PROCEEDINGS OF SCIENCE

The Universal Fake Factor method: an innovative approach to background estimation of jets misidentified as τ-leptons with the ATLAS detector

Federico Morodei on behalf of the ATLAS Collaboration

INFN Sezione di Roma,
Piazzale Aldo Moro 2, Rome, Italy
Dipartimento di Fisica, Università La Sapienza,
Piazzale Aldo Moro 5, Rome, Italy

E-mail: federico.morodei@uniroma1.it

Processes involving τ-leptons play a very important role in the physics program of the ATLAS experiment at the Large Hadron Collider, being fundamental both for Standard Model precision measurements and for new physics searches. The estimation of the background from jets misidentified as τ-leptons is a tough challenge for physics analyses due to the poor reliability of Monte Carlo simulations in modeling this background. This contribution describes an innovative technique to predict this background, the Universal Fake Factor method, which refines and generalizes the commonly used data-driven fake factors method. With this new method, fake factors are computed as a linear combination of fake factors measured in samples enriched in different sources of jets misidentified as τ-leptons. The Universal Fake Factor method presents several advantages: it does not require dedicated analysis control regions, provides a robust way to evaluate systematic uncertainties and can be easily used by a wide range of analyses.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

https://pos.sissa.it/

PoS(EPS-HEP2025)535

1. Introduction

Due to the very important role that τ-leptons play in a wide range of analyses from precision measurements of Standard Model parameters to searches for beyond Standard Model processes, the ATLAS experiment [1] at the Large Hadron Collider (LHC) is improving the methods used to reconstruct and identify this kind of particle and in general to deal with events involving τ-leptons [2, 3]. In this scenario, an innovative technique has been recently developed to obtain a precise estimate of the background from jets misidentified as τ-leptons, the Universal Fake Factor (UFF) method [4].

Because of their very short lifetime (2.9×10^{-13} s), τ-leptons can be detected by ATLAS only through their decay products, which are leptons (35% branching fraction) or hadrons (65% branching fraction). Analyses with hadronically decaying τ-leptons, which are reconstructed only through their visible decay products ($\tau_{\text{had-vis}}$), include, among the main sources of background, those deriving from jets misidentified as $\tau_{\text{had-vis}}$ (commonly labelled as *fakes*). The main sources producing fake $\tau_{\text{had-vis}}$ are jets originating from a b-quark (b), a light quark (q), a gluon (g) or pile-up (p fakes). Fakes from these categories have different properties and therefore different probabilities to be misidentified as a $\tau_{\text{had-vis}}$.

Since Monte Carlo (MC) simulations do not describe the fake background with high accuracy, data-driven techniques are usually preferred by analyses to estimate it. A data-driven method that has been successfully employed by several ATLAS and CMS analyses is the fake factor (FF) method [5]. For a given signal region (SR), the FF method predicts the fake background contribution using data from a control region (CR) through the application of transfer factors measured in a dedicated determination region (DR). Assuming a SR with exactly one $\tau_{\text{had-vis}}$ (ID SR), the CR is defined with the same selections except for an inverted requirement on the $\tau_{\text{had-vis}}$ identification, and therefore it is labelled the antiID SR. The number of fake events in the ID SR ($N_{\text{ID}}^{\text{fake}}$) can be obtained from the number of fake events in the antiID SR times a constant, called fake factor (FF), which depends on $\tau_{\text{had-vis}}$ properties like the transverse momentum (p_{T}) and the number of matched tracks (prongness). The number of fake events in the antiID SR is estimated as the number of data events ($N_{\text{antiID}}^{\text{data}}$) after subtracting the real τ-lepton contamination obtained from MC simulation ($N_{\text{antiID}}^{\text{real}}$), so that $N_{\text{ID}}^{\text{fake}}$ is given by

$$N_{\text{ID}}^{\text{fake}} = \left(N_{\text{antiID}}^{\text{data}} - N_{\text{antiID}}^{\text{real}}\right) \times \text{FF}. \tag{1}$$

The FF is measured in a fake-enriched DR, which is divided into two subregions, depending on whether the $\tau_{\text{had-vis}}$ passes the identification (ID DR) or not (antiID DR), and is given by

$$\text{FF} = \frac{N_{\text{ID DR}}^{\text{data}} - N_{\text{ID DR}}^{\text{real}}}{N_{\text{antiID DR}}^{\text{data}} - N_{\text{antiID DR}}^{\text{real}}}. \tag{2}$$

Also in Equation 2, the real τ-lepton contribution is subtracted from the data in the two DR subregions. The DR has to be defined to be as similar as possible to the SR in order to obtain the same composition of fake $\tau_{\text{had-vis}}$ sources ($q/g/b/p$).

PoS(EPS-HEP2025)535

2. The Universal Fake Factor method

The UFF method generalizes the FF method and remedies its main drawback, namely finding a DR with the same $q/g/b/p$ composition as the SR. Instead of finding a suitable DR for the considered SR, the UFF methods determines the fractions of the four fake $\tau_{\text{had-vis}}$ sources in the antiID SR. Then it computes the fake factor specific for the SR combining universal fake factors previously measured in four fake-enriched regions which are valid for all analyses. All the studies reported in this paper have been obtained using 140 fb^{-1} of proton-proton collision data recorded with the ATLAS detector at $\sqrt{s} = 13$ TeV.

Four regions are defined, each enriched in one fake $\tau_{\text{had-vis}}$ category. The $Z(\mu\mu)$ region, defined by exactly two opposite-charge muons with combined mass between 76 and 106 GeV, is enriched in q fakes. The $t\bar{t}$ region, defined by the presence of an electron and a muon of opposite charges and at least one b-jet, is enriched in b fakes. Two multijet regions are defined by requiring at least two jets which are back-to-back in transverse plane: the MJ hJVT region, requiring a $\tau_{\text{had-vis}}$ candidate with a jet-vertex tagger [6] score above 0.8, is enriched in g fakes, while the MJ lJVT region, with the opposite JVT request, is enriched in p fakes. Two further regions are defined to evaluate the systematic uncertainties of the UFF method: the $W(\mu\nu)$ region requires exactly one muon and one $\tau_{\text{had-vis}}$ with same charges, while the $\mu 3j$ region has the same selections requirements as the $W(\mu\nu)$ region plus a requirement of at least three jets, with at least one being b-tagged. The $q/g/b/p$ fake composition of the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT, MJ lJVT, $W(\mu\nu)$ and $\mu 3j$ antiID subregions, determined using MC simulations, is reported in Figure 1(a).

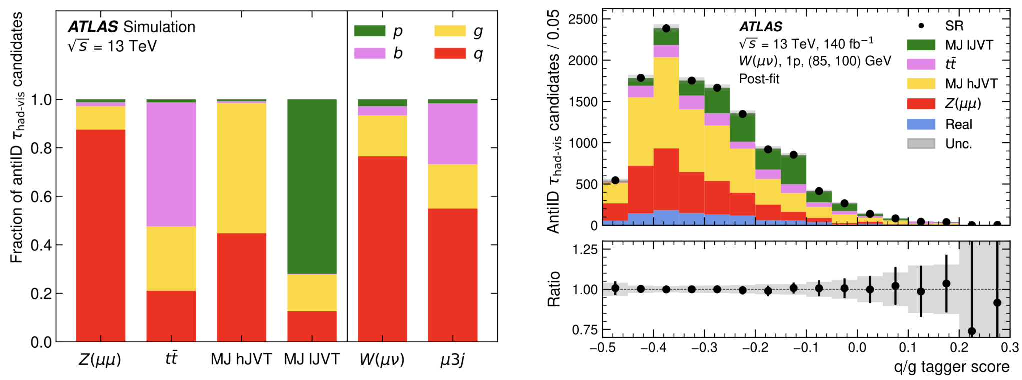

Figure 1: (a) The $q/g/b/p$ fake composition of the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT, MJ lJVT, $W(\mu\nu)$ and $\mu 3j$ antiID subregions. (b) Template fit results in the 1-prong and $p_{\text{T}} \in (85, 100)$ GeV bin of the $W(\mu\nu)$ region [4].

The UFF method estimates the mixture of the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT and MJ lJVT antiID subregions contributing to the antiID SR. The coefficients representing the relative contribution of each subregion are determined performing a template fit of the q/g tagger score [7] distributions. The template distributions are built using data from which the real τ-lepton contamination, estimated with MC simulations, is subtracted and are binned in p_{T} and prongness of the $\tau_{\text{had-vis}}$. The sum of the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT and MJ lJVT templates plus the histogram of the real τ-lepton con-

PoS(EPS-HEP2025)535

tamination in the antiID SR is then fitted to the data in the antiID SR using a maximum-likelihood method, independently in each p_T and prongness bin. Figure 1(b) gives as an example the results of the template fit in the 1-prong and $p_T \in (85, 100)$ GeV bin of the $W(\mu\nu)$ region, with the q/g tagger score templates of the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT and MJ lJVT antiID subregions scaled by the corresponding post-fit normalization factors.

The normalization factors μ_X, with $X = Z(\mu\mu)$, $t\bar{t}$, MJ hJVT or MJ lJVT, obtained from the template fit are then used as weights in a linear combination of the FFs measured in the four regions (shown in Figure 2(a)) to compute the combined fake factor ($\mathrm{FF_{UFF}}$) suitable for the SR:

$$\mathrm{FF_{UFF}} = \mu_{Z(\mu\mu)} \times \mathrm{FF}_{Z(\mu\mu)} + \mu_{t\bar{t}} \times \mathrm{FF}_{t\bar{t}} + \mu_{\text{MJ hJVT}} \times \mathrm{FF}_{\text{MJ hJVT}} + \mu_{\text{MJ lJVT}} \times \mathrm{FF}_{\text{MJ lJVT}} \tag{3}$$

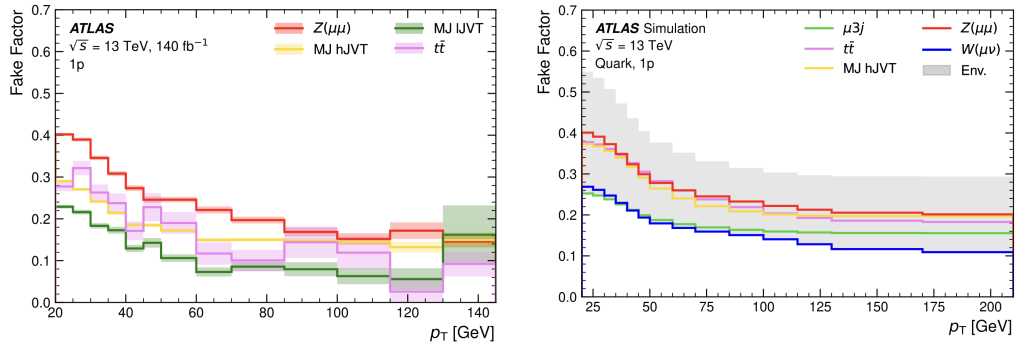

Figure 2: (a) FFs measured in the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT and MJ lJVT regions for 1-prong fake $\tau_{\text{had-vis}}$ as a function of the fake $\tau_{\text{had-vis}}$ p_T. (b) FFs for the 1-prong q fakes as a function of the fake $\tau_{\text{had-vis}}$ p_T [4].

3. Uncertainties

A statistical uncertainty is associated with the $\mathrm{FF_{UFF}}$, due to the finite size of the samples used to evaluate the FFs and the fit coefficients μ_X. The uncertainties on the FFs are considered uncorrelated in the uncertainty propagation to the $\mathrm{FF_{UFF}}$, while the uncertainties on μ_X are correlated and propagated using the Hessian covariance matrix V. Therefore, the total statistical uncertainty on $\mathrm{FF_{UFF}}$ is given by

$$\sigma_{\text{UFF, stat}} = \sqrt{\sum_X \mu_X^2 \sigma_{\mathrm{FF}_X}^2 + \sum_{X,Y} \mathrm{FF}_X \mathrm{FF}_Y V_{XY}}, \tag{4}$$

where X and Y run over the $Z(\mu\mu)$, $t\bar{t}$, MJ hJVT and MJ lJVT regions.

The systematic uncertainties for the UFF method are estimated as follows. In each of the six regions defined in the previous paragraph, the FFs for the q, g and b fakes are computed using MC samples, while the FFs for the p fakes are computed using data. As an example, Figure 2(b) shows the FFs for the q fakes in the different regions. The symmetrized uncertainty envelope shown in

Figure 2(b) is given by the largest difference between the $Z(\mu\mu)$ MC FFs and the others. This is then divided by the $Z(\mu\mu)$ MC FF to obtain the relative systematic uncertainty. A similar procedure is used to get the uncertainty of the $t\bar{t}$, MJ hJVT and MJ lJVT FFs. The systematic uncertainty is propagated to $\mathrm{FF_{UFF}}$ independently for each source. One FF of Equation 3 is varied at a time, both upwards and downwards, and each time the $\mathrm{FF_{UFF}}$ is recomputed obtaining the $\pm 1\sigma$ $\mathrm{FF_{UFF}}$ variations.

4. Validation

The $W(\mu\nu)$ region, highly enriched in fakes, can be used to test the validity of the UFF method. The $\mathrm{FF_{UFF}}$ are compared to the FFs evaluated directly as the ratio between ID and antiID events (Figure 3(a)). The $\mathrm{FF_{UFF}}$ agree with the reference FFs within the uncertainty band. For further validation, the fake-background predicted with the UFF method is compared with the data in the $W(\mu\nu)$ ID region. The sum of the real τ-lepton contribution estimated from MC samples and the fake-background agrees with the data within the uncertainties in all the tested variables. Figure 3(b) shows the data-prediction comparison for the transverse mass m_T.

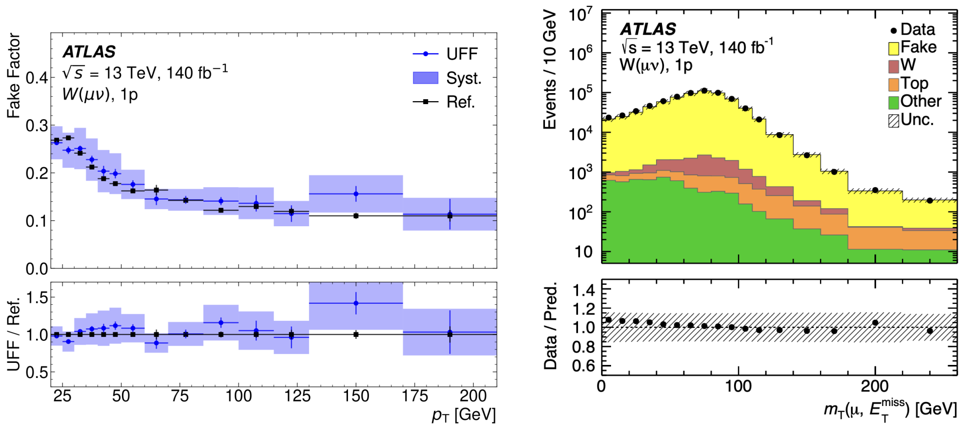

Figure 3: (a) Comparison between the $\mathrm{FF_{UFF}}$ and the reference FFs for 1-prong fakes in the $W(\mu\nu)$ region. (b) Comparison of the 1-prong fake-background prediction plus MC real τ-lepton contribution with the data in the $W(\mu\nu)$ region for the transverse mass between the muon and the missing transverse momentum [4].

5. Conclusion

The UFF method is a valid and easy-to-use technique developed by the ATLAS experiment at the LHC to estimate the fake-$\tau_{\text{had-vis}}$ background. In order to apply this method, the only input required is the q/g tagger score distributions and there is no need to find a suitable region similar to the SR to evaluate the FFs. The UFF method is best suited to analyses with τ-lepton p_T below 150 GeV. The UFF method has already been applied successfully by ATLAS analyses with both one $\tau_{\text{had-vis}}$ [8] and two $\tau_{\text{had-vis}}$ in the final state [9, 10].

References

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

[2] ATLAS Collaboration, *Reconstruction and identification of pairs of collimated τ-leptons decaying hadronically using $\sqrt{s}$ = 13 TeV pp collision data with the ATLAS detector*, Eur. Phys. J. C **85** (2025) 561, arXiv: 2411.09357 [hep-ex].

[3] ATLAS Collaboration, *Improved reconstruction of highly boosted τ-lepton pairs in the $\tau\tau \to (\mu\nu_\mu\nu_\tau)(hadrons + \nu_\tau)$ decay channels with the ATLAS detector*, Eur. Phys. J. C **85** (2025) 706, arXiv: 2412.14937 [hep-ex].

[4] ATLAS Collaboration, *Estimation of backgrounds from jets misidentified as τ-leptons using the Universal Fake Factor method with the ATLAS detector*, (2025), arXiv: 2502.04156 [hep-ex].

[5] K. Lehmann and B. Stelzer, *The Fake Factor Method and its relation to the Matrix Method*, Nucl. Instrum. Meth. A **1054** (2023) 168376.

[6] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector*, Eur.Phys. J. **C76** (2016) 581, arXiv: 1510.03823 [hep-ex].

[7] ATLAS Collaboration, *Performance and calibration of quark/gluon-jet taggers using* 140 *fb^{-1} of pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Chin. Phys. C **48** (2024) 023001, arXiv: 2058-6132 [hep-ex].

[8] ATLAS Collaboration, *Search for charged Higgs bosons decaying via $H^\pm \to \tau^\pm\nu_\tau$ in the τ+jets and τ+lepton final states with* 36 *fb^{-1} of pp collision data recorded at $\sqrt{s}$ = 13 TeV with the ATLAS experiment*, JHEP **09** (2018) 139, arXiv: 1807.07915 [hep-ex].

[9] ATLAS Collaboration, *Search for dark matter produced in association with a Higgs boson decaying to tau leptons at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, JHEP **09** (2023) 189, arXiv: 2305.12938 [hep-ex].

[10] ATLAS Collaboration, *A measurement of the high-mass $\tau\bar{\tau}$ production cross-section at $\sqrt{s}$ =* 13 *TeV with the ATLAS detector and constraints on new particles and couplings*, (2025), arXiv: 2503.19836 [hep-ex].