



# Estimation of fake and non-prompt electrons using the Matrix Method

Moshe Barboy, Tel Aviv University, Israel

Supervisors: Nedaa Alexandra Asbah, Paul Glaysheer, Judith Katzy

September 7, 2017

## Abstract

This note illustrates the use of the Matrix Method for the background estimation of misidentified electron in the  $t\bar{t}$  semi-leptonic process that was measured using the proton-proton collision at 13 TeV recorded by the ATLAS detector. The collected data corresponds to  $33.2\text{fb}^{-1}$  of luminosity. The main focus of this study is to optimize the choice of the dedicated control regions used for the estimation of the fakes and non-prompt electrons.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Fake background of the $t\bar{t}$ process . . . . .	3
1.2	Regions . . . . .	4
<b>2</b>	<b>Matrix Method</b>	<b>4</b>
<b>3</b>	<b>Calculating the Background</b>	<b>5</b>
3.1	Efficiency calculation . . . . .	5
3.2	Selections . . . . .	6
3.3	Choosing cuts . . . . .	6
3.4	Calculating Efficiencies . . . . .	7
<b>4</b>	<b>Results</b>	<b>7</b>
<b>5</b>	<b>Conclusions</b>	<b>8</b>
<b>6</b>	<b>Appendix</b>	<b>9</b>

# 1 Introduction

## 1.1 Fake background of the $t\bar{t}$ process

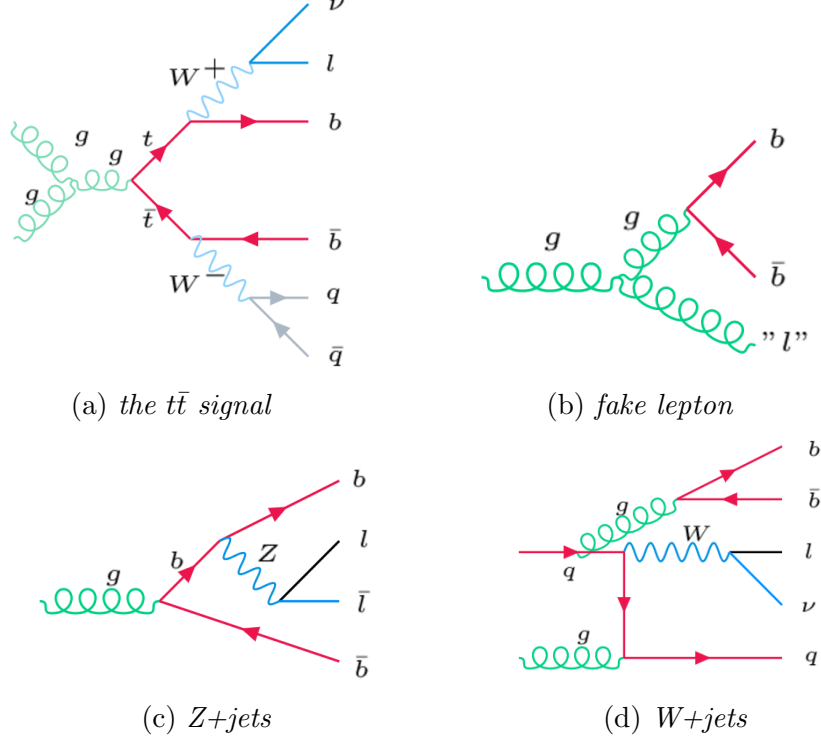


Figure 1: Diagram (a) shows the  $t\bar{t}$  process. Diagram (b) is the background coming from a gluon misidentified as an electron. The process generate the same final state even though It does not involve Top at all. Diagrams (c) and (d) are examples of the most dominant background in the regions of interest which were simulated in Monte Carlo.

Measuring the quark pair  $t\bar{t}$  production is important for checking the standard model. Feynman diagrams of the main process and some of its major backgrounds are illustrated in Fig. 1. Top quarks decay into W boson and bottom quark almost 100% of the time. The W decays into a lepton and a neutrino or into quarks. In this study we use events where one W dcays into an electron and a neutrino, and the other one decays into quarks, so called "semi-leptonic" decay. One of the backgrounds for the leptonic top decay, which is used to identify the top, is a misidentified lepton. The misidentification of an electron can occur because of either a wrongly classified jet, or an isolated electrons coming from semi-leptonic quark decays that are identified as part of the top decay. In the rest of this note, both cases of wrongly identifying an electron will be noted as fake. The fake background appears in a small corner of the phase space where it is poorly estimated by the Monte Carlo(MC). As a result, the estimation is done by data driven techniques such as the Matrix Method.

## 1.2 Regions

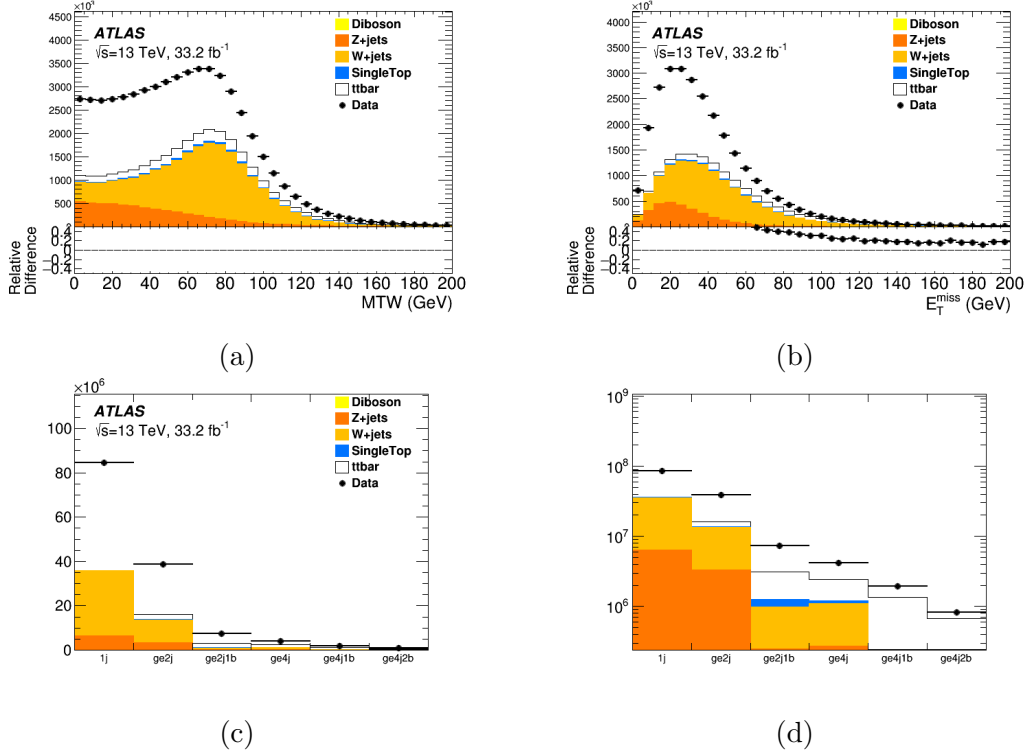


Figure 2: (a) and (b) show the distribution of the data in the  $ge2j$  region in MTW and MET. The gap between data and MC is assumed to come mostly from fake events. Diagram (c) presents the available statistics in different regions. Diagram (d) is a logarithmic scaled diagram (c). For the signal region,  $ge4j2b$ , all the backgrounds are greatly reduced.

The measured events can be divided into different regions, according to the number of jets measured in the final state and the number of b-jets (jets originated from b-quarks). In case of the  $t\bar{t}$  semi-leptonic process the main signal region is the final state of 4 jets, 2 of which are b-tagged (the other 2 jets come from the non-leptonic W decay). The main regions of focus for the analysis are 1 jet (denoted 1j) and more than 1 jet (denoted  $ge2j$ ). The histograms of the events in  $ge2j$  region with regards to Missing Transverse Energy (MET) and Transverse W Mass (MTW, defined  $M_T^W = \sqrt{E_T^{miss} P_T (1 - \cos \Delta\phi)}$ ) are shown in Fig. 2a, 2b. In Fig. 2c shows the number of events in each region. For regions 1j and  $ge2j$  there is a large gap between data and MC. We assume that these regions are rich in fake events.

## 2 Matrix Method

The Matrix Method is a data driven method used to estimate fake events. Using control regions where you know the behavior of the data, you can extrapolate your findings to

the signal's phase-space. The method assumes similar behavior of fake event fraction of the data in different regions.

Leptons in each event are identified using some quality requirements that reduce the background. The selection of the leptons using all these requirements is called tight, and the selection with less strict requirements is called loose. In this method the tight selection is a subset of the loose selection. For both selection, denoted by "t" for tight and "l" for loose, the number of leptons in the signal can be expressed as the sum of real leptons( $N_r$ ) and false (wrongly identified) leptons( $N_f$ ):

$$\begin{aligned} N^t &= N_r^t + N_f^t \\ N^l &= N_r^l + N_f^l = N_r^t/r + N_f^t/f \end{aligned} \quad (1)$$

Where we define fake and real efficiencies:  $f \equiv \frac{N_f^t}{N_f^l}$ ,  $r \equiv \frac{N_r^t}{N_r^l}$ . Assuming these are known quantities you can solve the equations for the fake background,  $N_f^t$ , and get:

$$N_f^t = \frac{f}{r - f}(rN^l - N^t) \quad (2)$$

If the efficiencies are event dependent, then the number of tight fake events is represented as a sum of weights:

$$N_f^t = \sum_i w_i = \sum_i \frac{f_i}{r_i - f_i}(r_i - \delta_i) \quad (3)$$

Where  $\delta$  is 1 for events that pass the tight selection and 0 otherwise. From equation 3 one can see that for known per-event efficiencies it is easy to calculate the fake background of the experiment.

In practice, the efficiencies,  $\epsilon$ , depend on many variables:  $\epsilon = \epsilon(\vec{x}, \vec{y})$ , where the  $x$  are the regions we want to consider and the  $y$  are continuous parameters we calculated the efficiencies for. Assuming no correlation between the  $y$  parameters and having the efficiencies given separately for each  $y$  (in 1d histograms). We calculate the efficiency of a given event using:

$$\epsilon(\vec{x}, \vec{y}) = \frac{1}{\epsilon(\vec{x})^{M-1}} \prod_i^M \epsilon(\vec{x}, y_i) \quad (4)$$

Where  $\epsilon(\vec{x})$  and  $\epsilon(\vec{x}, y_i)$  is the efficiency calculated in the  $\vec{x}$  regions. The real(fake) efficiencies are calculated by going to a Control Region(CR) in the phase-space where there are mostly real(fake) events, and taking the ratio between the number of tight and loose events.

## 3 Calculating the Background

### 3.1 Efficiency calculation

In our case, the control region taken for the fake electrons is low MET or MTW, because then its more probable that the neutrino from the W decay is absent and that the recognized leptons are fake. For real electrons the control region taken is the Z decay events,

where you identify one electron in loose selection and afterwards check if the second electron passes the tight selection, so that:  $r = \frac{\text{number of matching tight electrons found}}{\text{number of identified loose electrons in Z decay}}$ . The efficiency was calculated in regions 1j and ge2j for parameters:

- $P_T$  - leptons transverse momentum
- $\eta$  - longitudinal forward angle against the beam
- dR - angular distance between leading jet and lepton
- $JetP_T$  - leading jet Pt
- $\Delta\phi$  - angular azimuthal distance between MET and lepton

The y parameters from equation 4 are the different continuous parameters in the above list.

### 3.2 Selections

The experimental data used in the analysis corresponds to  $33.2\text{fb}^{-1}$  which was collected in the ATLAS detector at 13 TeV. Events are required to have at least 1 jet and 1 electron. Loose electrons are required to pass at least one of the triggers: HLT\_e26\_lhtight\_nod0\_1varloose, HLT\_e60\_lhmedium\_nod0, HLT\_e140\_lhloose\_nod0. Only events with electron  $P_T$  of at least 27 GeV are taken to compensate for the energy error of the lowest  $P_T$  trigger. Loose electrons also have MediumLH identification level. Tight electrons should also pass some additional isolation requirement, and have TightLH identification level instead of MediumLH.

### 3.3 Choosing cuts

In the histograms in Fig. 2 most of the fake events located in the lower spectrum of MET and MTW. Calculating the fake efficiency for electron was done taking different cuts on these parameters and assuming for this cuts we have no real events left. The different cuts are(all numbers are in GeV):

- $MTW < 20 \& MET + MTW < 60$
- $MET < 20$
- $MTW < 35$
- $MET < 15 \& MTW < 30$
- $MTW < 60$

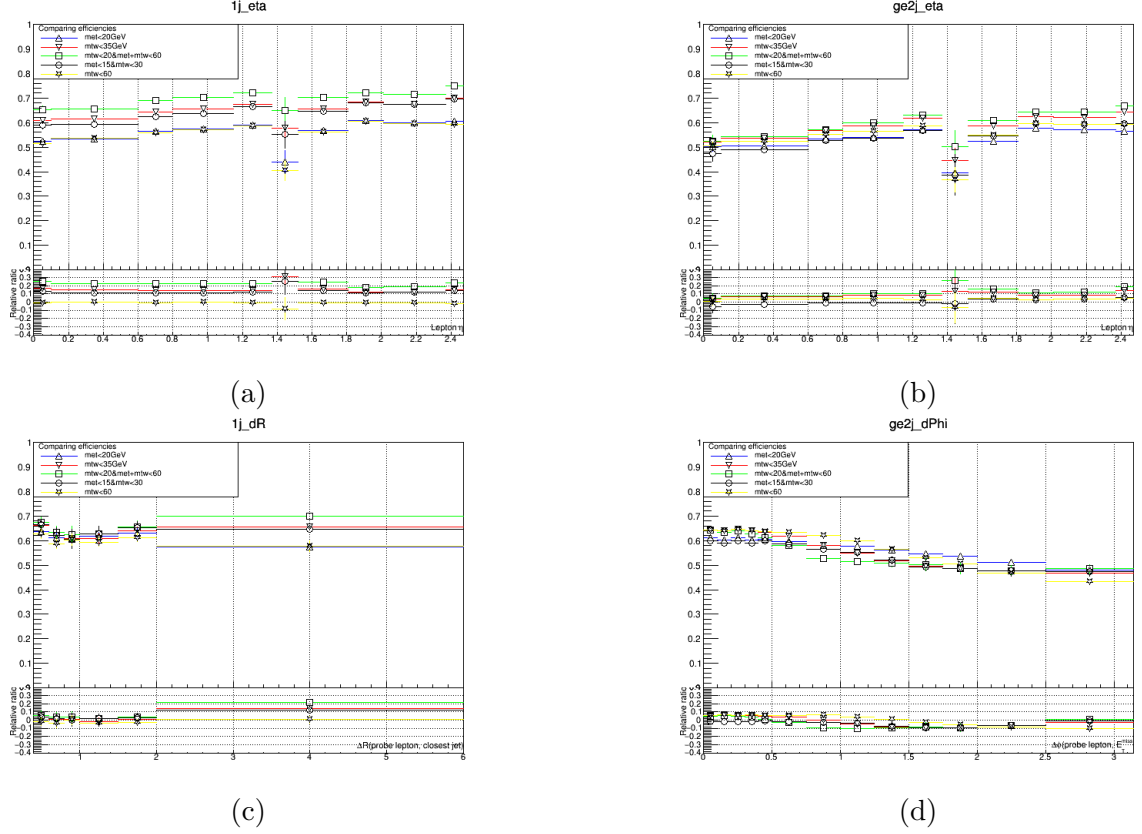


Figure 3: Fake efficiencies for  $\eta$  and  $dR$  parameters for regions  $1j$  and  $ge2j$ . In each histogram the efficiency for each cut and the differential ratio between different cuts and the  $MET < 20$  GeV cut. The region and parameter are written on top of the corresponding histogram.

### 3.4 Calculating Efficiencies

For each cut the efficiency was calculated with regard to every parameter on the list in Sec. 3.1. The results for some of the parameters are shown in Fig. 3. The efficiencies used in equation 3 to calculate the weights are derived from the calculated efficiencies using equation 4.

## 4 Results

The results in MET and MTW histograms are illustrated in Fig. 4. The histograms show the stack plots That are shown in Fig. 2, but with the fake estimation produced using the Matrix Method. In Fig. 5 there are also the results for different parameterization of the fake efficiency with  $MTW < 60$  GeV cut.

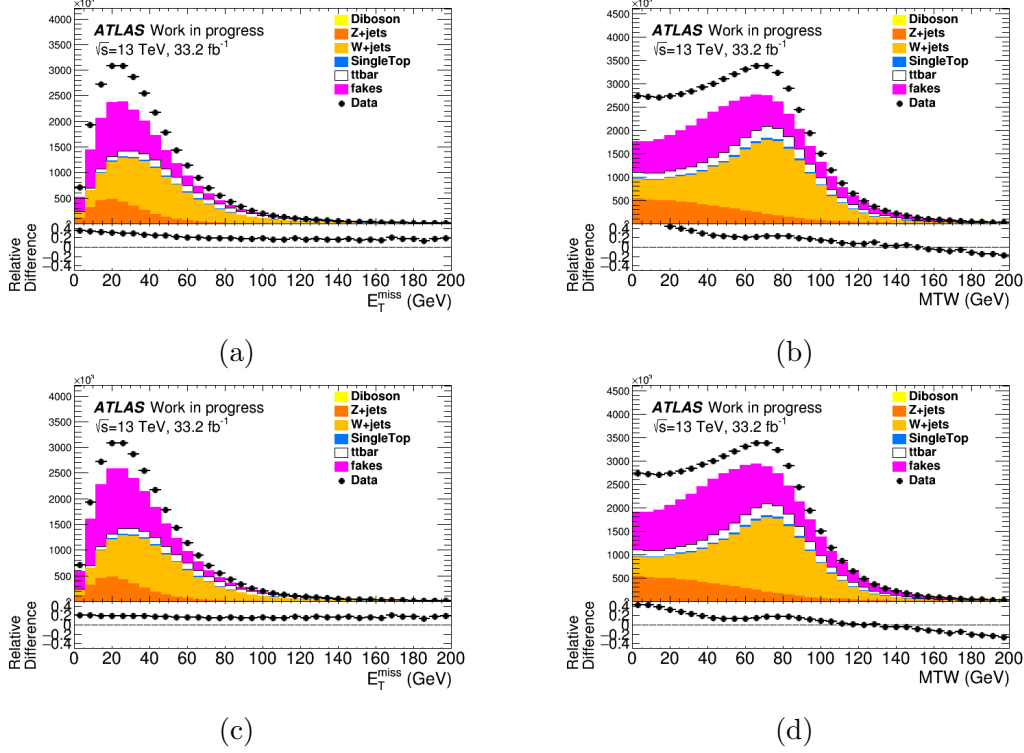


Figure 4: In (a) and (b) the estimation evaluated using the cut  $MET < 20$  GeV and fake and real parametrization with  $\eta, dR, jetP_T$ . In (c) and (d) the fake background was evaluated using the cut  $MTW < 60$  GeV and the same parametrization.

## 5 Conclusions

This work studied the behavior of the wrongly identified electron background of the  $t\bar{t}$  process. We estimated the background using different control regions and parametrizations for fake efficiency. Modifying these parameters introduced negligible changes compared to the uncertainties coming from the assumption of similar efficiency behavior in different regions when using the Matrix Method. Using the cut on MTW achieved a better behavior of the relative difference for MTW between 40 and 100 GeV. Using more parameters to evaluate the fake efficiency didn't change the resulting fake background estimation much (minimal differences between the plots reported in the appendix).

## References

- [1] ATLAS Collaboration, *Estimation of non-prompt and fake lepton backgrounds in final states with top quarks produced in proton-proton collisions at  $s=\sqrt{8}$  TeV with the ATLAS detector*, ATLAS-CONF-2014-058
- [2] xAODAnaHelpers library, <https://xaodanahelpers.readthedocs.io/en/master/>



[3] MatrixMethodLeptonEfficiencies library, <https://gitlab.cern.ch/jdandoy/MatrixMethodLeptonEfficiencies/>

## 6 Appendix

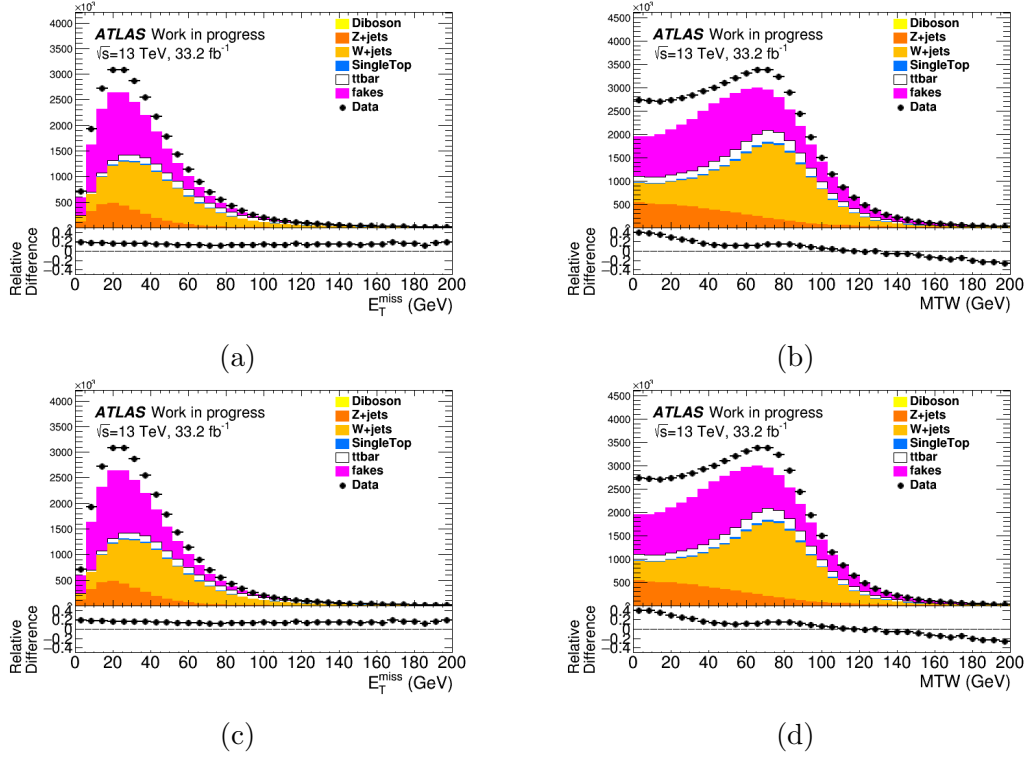


Figure 5: In (a) and (b) we see the estimation evaluated using fake parametrization with  $dR$ . In (c) and (d) the fake background was evaluated using the parametrization  $dR, \text{jet}P_T$ . All the histograms are for  $MTW < 60$  GeV cut with real parametrization  $\eta, dR, \text{jet}P_T$ .