ATLAS NOTE

ATLAS-CONF-2010-086

August 16, 2010

Tau Reconstruction and Identification Performance in ATLAS

The ATLAS collaboration

Abstract

Using LHC collisions at a centre-of-mass energy of $\sqrt{s} = 7$ TeV recorded with the ATLAS detector, the performance of the reconstruction and identification algorithms for hadronic τ decays is studied. Although the dataset used here, corresponding to an integrated luminosity of 244 nb^{-1}, contains a small number of real τ leptons, the background jets reconstructed as τ candidates can be used to assess performance aspects of these algorithms. Distributions of identification variables are compared in data and Monte Carlo samples, and the background efficiency of cut-based τ identification criteria is measured in a QCD dijet enriched data sample. Systematic effects on the background efficiency are also estimated. The rejection power of multi-variate τ identification discriminants such as boosted decision trees and projective likelihood methods is also investigated.

1 Introduction

Since March 2010 the ATLAS experiment at LHC has been collecting proton-proton collision data at a centre-of-mass energy of $\sqrt{s} = 7$ TeV. These data are analyzed to study the performance of the τ lepton reconstruction and identification algorithms.

The τ lepton decays 65% of the time to one or more hadrons and 35% of the time leptonically, in both cases with accompanying neutrinos. Because of their short lifetime, it is very difficult to separate τ leptons decaying to electrons or muons from prompt electrons and muons, and τ identification therefore focuses on reconstructing hadronically decaying τ candidates. While the number of true τ leptons present in the dataset considered in this study is expected to be small, the performance of the τ reconstruction and identification algorithms can be assessed using τ candidates from quark or gluon initiated jets in QCD events, which form the primary background to true τ leptons. Distributions of variables used in identifying τ leptons in these events can be compared to predictions from Monte Carlo (MC) simulation. The rejection power of identification algorithms, along with associated systematic uncertainties, can be evaluated on this background sample.

The data and MC samples used and the event selection are described in the following section. The reconstruction algorithm and kinematic and identification variables are presented in Section 3. The performance of τ identification in terms of background rejection is discussed in Section 4. The background efficiencies of τ identification criteria and their systematic uncertainties are measured with data and compared to the MC prediction.

2 Data Samples and Event Selection

The studies presented here are based on data collected with the ATLAS detector [1] at a centre-of-mass energy of $\sqrt{s} = 7$ TeV, corresponding to an integrated luminosity of approximately $\mathscr{L} = 244$ nb^{-1} [2]. The data considered are required to have been taken with 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:

- the Level 1 trigger requiring a τ trigger object passing a 5 GeV threshold [3] is satisfied,
- there are no "bad" jets in the event [4] 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 $|\eta| < 2.5$, as well as another τ candidate with $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.5$ (also fully calibrated). The 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).

This above cuts 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. In order to remove any bias due to the trigger requirement, the sample of τ candidates that are studied excludes the leading 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, where the allowed range of the transverse momenta of the outgoing partons in the rest frame of the hard interaction are restricted to be between 8 and 280 GeV. These samples are generated with Pythia [5] and passed through a Geant4 [6] simulation of the ATLAS detector [7]. In contrast to the MC samples used previously [8] which use the MC09 tune [9], the MC samples used here employ the DW tune [10] which uses virtuality-ordered showers

and 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 $Z \to \tau\tau$ MC sample with the MC09 tune is used.

3 Tau Reconstruction and Identification Variables

Hadronically decaying τ leptons are reconstructed starting from either calorimeter or track seeds [11]. *Track-seeded* candidates have a seeding track with $p_{\mathrm{T}} > 6$ GeV satisfying quality criteria on the number of associated hits in the silicon tracker ($N_{\mathrm{hit}}^{\mathrm{Si}} \geq 7$) and on the impact parameter with respect to the interaction vertex ($|d_0| < 2$ mm and $|z_0| \times \sin\theta < 10$ mm). *Calorimeter-seeded* candidates consist of calorimeter jets reconstructed with the anti-k_t algorithm [12] (using a distance parameter $D = 0.4$) starting from topological clusters (topoclusters) [13]. The candidate is required to have $p_{\mathrm{T}} > 10$ GeV, calibrated using the global cell energy-density weighting (GCW) calibration scheme [14]. Candidates are labelled *double-seeded* when a seed track and a seed jet are within a distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.2$ of each other. The p_{T} of the τ candidate is further adjusted by applying multiplicative factors derived from MC studies, in order to reconstruct the p_{T} of signal τ leptons accurately. In this note, double-seeded candidates and candidates with only a calorimeter-seed with at least one associated track are considered (and referred to generically as τ candidates), as there are very few candidates without a calorimeter seed or an associated track.

The reconstruction of τ candidates provides very little rejection against QCD jet backgrounds. Rejection comes from a separate identification step and is usually based on several discriminating variables. Identification methods for τ candidates include selections based on simple cuts, boosted decision trees, and projective likelihood methods [15, 16] (described in Section 4).

Identification variables used for these methods have shown good separation potential in MC studies. Since a hadronic τ decay is characterized by collimated energy deposits in the calorimeters and one or few collimated tracks, these properties are used to distinguish them from QCD jets. The variables that are used in the identification of τ leptons with early data include:

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:

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

where i runs over all tracks associated to the τ candidate, ΔR_i is defined relative to the τ jet seed axis and $p_{\mathrm{T},i}$ is the track transverse momentum.

Leading track momentum fraction:

$$f_{\mathrm{trk},1} = \frac{p_{\mathrm{T},1}^{\mathrm{track}}}{p_{\mathrm{T}}^{\tau}},$$

where $p_{\mathrm{T},1}^{\mathrm{track}}$ is the transverse momentum of the leading track of the τ candidate and p_{T}^{τ} is the transverse momentum of the τ candidate.

Electromagnetic radius: Transverse energy weighted shower width in the electromagnetic (EM) calorimeter:

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

where i runs over cells in the first three layers of the EM calorimeter associated to the τ candidate, ΔR_i is defined relative to the τ jet seed axis and $E_{\mathrm{T},i}^{\mathrm{EM}}$ is the cell transverse energy.

Core energy fraction: Fraction of transverse energy in the core ($\Delta R < 0.1$) of the τ candidate:

$$f_{\mathrm{core}} = \frac{\sum_i^{\Delta R<0.1} E_{\mathrm{T},i}}{\sum_i^{\Delta R<0.4} E_{\mathrm{T},i}},$$

where i runs over all cells associated to the τ candidate within ΔR_i of the τ jet seed axis.

Electromagnetic fraction: Fraction of GCW calibrated transverse energy of the τ candidate deposited in the EM calorimeter:

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

where $E_{\mathrm{T},i}$ ($E_{\mathrm{T},j}$) is the GCW calibrated transverse energy deposited in cell i (j), and i runs over the cells in the first three layers of the EM calorimeter, while j runs over the cells in all layers of the calorimeter.

Since the instantaneous luminosities for the data used here are low, pile-up effects are expected to be small for the distributions shown here. With higher luminosity, 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 anticipated higher instantaneous luminosities at the LHC.

After the selection described in Section 2, the number of τ candidates in MC samples are normalized to the number of τ candidates selected in data. The shapes from 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 these identification variables are shown in Figure 1 for τ candidates in data and MC samples. Compared to the MC-data comparison shown previously [8], the distributions for R_{EM} and f_{core} are more consistent with the data for the DW MC tune. The agreement of the distributions for data and MC samples is quite good for all identification variables.

4 Background Rejection in QCD events

The identification of τ leptons is applied after the reconstruction step, based on the reconstructed values of identification variables. Three independent identification algorithms have been investigated: simple cuts, boosted decision trees (BDT), and a projective likelihood (LL).

The performance for τ identification algorithms is expressed in terms of two quantities: signal efficiency and background efficiency. The signal efficiency is defined as

$$\varepsilon_{\mathrm{sig}} = \frac{N_{\mathrm{pass,match}}^{\tau}}{N_{\mathrm{match}}^{\tau}},$$

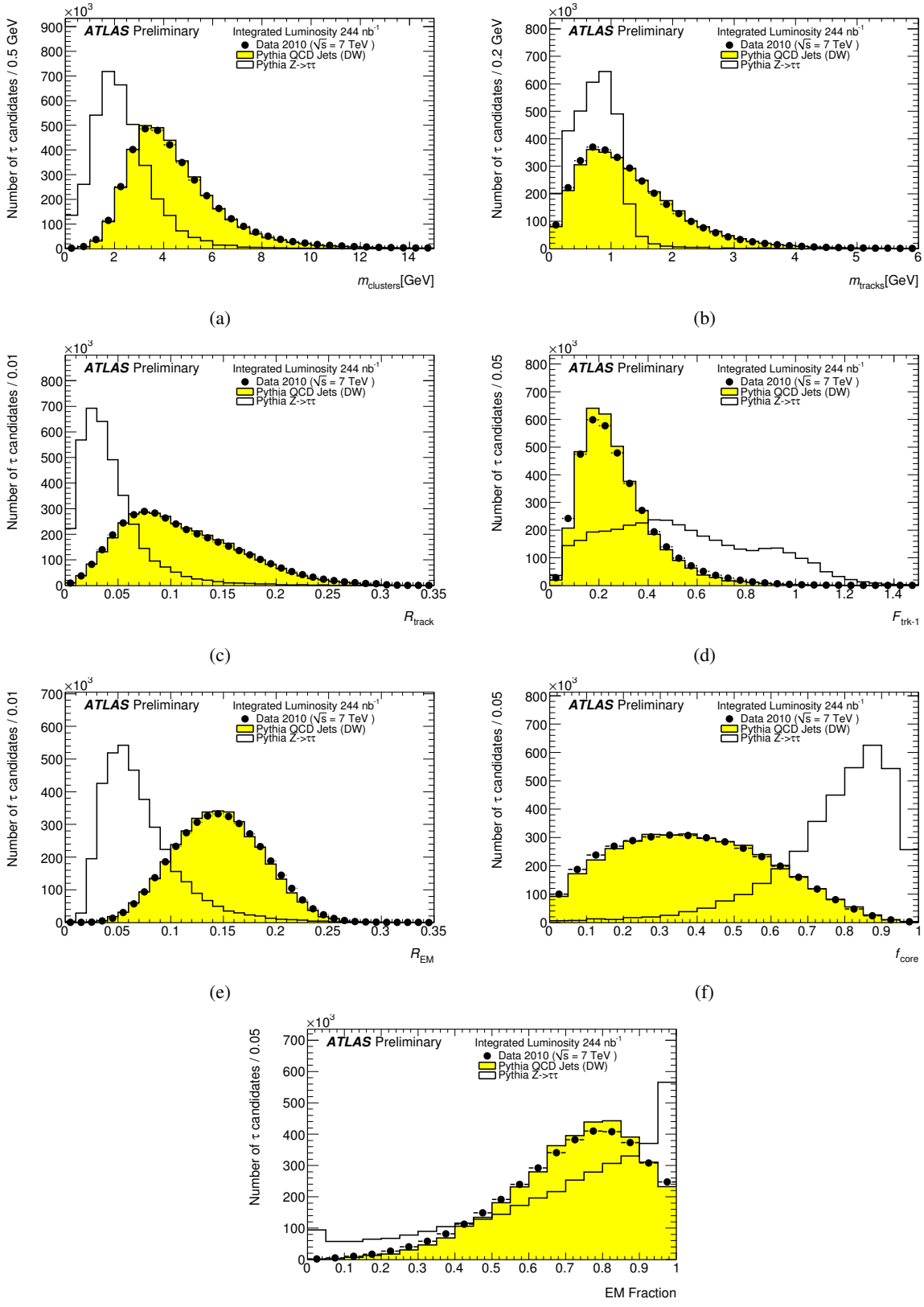


Figure 1: (a) Cluster mass, (b) track mass (c) track radius, (d) leading track momentum fraction, (e) EM radius, (f) core energy fraction, and (g) EM fraction of τ candidates. The number of τ candidates in MC samples are normalized to the number of τ candidates selected in data. The statistical errors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

where N^{τ}_{match} is the number of reconstructed τ candidates that are matched within $\Delta R < 0.2$ of a true, hadronically decaying τ lepton with visible transverse momentum $p_{\text{T}}^{\text{vis}} > 15$ GeV and visible pseudo-rapidity $|\eta^{\text{vis}}| < 2.5$, reconstructed with the correct number of associated tracks; while $N^{\tau}_{\text{pass,match}}$ is the number of these reconstructed candidates that pass the identification criteria. A $Z \to \tau\tau$ MC sample is used to evaluate the signal efficiency.

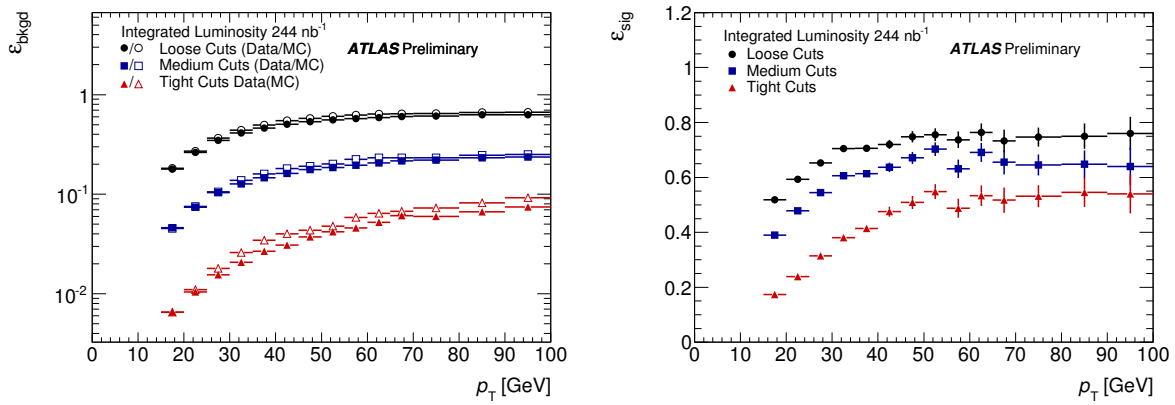
The background efficiency is defined as

$$\varepsilon_{\text{bkgd}} = \frac{N^{\text{bkgd}}_{\text{pass}}}{N^{\text{bkgd}}_{\text{total}}},$$

where $N^{\text{bkgd}}_{\text{pass}}$ is the number of the τ candidates that pass the identification criteria, and $N^{\text{bkgd}}_{\text{total}}$ is the number of τ candidates.

The cut-based identification only uses three relatively uncorrelated variables: R_{EM}, R_{track}, and $f_{\text{trk,1}}$. Cuts are optimized on signal and background MC samples for minimum background efficiency and tuned for roughly 30% signal efficiency (tight), 50% efficiency (medium), and 60% efficiency (loose) on the $Z \to \tau\tau$ MC sample for true, hadronically decaying τ leptons. Different cuts are applied for τ candidates with $n_{\text{track}} = 1$ and those with $n_{\text{track}} \geq 2$.

Figure 2 shows the background efficiencies obtained for data and MC samples as a function of the reconstructed p^{τ}_{T} as well as the signal efficiencies from the $Z \to \tau\tau$ MC sample for loose, medium, and tight selections. The background efficiencies measured in data agree well with the MC prediction, showing the good perfomance with data of the cut-based identification that was optimized with MC samples.



(a) (b)

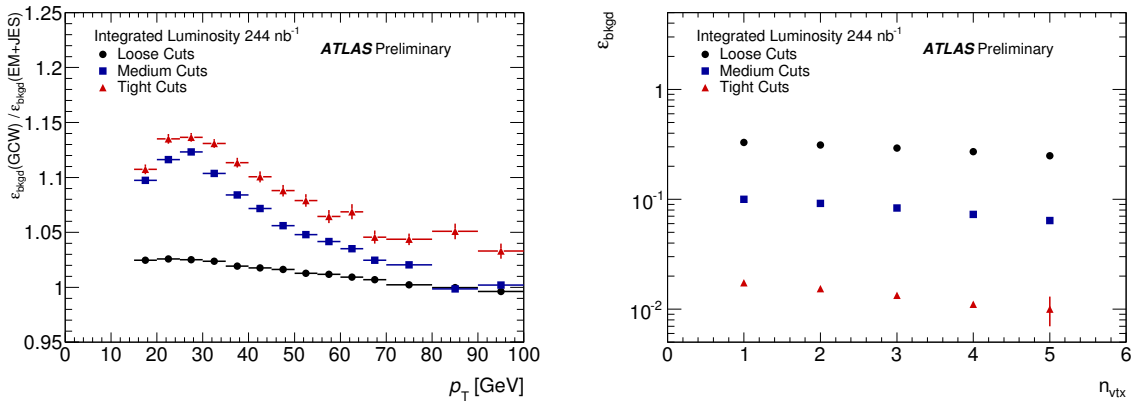
Figure 2: (a) Background efficiencies obtained for data and MC samples as a function of the reconstructed p^{τ}_{T}. (b) Signal efficiencies predicted by a $Z \to \tau\tau$ MC sample as a function of the reconstructed p^{τ}_{T}.

Two 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^{τ}_{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) [17]. 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 of the seven identification variables: m_{clusters}, f_{EM}, and $f_{\text{trk,1}}$, where the cut-based identification only uses the variable $f_{\text{trk,1}}$.

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 assigned as a systematic uncertainty. The relative difference in efficiency between the EM+JES and GCW calibrations is shown in Figure 3(a) as a function of p_{T}^{τ}.



(a) (b)

Figure 3: (a) Ratio of background efficiencies using EM+JES and GCW calibrations as a function of p_{T}^{τ}. (b) Background efficiencies as a function of n_{vtx}.

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 Figure 3(b).

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.

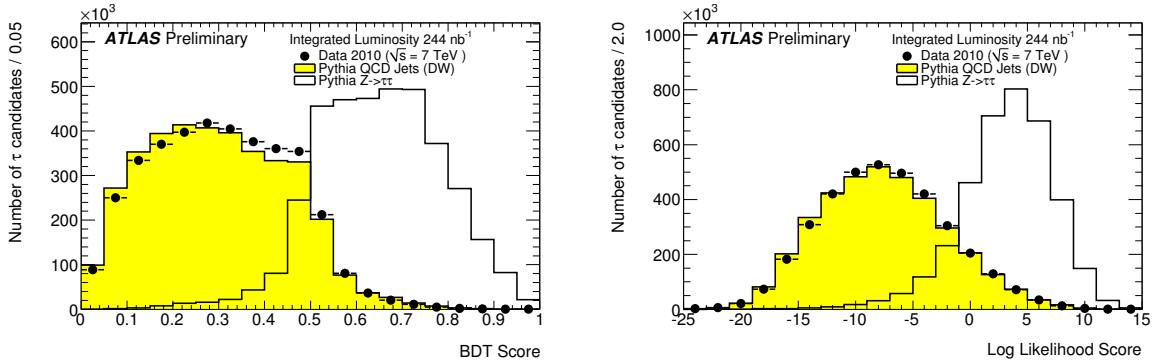
The measured background efficiency for the given data sample is listed in Table 1 along with the corresponding MC DW tune prediction. An alternative background efficiency, $\varepsilon'_{\mathrm{bkgd}}$, is also shown, that requires in addition that τ candidates must have $n_{\mathrm{track}} = 1$ or $n_{\mathrm{track}} = 3$, since many analyses with hadronic τ lepton final states may require this in addition. The uncertainties shown in this table for the data are from the two systematic effects discussed earlier, which are treated as independent. The statistical uncertainties from the data sample are negligible.

Selection	$\varepsilon_{\mathrm{bkgd}}$ (data)	$\varepsilon_{\mathrm{bkgd}}$ (MC)	$\varepsilon'_{\mathrm{bkgd}}$ (data)	$\varepsilon'_{\mathrm{bkgd}}$ (MC)
loose	$(3.2 \pm 0.2) \times 10^{-1}$	3.4×10^{-1}	$(9.4 \pm 0.6) \times 10^{-2}$	10×10^{-2}
medium	$(9.5 \pm 1.0) \times 10^{-2}$	9.9×10^{-2}	$(3.1 \pm 0.4) \times 10^{-2}$	3.3×10^{-2}
tight	$(1.6 \pm 0.3) \times 10^{-2}$	1.9×10^{-2}	$(5.6 \pm 0.9) \times 10^{-3}$	6.8×10^{-3}

Table 1: Background efficiencies for loose, medium, and tight selection cuts. The measured background efficiencies in data are compared to the MC DW tune prediction. Uncertainties for the background efficiencies in data are from transverse momentum calibration and pile-up effects.

The BDT uses all seven variables listed in Section 3, while the LL uses all variables except f_{core}

(due to correlations with other variables), to provide discrimination of τ leptons against jets. The LL is trained using signal and background MC samples split in five separate p_{T} bins, while the BDT does not split its training samples. The distributions for the BDT score and LL score are shown in Figure 4 for τ candidates in data and in the MC samples. The distributions for τ candidates matched to true τ leptons in a $Z \to \tau\tau$ MC sample are also overlaid. The distributions agree reasonably well between data and the MC samples, and demonstrate the strong separation power of these multi-variate discriminants.



(a) (b)

Figure 4: The (a) BDT jet score and (b) LL score for τ candidates in data and MC samples. The number of τ candidates in MC samples is normalized to the number of τ candidates in the data.

The background efficiencies obtained for the BDT and LL are also compared to those obtained for the cut-based identification for medium and tight criteria in Figure 5. The signal efficiencies of the BDT and LL as a function of p_{T}^{τ} are also shown in comparison with the cut-based identification. The increased rejection power against fake τ candidates of the more sophisticated BDT and LL discriminants is evident, although not accurately described by the MC prediction. Differences between MC and data are being investigated.

5 Conclusions

The variables used for τ identification have been investigated in a QCD jet enriched sample using 244 nb^{-1}of data collected by the ATLAS experiment at the LHC. All variables are well described by MC predictions and show good separation power between τ leptons and fake τ candidates from QCD jets. The background efficiency for three cut-based selections (loose, medium, tight) was measured as a function of p_{T}^{τ}, and found to be in good agreement with MC predictions. 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 LL identification algorithms increase the background rejection power significantly over 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] ATLAS Collaboration, G. Aad et al., *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

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

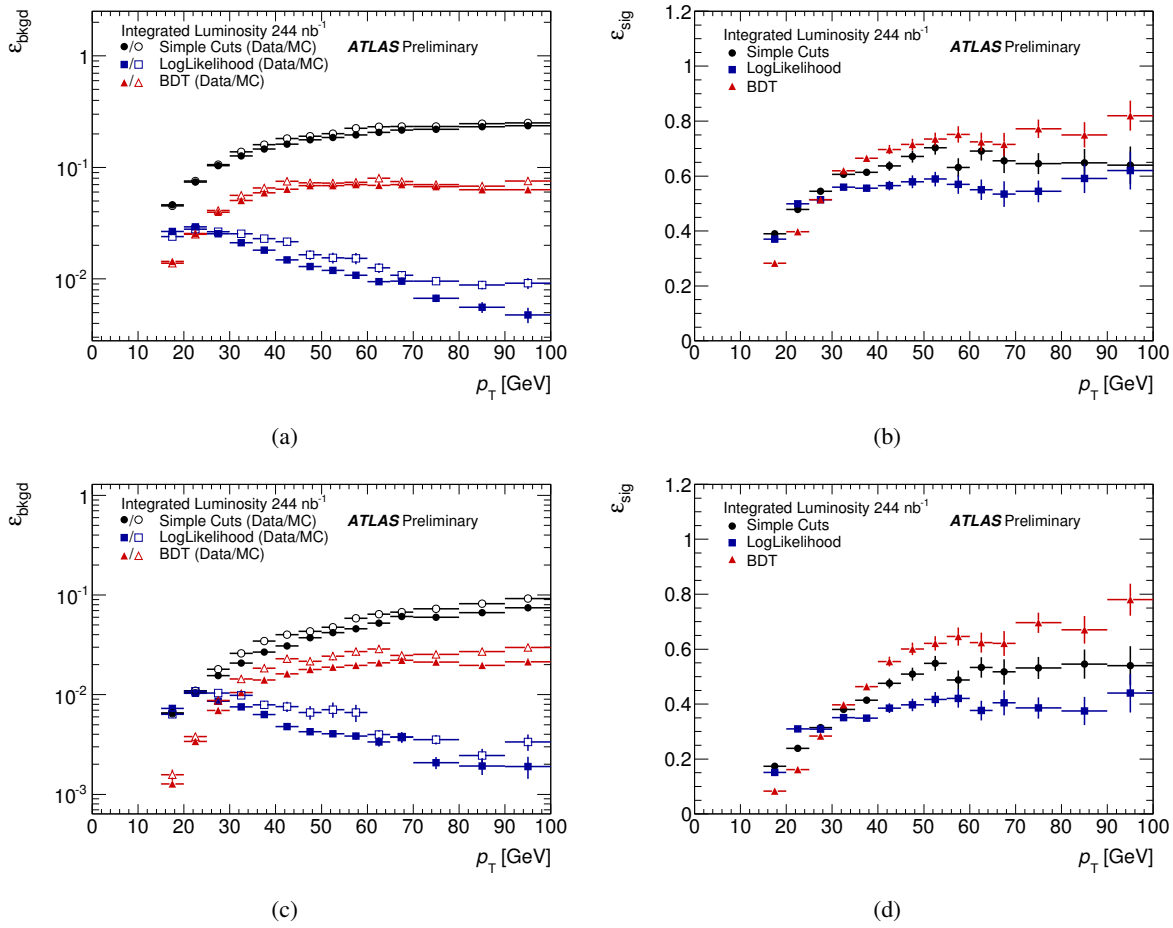


Figure 5: (a) Background efficiencies in data and MC as a function of p_{T}^{τ} with the medium selection for cut-based, BDT, and LL identification. (b) Signal efficiencies from MC as a function of p_{T}^{τ} with the medium selection for cut-based, BDT, and LL identification. (c) Background efficiencies in data and MC as a function of p_{T}^{τ} with the tight selection for cut-based, BDT, and LL identification. (d) Signal efficiencies from MC as a function of p_{T}^{τ} with the tight selection for cut-based, BDT, and LL identification.

[3] ATLAS Collaboration, *Performance of the ATLAS tau trigger in p-p collisions at* $\sqrt{s}$ = 7 TeV, in preparation.

[4] 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 (2010).

[5] T. Sjostrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **05** (2006) 026.

[6] GEANT4 Collaboration, S. Agostinelli et al., *GEANT4: A simulation toolkit*, Nucl. Instrum. Meth. **A506** (2003) 250–303.

[7] ATLAS Collaboration, G. Aad et al., *The ATLAS Simulation Infrastructure*, `arXiv:1005.4568v1 [physics.ins-det]`. Submitted to Eur. Phys. J. C.

[8] ATLAS Collaboration, *Reconstruction of hadronic tau candidates in QCD events at ATLAS with 7 TeV proton-proton collisions*, ATLAS-CONF-2010-059 (2010).

[9] ATLAS Collaboration, *ATLAS Monte Carlo tunes for MC09*, ATL-PHYS-PUB-2010-002 (2010).

[10] 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`.

[11] ATLAS Collaboration, G. Aad et al., *Expected Performance of the ATLAS Experiment – Detector, Trigger and Physics*, CERN-OPEN-2008-020 (2008).

[12] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063.

[13] ATLAS Collaboration, *Calorimeter clustering algorithms: Description and performance*, ATL-LARG-PUB-2008-002 (2008).

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

[15] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. Chapman & Hall, 1984.

[16] Y. Freund and R. Shapire, *Experiments with a New Boosting Algorithm*, in *Proceedings 13th International Conference on Machine Learning*. 1996.

[17] ATLAS Collaboration, *Jet energy scale and its systematic uncertainty for jets produced in proton-proton collisions at* $\sqrt{s}$ = 7 TeV *and measured with the ATLAS detector*, ATLAS-CONF-2010-056.