Data-driven estimation of fake τ background in Higgs searches in ATLAS

On behalf of ATLAS collaboration

Marzieh Bahmani

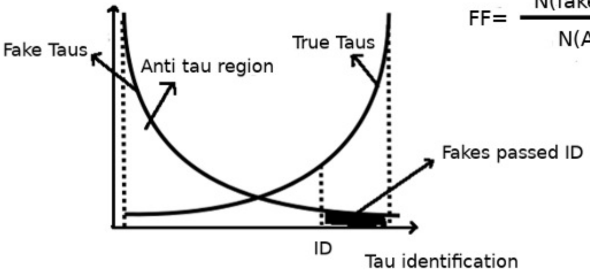
IFJ-PAN Krakow

Sep 2018

Introduction

- Motivation, i.e. jets genuinely fake hadronic τ and all the τ-related analyses suffer from such backgrounds.
- The fake tau background is not well modeled by MC therefore, we developed data driven techniques.
- There are different approaches for estimation of jet-to-tau misidentified hadronic τ decays:
  1. Fake factor method:
     - Fully data-driven.
     - Example: $H^{\pm} \to \tau\nu$ analysis (arXiv:1807.07915).
  2. Fake rate method:
     - Data driven efficiency factor applied to MC
     - Examples: BSM A/H/Z$\to \tau\tau$ (10.1007/JHEP11(2014)056).

Fake-Factor method: (FF determination)



$$FF = \frac{N(\text{fakes passed ID})}{N(\text{Anti tau})}$$

- The *tau* candidate matching a true hadronic *tau* dacay, an electron or a muon at generator level must be subtracted.
- Fake-factors are usually measured in bins.(e.g. p_T, number of tracks) They can also be measured in opposite- or same-sign regions, with or without b-jets, depending of the topology of interest in the analysis.

Considering q/g jet composition

- There can be one or several CR(s) where FFs are measured, for one CR, one must ensure that the fake τ composition is close to the one in the signal region (SR). Otherwise, one should measure FFs in several CRs that have different fake τ compositions and then combine them.
- Usually, FFs are measured in CRs enriched in either gluon-initiated or quark-initiated jets, as the probability for a hadronic jet to fake a τ depends on its origin.($\tau_{had-vis}$ jet width and Charged track multiplicity)
- In the case where two (or more arXiv:1808.00336) CRs are used, and if one is enriched in gluon-initiated jets, FF for each bin:

$$FF = \alpha_g \times FF(g) + [1 - \alpha_g] \times FF(other(s))$$

In that case, one only needs to compute the fraction of gluon-initiated jet events in the SR-like anti-τ region

Application of FFs

- Define an anti-τ region, which is similar to the signal region but where a τ candidate fails the ID-requirement, instead of fulfilling it.
- In a bin i, the number of events with a j$\rightarrow \tau$ fake is

$$N^{\tau}_{fakes}(i) = N^{anti-\tau}_{fakes}(i) \times FF(i),$$

$$N^{anti-\tau}_{fakes}(i) = N^{anti-\tau}_{fakes}(data, i) - N^{anti-\tau}_{fakes}(MC, \tau \neq j, i)$$

Example of $H^{\pm} \rightarrow \tau\nu$ analysis

- Two control region with different jet compositions are used in order to determine the rate of the fake $\tau_{had-vis}$ objects.
  1. Multi-jet CR (dominated by gluon-initiated jets)
  2. W+jet CR (dominated by quark-initiated jets)
- In the anti-$\tau_{had-vis}$ regions, the fractions of quark- and gluon-initiated jets misidentified as $\tau_{had-vis}$ candidates are measured using a template-fit approach, based on variables that are sensitive to the difference in quark- and gluon-fractions between these two types of jets

Combined Fake Factor in $H^{\pm} \rightarrow \tau\nu$ analysis

- Chosen variables : $\tau_{had-vis}$ identification BDT output score for 3-track and the $\tau_{had-vis}$ jet width for 1-track
- For each bin, two binned templates, denoted f_{MJ} (Multijet CR) and f_W (W+jet CR) , are obtained in their corresponding CRs.
- Their fractional contribution in the SR is determined using a template fit to the respective distributions in the anti tau SR:

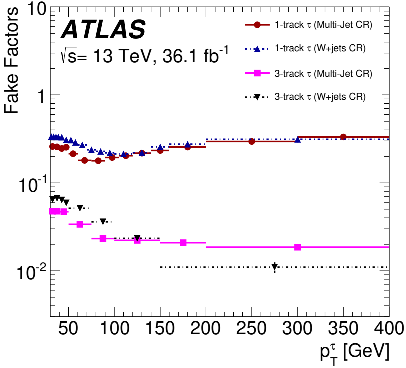
$$f(x|\alpha_{MJ}) = \alpha_{MJ} \times f_{MJ} + (1 - \alpha_{MJ}) \times f_W$$

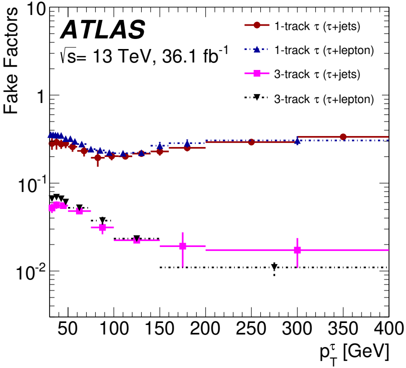
- α_{MJ} is a free parameter.
- From the best fit values of α_{MJ} , combined FF are given by :

$$FF^{comb}(i) = \alpha_{MJ}(i) \times FF^{MJ} + (1 - \alpha_{MJ}(i)) \times FF^{W}$$

Fake Factors from $H^{\pm} \rightarrow \tau\nu$ analysis

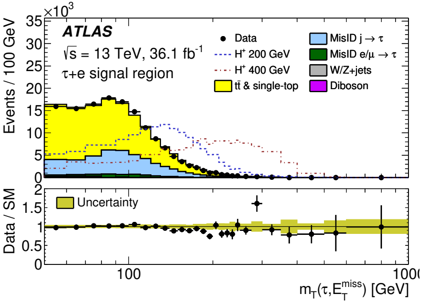
Fake factors parameterized as a function of p_T^{τ} and number of tracks, in the left plot in the multi-jet and w+jet CRs and errors represent the statistical uncertainties, in the right plot after reweighting by α_{MJ} in the $\tau_{had-vis}$+jets and $\tau_{had-vis}$+lepton channel, and it is with additional systematic uncertainties obtained from the combination in a given p_T^{τ} bin.

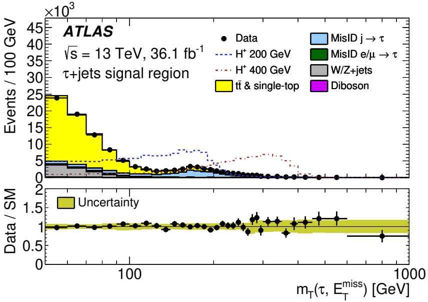




Validation of the background modelling in $H^{\pm} \rightarrow \tau\nu$ analysis

Distribution of m_T ($\tau_{had-vis}$,E_T^{miss}) in the two signal regions, (a) $\tau_{had-vis}$+electron, (b) $\tau_{had-vis}$+jets





Fake-Rate method

- The fake rates are defined as ratios of event yields with identified τs and the ones with τ candidates without identification applied. They are applied to non-true τ objects in a signal-like region in MC.
- This is a semi-data-driven method as fake rates are applied to simulated events.
- After subtraction of events where $\tau \neq$j, fake rates are measured in dedicated CRs as:

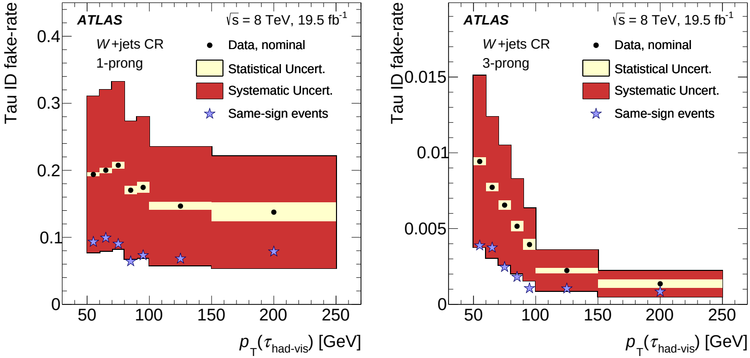
$$FR = \frac{N_{\tau-ID}(data) - N_{\tau-ID}(MC, \tau \neq j)}{N_{\tau-noID}(data) - N_{\tau-noID}(MC, \tau \neq j)}$$

- Usually parameterized in bins of number of tracks, p_T, and η.

Fake-Rate in high-mass resonances decaying to $\tau\tau$ analysis

10.1007/JHEP07(2015)157

Tau-ID fake-rate measured in W($\mu\nu$)+jets data events for the BDT loose, The fake-rate is parameterized in the charge product of the muon and fake tau candidate.
Opposite-sign events are depicted by black circles and same-sign events by blue stars.
The systematic uncertainty covers differences due to jet composition and is added to the statistical uncertainty in quadrature identification working point.



Systematic Uncertainties $H^{\pm} \to \tau\nu$

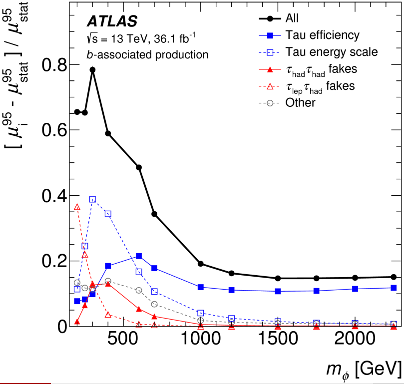
Source of systematic uncertainty	Impact on the expected limit (stat. only) in % $m_{H^+} = 170\ GeV$	$m_{H^+} = 1000\ GeV$
Experimental		
luminosity	2.9	0.2
trigger	1.3	<0.1
$\tau_{\text{had-vis}}$	14.6	0.3
jet	16.9	0.2
electron	10.1	0.1
muon	1.1	<0.1
E_T^{miss}	9.9	<0.1
Fake-factor method	20.3	2.7
Υ modelling	0.8	–
Signal and background models		
$t\bar{t}$ modelling	6.3	0.1
W/Z+jets modelling	1.1	<0.1
cross-sections ($W/Z/VV/t$)	9.6	0.4
H^+ signal modelling	2.5	6.4
All	52.1	13.8

The dominant sources of systematic uncertainty of Fake factor method:

- The requirement $\tau_{had-vis}$ BDT output score in the anti-$\tau_{had-vis}$ definition.
- The contamination of true $\tau_{had-vis}$ candidates fulfilling the anti-$\tau_{had-vis}$ selection (varied by 50%).
- The statistical uncertainty of the control sample.
- The statistical error on the best-fit value of α_{MJ}

Systematic Uncertainties in high-mass resonances decaying to $\tau\tau$ analysis (JHEP 01(2018)055)

- The uncertainty in the fake-rates used to weight simulated non-multijet events in the $\tau_{had}\tau_{had}$ channel is dominated by the limited size of the fakes regions and can reach 40%



Pros and cons

- Fake factor method
  - It is universal and precise. (estimate entire background from all sources)
  - The q/g jet composition needs to be known in CRs and SR.
- Fake rate method
  - In the FR method, the statistical precision of the estimate is enhanced. (since all the events are considered in the estimation)
  - It is only applied to the background modeled by MC.

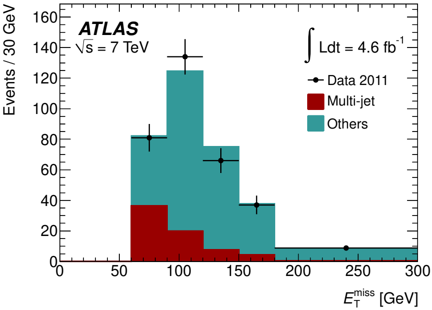
Optimal strategy depends on the specific analysis! Simulation or data driven, or a combination? Which data driven method.

Summary

- Two most commonly used methods for estimation of misidentified hadronic τ decays in ATLAS analyses have been presented: fake-factor method, fake-rate. (There are also other methods: ABCD and template-fit method)
- Example application of $H^{\pm} \rightarrow \tau\nu$ and BSM A/H/Z $\rightarrow \tau\tau$ were shown
- While the Fake Factor method appears the most generic, the actual choice depends on the type of background dominating a prior analysis.

Template-fit method

- The multi-jet component in the SR estimated by fitting MJ template shape to data.
- The template shape extracted from data in MJ-enriched CR.
- In the $H^{\pm} \rightarrow \tau\nu$ analysis(arXiv:1204.2760) the multi-jet background were estimated by fitting its E_T^{miss} shape(and the E_T^{miss} shape of other backgrounds) to data.



Backup slides

Regions for fake-factor measurements in $H^{\pm} \rightarrow \tau\nu$ analysis

Multi-jet CR
number of jet at least 2
$E_{Tmiss} < 80$ GeV
bjets veto
electron and muon veto
p_T of $\tau > 30$ GeV
The transverse mass m_T of τ(τ_met_mt) >50 GeV
$BDTJetSigTrans\ score > 0.02$

W+jets CR
one electron or muon
at least one reconstructed τ_{hadvis} candidate
p_T of electron and muon > 30 Gev
bjets veto
60< $m_T(l, missingET) < 160$ Gev
$BDTJetSigTrans\ score > 0.02$

τ jet width definition

$$w_{\tau} = \frac{\Sigma[p_T^{track} \times \Delta R(\tau_{had-vis}, track)]}{\Sigma p_T^{track}} \tag{1}$$

ABCD method

- One needs two uncorrelated variables (Var1 and Var2), each passing or failing a specific cut, e.g. pass or fail the tau-ID, the charge correlation (OS or SS), below or above a transverse mass threshold, etc.
- The data-set is divided in four regions depending on whether or not each variable passes or fails its cut
- Let B be the signal region. A and B differ from the cut on Var1, C and D differ from the cut on Var2. Each region contains a different fraction of signal, which must be subtracted (the same applies to all other backgrounds if they are estimated with other methods, e.g. simulation)
- The fake-tau background in region B can be computed as

$$N_B^{bkg} = N_A^{bkg} \times \frac{N_D^{bkg}}{N_C^{bkg}} \tag{2}$$

Analyses using the fake-rate method:

- A/H/$Z' \rightarrow \tau\tau$(hadhad): multi-jet background estimated with fake factors, others (W+jets,t$\bar{t}$) fake rates applied to simulation.
- The hh $\rightarrow bb\tau\tau$(hadhad): a fake-rate method is used to estimate t$\bar{t}$ where at least one of the taus is fake.

Validation of the background modelling in $H^\pm \to \tau\nu$ analysis(2)

Distribution of m_T $(\tau_{had-vis}, E^T_{miss})$ in the signal region : $\tau_{had-vis}$+muon