PROCEEDINGS OF SCIENCE

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

Marzieh Bahmani, on behalf of the ATLAS Collaboration[*]

Institute of Nuclear Physics, Polish Academy of Sciences.

E-mail: marzieh.bahmani@ifj.edu.pl

In analyses using reconstructed τ leptons, estimation of backgrounds arising for jets misidentified as hadronically decaying τs becomes a crucial issue. This paper presents two methods employed by the ATLAS experiment at the LHC, the fully data-driven fake factor method and the semi-data-driven fake rate method. Example of fake factor method in the background modelling of $H^{\pm} \rightarrow \tau\nu$ analysis and the fake rate method applied to the search for high mass resonances decaying to $\tau\tau$, in the $\tau_{\text{had}}\tau_{\text{had}}$ channel, are presented. The systematics associated with the methods are also discussed.

Prospects for Charged Higgs Discovery at Colliders - CHARGED2018
25-28 September 2018
Uppsala, Sweden

[*] Speaker.

1. Introduction and motivation

An important source of background for physics analyses using particle-level identification criteria, is misidentification. One of the primary motivation for estimating the fake τ background in the τ related analyses is that jets genuinely fake hadronically decaying τ leptons, and these analyses suffer from such a background. The fake τ background is not well modeled by Monte Carlo (MC) therefore, data driven techniques had to be developed. There are different approaches to the estimation of jet-to-τ misidentified contribution to a τ selection. In this paper, we focus on two main methods used in the ATLAS experiment [1]: the fake factor method and the fake rate method. The former method is fully data-driven, while the latter is semi-data-driven since the data-driven efficiency factor is applied to MC. In section 2 the general ideas behind the fake factor method are presented and the example analysis searching for $H^{\pm} \rightarrow \tau\nu$ [2] using the fake factor method is shown. Section 3 describes the fake rate method and an example of its application in the search for high mass resonances decaying to $\tau\tau$ [4], in the $\tau_{\mathrm{had}}\tau_{\mathrm{had}}$ channel is shown.

2. Fake factor data-driven method

The dominant background processes can be categorized based on the object that gives rise to reconstructed and identified hadronically decaying τ candidate [3]. These are mostly quark- or gluon-initiated jets fulfilling selection criteria of the signal region. These backgrounds are poorly modelled due to the statistical limitations in the sample of simulated events (e.g. multi-jet processes). Also the systematic uncertainties related to object misidentified as τ are not well known. Therefore, a data-driven approach is used to estimate this background. In the fake factor (FF) method, background processes where a quark- or gluon-initiated jet is reconstructed and identified as a τ candidate are estimated from data. For this purpose, an anti-τ selection is defined by requiring the τ candidate to fail the identification criteria of the nominal selection. The fake factor is defined as the ratio between the number of jets reconstructed as τ candidates and fulfilling the nominal τ identification criteria to the number of corresponding candidates failing the identification criteria (anti-τ) and is measured in a dedicated control region (CR) enriched with fake τs:

$$\mathrm{FF} = \frac{\mathrm{N}^{\mathrm{CR}}_{\tau-\mathrm{ID}}(\mathrm{data}) - \mathrm{N}^{\mathrm{CR}}_{\tau-\mathrm{ID}}(\mathrm{MC}, \tau \neq \mathrm{j})}{\mathrm{N}^{\mathrm{CR}}_{\mathrm{anti}-\tau-\mathrm{ID}}(\mathrm{data}) - \mathrm{N}^{\mathrm{CR}}_{\mathrm{anti}-\tau-\mathrm{ID}}(\mathrm{MC}, \tau \neq \mathrm{j})} \tag{2.1}$$

In order to obtain the $\mathrm{N}^{\mathrm{CR}}_{\mathrm{anti}-\tau}$ and $\mathrm{N}^{\mathrm{CR}}_{\tau}$, contribution from true τ events in either categories are subtracted using simulation. The fake factors are usually measured in bins of p_{T} or number of associated tracks in the τ hadronic decay (1-prong, 3-prong), they can also be measured in opposite- or same-sign regions, with or without b-jets, depending on the topology of interest for the analysis.

2.1 Considering quark-gluon jet composition

The fake factors are usually extracted in control regions enriched in either gluon-initiated or quark-initiated jets, as the probability for a hadronic jet to fake a τ depends on its origin. Depending on the analysis, there can be one or several control regions where the fake factors are measured. In case there is only one control region, one must ensure that the origin of fake τ composition is

Multi-jet CR	W+jets CR
number of jet at least 2	one electron or muon
$E_{\mathrm{T}}^{miss} < 80$ GeV	at least one reconstructed $\tau_{\mathrm{had-vis}}$ candidate
bjets veto, electron and muon veto	bjets veto
p_{T} of $\tau > 30$ GeV	p_{T} of electron and muon > 30 Gev
$m_{\mathrm{T}}(\tau\ , E_{\mathrm{T}}^{miss})$ >50 GeV	60< $m_{\mathrm{T}}(\ell, E_{\mathrm{T}}^{miss}) < 160$ Gev
τ identification score > 0.02	τ identification score > 0.02

Table 1: Control regions for fake factor measurement in $H^{\pm} \to \tau\nu$ analysis [2].

close to the one in the signal region. When two (or more [6]) control regions are used, and one is enriched in gluon-initiated jet, the FF for each bin is calculated as follows:

$$\mathrm{FF} = \alpha_{\mathrm{g}} \times \mathrm{FF(g)} + [1 - \alpha_{\mathrm{g}}] \times \mathrm{FF(other(s))} \tag{2.2}$$

where FF(g) and FF(other(s)) are the fake factor for gluon-initiated jet and other control regions, and α is the fraction of reflecting composition in the signal like anti-τ region. Therefore, one needs to estimate the α.

In order to estimate the yield of fake τ background in the signal region, an anti-τ region is defined identical to the signal region but where τ candidate fails the τ identification requirement, instead of fulfilling it. Then in a bin i, the number of events with a jet misidentified as τ is given by :

$$\mathrm{N}^{\tau}_{\mathrm{fakes}}(\mathrm{i}) = \mathrm{N}^{\mathrm{anti-\tau-ID}}_{\mathrm{fakes}}(\mathrm{i}) \times \mathrm{FF(i)}, \tag{2.3}$$

2.2 Fake factor method in $H^{\pm} \to \tau\nu$

In the $H^{\pm} \to \tau\nu$ analysis [2], in order to account for different sources of misidentified hadronically decaying τ lepton ($\tau_{\mathrm{had-vis}}$) [3] in the signal region, fake factors are measured in two control regions of the data with different fractions of quark- and gluon-initiated jets, and then they are combined. The first control region with a significant fraction of gluon-initiated jets (multi-jet CR) is defined as shown in Table 1(left), Such events are collected using a combination of multi-jet triggers. The other control region enriched in quark-initiated jets (W+jets CR) is defined as shown in Table 1(right), and using single lepton trigger. The transverse mass m_{T} of the τ candidate is obtained, as a function of the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$ and the reconstructed τ momentum by eq. 2.4:

$$m_{\mathrm{T}} = \sqrt{2 p_{\mathrm{T}}^{\tau} E_{\mathrm{T}}^{miss} (1 - cos\Delta\phi_{\tau,miss})} \tag{2.4}$$

In the second control region the transverse mass variable $m_{\mathrm{T}}(\ell, E_{\mathrm{T}}^{miss})$ is computed as in the previous case using the lepton p_{T} and separation in azimuthal angle from the missing transverse momentum of the event. The fake factors measured in these two control regions are shown in Figure 1 left plot. In the anti-$\tau_{\mathrm{had-vis}}$ regions corresponding to the nominal event selections, the fraction of quark- and gluon-initiated jets misidentified as $\tau_{\mathrm{had-vis}}$ candidates are then measured using a template-fit approach, based on variables that are sensitive to the difference in quark- and gluon-induced jets. For 3-prong $\tau_{\mathrm{had-vis}}$ candidates, the τ identification score (based on the multivariate BDT approach) is used as a template. For 1- prong $\tau_{\mathrm{had-vis}}$ candidates, the $\tau_{\mathrm{had-vis}}$ jet width is used which is defined as follow:

$$w_\tau = \frac{\Sigma[p_{\mathrm{T}}^{track} \times \Delta R(\tau_{\mathrm{had-vis}}, track)]}{\Sigma p_{\mathrm{T}}^{track}} \tag{2.5}$$

where the sum runs over the tracks satisfying $\Delta R(\tau_{\mathrm{had-vis}}, track) < 0.4$. In order to account for unknown gluon- and quark-initiated jets composition in the signal region, a linear combination of the two templates is defined as :

$$\mathrm{f}(\mathrm{x}|\alpha_{\mathrm{MJ}}) = \alpha_{\mathrm{MJ}} \times \mathrm{f}_{\mathrm{multi-jet}}(\mathrm{x}) + (1 - \alpha_{\mathrm{MJ}})\mathrm{f}_{\mathrm{W+jets}}(\mathrm{x}) \tag{2.6}$$

with a free parameter α_{MJ} and the f(x) is the $\tau_{\mathrm{had-vis}}$ jet width or the τ identification score. $\mathrm{f}_{\mathrm{multi-jet}}$ and $\mathrm{f}_{\mathrm{W+jets}}$ are two binned templates obtained in the multi-jet and W+jets control regions defined above, respectively. This linear combination is fitted to the normalized distribution measured in the signal region, by varying the α_{MJ} and in every bin of p_{T} minimizing the χ^2 distribution for each channel separately. Finally, from the best fit values of α_{MJ} , combined fake factors are obtained by:

$$\mathrm{FF}^{\mathrm{comb}}(\mathrm{i}) = \alpha_{\mathrm{MJ}}(\mathrm{i}) \times \mathrm{FF}^{\mathrm{multi-jet}}(\mathrm{i}) + (1 - \alpha_{\mathrm{MJ}}) \times \mathrm{FF}^{\mathrm{W+jets}}(\mathrm{i}) \tag{2.7}$$

where i refers to each bin in the parametrization of fake factor. The combined fake factors, used in the $\tau_{\mathrm{had-vis}}$+jets and $\tau_{\mathrm{had-vis}}$+lepton signal regions are shown in Figure 1 right plot.

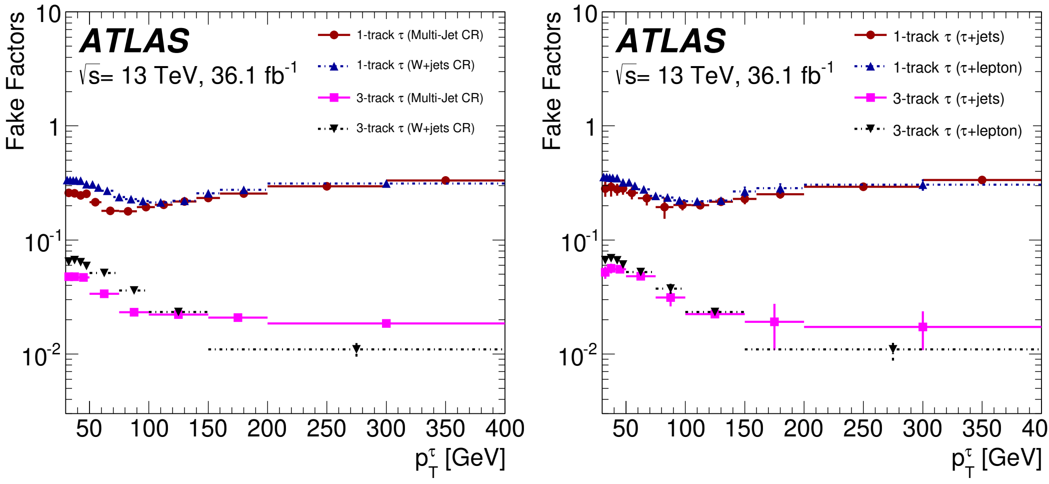

Figure 1: Fake factors parameterized as a function of p_{T}^{τ} and number of tracks. The left plot shows the fake factor in the multi-jet and w+jet CRs. Errors represent the statistical uncertainties. The right plot shows fake factors after reweighting by α_{MJ} in the $\tau_{\mathrm{had-vis}}$+jets and $\tau_{\mathrm{had-vis}}$+lepton channel. [2].

The dominant sources of systematic uncertainty of fake factor method are coming from a) the range of the $\tau_{\mathrm{had-vis}}$ identification score in the anti-$\tau_{\mathrm{had-vis}}$ definition of control samples, which modifies the corresponding fraction of quark- and gluon-initiated jets, as well as the event topology. b) the contamination of true $\tau_{\mathrm{had-vis}}$ candidates fulfilling the anti-$\tau_{\mathrm{had-vis}}$ selection. c) the statistical uncertainty of the control sample. d) the statistical error on the best fit value of α_{MJ}. The impact of systematic uncertainty is different according to the H^+ mass. For low and intermediate mass of H^+ the dominant source of systematic uncertainty is caused by fake factor method, while in the high mass range it is caused by signal modelling as the contribution of background is smaller.

3. Fake rate method

Fake rates are defined as ratios of event yields with identified τs to the yields of all 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, since fake rates are applied to simulated events. Fake rates are measured in dedicated control region as:

$$\mathrm{FR} = \frac{\mathrm{N}_{\tau-\mathrm{ID}}(\mathrm{data}) - \mathrm{N}_{\tau-\mathrm{ID}}(\mathrm{MC}, \tau \neq \mathrm{j})}{\mathrm{N}_{\mathrm{noID}-\tau}(\mathrm{data}) - \mathrm{N}_{\mathrm{noID}-\tau}(\mathrm{MC}, \tau \neq \mathrm{j})} \tag{3.1}$$

where MC events in which a reconstructed τ is associated with a true τ at the generator level are subtracted. Fake rates are usually parameterized in bins of number of tracks, p_{T} and η. Examples of analysis using fake rate method are $A/H/Z' \rightarrow \tau\tau$(had had) [4], hh$\rightarrow bb\tau\tau$ (had had) [6] and LFV $Z' \rightarrow l\tau$ [7].

3.1 Fake rate in high mass resonances decaying to $\tau\tau$

In the search for high mass resonances decaying to $\tau\tau$ [4], in the $\tau_{\mathrm{had}}\tau_{\mathrm{had}}$ channel, background originating from quark- and gluon-initiated jets that are misidentified as hadronic τ decays in processes other than multi-jet production are estimated using simulation. Simulated jets misidentified as hadronically decaying τ are weighted by the fake rates. This not only ensures the correct fake yield, but enhances the statistical precision of the estimate, since the events which were failing the identification are not removed. The fake rate $\mathrm{FR}_{\mathrm{tau-ID}}$, for both leading and sub-leading τ candidate, is defined as the ratio of number of τ candidates that pass a τ identification score cut, $\mathrm{N}^{\mathrm{pass,tau-ID}}$, to the total number of τ candidates, $\mathrm{N}^{\mathrm{total}}$, the only difference is that leading τ candidates should also pass the single τ trigger requirement. It is measured in W+jets events as:

$$\mathrm{FR}_{\mathrm{tau-ID}}(p_{\mathrm{T}}, \mathrm{N}_{\mathrm{track}}) = \frac{\mathrm{N}^{\mathrm{pass,tau-ID}}(p_{\mathrm{T}}, \mathrm{N}_{\mathrm{tarck}})}{\mathrm{N}^{\mathrm{total}}(p_{\mathrm{T}}, \mathrm{N}_{\mathrm{tarck}})}\Big|_{W+jets} \tag{3.2}$$

Events in the W+jets control region are selected by a single-muon trigger with a p_{T} threshold of 36 GeV. They are required to contain one isolated muon with $p_{\mathrm{T}} > 40$ GeV matched to the object that passed the trigger. There must be no additional muons or electrons and at least one τ candidate with opposite charge to the muon. Later all simulated events for background processes other that multi-jet events, processed by the search analysis, are assigned a weight given by:

$$w_{MC} = \prod_{i \in \{lead, sub-lead\}} (1 - \delta^i[1 - FR^i_{\mathrm{tau-ID}}(p^i_{\mathrm{T}}, \mathrm{N}^{\mathrm{i}}_{\mathrm{track}}]) \tag{3.3}$$

where δ^i is 1 if the τ candidate originates from a jet and 0 otherwise.

The uncertainty in the fake rates used to weight simulated non-multi-jet events in the $\tau_{\mathrm{had}}\tau_{\mathrm{had}}$ channel is dominated by the limited statistics of the fake regions and can reach 40%. The requirement of opposite charge between muon and τ candidate enhance the quark-initiated jet composition. To evaluate the systematic uncertainty from applying these fake rates to simulate samples with different jet origin, the fake rate are also calculated for the same sign events which have higher fraction of gluon initiated jets, resulting in lower fake rates as shown in Figure 2. A relative uncertainty of 60% is assigned to cover the range of the measured fake rates for events with opposite- or same-sign τ candidates. The uncertainty is omitted for W+jets events as they are expected to have the same jet composition as events in the control region.

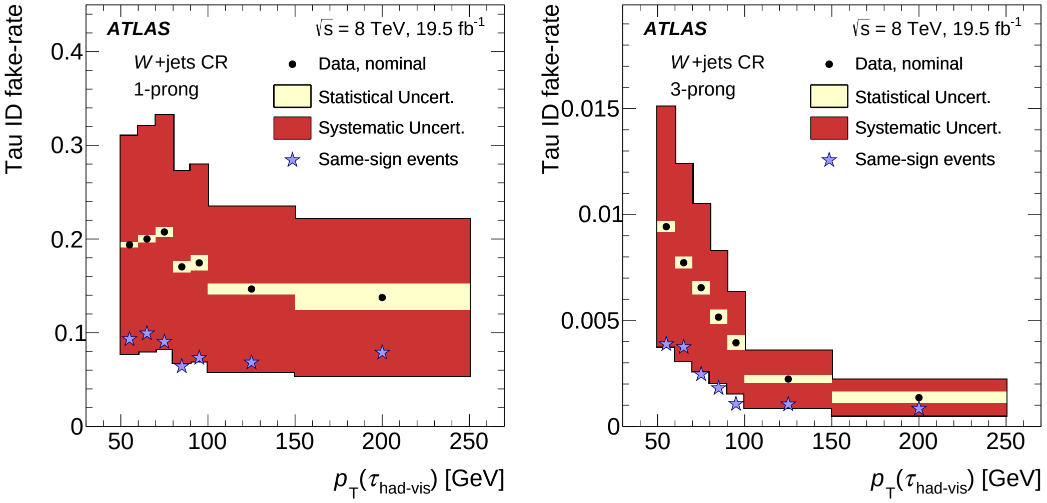

Figure 2: Tau-ID fake-rate measured in W($\mu\nu$)+jets data events, shown separately for muons of the same sign and opposite sign to the reconstructed τ candidate. Opposite-sign events are depicted by black circles and same-sign events by blue stars. The systematic uncertainty covers differences due to jet composition and is added to the statistical uncertainty [5].

4. Conclusion

The most commonly used approaches for the estimation of the contamination from misidentified hadronic τ decays in ATLAS analyses are the fake factor and fake rate methods. The fake factor method is universal and precise. It estimates entire background from all sources, however in this method the relative quark/gluon composition of jets in control and signal regions needs to be known. The fake rate method is a semi-data-driven approach and is applied to Monte Carlo samples, hence it can be used for estimating backgrounds which are modelled by MC. In this method the statistical precision of the estimate is enhanced, since the events failing the τ identification are not discarded. The choice of the optimal strategy for the determination of the background, among the two described above, depends on the specific analysis and it can even be a combination of them.

References

[1] ATLAS Collaboration, 2008 JINST 3 S08003.

[2] ATLAS Collaboration, JHEP 1809 (2018) 139.

[3] ATLAS Collaboration, Eur.Phys.J. C75 (2015) no.7, 303

[4] ATLAS Collaboration, JHEP 1801 (2018) 055.

[5] ATLAS Collaboration, JHEP 1507 (2015) 157.

[6] ATLAS Collaboration, Phys. Rev. Lett. 121, 191801 (2018).

[7] ATLAS Collaboration, Phys. Rev. D 98, 092010 (2018).