Data-driven background estimation in $H \rightarrow WW$ searches

Sergey Kotov

MPI für Physik, Munich

ATLAS seminar, MPI, 26.11.2008

Searches for new particles at LHC

- The primary goal of the LHC project is to find direct evidence of the existence of a Higgs boson and SUSY (SuperSymmetry) particles.
- The Standard Model and SUSY theories predict
  - how the new particles interact with already known particles,
  - in what type of processes they can be produced.
- A search for a new particle, such as Higgs boson, can be successful only if
  - the production rate of this particle is large enough,
  - the products of the particle decay can be detected with reasonable efficiency.
- So, a high energy physicist needs some tools for searches of new particles:
  - an accelerator (the bigger the better) copiously producing the particles of interest,
  - a detector registering the particle decays,
  - a computing system for analysis of data collected from the detector.

That is the LHC project – the biggest tool ever created.

- Before embarking on a real data analysis, one has to rely on theoretical predictions (production cross-sections and decay branching ratios) and detector simulation to estimate “hopefulness” or “hopelessness” of a particular search for the new particle.

Signal and background events

- The measurable quantity in a typical search for a new particle is the number of a particular type of events observed by the detector.
- The type of events (signal events) is defined by the process that contains the new particle and described by:
  - types of particles in the final state,
  - event topology and kinematics.
- Unfortunately, particle interactions in colliders is a random process $\Rightarrow$ One cannot produce on demand only signal events, a plethora of other processes occur in the beam collisions producing background events.
- Physicists have to use some intricate event selection requirements trying to select only signal events and to reduce contamination from background processes.
- The total number of observed events can be written as:

$$N^{obs} = N_s^{obs} + N_b^{obs} = \epsilon_s \cdot N_s + \epsilon_b \cdot N_b$$

  - N_s and N_b are the numbers of signal and background events produced in beam collisions $\Rightarrow$ depend on the collider parameters (luminosity and center-of-mass energy, $\mathcal{L}$ and $\sqrt{s}$),
  - ϵ_s and ϵ_b are the efficiencies of detection of signal and background events $\Rightarrow$ depend on the detector design and implementation.

Event detection efficiencies

Efficiencies can usually be factorized as:

$$\epsilon = \epsilon^{trig} \cdot \prod_i \epsilon_i^{acc} \cdot \prod_i \epsilon_i^{reco} \cdot \epsilon^{sel}$$

- ϵ^{trig} – trigger efficiency for the particular final state:
  - defined by available signatures in the trigger menu,
  - the trigger menu is configurable but limited in functionality by the trigger system hardware,
  - the trigger menu is configured on the basis of consensus from data analysis groups.
- ϵ_i^{acc} – detector acceptance for the i-th particle (i runs over all particles in the final state):
  - describes hermeticity of the detector (solid angle coverage),
  - for different types of particles corresponds to the geometrical layout of the detector subsystems.
- ϵ_i^{reco} – reconstruction efficiencies of different types of particles:
  - determined by the performance of reconstruction algorithms,
  - reflect inhomogeneities in the layout of readout elements, inhomogeneities in the detector dead material, failed/broken detector modules, etc.,
  - specifically studied by the detector performance groups.
- ϵ^{sel} – efficiency of selection requirements in the data analysis.
- All efficiency terms are functions of the final state particle parameters (transverse momentum, pseudorapidity, azimuthal angle).
- There are usually correlations between ϵ^{acc} and ϵ^{sel} and some minor correlations or overlaps between other terms are possible.

Signal significance

The main goal for a physicist in searches for new particles is to find selection requirements which:

- maximize the signal efficiency ϵ_s,
- and minimize the background efficiency ϵ_b.

Hard to do that simultaneously $\Rightarrow$ Instead the physicist requires the highest signal significance:

$$S/\sqrt{B} = N_s^{obs}/\sqrt{N_b^{obs}} = \text{max}$$

- $S/\sqrt{B} > 3$ - “evidence” for a new particle,
- $S/\sqrt{B} > 5$ - “discovery” of a new particle,
- an alternative definition is $S/\sqrt{S+B}$ (when $S \sim B$).

The Big Question

How to decompose the total number of observed events into signal and background contributions?

Partitioning of observed events: "Simple event counting"

Approach I: "Simple event counting".

- Calculate the number of background events produced in beam collisions:

$$N_b = \sigma_b \cdot \mathcal{L}$$

  - σ_b – the background cross-section known from theoretical predictions,
  - $\mathcal{L}$ – the integrated luminosity delivered by the collider.

- Derive efficiencies ϵ_s and ϵ_b from the detector full Monte Carlo simulation.
- Compute "excess of observed events over expected background":

$$N_s^{obs} = N^{obs} - N_b^{exp} = N^{obs} - \epsilon_b \cdot N_b = N^{obs} - \epsilon_b \cdot \sigma_b \cdot \mathcal{L}$$

- Assess the signal significance

$$S/\sqrt{B} = N_s^{obs} / \sqrt{N_b^{exp}}$$

  and claim an "evidence", a "discovery", or an "exclusion upper limit".

- Finally, the signal process cross-section can be derived and compared with theoretical predictions:

$$\sigma_s = N_s^{obs} / (\epsilon_s \cdot \mathcal{L})$$

Mission completed.

Difficulties in calculations of background rates

At LHC uncertainties in calculations of expected number of background events are very large!

$$N_b = \sigma_b \cdot \mathcal{L}$$

- Theoretical predictions for cross-sections have large uncertainties:
  - 10-30% for most of QCD processes (for some processes NLO calculations are not available and uncertainties are above 50%),
  - $\sim 5\%$ for EW processes.
- Uncertainties in Parton Density Functions (fraction of a proton momentum carried by its constituent quarks and gluons) are $\sim 10\%$.
- Luminosity at LHC will be initially measured with $\sim 20\%$ uncertainty $\Rightarrow$ will be reduced to $\sim 3\%$ after installation of luminosity monitors.
- Total initial uncertainties at LHC in background estimations are at the level of 25-40%.

Miscalculation of efficiencies

- Calculation of efficiencies from the full Monte Carlo simulation of the detector can carry a considerable uncertainty due to several factors (in the order of severeness):
  - inadequate calibration and/or alignment of the detector,
  - overestimation of the performance of reconstruction algorithms,
  - limited statistics of the detector Monte Carlo simulation,
  - discrepancies between the real detector geometry and its description in the simulation.
- The overall level of initial detector uncertainties for different analyses with the ATLAS detector is about 10-20%.
- These uncertainties will be decreasing to the level of 3-5% with improved understanding of the detector performance.

Partitioning of observed events: “Fitting curves”

Approach II: “Fitting curves”.

- From observed events plot values of some variable x (usually invariant mass of some particles) *apriori* having different distributions for signal and background events.
- Fit the obtained distribution $F(x) = F_s(x) + F_b(x)$ with a proper function $f(x)$ reflecting the shapes of x distribution in signal and background events:

$$f(x; \mathtt{Norm}_s, \mathtt{Norm}_b) = \mathtt{Norm}_s \cdot f_s(x) + \mathtt{Norm}_b \cdot f_b(x)$$

  - $\mathtt{Norm}_s, \mathtt{Norm}_b$ are the signal and background normalizations,
  - $f_s(x)$ is the shape of x distribution in signal events (passing all selection requirements),
  - $f_b(x)$ is the shape of x distribution in background events (passing all selection requirements).
- Deduce the number of observed signal events N_s^{obs} from the area of the signal part of the full distribution $F_s(x) = \mathtt{Norm}_s \cdot f_s(x)$ after the fit.
- Then follow the chain: signal significance assessment $\Rightarrow$ claim of discovery $\Rightarrow$ signal cross-section calculation $\Rightarrow$ write a paper.
- Where from one can get the necessary ingredients $\mathtt{Norm}_s, \mathtt{Norm}_b, f_s(x)$, and $f_b(x)$?
  - Monte Carlo simulations,
  - real data,
  - combination of both.

Overview of data-driven background estimation methods

To reduce dependencies of an analysis on Monte Carlo simulations and theoretical predictions one can employ data-driven background estimation techniques.

- The goal of a data-driven background estimation is to obtain the background normalization $\mathtt{Norm}_b$ and the function shape $f_b(x)$ used in the fitting function of some distribution measured in data, $f(x) = \mathtt{Norm}_s \cdot f_s(x) + \mathtt{Norm}_b \cdot f_b(x)$, using other distributions derived from data itself.
- This goal is quite difficult to achieve $\Rightarrow$ simplified procedures are used:
  - quite often, the background shape can be rather well predicted using a Monte Carlo simulation, leaving only the background normalization to be determined from data,
  - sometimes it is possible to determine the ratio of the signal and background normalizations $\Rightarrow$ the number of free parameters in the final fit is decreased.

Data-driven methods for background estimation

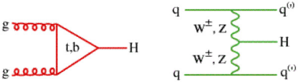
- Method of control samples.
- Particle replacement method.
- Side-band subtraction method.
- Same/opposite sign subtraction method.

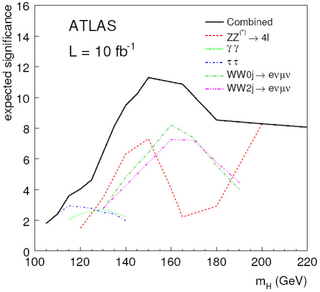
Method of control samples

Data-driven background estimation: “Method of control samples”.

- After applying all analysis selection requirements to data a signal sample is obtained.
- By varying analysis selection requirements (choosing different subsets or changing values of cuts) a background control sample with highly enriched content of background events can be produced (usually a set of control samples is produced).
- The background shape $f_b(x)$ is extracted by fitting the distribution from events in the background control sample.
- One has to check that the background function shape is preserved after the full set of selection requirements is applied $\Rightarrow$ can be done either by using Monte Carlo or by using a different control sample.
- The $f_b(x)$ is used in the fit of the variable x distribution from the signal sample with signal and background normalizations as free parameters.
- By using a different background control sample the background normalization $\mathtt{Norm}_b$ can be determined in a similar way and the number of free parameters in the final fit decreased.

- $H \to WW^* \to e\nu\mu\nu$ is a possible early discovery channel for a SM Higgs boson in the medium range of the Higgs mass: 140 GeV $< M_H <$ 190 GeV
- The gluon fusion (GF) and vector boson fusion (VBF) Higgs boson production modes are considered.





- The major backgrounds for $H \to WW^* \to e\nu\mu\nu$ searches are:
  - $gg \to WW$ and $qq/qg \to WW$ diboson production (irreducible background for the GF mode),
  - $t\bar{t}$ production with $t\bar{t} \to bWbW \to be\nu b\mu\nu$ (reducible),
  - W+jets production with $W \to \mu\nu$ and a jet faking an electron (reducible),
  - Z+jets production with $Z \to \tau\tau \to e\nu\nu\mu\nu\nu$ (reducible).
- Searches in the same lepton flavor channels $H \to WW^* \to ee\nu\nu$ and $H \to WW^* \to \mu\mu\nu\nu$ are also being persuited $\Rightarrow$ more difficult due to $Z \to ee/\mu\mu$ backgrounds.

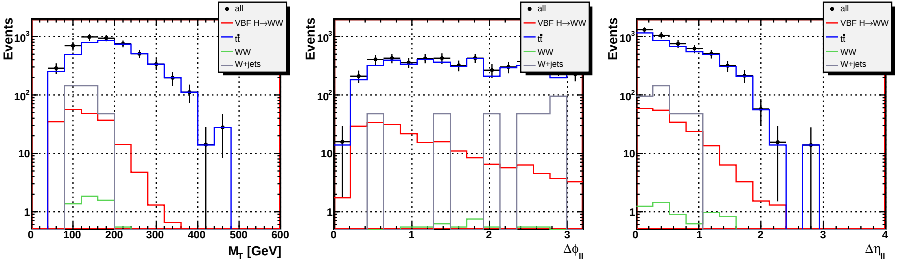
Event selection in $H \to WW^* \to e\nu\mu\nu$ VBF searches

Basic event selection requirements:

- a pair of oppositely charged isolated electron and muon with $p_T > 15$ GeV and $|\eta| < 2.5$,
- at least two jets with $p_T > 20$ GeV and $|\eta| < 4.8$ (tag jets),
  - leading and sub-leading jets should be in opposite hemispheres with $|\Delta\eta| > 3$,
  - leptons should be between the jets in pseudorapidity,
- $Z \to \tau\tau$ reconstruction using collinear approximation, $|M_{\tau\tau} - M_Z| > 25$ GeV.

The variables of interest for data-driven background estimation are:

- the dilepton opening angle in the transverse plane $\Delta\phi_{ll}$,
- the transverse mass $M_T = \sqrt{2p_T^{ll}E_T^{miss}(1 - \cos(\Delta\phi_{ll}))}$,
- the pseudorapidity gap between the leptons $\Delta\eta_{ll}$.



Distributions after basic selection cuts (normalized to $\mathcal{L} = 10$ fb^{-1})

Scheme of data-driven $t\bar{t}$ background estimation

Partitioning of events passing basic preselection (plus some additional cuts) into signal and background control samples

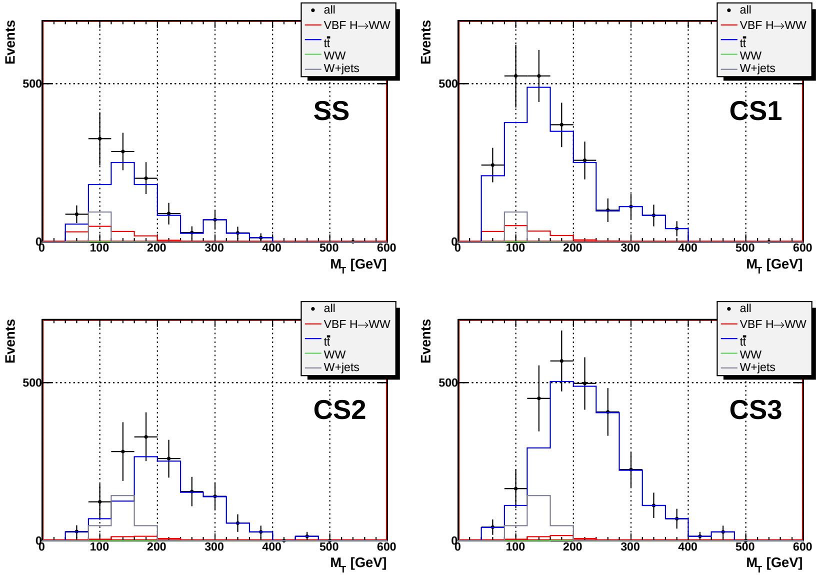
- An event belongs to the signal sample (SS) if $\Delta\phi_{ll} < 1.5$ and it passes b-jet veto (b-tag of jets < 3).
- An event belongs to the control sample 1 (CS1) if $\Delta\phi_{ll} < 1.5$ and no b-jet veto applied.
- An event belongs to the control sample 2 (CS2) if $\Delta\phi_{ll} > 1.5$ and it passes b-jet veto.
- An event belongs to the control sample 3 (CS3) if $\Delta\phi_{ll} > 1.5$ and no b-jet veto applied.

Number of expected events in the signal and control samples according to the ATLAS CSC note on $H \to WW$ searches (for an integrated luminosity of 10 fb^{-1})

Sample	$H \to WW$	$t\bar{t}$	WW	$W + jets$
Signal	93	285	48	43
CS1	96	1140	50	60
CS2	30	890	98	79
CS3	31	3117	103	79

- Take as the background shape $f_b(x)$ the shape of the distribution from CS1.
- Take as the background normalization $\mathtt{Norm}_b$ the ratio between the numbers of events in CS2/CS3.
- Fit the distribution in the signal sample with $f(x) = \mathtt{Norm}_s \cdot f_s(x) + \mathtt{Norm}_b \cdot f_b(x)$, where $f_s(x)$ is a Bifurcated Gaussian (with fixed widths obtained from MC).

M_T distributions for different control samples



Summary

- Partitioning of observed events into signal and background contributions is a complicated analysis issue.
- Monte Carlo simulation is a necessary tool helping to separate signal and background events.
- Initial uncertainties at LHC in estimation of background rates and detector efficiencies are quite large.
- Data-driven background estimation can help to considerably reduce these uncertainties.
- Procedures for data-driven background estimation are rather tricky and can carry uncertainties comparable with MC simulation uncertainties.