Optimization of τ identification safe variables with TMVA

9/9/2009

Fuquan Wang

Abstract:

During the summer student program I optimized two cut based approaches for the identification and reconstruction of τ leptons in ATLAS.

In ATLAS, hadronically decaying τ leptons are reconstructed by a calorimeter-seeded algorithm and a track-seeded one. For optimization, two cut-based approaches will be used with "safe variables", which will be well understood at the 1st stage of data taking. The optimization and performance of these approaches in terms of efficiency and rejection against jets are discussed.

1. Introduction:

τ leptons will play an important role at the LHC. They will provide an excellent probe in searches for new phenomena: the Standard Model Higgs boson at low masses, the MSSM Higgs boson or Supersymmetry (SUSY).

The reconstruction of τ leptons is usually understood as a reconstruction of hadronic decay modes in ATLAS, since it would be difficult to distinguish leptonic decaying τs from primary electrons and muons. Moreover, their reconstruction at hadron colliders remains a very difficult task in terms of distinguishing interesting events from background processes dominated by QCD multi-jet production.

Hadronically decaying τ leptons are distinguished from QCD jets on the basis of low track multiplicities contained in a narrow cone, characteristics of the track system and shapes of the calorimetric showers. Isolation from the rest of the event is usually required both in the Inner Detector and the Calorimeter. Other properties like the ratio of EM energy deposits to hadronic energy deposits, or the ratio of the track transverse momenta to the energy deposits in the calorimeter can also be exploited.

Cut-based approaches for the offline reconstruction are described in Section2. In Section 3 I document the performance of the cut-based τ lepton selection.

2. Cut-based approaches for τ identification

The purpose of these cut-based approaches is to provide a selection of "safe variables" for analysis with early data. The motivation to use only "safe variables" is to select only variables that are thought to be well understood in the early data taking phase and some other variables were

avoided. The description of variables that are chosen can be found in sections 2.1 and 2.2.

Currently three threshold for both cut selections are defined: *tight, medium* and *loose*, corresponding to efficiencies of *0.3, 0.5* and *0.7* with respect to reconstructed candidates with the correct reconstructed track multiplicity matched to true hadronically decaying τ leptons.

These cuts varied for multi-prong τ candidates and those with one or fewer tracks as well as for several p_T bins and for candidates with both seeds and those with only calo-seed. The following six p_T bins are defined: 0-20GeV, 20-30GeV, 30-45GeV, 45-60GeV, 60-100GeV and >100GeV.

2.1 Calorimeter-based approach

This approach applies selection criteria on calorimeter-seeded τ candidates. It uses a selection of four calorimeter variables which are claimed to be safe by calorimeter experts and which are not highly correlated. This approach avoids the use of identification with tracking variables, in case the tracker is not yet well understood. The following variables are used for identification for τ lepton decay products:

- **The electromagnetic radius R_{em} (EMRadius):**
  To exploit the smaller transverse shower profile in τ decays, the electromagnetic radius R_{em} is used, defined as:
  $$R_{em} = \frac{\sum_{i=1}^{n} E_{T,i}\sqrt{(\eta_i - \eta_{cluster})^2 + (\phi_i - \phi_{cluster})^2}}{\sum_{i=1}^{n} E_{T,i}}$$
  where i runs over all cells in the electromagnetic calorimeter that are associated to the τ candidate within ΔR<0.4, where $\Delta R = \sqrt{(\eta_i - \eta_{cluster})^2 + (\phi_i - \phi_{cluster})^2}$. The quantities η_i, ϕ_i and $E_{T,i}$ denote their position and transverse energy in cell i.
- **Transverse energy width in the η strip layer (stripWidth2):**
  The transverse energy width Δη is defined as
  $$\Delta\eta = \sqrt{\frac{\sum_{i=1}^{n} E_{T,i}^{strip} (\eta_i - \eta_{cluster})^2}{\sum_{i=1}^{n} E_{T,i}^{strip}}}$$
  where the sum runs over strip cells in a cone with ΔR<0.4 associated to the τ candidate around the cluster axis and $E_{T,i}^{strip}$ is the corresponding strip transverse energy.
- **Isolation in the calorimeter (IsoFrac):**
  Clusters built from hadronic τ decays are well collimated and therefore tight isolation criteria can be exploited. Here a ring between 0.1<ΔR<0.2 around the τ candidate was chosen as the isolation region and the quantity
  $$\Delta E_T^{12} = \frac{\sum_i E_{T,i}}{\sum_j E_{T,j}}$$
  is calculated, where the indices i and j run over the electromagnetic calorimeter cells in a cone around the cluster axis with 0.1<ΔR<0.2 and ΔR<0.4 respectively associated to the τ candidate, and $E_{T,i}$ and $E_{T,j}$ denote the transverse cell energies.
- **Ratio of EM energy and total energy at the EM scale (EtEMEt):**

$$\frac{E_T^{EM}}{E_T^{total}} = \frac{\sum_i E_{T,i}^{EM}}{\sum_i E_{T,i}^{EM} + \sum_j E_{T,j}^{Had}}$$

where the sums run over all cells in a cone with ΔR<0.4 associated to τ candidate, $E_{T,i}^{EM}$ is the energy in the electromagnetic calorimeter and $E_{T,j}^{Had}$ the energy in the hadronic calorimeter, both taken at EM scale.

2.2 Calorimeter+Track-based approach

The calorimeter+track-based approach is a more aggressive approach. It applies selection criteria on τ candidates seeded by both the calorimeter and tracking, and combines the four variables of the calorimeter-based method with five variables that involve tracking. This approach is expected to have a better performance due to a larger number of variables providing additional information.

- **Width of track momenta (RWidth2Trk3P):**
  The variance of tracks in η, ϕ-space, weighted with their transverse momenta, W_{tracks}^{τ}, for multi-track candidates is taken:
  $$W_{tracks}^{\tau} = \frac{\sum(\Delta\eta^{\tau_{1p3p,track}})^2 \cdot p_T^{track}}{\sum p_T^{track}} - \frac{(\sum \Delta\eta^{\tau_{1p3p,track}} \cdot p_T^{track})^2}{(\sum p_T^{track})^2}$$
- **E_T over p_T of the leading track (etOverPtLeadLooseTrk):**
  $$\frac{E_T^{total}}{p_{T,1}} = \frac{\sum_i E_{T,i}^{EM} + \sum_j E_{T,j}^{Had}}{p_{T,1}}$$
  where the calibrated transverse energies of the cells associated to the τ candidate in both EM and hadronic calorimeters is taken.
- **Fraction of EM energy after H1 style calibration and sum of total p_T of tracks (EtEMsumpT):**
  $$\frac{E_T^{EM}}{p_T^{total}} = \frac{\sum_i E_{T,i}^{EM}}{\sum_{j=1}^{n} p_{T,j}^{track}}$$
  where the sum in the numerator runs over all EM calorimeter cells associated to the τ candidate in a cone with ΔR and $E_{T,i}^{EM}$ is the energy after H1 style calibration in cell i and sum in the denominator runs over the transverse momenta p_T^{track} of all tracks associated to the τ candidate.
- **Fraction of hadronic energy after H1 style calibration and sum of total p_T of tracks (EtHadsumpT):**
  $$\frac{E_T^{Had}}{p_T^{total}} = \frac{\sum_i E_{T,i}^{Had}}{\sum_{j=1}^{n} p_{T,j}^{track}}$$
  where the sum in the numerator runs over all hadronic calorimeter cells associated to the τ candidate in a cone with ΔR and $E_{T,i}^{Had}$ is the energy after H1 style calibration in cell i and sum in the denominator runs over the transverse momenta p_T^{track} of all tracks associated to the τ candidate.
- **Fraction of sum of total p_T of tracks and total energy after H1 style calibration (sumpTEt):**

$$\frac{p_T^{total}}{E_T^{total}} = \frac{\sum_{k=1}^{n} p_{T,k}^{track}}{\sum_i E_{T,i}^{EM} + \sum_j E_{T,j}^{Had}}$$

where the sum in the numerator runs over the transverse momenta p_T^{track} of track associated to τ candidates and the sums in the denominator runs over all cells associated to the τ candidate in a cone with ΔR<0.4, $E_{T,i}^{EM}$ is the cell energy in the electromagnetic calorimeter and $E_{T,j}^{Had}$ the cell energy in the hadronic calorimeter, with cell energies being taken after H1 style calibration.

2.3 Optimization procedure

The optimization for the cut-based selection was done in multiple steps and has been performed for six p_T-bins (0-20 GeV, 20-30 GeV, 30-45 GeV, 45-60 GeV, 60-100 GeV and >100 GeV) and 1-prong and 3-prong τ-lepton candidates respectively. All data samples listed in appendix A were used as input.

TMVA, a toolkit for Multivariate Data Analysis was used. All variables picked up as "safe variables" were used as input for a rectangular cut optimization. This classifier returns a binary response (signal or background) and maximizes the background rejection at a given signal efficiencies (0.1…0.9). For the optimization a so-called "constrained method" was used: optimal cut values of the previous run were used as upper limits to the next optimization step. Five iterative steps from 0.9→0.7→0.5→0.3→0.1 have been performed and results of the *0.7*, *0.5* and *0.3* run were used as cut values for the selections *loose*, *medium* and *tight*.

3. Performance of the cut-based selections

This section discusses and compares the performance of the Calorimeter-based and he Calorimeter+Track-based approach. The Calorimeter-based approach uses four variables and its performance is expected to be lower than for the Calorimeter+Track-based approach with eight (nine) variables for the 1-prong (3-prong) decay modes. To assess the performance of the cut-based identification, both approaches will be compared with a projective logarithmic likelihood method and a neural network method for the calorimeter+track-based approach, with both the likelihood method and neural network method using a broader selection of discriminate variables that are not restricted to the "safe variables".

3.1 τ/jet separation

The performance is presented as curves in the signal efficiency/background rejection plane. For the signal efficiency, ϵ^{sig}, τ-candidates were matched to Monte Carlo τ within a cone of ΔR=0.2. The efficiency was calculated separately for all six p_T bins and decay modes:

$$\epsilon^{sig} = \frac{\text{number of matched rec } 1(3) \text{ prong taus passing cuts}}{\text{number of MC } 1(3) \text{ prong taus in kinematic range with stable daughters}}$$

In case of background, the efficiency calculation uses truth matching of Monte Carlo jets to

reconstructed τ candidates with 1(3) associated tracks, within a cone of ΔR=0.2. The background efficiency, ϵ^{bkg} is defined as:

$$\epsilon^{\mathrm{bkg}} = \frac{\text{number of matched rec } 1(3) \text{ prong taus passing cuts}}{\text{number of MC jets matched to rec } 1(3) \text{ prong taus in kinematic range}}$$

The background rejection, r, is defined as:

$$\mathrm{r} = \frac{1 - \epsilon^{\mathrm{bkg}}}{\epsilon^{\mathrm{bkg}}}$$

3.2 Performance of Calorimeter-based variables

The four variables used for the Calorimeter-based approach are described in section 2.1. The distribution for those four variables is shown in Figure 1. The position of the cut values are displayed with a solid line (tight), a dashed line (medium) and a dotted line (loose) in the p_T bin 30-45GeV.

Figure 3 shows the performance of the Calorimeter-based approach for one prong events, the performance for three prong events is shown in Figure 4. In both cases the τ candidates were reconstructed from the calorimeter seed. Table 1 shows the efficiencies and rejections for the loose, medium and tight cuts for 1-prong and 3-prong candidates using the Calo Only approach.

One prong candidates show a better efficiency/rejection ratio than three prong candidates compared to each other and the logarithmic likelihood method.

3.3 Performance of Calorimeter+Track-based variables

The Caloremeter+Track-based approach uses four calorimeter variables and four additional tracking variables for one prong τ candidates and five tracking variables for three prong τ candidates. Distributions for the Calorimeter+Track variables are shown in Figure2. The position of the cut values are displayed with a solid line (tight), a dashed line (medium) and a dotted line (loose) in the p_T bin 30-45 GeV.

The performance for the Calorimeter+Track-based approach is shown in Figure 5 for one prong candidates, the performance for three prong candidates is shown in Figure 6. For both cases the τ candidates were reconstructed from both calorimeter and track seeds. Table 2 shows the efficiencies and rejections for the loose, medium and tight cuts for 1-prong and 3-prong candidates using the Calo Only approach.

As expected, the Calorimeter+Track approach has a rejection considerably higher for a given efficiency than the Calorimeter approach. In both approaches, the performance for one prong candidates is also much higher than for three prong candidates. Compared to a complex logarithmic likelihood and neural network method, the Calorimeter+Track-based approach is able to reject jets at a reasonable rate for both one prong and three prong τ candidates.

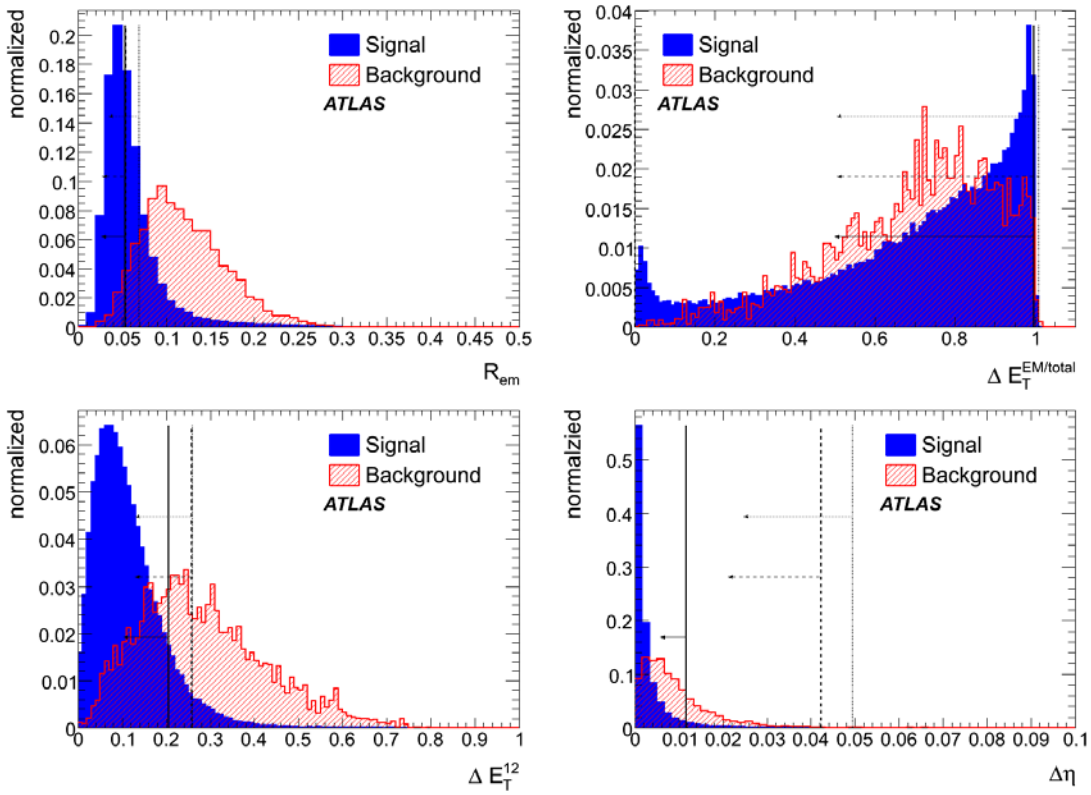


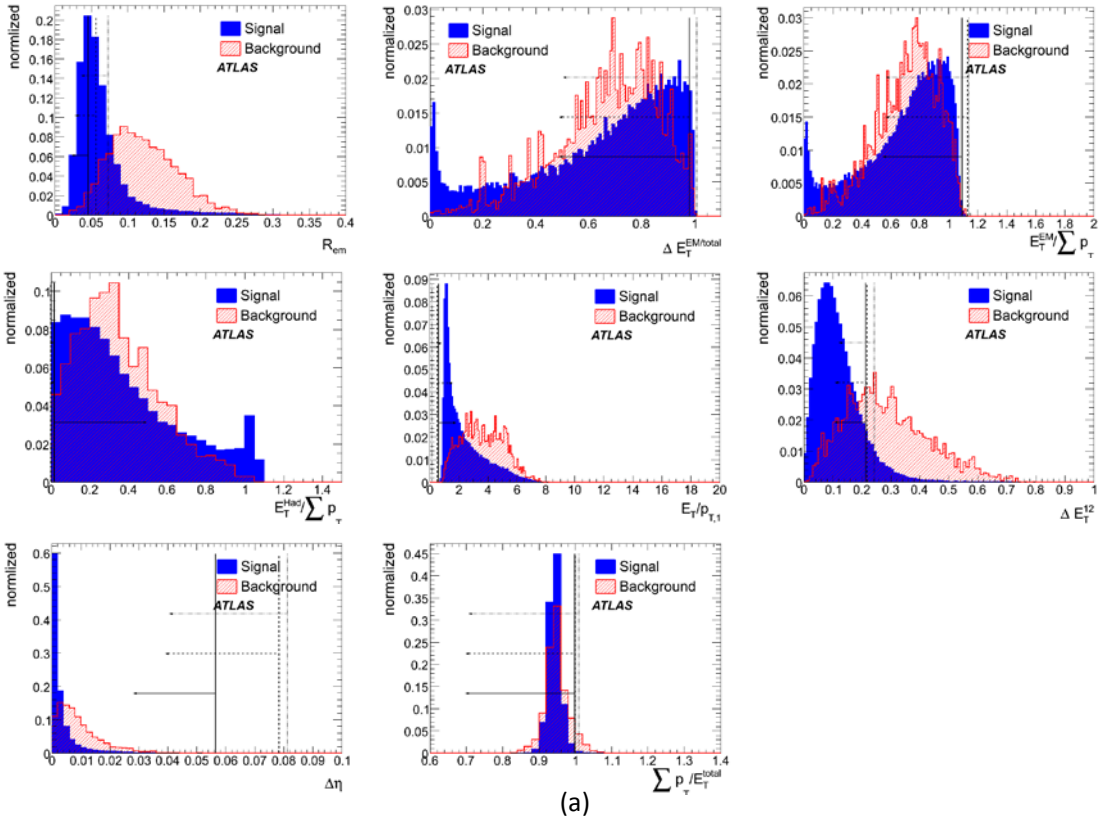
Figure 1: Distribution of calorimeter variables for 1-prong τ candidates within a p_T range of 30-45 GeV.

		1-prong		3-prong	
Selection	p_T-range(GeV)	cut-eff.	cut-rej.	cut-eff.	cut-rej.
Loose	0-20	0.748082	16.3608	0.748866	4.97824
	20-30	0.73828	41.0383	0.72719	6.50494
	30-45	0.733529	79.4153	0.719744	8.45063
	45-60	0.729195	183.153	0.731184	13.4065
	60-100	0.730651	219.488	0.721822	9.31924
	>100	0.614201	491.122	0.539105	23.9907
Medium	0-20	0.540367	39.6406	0.538608	13.4037
	20-30	0.52258	97.9918	0.505022	18.044
	30-45	0.512155	200.825	0.501029	23.0953
	45-60	0.513057	549.884	0.517659	39.9325
	60-100	0.498244	592.008	0.501492	37.7035
	>100	0.367626	1577.15	0.300499	106.148
Tight	0-20	0.337775	95.3431	0.320887	38.0657
	20-30	0.310906	234.121	0.289141	54.0301
	30-45	0.313311	532.922	0.291916	62.1821
	45-60	0.330141	1233.01	0.306993	118.516
	60-100	0.307917	1653.52	0.279033	147.75
	>100	0.199695	6169.12	0.124376	453.019

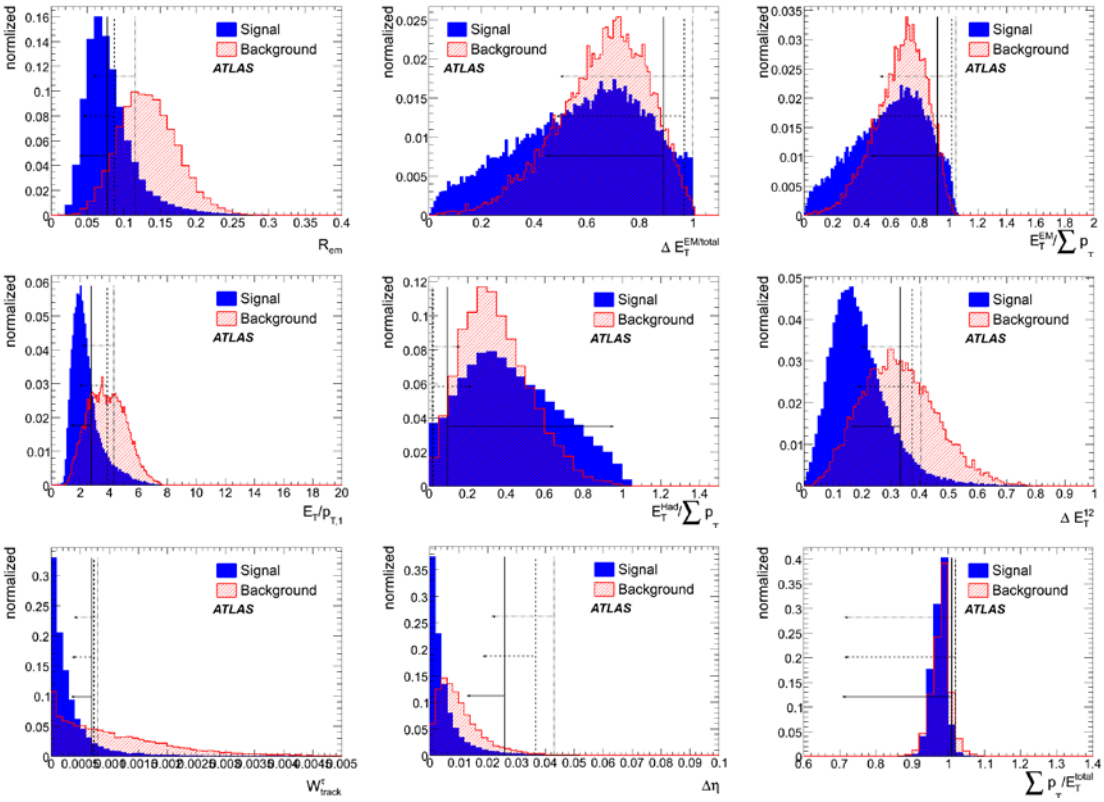
Table1: Efficiency and rejection for the cut-based selection using the Calo Only approach.

		1-prong		3-prong	
Selection	p_T-range(GeV)	cut-eff.	cut-rej.	cut-eff.	cut-rej.
Loose	0-20	0.768999	42.1321	0.742303	29.0793
	20-30	0.763074	67.2367	0.744189	14.4269
	30-45	0.759276	109.938	0.742171	15.706
	45-60	0.750487	216.676	0.850859	11.8248
	60-100	0.731813	311.625	0.739424	26.7263
	>100	0.638675	544.089	0.582715	86.788
Medium	0-20	0.567632	91.2942	0.53434	56.8214
	20-30	0.559535	146.714	0.539316	37.8027
	30-45	0.544257	291.145	0.539889	47.7368
	45-60	0.534337	736.667	0.536147	78.8746
	60-100	0.496681	946.693	0.49745	95.2579
	>100	0.406061	1695.64	0.292604	441.677
Tight	0-20	0.363788	191.496	0.325232	131.978
	20-30	0.346147	383.587	0.329076	108.911
	30-45	0.332906	692.138	0.326374	138.817
	45-60	0.325498	1878.97	0.32259	277.012
	60-100	0.281118	3238.27	0.181944	801.455
	>100	0.204262	8442.15	0.124382	3083.47

Table2: Efficiency and rejection for the cut based selection using the Calo+Track approach.



(a)

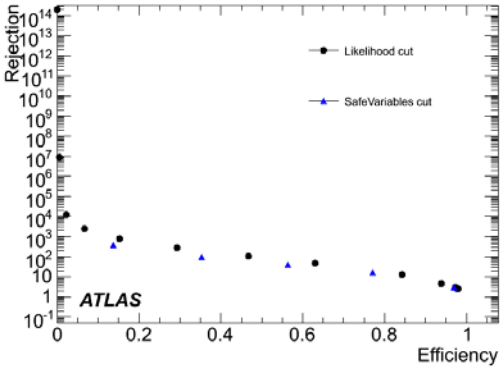


(b)

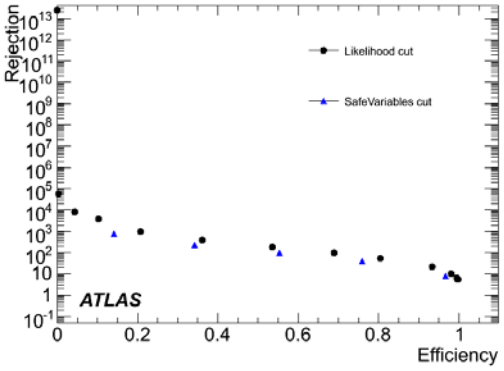
Figure 2: Distribution of calorimeter and tracking variables for one prong τ candidates (4(a)) and three prong τ candidates (4(b)) within a range of 30-45 GeV for signal and background.

4 Acknowledgments

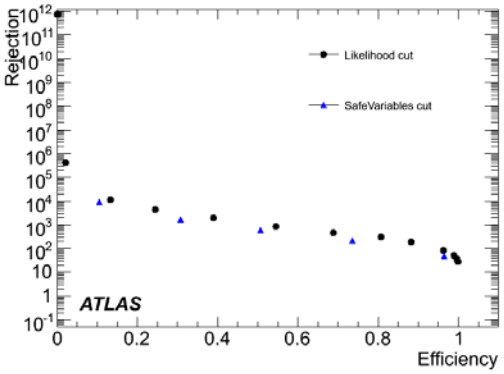
I would like to thank my supervisor, Mr Björn Gosdzik, for instruction and lots of helpful advices, and I would like to thank the DESY Summer Student Program to offer me this opportunity to work in this energetic and creative environment.



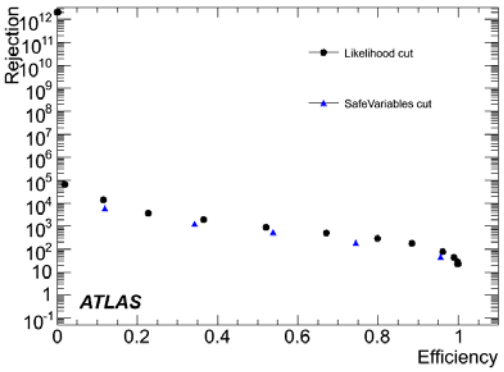
(a)



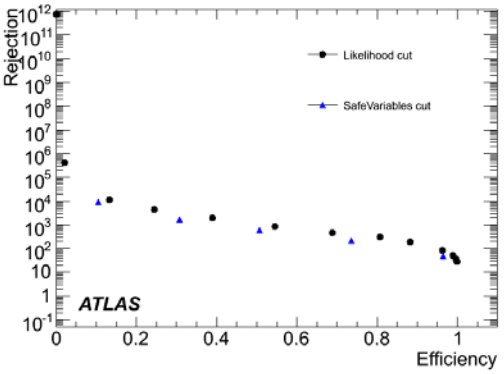
(b)



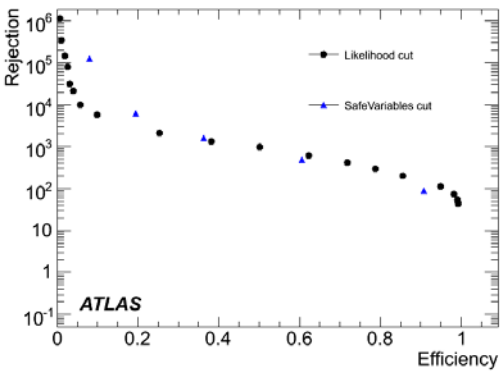
(c)



(d)

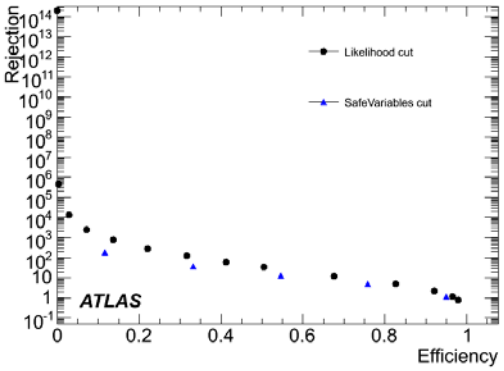


(e)

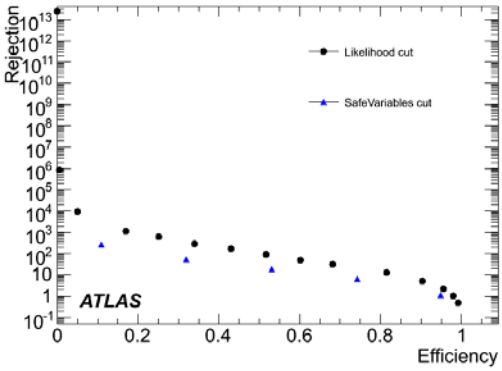


(f)

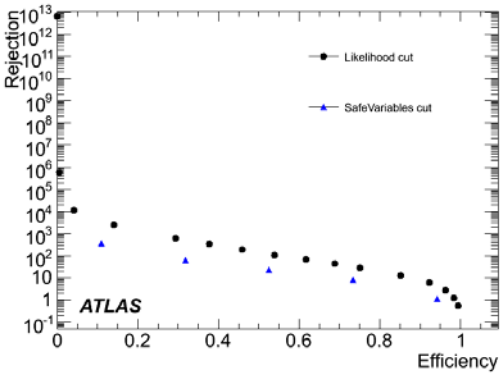
Figure 3: Performance of calorimeter-based approach (triangle) compared with log. likelihood (circle) for 1-prong τ candidates for 0-20 GeV (3(a)), 20-30 GeV (3(b)), 30-45 GeV(3(c)), 45-60 GeV(3(d)), 60-100 GeV(3(e)) and >100 GeV (3(f)).



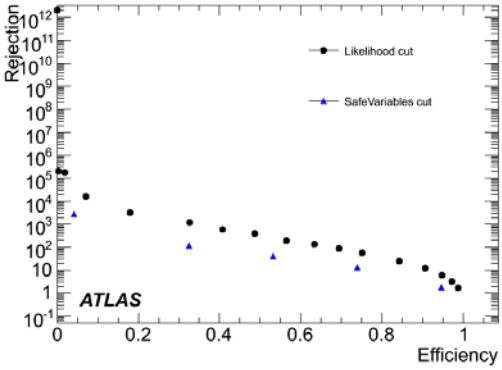
(a)



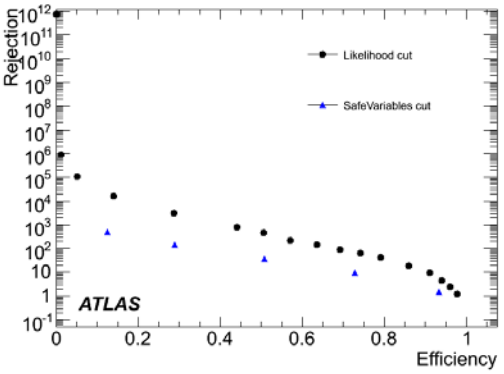
(b)



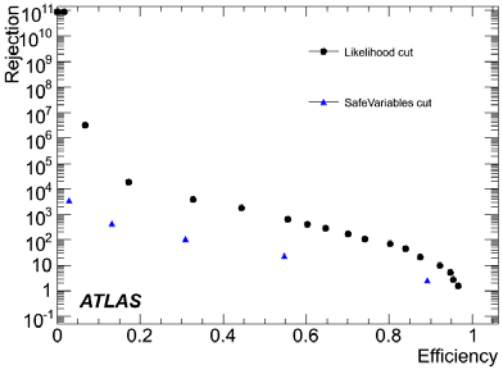
(c)



(d)

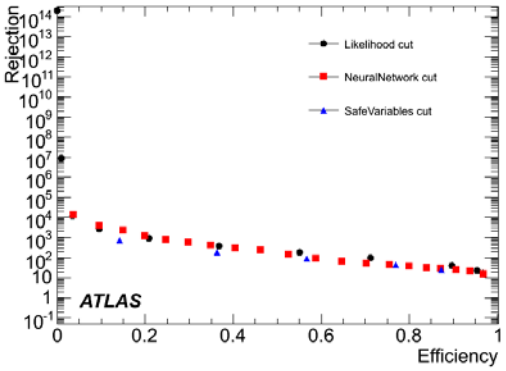


(e)

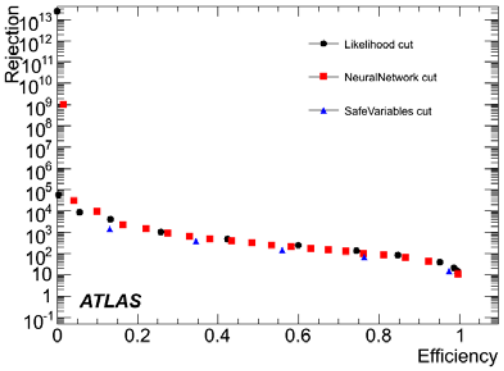


(f)

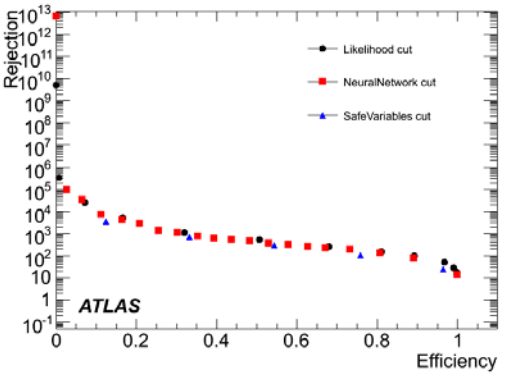
Figure 4: Performance of calorimeter-based approach (triangle) compared with log. likelihood (circle) for 3-prong τ candidates for 0-20 GeV (4(a)), 20-30 GeV (4(b)), 30-45 GeV(4(c)), 45-60 GeV(4(d)), 60-100 GeV(4(e)) and >100 GeV (4(f)).



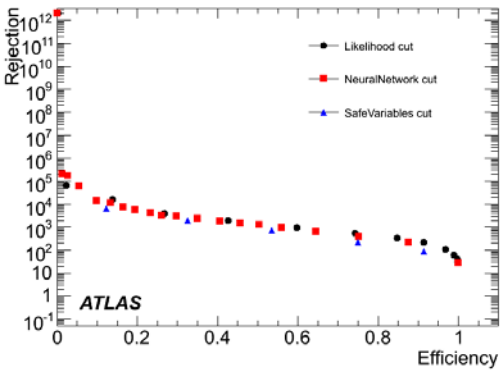
(a)



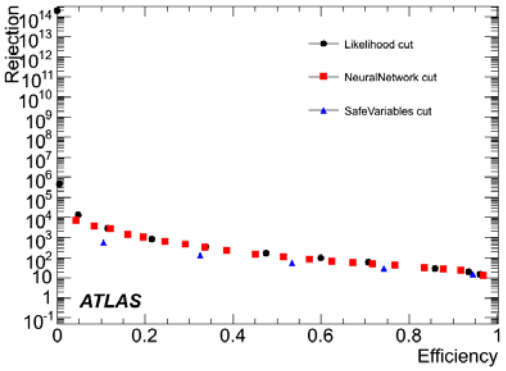
(b)



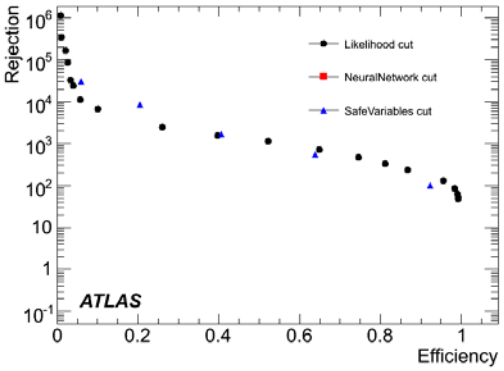
(c)



(d)

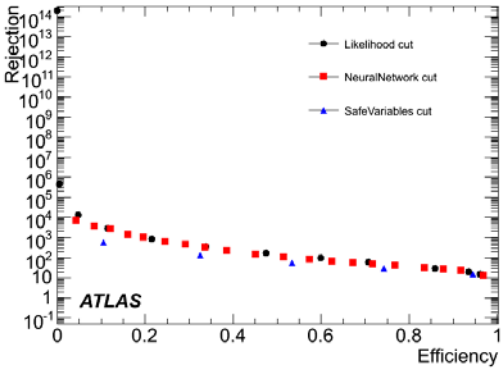


(e)

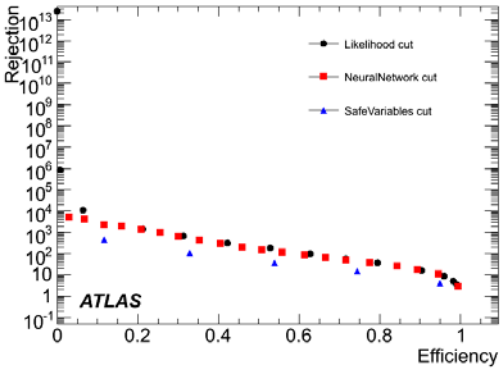


(f)

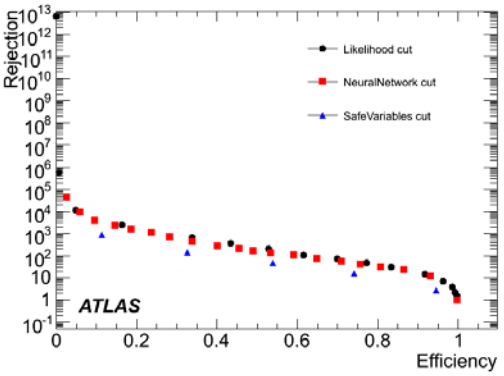
Figure 5: Performance of Calorimeter+Track-based approach (triangle) compared with log. likelihood (circle) and neural network (square) for 1-prong τ candidates for 0-20 GeV (5(a)), 20-30 GeV (5(b)), 30-45 GeV(5(c)), 45-60 GeV(5(d)), 60-100 GeV(5(e)) and >100 GeV (5(f)).



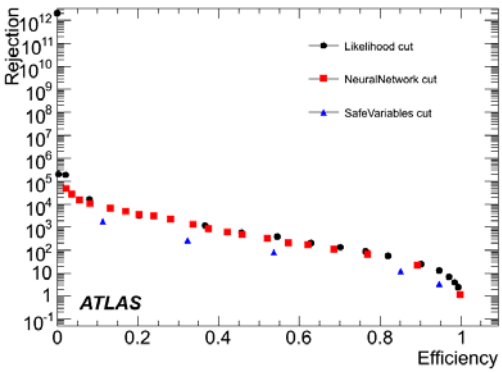
(a)



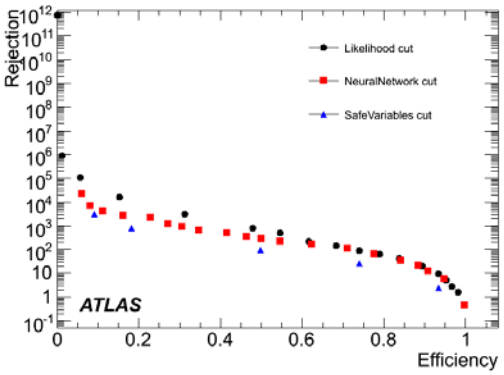
(b)



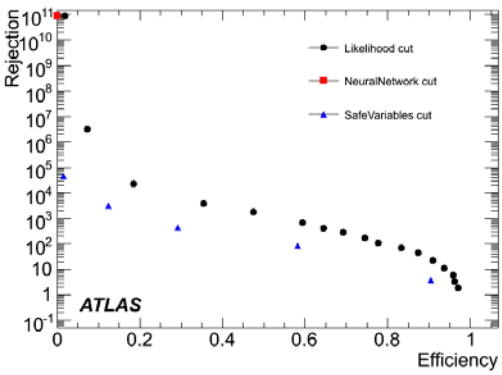
(c)



(d)



(e)



(f)

Figure 6: Performance of Calorimeter+Track-based approach (triangle) compared with log. likelihood (circle) and neural network (square) for 3-prong τ candidates for 0-20 GeV (6(a)), 20-30 GeV (6(b)), 30-45 GeV(6(c)), 45-60 GeV(6(d)), 60-100 GeV(6(e)) and >100 GeV (6(f)).

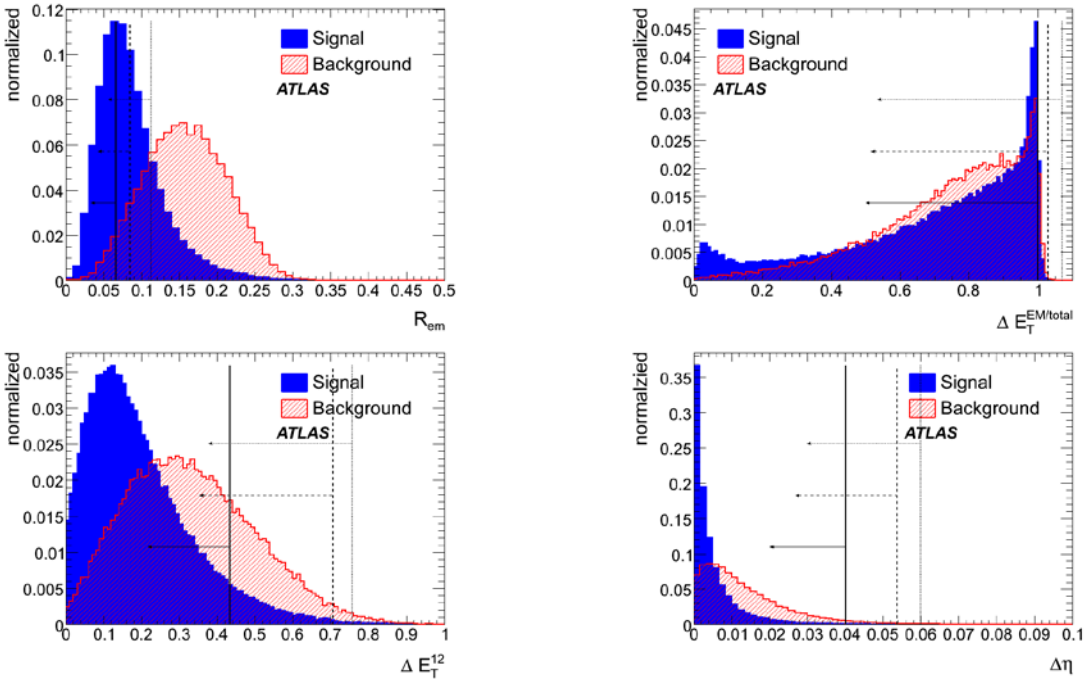
A Data Samples used

The following physics event samples were used:

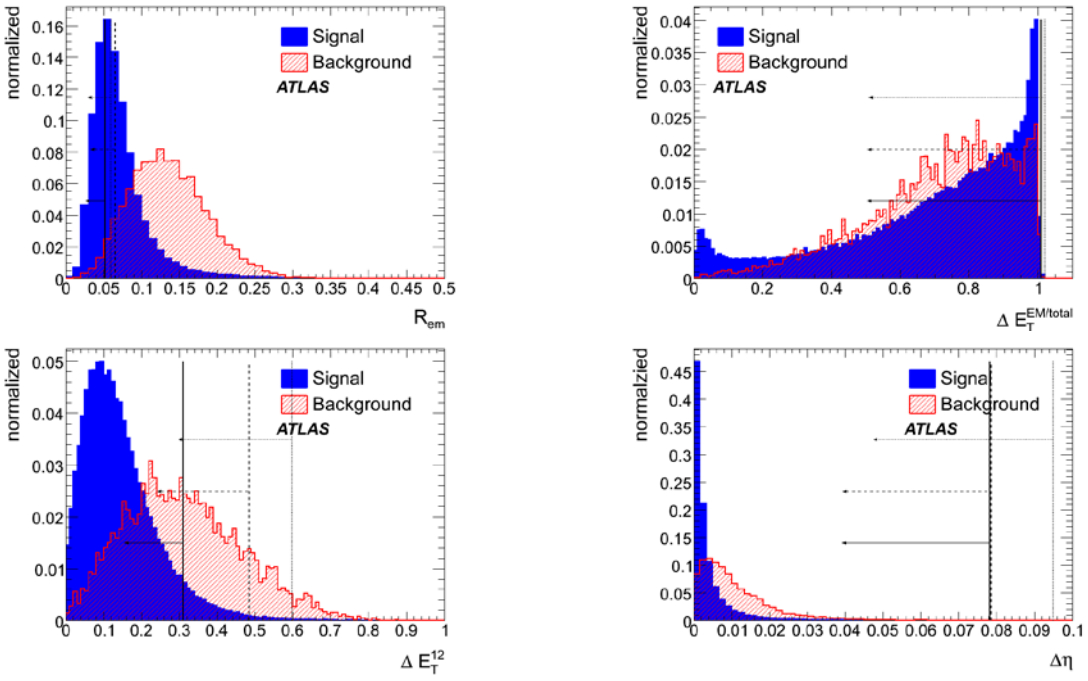
- Signal (718806 events)
  - Z→ τ τ (106052)
  - A → τ τ (109126)
  - bbA → ττ (106573)
- Background (2930997 events)
  - QCD di-jets (105009) 8-17 GeV
  - QCD di-jets (105010) 17-35 GeV
  - QCD di-jets (105011) 35-70 GeV
  - QCD di-jets (105012) 70-140 GeV
  - QCD di-jets (105013) 140-280 GeV
  - QCD di-jets (105014) 280-560 GeV

B Distribution of calorimeter variables

B.1 1-prong truth matched τ candidates

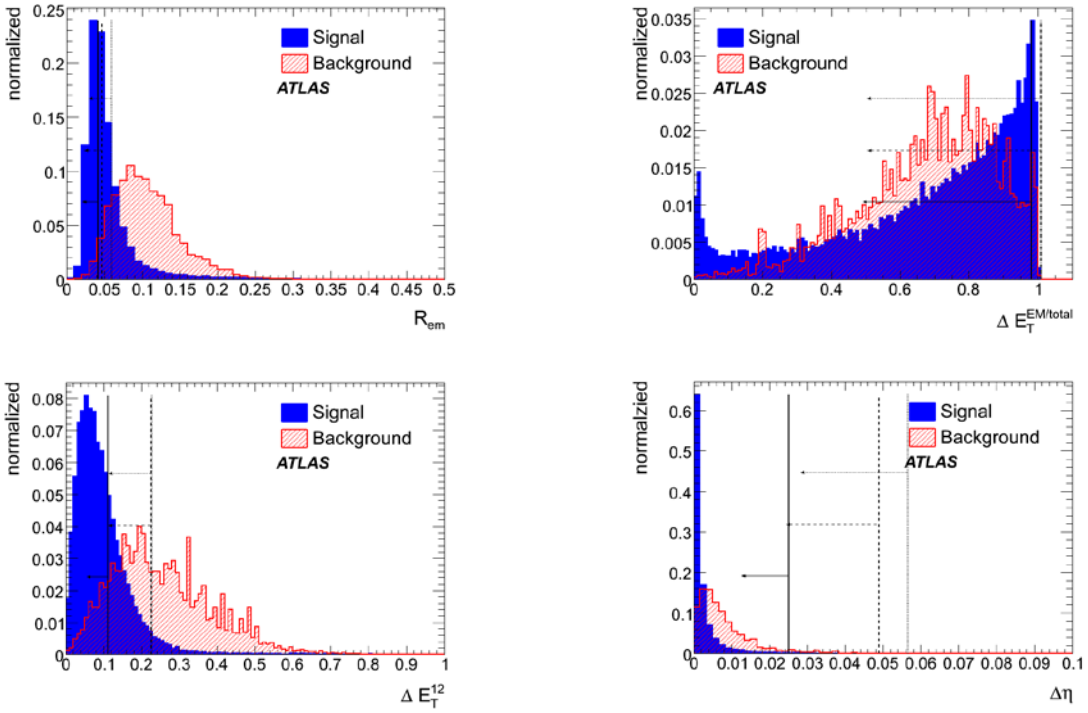


(a)

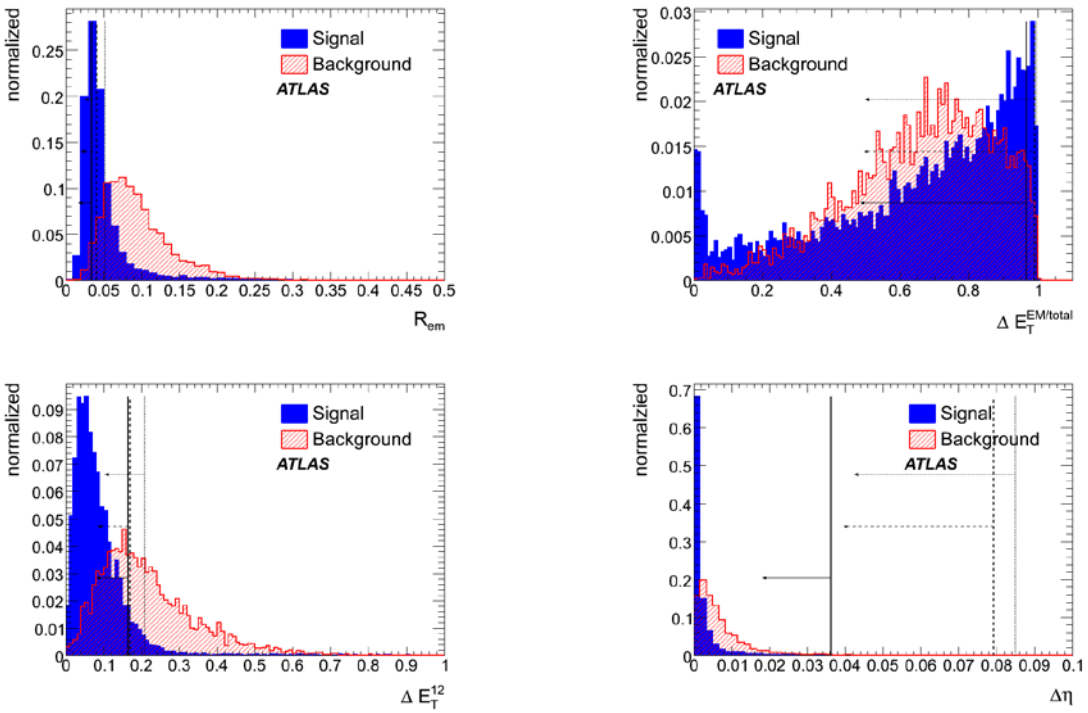


(b)

Figure 7: Distribution of calorimeter variables for one prong τ candidates within a p_T range of 0-20 GeV (7(a)) and 20-30 GeV(7(b)) for signal and background.



(a)



(b)

Figure 8: Distribution of calorimeter variables for one prong τ candidates within a p_T range of 45-60 GeV (8(a)) and 60-100 GeV(8(b)) for signal and background.

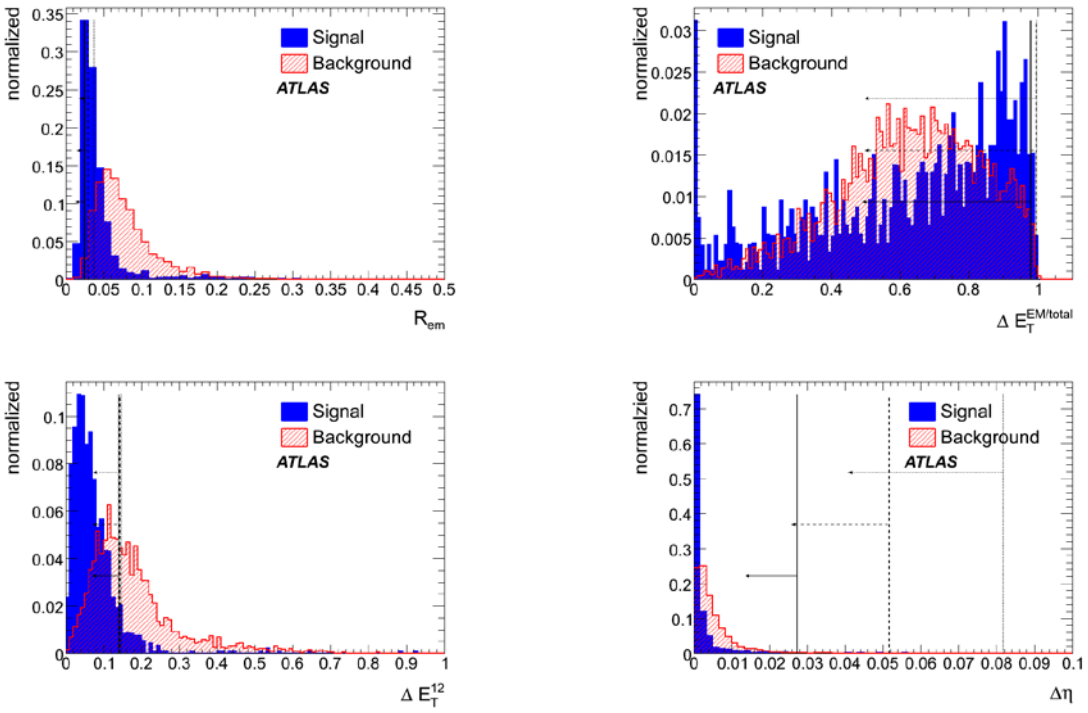
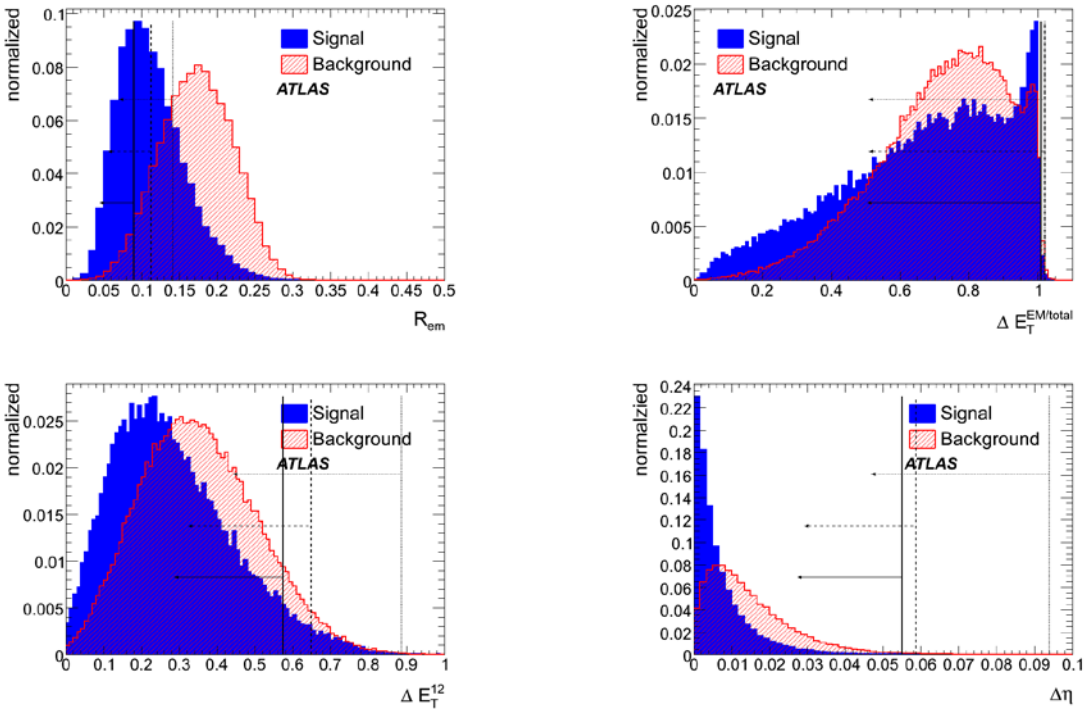
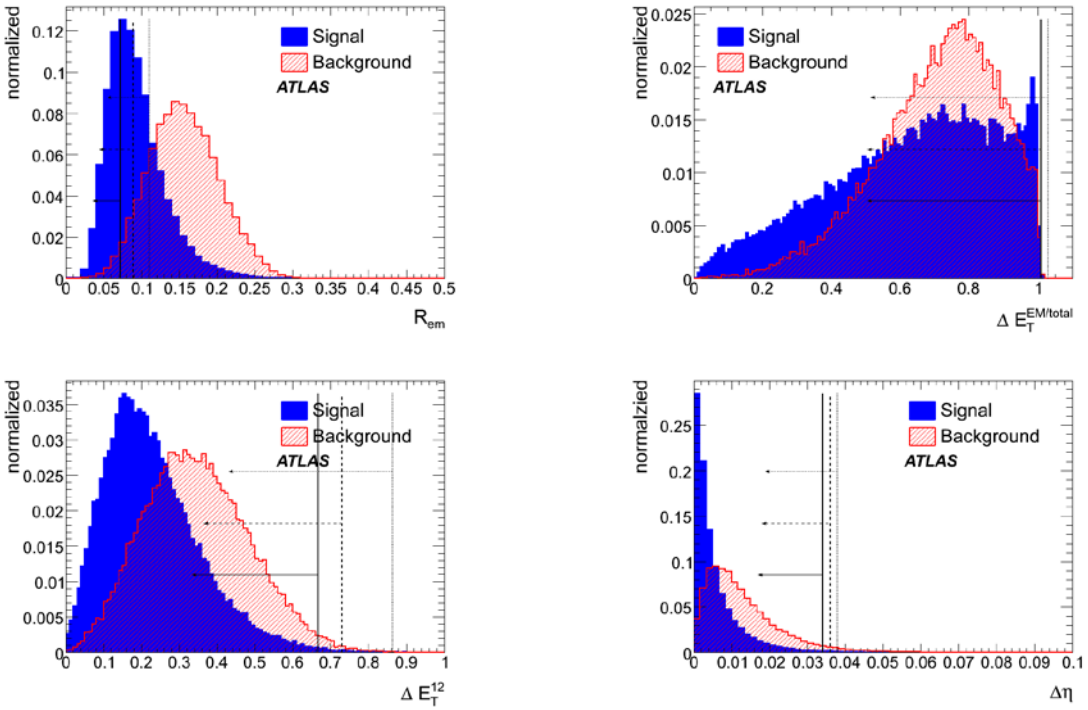


Figure 9: Distribution of calorimeter variables for one prong τ candidates within a p_T range of >100 GeV for signal and background.

B.2 3-prong truth matched τ candidates

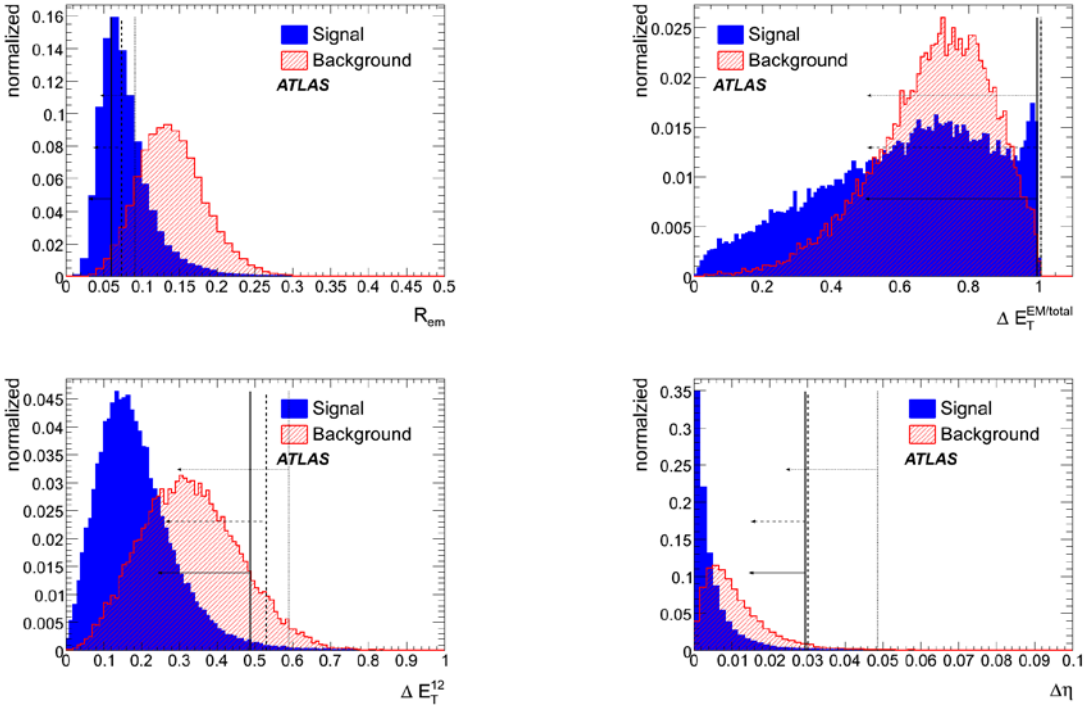


(a)

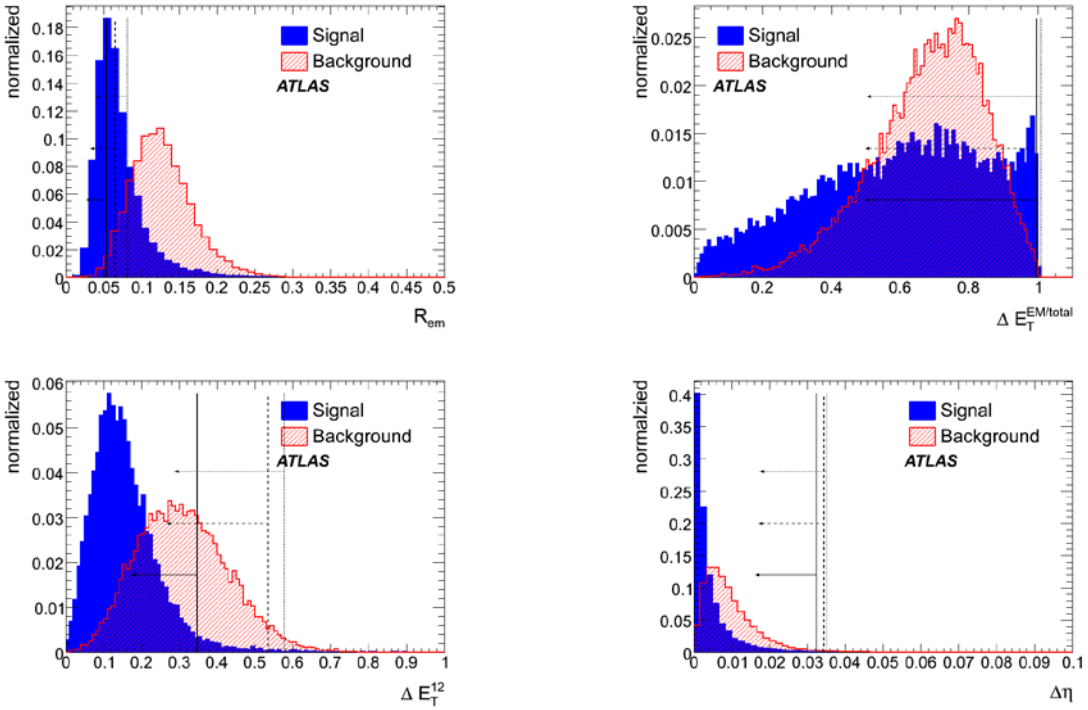


(b)

Figure 10: Distribution of calorimeter variables for three prong τ candidates within a p_T range of 0-20 GeV (10(a)) and 20-30 GeV(10(b)) for signal and background.

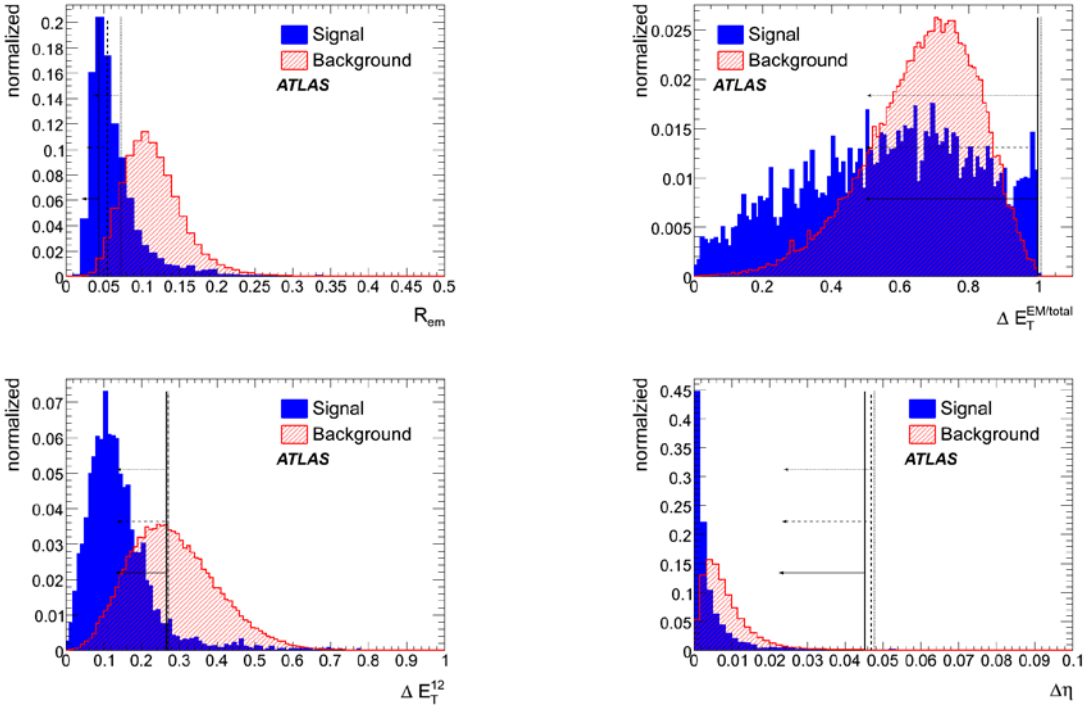


(a)

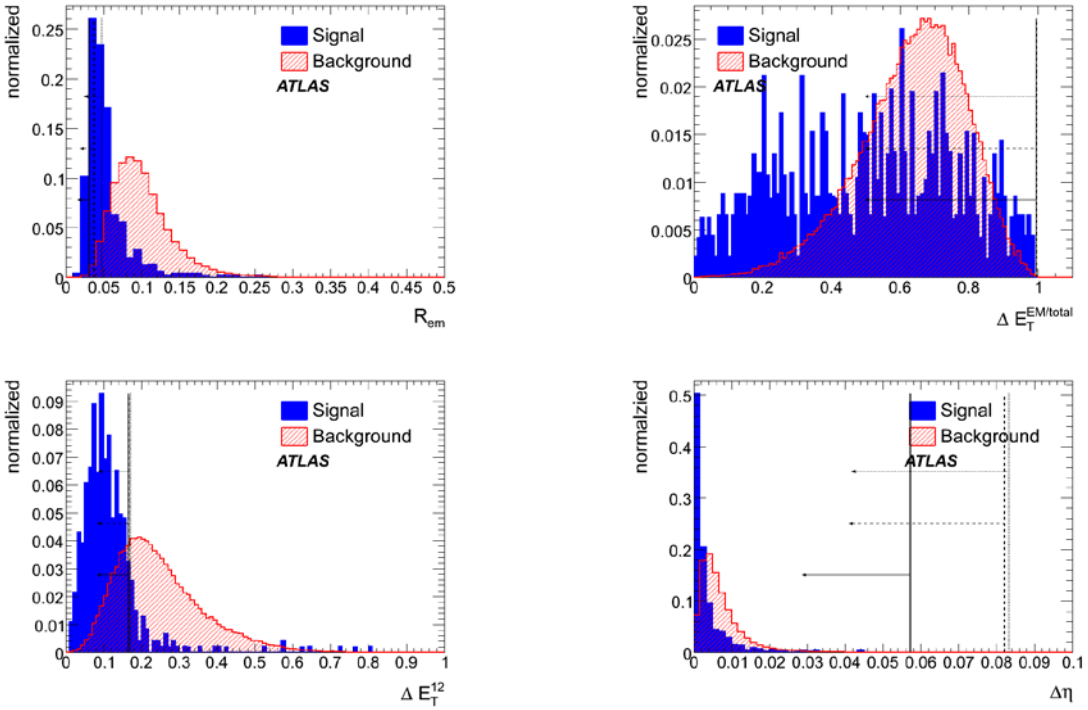


(b)

Figure 11: Distribution of calorimeter variables for three prong τ candidates within a p_T range of 30-45 GeV (11(a)) and 45-60 GeV(11(b)) for signal and background.



(a)

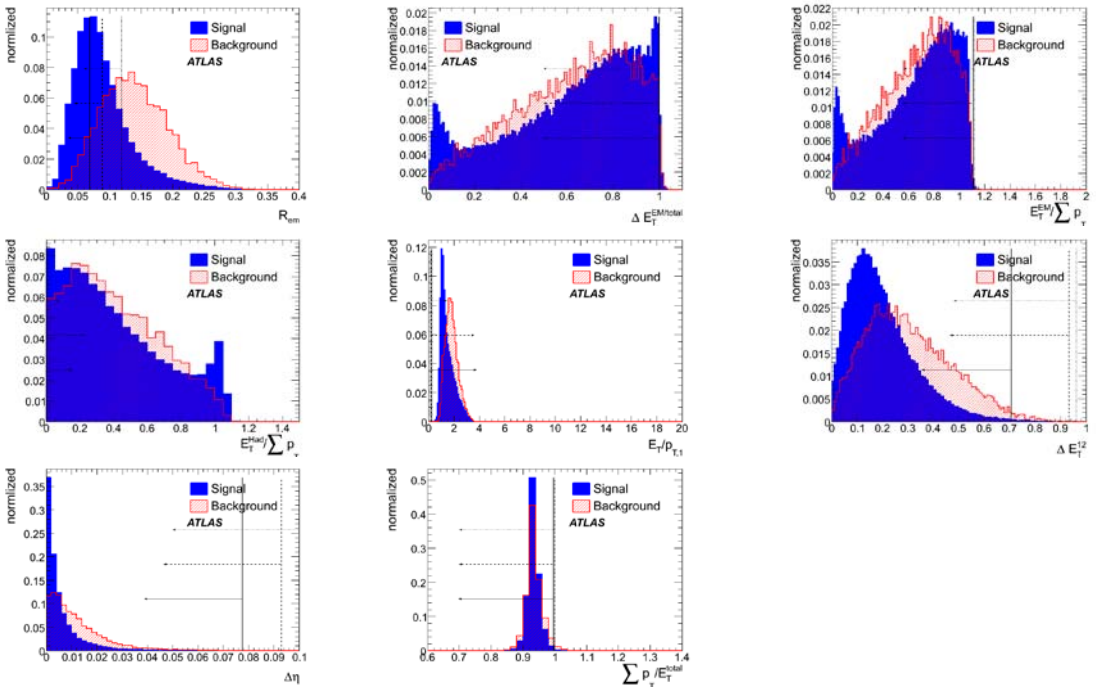


(b)

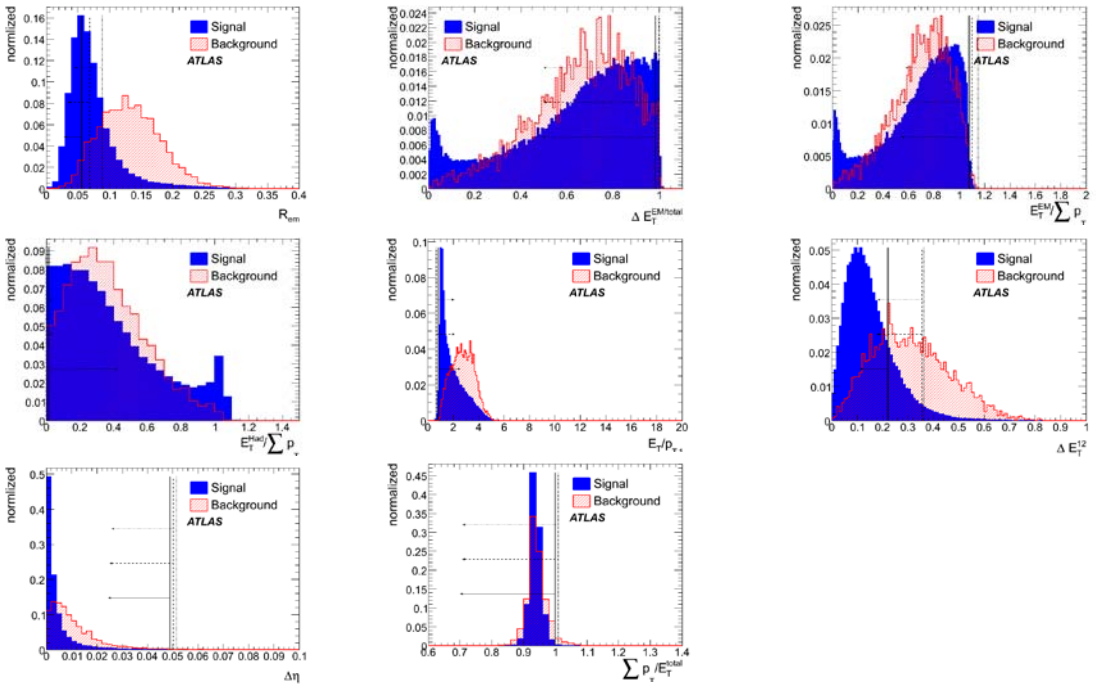
Figure 12: Distribution of calorimeter variables for three prong τ candidates within a p_T range of 60-100 GeV (12(a)) and >100 GeV(12(b)) for signal and background.

C Distribution of calorimeter and tracking variables

C.1 1-prong truth matched τ candidtes

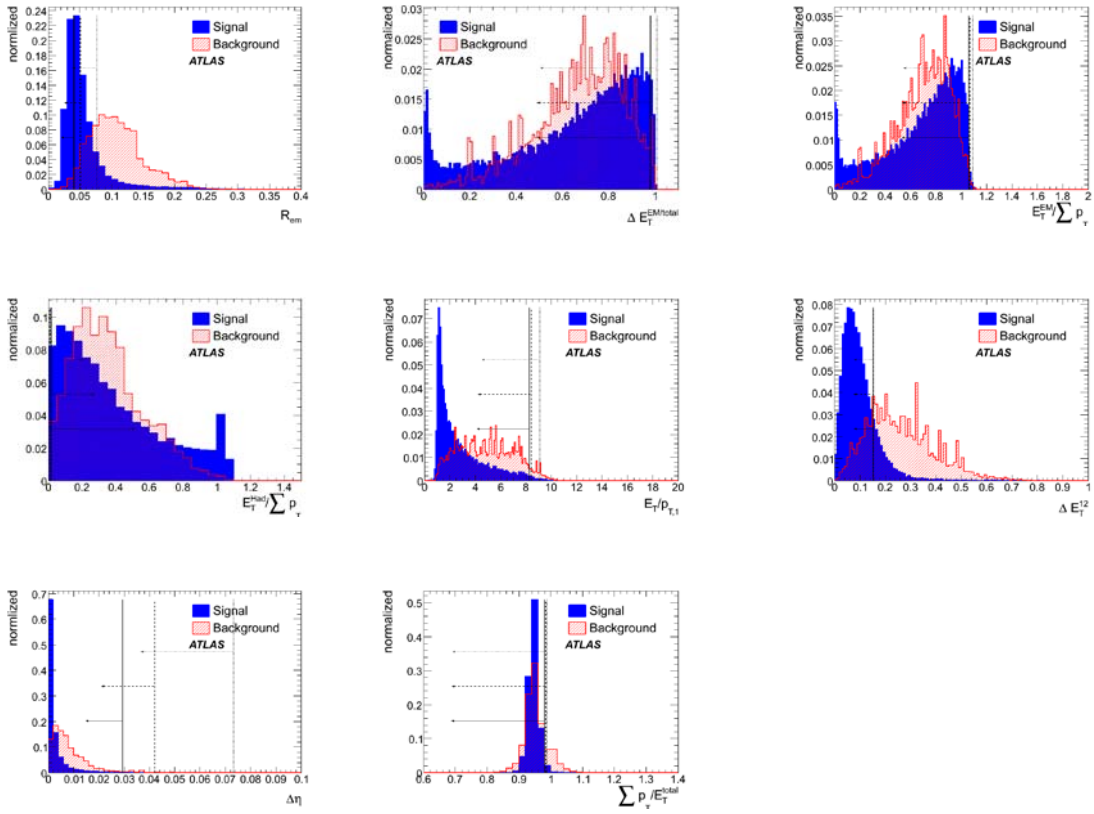


(a)

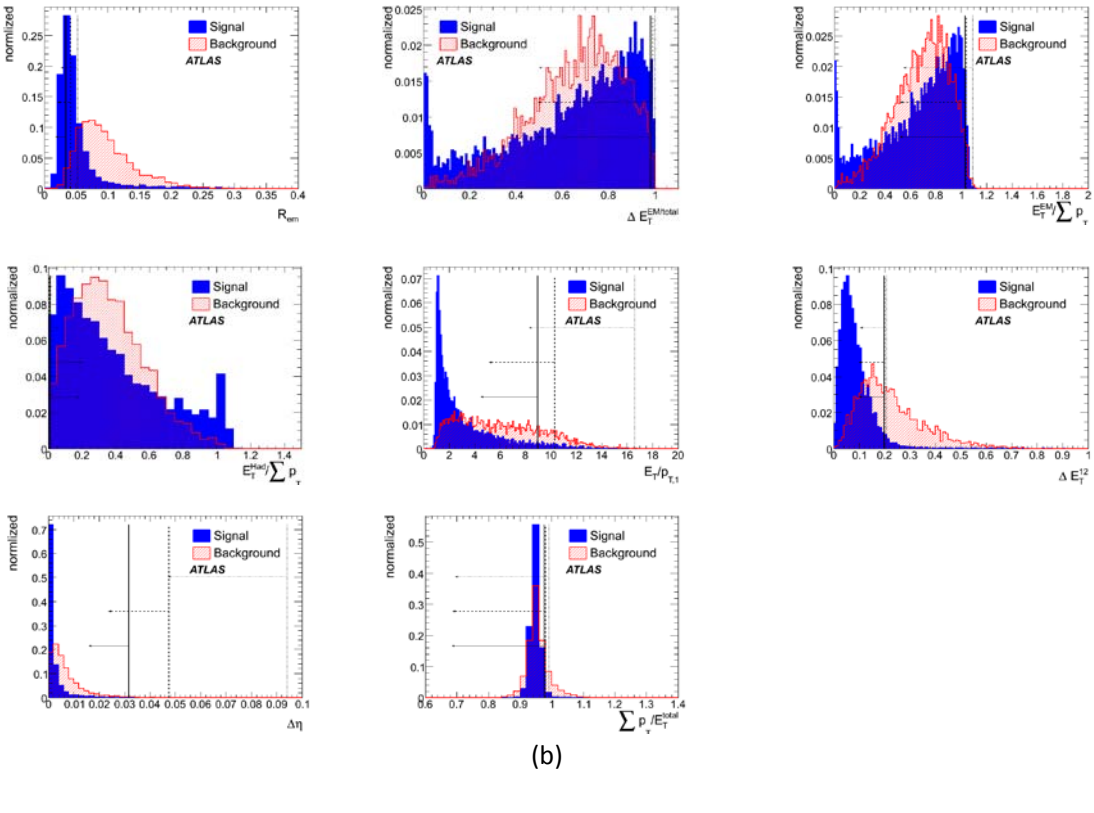


(b)

Figure 13: Distribution of calorimeter and tracking variables for one prong τ candidates within a p_T range of 0-20 GeV (13(a)) and 20-30 GeV(13(b)) for signal and background.



(a)



(b)

Figure 14: Distribution of calorimeter and tracking variables for one prong τ candidates within a p_T range of 45-60 GeV (14(a)) and 60-100 GeV(14(b)) for signal and background.

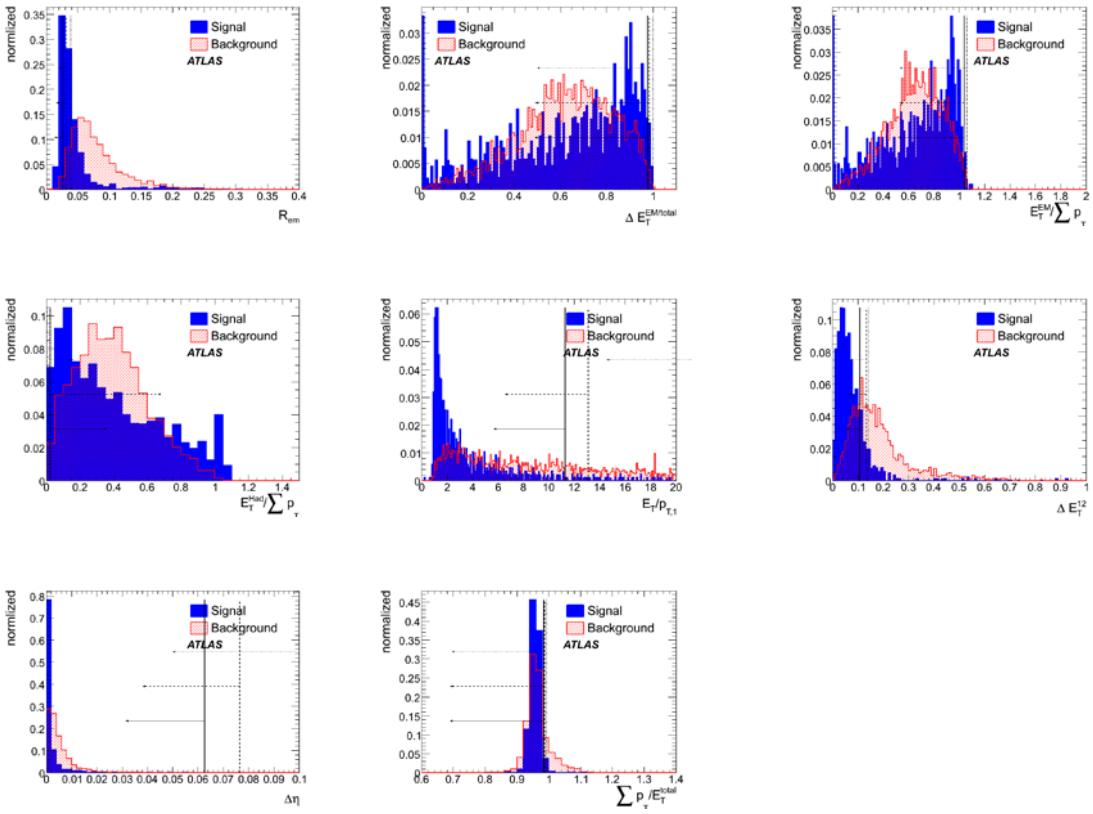
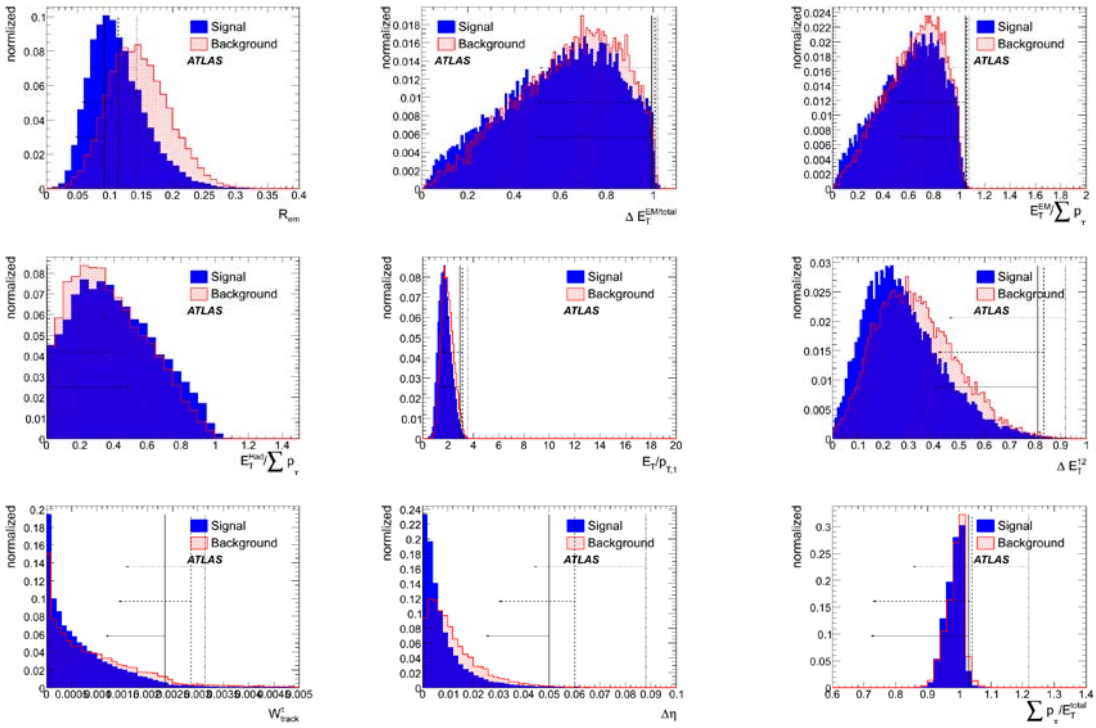
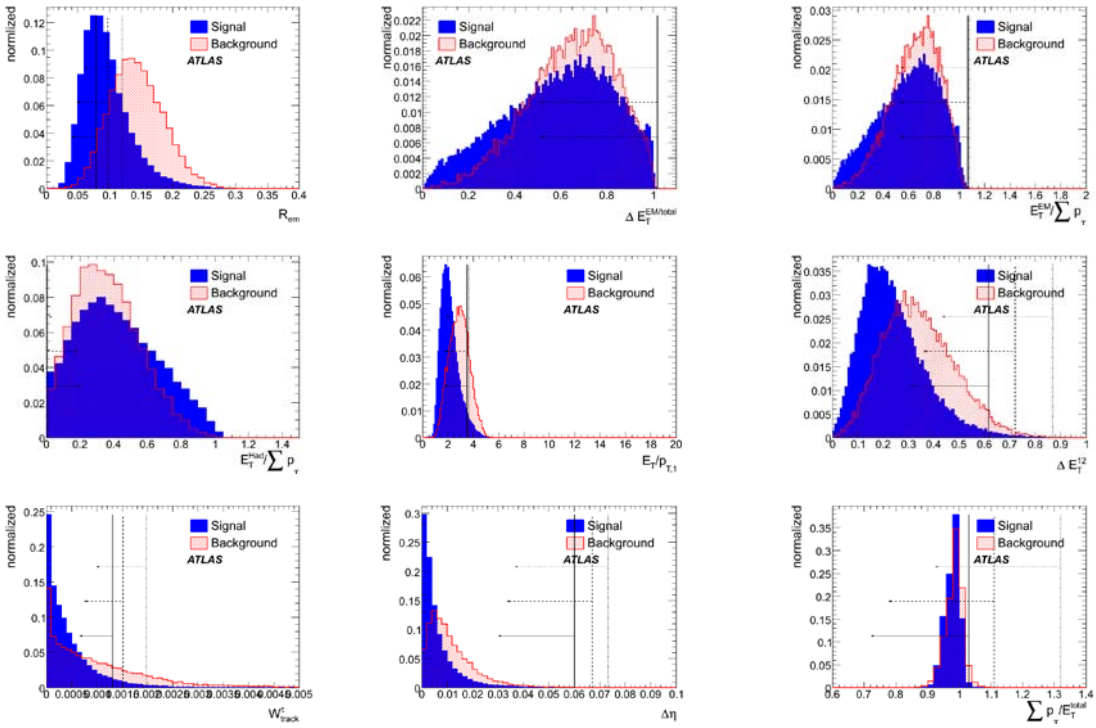


Figure 15: Distribution of calorimeter and tracking variables for one prong τ candidates within a p_T range of >100 GeV for signal and background.

C.2 3-prong truth matched τ candidates

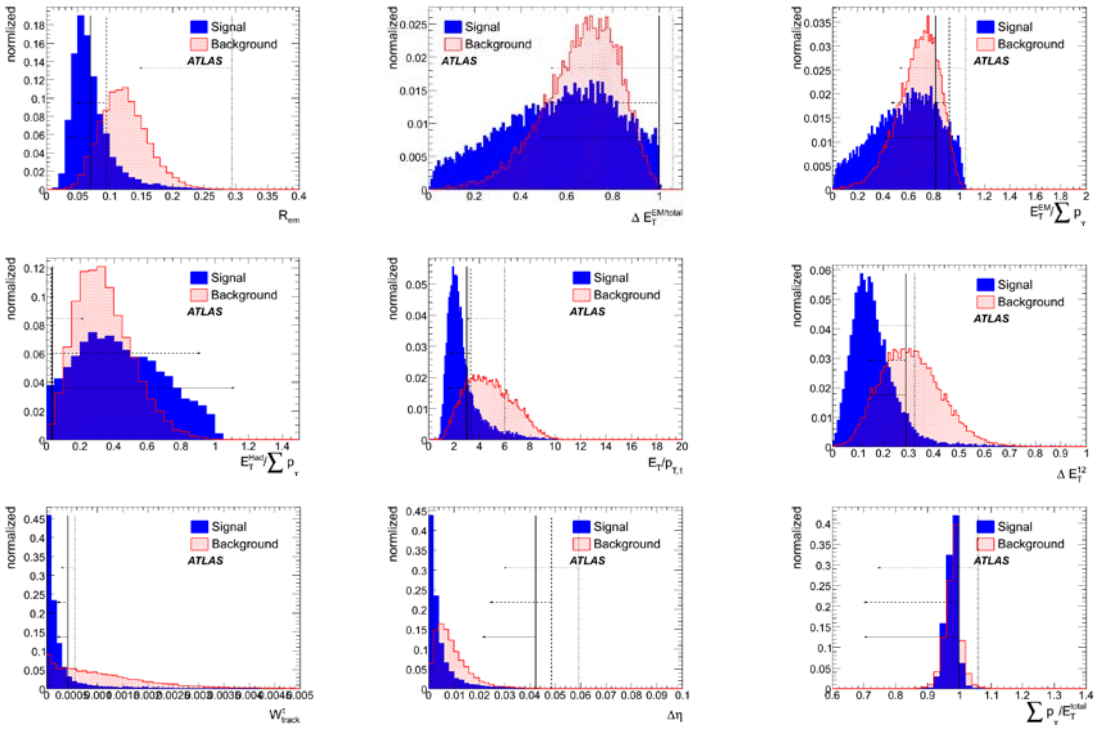


(a)

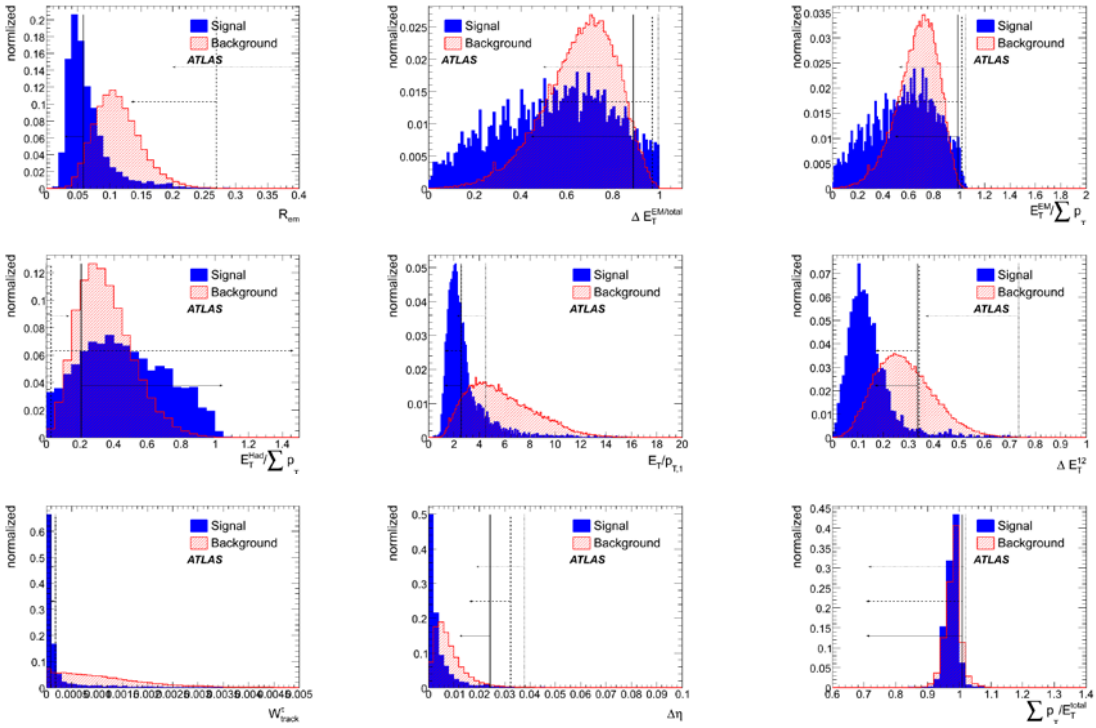


(b)

Figure 16: Distribution of calorimeter and tracking variables for three prong τ candidates within a p_T range of 0-20 GeV (16(a)) and 20-30 GeV(16(b)) for signal and background.



(a)



(b)

Figure 17: Distribution of calorimeter and tracking variables for three prong τ candidates within a p_T range of 45-60 GeV (17(a)) and 60-100 GeV(17(b)) for signal and background.

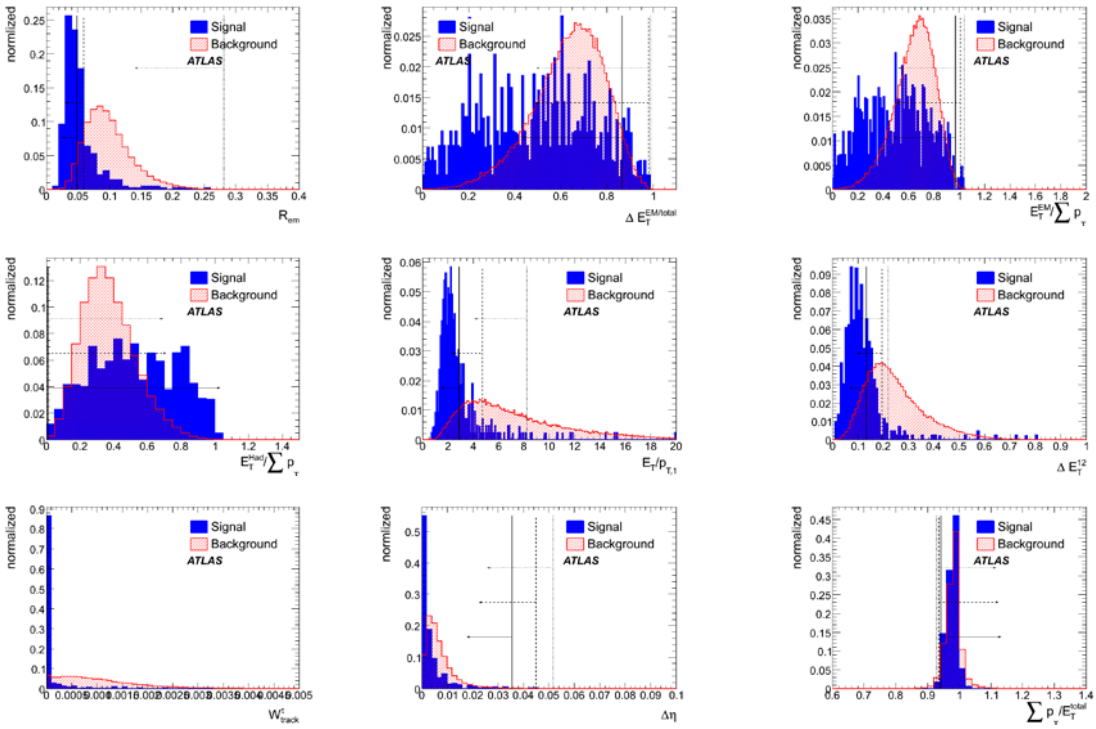


Figure 18: Distribution of calorimeter and tracking variables for three prong τ candidates within a p_T range of >100 GeV for signal and background.