PROCEEDINGS OF SCIENCE

Data-driven background estimations in H^+ analyses in ATLAS

Allison Mc Carn*, **on behalf of the ATLAS Collaboration**
University of Michigan
E-mail: allison.renae.mc.carn@cern.ch

Included are proceedings for "Prospects for Charged Higgs Discovery at Colliders - CHARGED 2014" at Uppsala University, for a contributed talk describing the major data-driven background estimation techniques used in $H^+ \to \tau\nu$ searches conducted on collisions produced by the LHC and recorded by the ATLAS detector. These techniques include true τ_{had} embedding, the use of correction factors to estimate backgrounds with a jet misidentified as a τ_{had}, and matrix methods used to predict contributions where jets are misidentified as a τ_{had} or other lepton.

Prospects for Charged Higgs Discovery at Colliders - CHARGED 2014
16-18 September 2014
Uppsala University Sweden

*Speaker.

PoS(Charged2014)009

1. Introduction

Charged Higgs bosons are predicted by beyond Standard Model scenarios with extended Higgs sectors, of which the Minimal Supersymmetric Standard Model (MSSM), which has a second Higgs doublet field, is one example [1, 2, 3, 4, 5]. These proceedings include information from searches that have been conducted for the existence of a charged Higgs boson using collisions produced at $\sqrt{s}$ = 7 and 8 TeV by the LHC and recorded by the ATLAS detector [6, 7, 8, 9].

1.1 Charged Higgs boson production and decay

The dominant production mechanism for charged Higgs bosons differs depending on its mass. When $m_{H+} < m_{top}$, it is produced in $t\bar{t}$ decays. When $m_{H+} > m_{top}$, it is dominantly produced in association with a top quark.

For $m_{H+} < m_{top}$, the dominant decay is $H^+ \to \tau\nu$, and this decay channel remains sizeable even for $m_{H+} > m_{top}$. Since this decay mode provides relatively clean signatures, it has been a focus for run-1 searches. As a result, modelling of τ leptons is very important for run-1 searches, and many data-driven methods have been developed to estimate events with τ leptons and their backgrounds.

1.2 τ reconstruction and identification

τ leptons can decay either hadronically or into an electron or muon. When the decay of the τ is to an electron or muon, it is identified as an electron or muon. When a τ decays hadronically, it can be identified by a characteristic signature of one or three highly collimated charged tracks. The visible decay product of the hadronically-decaying τ lepton, $\tau_{\text{had-vis}}$, is reconstructed and identified with dedicated algorithms.

Candidates for identification as $\tau_{\text{had-vis}}$ arise from jets reconstructed from energy deposits in calorimeters, using the anti-k_t algorithm [10] with a radius parameter of $R = 0.4$, which have $p_{\text{T}} >$ 10 GeV and one or three charged particle tracks within a narrower radius of $\Delta R < 0.2$ around the intermediate $\tau_{\text{had-vis}}$ axis [11]. The output of boosted decision tree algorithms [12] are used to distinguish $\tau_{\text{had-vis}}$ from jets not initiated by τ leptons, separately for τ_{had} decays with one or three charged tracks. The requirement on this algorithm is referred to as the $\tau_{\text{had-vis}}$ ID requirement. Other dedicated algorithms are used to reject electrons and muons that could be incorrectly identified as $\tau_{\text{had-vis}}$ [11]. After these algorithms are applied, the backgrounds arising from muons and electrons misidentified as $\tau_{\text{had-vis}}$ are generally very small, though there is still a sizeable jet background.

1.3 Searches and their backgrounds

This document covers background estimation techniques used in three separate searches. The first of these is the "τ_{had} + jets search [7], which is a search for a final state with one τ_{had}, $E_{\text{T}}^{\text{miss}}$, jets, at least 1 b-tagged jet, and no electrons or muons. This search is performed for m_{H^+} in the ranges of 80-160 GeV and 180-1000 GeV. The second is the "Ratio Method" search [8], which investigates the ratio between the yields of "τ_{had}+ e or μ" and "e+μ" final states. Deviations from the expected ratio would indicate the presence of new physics, such as H^+, and the search was performed for m_{H+} = 90-160 GeV. The third is the "τ_{lep} + jets" search, which is for a final state

with one electron or muon, $E_{\mathrm{T}}^{\mathrm{miss}}$, jets, at least 1 b-tagged jet, and no τ_{had}[9]. The backgrounds and methods of estimation for each analysis are described in Table 1.

Backgrounds with...	"τ_{had} + jets"	"Ratio Method"	"τ_{lep} + jets"
...correctly reconstructed τ_{had} ("True τ_{had}")	τ_{had} embedding	Simulation	N/A
...jets misidentified as τ_{had}	Simulation	Simulation	N/A
...e or μ misidentified as τ_{had} ("Fake τ_{had}")	Matrix method	Correction factors	N/A
...jets or non-prompt e/μ misidentified as prompt e/μ ("Fake e/μ")	N/A	Matrix method	Matrix method

Table 1: Background categorization and estimation techniques for the three searches covered in these proceedings, "τ_{had} + jets", "Ratio Method", and "τ_{lep} + jets".

2. True τ_{had}: embedding

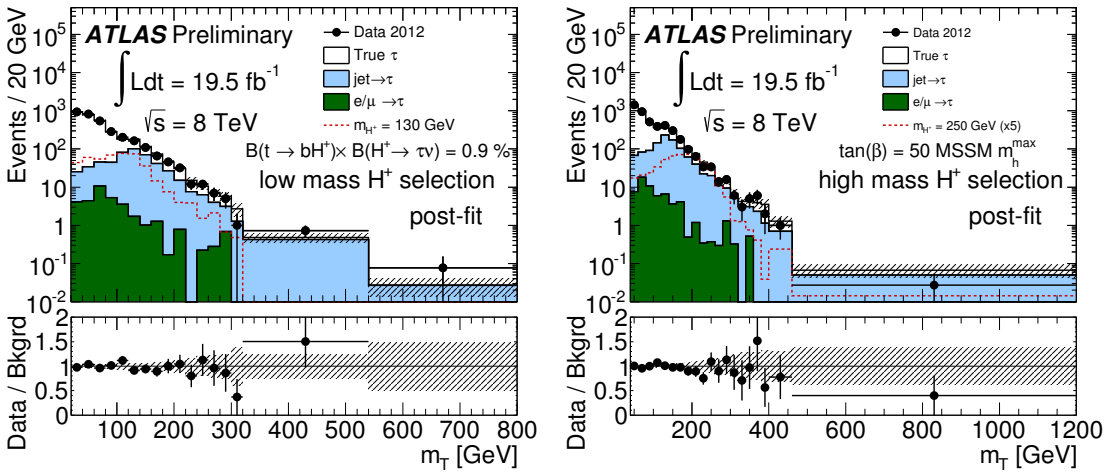
In the "Ratio Method" search, backgrounds with true τ_{had} are estimated from simulation, but in the "τ_{had} + jets" analysis, it is possible to estimate this background using a method known as τ_{had} embedding. This method takes the full event, except for the τ_{had}, directly from data, so no uncertainties due to theoretical cross sections or simulation need to be taken into account. While there are systematic uncertainties associated with the embedding process, these are generally much smaller than the total systematic uncertainties included when using simulation.

The embedding method uses events from a "μ + jets" region in data, which is dominated by $t\bar{t}$ decays, as is also expected in the "τ_{had}+ jets" signal region. Data events that pass a "τ_{had} + jets" selection cannot be used to estimate the background, since they cannot be selected in a data region that does not also have substantial potential signal contamination. On the other hand, "μ + jets" events can be selected with high purity and low signal contamination.

The "μ + jets" events are required to fire a single muon trigger, have exactly one μ with $p_{\mathrm{T}} >$ 25 GeV, no electrons with $E_{\mathrm{T}} > 25$ GeV, at least 4(3) jets, for the low (high) m_{H^+} regions, at least 1 b-tagged jet, and $E_{\mathrm{T}}^{\mathrm{miss}} > 25$ (40) GeV for the low (high) m_{H^+} regions. In these selected events, the μ is then entirely removed, and a simulated τ_{had} is embedded in its place. Then, the rest of the "τ_{had} + jets" signal region selection is applied to this embedded sample.

A strength of the τ_{had} embedding is that the entire event is taken from data, with the exception of the simulated τ_{had}. Due to lepton universality, these embedded events provide the normalization for the "τ_{had} +jets" signal region, after corrections are made for $\tau \to \mu$ decays in data, the μ and τ_{had} trigger and identification efficiencies, and the branching fraction for τ to decay hadronically. The results of τ_{had} embedding can be seen in the final distributions for the "τ_{had} + jets" search, shown in Figure 1.

While the τ_{had} embedding method is useful for reducing the systematic uncertainties and removing the reliance of the background estimation on theoretical cross sections, it also has several important sources of potential bias that one needs to check. One of these is a bias from the $E_{\mathrm{T}}^{\mathrm{miss}}$ requirement in the "μ + jets" event selection, since it is possible, in embedded events, for most of the $E_{\mathrm{T}}^{\mathrm{miss}}$ of the event to arise from the hadronic τ decay. Another potential source of bias is signal contamination in the data region, through $H^+ \to \tau_{\mathrm{lep}}\nu$. The m_{T} distribution of these potential contaminating events has been observed to be much softer than signal, such that the contamination has a negligible effect on exclusion limits in the current dataset.



(a) Low mass H^+ selection

(b) High mass H^+ selection

Figure 1: Distributions of m_{T} after all selection criteria [7]. The hatched area shows the total post-fit uncertainty for the SM backgrounds. For the low mass selection (a), bins are 20 GeV wide up to m_{T} = 320 GeV, then 320–540 GeV and $>$ 540 GeV. For the high mass selection (b), bins are 20 GeV wide up to m_{T} = 400 GeV then 400–460 GeV and $>$ 460 GeV. All bins are normalised to 20 GeV bin width. For the low mass search (a), a possible signal contribution with $m_{H^+} = 130$ GeV, and $\mathscr{B}(t \to bH^+) \times \mathscr{B}(H^+ \to \tau^+\nu) = 0.9\%$ is overlaid on the SM contributions. For the high mass search (b), a possible signal contribution with $m_{H^+} = 250$ GeV and $\tan\beta = 50$ in the m_h^{max} scenario of the MSSM [13], where the corresponding cross section is scaled up by a factor of 5, is overlaid on the SM contributions.

3. Fake τ_{had} and fake e/μ

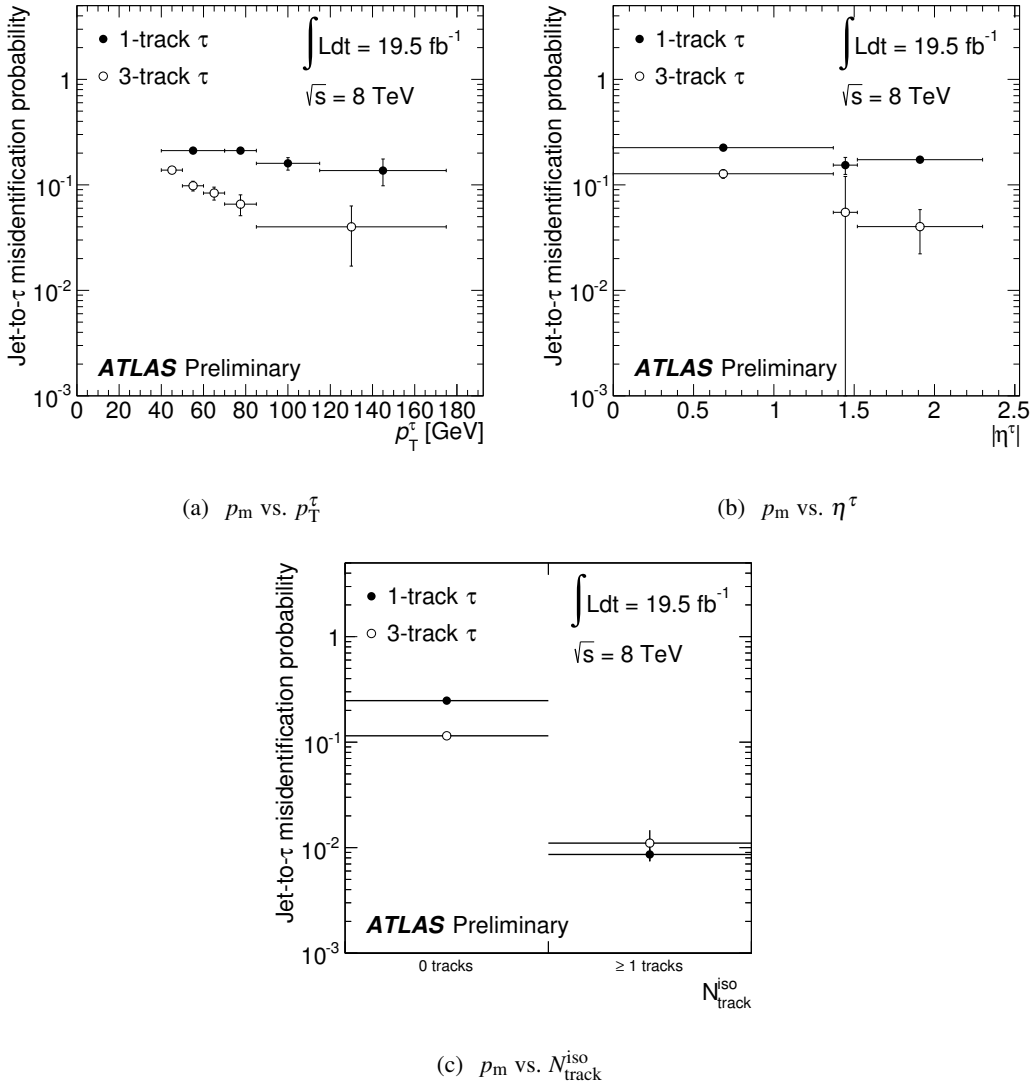
Estimation of backgrounds with fake τ_{had} or fake e/μ meet with several difficulties. First of all, they generally arise from multi-jet processes, which have massive cross sections and tiny acceptance for signal regions. This makes it extremely difficult to generate sufficient numbers of simulated events to model the background. In addition there are problems of mismodeling in simulation, particularly in the case with a jet that is misidentified as a τ_{had}. Due to these issues, these backgrounds are generally estimated in data-driven ways.

3.1 Correction factors

In the "Ratio Method" analysis, the approach to estimating the fake τ_{had} backgrounds is to apply correction factors to simulation. This approach addresses the issues related to the mismodeling of simulated fake τ_{had}. Since this analysis searches for final states with an e/μ as well as a τ_{had}, the backgrounds with fake τ_{had} arise dominantly from electroweak processes, instead of multi-jet, so there is no issue due to the ability to generate sufficient numbers of simulated events for the dominant source of background.

To correct the modeling of the fake τ_{had} events, correction factors are applied to the track distribution (0.71 for 1-prong and 0.92 for 3-prong τ_{had}), and then τ_{had} ID efficiency factors measured from data are applied to simulated events, instead of using the simulated τ_{had} ID. A drawback to this

approach was that it still relies on simulated events, and thus includes all systematic uncertainties related to simulation.



(a) p_{m} vs. p_{T}^{τ}

(b) p_{m} vs. η^{τ}

(c) p_{m} vs. $N_{\mathrm{track}}^{\mathrm{iso}}$

Figure 2: Efficiency p_{m} for a misidentified object in the loose sample to pass the tight selection, as defined in Eq. 3.1 [7]. p_{m} is measured in a W+jets control region in data, and shown as a function of p_{T}^{τ} (a), $|\eta^{\tau}|$ (b), and $N_{\mathrm{track}}^{\mathrm{iso}}$ (c), separately for 1-track and 3-track $\tau_{\text{had-vis}}$ candidates. $N_{\mathrm{track}}^{\mathrm{iso}}$ gives the number of charged particle tracks in a cone $0.2 < \Delta R < 0.4$ around the axis of the $\tau_{\text{had-vis}}$.

3.2 Matrix method

In order to remove reliance on simulation, and therefore remove the need for systematic uncertainties associated with simulation, a new class of background estimation methods, "matrix methods", have been developed. These methods can be used to estimate backgrounds with fake τ_{had} as well as fake e/μ. They are used to estimate the fake τ_{had} background in the "τ_{had} + jets" search (14-22% of background), the fake e/μ background in the "τ_{lep} + jets" search ($\sim$ 5%), and

events where jets are misidentified as both τ_{had} and e/μ in the "Ratio Method" search ($< 1\%$ for τ_{had}+e/μ, $\sim 9\%$ for $e+\mu$).

For the matrix method, two data samples are defined: the *tight* sample, which contains a larger fraction of events with a true object, and the *loose* sample, which contains a larger fraction of events with a misidentified object. By construction, the *tight* data sample is a subset of the *loose* data sample. The number of *loose* and *tight* events are denoted by N_{L} and N_{T}, while the number of events with true ("real") or misidentified objects are denoted as N_{r} and N_{m}. In terms of the efficiencies for a true or misidentified object in the *loose* sample to pass the *tight* selection (p_{r} and p_{m}, respectively), the following relation can be established:

$$\begin{pmatrix} N_{\text{T}} \\ N_{\text{L}} \end{pmatrix} = \begin{pmatrix} p_{\text{r}} & p_{\text{m}} \\ (1-p_{\text{r}}) & (1-p_{\text{m}}) \end{pmatrix} \times \begin{pmatrix} N_{\text{r}} \\ N_{\text{m}} \end{pmatrix}. \tag{3.1}$$

Inverting the 2×2 matrix above, the number of events in which the misidentified *loose* object passes the *tight* selection can be written as:

$$N_{\text{m}}^{\text{T}} = p_{\text{m}} N_{\text{m}} = \frac{p_{\text{m}} p_{\text{r}}}{p_{\text{r}} - p_{\text{m}}} N_{\text{L}} + \frac{p_{\text{m}}(p_{\text{r}} - 1)}{p_{\text{r}} - p_{\text{m}}} N_{\text{T}}. \tag{3.2}$$

For the "Ratio Method" and "τ_{lep} + jets" searches, the *loose* selection loosens identification and isolation requirements from the standard object selection of the search, which defines the *tight* selection. For the "τ_{had}+jets" search, the *loose* and *tight* selections are determined by not requiring or requiring the τ_{had} candidate of the event to pass a τ_{had} ID requirement. For this search, the values for p_{m} are measured from a control region in data that is dominated by W+jets events, and the p_{r} efficiency is measured in simulation, with standard data-driven correction factors applied. The values for p_{m} measured from data and used in this search are shown in Figure 2, and the final results of the method in the signal region can be seen in Figure 1.

4. Conclusions

Many data-driven techniques have been developed and used in the context of the latest H^+ searches, including τ_{had} embedding and matrix methods. These methods rely on data events to estimate backgrounds, thus avoiding modeling problems, many systematic uncertainties, and reliance on theoretical cross sections. These data-driven methods will need to be re-examined and re-commissioned for run-2 analyses, but the knowledge gained from backgrounds studies in run-1 will be a great help in future development.

References

[1] P. Fayet, Supersymmetry and Weak, Electromagnetic and Strong Interactions, Phys.Lett. B64 (1976) 159.

[2] P. Fayet, Spontaneously Broken Supersymmetric Theories of Weak, Electromagnetic and Strong Interactions, Phys.Lett. B69 (1977) 489.

[3] G. R. Farrar and P. Fayet, Phenomenology of the Production, Decay, and Detection of New Hadronic States Associated with Supersymmetry, Phys.Lett. B76 (1978) 575-579.

[4] P. Fayet, Relations Between the Masses of the Superpartners of Leptons and Quarks, the Goldstino Couplings and the Neutral Currents, Phys.Lett. B84 (1979) 416.

[5] S. Dimopoulos and H. Georgi, Softly Broken Supersymmetry and SU(5), Nucl.Phys. B193 (1981) 150.

[6] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 (2008) S08003.

[7] ATLAS Collaboration, Search for charged Higgs bosons decaying via $H^{\pm} \to \tau^{\pm}\nu$ in hadronic final states using pp collision data at $\sqrt{s} = 8$ TeV with the ATLAS detector, ATLAS-CONF-2014-050 (2014), https://cds.cern.ch/record/1756361.

[8] ATLAS Collaboration, Search for charged Higgs bosons through the violation of lepton universality in $t\bar{t}$ events using pp collision data at $\sqrt{s}$ = 7 TeV with the ATLAS experiment, JHEP 1303 (2013) 076, [arXiv:1212.3572].

[9] 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 (2012) 039, [arXiv:1204.2760].

[10] M. Cacciari, G. P. Salam, and G. Soyez, The anti$-k_t$ jet clustering algorithm, JHEP 0804 (2008) 063, [arXiv:0802.1189].

[11] ATLAS Collaboration, Identification of the Hadronic Decays of Tau Leptons in 2012 Data with the ATLAS Detector, ATLAS-CONF-2013-064 (2013), https://cds.cern.ch/record/1562839.

[12] Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1997), no. 1 119.

[13] M. Carena et al., MSSM Higgs Boson Searches at the LHC: Benchmark Scenarios after the Discovery of a Higgs-like Particle, Eur. Phys. J. C73 (2013) 2552, [arXiv:1302.7033].