# Light Non-prompt (Fake) Lepton Background Estimation with the Matrix Method for $HH$ Production with a Signature of Two Same Electric Charge Light Leptons Associated with a Tau – Summer Student Report

Josefine Bjørndal Robl

*CERN Summer Student 2021*

Supervisor: André Sopczak

*IEAP CTU in Prague*

**Abstract**

The search for di-Higgs boson production $HH$ probes the tri-linear coupling $HHH$ predicted by the Higgs mechanism. In this search with two same electrically charged light leptons and one hadronically decaying tau (2lSS1tau channel), the $t\bar{t}$ background is dominant and the estimation of this background by a data-driven method is studied. For the data-driven $t\bar{t}$ background determination, the Matrix Method is used with the implementation of ATLAS Isolation and Fake Forum (IFF) tools. The IFF tools are used both for the real and fake efficiency determination of light leptons, and the Matrix Method event weight determination. For the display of signal, background and data the TRexFitter package is used. The study shows that the Matrix Method is well suited for the data-driven background determination of expected $t\bar{t}$ events, and closure tests are successfully completed. They indicate the expected systematic uncertainty of the method to be below 10%. The data-driven efficiencies are applied in the Matrix Method for a test in the control region and in the signal region.

# 1 Introduction

The trilinear coupling $HHH$ leads to non-resonant pair production of Higgs bosons. An off-shell Higgs boson decays to a pair of Higgs bosons, the leading production mechanism being gluon-fusion. Direct observation of Higgs boson pair-production $HH$ would lead to a direct measurements of the trilinear coupling $\lambda$ in the Higgs boson theory. Figure 1 shows the Feynman diagrams for the $HH$ production mechanisms.

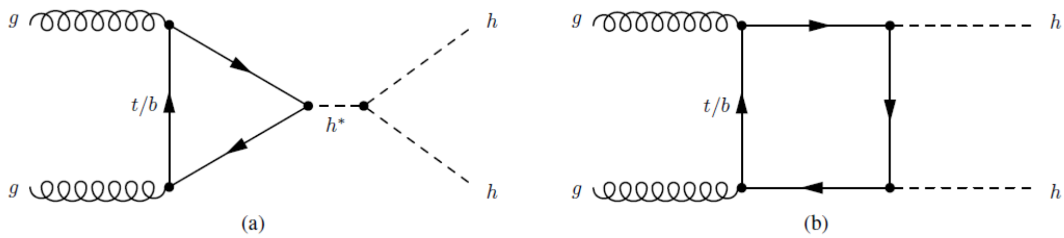


Figure 1: Representative diagrams that contribute to non-resonant $HH$ production. (a) Diagram that is sensitive to the trilinear coupling, $\lambda$. (b) Box diagram that interferes destructively with the $\lambda$-sensitive process.

As a result of the destructive interference, and the already relatively large Higgs mass of 125 GeV, the SM $HH$ production has a total cross-section of $\sim 33.4$ fb [1] at a $pp$ center-of-mass collision energy of 13 TeV. The inclusive cross-section for the pair-production of top quarks, which will be one of the dominant SM backgrounds is nearly 1000 pb, or $1 \times 10^6$ fb [2, 3]. It is therefore essential for this analysis to determine this background with high precision.

A detailed description is given of the estimation of the light non-prompt lepton background (also referred to as fakes) in the $ee$, $e\mu$ and $\mu\mu$ same-sign plus one tau-hadron channel using the Matrix Method (MM). An outline of the MM is given in Section 2.

Fake light lepton events are defined as events where at least one of the light leptons is not produced in decays of electroweak bosons but rather result from another source. These are heavy-flavor hadron decays, photon conversions, misreconstruction of particles such as hadronic jets or prompt light leptons whose charge is misidentified. Fake leptons are often also referred to as electrons or muon which are not originating from the primary Interaction Point (IP). Most likely they are resulting from semi-leptonic B-meson decays. The B-meson travels a distance before it decays, and the $b$-quark converts into a $c$-quark and a lepton and corresponding neutrino. Thus, these leptons appear to originate not from the primary IP. The dominant origin of these fake lepton events is $t\bar{t}$ production as each $t$-quark decay into a $b$-quark in this analysis with a signal region (SR) of 2 light leptons with the same electric charge and one hadronically-decaying tau lepton (2lSS1tau). Further aspects of fake lepton events are discussed in Babar Ali's thesis [4] (Section 5.4.2).

In this analysis, the signal region (SR) is defined as having exactly two light leptons with the same electric charge and one hadronically-decaying tau lepton (2lSS1tau).

The estimation of fake light leptons (electrons and muons) is very important for the analysis channel requiring two same sign light leptons. The fake rate determination cannot rely on Monte-Carlo (MC) simulation alone, therefore a data-driven method is chosen. This method is the Matrix Method [5].

The most important contribution to events with fake leptons is due to the $t\bar{t}$ production process.

The Fake Efficiency Tool is an IFF standalone package to derive lepton fake/non-prompt (and real) efficiencies for common usage [5].

The IFF tools are the recommended software for the fake rate determination for the Matrix Method (MM). “All the PA groups are strongly suggested to use these IFF fake leptons tools in their analysis. We (IFF) as well as Physics Coordination are strongly suggesting to all groups to use the FakeBkgTool tool to estimate the amount of fake lepton background in their analysis.” [5]

In order to use the Matrix Method, first an understanding of the origin and types of the fakes is needed. In particular there should be similarity between the signal and control region.

As a prerequisite for the Matrix Method, the real and fake efficiencies have to be determined. This determination is also performed with the IFF tools [5] based on a tag-and-probe method using ee, $\mu\mu$, and $e\mu$ selection.

An important step is to verify the performance of the Matrix Method and to determine the related systematic uncertainty. For this purpose, closure tests are performed where the Matrix Method estimation of the fake rates are determined from simulations and the MM results are compared to MC expectations of $t\bar{t}$ events. Ideally, the MM and MC rates should agree, as the same data sets are used for the MM and MC results. The obtained difference between the MM and MC results are interpreted as systematic uncertainties. The uncertainties are taken as the difference in ratio, or the uncertainty on the ratio, whichever value is larger.

After successful closure tests, the MM is applied to the recorded data. The $t\bar{t}$ background is determined as the MM expectations using MM weights determined from the recorded data. The MM $t\bar{t}$ background result is compared with the MC expectation.

For the technical details of the real and fake efficiency determination and the application of the Matrix Method, the IFF instructions have been followed [5].

The control region which was used for the efficiency determination was also used as test region. The results in the control and signal regions are determined without and with background subtraction for the efficiency determination.

## 2 Matrix Method

A brief description of the Matrix Method (MM) estimation of the light non-prompt lepton background (also referred to as fakes) is given.

The matrix method is a data driven technique used to estimate the contamination of fake physics objects (electrons and muons in the context of this analysis) which pass a given selection corresponding to the one used for objects in the signal region definition, referred to in the following as “tight” selection. The basic idea underlying the matrix method can be outlined first in a simplified scenario where only one lepton is taken into account. The number of events with a tight lepton (denoted with $T$) and that with a lepton which fails the tight selection (referred to as anti-tight, $\overline{T}$) can be expressed in terms of efficiencies and inefficiencies for “loose” (denoted $L$ such that $L \equiv T \cup \overline{T}$) real[1] (prompt) or fake leptons to pass the tight selection via a system of two equations:

$$\begin{aligned} N^T &= \epsilon_r N^r + \epsilon_f N^f \ , \quad \text{and} \\ N^{\overline{T}} &= \bar{\epsilon}_r N^r + \bar{\epsilon}_f N^f \ , \end{aligned} \tag{1}$$

or, in a matrix form:

$$\begin{pmatrix} N^T \\ N^{\overline{T}} \end{pmatrix} = \begin{pmatrix} \epsilon_r & \epsilon_f \\ \bar{\epsilon}_r & \bar{\epsilon}_f \end{pmatrix} \begin{pmatrix} N^r \\ N^f \end{pmatrix} , \tag{2}$$

[1] In the context of this analysis, “real” denotes prompt and isolated leptons coming from the primary interaction vertex

where $\epsilon_r$ ($\epsilon_f$) represents the efficiency for a real (fake) lepton to pass tight selection, and $\bar{\epsilon}_r \equiv (1-\epsilon_r)$ ( $\bar{\epsilon}_f \equiv (1-\epsilon_f)$) represents the probability for a real (fake) lepton to fail tight but still pass the loose selection.

Inverting the equation above allows us to access the unknown number of events with real and fake leptons in a region of interest, through observable quantities i.e the number of events with tight and anti-tight leptons and the efficiencies of passing the tight selection. The real and fake lepton efficiencies can be measured directly in dedicated control regions using data as discussed later on.

In this analysis case, the dilepton case of the matrix method formalism is employed. Depending on whether or not each lepton passes the tight selection, each i-th event populates only one of the following four orthogonal (side-bands) regions:

- $TT_i$: an event where both lepton candidates pass the tight selection. The total number of events in this region is labeled $N^{TT}$.
- $T\overline{T}_i$: an event where the leading (most energetic) lepton passes the tight selection whereas the subleading (second most energetic) lepton fails it. The total number of events in this region is labeled $N^{T\overline{T}}$
- $\overline{T}T_i$: an event where the leading lepton fails the tight selection while the subleading lepton passes it. $N^{\overline{T}T}$ denotes the total number of events in this region.
- $\overline{TT}_i$: an event where both lepton candidates fail the tight selection. $N^{\overline{TT}}$ denotes the total number of events in this region.

As for the "single lepton" simplified scenario, one can define an efficiency matrix of dimension 4 for the dilepton case, mapping the observed total number of events above to that in four orthogonal regions with different real and fake lepton composition, as follows:

- $rr_i$: corresponds to an event where both leptons are real (prompt). $N^{rr}$ denotes the total number of events in this region.
- $rf_i$: an event where the leading lepton is real while the subleading lepton is fake. The total number of events in this region is labeled $N^{rf}$.
- $fr_i$: an event where the leading lepton is fake and the subleading lepton is real. $N^{fr}$ corresponds to the total number of events in this region.
- $ff_i$: an event where both leptons are fake. The total number of such events is labeled $N^{ff}$

The corresponding 4×4 matrix equation can be written as:

$$\begin{pmatrix} N^{TT} \\ N^{T\bar{T}} \\ N^{\bar{T}T} \\ N^{\bar{T}\bar{T}} \end{pmatrix} = \begin{pmatrix} \epsilon_{r,1}\epsilon_{r,2} & \epsilon_{r,1}\epsilon_{f,2} & \epsilon_{f,1}\epsilon_{r,2} & \epsilon_{f,1}\epsilon_{f,2} \\ \epsilon_{r,1}\bar{\epsilon}_{r,2} & \epsilon_{r,1}\bar{\epsilon}_{f,2} & \epsilon_{f,1}\bar{\epsilon}_{r,2} & \epsilon_{f,1}\bar{\epsilon}_{f,2} \\ \bar{\epsilon}_{r,1}\epsilon_{r,2} & \bar{\epsilon}_{r,1}\epsilon_{f,2} & \bar{\epsilon}_{f,1}\epsilon_{r,2} & \bar{\epsilon}_{f,1}\epsilon_{f,2} \\ \bar{\epsilon}_{r,1}\bar{\epsilon}_{r,2} & \bar{\epsilon}_{r,1}\bar{\epsilon}_{f,2} & \bar{\epsilon}_{f,1}\bar{\epsilon}_{r,2} & \bar{\epsilon}_{f,1}\bar{\epsilon}_{f,2} \end{pmatrix} \begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix}, \tag{3}$$

where the indices of $\epsilon_r$ and $\epsilon_f$ correspond to the $p_T$ order of the lepton.

The number of fake events in the signal region can be accessed through observable quantities, by inverting the above equation:

$$\begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix} = \begin{pmatrix} \epsilon_{r,1}\epsilon_{r,2} & \epsilon_{r,1}\epsilon_{f,2} & \epsilon_{f,1}\epsilon_{r,2} & \epsilon_{f,1}\epsilon_{f,2} \\ \epsilon_{r,1}\bar{\epsilon}_{r,2} & \epsilon_{r,1}\bar{\epsilon}_{f,2} & \epsilon_{f,1}\bar{\epsilon}_{r,2} & \epsilon_{f,1}\bar{\epsilon}_{f,2} \\ \bar{\epsilon}_{r,1}\epsilon_{r,2} & \bar{\epsilon}_{r,1}\epsilon_{f,2} & \bar{\epsilon}_{f,1}\epsilon_{r,2} & \bar{\epsilon}_{f,1}\epsilon_{f,2} \\ \bar{\epsilon}_{r,1}\bar{\epsilon}_{r,2} & \bar{\epsilon}_{r,1}\bar{\epsilon}_{f,2} & \bar{\epsilon}_{f,1}\bar{\epsilon}_{r,2} & \bar{\epsilon}_{f,1}\bar{\epsilon}_{f,2} \end{pmatrix}^{-1} \begin{pmatrix} N^{TT} \\ N^{T\bar{T}} \\ N^{\bar{T}T} \\ N^{\bar{T}\bar{T}} \end{pmatrix} . \tag{4}$$

The total number of fake events (where at least one of the leptons is fake) in the signal region i.e where both leptons pass the tight selection, $N^f_{TT}$, can then be derived as follows:

$$N^f_{TT} = N^{rf}_{TT} + N^{fr}_{TT} + N^{ff}_{TT} = \epsilon_{r,1}\epsilon_{f,2}N^{rf} + \epsilon_{r,2}\epsilon_{f,1}N^{fr} + \epsilon_{f,1}\epsilon_{f,2}N^{ff} . \tag{5}$$

When combining the matrix equation (4) with the above formula and making the weighted sum over the events explicit, we finally obtain:

$$\begin{aligned} N^f_{TT} &= \sum_i (\epsilon_{r,1}\epsilon_{f,2}rf)_i + (\epsilon_{r,2}\epsilon_{f,1}fr)_i + (\epsilon_{f,1}\epsilon_{f,2}ff)_i \\ &= \sum_i^{\{TT\}} (w^{MM}_{TT}TT)_i + \sum_i^{\{T\overline{T}\}} (w^{MM}_{T\overline{T}}T\overline{T})_i + \sum_i^{\{\overline{T}T\}} (w^{MM}_{\overline{T}T}\overline{T}T)_i + \sum_i^{\{\overline{TT}\}} (w^{MM}_{\overline{TT}}\overline{TT})_i . \end{aligned} \tag{6}$$

Where the matrix method weights, $w^{MM}$, are defined as follows:

$$\begin{aligned} w^{MM}_{TT}{}_i &= (1 - \beta\epsilon_{r,1}\epsilon_{r,2}\bar{\epsilon}_{f,1}\bar{\epsilon}_{f,2})_i , \\ w^{MM}_{T\overline{T}}{}_i &= (\beta\epsilon_{r,1}\epsilon_{r,2}\epsilon_{f,2}\bar{\epsilon}_{f,1})_i , \\ w^{MM}_{\overline{T}T}{}_i &= (\beta\epsilon_{r,1}\epsilon_{r,2}\epsilon_{f,1}\bar{\epsilon}_{f,2})_i , \\ w^{MM}_{\overline{TT}}{}_i &= -(\beta\epsilon_{r,1}\epsilon_{r,2}\epsilon_{f,1}\epsilon_{f,2})_i , \quad \text{and} \\ \beta &= \frac{1}{(\epsilon_{r,1_i} - \epsilon_{f,1_i})(\epsilon_{r,2_i} - \epsilon_{f,2_i})} . \end{aligned} \tag{7}$$

Evidently, each i-th event will contribute to only one of the four sums on the right side of equation (6).

### 2.1 Real and fake lepton efficiencies measurement

The measurement of the efficiency for real and fake leptons to pass the tight selection, which are key ingredients for the MM formalism, is performed in dedicated control regions referred to as CR. This control region is designed to be representative of the signal region in terms of kinematics and background composition while being completely orthogonal to the SR and retaining sufficiently large statistics.

In order to provide a fake prediction dependent on the lepton kinematics, the efficiencies have been parametrized in bins of $p_T$.

## 3 Real and Fake Efficiencies

For the efficiency generation, bin ranges of transverse momenta {7, 12, 20, 35, 50, 1000} GeV are used in order to have sufficient statistics in each bin.

The real efficiency is defined as the ratio of the number of events with real leptons passing the tight selection over the number of events with real leptons passing the loose selection. The fake efficiency is defined as the ratio of the number of events with fake leptons passing the tight selection over the number of events with fake leptons passing the loose selection.

For the recorded data, real and fake are not defined by the truth information, thus the following approach is used. For the real efficiency, a high-statistics $Z \to \ell\ell$ sample is used and the real efficiency is determined by the tag-and-probe method. One lepton is required to pass the tight selection, and the second lepton passing the loose selection is used as probe. The efficiency is defined as the ratio of the probe lepton passing the tight selection to all events in the $Z \to \ell\ell$ sample with one tagged lepton.

For the fake efficiency determination, the CR sample is used. In this case the tagged lepton is the leading lepton fulfilling the tight selection and the probe lepton is the subleading lepton. The efficiency is defined as the ratio of the probe lepton passing the tight selection to all events in the CR with one tagged lepton.

# 4 Results

Technical details of the analysis are given the internal ATLAS note [6]. The main results of this study are:

- The IFF tools recommended by the ATLAS physics analysis coordination are used to determine the fake rate from $t\bar{t}$ production in the 2lSS1tau channel.
- The type and origin of true fakes are determined from simulated events and the $t\bar{t}$ rates are dominant for the 2lSS1tau analysis.
- Good similarity of type and origin are observed in the SR and CR.
- The real electron and muon efficiencies are determined with a tag-and-probe method using the IFF tools for mcA (2015, 2016), mcD (2017) and mcE (2018) detector simulations.
- The fake electron and muon efficiencies are determined with a tag-and-probe method using the IFF tools, where the efficiency is defined as ratio between Loose and Tight objects.
- Using the real and fake efficiencies, the Matrix Method is applied. As input standard group analysis ntuples are used and these ntuples have been extended by a variable of MM weights for each event.
- These extended ntuples then serve as input for the TRexFitter tool to illustrate the event distributions in a standard format.
- Closure tests were performed for di-electron, di-muon, and electron-muon selections of events with same electric charge. The closure tests completed with an agreement within 1 standard deviation for the di-electon selection, within 3-4 standard deviations for the di-muon selection, and within 1-2 standard deviations for the electron-muon selection, both for the SR and CR.
- For the combined closure test of three di-lepton channels, the agreement between MC and MM predictions is within standard deviation, and the uncertainty is 8%.
- The real and fake efficiencies were also determined for the recorded CR data. Owing to the low statistics in each year of data-taking for the fake efficiency determination, the data of each year were combined, and thus no differences in the efficiencies for individual years were assumed. For the real efficiency determination, high statistics were available and the data recorded in 2016 were used.

- The Matrix Method was applied in the CR as validation, using the efficiencies determined in the CR.
- The Matrix Method was applied in the SR, using the efficiencies determined in the CR. Thus, the MC $t\bar{t}$ background expectation is replaced by the data-driven estimate from the Matrix Method.
- Some excess of MM events plus simulated background events compared to the data are noted in the validation.
- The comparison between data and prediction was improved by applying a background subtraction of $ttH$, $ttW$ and $ttZ$ events for the data-driven efficiency determinations.
- Using the real and fake efficiencies determined from the CR, the MM is applied in the SR and the simulated $t\bar{t}$ events are replaced by the MM expectation.

# 5 Conclusion

In this summer student project, a contribution to a challenging analysis of di-Higgs boson production was made. The specific analysis channel contains two same-sign electrically charged light leptons and one hadronically decaying tau. Two selections of events were used, one Signal Region with the requirement of no $b$-tagged jet, and one Control Region with the requirement of at least one $b$-tagged jet. For both signal and control regions, the type and origin of expected real and fake light leptons was studied and found to be similar. A Matrix Method was used to determine the number of expected events with fake leptons for different cases. The analysis was performed for a di-electron, di-muon, electron-muon and a combined di-lepton selection. In order to test the Matrix Method fake rate determination, a closure tests were performed for the leading $t\bar{t}$ background. The method proved to meet the expectations, and an uncertainty of 8% was determined by taking the maximum between the deviation of the Matrix Method result and the simulation expectation, and the statistical uncertainty on the comparison. A test in the CR was performed using the real and fake efficiencies determined from the control region data. Using the same efficiencies, the Matrix Method was applied on recorded signal region data for a data-driven determination of the $t\bar{t}$ background expectation. Background subtraction in the data-driven efficiency determination improved the agreement in the control region, and the agreement between data and simulation in the signal region. As a follow-up, a validation region orthogonal to the control region used for the efficiency determination will be defined.

## Acknowledgements

The author would like to thank CERN for making this summer student project possible, André Sopczak for supervising the project, Arthur Chomont for fruitful discussions and troubleshooting of the tools used during the project, Bartlomiej Zabinski for providing an analysis ntuple production, and the $HH$ analysis team for feedback.

## References


1. *Higgs cross sections for HL-LHC and HE-LHC* https://twiki.cern.ch/twiki/bin/view/LHCPhysics/HiggsEuropeanStrategy#SM_Higgs_production_cross_se_AN2 (2021).

2. ATLAS Collaboration. Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with $b$-tagged jets in $pp$ collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector. *Phys. Lett. B* **761,** 136. arXiv: 1606.02699 [hep-ex] (2016). Erratum: *Phys. Lett. B* **772,** 879 (2017).

3. ATLAS Collaboration. *Measurements of the $t\bar{t}$ production cross-section in the dilepton and lepton-plus-jets channels and of the ratio of the $t\bar{t}$ and $Z$ boson cross-sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector* ATLAS-CONF-2015-049. 2015. https://cds.cern.ch/record/2052605.

4. Ali, B. *Search for the Standard Model Higgs boson produced in association with a top quark pair in multilepton final states at $\sqrt{s}$ = 13 TeV with the ATLAS detector* Presented 28 Feb 2020 (Dec. 2019). https://cds.cern.ch/record/2774418.

5. *FakeLeptonBackgrounds* https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/FakeLeptonBackgrounds (2021).

6. Robl, J. B. & Sopczak, A. *Light Non-prompt (Fake) Lepton Background Estimation with the Matrix Method for HH Production with a Signature of Two Same Electric Charge Light Leptons Associated with a Tau - Summer Student Project* tech. rep. ATL-COM-PHYS-2021-974 (CERN, Geneva, Nov. 2021). https://cds.cern.ch/record/2789839.