ZAi Notes

Lukas Bayer

November 17, 2023

1 Fake-Factors

1.1 Nomenclature

We use the following samples to estimate the amount of fake electrons in our data:

data is a sample of measured data from the ATLAS experiment. It includes real as well as fake electrons.

MJ is the subset of the data sample that consists only of fake electrons from multi-jet events. A priori, its size and distribution are unknown.

prompt is a sample of Monte-Carlo generated events that consists only of simulated real electrons. It is normalized to the luminosity, such that, for a perfect simulation, it would include the same number of real electrons as the data sample. In reality, they differ by a factor of $\hat{\mu}$, which is to be determined.

$$N_{\mathrm{MJ}} = N_{\mathrm{data}} - \hat{\mu} \cdot N_{\mathrm{prompt}} \tag{1}$$

Additionally we define the identification regions:

ID are events passing tight identification criteria

nL are events failing loose identification criteria[1]

and invariant mass regions:

on are events with an invariant mass close to the Z pole $m_{\ell\ell} \approx m_Z$

off are events with an invariant mass far away from the Z pole $m_{\ell\ell} \gg m_Z$

thus dividing each sample into a total of four regions. $\mathsf{SR} = \mathsf{IDon}$ is the signal region of interest, while the remaining three regions nLon, nLoff and IDoff are enriched in multi-jet experiments and used as control regions CR.

1.2 Calculation of $\hat{\mu}$

Assume that the ratio of fake electrons passing to failing identification is the same on and off the Z pole. Then the amount of fake electrons in the signal region is given by

$$N_{\mathrm{MJ}}^{\mathrm{IDon}} = N_{\mathrm{MJ}}^{\mathrm{nLon}} \cdot \frac{N_{\mathrm{MJ}}^{\mathrm{IDoff}}}{N_{\mathrm{MJ}}^{\mathrm{nLoff}}} \tag{2}$$

Rearranging equation 2 and inserting equation 1 leads to:

$$\begin{aligned} 0 &= N_{\mathrm{MJ}}^{\mathrm{nLon}} \cdot N_{\mathrm{MJ}}^{\mathrm{IDoff}} - N_{\mathrm{MJ}}^{\mathrm{IDon}} \cdot N_{\mathrm{MJ}}^{\mathrm{nLoff}} \\ &= \left(N_{\mathrm{data}}^{\mathrm{nLon}} - \hat{\mu} N_{\mathrm{prompt}}^{\mathrm{nLon}}\right) \cdot \left(N_{\mathrm{data}}^{\mathrm{IDoff}} - \hat{\mu} N_{\mathrm{prompt}}^{\mathrm{IDoff}}\right) - \left(N_{\mathrm{data}}^{\mathrm{IDon}} - \hat{\mu} N_{\mathrm{prompt}}^{\mathrm{IDon}}\right) \cdot \left(N_{\mathrm{data}}^{\mathrm{nLoff}} - \hat{\mu} N_{\mathrm{prompt}}^{\mathrm{nLoff}}\right) \\ &= A\hat{\mu}^2 + B\hat{\mu} + C \end{aligned}$$

[1] the identification criteria employed here are not necessarily the same as for the scale-factor calculation

with

$$
\begin{aligned}
A &= N^{\text{nLon}}_{\text{prompt}} \cdot N^{\text{IDoff}}_{\text{prompt}} - N^{\text{IDon}}_{\text{prompt}} \cdot N^{\text{nLoff}}_{\text{prompt}} \\
B &= N^{\text{IDon}}_{\text{data}} \cdot N^{\text{nLoff}}_{\text{prompt}} - N^{\text{nLon}}_{\text{data}} \cdot N^{\text{IDoff}}_{\text{prompt}} + N^{\text{IDon}}_{\text{prompt}} \cdot N^{\text{nLoff}}_{\text{data}} - N^{\text{nLon}}_{\text{prompt}} \cdot N^{\text{IDoff}}_{\text{data}} \\
C &= N^{\text{nLon}}_{\text{data}} \cdot N^{\text{IDoff}}_{\text{data}} - N^{\text{IDon}}_{\text{data}} \cdot N^{\text{nLoff}}_{\text{data}}
\end{aligned}
$$

Solving for (real, positive values of) $\hat{\mu}$ gives[2][3]:

$$\hat{\mu} = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{3}$$

1.3 Estimation of Multi-Jet Content

After calculating $\hat{\mu}$ with equation 3, it can be used to determine estimates for the multi-jet contents of the control regions using equation 1. Finally, these are inserted back into equation 2 to estimate the multi-jet content of the signal region.

The main contribution to the statistical uncertainty is expected to come from the data counts (not from $\hat{\mu}$ or prompt counts). Hence, for the control regions:

$$\Delta N^{\text{CR}}_{\text{MJ}} \approx \sqrt{N^{\text{CR}}_{\text{data}}}$$

and for the signal region:

$$
\begin{aligned}
\Delta N^{\text{IDon}}_{\text{MJ}} &= \sqrt{\sum_{\text{CR}} \left(\Delta N^{\text{CR}}_{\text{MJ}} \cdot \frac{\partial N^{\text{IDon}}_{\text{MJ}}}{\partial N^{\text{CR}}_{\text{MJ}}} \right)^2} \\
&\approx \sqrt{N^{\text{nLon}}_{\text{data}} \cdot \left(\frac{N^{\text{IDoff}}_{\text{MJ}}}{N^{\text{nLoff}}_{\text{MJ}}} \right)^2 + N^{\text{IDoff}}_{\text{data}} \cdot \left(\frac{N^{\text{nLon}}_{\text{MJ}}}{N^{\text{nLoff}}_{\text{MJ}}} \right)^2 + N^{\text{nLoff}}_{\text{data}} \cdot \left(\frac{N^{\text{nLon}}_{\text{MJ}} N^{\text{IDoff}}_{\text{MJ}}}{(N^{\text{nLoff}}_{\text{MJ}})^2} \right)^2}
\end{aligned}
$$

1.4 Calculation of Fake-Factors

Let's define the fake efficiency as the probability of a fake electron passing identification criteria:

$$\epsilon_f := \frac{N^{\text{ID}}_{\text{MJ}}}{N_{\text{MJ}}}$$

According to [1] the fake-factor is defined as:

$$F := \frac{\epsilon_f}{1 - \epsilon_f} = \frac{N^{\text{ID}}_{\text{MJ}}}{N^{\text{nID}}_{\text{MJ}}}$$

But in our implementation it is instead calculated as:

$$F = \frac{N^{\text{IDoff}}_{\text{MJ}}}{N^{\text{nLoff}}_{\text{MJ}}}$$

Why is this different?

---

[2] In the highly improbable case that $A = 0$ the result becomes $\hat{\mu} = -C/A$ instead.

[3] Is this sqrt really always positive?

1.5 applyFakeFactor

So far I have described how fake-factors are calculated when the makeFakeFactor flag is set. However, I don't understand, what is being done when applyFakeFactor is set. Can anyone explain?

References

[1] ATLAS Collaboration. Tools for estimating fake/non-prompt lepton backgrounds with the atlas detector at the lhc, 2022.