Data Driven Background Estimation

Christian Sander (Universität Hamburg)

Hamburg – SFB Block Meeting - 23rd February 2012
Outline

• Background from Monte Carlo
• “Bump hunts”
• Motivation for data driven backgrounds
• Control sample weighting
  • Factorization / “ABCD” Methods
• Control events weighting
  • Inversion techniques
  • Fake methods
  • Replacement/Removal techniques
• QCD Background – Rebalance and Smear
• Application: hadronic SUSY search with jets and MET
Monte Carlo (event generation + full detector simulation) widely used in HEP experiments

- Excellent agreement for most of the observables
- Evaluation of systematic uncertainties
  - Scale for Data/MC differences (trigger, efficiencies ...)
  - Simulation with different settings (scale variations, pdf, fragmentation model ...)
  - Systematic variation of properties of reconstructed particles (jet energy scale uncertainties ...)

→ Valid approach for final states which are well understood
Scaled Monte Carlo Simulation

- If shape of simulated distribution is well simulated
  → Use background dominated region to normalize prediction

  **Example:** Normalize DY background in invariant mass window around $Z$ peak

- Don't forget about remaining contributions from other SM processes

- **Important:**
  Assign uncertainty to model assumption:
  
  "Same shapes in data and simulation?"

**Example:** Search for ADD models
Invariant $e^+e^-$ mass

**CMS PAS EXO-11-087**
“Bump Hunts”

- Search for narrow resonances over smooth background (shape not well predicted by theory)
- Limit depends on model assumption (fit function)
  - Statistical uncertainties from fit variance
  - Systematic uncertainties on fit function is challenging to assign, if function is not based on physics principles

→ **Possibility**: validate in control region, where no excess is expected (not always possible).

**Example**: Search for resonances in di-jet mass

CMS PAS EXO-11-015
Data Driven Methods in MET Searches

Searches for NP → excess in MET tails

→ Plain simulation at the limit

- SM backgrounds with large MET:
  - Backgrounds from out-of-acceptance, isolation, or reconstruction
    example: \((W \rightarrow e/\mu+\nu)+\text{jets}\)
  - Fake backgrounds
    example: \((W \rightarrow \tau_{\text{had}}+\nu)+\text{jets}\)
  - Irreducible backgrounds
    example: \((Z \rightarrow \nu\nu)+\text{jets}\)
  - Mismeasurements
    example: QCD multi-jet production

→ Each bg requires special methods

Example: Search for supersymmetry in all-hadronic events with missing energy

CMS PAS SUS-11-004
Factorization / “ABCD”-Method

- If search region is defined by sequential cuts, e.g. on \( \text{var}_1 \) and \( \text{var}_2 \) (with discriminative power, e.g. MET and HT)

\[
A = (*) \quad \text{D = signal region}
\]

\[
B = (*) \quad \text{C = bg dominated control region (*)}
\]

- If variables are uncorrelated (has to be verified)
  \[\text{shape of var}_1 \text{ distribution independent of choice of var}_2\]
  \[\text{background in signal region predicted by scaling of control sample}\]

\[N_D = N_C \cdot N_A / N_B\]

- Modification for correlated variables possible, however challenging!
Inversion Techniques

Example: full-hadronic search requires explicitly no isolated lepton! Events with leptons contribute to bg, because not reconstructed, not isolated or out of acceptance!

- Select events with exactly one isolated lepton
- **Reweight events** according to isolation and reconstruction efficiencies (probabilities that a lepton is reconstructed or isolated)
  - Efficiencies depend on event kinematics (e.g. high jet multiplicity → low isolation efficiency)
  - Obtain efficiencies from data in sufficient binning in relevant variables (e.g. distance of lepton to nearest jet)
- **Apply acceptance correction from MC**
Tag and Probe – Efficiencies from Data

Tag: well defined muon
Probe: loose definition of muon

\[ m_{\mu\mu} \text{ compatible with } Z \text{ mass } \rightarrow \text{ Probe is muon} \]

\[ \frac{\text{pass}}{\text{pass} + \text{fail}} \text{ ratio of probe is efficiency} \]

Muon \( p_T = 27.3, 20.5 \text{ GeV/c} \)
Inv. mass = 85.5 GeV/c^2
Fake Techniques

**Example: full-hadronic search requires explicitly no isolated lepton! Events with leptons contribute to bg, because of taus decaying hadronically and thus faking a jet!**

- Select a control sample with exactly one isolated light lepton (e.g. muon)

- **Weight events** according to muon reconstruction and isolation efficiencies

- Use universality of leptons
  → same number of taus

- **Weight events** according to hadronic branching fraction and tau reconstruction efficiencies

- **Replace muon by hadronic tau-jet** (draw random $p_T$ from templates) and recalculate MHT, HT ...

![CMS Simulation](image)
Replacement/Removal Techniques

**Example:** irreducible background to full-hadronic search is \( (Z \rightarrow \nu\nu) + \text{jets} \!

- Select events with same kinematics but light leptons instead of neutrinos

\[
\sigma \cdot \text{Br}(Z \rightarrow \nu\nu) = \frac{\text{Br}(Z \rightarrow \nu\nu)}{\text{Br}(Z \rightarrow e^+e^-)} \cdot \frac{N_{Z}^{\text{observed}} - N_{Z}^{\text{background}}}{\text{acceptance}_Z \cdot \text{efficiency}_Z \cdot \int L dt}
\]

- Remove leptons from event
- Recalculate MHT
- Alternatively one can use \( W + \text{jets} \) or \( \text{photon} + \text{jets} \) events
  - Similar behavior at high transverse momentum of the boson
  - **Advantage:** higher statistics
  - **Disadvantage:** systematic uncertainties more difficult to estimate
Rebalance and Smear

- Large uncertainties in QCD multi jet prediction and detector simulation of tails
  → **Obtain seed sample directly from data**, by stripping off transverse momentum imbalance by kinematic fit (method intrinsically safe against non-QCD contributions)
  → **Smear $p_T$ of rebalanced jets by jet response** (corrected for data/MC differences)
  → **Smeared sample describes full kinematic properties of QCD events**
All data driven methods have to be validated on simulated samples! → “Closure Tests”

→ Good performance on simulated events (here: QCD Pythia)
Results – Hadronic SUSY Search

- Put together all data driven background predictions

→ Good agreement with data! → Set very sensitive limits!

CMS preliminary, \( L = 1.1 \text{ fb}^{-1}, \sqrt{s} = 7 \text{ TeV} \)

- Data
- Bkg. predicted from data:
  - \( Z \rightarrow \nu \bar{\nu} + \text{jets} \)
  - \( W/\tau \rightarrow \tau_h + X \)
  - \( W/\tau \rightarrow e/\mu + X \)
  - QCD

\( m_{1/2} \) vs. \( m_0 \)

CMS Preliminary

\( L_{\text{int}} = 1.1 \text{ fb}^{-1}, \sqrt{s} = 7 \text{ TeV} \)

- Observed
- Expected ± 1σ

CMS PAS SUS-11-004
Summary

• Data driven methods are crucial for many searches for which the uncertainties from simulation are large or simply not known

• Often many techniques are used in one analysis
  • To estimate different backgrounds
  • To have independent cross checks

• Each method has to be validated on simulated events ("Closure tests")

• Most of the work: evaluation of systematic uncertainties

→ But it's worth the effort, since we would like to prepare for discoveries ... in 2012?
**Detector effects:** Jet resolution, dead ECAL cells, Punch through ...

**Physics:** Leptonic heavy flavor decays

→ **mismeasured jets** (→ large MHT)

Full jet response (incl. tails) measured from data

Select or mimic particle jets (rebalance seed events):

Smear seed events according to measured jet resolutions

Smeared sample resembles full kinematic