Chapter 5

ESTIMATING THE RATES OF ELECTRON CHARGE MIS-IDENTIFICATION

Many physics analyses involve charged leptons in their final states, where leptons typically refer to electrons or muons. Such an electron or a muon leaves a track in the detector, and its charge, or more specifically the sign of the charge, is determined from the curvature — due to the installed magnetic fields — of the track. Because of some factors which will be discussed below, this determination could occasionally be erroneous, leading to what is called charge misidentification.

Electron charge mis-identification is important for analyses that involve same-sign electrons in the final state. Examples of such analyses include measurements of same-sign WW scattering [47], analyses that involve the production of a Higgs in association with a $t\bar{t}$ pair ($t\bar{t}H$), and supersymmetry search with two same-sign leptons [49]. In general, charge mis-identification rates occur on the order of O(1%), while Standard Modle processes that provide opposite-sign dileptons (dominantly $Z \to e^+e^-$ bosons) occur approximately 10^3 times more commonly than genuine Standard Model sources of same-sign leptons (dominantly WZ production). As a result, opposite-sign sources of dileptons suffering from charge mis-identification can constitute a large background in these searches, and so it is crucial to estimate the charge mis-identification background precisely.

This chapter describes a method for estimating the rate of charge mis-identification using a likelihood function. Section 5.1 discusses briefly how electron charge mis-identification might arise at ATLAS. Section 5.2 discusses the likelihood method, including the Poisson likelihood used as well as how it is applied to $Z \to e^+e^-$ events to measure the charge mis-identification rates. Finally, Section 5.3 provides some conclusions.

It is to be noted that muon charge mis-identification is known to be negligible, except at very high p_T. Indeed, the magnetic field in the Muon Spectrometer (see Section 3.3.2.3) allows the measurement of the track curvature over a larger radius, thereby reducing the chance the charge could be mis-identified. Analyses such as the supersymmetry search with two same-sign leptons mentioned above have found that muon charge mis-identification is in indeed negligible (cite).

5.1 Electron Charge Mis-identification

At ATLAS, the sign of the charge of an electron is determined from its track in the Inner Detector (see Section 3.3.2.1). Indeed, as the electron passes through the Inner Detector , its track is bent by the installed magnetic fields. The direction of the curvature of the track determines the charge of the electron.

Charge mis-identification, where the charge of the electron is identified incorrectly, occurs mainly because of two reasons:

- ❏ As the electron passes through the detector and interacts with the materials in the detector, it may radiate photons. These radiated photons may in turn convert to electron-positron pairs. A charge mis-identification occurs when the electron candidate is matched to the wrong track.
- ❏ The reconstructed track of the electron appears rather straight, i.e. the curvature of the track is small, at very high momentum or at large pseudorapidity, the latter because the lever arm of the tracker is limited.

5.2 The Likelihood Method

Since it is impossible to know with absolute certainty, after the charge of an electron has been measured, that a charge mis-identification has occured or not, we seek instead to determine the rates of charge mis-identification for an ensemble of electrons. Essentially, we start from a sample of electrons for which the true charges are known. Charge measurements on the sample will result in another sample that consists of electrons with the original charges as well as elecrons whose charges have switched. A key step is to write down the probababilistic distribution of charge assuming a rate of charge mis-identification, and then seek to determine this rate from the actual data. This method is called the likelihood method and will be discussed in section.

This chapter uses $Z \to e^+e^-$ events because these provide a good source of clean, high-statistics sample of electrons.

5.2.1 The Poisson Likelihood

Consider a pair of electrons e^+e^- (an opposite-sign pair) at truth level. The charges of the electrons are opposite of each other, but because of charge mis-identification there is a chance of this pair being identified as having the same charge (a same-sign pair). Assuming a probability p of such a chance, then in considering n pairs e^+e^-, the probability of seeing exactly n_{ss} same-sign pairs is given by the binomial distribution

$$P(n_{ss}) = \binom{n}{n_{ss}} p^{n_{ss}} (1-p)^{n-n_{ss}}.$$

Since it is known that the charge mis-identification probability p is typically small while the number of pairs considered n is typically very large , the Poisson distribution may be used to approximate the binomial distribution. Thus, let

$$m_{ss} = np \tag{5.1}$$

denote the expected number of same-sign pairs, it follows that

$$P(n_{ss}) = \frac{m_{ss}^{n_{ss}} e^{-m_{ss}}}{n_{ss}!} \tag{5.2}$$

is the Poisson probability of seeing n_{ss} same-sign pairs, given the average number of same-sign pairs m_{ss}. This will also be called a likelihood function.

Consider again an opposite-pair e^+e^- with the probability p of being identified as a same-sign pair. We may speak directly of a probability ϵ of an electron in the pair having its charge incorrectly identified. Then a same-sign pair results if only one of the electrons in the original opposite-sign pair has its charge identified incorrectly, which means we may write

$$p = (1-\epsilon)\epsilon + \epsilon(1-\epsilon). \tag{5.3}$$

The Poisson likelihood of Equation 5.2 may now be written to depend explicitly on ϵ:

$$P(n_{ss}|\epsilon) = \frac{m_{ss}^{n_{ss}} e^{-m_{ss}}}{n_{ss}!}, \quad m_{ss} = np = n(1-\epsilon)\epsilon + \epsilon(1-\epsilon). \tag{5.4}$$

In an actual measurement of the rates of charge mis-identification, the individual rates ϵ's are in general different for the electrons in the pair, if for example the dependence of ϵ on the transverse momenta p_T is taken into account. We may introduce then a number of bins in p_T and may speak of a charge mis-identification probability associated with a bin i. Consequently an electron pair is associated with a pair of bins (i, j), and we have the following quantities:

- The probability

  $$p_{ij} = (1-\epsilon_i)\epsilon_j + \epsilon_i(1-\epsilon_j) \tag{5.5}$$

  in place of the probability p in Equation 5.3. This is the probability an opposite-sign pair may be seen as a same-sign pair in the bin pair (i, j)

- The number of electron pairs considered, n_{ij}, in the bin pair (i, j)

- The expected number of same-sign pairs

  $$m_{\mathrm{ss,ij}} = n_{ij} p_{ij} \tag{5.6}$$

  in place of the expected number of same-sign pairs in Equation 5.1

- The Poisson likelihood

  $$P(n_{ss,ij}|\epsilon_i, \epsilon_j) = \frac{m_{ss,ij}^{n_{ss,ij}} e^{-m_{ss,ij}}}{n_{ss,ij}!} \tag{5.7}$$

in place of the Poison likelihood in Equation 5.4. This will also be denoted simply as L_{ij}

These equations remain the same if instead of one-dimensional bins, two dimensional bins are used. Indeed, suppose the dependency of the charge mis-identification rates on, say p_T and η, needs to be taken into account. If there are a bins in p_T and b bins in η, the total number of bins ab could be labelled $1, 2, \cdots, ab$ and treated as an ordered sequence of one-dimensional bins. The actual p_T and η binning together with the estimation of the charge mis-identificaion rates will be discussed in the following section.

All the possible bin pairs (i, j) need to be used and therefore, assuming statistically-independent rates, we will maximize the likelihood function

$$L = \prod_{i,j} L_{ij}$$

to find the rates ϵ_i, the data being n_{ij}, the numbers of electrons observed in the bin pair (i, j), and $n_{ss,ij}$, the number of same-sign electron pairs observed in the bin pair (i, j).

5.2.2 Estimation of the Rates on $Z \rightarrow e^+e^-$ sample

The Poisson likelihood is applied on a $Z \rightarrow e^+e^-$ data sample to determine the charge mis-identification rates as follows. First, several selections are applied, including

- Logical OR between two single-electron triggers, one with $E_T > 24$ GeV plus Medium identification, one with $E_T > 60$ GeV plus Loose identification
- At least two electron candidates with $|\eta| < 2.47$
- One electron is required to pass the Tight identification requirement, and to have $E_T > 25$ GeV. The other electron must have $E_T > 10$ GeV and must satisfy the track quality criteria (the tracks associated with the electron must have at least one hit in the pixel detector and at least seven hits in the pixel and SCT detectors)
- The invariant mass is within ± 15 GeV of the Z mass

The invariant mass of the pair of electrons will play an important role in the following discussion. Figure 5.1 [37] shows the invariant mass distribution m_{ee} for two different η range, each electron having $0.0 < \eta < 0.8$ and $2.0 < \eta < 2.47$; in both figures the each electron in the pairs is selected to have E between 25 GeV and 50 GeV. Due to charge mis-identification same-sign electron pairs exist in addition to opposite-sign pairs, and they are plotted alongside opposite-sign pairs. It is seen that same-sign pairs have a broader peak which is also slightly shifted to lower values, consistent with the fact that radiation which causes charge-misidentification also causes energy loss. It is also seen that charge-misidentification is higher at higher η, as have been commented previously.

To continue, an invariant mass interval (m_l, m_h) is selected, where $m_l = 15$ GeV is the low mass point and $m_h = 15$ GeV the high mass point around the Z mass peak. Then electrons in the events are binned according to their η and p_T values. Each electron pair then is associated to a pair of bin (i, j), and all such bin pairs need to be taken into account. The quantities needed are (see Section 5.2.1):

- n_{ij}, the number of electrons counted in the bin pair (i, j)
- $n_{ss,ij}$, the number of same-sign electron pairs counted in the bin pair (i, j)

There are non-signal events contamination in these quantities and to deal with them we assume that the two sides of the interval (m_l, m_h) are dominated by background contributions, and adopt following method. To begin, in addition to the original invariant mass interval (m_l, m_h), we consider the interval $(m_l - w_l, m_h + w_h)$ where $w_l = 15$ GeV and $w_h = 15$ GeV are some widths. The latter interval is made up of three intervals:

- (m_l, m_h). This is the original invariant mass interval.
- $(m_l - w_l, m_l)$. This is the interval that lies to the left of the original interval
- $(m_h, m_h + w_h)$. This is the interval that lies to the right of the original interval

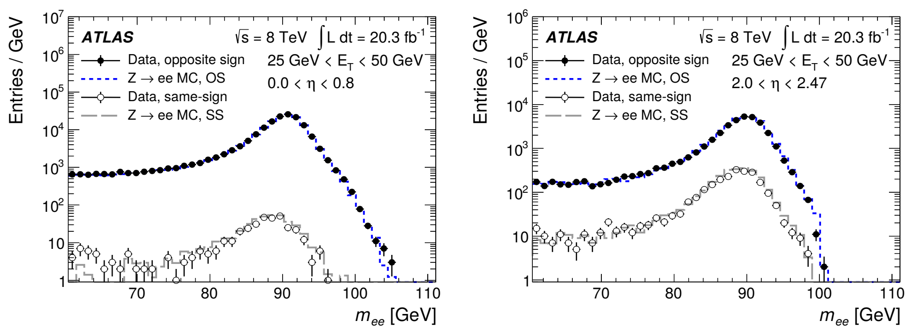

Figure 5.1: Distribution of the invariant mass m_{ee} for E_T between 25 and 50 GeV and $|\eta|$ between 0.0 and 0.8 [37]. Due to charge mis-identification same-sign pairs as well as opposite-sign pairs are seen.

Then, in addition to the quantities n_{ij} and $n_{ss,ij}$ in the original central interval (m_l, m_h), we will consider the corresponding quantities in the two new intervals, to be denoted n^l_{ij} and $n^l_{ss,ij}$ in the left interval and n^h_{ij} and $n^h_{ss,ij}$ in the right interval. We assume the left and right intervals are background intervals and subsequently compute the weighted quantities $b(n_{ij})$, to mean the background contamination in n_{ij}, and $b(n_{ss,ij})$, to mean background contamination in $n_{ss,ij}$:

$$b(n_{ij}) = \frac{w_l \times n^l_{ij} + w_h \times n^h_{ij}}{w_l + w_h}, \qquad b(n_{ss,ij}) = \frac{w_l \times n^l_{ss,ij} + w_h \times n^h_{ss,ij}}{w_l + w_h}$$

which will be taken as the backgrounds in n_{ij} and $n_{ss,ij}$ in the central interval respectively.

The terms n_{ij} and $n_{ss,ij}$ and the background terms $b(n_{ij})$ and $b(n_{ss,ij})$ are to be used as follows. According to Equation 5.7 the Poisson likelihood to be fitted is

$$P(n_{ss,ij}|\epsilon_i,\epsilon_j) = \frac{m_{ss,ij}^{n_{ss,ij}} e^{-m_{ss,ij}}}{n_{ss,ij}!}$$

The background terms make a contribution to the expected number of same-sign $m_{ss,ij}$ in the likelihood, modifying it from $m_{ss,ij} = n_{ij}p_{ij}$ (see Equation 5.6) to

$$m_{ss,ij} = (n_{ij} - b(n_{ij})) \times p_{ij} + b(n_{ss,ij})$$

The first quantity on the right in the equation above is the same-sign contribution from signal events where the background has to be subtracted, and the second quantity is the contribution from background events.

5.2.3 Charge Mis-identification Rates and Uncertainties

The rates are obtained upon the maximization of the likelihood function discussed in the previous section. The statistical uncertainties associated with the estimated rates depend on the statistics of the data, and are given by the statistical tool that maximizes the Poisson likelihood.

The following sources of systematic uncertainties are evaluated:

- ❏ Systematic uncertainty that comes from background subtraction, which is evaluated by determining the rates with and without background subtraction. The inclusion of this uncertainty ensures a conservative figure of systematic uncertainty in the charge mis-identification rates; it has a small impact because the background is small.
- ❏ The invariant mass interval (m_l, m_h) may be varied, from 15 GeV around the Z mass to 10 and 20 GeV additionally. In this way an idea of how the selection of an interval may affect the rates may be obtained.
- ❏ The invariant mass widths w_l and w_h may be varied, taking values 20, 25, or 30 GeV. This takes into account the uncertainty on the rates due to the choice of a mass width.

The actual rates are estimated for the following three sets of requirements:

- ❍ Medium: Medium identification requirements
- ❍ Tight + isolation: Tight identification requirements plus track isolation cut $p_T^{\mathrm{cone}\ 0.2}/E_T < 0.14$.
- ❍ Tight + isolation + impact parameter: Tight identification plus $E_T^{\mathrm{cone}\ 0.3}/E_T < 0.14$ and $p_T^{\mathrm{cone}\ 0.2}/E_T < 0.07$ and additionally $|z_0| \times \sin\theta < 0.5$ mm and $|d_0|/\sigma_{d_0} < 5.0$

Figure 5.2 [37] show the estimated rates in data and simulation. The dashed lines indicate the bins in which the rates are calculated. Total uncertainty, which is computed as the sum in quadrature of statistical and systematic uncertainties, is also showed. In most bins, simulation over-estimates the rates as compared to the data by 5-20% depending on η and electron requirements.

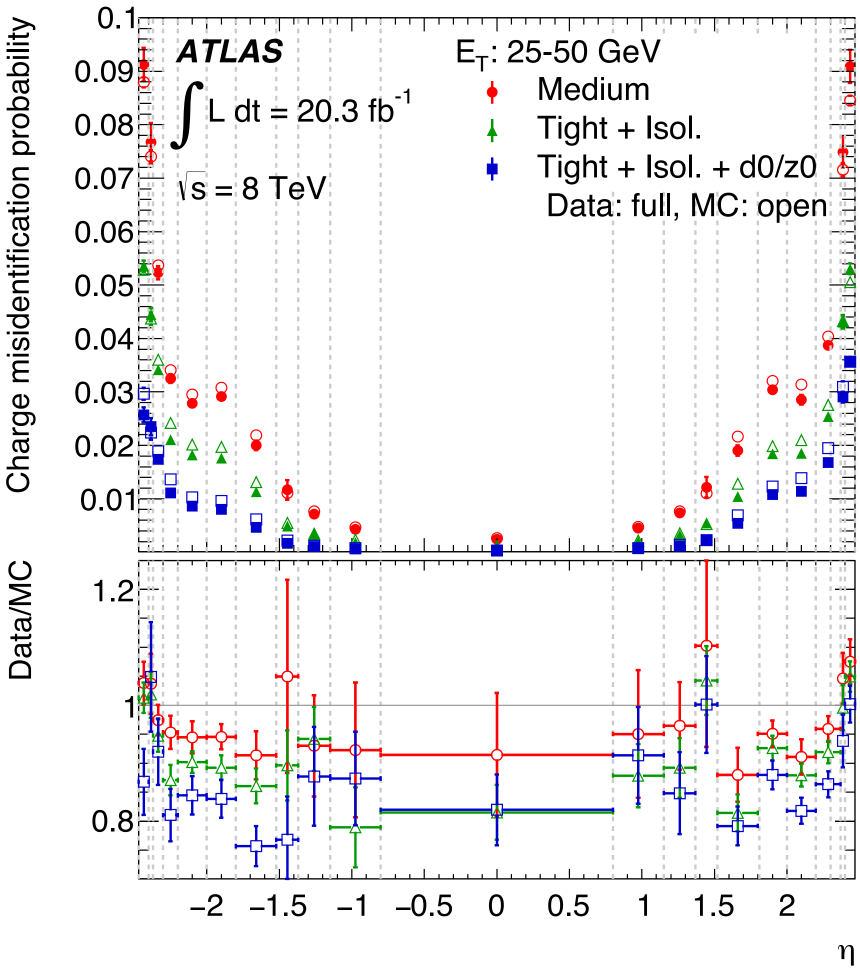

Figure 5.2: Charge mis-identification probabilities in η bins, E_T between 25 GeV and 50 GeV [37]. Three different sets of selection requirements (Medium, Tight + Isolation, and Tight + Isolation + impact parameter) are shown, along with simulation expectations. Displayed in the lower panel is the data-to-simulation ratios. The uncertainties are the total uncertainties from the sum in quadrature of statistical and systematic uncertainties. The dashed lines indicate the bins in which the rates are calculated.

5.2.4 Estimating Charge Mis-identification Background from the Charge Mis-identification Rates

In this section we give an example of how the charge mis-identification rates may be used to estimate the charge mis-identificaion background in analysis with a same-sign lepton pair signature. Suppose a sample of same-sign electron pairs has been selected in the bin pairs (i, j) (see Section 5.2). Let there be $n_{\mathrm{ss,ij}}$ of such pairs, and we wish to determine the charge misidentification contribution to this number.

To begin, we have to distinguish between the number of same-sign electron pairs $n_{\mathrm{ss,ij}}$ that has been selected and the number of genuine same-sign electron pairs. The latter is what would be counted if there were no charge misidentification. Denote it by $\bar{n}_{\mathrm{ss,ij}}$.

A charge misidentification contribution occurs whenever there is an opposite-sign pair of electrons in which one of the electron has its charge mis-identified. The probability for this to happen is, according to Equation 5.5,

$$p_{ij} = (1 - \epsilon_i)\epsilon_j + \epsilon_i(1 - \epsilon_j,)$$

where ϵ_i and ϵ_j are the charge mis-identification rates in the bins. This probability has to be multiplied by the real number of opposite-sign pairs, and not the number of opposite-sign pairs counted in the bin pair, because the latter involves contribution from same-sign pairs $\bar{n}_{\mathrm{ss,ij}}$ as well.

Denote the number of opposite-sign pairs counted in the bin pair (i, j) by $n_{\mathrm{os,ij}}$, and the corresponding real quantity by $\bar{n}_{\mathrm{os,ij}}$. The quantities available are $n_{\mathrm{ss,ij}}$, $n_{\mathrm{os,ij}}$, and the mis-identification rates ϵ_i and ϵ_j. The unknown are $\bar{n}_{\mathrm{ss,ij}}$ and $\bar{n}_{\mathrm{os,ij}}$, but they are needed to determine charge mis-identification contribution. Regarding this, the following relation holds

$$n_{\mathrm{os,ij}} = \bar{n}_{\mathrm{os,ij}} - \bar{n}_{\mathrm{os,ij}} \times p_{ij} + \bar{n}_{\mathrm{ss,ij}} \times p_{ij},$$

which says that the number of opposite-sign lepton pairs counted in the bin pair (i, j) is the corresponding real number minus the portion that is identified as same-sign plus the contribution from real same-sign pairs. This may be re-written as

$$n_{\mathrm{os,ij}} = \bar{n}_{\mathrm{os,ij}} \times (1 - p_{ij}) + \bar{n}_{\mathrm{ss,ij}} \times p_{ij}.$$

Similarly we have the following relation

$$n_{\mathrm{ss,ij}} = \bar{n}_{\mathrm{ss,ij}} \times (1 - p_{ij}) + \bar{n}_{\mathrm{os,ij}} \times p_{ij}$$

These two relations form a system of equations from which the unknown $\bar{n}_{\mathrm{os,ij}}$ and $\bar{n}_{\mathrm{ss,ij}}$ may be solved. Then the charge mis-identification contribution to $n_{\mathrm{ss,ij}}$ is simply $\bar{n}_{\mathrm{os,ij}} \times p_{ij}$.

The method just discussed is to be contrasted with the scale factor method, where scale factors that adjust the charge mis-identification rates in simulations to match the data are provided to different analyses. The former method excludes the need for the use of all systematic uncertainties that are associated with the use of simulation samples.

5.3 Conclusions

This chapter described the electron charge mis-identification problem at ATLAS and how the charge mis-identification rates are measured by fitting a Poisson likelihood function using the $Z \to e^+e^-$ data sample (data set). Three sets of charge mis-identification rates are measured and provided to ATLAS analyses, corresponding to three different sets of selection requirements (Medium, Tight + Isolation, and Tight + Isolation + impact parameter) (the range of flip rates observed). In general, simulation underestimates the charge mis-identificaion rates as compared to those in the data.

It is to be noted that in addition to measuring the charge mis-identification rates, a separate effort was started by the physics team at Université de Montréal aiming at reducing charge mis-identification. The technique relies on the output of a boosted decision tree (BTD) using a simulated sample of single electrons (see Reference [37]) (fix reference).