Data Driven Background Estimation in $t\bar{t}$ Analyses at CMS

Holger Enderle

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

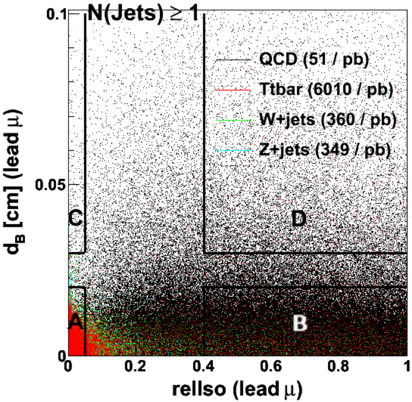
LHC Physics Discussions: Top Physics
15th November 2010

Outline

ABCD - Method

Idea of the method:

- Count QCD events in pure QCD control regions
- Translate into number of QCD events in signal region



Procedure:

- Find 2 uncorrelated variables providing pure QCD regions
  - example for semileptonic (μ)
  - relIso: relative isolation of muon
  - d_B: muon impact parameter
- Define 4 regions in the 2-dim space spanned by these variables
  - signal region A(final selection)
  - 3 QCD regions B,C,D

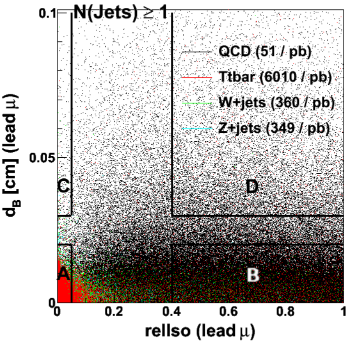
Formula:

- $\frac{N_A}{N_B} = \frac{N_C}{N_D} \Rightarrow N_A = N_B \cdot \frac{N_C}{N_D}$

ABCD - Uncertainties

Signal contamination in regions B,C,D:

- How large is the effect?
  - our example: B($\sim$ 0.2%) and C($\sim$ 2%)
  - $N'_A = 1.002 \cdot N_B \cdot \frac{1.02 \cdot N_C}{N_D} \approx 1.022 \cdot N_A$
  - results in: $\sim$ 2.2% higher QCD estimate
- Relative uncertainty under control
  - systematic overestimation
  - proportional to relative contamination



Correlated variables:

- More complicated
  - relation not necessarily satisfied
- Correlation can be estimated
  - determine $\frac{N_C}{N_D}$ in different slices of control region
  - extrapolate to signal region

Matrix Method

Idea of the method:

- Factorize efficiencies of real and fake leptons for some isolation criteria
- Knowing these efficiencies and varying the isolation criteria
  - define a system of 2 equations with 2 unknowns
  - estimate QCD-like background (no real isolated lepton)
- Two leptons in the dileptonic case
  - define a system of 3 equations with 3 unknowns
  - estimate QCD- and W+jets-like background at the same time

System of equations:

- N^l – relaxed isolation criteria for both leptons
- N^m – at least one lepton fulfills tight isolation criteria
- N^t – tight isolation criteria for both leptons

$$
\begin{aligned}
N^l &= N^l_\mathrm{S} + N^l_\mathrm{W} + N^l_\mathrm{QCD} \\
N^m &= (2\epsilon_\mathrm{real} - \epsilon^2_\mathrm{real})N^l_\mathrm{S} + (\epsilon_\mathrm{real} + \epsilon_\mathrm{fake} - \epsilon_\mathrm{real}\epsilon_\mathrm{fake})N^l_\mathrm{W} + (2\epsilon_\mathrm{fake} - \epsilon^2_\mathrm{fake})N^l_\mathrm{QCD} \\
N^t &= \epsilon^2_\mathrm{real} N^l_\mathrm{S} + \epsilon_\mathrm{real}\epsilon_\mathrm{fake} N^l_\mathrm{W} + \epsilon^2_\mathrm{fake} N^l_\mathrm{QCD}
\end{aligned}
$$

Matrix Method - Uncertainties

- ABCD method is essentially a simplified version of matrix method
  - $N^{t}_{\mathrm{QCD}} = \epsilon_{fake} \cdot N^{l}_{\mathrm{QCD}}$
  - $N_A = N^{t}_{\mathrm{QCD}}, \quad N_C / N_D = \epsilon_{fake}, \quad N_B = N^{l}_{\mathrm{QCD}}$
- Advantages of matrix method
  - accounts for signal contamination
  - calculates number of signal and background events at the same time
- Disadvantage of matrix method
  - need to know efficiencies of real and fake leptons

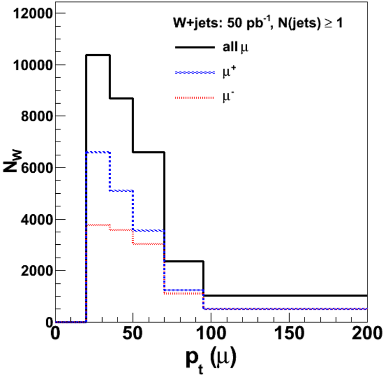
- Main difficulty and largest source of uncertainty
  - evaluation and parametrisation of efficiencies
    - ★ relative uncertainty on electron efficiency $\sigma(\epsilon_{\mathrm{real}}) \sim 5\%$
    - ★ relative uncertainty on fake rate $\sigma(\epsilon_{\mathrm{fake}}) \sim 10\%$
- Study with $100\,\mathrm{pb}^{-1}$ (dileptonic)
  - signal-like events: $149 \pm 13(\mathrm{stat}) \pm 1(\mathrm{syst})$ (9% comb uncert)
  - W+jets-like events: $5.34 \pm 0.69(\mathrm{stat}) \pm 0.91(\mathrm{syst})$ (21% comb uncert)
  - QCD-like events: negligible

Charge Asymmetry Method

Idea of the method:

- pp-collider: $N(W^+) > N(W^-)$
- W+jets: $N(\mu^+) > N(\mu^-)$
- $t\bar{t}$: $N(\mu^+) = N(\mu^-)$

▶ Calculate $N_{\mathrm{W+jets}}$ from charge asymmetry of isolated muons



Formula:

- Define asymmetry parameter: $R = \frac{(N^+ + N^-)_{\mathrm{W+jets}}}{(N^+ - N^-)_{\mathrm{W+jets}}}$ (from MC)
- $(N^+ - N^-)_{\mathrm{W+jets}} = (N^+ - N^-)_{\mathrm{all\ data}}$

▶ $N_{\mathrm{W+jets}} = R \cdot (N^+ - N^-)_{\mathrm{all\ data}}$

Charge Asymmetry - Uncertainties

- Uncertainty on the charge Asymmetry in the simulation
  - dominated by PDF uncertainties
- Other events leading to charge asymmetry
  - e.g. single top

► Systematic uncertainties are in the order of 10%

- With 50 pb^{-1} statistical uncertainties dominate
  - combined relative uncertainty: $\sim$ 40%
- With 1 fb^{-1} systematic uncertainties start to dominate
  - combined relative uncertainty: $\sim$ 15%

- Completely different approach than all other methods

► Valuable cross check

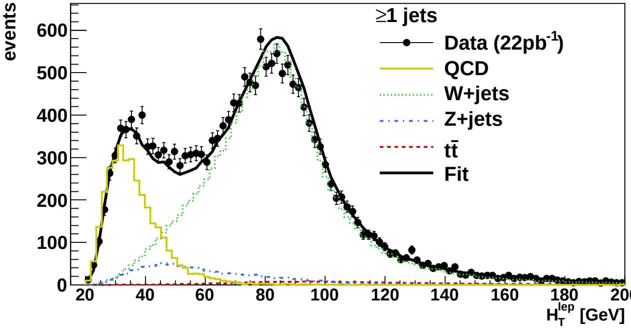
Template Fit

Idea of the method:

- Find variable with different shapes for signal and background distribution
  - determine template for each distribution
- Estimate event composition by varying the template scales to fit the data

Example for QCD estimation in semileptonic $t\bar{t}$ channel:

- $H_T^{\mathrm{lep}} = p_T(\mu) + \mathrm{MET}$



- Main difficulty
  - getting reliable background shapes

Template Fit - Distribution Shapes

V+jets and signal:

- Trust simulation and simply take simulated distributions

QCD:

- Trust shape of QCD distribution?
- Cross check with distribution from control region in data
  - same problems like in ABCD method
    - signal contamination and correlation
  - signal contamination results in overestimation of background
  - correlation can be estimated
    - determine template from different slices of control region
- Quality of the fit can be used to validate shapes

Summary

Background estimation is most important systematic uncertainty

- top cross section measurement
- top mass determination

Different methods for data driven background estimation are available:

- ABCD method - mainly QCD estimation
- Matrix method - all backgrounds with different number of isolated leptons
- Charge Asymmetry - charge asymmetric backgrounds like W+jets
- Template Fits - all backgrounds

► Cross checks between methods possible

Should get reasonable results already for the full 2010 data set

- Statistical uncertainties still dominate