Tau identification using multivariate techniques in ATLAS

D C O'Neil on behalf of the ATLAS collaboration
Simon Fraser University
E-mail: doneil@sfu.ca

Abstract. Tau leptons play an important role in the physics program of the LHC. They are being used in electroweak measurements, in detector related studies and in searches for new phenomena like the Higgs boson or Supersymmetry. In the detector, tau leptons are reconstructed as collimated jets with low track multiplicity. Due to the background from QCD multijet processes, efficient tau identification techniques with large fake rejection are essential. Since single variable criteria are not enough to efficiently separate them from jets and electrons, modern multivariate techniques are used. In ATLAS, several advanced algorithms are applied to identify taus, including a projective likelihood estimator and boosted decision trees. All multivariate methods applied to the ATLAS simulated data perform better than the baseline cut analysis. Their performance is shown using high energy data collected at the ATLAS experiment. The improvement ranges from a factor of 2 to 5 in rejection for the same efficiency, depending on the selected efficiency operating point and the number of prongs in the tau decay. The strengths and weaknesses of each technique are also discussed.

1. Introduction

Tau leptons (τ) play an important role in the physics program of the LHC. In addition to measurements of standard model processes involving τ, they provide an important signature of several types of new physics. Notably, both low mass Standard Model Higgs boson searches and several new particle searches in Supersymmetric models involve decays to τ. Figure 1 shows the branching ratio for Higgs boson decay in the Standard Model as a function of Higgs mass [1]. The two-τ channel is an important signature for Higgs masses below about 140 GeV.

The challenge in exploiting τ signatures in the LHC environment is separation from background. The τ lifetime is too short to observe them directly. Instead, the τ decay products are observed. Figure 2 shows a pie-chart of the decay branching ratio for τ. The leptonic decays, representing about 35% of the total, are usually considered to be indistinguishable from prompt lepton production and are therefore included in other signal channels. For this reason, τ identification concentrates on identifying the hadronic decay modes, representing about 65% of the total. The difficulty with the hadronic modes is distinguishing them from an overwhelming background of quark and gluon jets at the LHC. To overcome this difficulty, a multivariate approach is required.

This paper presents the status of τ identification in the ATLAS experiment at the time of the 14th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2011).

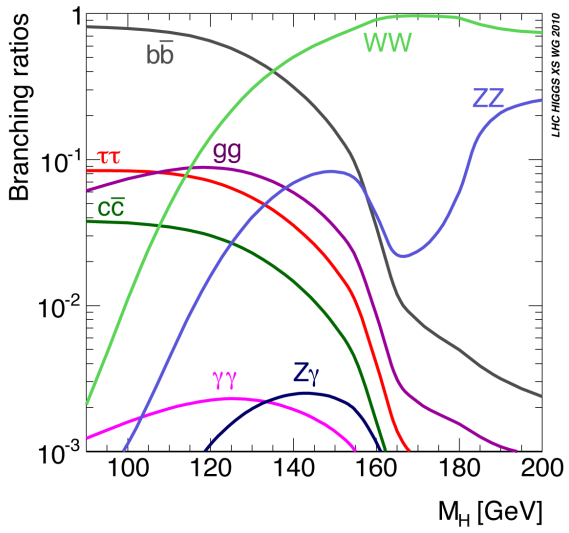


Figure 1: Standard Model Higgs branching ratios as a function of the mass of the Higgs [1].

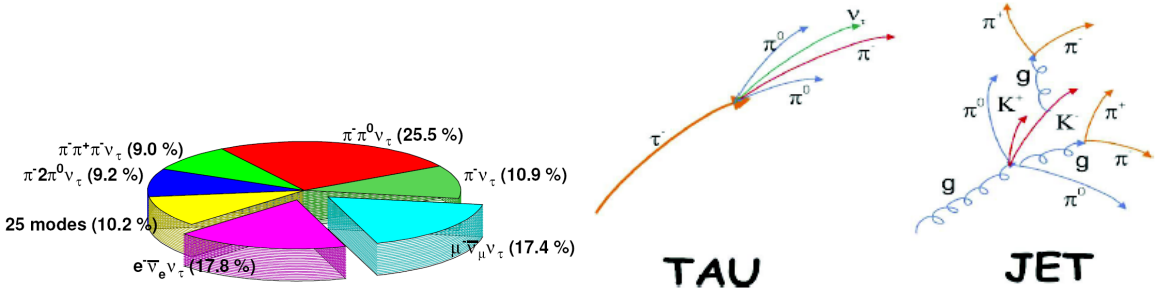


Figure 2: Left: pie-chart of hadronic tau decay branching ratios. Right: artist rendering of a τ jet and a gluon jet.

2. τ reconstruction and identification

The reconstruction and identification of hadronic τ in ATLAS are described in detail in [2]. Separation of hadronic τ decays from jet backgrounds happens in two steps. The first is to reconstruct a list of τ candidates in each event. Jets reconstructed with the anti-K_T algorithm [3] with distance parameter $R = 0.4$ are used to "seed" the reconstruction. Tracks are then associated to each seed and variables useful in distinguishing τ from quark and gluon jets are calculated and stored for each candidate. Very little separation between jets and τ is gained by this reconstruction step. An incomplete list of distinguishing variables includes:

- Track Radius (R_{track}): the p_T-weighted average track radius measured from the tau candidate axis.
- Electromagnetic Radius (R_{EM}): the transverse energy weighted average calorimeter radius measured from the tau candidate axis. Only cells in the EM calorimeter with $\Delta R < 0.4$ are used[1].
- Core Energy Fraction (f_{core}): the fraction of transverse energy in the core of the τ candidate. The core is defined as being within $\Delta R < 0.1$ of the tau axis.
- Cluster Mass ($m_{clusters}$): Invariant mass computed from the constituent clusters of the jet seed.

[1] $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$

- Track Mass (m_{track}): Invariant mass computed from the constituent tracks of the τ candidate.
- Transverse Flight Path Significance (S_T^{flight}): The decay-length significance of the secondary vertex for multi-track τ candidates in the transverse plane.

Figure 3 shows these distinguishing variables for both simulated candidates and real candidates extracted from the 2010 ATLAS dataset. Most of these variables exploit the narrower energy deposits and lower particle multiplicity in τ decays compared to quark or gluon jets.

The second step is to use the distinguishing variables calculated during reconstruction to separate τ from jets. ATLAS has used three different approaches for this step: simple cuts, log likelihood (LLH), and boosted decision trees (BDT). The latter two will be described herein.

Both the LLH and BDT techniques were "trained" using monte carlo generated events ($Z \to \tau\tau$) as signal and dijet ATLAS data as background. The two techniques are then compared on common samples using common definitions of signal efficiency and background rejection. They are trained and evaluated separately for 1-track (1-prong) and multi-track (3-prong) τ candidates.

2.1. LLH ID technique

The likelihood function is defined as the product of the distributions of the identification variables:

$$L_{S(B)} = \prod_{i=1}^{N} p_i^{S(B)}(x_i)$$

where $S(B)$ refers to signal (background), and $p_i^{S(B)}$ is the signal (background) probability density function of variable x_i. The likelihood function neglects correlations between variables and represents the joint probability distribution for the input variables.

A discriminant is built from the log-likelihood ratio between signal and background:

$$d = \ln\left(\frac{L_S}{L_B}\right) = \sum_{i=1}^{N} \ln\left(\frac{p_i^S(x_i)}{p_i^B(x_i)}\right).$$

A representative discriminant distribution is shown in figure 4. Good separation between signal (τ) and background (jets) is seen.

2.2. BDT ID technique

Boosted decision trees [4] turn a simple cut-based approach into an advanced multivariate technique. A simple decision tree makes a series of selections (cuts) to classify each τ candidate as either signal or background. However, candidates which fail each cut are not discarded, rather they are further processed by the algorithm. By successively applying the best selection on the best available variable to achieve optimal classification of each candidate, a simple decision tree can effectively make a piecewise continuous cut through the multidimensional space defined by the input variables, as illustrated in figure 5. The process of "boosting" then creates a second decision tree optimized to work well on the candidates which were misclassified by the first tree. A large number of trees can be created in this way, each optimized to work well on the misclassified candidates in the previous tree. Boosting improves the stability of the training and the overall performance of the signal and background classification. At the end of the process, the final BDT "score" is determined by a weighted average of the scores from each of the trees in the BDT set.

In this case, the BDT was trained using the TMVA [5] program and 50 trees were used. The boosting was performed by the adaBoost [4] algorithm with boosting strength=0.2. Figure 5

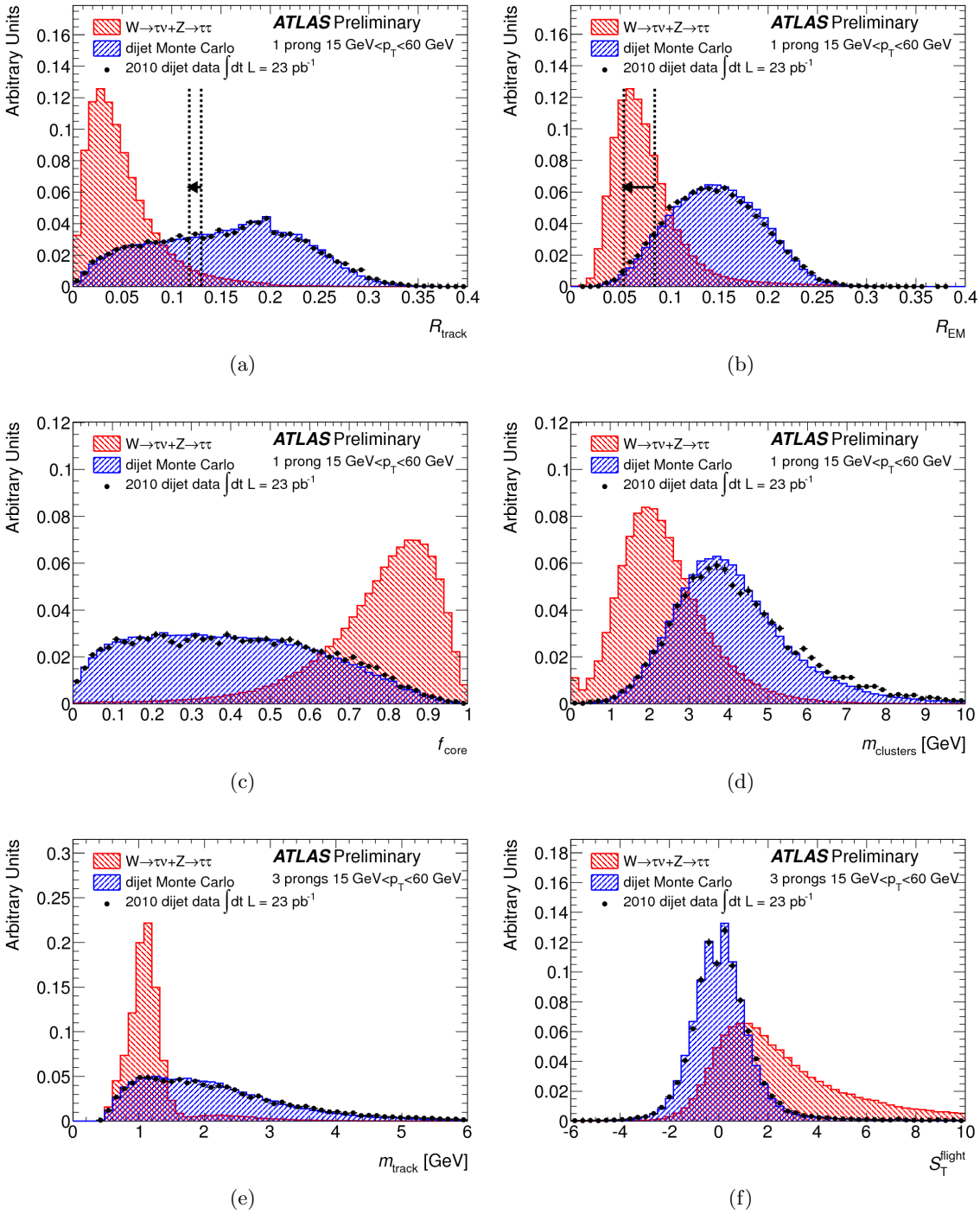
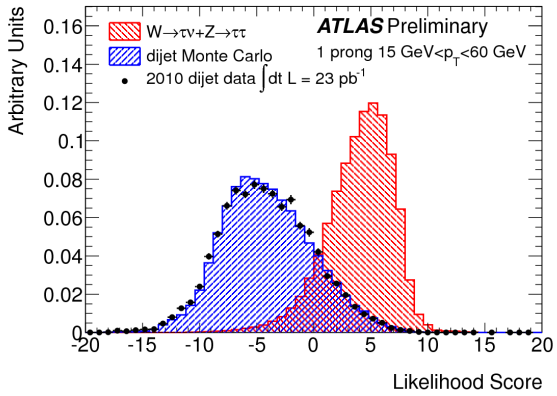
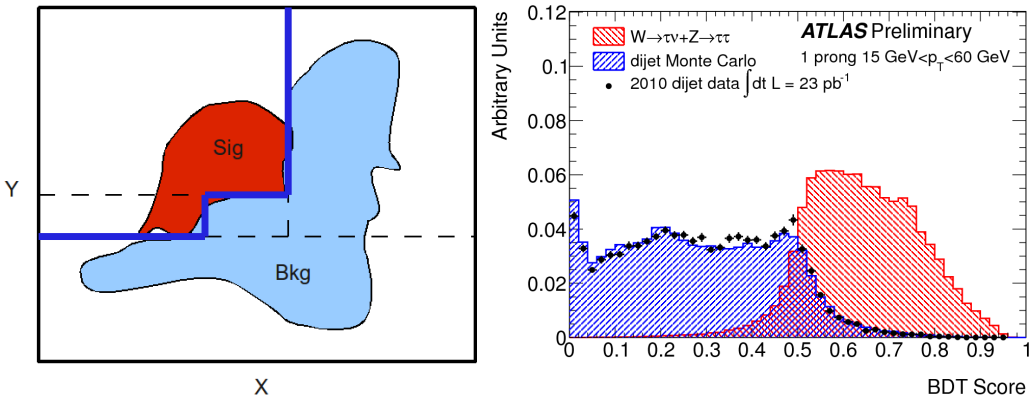


Figure 3: Variable distributions used to separate hadronic τ from quark and gluon jets [2]. (a) Track radius (b) Electromagnetic radius (c) Core energy fraction (d) Cluster mass (e) Track mass (f) Transverse flight path significance.



(a)

Figure 4: The log-likelihood-ratio for 1-prong τ candidates [2].



(a) (b)

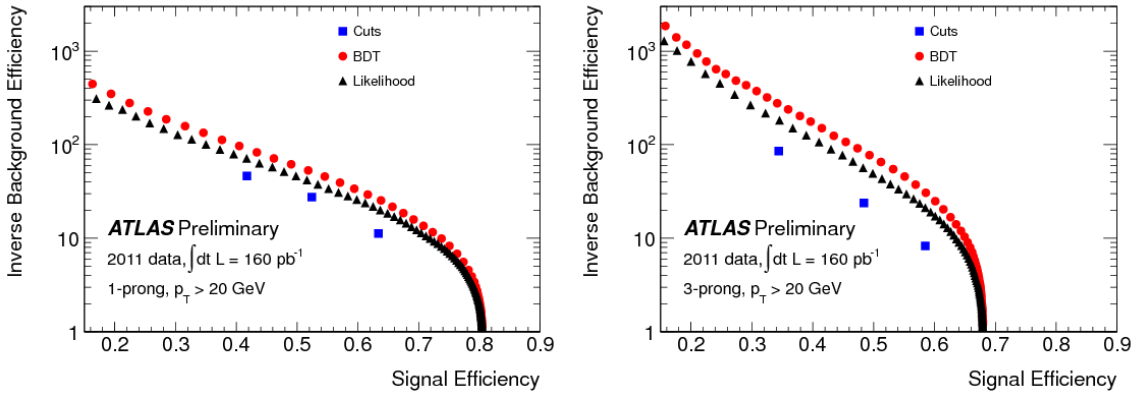
Figure 5: (a) Illustration of a piecewise continuous cut through a 2D plane made by a simple decision tree. Signal is in red, background in blue. (b) BDT discriminant output scores for signal and background for 1-prong τ candidates [2].

shows a representative discriminant distribution for BDT using signal Monte Carlo and dijet ATLAS data. Good separation between signal (τ) and background (jets) is seen.

In addition to its good performance, the BDT technique has several distinct advantages over other approaches. Correlated variables are handled easily by the technique. At each decision point, the best variable is chosen to separate signal from background. Adding a well-modeled variable does not worsen the performance of the algorithm, regardless of the correlation with existing variables. The training and optimization of BDT is straightforward and fast, with very few tunable parameters to optimize. A disadvantage of BDT compared to simple cuts or LLH is the number of candidates needed in the training sample. The power of the technique makes it necessary to use larger training samples.

2.3. Performance of MV techniques

Figure 6 shows the final performance of three τ identification techniques in ATLAS: simple cuts, LLH and BDT. The figure shows inverse background efficiency (rejection) vs. signal efficiency.



(a) (b)

Figure 6: Performance summary for three different techniques to separate τ from jet backgrounds: cuts, LLH and BDT. (a) 1-prong candidates (b) 3-prong candidates. Taken from [6].

As expected, the least performant solution is the simple cuts approach, the next solution is the LLH and the best is BDT.

3. Physics Applications

The ATLAS τ identification algorithms have been applied in several public physics results in 2010 and 2011. In this section, two such results are described. Each of these results uses the BDT τ identification technique.

3.1. $W \to \tau\nu$ cross section

Rediscovery of the W boson and measurement of its cross section was one of the most important milestones acheived in the first months of LHC running. Rediscovery of the τ channel in the same dataset was challenging. The first paper published in this channel at the LHC is the ATLAS result using the BDT τ identification [7]. In this analysis, a cut was made on the BDT output score to purify the sample. This cut accepted roughly 30% of signal events while applying a jet rejection factor of about 100 for 1-prong candidates. For 3-prong candidates, the signal efficiency was 35% for a jet rejection factor of about 300.

Figure 7 shows the number of charged tracks distribution after application of all selection cuts, including the BDT τ identification. The supression of the 2-track bin relative to the 3-track bin is a strong indication of a signal. The final inclusive cross section measurement is

$$\sigma^{tot}_{W\to\tau\nu} = 11.1 \pm 0.3(stat) \pm 1.7(syst) \pm 0.4(lumi)nb.$$

3.2. $t\bar{t} \to \mu + \tau + X$ cross section

Measurement of the branching ratio of top quarks to τ is of interest in probing for physics beyond the Standard Model. As an example, if the charged Higgs boson of Supersymmetry has a mass less than the top quark mass minus the b-quark mass, the top quark can decay to it. If this decay is possible, the charged Higgs will then favour decays to $\tau + \nu$ over a wide range of SUSY parameter space. Therefore, a measurement of an unexpectedly high cross section in the $t\bar{t} \to \tau + X$ channel would be an indication of new physics.

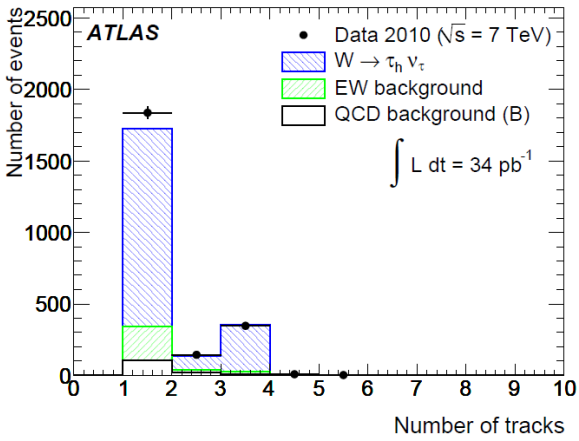


Figure 7: The number of charged tracks reconstructed as part of the τ candidate. True τ should have an odd number of associated tracks (charge conservation) [7].

In the summer of 2011, the most precise measurement of the $t\bar{t} \to \tau$ cross section came from the ATLAS experiment in the $\mu + \tau$ channel [8]. After preselection, the dominant background to this channel is $t\bar{t} \to \mu + jets$ with one of the jets faking a τ. The most powerful remaining feature to exploit to remove this background is BDT τ identification.

This analysis used BDT τ identification in a different way than the W analysis described earlier. Rather than cutting on the BDT output score, the full score distribution was used. Since the jets faking τ are from $t\bar{t}$ events, they are dominantly quark jets. The BDT score distribution for a quark-rich data sample was obtained from $W + jets$ data. Figure 8 shows two different quark-rich templates, one extracted from data in which the muon and the jet had the opposite sign charge (OS) and another from the same sign (SS) sample. It also shows a true τ template extracted from simulation. Figure 9 shows the result of fitting these templates to Monte Carlo simulation (a) and data (b). The results in simulation (MC) agree well with expectation. The results in data yield

$$\sigma_{t\bar{t}}^{tot} = 142 \pm 21(stat) \pm^{20}_{16} (syst) \pm 5(lumi)pb$$

in agreement with the Standard Model.

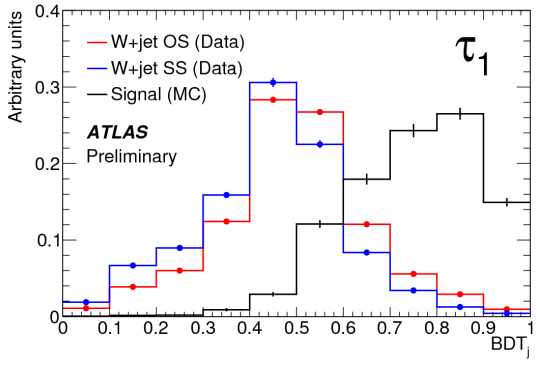
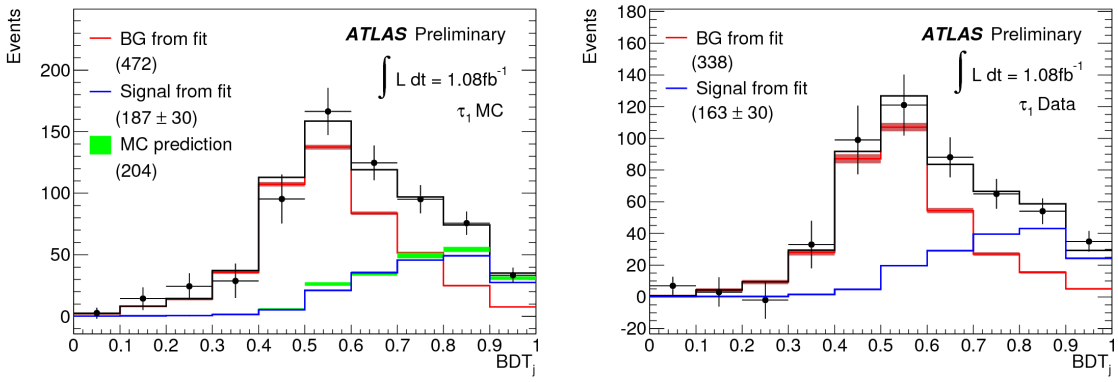


Figure 8: Templates used to extract signal cross section in $t\bar{t}$ analysis [8].

4. Summary and Conclusions

The ATLAS τ identification relies on multivariate techniques to separate signal from quark and gluon jet backgrounds. The perfomance gain compared to a simple cut-based approach is substantial (factor of 2-5 in background rejection for the same signal efficiency). These techniques has been applied to physics measurements in challenging channels, allowing cross section measurements in τ channels for the first time at the LHC.



(a) (b)

Figure 9: Results of template fits from the $t\bar{t} \to \mu + \tau + X$ analysis. (a) The fits performed on simulation only, (b) The fits performed on data [8].

References

[1] Dittmaier S, Mariotti C, Passarino G and Tanaka R (editors) 2011 CERN-2011-002, *Preprint* arXiv:1101.0593v3

[2] The ATLAS Collaboration 2011 ATLAS-CONF-2011-077, https://cdsweb.cern.ch/record/1353226

[3] Cacciari M, Salam G P, and Soyez G 2008 *Journal of High Energy Physics* **04**, 063

[4] Breiman L, Friedman J, Stone C, and Olshen R 1984 *Classification and Regression Trees* (Chapman and Hall)

Freund Y and Shapire R 1996 *Proceedings 13th International Conference on Machine Learning* (July 3-6, 1996, Bari, Italy)

[5] *Toolkit for Multivariate Data Analysis* 2011 http://tmva.sourceforge.net/, Jan, 2011. version 4.0.4 (available as part of ROOT version 5.26).

[6] The ATLAS Collaboration 2011
Tau Public Results https://twiki.cern.ch/twiki/bin/view/AtlasPublic/TauPublicCollisionResults

[7] The ATLAS Collaboration 2011 Accepted by *Physics Letters B.*, CERN-PH-EP-2011-122, (*Preprint* arXiv:1108.4101v1)

[8] The ATLAS Collaboration 2011 ATLAS-CONF-2011-119 https://cdsweb.cern.ch/record/1376411