"Shape-based" Scale Factors for Electrons

Martín Alcalde Martínez

September 2022

1 Introduction

Whenever we compare the data obtained from a real physics experiment with a simulation, we may see that the simulation does not match the data perfectly. Simulations are not a perfect replica of reality, they are the best approximation we have to understanding the underlying processes.

In this project, we will compare data and simulated events from the CMS experiment at CERN, in particular, from the year 2018 (included in the Run-2).

In high energy physics, a large amounts of collision data are recorded in order to have high enough statistics to allow us to carry out physical analyses. To filter the data produced in order to select data of interest, a series of selection criteria are applied to different values and magnitudes derived from the events. When the simulation is compared with the resulting data, we may see disagreements between them and in order to reconcile data and simulation a series of weights called *scale factors* (SFs) are calculated as the ratio between the efficiency of the data and the efficiency of the simulation.

For the purpose of extracting the values of the scale factors, we will use a framework known as *spark_tnp*. So far, this framework has only been used for muons; in this project, we will focus on the viability of the application of this framework for the extraction of scale factors for electrons produced in decays of Z-boson using data recorded at the CMS Experiment in the year 2018.

2 Efficiency. The *Tag And Probe* method

As previously stated, the scale factors are calculated using a rather simple formula:

$$\mathrm{SF} = \frac{\varepsilon_{data}}{\varepsilon_{MC}} \tag{2.1}$$

Nevertheless, extracting the values is no trivial matter. In order to calculate the value of the scale factors, one first needs to know the value of the efficiencies of the reconstructed leptons. In order to determine the efficiency, a method known as *Tag And Probe* is applied.

The first step of this method consists in applying a very restrictive (tight) cut to the set of events. The electrons that manage to pass this filter are then categorized as **tags**. This step allows us to make sure that one lepton of the pair is going to be a real lepton. This way, we can make sure that the background is suppressed and ensure that the obtained event sample largely consists of actual electron pairs.

To find the other lepton for the pair, a less restrictive criterion is applied to the original set of events. The passing events are then known as **probes**.

In order to reconstruct the initial state, we pair tags and probes (passing probes and failing probes, separately) so that we have a set of passing pairs and a set of failing pairs. With each of these sets the resonance is reconstructed and the yields are determined from a fit. This fit is used to determine the number of events with real electrons coming from the resonance, while the non-resonant background could still contain candidates that are wrongly identified using the *Tag Probe* method and would come from other particles.

Finally, to calculate the efficiency we just need to apply this next formula:

$$\varepsilon = \frac{P_{pass}}{P_{pass} + P_{fail}} \tag{2.2}$$

where P_{pass} and P_{fail} are the yields for the passing and failing pairs sets, respectively. Using this method, we can extract the values of the scale factors applying the formula (2.1).

3 Scale Factors

There are two types of methods to identify leptons: cut-based or MVA-based; in this project, we will focus on the latter, developing scale factors that aim to correct the full shape of the MVA output. The MVA (**M**ulti**V**ariate **A**nalysis) is a discriminant (with a value range from -1 to 1) that is used to set regions or *working points* (WPs) for which the scale factors are extracted. It is the output of a neural network used to identify whether a lepton is a prompt or non-prompt/fake lepton; the more confident the neural network is that the input is a real lepton, the closer the MVA value is to one.

The scale factors are extracted in kinematic regions (or bins) of the transverse momentum and pseudorapidity. In this case, we aim to produce "shape-based" SFs, this means that the values extracted for the WPs will be interpolated to obtain a function that given the value fo p_T, η and MVA will output the value of the corresponding SF. These continuous SFs allow us to use the MVA values for physical analysis; we can only use the values once they have been corrected using the SFs.

4 The framework: *spark_tnp*

This is a framework that is currently being used for extracting the values of the SFs for muons using the Tag And Probe method, it has so far been used for muons only.

The framework uses very few commands. In the scope of this project, the use of this framework can be summed up in four steps:

1. **Flatten:** This is the first step. It will allow us to convert the histograms saved in the parquet files to a format that can be used in the next step.

2. **Fit (I):** In this step, the framework takes the flattened histograms and proceeds to fit the passing and failing pairs. This step is run in a computing cluster.

3. **Fit (II):** Once all the jobs sent to the cluster in the previous step have finished, we run a command to recover the output and save it locally.

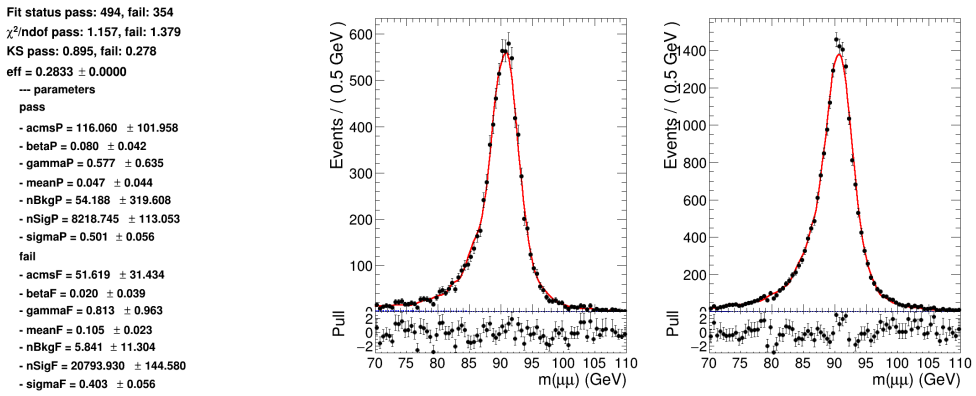
4. **Prepare:** In this final step, the framework produces a series of graphs plotting the efficiencies, the scale factors and their uncertainties.

We need to run these four steps for every working point, but the bins for the kinematic regions are set in a configuration file so the SFs are calculated for all of them in the given WP. The higher the number of kinematic regions, the more time the framework will take to finish the process.

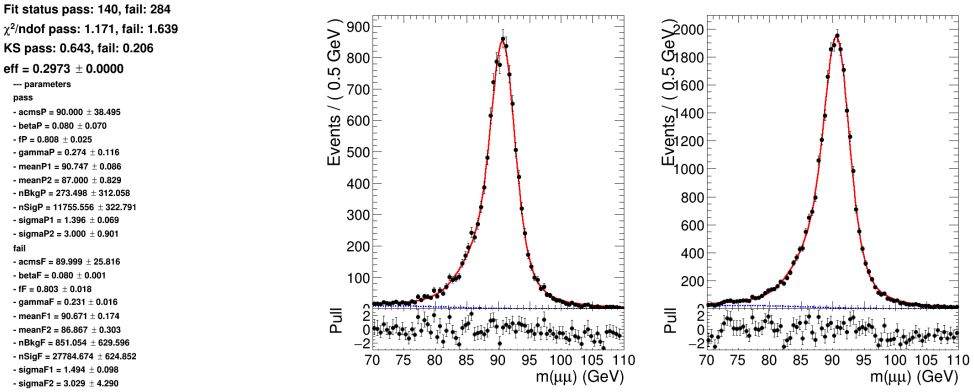
5 Results

5.1 Scale Factors for electrons

The *spark_tnp* framework has proved to work excellently with electrons, being able to produce the SFs without problems. In Figure 1a we show how the framework performs fitting the mass distributions of passing and failing pairs, for data. In Figure 1b we do the same for *Monte Carlo* (MC):



(a) *Mass distribution fits for passing and failing pairs of data events.*



(b) *Mass distribution fits for passing and failing pairs of simulated events.*

Figure 1: *Output from the the spark_tnp fitting process. On the left, the different parameters for the fits. On the right, two plots showing the reconstructed mass for passing probes (left) and failing probes (right).*

This comparison has been done in a high statistics region ($15\ \text{GeV} \leq p_T < 45\ \text{GeV}, 0 \leq \eta < 0.5$). As shown in these plots, the framework performs exceptionally both for data and simulated events.

The figures containing the efficiency values for data and MC for every working point and kinematic region combination can be seen in the Appendix A Electron Efficiencies.

The figures containing the scale factor values and their uncertainties for every working point and kinematic

region combination can be seen in the Appendix B Electron SFs.

The **working points** employed are:

1. $0.70 \leq MVA < 0.85$
2. $0.85 \leq MVA < 0.95$
3. $0.95 \leq MVA$

The kinematic regions for the SFs used in these results are:

- **Transverse momentum:**
  1. $15\text{GeV} \leq p_T < 45\text{GeV}$
  2. $45\text{GeV} \leq p_T < 120\text{GeV}$
- **Pseudorapidity:**
  1. $0.0 \leq \eta < 0.5$
  2. $0.5 \leq \eta < 1.0$
  3. $1.0 \leq \eta < 1.5$
  4. $1.5 \leq \eta < 2.0$
  5. $2.0 \leq \eta < 2.5$

5.2 Electron MVA distribution

In addition to the previous result, it was found that the distribution of the MVA value for electrons is quite different for that of muons. For muons, the majority of the events with positive MVA were concentrated in the region with $MVA \geq 0.85$. However, for electrons, the distribution is broader, extending to lower values of the MVA discriminant as shown in Figure 2.

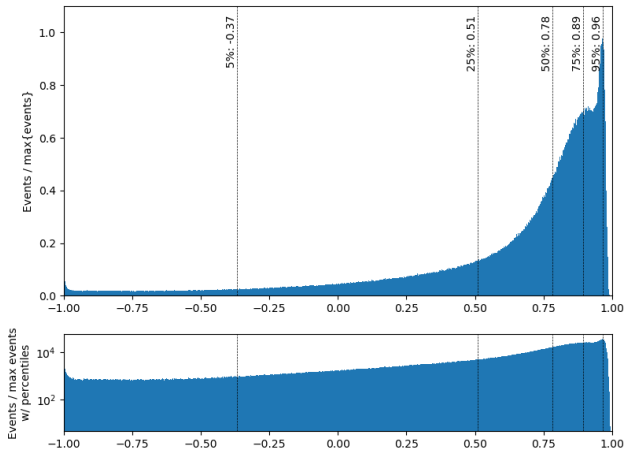


Figure 2: *The x axis in this graph represents the value of the MVA discriminant; in the y axis, the number of events. The different quantiles are marked by dashed lines. The lower (upper) panel shows the same distribution in logarithmic (linear) scale.*

Due to this change of shape, the working points used for muons are not valid for electrons.

6 Conclusions

Finally, we can conclude that the *spark_tnp* framework can be used successfully to extract the scale factors for electrons. Additionally, we have seen that the shape of the MVA distribution in electrons is different than that of muons, leading to a new set of working points.

Next, a comparison among different binnings and working points should be made to determine what are the optimal parametres. Also an interpolation with the extracted scale factors should be done in order to have continuous ("shape-based") scale factors. These scale factors should then be evaluated using the full Run-2 data in order to test their accuracy.

Moreover, being able to have scale factors as continuous functions after the interpolation will allow us to use the information contained in the MVA shape output in physical analyses for the first time.

A Electron Efficiencies

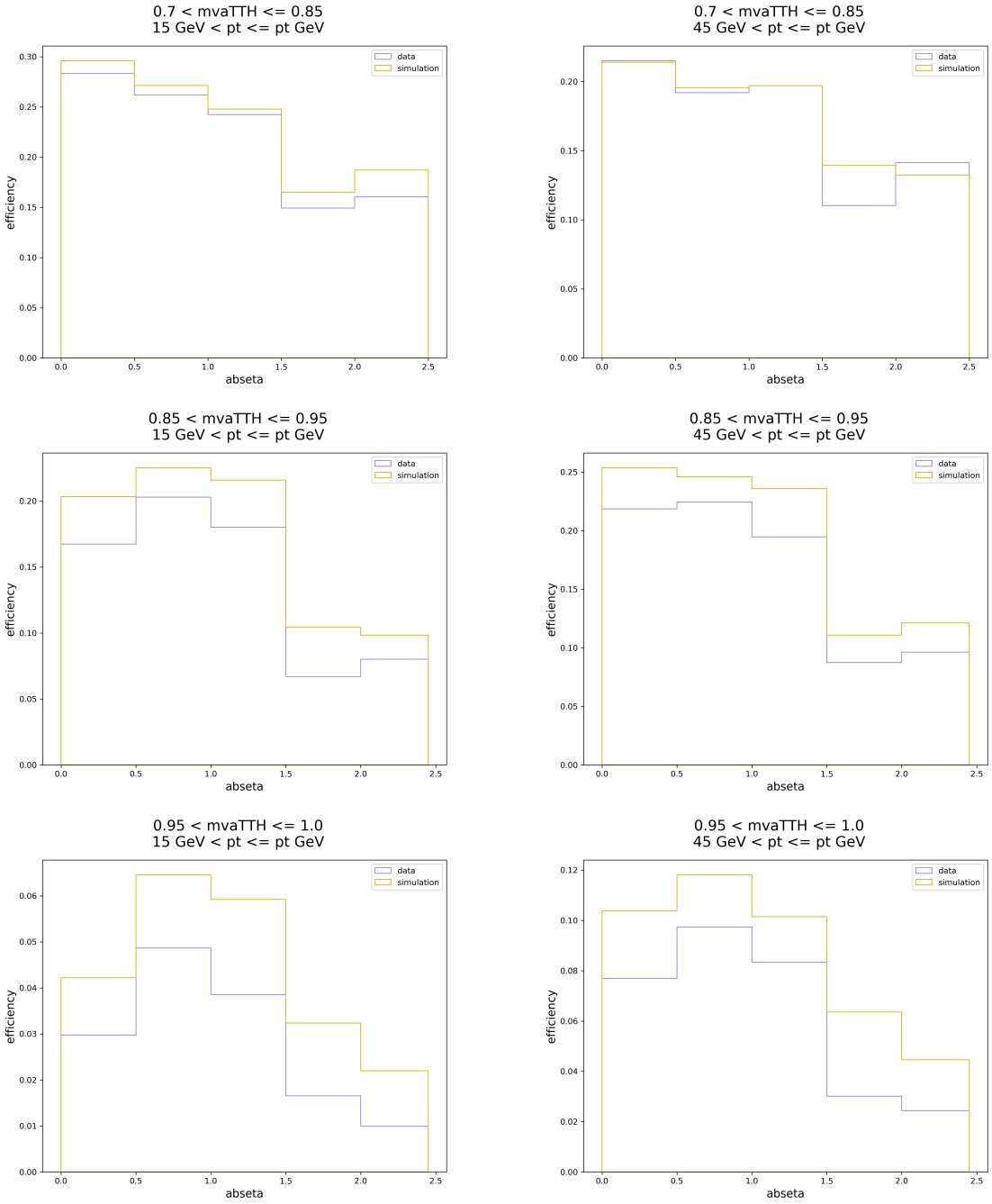


Figure 3: *Efficiencies for electrons in the six WP/transverse momentum regions. In each of the subfigures, the values of the data efficiency and MC efficiency are plotted for each of the pseudorapidity bins.*

B Electron SFs

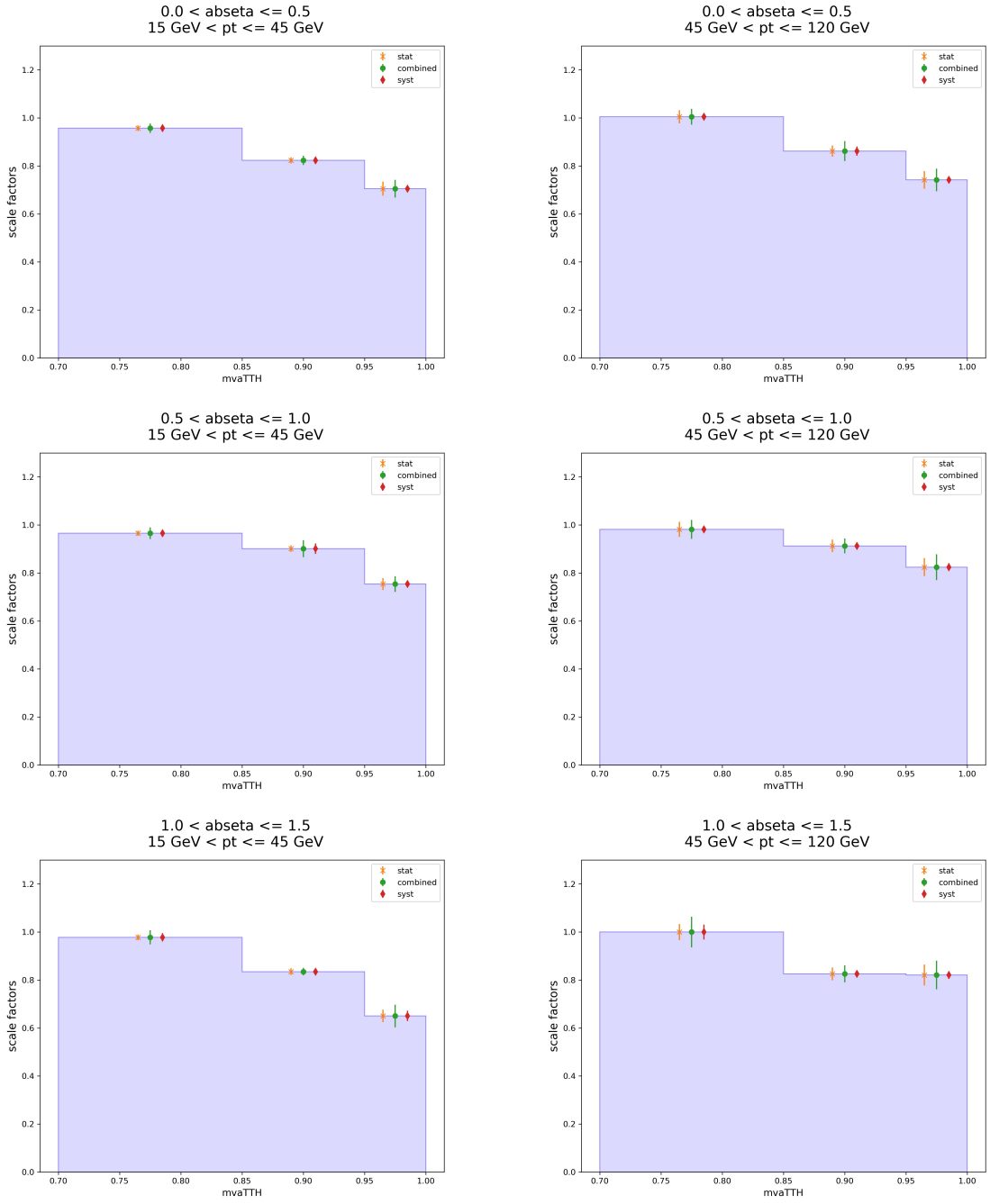


Figure 4: *SFs for electron in the ten different kinematic regions. In each of the subfigures, the values of the SFs and their uncertainties are plotted for each of the WPs.*

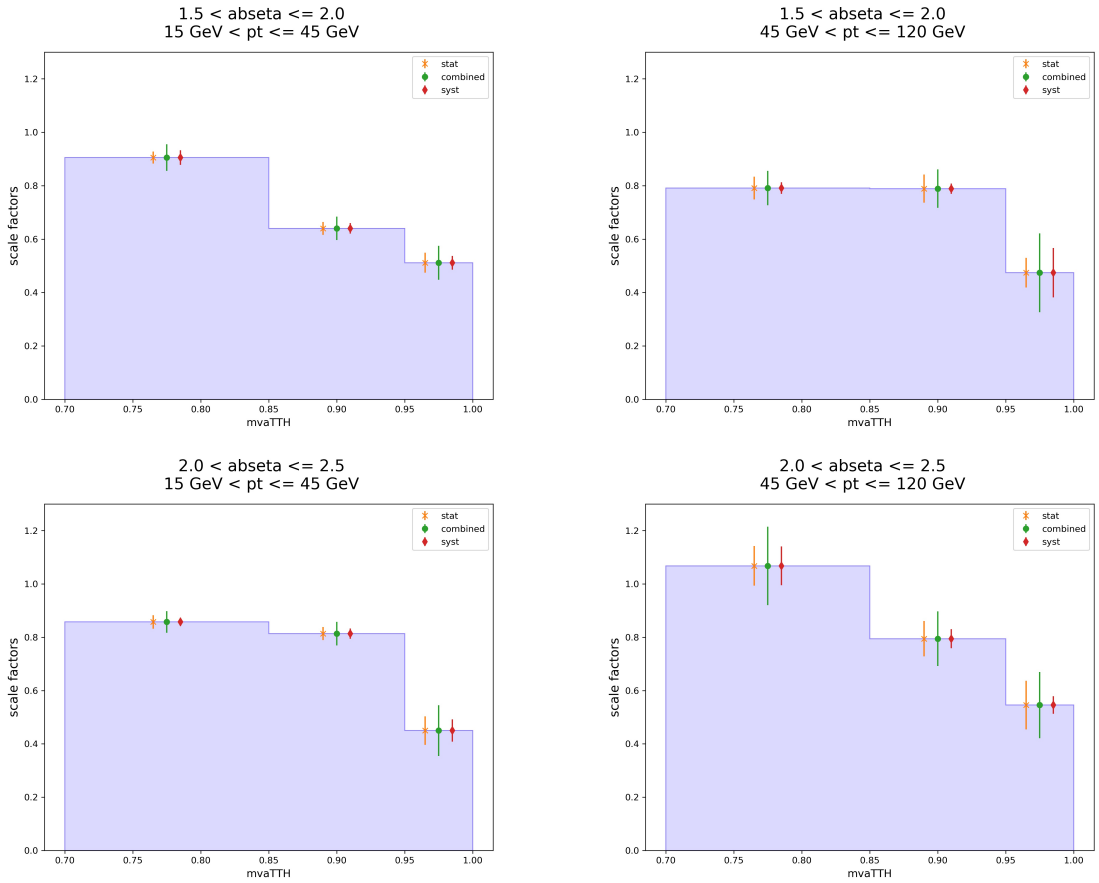


Figure 4: *SFs for electrons in the ten different kinematic regions. In each of the subfigures, the values of the SFs and their uncertainties are plotted for each of the WPs.*