Background studies for Angular Coefficient Analysis

Valerio Bertacchi
Lorenzo Bianchini
Elisabetta Manca
Gigi Rolandi
Suvankar Roy Chowdhury

Kick-off meeting
PRIN 2017F28R78
7 February 2020

INFN

SCUOLA NORMALE SUPERIORE

CMS Compact Muon Solenoid

Outlook

- Background of the analysis
- «ABCD» method
- Preliminary results
- Systematic uncertainty treatment
- Future upgrades

Background of angular coefficient analysis

Signal:

Events with exactly one Muon with:

- p_T>25 GeV, |η|<2.4
- Small impact parameter
- Isolated
- M_T>40 GeV

EW background:

Prompt muons from EW processes:

- Drell-Yan events with a missed muon
- Top events (single or tt), via $t \rightarrow W\ (\rightarrow\ \mu\nu)b$
- Diboson events (WW,WZ,ZZ), with a muonic decay
- $Z \rightarrow\ \bar{\tau}\tau$, W $\rightarrow\ \tau\nu$

Evaluated using MC

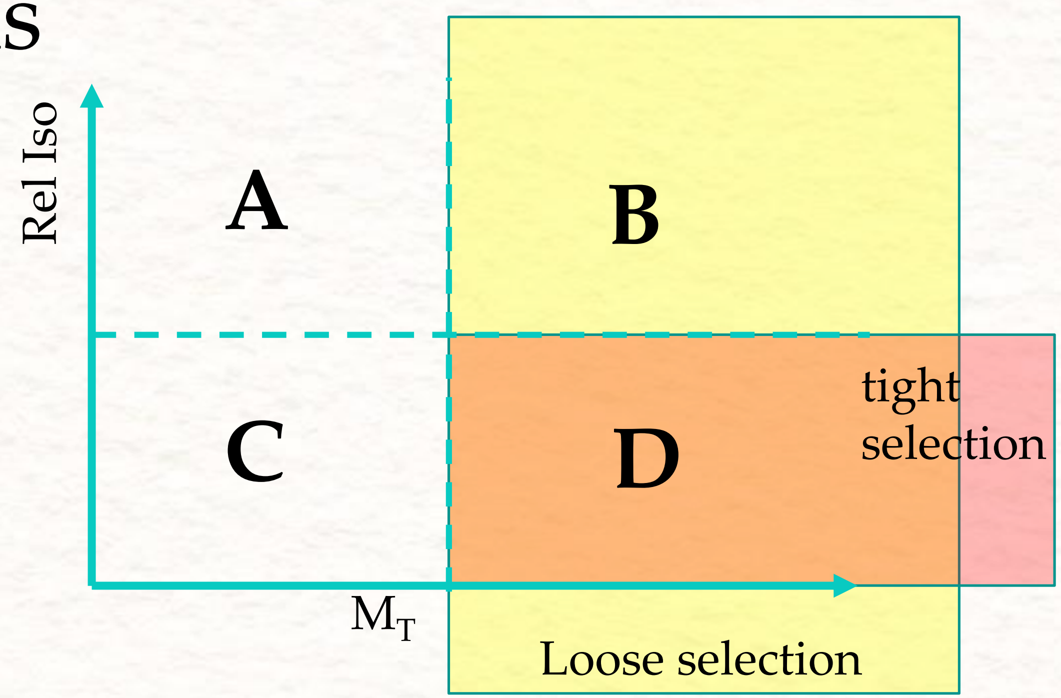
QCD background:

High-energy muons from multijet production, isolated “by chance”

- Muonic decays of heavy flavor mesons

Data-Driven estimation method: “ABCD”

ABCD method – definitions



- Define the four region based on 2 cuts:
  - Loose cut (high Mt): B+D region
  - Tight cut (loose AND isolated): D region

- The D region is the signal region, most of W are there

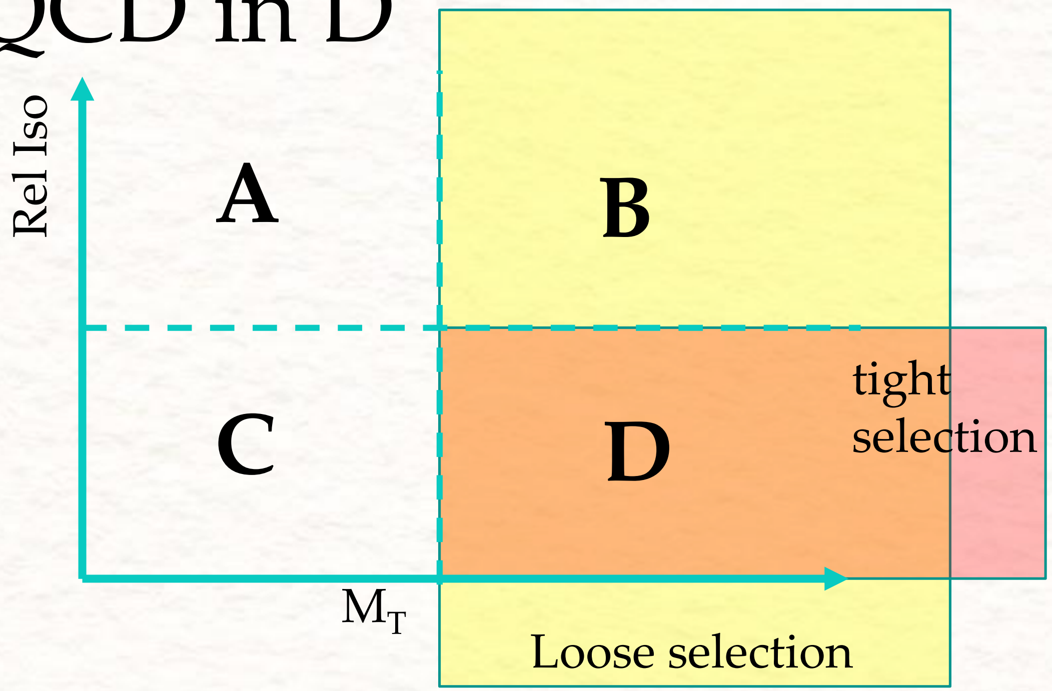
Prompt rate: $p = \frac{D}{D+B}\Big|_{EW}$ (Ideally 1)

- How large is the probability that a QCD event that pass the loose selection, pass also the tight one?

Fake rate: $f = \frac{D}{D+B}\Big|_{QCD}$ (Ideally 0)

ABCD method – measure of QCD in D

- Goal: measure QCD in Signal region (D) using :
  - Data
  - EWK MC



- Simplified case: assume p=1

$$D_{\mathrm{QCD}} = (B+D)_{\mathrm{QCD}} \cdot f = \frac{f}{1-f} B_{\mathrm{QCD}} = \frac{f}{1-f} B_{\mathrm{data}}$$

Prompt rate: $p = \frac{D}{D+B}\Big|_{EW}$

Fake rate: $f = \frac{D}{D+B}\Big|_{QCD}$

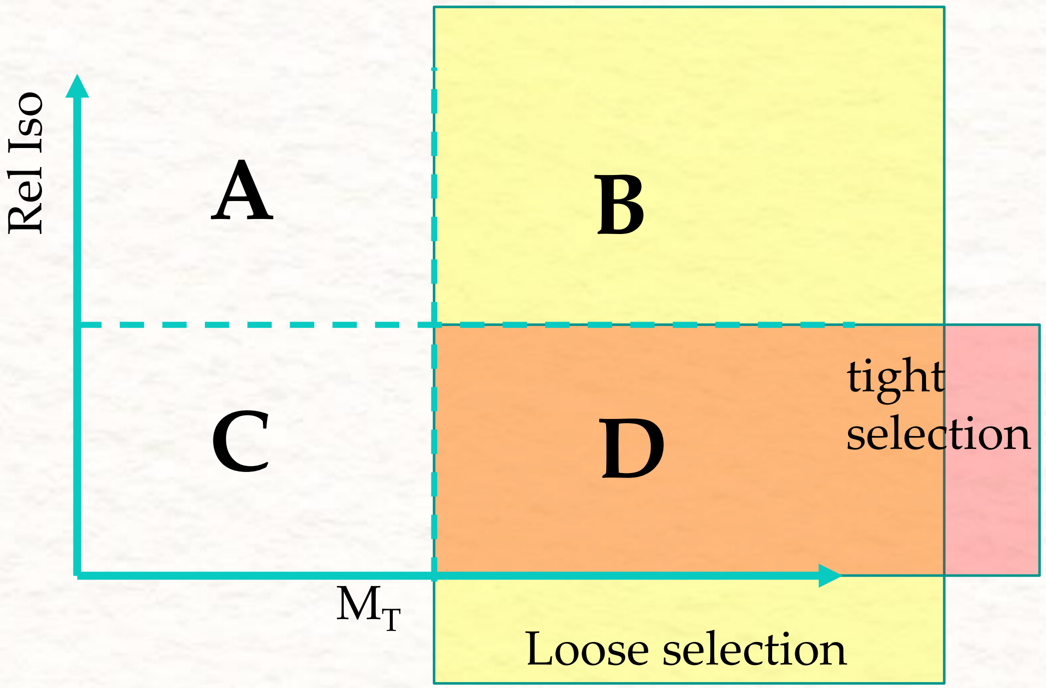
- Real case: p<1→also W in B region

$$D_{\mathrm{QCD}} = \frac{f}{p-f}\left[pB_{\mathrm{data}} - (1-p)D_{\mathrm{data}}\right]$$

(see backup for details)

ABCD method - workflow

1. Measure **p** and **f** *(described in following slides)*
2. Bin in η and p_T of the muon
3. Apply the loose selection only
4. For each data event weight the event if it falls in B or D region:
   - Anti-Iso weight (B): $\frac{fp}{p-f}$
   - Iso weight (D): $-\frac{f(1-p)}{p-f}$
5. The distribution of the weighted events is the number of QCD events in signal region (D), in bin of η and p_T



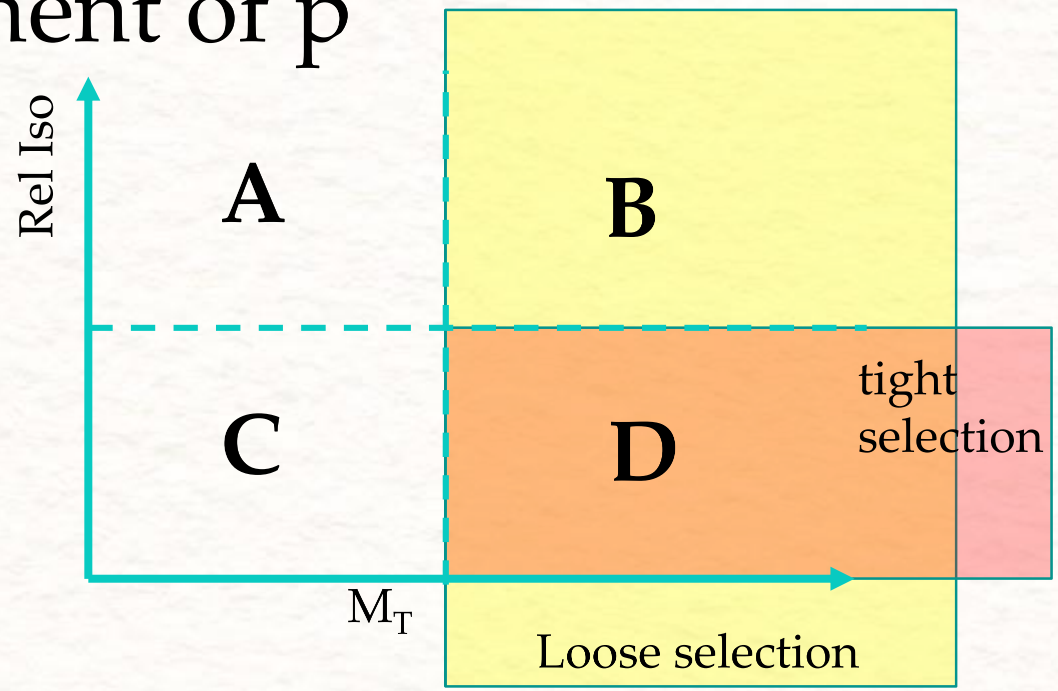
Prompt rate: $p = \frac{D}{D+B}\Big|_{EW}$

Fake rate: $f = \frac{D}{D+B}\Big|_{QCD}$

ABCD method – measurement of p

1. Bin in η and p_T of the muon
2. Loose selection (Mt>40)
3. From EWK MC events evaluate $p = \frac{D}{D+B}$
4. For each bin of η, fit p in function of p_T with and error function:

$$p_\eta(p_T) = A\frac{2}{\sqrt{\pi}}\int_0^{p_T} e^{-(Bx+C)^2}dx$$



$$\text{Prompt rate: } p = \frac{D}{D+B}\Big|_{EW}$$

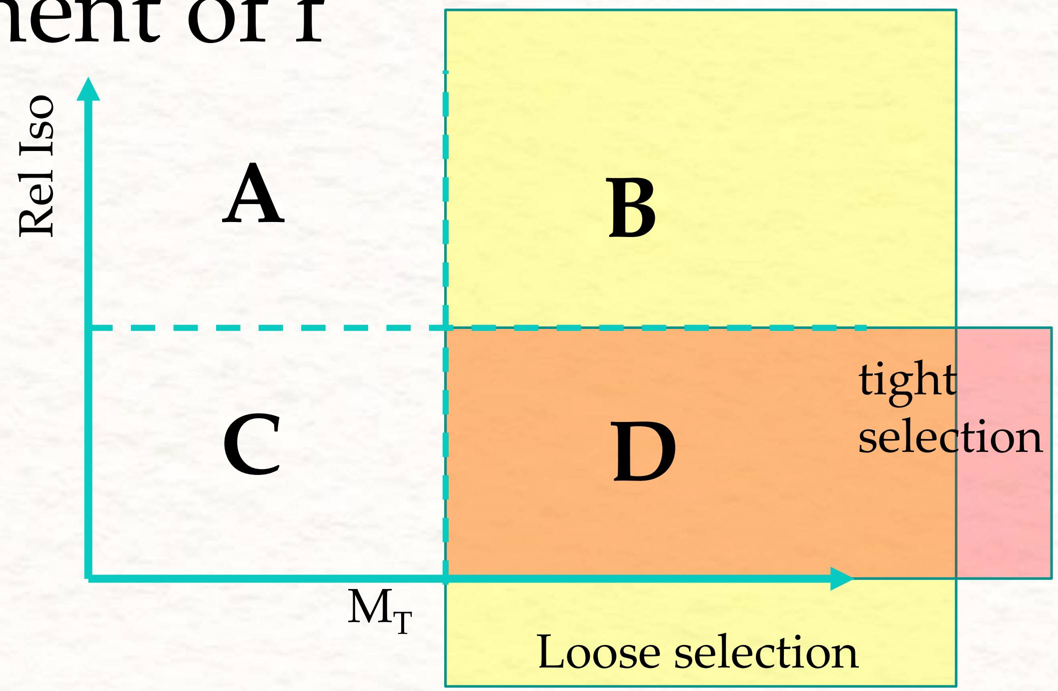
$$\text{Fake rate: } f = \frac{D}{D+B}\Big|_{QCD}$$

ABCD method – measurement of f

1. Bin in η and p_T of the muon
2. Anti-loose selection (M_T<40): **fake rate hypotesis**:

$$\frac{D}{D+B}\Big|_{QCD} = \frac{C}{A+C}\Big|_{QCD}$$

1. Subtraction of EWK to data: $\text{data} \Rightarrow \text{data} - \text{EWK}$
2. Evaluation of $f = \frac{C}{A+C}$ on this sample
3. For each bin of η, linear fit of f function of p_T



Prompt rate: $p = \frac{D}{D+B}\Big|_{EW}$

Fake rate: $f = \frac{D}{D+B}\Big|_{QCD}$

ABCD method – measurement of f

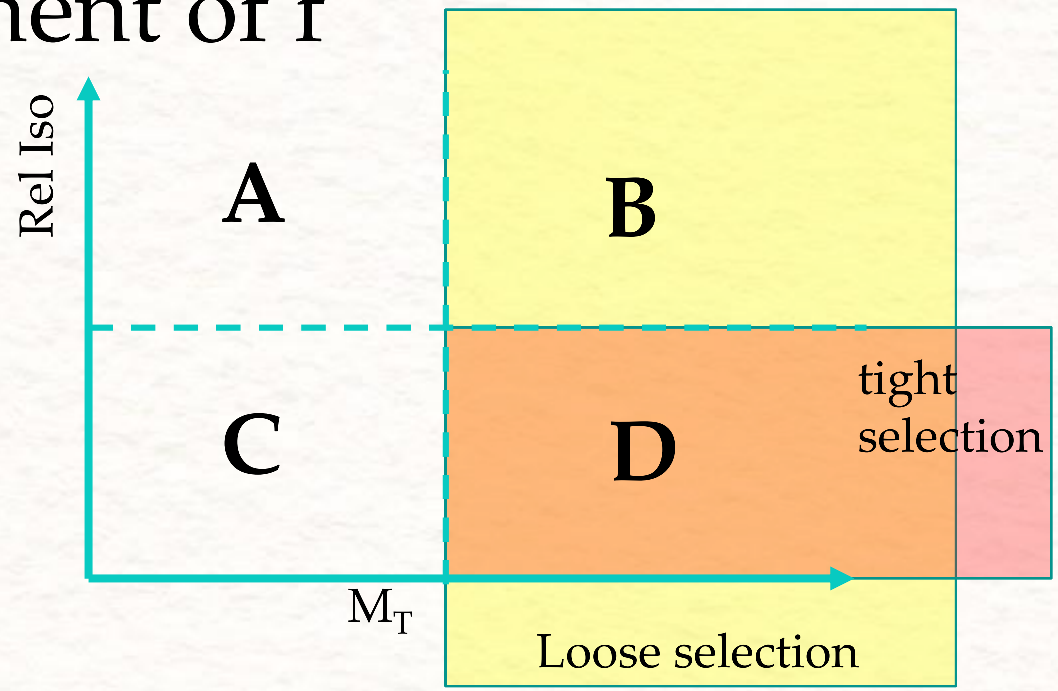
1. Bin in η and p_T of the muon
2. Anti-loose selection (M_T<40)**: fake rate hypotesis**:

$$\frac{D}{D+B}\Big|_{QCD} = \frac{C}{A+C}\Big|_{QCD}$$

3. Subtraction of EWK to data: $\text{data} \Rightarrow \text{data} - \text{EWK}$
4. Evaluation of $f = \frac{C}{A+C}$ on this sample
5. For each bin of η, linear fit of f function of p_T

Why is true?
- Because QCD isolation is not M_T-dependent

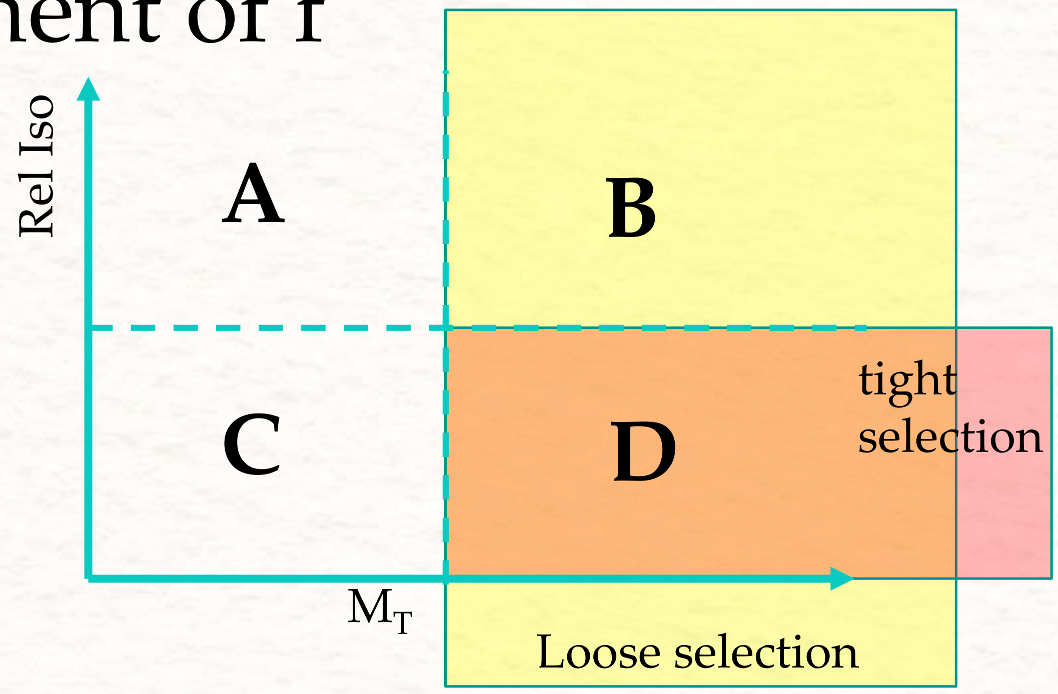
Why is needed?
- Because of the subtraction of point 3 (in signal region QCD<<EWK)



Prompt rate: $p = \frac{D}{D+B}\Big|_{EW}$

Fake rate: $f = \frac{D}{D+B}\Big|_{QCD}$

ABCD method – measurement of f



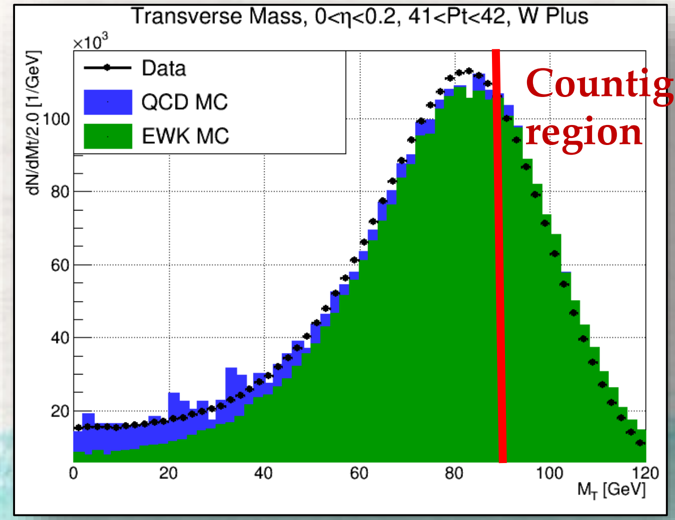
1. Bin in η and p_T of the muon
2. Anti-loose selection (M_T<40): **fake rate hypotesis**:

$$\frac{D}{D+B}\Big|_{QCD} = \frac{C}{A+C}\Big|_{QCD}$$

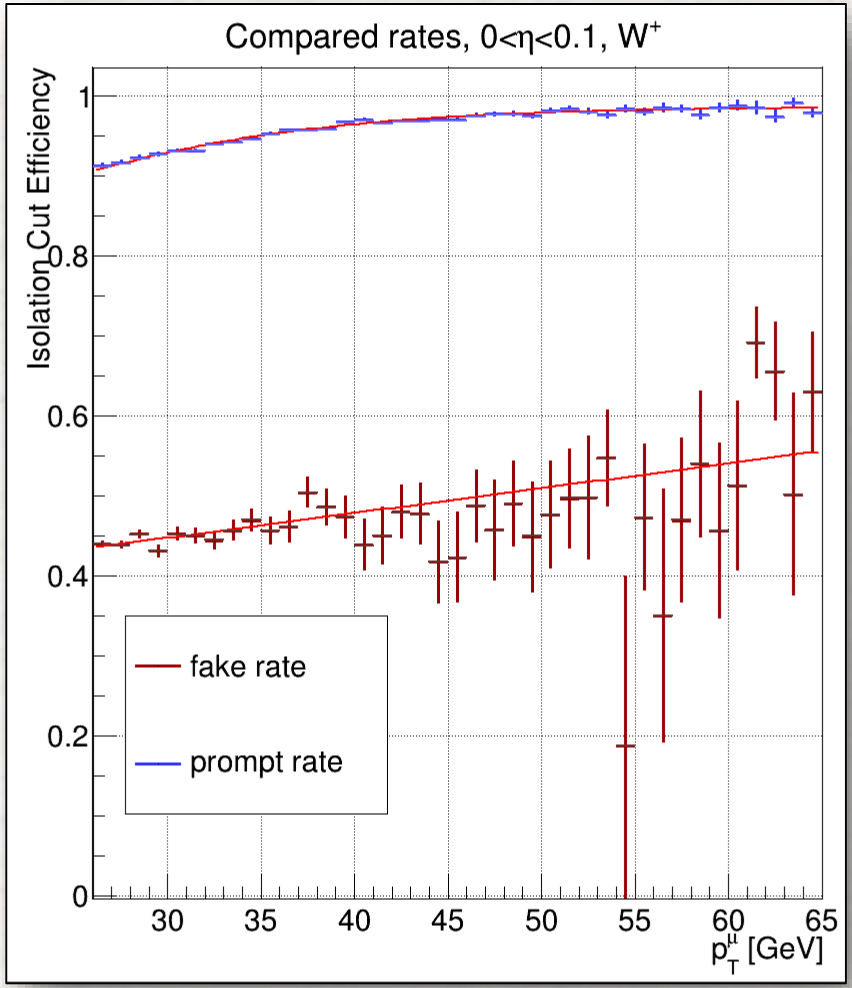
3. Subtraction of EWK to data: $\text{data} \Rightarrow \text{data} - \text{EWK}$
4. Evaluation of $f = \frac{C}{A+C}$ on this sample
5. For each bin of η, linear fit of f function of p_T

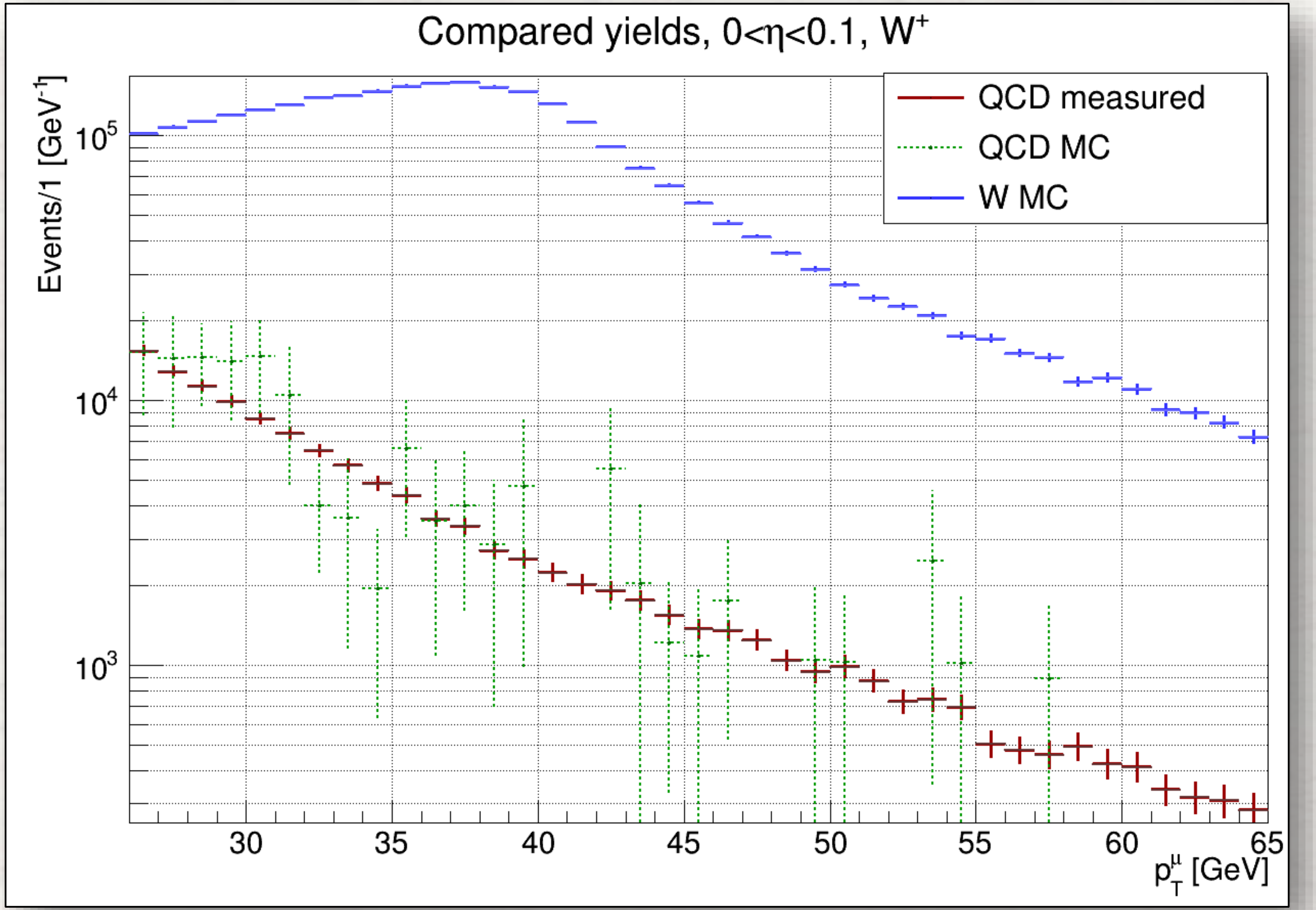
EWK MC must be fine tuned!
Developed and EWK Scale Factor:

- Cut at M_T>90
- $N_{EWK}^{data} = N^{data} - N^{MC}_{QCD}$
- $S(p_T, \eta) = N_{EWK}^{data} / N_{EWK}^{MC}$
- $\text{EWK} \Rightarrow S \times (EWK)$, $\quad \text{data} \to \text{data} - s \times (\text{EWK})$



Examples of preliminary results





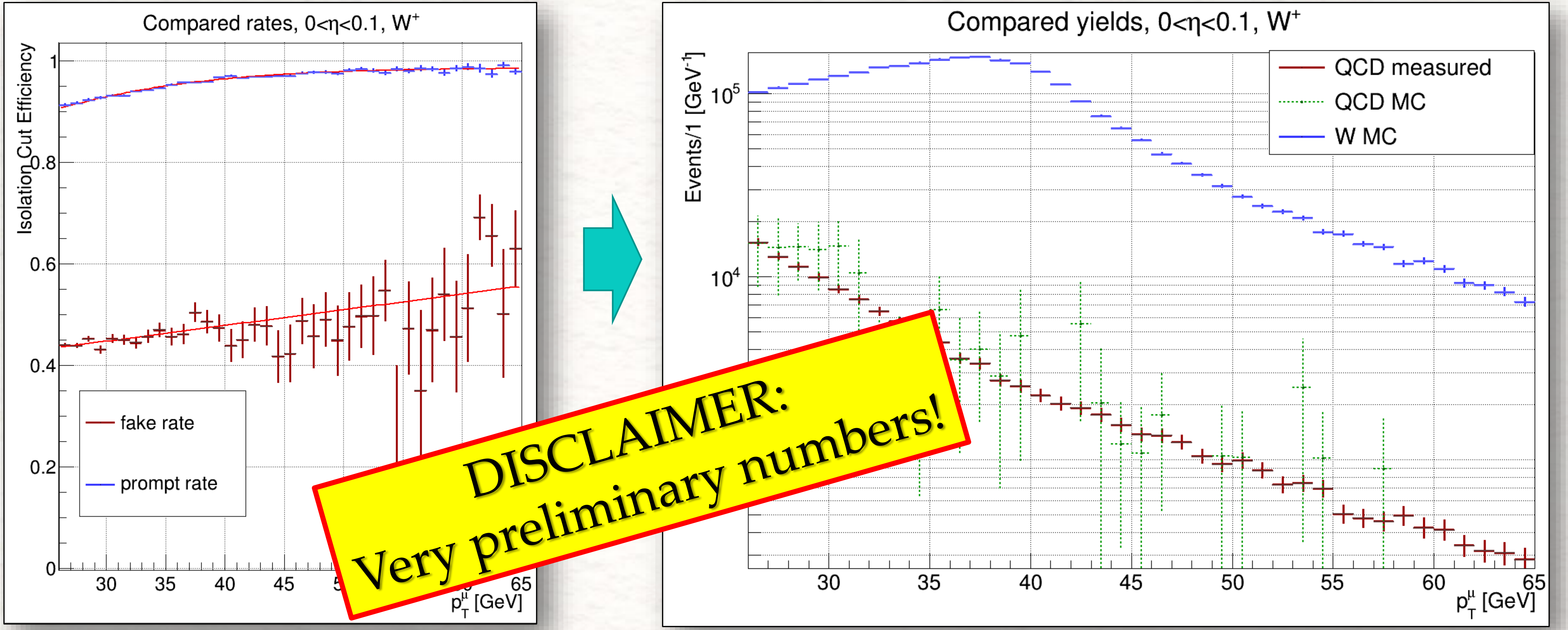
Fake rate resolution: $\frac{\sigma_f}{f} \simeq 1\% - 5\%$

Prompt rate resolution: $\frac{\sigma_p}{p} \simeq 0.1\% - 1\%$

QCD resolution: $\frac{\sigma_{N_{\rm QCD}}}{N_{\rm QCD}} \simeq 4\% - 15\%$

QCD uncertainty weight: $\frac{\sigma_{N_{\rm QCD}}}{\sigma_{N_{\rm W}}} = 5\% - 20\%$

Examples of preliminary results



Fake rate resolution: $\frac{\sigma_f}{f} \simeq 1\% - 5\%$

Prompt rate resolution: $\frac{\sigma_p}{p} \simeq 0.1\% - 1\%$

QCD resolution: $\frac{\sigma_{N_{\mathrm{QCD}}}}{N_{\mathrm{QCD}}} \simeq 4\% - 15\%$

QCD uncertainty weight: $\frac{\sigma_{N_{\mathrm{QCD}}}}{\sigma_{N_{\mathrm{W}}}} = 5\% - 20\%$

Systematic uncertainties treatment

- Variation of input vatiables (M_T,RelIso,Iso,P_T, η)
- 20 variation
  - SF stat and syst (Trigger, ID, Iso), (12)
  - PT Rochester Correction (2)
  - JER, JES, unclustered Energy (6)
- Run the analysis again for each variation and produce a:
  - «varied» prompt and fake rate (in bin of p_T and η)
  - «varied» QCD template

Advantages of these approach:

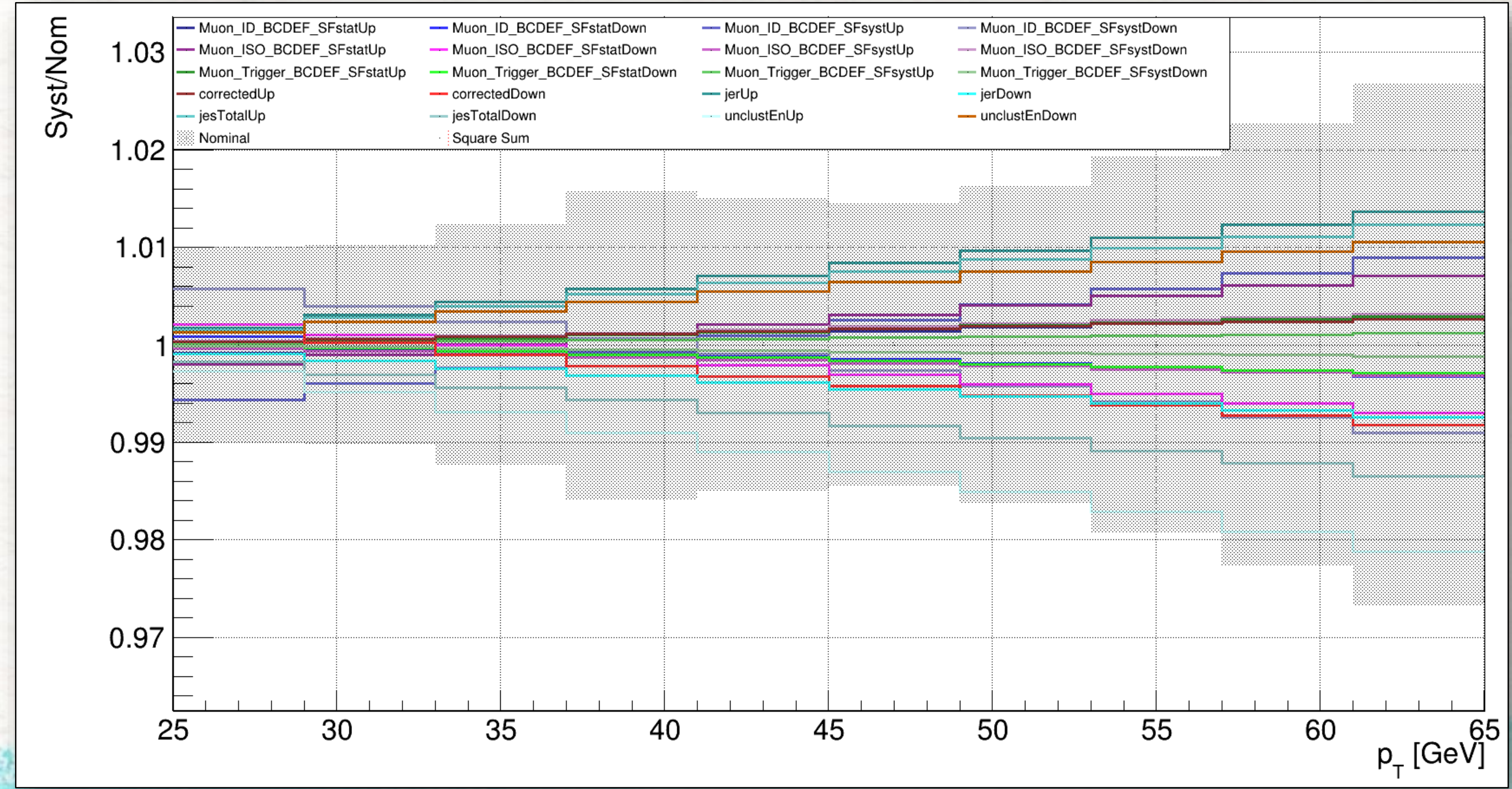
1. It is possible to follow the impact of the single source of systematic uncertainties

2. Provide a set of varied template comparable with the same variation in the rest of the analysis

3. *Extra*: the systematic variation can be used to describe nonlinearities of the fakerate *(see backup slides)*

Example of systematics analysis

Ratio varied/nominal QCD template, $0<\eta<0.4$, W^+



Conclusions

Current status:

- The strategy to extract the bkg template is ready
- The bkg template are ready to be tested

Next steps:

- Optimize the current implementation (Scale Factors, Cuts, Binning, Fit procedure, systematic approach)
- Do closures (on data) and robustness checks

Future Upgrade:

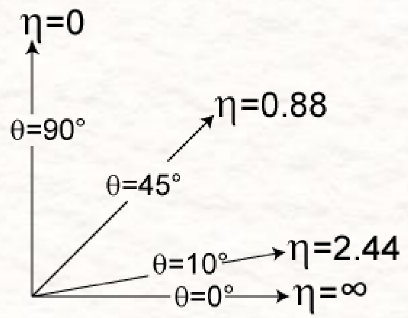
- Implement a simultaneous fit of fake rate and QCD yields in signal region, assuming prompt rate and data isolation efficiency only *(see backup)*

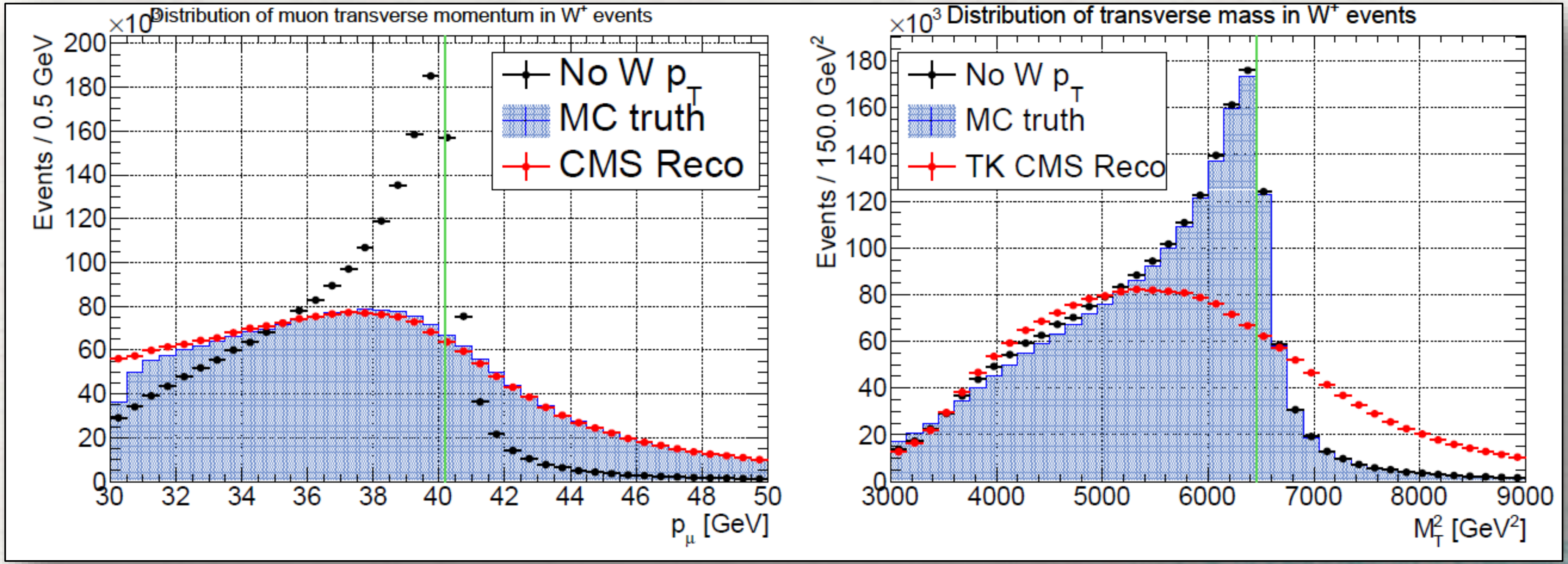
BACKUP SLIDES

Useful variables

Transverse Mass: $M_{\perp}^{2} = 2p_l^{\perp}p_{\nu}^{\perp}(1-\cos(\Delta\phi)) = 2(|\vec{p_l}^{\perp}||\vec{p_l}^{\perp}+\vec{h}^{\perp}|+(\vec{p_l}^{\perp})^2+\vec{p_l}^{\perp}\cdot\vec{h}^{\perp})$

Pseudorapidity: $\eta = -\ln(\tan(\frac{\theta}{2}))$





Event selection

```python
def fiducial_muon(mu):
    return (abs(mu.eta)<2.4 and mu.pt>10 and abs(mu.dxy)<0.05 and abs(mu.dz)<0.2)
def loose_muon_id(mu):
    return (fiducial_muon(mu) and mu.isPFcand and mu.pfRelIso04_all< 0.25 and mu.pt>10)
def medium_muon_id(mu):
    return (fiducial_muon(mu) and mu.mediumId and mu.pfRelIso04_all<=0.15 and mu.pt>20)
def medium_aiso_muon_id(mu):
    return (fiducial_muon(mu) and mu.mediumId and mu.pfRelIso04_all> 0.15 and mu.pt>20)
event_flag = -1
(idx1, idx2) = (-1, -1)
# Z-like event
if len(loose_muons)>=2:
    if len(loose_muons)==2:
        (idx1, idx2) = (loose_muons[0][1], loose_muons[1][1])
        event_flag = 2 if (loose_muons[0][0].charge+loose_muons[1][0].charge)==0 else 3
    else: event_flag = -1
# W-like event: 1 loose, 1 medium
elif len(medium_muons)==1:
    event_flag = 0
    (idx1, idx2) = (medium_muons[0][1], -1)
# Fake-like event
elif len(medium_muons)==0 and len(medium_aiso_muons)==1:
    event_flag = 1
    (idx1, idx2) = (medium_aiso_muons[0][1], -1)
# anything else
else:
    event_flag = -1
```

Isolation variable definition

- Most common isolation definition during Run-1 (*"δβ-corrected"*) is :

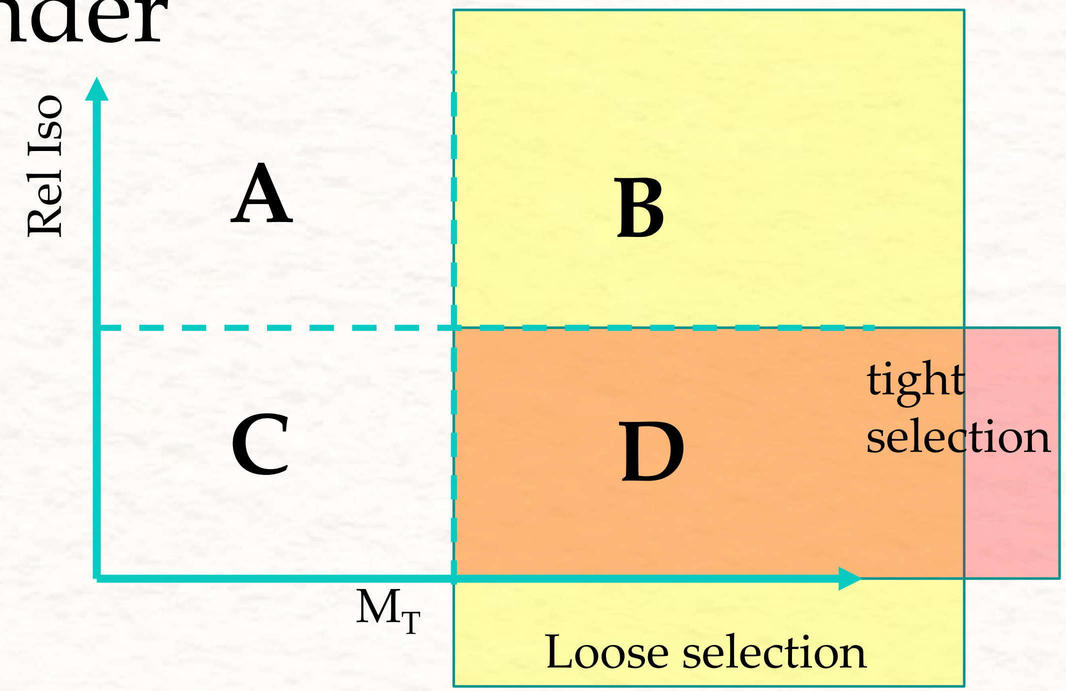
$$\mathrm{Iso}_{\delta\beta} = \Big(\sum_{CH} P_T + \max(\sum_{NH} P_T + \sum_{PH} P_T - 0.5 \sum_{PU} P_T, 0)\Big)/P_T^{\mu}$$

  - P_T sum of **charged hadrons** from LV arounds the muon
  - As **neutral component**, add following if it is grater than 0 :
    - "P_T sum of neutral hadrons and photons arounds the muon"
      minus "estimated contribution from PU neutral particles ($\delta\beta$)"
      - $\delta\beta$ = **0.5** x "PT sum of **charged hadrons from PU**"
        0.5 comes from the ratio between charged/neutral particles in isospin limit.

Background strategy reminder

Prompt rate: $p = \frac{D}{D+B}\Big|_{EW}$ Fake rate: $f = \frac{D}{D+B}\Big|_{QCD}$

QCD yeld in signal region: $D_{QCD} = \frac{f}{p-f}(pB - (1-p)D)$



Approach:

1. Measure p and f
2. Bin in η and p_T of the muon
3. Apply the loose selection
4. For each data event weight the event if it falls in B or D region
5. The distribution of the weighted events is the number of QCD events in signal region (D), in bin of η and p_T

Measurement of f

1. Binning in η of the muon and evaluation of EWK scale factor (high M_T counting)
2. Bin in η and p_T of the muon
3. Anti-loose selection (es M_T<40): **fake rate hypotesis:** $\frac{D}{D+B}\Big|_{QCD} = \frac{C}{A+C}\Big|_{QCD}$
4. Subtraction of EWK rewighted to data: $\text{data} \Rightarrow \text{data} - s(\text{EWK})$
5. Evaluation of f=C/(A+C) on this sample
6. For each bin of η, linear fit of f function of p_T

Measurement of p

1. Bin in η and p_T of the muon
2. Loose selection (es Mt>40)
3. From W MC evaluated p=D/(D+B) events
4. For each bin of η, linear fit of p function of p_T

Derivation of QCD yield formula

$A, B, C, D =$number of events in the region.
$q =$QCD, $q =$elwctroweak, $d =$Data

$$f = \frac{D_q}{D_q + B_q}, \quad \frac{D_w}{D_w + B_w}$$

$$\begin{aligned} D_q &= f(D+B)_q \\ &= f(D+B)_d - f(D+B_w) \\ &= f(D+B)_d - \frac{f}{p}D_w \\ &= f(D+B)_d - \frac{f}{p}(D_d - D_q), \end{aligned}$$

$$D_q(\frac{p-f}{p}) = f\left[B_d + D_d(\frac{p-1}{p})\right],$$

$$D_q = \frac{f}{p-f}\left[pB_d - (1-p)D_d\right]$$

ElectroWeak Scale Factor approach comparison

- We want to bin both in η and $_{pT}$
  1. **Very high MT counting**:
     - Cut at Mt>90
     - $N_{EWK}^{data}=N^{data}-N^{MC}_{QCD}$
     - $EWSF(p_T, \eta)=N_{EWK}^{data}/N_{EWK}^{MC}$
  2. **Strong Isolation counting**:
     - Cut at first isolation bin (RelIso<0.01)
     - $N_{EWK}^{data}=N^{data}-N^{MC}_{QCD}$
     - $EWSF(p_T, \eta)=N_{EWK}^{data}/N_{EWK}^{MC}$

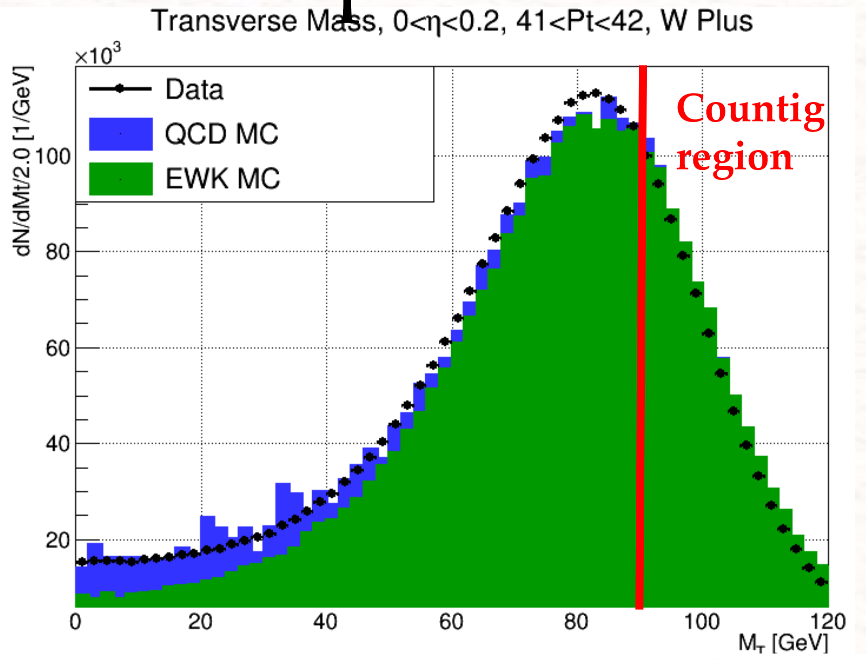
Pro: well calibrated

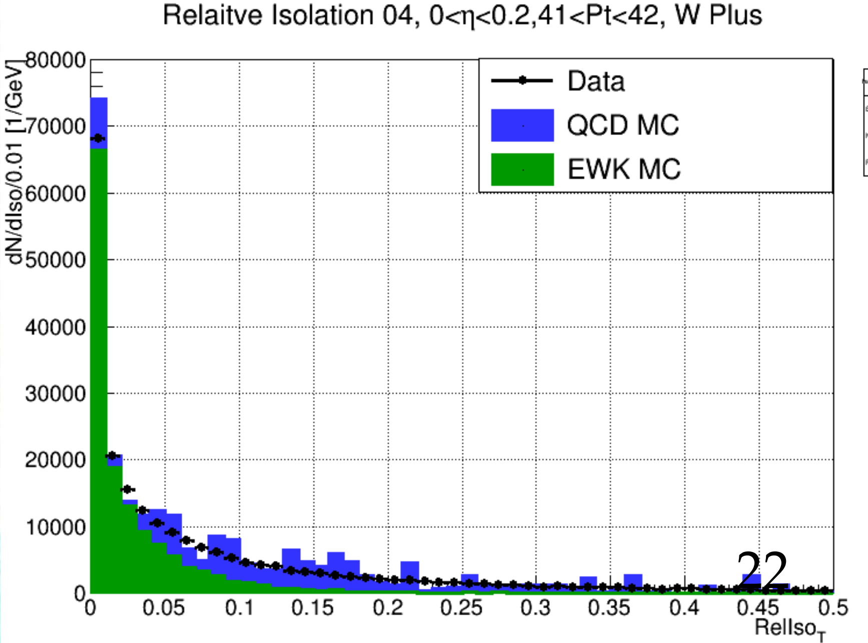
Cons: Add a cut in M_T (JER/JES dependent)

Pro: Indipendent from JER/JES

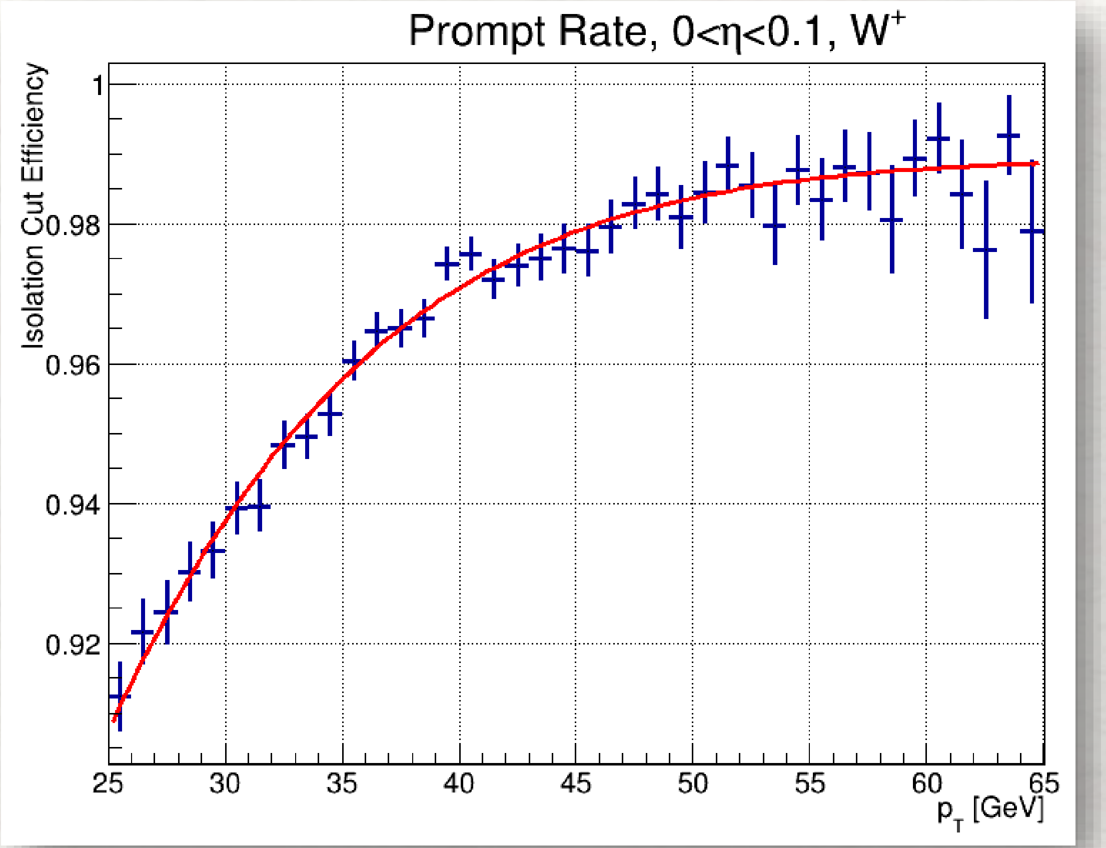
Cons: calibration of fist bin must be checked

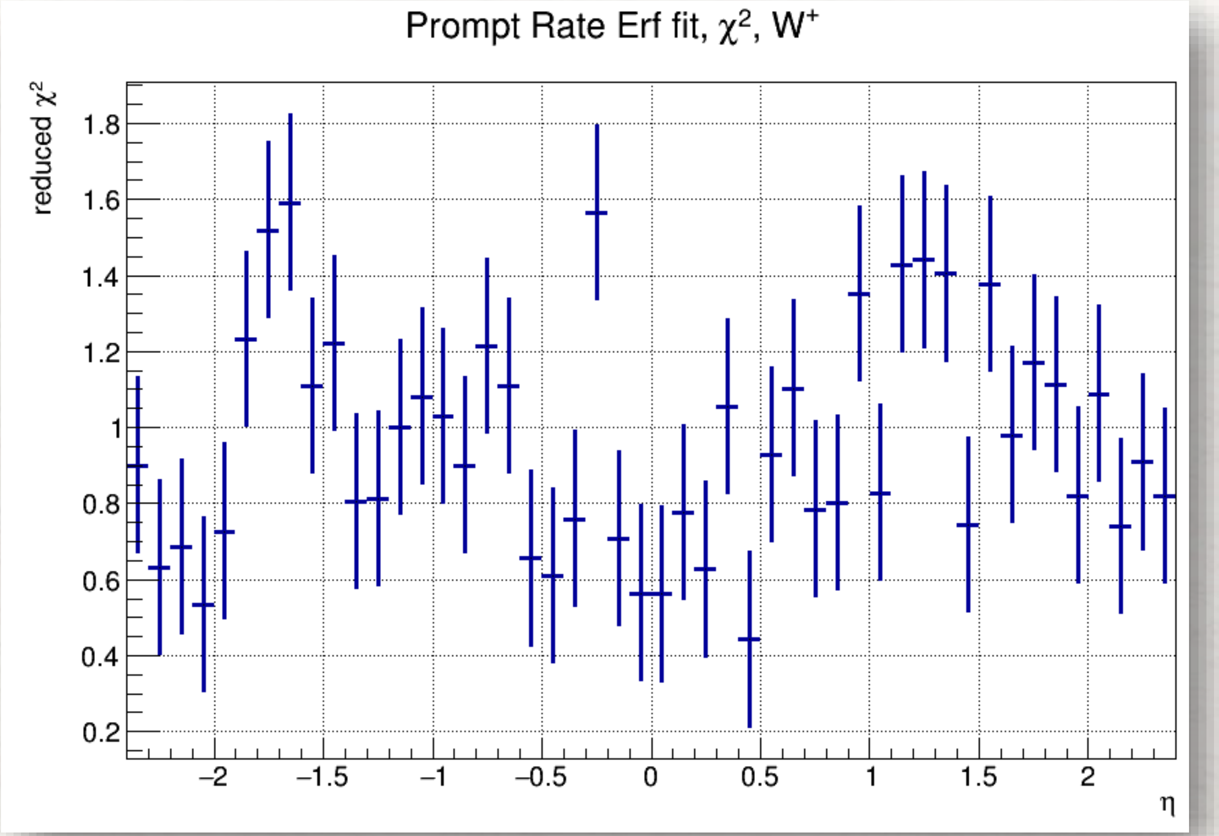
(see backup)





Prompt rate detailed plots





Systematics to describe nonlinearities: the *Correlated Fit*

The idea:

Taking into account p_T-correlation of several systematic variation to better descrive fakerate non-linearties

- Build a correlation matrix of the fake rate data with the bin-to-bin correlation in p_T
- Do a linear fit, for each bin of η, using the correlation matrix

Implementation:

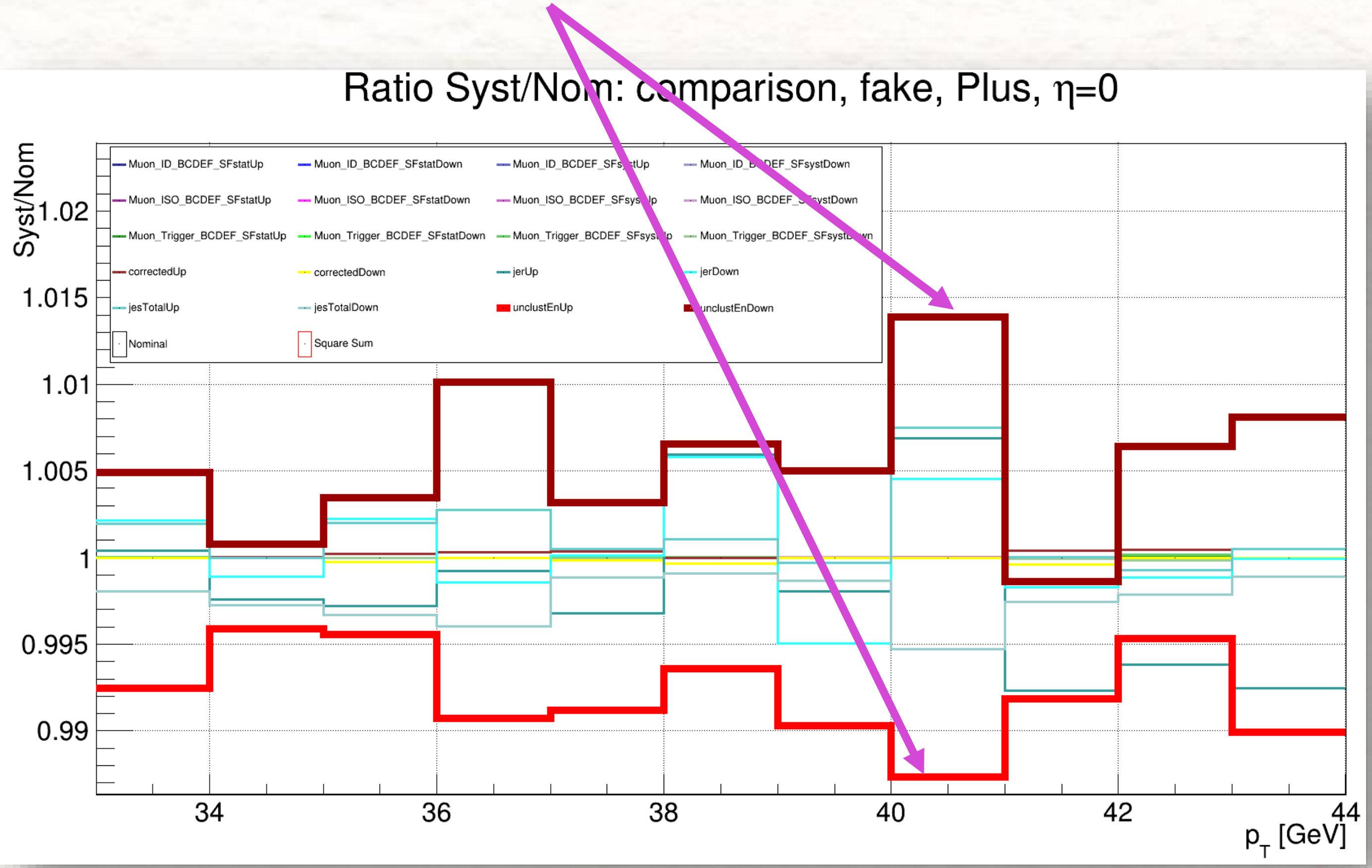
- correlation matrix: $\mathrm{C} = M_{\mathrm{stat}} + \sum_{s=\mathrm{syst0}}^{\mathrm{systM}} M_{\mathrm{s}}, \quad \mathrm{con:} \quad M_{\mathrm{stat}}^{i,j} = \delta_{i,j}\sigma_{i,j}^2, \quad M_{\mathrm{s}}^{i,j} = (f_i^{\mathrm{nom}} - f_i^s)(f_j^{\mathrm{nom}} - f_j^s) \quad \mathrm{e\ con}\ i,j \in \{p_T^0, \ldots, p_T^N\}$

- Simmetrization: $|f_i^{\mathrm{nom}} - f_i^s| = \frac{1}{2}(|f_i^{\mathrm{nom}} - f_i^{\mathrm{s,up}}| + |f_i^{\mathrm{nom}} - f_i^{\mathrm{s,down}}|) \qquad \mathrm{sign}(f_i^{\mathrm{nom}} - f_i^s) = \mathrm{sign}(f_i^{\mathrm{nom}} - f_i^{s,up})$

- Minimization: $\chi^2 = \vec{v}^T C^{-1} \vec{v}, \quad \vec{v} = p\vec{x} + q$

Systematics to describe nonlinearities: the *Correlated Fit*

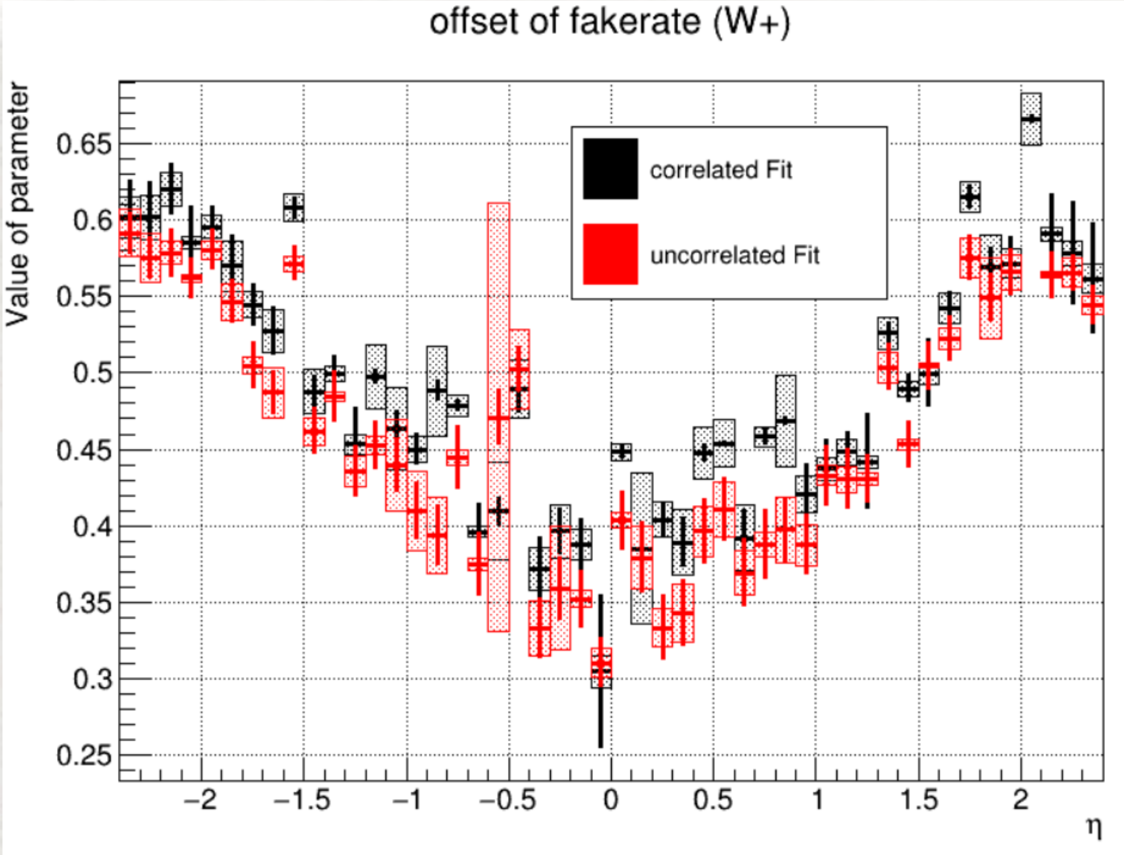
Example of the correlation in p_T bins in the fakerate with the systematic ratios

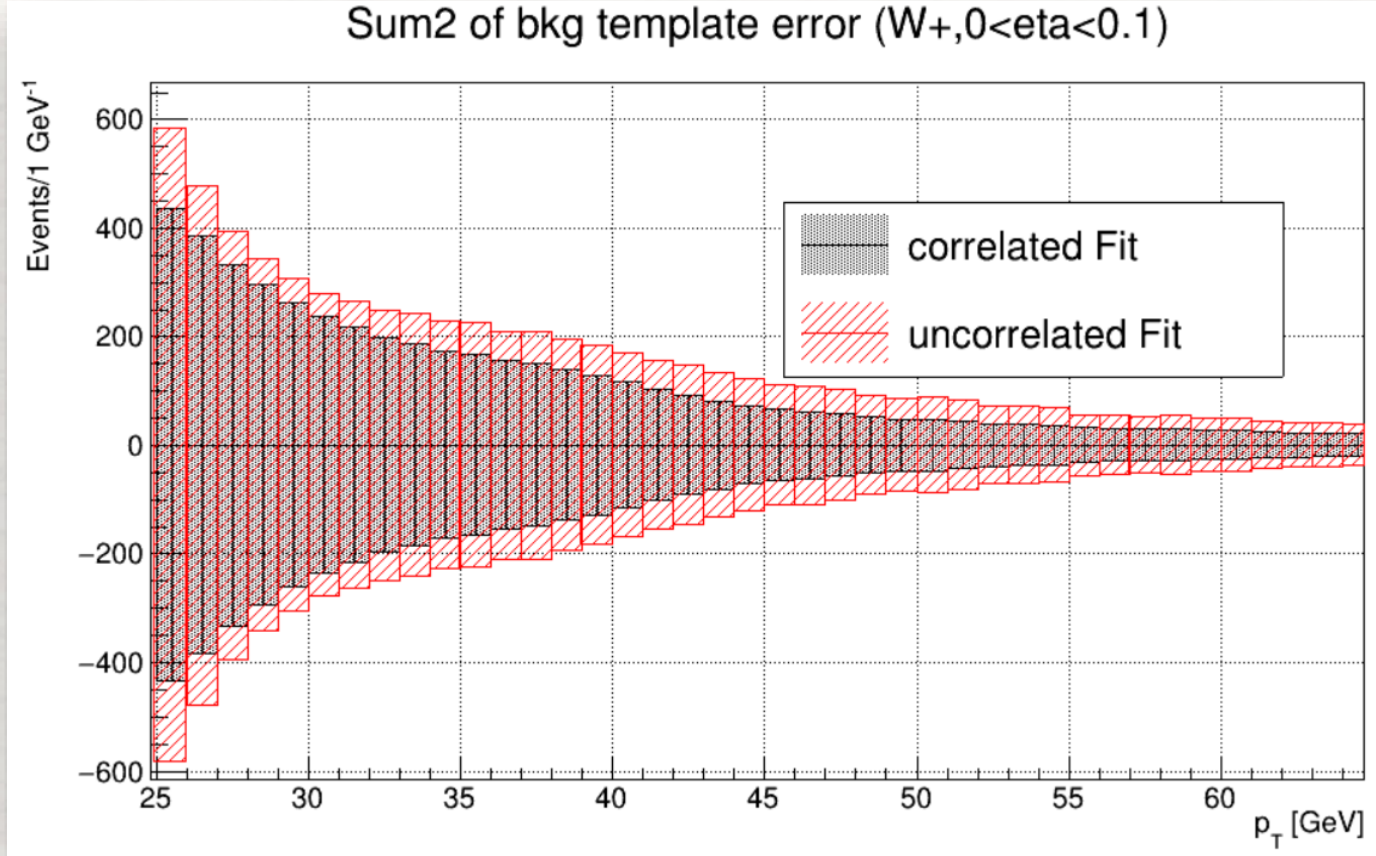


How to use the "correlated fit"?

- Using correlatedFit the "statistical" uncertainty on parameters contains also systematic variation uncertainty
- To properly evaluate the statistical uncertainty has been used 1k toys (independently for each η bin) varying the central values of nominal fake rate
- The correlatedFit has been repeated for each systematic variation, changing the central value only (same correlation matrix!)

Systematics to describe nonlinearities: the *Correlated Fit*





- Smaller statistical error
- Almost the same systematic error

Future Upgrade – Simultaneous Fit

The Idea

Using the p_T binning there are enough degree of freedom to fit together the normalization and the fake rate.

Advantages

- No EWKSF tuning
- No transfer between low M_T and high M_T region for fakerate

Implementation

- Each η bin a separate fit
- Free parameters: fakerate slope and offset, normalization of each eta bins for W and QCD (2+2N)
- Measurement (fixed parameters): data yield and isolation efficiency of data for each p_T bin (2N), prompt rate for each p_T bin (N)
- Degree of freedom: N-2≈37