

Basic selection before b-tagging

- jets
- missing tranverse energy
- isolated lepton

Composition of the selected data

W+jets = real isolated leptons

Multijet = fake isolated lepton

fake **electron** : jet

fake **muon** : real muon from a jet (HF, decay in flight)

Multijet backgrounds comes from misreconstruction or very electromagnetic jets -> very difficult to get from MC

Use data with standard selection and reverse id cut (fake enriched sample)

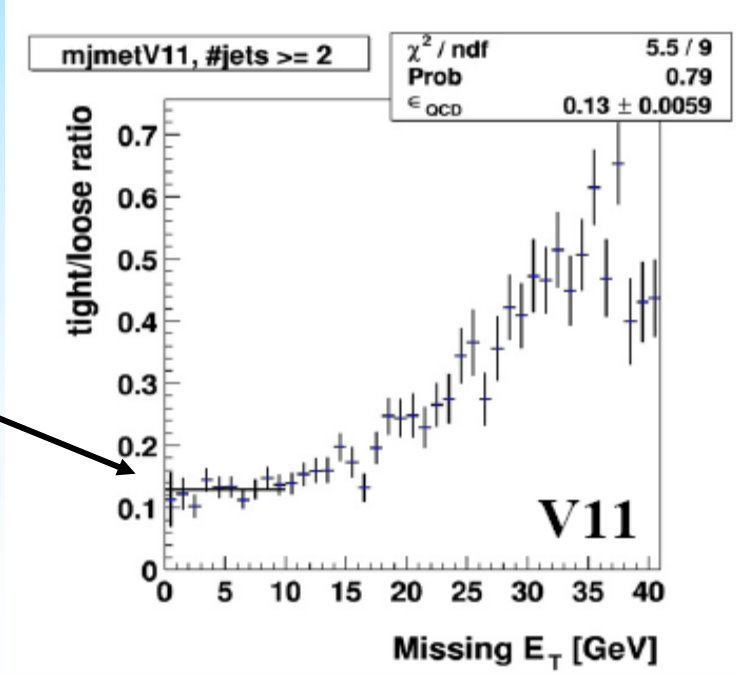
Multijet and W+jet normalization

Relax one lepton identification cut. Tight sample -> Loose sample
Determine selection efficiencies ε_{real}, ε_{fake} separatly

$$\begin{cases} N_{Loose} = N_{real} + N_{fake} \\ N_{Tight} = \varepsilon_{real} \cdot N_{real} + \varepsilon_{fake} \cdot N_{fake} \end{cases} \Rightarrow \begin{cases} N_{real} = \dfrac{N_{Tight} - \varepsilon_{fake} \cdot N_{Loose}}{\varepsilon_{real} - \varepsilon_{fake}} \\ N_{fake} = \dfrac{N_{Tight} - \varepsilon_{real} \cdot N_{Loose}}{\varepsilon_{fake} - \varepsilon_{real}} \end{cases}$$

ε_{real} is measured at the Z peak (tag and probe).

ε_{fake} is estimated on selected data with reverse $\not{E}_T$ cut.

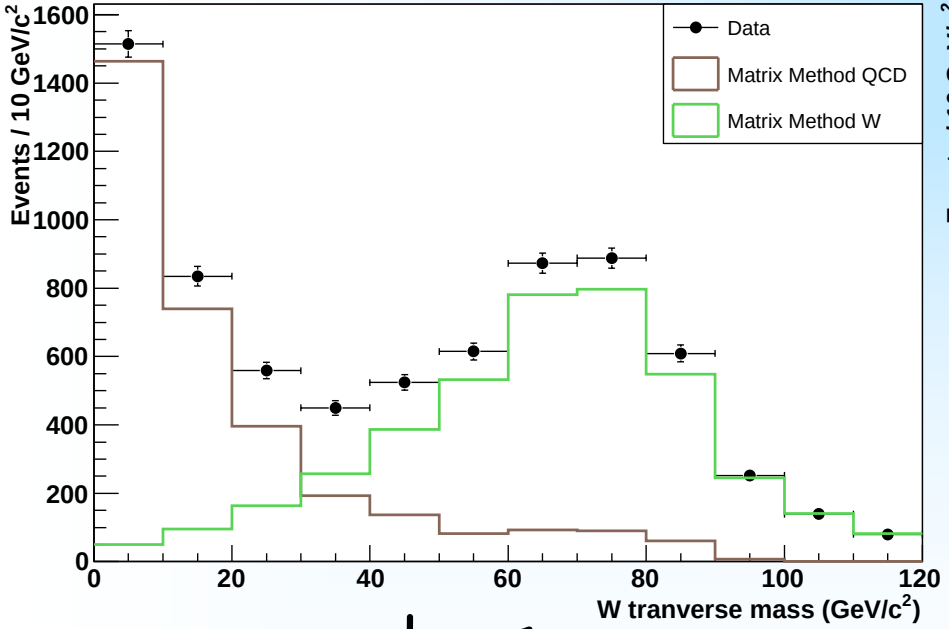


$\varepsilon_{real} \cdot N_{real}$ -> W+jets normalization.
$\varepsilon_{fake} \cdot N_{fake}$ -> Multijet normalization.

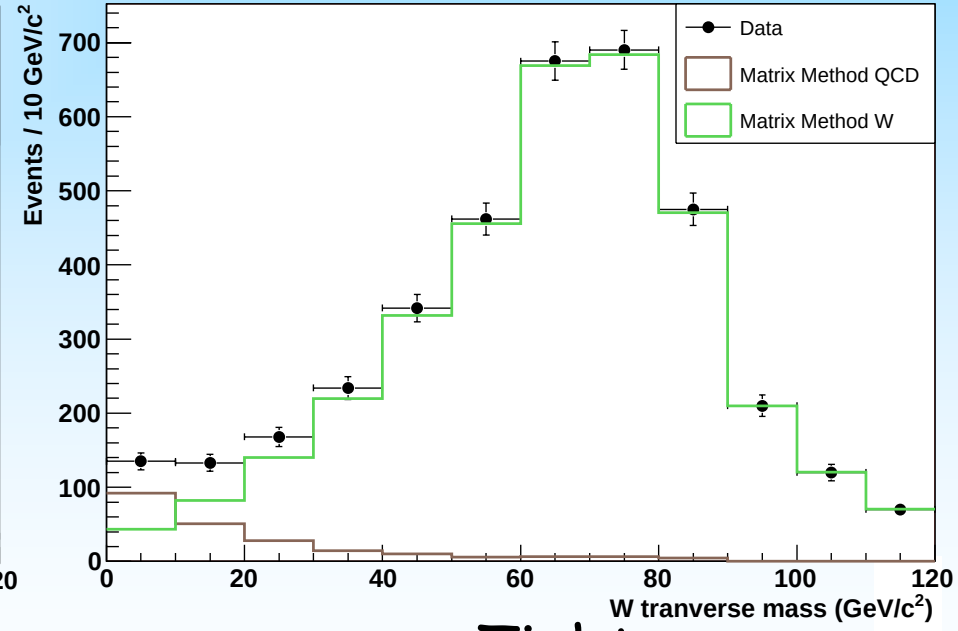
Multijet and W+jet normalization

W tranverse mass distribution split into :

- real lepton data (W+jet)
- fake lepton data (Multijet)



Loose



Tight

Summary

Several uses of real data to « fix » MC/data discrepencies

-Degrading MC resolution : Smearing
-Correcting selection cut efficiencies

- Scaling factors
- Event weighting

When using these methods

- Taking care not to introduce bias in the analysis

- Systematics associated to every correction have to be carefully estimated.

- Weighting procedure may need extensive cross-checks.

And try to **fix discrepencies at the source**

A bit of maths

Assuming s independent data samples containing b different signal or backgrounds, one can write for each sample j (j=1..s) an equation:

$$1 = \sum_{i=1}^{s} n_i^j$$

with n^j_i: fraction of signal i in sample j.

Moreover supposing t different selection criteria (cuts on various variables), one can write for each sample j (j=1..s) and each criteria k (k=1..t) another equation:

$$q_j^k = \sum_{i=1}^{s} \varepsilon_i^k n_i^j$$

with ε^k_i: efficiency of cut k on signal i.

q^k_j: fraction of remaining events in sample j after cut k.

A bit more

If the selection criteria are **independent**, the efficiencies of applying x criteria $k_{i,i=1..x}$ can be written :

$$\varepsilon_i^{k_1..k_x} = \prod_{u=1}^{x} \varepsilon_i^{k_u}$$

And then, new equations applying several criteria :

$$q_j^{k_1..k_x} = \sum_{i=1}^{s} (n_i^j \prod_{u=1}^{x} \varepsilon_i^{k_u})$$

And finally

The various q_j are easily mesured on data.

Selections and samples bring :

$2^t.s$ equations,

$b.(s+t)$ unkowns (fractions and efficiencies).

You need as many equations as unknowns :

$$2^t.d = b.(d+t)$$

Simplest (???) solutions are :

(a) $s = 1$, $t = 3$, $b = 2$ and (b) $s = 2$, $t = 2$, $b = 2$.

=> 8 equations, 8 unknowns : System8.