PROCEEDINGS OF SCIENCE

Data-driven background estimation in ATLAS

Yoram ROZEN*
On behalf of the ATLAS Collaboration.

Technion - Israel Institute of Technology (IL)
E-mail: yoram.rozen@cern.ch

The search for a charged Higgs with its mass below the top quark mass in ATLAS assumes a production of this boson in $t\bar{t}$ events. This leads to a significant background from a variety of sources common to $t\bar{t}$ events. As the knowledge of many of these sources is limited, alternative methods were developed to tackle these systematic uncertainties. Data-driven methods are used to estimate most of the background of the charged Higgs searches with the ATLAS detector at the LHC.

Prospects for Charged Higgs Discovery at Colliders
October 8-11, 2012
Uppsala University Sweden

*Speaker

PoS(CHARGED 2012)013

1. Introduction

ATLAS [1] has reported a number of searches for the charged Higgs at masses below the top quark mass [2]-[6]. These searches can be divided into $H^+ \to \tau\nu_\tau$ and $H^+ \to c\bar{s}$. The former are divided into three analyses with two of them searching for the tau in its hadronic decay mode and one analysis with a leptonic decay of the tau. The hadronic channels differ in their decay of the other top quark in the event. One analysis assumes a hadronic decay of the W boson while the other uses the leptonic decay. The leptonic decaying tau channel is searched for in events where the other top quark decays to hadrons. Hence, now also taking into account $H^+ \to c\bar{s}$, the four analyses share many common issues and backgrounds. In this article, we will describe the different sources of background and the way ATLAS is estimating them.

While the four analyses mentioned above suffer from different backgrounds due to different physical processes, it is easier to use a different approach in which we look at the actual particles that are producing the background. Hence, a lot of common issues between the analyses can be estimated by using the same methods.

As tau leptons are the key for three of these searches, some of the background sources can be collected under the category of fake tau candidates. Such fake tau candidates can originate from two sources: electrons and hadronic jets. Section 2 describes the methods to estimate these backgrounds. The remainder of the environment can also be misidentified and this is described in section 3. Finally, true taus from the $t\bar{t}$ decay may also pass the selection criteria and these are dealt with a technique named "embedding" which is described in section 4.

Data-driven methods are also used to estimate the fake rate of electrons and muons where the final state of the decay process includes such leptons. Fake lepton backgrounds are not presented here, however details can be found in [3]

2. Fake τ candidates

The algorithm for identifying tau leptons is described in [3]. However, regardless of the identification method used, there will be some candidates which did not originate from tau leptons. Two such sources are electrons and hadronic jets. These two categories produce a signature which overlap that of the tau leptons to some extent. The method to estimate their amount are described below.

2.1 Electrons misidentified as τ jets

A technique to derive this fake rate from data is the so-called tag-and-probe method. The process $Z \to ee$ allows the selection of an unbiased and clean sample of electrons from data. While the tag electron is required to satisfy a tight electron selection, the other, if it is reconstructed as a τ jet candidate, is then used as the probe. Figure 1 shows the dielectron invariant mass with a clear Z signal.

Only those probe τ jet candidates with exactly one associated track are considered, as the rate of electrons faking 3-track τ jets is negligible. As an illustration, the fake rates measured with the 2010 data are shown in Fig. 2. Within uncertainties, the fake rates modeled in simulation agree with those obtained from data.

PoS(CHARGED 2012)013

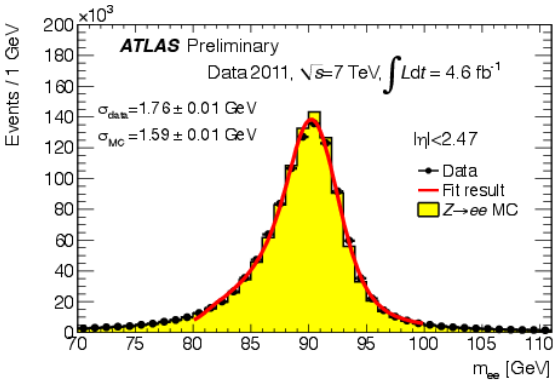

Figure 1: Dielectron invariant mass.

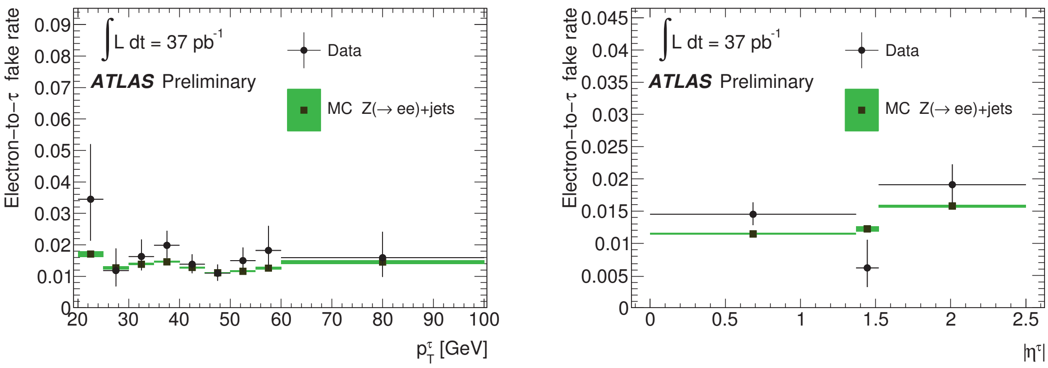

Figure 2: The fake rate for probe objects passing the τ identification, the electron veto, and overlap removal with reconstructed electrons is shown parametrized in p_T and $|\eta|$. The uncertainties indicated are statistical only.

The electron-to-τ fake background is estimated the following way: in simulated events, any true electron matched to a τ jet candidate is labeled as an identified τ jet and the event is given a weight equal to the fake rate probability as described above, instead of performing the usual τ identification (i.e. the τ identification part is taken from data instead of simulation). The baseline selections of both the τ+jets and the τ+lepton channels are then applied and the number of events surviving is counted (summing the weights of these events).

Systematic uncertainties

Three main sources of systematic uncertainties on the electron-to-τ fake rate have been studied. The largest contribution originates from the background contamination with QCD jets (after the application of the electron veto on the probe object, QCD jets are enhanced with respect to electrons among the τ jet candidates) and gives an uncertainty of about 30%. The choice of the mass window size around the Z boson mass applied to the tag-and-probe objects introduces another uncertainty (13%). The uncertainty of the electron energy scale (via the cut on the tag electron energy) only gives a small contribution (2%). The total systematic uncertainty varies slightly with p_T and η and is estimated to be 33%.

2.2 Jets misidentified as τ jets

A data-driven method based on a control sample enriched in W+jets events is used to measure the probability for a jet to be misidentified as a hadronically decaying τ lepton. Like jets from the hard process in the dominant $t\bar{t}$ background, jets in the control sample originate predominantly from quarks instead of gluons. The measured probability is used to predict the yield of events due to jet$\rightarrow$ τ misidentification from the most important SM backgrounds.

The main difference between $t\bar{t}$ and W+jets events is the different fraction of b jets, which is smaller in W+jets events. However, the probability for a b jet to be misidentified as a τ jet is smaller than the corresponding probability for a light-quark jet, because the average track multiplicity is higher for b jets. Also, variables measuring the mass enter the τ reconstruction, providing further discrimination between b jets and τ jets. Differences in jet composition (e.g. the ratio of gluons to quarks) between $t\bar{t}$ and W+jets, assessed using simulation, are taken into account as systematic uncertainties. These also cover the dependence of the probability on whether a b jet or a light-quark jet is misidentified a τ jet. Events in the control region are required to pass the same single-lepton trigger, data quality and lepton requirements as in the τ+lepton event selection. Additionally, the presence of a τ candidate and $E_T^{miss} > 40$ GeV are required, and events with b-tagged jets are vetoed. Events with a true τ contribute at a level below 0.5%. The τ candidates are required to have $p_T > 20$ GeV, $|\eta| < 2.3$, and cannot be within $\Delta R = 0.2$ of any electron or muon; they are not required to pass τ identification.

The jet$\rightarrow$ τ misidentification probability is defined as the number of objects passing the full τ identification divided by the number prior to requiring identification. This misidentification probability is measured as a function of both p_T and η. In addition, it is evaluated separately for τ candidates with 1 or 3 associated tracks. In order to predict the background for the charged Higgs boson search, the measured jet$\rightarrow$ τ misidentification probability is applied to simulated $t\bar{t}$, single top quark, W+jets, Z/γ^*+jets and diboson events, which are required to pass the full event selection except for the τ identification. For these events, τ candidates not overlapping with a true τ lepton or a true electron, but otherwise fulfilling the same requirements as in the denominator of the misidentification probability, are identified. Each of them is considered separately to be potentially misidentified as a τ jet. In order to avoid counting the same object twice, each jet that corresponds to a τ candidate is removed from the event, affecting the number of reconstructed jets and the number of b-tagged jets. If, after taking this into consideration, the event passes the τ+lepton selection, it is counted as a background event with a weight given by the misidentification probability corresponding to the p_T and η of the τ candidate. Figure 3 show the distribution of the E_T^{miss} of the τ+lepton channel. The largest contribution to this channel is from jets misidentified as taus.

3. Multi-jet background (τ + jets only)

The multi-jet background is estimated by fitting its E_T^{miss} shape (and the E_T^{miss} shape of other backgrounds) to data. In order to study this shape in a data-driven way, a control region is de-

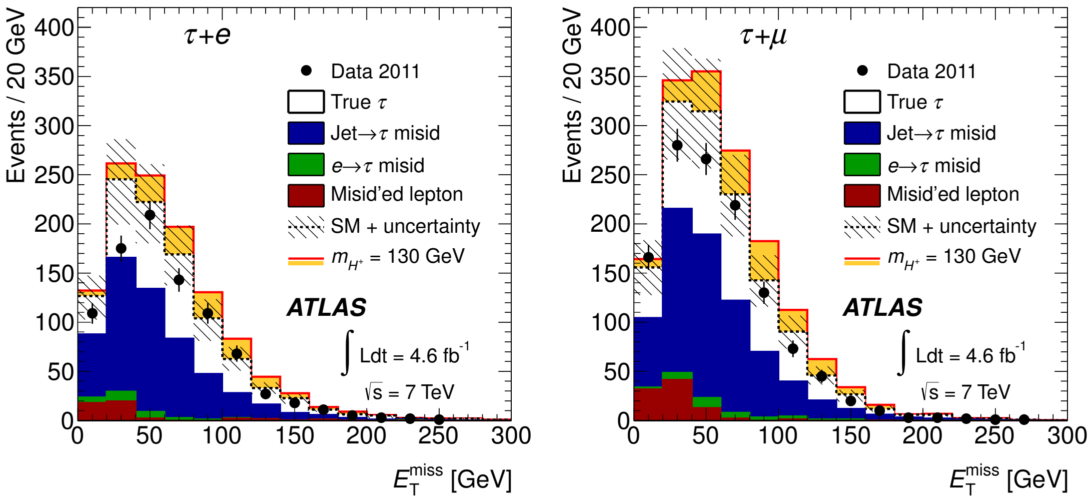

Figure 3: E_T^{miss} distribution after all selection cuts in the τ+lepton channel, for (a) τ+electron and (b) τ+muon final states. The hatched area shows the total uncertainty for the SM backgrounds. The stacked histogram shows the predicted contribution of signal+background in the presence of a 130 GeV charged Higgs boson with $B(t \to bH^+) = 5\%$ and $B(H^+ \to \tau\nu) = 100\%$. The contributions of $t\bar{t} \to b\bar{b}W^+W^-$ events in the backgrounds with true or misidentified τ jets are scaled down accordingly.

fined where the τ identification and b-tagging requirements are modified, i.e. τ candidates must pass a loose τ identification but fail the tight τ identification used in the signal selection, and the event is required not to contain any b-tagged jet. Assuming that the shapes of the E_T^{miss} and m_T distributions are the same in the control and signal regions, the E_T^{miss} shape for the multi-jet background is measured in the control region, after subtracting the simulated background contributions from other processes. These other processes amount to less than 1% of the observed events in the control region. The E_T^{miss} shapes obtained with the τ+jets selection or in the control region are compared early in the selection cut flow in Fig. 4(a). The differences between the two distributions are accounted for as systematic uncertainties. For the baseline selection, the E_T^{miss} distribution measured in the data is then fit using two shapes: the multi-jet model and the sum of other processes (dominated by $t\bar{t}$ and W+jets), for which the shape and the relative normalisation are taken from simulation, as shown in Fig. 4 (b).

4. True tau background - Embedding (τ + jets only)

An embedding method is used to estimate the backgrounds that contain correctly reconstructed τ jets. The method consists in selecting a control sample of $t\bar{t}$-like μ+jets events and replacing the detector signature of the muon by a simulated hadronic τ decay. These new hybrid events are then used for the background prediction. In order to select this control sample from the data, the following event selection is applied:

- event triggered by a single-muon trigger (p_T threshold of 18 GeV);
- exactly one isolated muon with $p_T > 25$ GeV, no isolated electron with $E_T > 20$ GeV;

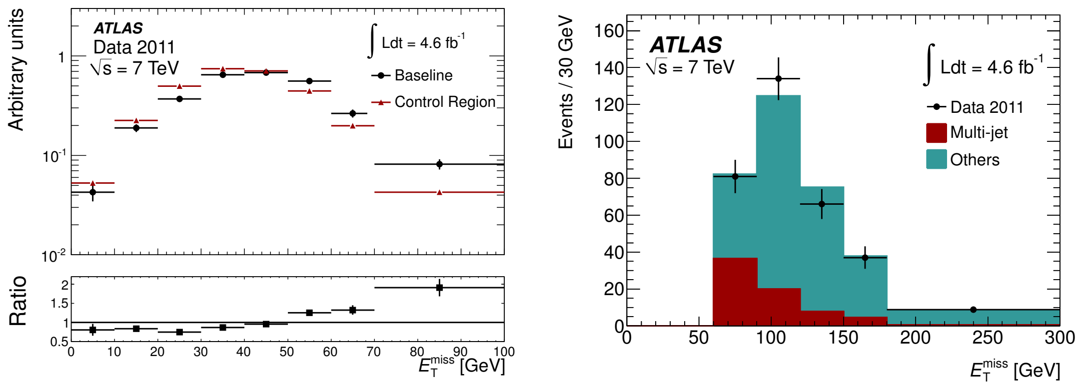

Figure 4: (a) Shape of E_T^{miss} in a control region of the data using the baseline selection, after subtracting the expectation from $t\bar{t}$, W+jets, and single top quark processes estimated from simulation. The distributions are compared just before the E_T^{miss} requirement in the baseline selection, with the exception that, in the control region, the τ selection and the b-tagging requirements are modified, see text. (b) Fit of the E_T^{miss} template to data, in the signal region. Only statistical uncertainties are shown.

- at least four jets with $p_T > 20$ GeV and $|\eta| < 2.4$, at least one of which is b-tagged;
- $E_T^{miss} >$ 35 GeV.

This selection is looser than the selection of the analysis in order not to bias the control sample. The τ+jets event selection is then applied to the embedded events. The impurity from the background with muons produced in τ decays and non-isolated muons (dominantly $b\bar{b}$ and $c\bar{c}$ events) is about 10%. However, this contribution is greatly reduced as these events are much less likely to pass the τ+jets selection, in particular the jet p_T requirement. The shape of the m_T distribution for the backgrounds with true τ jets is taken from the distribution obtained with the embedded events. The normalisation is then derived from the number of embedded events:

$$N_\tau = N_{embedded}(1 - c_{\tau\to\mu})\frac{\varepsilon^{\tau+E_T^{miss}-trigger}}{\varepsilon^{\mu-ID,trigger}}B(\tau \to hadrons + \nu)$$

where N_τ is the estimated number of events with correctly reconstructed τ jets, $N_{embedded}$ is the number of embedded events in the signal region, $c_{\tau\to\mu}$is the fraction of events in which the selected muon is a decay product of a τ lepton (taken from simulation), $\varepsilon^{\tau+E_T^{miss}-trigger}$ is the $\tau + E_T^{miss}$ trigger efficiency (as a function of p_τ and E_T^{miss}, derived from data), $\varepsilon^{\mu-ID,trigger}$ is the muon trigger and identification efficiency (as a function of p_T and η, derived from data) and $B(\tau \to hadrons + \nu)$ is the branching ratio of the τ lepton decays involving hadrons. The m_T distribution for correctly reconstructed τ jets, as predicted by the embedding method, is shown in Fig. 5 and compared to simulation.

5. Summary

In the search for charged Higgs at masses below the top quark mass, many background sources can be accounted for and estimated using data. The level of simulation input into these estimation

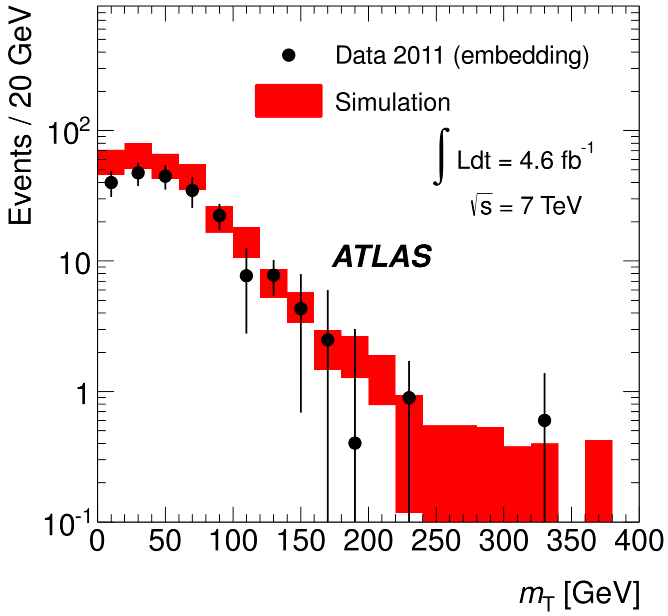

Figure 5: Comparison of the m_T distribution for correctly reconstructed τ jets, predicted by the embedding method and simulation. Combined statistical and systematic uncertainties are shown.

varies from minimal as is the case of the electron fake rate to high dependency as in the case of the multi-jet background. Figure 6 shows the transverse mass distribution with a breakdown of the background sources in the case of the τ + jets channel.

References

[1] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," 2008 JINST 3 S08003.

[2] ATLAS Collaboration, "Search for a charged Higgs boson decaying via H+ to tau(lep)+nu in $t\bar{t}$ events with one or two light leptons in the final state using 1.03/fb of pp collision data recorded at $\sqrt{s} = 7$ TeV with the ATLAS detector," ATLAS-CONF-2011-151.

[3] ATLAS Collaboration, "Search for charged Higgs bosons decaying via $H^+\tau\nu$ in $t\bar{t}$ events using 4.6 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector," ATLAS-CONF-2012-011.

[4] ATLAS Collaboration, "A Search for a light charged Higgs boson decaying to $c\bar{s}$ in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," ATLAS-CONF-2011-094.

[5] ATLAS Collaboration, "Search for Charged Higgs Bosons in the τ+jets Final State in $t\bar{t}$ Decays with 1.03 fb^{-1} of pp Collision Data Recorded at $\sqrt{s} = 7$ TeV with the ATLAS Experiment," ATLAS-CONF-2011-138.

[6] ATLAS Collaboration, "Search for charged Higgs bosons decaying via $H^+ \to \tau\nu$ in top quark pair events using pp collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector," JHEP **1206**, 039 (2012) [arXiv:1204.2760 [hep-ex]].

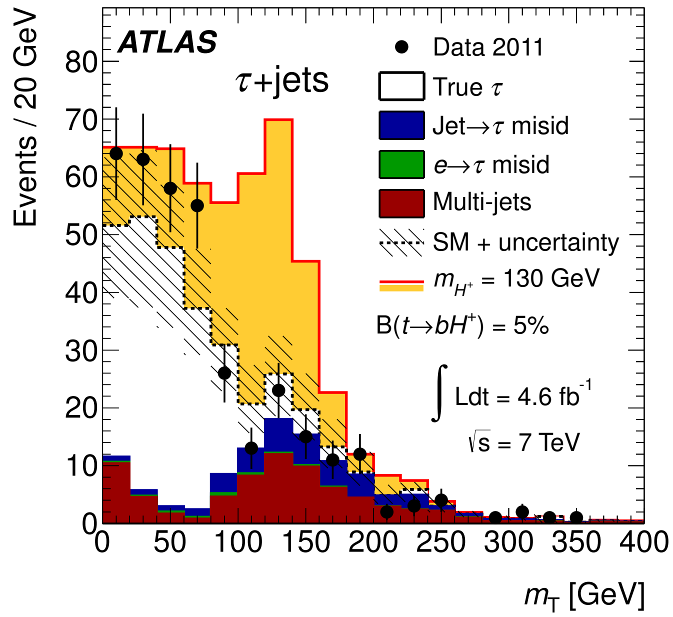

Figure 6: Distribution of m_T after all selection cuts in the τ+jets channel. The hatched area shows the total uncertainty for the SM backgrounds. The stacked histogram shows the predicted contribution of signal+background in the presence of a charged Higgs boson with $m_{H^+} = 130$ GeV, assuming $B(t \to bH^+) = 5\%$ and $B(H^+ \to \tau\nu) = 100\%$. The contributions of $t\bar{t} \to b\bar{b}W^+W^-$ events in the backgrounds with misidentified objects are scaled down accordingly.