PERFORMANCE OF τ LEPTON IDENTIFICATION IN ATLAS 7 TeV DATA* **

Andrzej Zemla

on behalf of the ATLAS Collaboration

The Henryk Niewodniczański Institute of Nuclear Physics
Polish Academy of Sciences
Radzikowskiego 152, Kraków, Poland

(Received May 6, 2011)

The reconstruction and identification of τ leptons are important in physics analysis of Standard Model (SM) processes and searches for new phenomena. In this paper the comparison of distributions of identification variables between data and Monte Carlo samples is done, using the initial dataset collected by the ATLAS detector at a centre-of-mass energy of $\sqrt{s} = 7$ TeV, corresponding to an integrated luminosity of 244 nb^{-1}. As the number of recorded τ leptons is small, the background jets reconstructed as τ-lepton candidates can be used to check the algorithm performance. The background efficiency suppression is tested on a QCD dijet data sample using cut-based τ identification criteria. Multivariate methods, such as boosted decision trees and projective likelihood, are also tested to check their background rejection performance.

DOI:10.5506/APhysPolB.42.1717
PACS numbers: 14.60.Fg, 29.90.+r, 07.05.Kf, 07.05.Tp

1. Introduction

The τ lepton, with a mass of $\tau = 1776.84 \pm 0.17$ MeV [1], is the only SM lepton heavy enough to decay both leptonically and hadronically. It decays approximately 65% of the time to one or more hadrons and 35% of the time leptonically. The reconstruction and identification of τ leptons are important in the analysis of Standard Model processes and searches for

* Presented at the Cracow Epiphany Conference on the First Year of the LHC, Cracow, Poland, January 10–12, 2011.

** Work supported in parts by the Polish Government grant 2899/B/H03/2010/38 (years 2010–2012), the Polish–DAAD collaboration 764/N-DAAD/2010/0 and the Polish Government grant NN202127937 (years 2009–2011).

new phenomena. They can appear in final states in the production of Higgs bosons, supersymmetric (SUSY) particles, and other particles not described by the SM [2]. Standard Model processes, such as W, Z boson and $t\bar{t}$ production, can also result in signatures with τ leptons, and events from these processes can be used to study τ lepton reconstruction in detail and establish τ lepton identification efficiency. Particularly challenging in identifying hadronically decaying τ leptons is to distinguish them from hadronic jets which are produced in processes with very large cross-sections. However, some properties of τ lepton decays can be used to differentiate them from jets. Hadronically decaying τ leptons decay to one charged pion (1-prong) with a branching ratio (BR) of approximately 77% and to three charged pions (3-prong) with a BR of approximately 23%. The decay products tend to be well collimated and the invariant mass of the visible decay products is usually smaller than those of jets. The proper decay length of the τ lepton is 87 μm, so decay vertices can be resolved from the primary vertex by the silicon tracker.

2. Datasets and event selection

The studies presented here are based on data collected with the ATLAS detector [3] at a centre-of-mass energy of $\sqrt{s} = 7$ TeV, corresponding to an integrated luminosity of approximately $\mathcal{L} = 244$ nb^{-1} [4, 5]. It was required that the data used in the analysis have been taken during stable LHC beam conditions, and pass data quality requirements for the inner detector (tracker) and the calorimeter. Furthermore, all events must satisfy the following criteria:

- at the Level 1 trigger it is required that the τ trigger object passes a 5 GeV threshold cut [6],
- there are no "bad" jets in the event [7] caused by out-of-time cosmic events or sporadic noise effects in the calorimeters,
- at least one vertex reconstructed with more than four tracks is present,
- at least one τ candidate with $p_{\mathrm{T}} > 30$ GeV (fully calibrated, as described in Section 3) and $|\eta| < 2.5$, and a second τ candidate with $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.5$ (also fully calibrated) are present. Those two candidates are required to be separated by at least 2.7 radians in azimuth (the angle in the plane transverse to the beam pipe).

The cuts listed above aim at selecting events with back-to-back jets, enriching the sample with fake τ candidates from QCD jet processes that form the primary background to signatures such as $Z \to \tau\tau$, in order to study fake τ candidate properties. To minimize the bias due to the trigger requirement, the sample of studied τ candidates excludes the leading (with

higher p_{T}) τ candidate. With these requirements the selected data sample contains about 2.9 million events and 3.9 million τ candidates. QCD dijet MC samples are used for comparison. These samples are generated with PYTHIA [8] and passed through a GEANT4 [9] simulation of the ATLAS detector [10]. The MC samples used here employ the DW tune [11] which uses virtuality-ordered partonic showers and which was derived to describe the CDF II underlying event and Drell–Yan data. The DW tune seems to model the forward activity of the underlying event better than the MC09 tune, and describes jet shapes and profiles in data more accurately. When showing distributions for true τ lepton candidates, a $Z \rightarrow \tau\tau$ MC sample with the MC09 tune is used.

3. τ reconstruction and identification algorithms

For more detailed description of τ reconstruction see [12] or [13]. Hadronically decaying τ leptons are reconstructed starting from either calorimeter or track seeds. *Track-seeded* candidates have a seeding track with $p_{\mathrm{T}} > 6$ GeV satisfying quality criteria. *Calorimeter-seeded* candidates consist of calorimeter jets reconstructed with the anti-k_{t} algorithm [14] starting from topological clusters (topoclusters) [15]. The candidate is required to have $E_{\mathrm{T}} > 10\,\mathrm{GeV}$.

Efficient and robust identification methods are necessary to discriminate between the overwhelming QCD background and real, hadronically decaying τ leptons. It is also necessary to use identification algorithms to reject true electrons and muons reconstructed as τ candidates. The ATLAS Collaboration has developed a number of identification methods, including a simple cut-based method, and two multivariate methods: projective likelihood and boosted decision trees (BDT) [16]

4. Discriminating variables

There are several different classes of discriminating variables employed in the ATLAS τ identification algorithms. They are based on three groups of physics properties:

- Shower width (shower radius and isolation);
- Particle multiplicity (*e.g.* number of tracks, clusters);
- Fractions of τ jet candidate energy deposited in the electromagnetic and hadronic calorimeters.

From the detector perspective, it is more appropriate to reclassify those variables according to which feature of the detector is used to construct the discriminating quantity:

- Calorimeter cluster-based variables (*e.g.* number of clusters, mass);
- Tracking variables (*e.g.* track width, track mass);
- Variables which combine calorimeter and tracking information (*e.g.* E/p).

This leads to a relatively long list of variables which could be considered for use in multivariate techniques. However, for early data the priority is to use robust, relatively uncorrelated and well understood variables only:

- **Cluster mass:** Invariant mass computed from associated topoclusters: m_{clusters} ;
- **Track mass:** Invariant mass of the track system: m_{tracks} ;
- **Track radius:** p_{T} weighted track width

$$R_{\text{track}} = \frac{\sum_i^{\Delta R_i<0.2} p_{\text{T},i} \Delta R_i}{\sum_i^{\Delta R_i<0.2} p_{\text{T},i}} ,$$

  where ΔR_i denotes

$$\Delta R_i = \sqrt{(\eta_i - \eta_{\text{clusters}})^2 + (\phi_i - \phi_{\text{clusters}})^2} .$$

- **Leading track momentum fraction:** the ratio between the p_{T} of the leading track and the total visible transverse momentum of the τ candidate, determined from associated calibrated calorimeter clusters

$$F_{\text{trk,1}} = \frac{p_{\text{T,1}}^{\text{track}}}{p_{\text{T}}^{\tau}} .$$

- **Electromagnetic radius:** To exploit the smaller transverse shower profile in τ decays, the electromagnetic radius R_{EM} is used, defined as

$$R_{\text{EM}} = \frac{\sum_i^{\Delta R_i<0.4} E_{\text{T},i}^{\text{EM}} \Delta R_i}{\sum_i^{\Delta R_i<0.4} E_{\text{T},i}^{\text{EM}}} ,$$

  where i runs over all cells in the electromagnetic calorimeter associated to the τ candidate, ΔR is a cone defined relative to the τ-jet seed axis and $E_{\text{T},i}$ is the cell transverse energy.
- **Core energy fraction:** Fraction of transverse energy in the core ($\Delta R < 0.1$) of the τ candidate

$$f_{\text{core}} = \frac{\sum_i^{\Delta R_i<0.1} E_{\text{T},i}^{\text{EM}}}{\sum_i^{\Delta R_i<0.4} E_{\text{T},i}^{\text{EM}}} .$$

- **Electromagnetic fraction:** Fraction of GCW (Global Cell Weighting) [17] calibrated transverse energy of the τ candidate deposited in the EM calorimeter

$$f_{\mathrm{EM}} = \frac{\sum_i^{\Delta R_i<0.4} E_{\mathrm{T},i}^{\mathrm{GCW}}}{\sum_j^{\Delta R_j<0.4} E_{\mathrm{T},j}^{\mathrm{GCW}}} .$$

  The GCW calibration scheme attempts to compensate for the different calorimeter response to hadronic and electromagnetic energy depositions.

Since the instantaneous luminosities for the data used here are low, pile-up effects are expected to be small. With higher luminosity, the pile-up will affect the distributions of these variables for both fake and true τ candidates, thus reducing their separation power. Variables that are more robust under pile-up conditions are also being studied, in preparation for the expected higher instantaneous LHC luminosities. After the selection described in Sec. 2, the number of τ candidates in MC samples is normalized to the number of τ candidates selected in the data. The variable distributions of τ candidates reconstructed in a signal $Z \to \tau\tau$ MC sample and matched to true hadronically decaying τ leptons are also overlaid to show the expected distributions of real τ leptons.

The distributions for the discussed identification variables are shown in Fig. 1 for τ candidates in data and MC samples. The agreement of the distributions for data and MC samples is quite good for all identification variables.

5. Efficiency definitions

The identification algorithms are optimized on signal and background MC samples for minimum background efficiency and tuned for roughly 30% (tight), 50% (medium), and 70% (loose) signal efficiency on the $Z \to \tau\tau$ MC sample for true, hadronically decaying τ leptons. Candidates with exactly one reconstructed track are considered as 1-prong candidates and candidates with two or more reconstructed tracks are considered as 3-prong candidates. Candidates without tracks are excluded. We define 1-prong signal and background efficiencies as

$$\varepsilon_{\mathrm{sig}}^{1-\mathrm{prong}} = \frac{\#\text{matched candidates with 1 track passing cut}}{\#\text{visible hadronic 1-prong Monte Carlo } \tau\text{s}} ,$$

$$\varepsilon_{\mathrm{bkg}}^{1-\mathrm{prong}} = \frac{\#\text{candidates with 1 track passing cut}}{\#\text{candidates with 1 reconstructed track}} ,$$

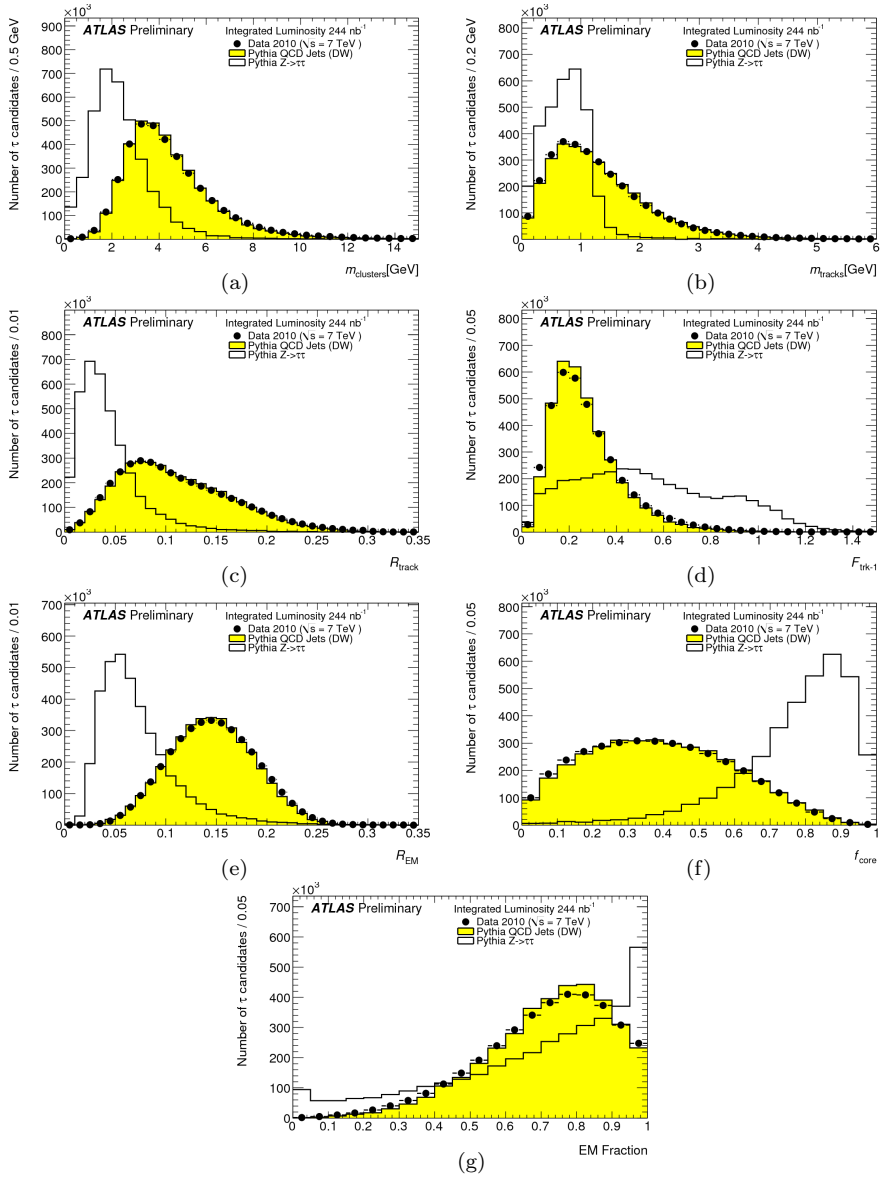

Fig. 1. Distribution of discriminating variables: cluster mass (a), track mass (b), track radius (c), leading track momentum fraction (d), EM radius (e), core energy fraction (f) and electromagnetic fraction of τ candidates (g). The number of τ candidates in MC samples is normalized to the number of τ candidates selected in data. The statistical errors on the MC are negligible.

where signal candidates are matched to visible 1-prong hadronic Monte Carlo τ within ΔR <0.2. A visible Monte Carlo τ has $|\eta|$ <2.5 with $E_{\mathrm{T}}^{\mathrm{vis}}$ > 10 GeV (excluding neutrinos). Signal and background efficiencies for a 3 prong candidate are defined as

$$\varepsilon_{\mathrm{sig}}^{3-\mathrm{prong}} = \frac{\#\text{matched candidates with 3 tracks passing cut}}{\#\text{visible hadronic 3-prong Monte Carlo } \tau\text{s}},$$

$$\varepsilon_{\mathrm{bkg}}^{3-\mathrm{prong}} = \frac{\#\text{candidates with 3 tracks passing cut}}{\#\text{candidates with 3 reconstructed tracks}},$$

where again signal candidates are matched to visible 3-prong hadronic Monte Carlo τ leptons. Since the reconstruction efficiency for 3-prong candidates is only about 70%, defining the loose cut value may be impossible, since the loose is defined to have 70% signal efficiency. In that case, some analyses presented here in (*e.g.* BDT) loosen the 3-prong match requirement to 2 or more tracks when determining the cut values. However, the mathematical definitions given above are strictly enforced when calculating rejections or comparing identification efficiences.

In addition to these prong-specific definitions, we define global efficiencies which do not place constraints on the number of reconstructed tracks or require equality between track and prong multiplicity. The global signal and background efficiencies are defined as

$$\varepsilon_{\mathrm{sig}}^{\mathrm{global}} = \frac{\#\text{matched candidates with at least one track passing cut}}{\#\text{visible hadronic Monte Carlo } \tau\text{s}},$$

$$\varepsilon_{\mathrm{bkg}}^{\mathrm{global}} = \frac{\#\text{candidates with at least one track passing cut}}{\#\text{candidates with at least one track}},$$

where the matched signal candidate lies within $\Delta R < 0.2$ of any visible hadronic Monte Carlo τ leptons, regardless of the number of prongs.

6. Cut based identification performance

Future τ identification in ATLAS will rely on sophisticated multivariate techniques to achieve the necessary rejection of quark and gluon jets. However, in early data, the certified τ identification is based on a simple cut-based approach. This optimization has been done in 7 TeV simulations using the same τ reconstruction version as it was used for data. In the cuts optimization only three variables are used: electromagnetic radius, track radius and the leading track momentum fraction. The optimization procedure uses a cross-section weighted combination of $W \to \tau\nu$ and $Z \to \tau\tau$ Monte Carlo for signal and a cross-section weighted combination of the dijet Monte Carlo samples with p_{T} of the leading outgoing partons in range 8–280 GeV

for background. With only those three variables, the cut procedure explores every combination of cuts with reasonable granularity, calculating the signal and background efficiencies for each combination. Then, it searches for such combinations that gave signal efficiencies at the level of about 70%, 50%, and 30%, and for those background efficiencies are calculated (Fig. 2).

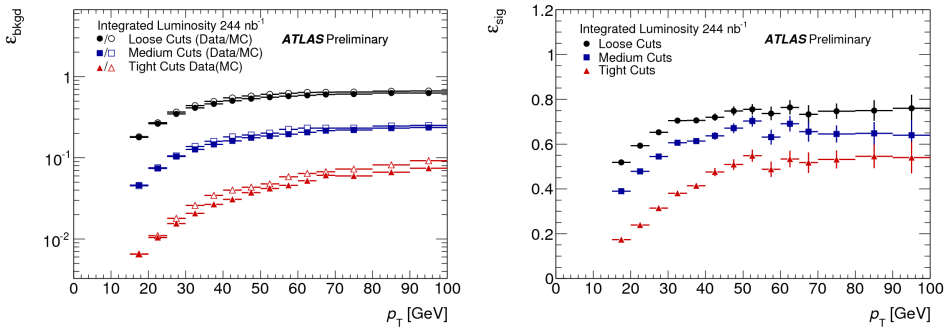

Fig. 2. Background efficiencies obtained for data and MC samples (left) and signal efficiencies predicted by a $Z \to \tau\tau$ MC sample for loose, medium and tight efficiency level (right) as a function of the reconstructed p_{T}^{τ}.

7. Systematic uncertainties for cuts optimization

The systematic uncertainties on the measured background efficiencies are considered from two different effects: the transverse momentum calibration for τ candidates and pile-up effects due to varying beam conditions. These uncertainties may partly account for some of the data-MC discrepancies observed in the high-p_{T} region. The current transverse momentum calibrations are based on the GCW calibration scheme. Different calibration schemes have also been studied, including a simple p_{T} and η dependent calibration (EM + JES) [18]. The variation of the background efficiency was studied by comparing the calibration of τ candidates using the GCW scheme with the EM + JES scheme.

This calibration affects the reconstruction of three identification variables: m_{clusters}, f_{EM}, and $f_{\mathrm{trk,1}}$. Of these, only $f_{\mathrm{trk,1}}$ is used in the cut-based identification. When using the EM + JES calibration, the background efficiency for the cut-based identification decreases by 2.1%, 8.5%, and 9.6% for loose, medium, and tight selections, respectively, and is assigned as a systematic uncertainty. The relative difference in efficiency between the EM+JES and GCW calibrations is shown in Fig. 3, left as a function of p_{T}^{τ}.

Another systematic effect considered is the effect of pile-up due to varying beam conditions. Over the course of the data taking period, relevant for this analysis, the beam intensity increased by a factor of three. Increased beam

intensities lead to different pile-up conditions that affect the distributions of the identification variables. Since the number of vertices n_{vtx} is highly correlated with pile-up activity, the background efficiency was evaluated as a function of n_{vtx}. This is shown in Fig. 3.

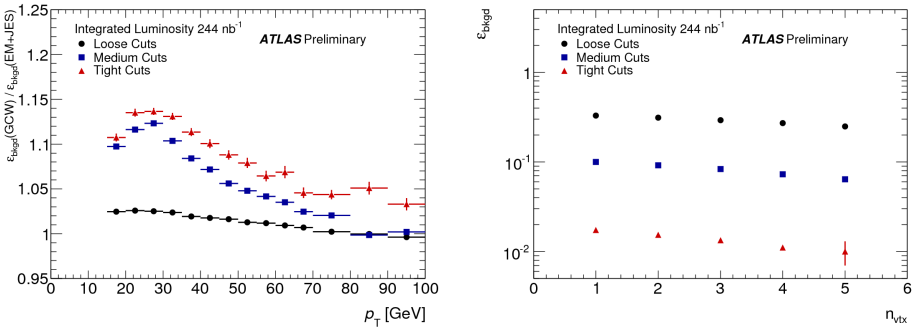

Fig. 3. Ratio of background efficiencies using EM + JES and GCW calibration as a function of p_{T}^{τ} (left) and background efficiencies as a function of n_{vtx} (right).

A systematic uncertainty is determined by taking the mean difference of the background efficiency for τ candidates in events with $n_{\mathrm{vtx}} = 1$ and $n_{\mathrm{vtx}} > 1$ with the background efficiencies obtained from the entire sample. The resulting uncertainty is 5.7% for the loose cut selection, 9.3% for the medium cut selection, and 14.5% for the tight cut selection. Other sources of systematic uncertainties such as beam spot variations, the impact of calorimeter noise, and detector alignment effects were investigated, but found to be small.

8. τ identification performance using multivariate methods

Multivariate techniques of data mining were also used for hadronic τ identification in the ATLAS experiment. Currently, two of them, the Projective Likelihood (LLH) and Boosted Decision Trees (BDT) [16] are optimized for τ leptons identification.

8.1. Projective likelihood

The input probability density functions (PDFs) have been obtained from a mixture of PYTHIA $W \to \tau\nu$, $Z \to \tau\tau$ and $A \to \tau\tau$ events for the signal, and PYTHIA dijet events for the background. The events were simulated in the context of the MC09 production with the default tune. The evaluation of the jet rejection performance on MC is done with PYTHIA dijet samples

using the DW tune. The PDFs are created for 9 bins in p_{T} (in GeV: 10–20; 20–30; 30–45; 45–70; 70–100, 100–150; 150–200; 200–300; and greater than 300) and produced separately for 1-prong and 3-prong candidates; the 1-prong PDFs are further divided by the number of associated π^0 clusters [5].

8.2. Boosted decision trees

This BDT was trained with signal represented by a mixture of $Z \to \tau\tau$ and $A \to \tau\tau$ MC simulation and background represented by PYTHIA dijet MC (default MC09 tune). For the training procedure all variables presented in Sec. 4 are used.

The performance of those methods comparing to rectangular cuts is shown in Fig. 4.

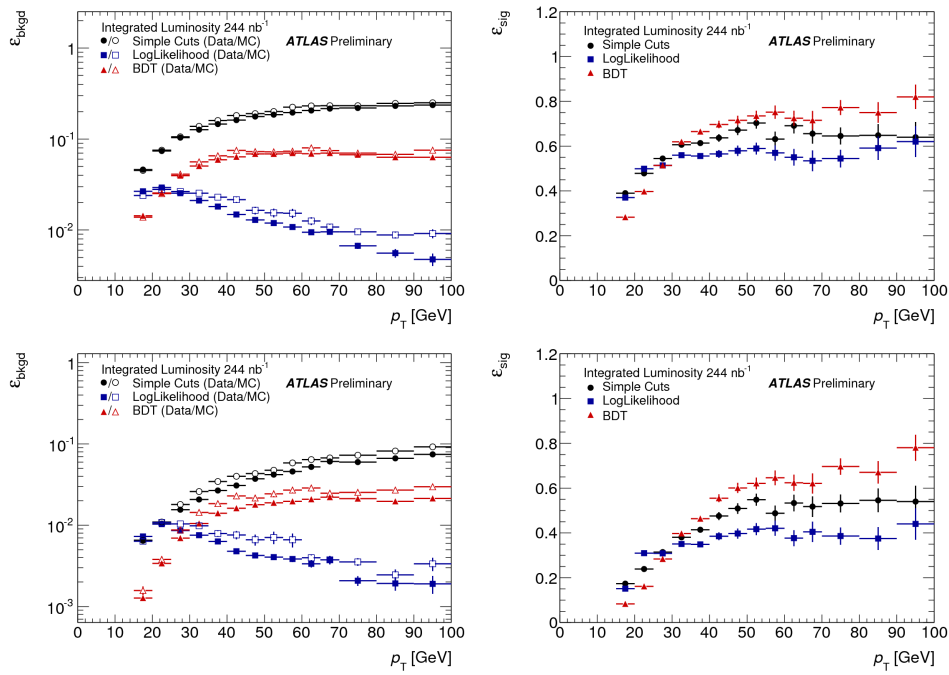

Fig. 4. Background efficiencies in data and MC with medium selection for cut-based, BDT, and LLH identification (top left), signal efficiencies from MC with medium selection for cut-based, BDT, and LLH identification (top right), background efficiencies in data and MC with tight selection for cut-based, BDT, and LLH identification (bottom left) signal efficiencies from MC with tight selection for cut-based, BDT, and LLH identification (bottom right) as a function of p_{T}^{τ}.

9. Summary and conclusions

τ identification has been shown using three different MV approaches: simple cuts, BDT and LLH. In each case, the signal efficiencies were calculated based on MC samples, while the rejection was calculated on both data and simulated dijet samples. Good performance was seen for all three techniques. Systematic uncertainties on the background efficiencies from transverse momentum calibration and pile-up effects were determined. Both data and MC predictions show that the BDT and LLH identification algorithms increase the background rejection power significantly over the cut-based identification. These methods will be tested on $W \to \tau\nu$ and $Z \to \tau\tau$ events in data in the near future.

REFERENCES

[1] C. Amsler *et al.* [Particle Data Group], *Phys. Lett.* **B667**, 1 (2008) and 2009 partial update for the 2010 edition.

[2] [ATLAS Collaboration], Expected Performance of the ATLAS Experiment, Detector, Trigger, and Physics, CERN-OPEN-2008-020, Geneva 2008, pp. 1167–1828.

[3] G. Aad *et al.* [ATLAS Collaboration], *JINST* **3**, S08003 (2008).

[4] [ATLAS Collaboration], Luminosity Determination Using the ATLAS Detector, ATLAS-CONF-2010-060 (2010).

[5] [ATLAS Collaboration], Tau Reconstruction and Identification Performance in ATLAS, ATLAS-CONF-2010-086 (2010).

[6] [ATLAS Collaboration], Performance of the ATLAS τ Trigger in p–p Collisions at $\sqrt{s} = 7$ TeV, in preparation.

[7] [ATLAS Collaboration], Data-Quality Requirements and Event Cleaning for Jets and Missing Transverse Energy Reconstruction with the ATLAS Detector in Proton–Proton Collisions at a Center-of-Mass Energy of $\sqrt{s} =$ 7 TeV, ATLAS-CONF-2010-038.

[8] T. Sjostrand, S. Mrenna, P.Z. Skands, *J. High Energy Phys.* **05**, 026 (2006).

[9] S. Agostinelli *et al.* [GEANT4 Collaboration], *Nucl. Instrum. Methods* **A506**, 250 (2003).

[10] [ATLAS Collaboration], `arXiv:1005.4567v1 [physics.ins-det]`, submitted to *Eur.Phys. J.* **C**.

[11] R. Field, Min-Bias and Underlying Event at the Tevatron and the LHC, talk presented at the Fermilab MC Tuning Workshop (Oct 2002), `http://www-cdf.fnal.gov/physics/conferences/cdf8547/_RDF/_TeV4LHC.pdf`

[12] See talk of M. Wolter, *Acta Phys Pol. B* **42**, 1689 (2011), this issue.

[13] [ATLAS Collaboration], Reconstruction and Identification of Hadronic τ Decays, ATL-PHYS-PUB-2009-000 (2009).

[14] M. Cacciari, G.P. Salam, G. Soyez, *J. High Energy Phys.* **04**, 063 (2008).

[15] [ATLAS Collaboration], Calorimeter Clustering Algorithms: Description and Performance, ATL-LARG-PUB-2008-002 (2008).

[16] Y. Freund, R.E. Schapire, Expermients with a New Boosting Algorithm in: Machine Learning, Proceedings of the Thirteenth International Conference, pp. 148–156, 1996.

[17] [ATLAS Collaboration], Properties of Jets and Inputs to Jet Reconstruction and Calibration with the ATLAS Detector Using Proton–Proton Collision at $\sqrt{s} = 7$ TeV, ATLAS-CONF-2010-053 (2010).

[18] [ATLAS Collaboration], Jet Energy Scale and Its Systematic Uncertainty in ATLAS for Jets Produced in Proton–Proton Collisions at $\sqrt{s} = 7$ TeV, ATLAS-CONF 2010-056 (2010).