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 $H^+ \rightarrow \tau \nu$ searches conducted on collisions produced by the LHC and recorded by the ATLAS detector. These techniques include true $\tau_{\text{had}}$ embedding, the use of correction factors to estimate backgrounds with a jet misidentified as a $\tau_{\text{had}}$, and matrix methods used to predict contributions where jets are misidentified as a $\tau_{\text{had}}$ or other lepton.

*Speaker.

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike Licence.
http://pos.sissa.it/
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 \( m_{H^+} < m_{\text{top}} \), it is produced in \( t\bar{t} \) decays. When \( m_{H^+} > m_{\text{top}} \), it is dominantly produced in association with a top quark.

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

1.2 \( \tau \) reconstruction and identification

\( \tau \) leptons can decay either hadronically or into an electron or muon. When the decay of the \( \tau \) is to an electron or muon, it is identified as an electron or muon. When a \( \tau \) 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 \( \tau \) 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 \( p_T > 10 \) 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 \( \tau \) leptons, separately for \( \tau_{\text{had}} \) 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 "\( \tau_{\text{had}} + \text{jets} \) search [7], which is a search for a final state with one \( \tau_{\text{had}}, E_T^{\text{miss}}, \) 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 "\( \tau_{\text{had}} + e \) or \( \mu \)" and "\( e+\mu \)" 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 "\( \tau_{\text{lep}} + \text{jets} \) search, which is for a final state
Data-driven background estimations in H+ analyses in ATLAS

Allison Mc Carn

with one electron or muon, \(E_T^{\text{miss}}\), jets, at least 1 \(b\)-tagged jet, and no \(\tau_{\text{had}}\) [9]. The backgrounds and methods of estimation for each analysis are described in Table 1.

![Table 1: Background categorization and estimation techniques for the three searches covered in these proceedings, \(\tau_{\text{had}} + \text{jets}\), "Ratio Method", and \(\tau_{\text{lep}} + \text{jets}\).]

<table>
<thead>
<tr>
<th>Backgrounds with...</th>
<th>(\tau_{\text{had}} + \text{jets})</th>
<th>&quot;Ratio Method&quot;</th>
<th>(\tau_{\text{lep}} + \text{jets})</th>
</tr>
</thead>
<tbody>
<tr>
<td>...correctly reconstructed (\tau_{\text{had}}) (&quot;True (\tau_{\text{had}})&quot;)</td>
<td>(\tau_{\text{had}}) embedding</td>
<td>Simulation</td>
<td>N/A</td>
</tr>
<tr>
<td>...jets misidentified as (\tau_{\text{had}})</td>
<td>Simulation</td>
<td>Simulation</td>
<td>N/A</td>
</tr>
<tr>
<td>...(e) or (\mu) misidentified as (\tau_{\text{had}}) (&quot;Fake (\tau_{\text{had}})&quot;)</td>
<td>Matrix method</td>
<td>Correction factors</td>
<td>N/A</td>
</tr>
<tr>
<td>...jets or non-prompt (e/\mu) misidentified as prompt (e/\mu) (&quot;Fake (e/\mu)&quot;)</td>
<td>N/A</td>
<td>Matrix method</td>
<td>Matrix method</td>
</tr>
</tbody>
</table>

2. True \(\tau_{\text{had}}\): embedding

In the "Ratio Method" search, backgrounds with true \(\tau_{\text{had}}\) are estimated from simulation, but in the "\(\tau_{\text{had}} + \text{jets}\)" analysis, it is possible to estimate this background using a method known as \(\tau_{\text{had}}\) embedding. This method takes the full event, except for the \(\tau_{\text{had}}\), 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 "\(\mu + \text{jets}\)" region in data, which is dominated by \(t\bar{t}\) decays, as is also expected in the "\(\tau_{\text{had}} + \text{jets}\)" signal region. Data events that pass a "\(\tau_{\text{had}} + \text{jets}\)" 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, "\(\mu + \text{jets}\)" events can be selected with high purity and low signal contamination.

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

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

While the \(\tau_{\text{had}}\) 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 \(E_T^{\text{miss}}\) requirement in the "\(\mu + \text{jets}\)" event selection, since it is possible, in embedded events, for most of the \(E_T^{\text{miss}}\) of the event to arise from the hadronic \(\tau\) decay. Another potential source of bias is signal contamination in the data region, through \(H^+ \rightarrow \tau_{\text{lep}}V\). 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.
Data-driven background estimations in $H^+$ analyses in ATLAS

Allison Mc Carn

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 $m_{H^\pm} = 130$ GeV, and $\mathcal{B}(t \rightarrow bH^\pm) \times \mathcal{B}(H^\pm \rightarrow \tau^+\nu) = 0.9\%$ is overlaid on the SM contributions. For the high mass search (b), a possible signal contribution with $m_{H^\pm} = 250$ 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 $\tau_{\text{had}}$ and fake $e/\mu$

Estimation of backgrounds with fake $\tau_{\text{had}}$ or fake $e/\mu$ 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 $\tau_{\text{had}}$. 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 $\tau_{\text{had}}$ backgrounds is to apply correction factors to simulation. This approach addresses the issues related to the mismodeling of simulated fake $\tau_{\text{had}}$. Since this analysis searches for final states with an $e/\mu$ as well as a $\tau_{\text{had}}$, the backgrounds with fake $\tau_{\text{had}}$ 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 $\tau_{\text{had}}$ events, correction factors are applied to the track distribution (0.71 for 1-prong and 0.92 for 3-prong $\tau_{\text{had}}$), and then $\tau_{\text{had}}$ ID efficiency factors measured from data are applied to simulated events, instead of using the simulated $\tau_{\text{had}}$ ID. A drawback to this
Data-driven background estimations in $H^+$ analyses in ATLAS

Allison Mc Carn

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

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

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 $\tau_{\text{had}}$ as well as fake $e/\mu$. They are used to estimate the fake $\tau_{\text{had}}$ background in the "$\tau_{\text{had}} + \text{jets}$" search (14-22% of background), the fake $e/\mu$ background in the "$\tau_{\text{lep}} + \text{jets}$" search (~5%), and
events where jets are misidentified as both $\tau_{\text{had}}$ and $e/\mu$ in the "Ratio Method" search ($< 1\%$ for $\tau_{\text{had}}+e/\mu$, $\sim 9\%$ 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:

$$
\begin{pmatrix}
N_T \\
N_L
\end{pmatrix} =
\begin{pmatrix}
p_r & p_m \\
1-p_r & 1-p_m
\end{pmatrix}
\times
\begin{pmatrix}
N_r \\
N_m
\end{pmatrix}.
$$

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

$$
N_m^T = p_m N_m = \frac{N_m p_t}{p_t - p_m} N_L + \frac{p_m(p_t - 1)}{p_t - p_m} N_T.
$$

For the "Ratio Method" and "$\tau_{\text{lep}} + \text{jets}" searches, the loose selection loosens identification and isolation requirements from the standard object selection of the search, which defines the tight selection. For the "$\tau_{\text{had}}+\text{jets}" search, the loose and tight selections are determined by not requiring or requiring the $\tau_{\text{had}}$ candidate of the event to pass a $\tau_{\text{had}}$ ID requirement. For this search, the values for $p_m$ are measured from a control region in data that is dominated by $W+\text{jets}$ events, and the $p_t$ 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 $\tau_{\text{had}}$ 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


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

Allison Mc Carn


[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].


