Accurate Modeling of Electron Tagging Efficiency

Mentee: Haile Ayalneh (Wabash college)
Mentor: Nick Smith(Fermilab)

Abstract

Abstract: The Tag&Probe (T&P) method is an experimental procedure used commonly in particle physics that allows to measure process efficiencies directly from data, therefore not relying on the accuracy of simulations. An accurate model of true electron efficiency could be derived using a binned Tag and Probe approach. However, this method has its own flaws, namely: the curse of dimensionality makes it challenging to estimate the efficiency in a high-dimensional binning; each bin in the probe kinematics needs to separately be validated; and the binned efficiency estimates do not capture the expected smooth structure of the true efficiency. We propose to develop an unbinned efficiency measurement, and compare its performance with the classic binned T&P approach on CMS Open Data. The method relies upon $Z \rightarrow$ di-electron decays to provide an unbiased, high-purity, electron sample with which to measure the efficiency of a particular selection or trigger.

Overview

1. Introduction
2. Tag and Probe
3. Binned and unbinned Tag & Probe
4. Method

Introduction - Theory

- The Tag&Probe (T&P) method is an experimental procedure used commonly in particle physics that allows to measure process efficiencies directly from data, therefore not relying on the accuracy of simulations.

- Tag electron = well identified, triggered electron (tight selection criteria)

- Probe electron = unbiased set of electrons with a very loose selection criteria.

- This method uses resonances (e.g J/ψ,Y,Z) to confirm the probe electron is true electron.

Method

- Implement three efficiency measurement techniques:
  - Binned (in pT) T&P
  - Binned (in pT)Cut&Count
  - Unbinned (in pT) Cut&Count

- Compare performance of the three methods using CMS 2015 open simulation (Z->ee)

2015 open data NanoAod root file

Corresponding 2015 open simulation

Tag & probe

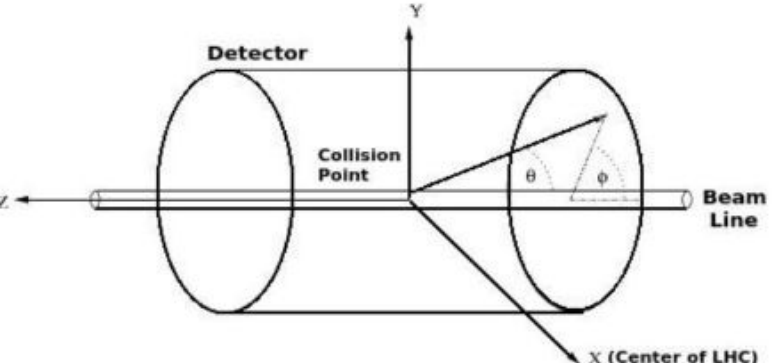
Efficiency (ε)

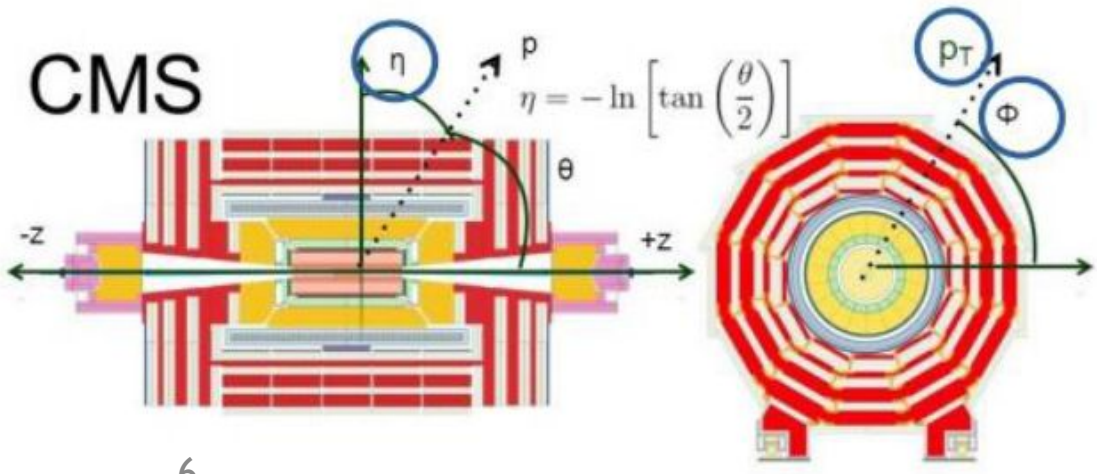
As a function of

- Transverse momentum (p_T)

Method

- Produce opendata NanoAOD using workflow from 2022 PURSUE program
- Use coffea analysis tools & scipy to perform binned T&P fits





Binned Tag and Probe

- Binned T&P efficiency is a method where the data is divided into bins based on some observable (e.g., transverse momentum), and the efficiency is estimated for each bin separately.

- It provides efficiency estimates for discrete intervals, which can be useful for comparing efficiency across different kinematic regions.

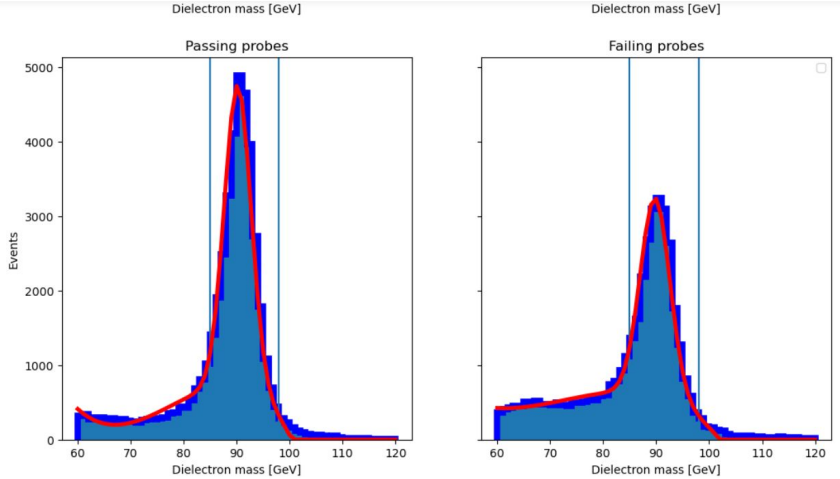
- This approach has its own limitations such as having too many fits and jagged function.

```python
ptbins = np.array([5., 10., 15., 20., 25., 30., 50., 100., 200.]) # np.linspace(20, 100, 5)
print(ptbins)
eff_per_bin = [
    tag_eff(
        zcands.mass[zcands.goodprobe & (zcands.probe_pt >= lo) & (zcands.probe_pt < hi)],
        zcands.mass[~zcands.goodprobe & (zcands.probe_pt >= lo) & (zcands.probe_pt < hi)],
    )
    for lo, hi in zip(ptbins[:-1], ptbins[1:])
]
```

```python
eff_per_bin
```

```
[0.29659522187813986+/-0.02300097550851576,
 0.38387381129747944+/-0.011911475390449584,
 0.539779086880948+/-0.013774650534399361,
 0.5806001341863474+/-0.01156264650924916,
 0.6543769696436891+/-0.010016099259790563,
 0.7102235361706415+/-0.00680810509767564,
 0.7900737810826105+/-0.007531843694984588,
 0.7990502812035124+/-0.015600457008654723]
```

Binned Tag and Probe




```
def pdf(x, a, b, c, n, mu, sigma, d):
    # background = a + b*x + c*(x**2)
    background = np.polyval([a, b, c, d], x)
    signal = n*norm.pdf(x=x, loc=mu, scale=sigma)
    # np.exp(-0.5*(x-mu)**2/sigma**2)/np.sqrt(2*np.pi)/sigma
    return np.maximum(signal + background, 0)
```

- Two components to fit the graph, Signal and Background.
- The cut and count method takes events from 85 to 100 pT.

Cut and Count T&P

- Cut and Count method simplifies the data analysis by selecting specific regions or "cuts" in the data
- This method calculates the efficiency by counting events that are in the provided mass range (80-100 GeV)

```
import hist.intervals
from uncertainties import ufloat

def cut(pass_mass, fail_mass):
    n_pass = np.sum((pass_mass >= 80) & (pass_mass <= 101))
    n_fail = np.sum((fail_mass >= 80) & (fail_mass <= 101))
    n_pass_unc = np.sqrt(n_pass)
    n_fail_unc = np.sqrt(n_fail)
    x = ufloat(n_pass, n_pass_unc)
    y = ufloat(n_fail, n_fail_unc)
    efficiency = x/(x+y)
    print(efficiency*100,"%")
    uncertainty = hist.intervals.ratio_uncertainty(
        np.array(n_pass),
        np.array(n_pass+n_fail),
        uncertainty_type='efficiency'
    )
    alt_efficiency = ufloat(n_pass/(n_pass+n_fail), uncertainty[0])
    print("alt",efficiency*100,"%")
    return efficiency

# Calculate the efficiency for the current bin using the cut_and_count_efficiency function
eff_bin = cut(
        zcands.mass[zcands.goodprobe & (zcands.mass >= 80) & (zcands.mass <= 101)],
        zcands.mass[~zcands.goodprobe &   (zcands.mass >= 90) & (zcands.mass <= 95)],
    )

eff_per_bin_cut = [
    cut(
        zcands.mass[zcands.goodprobe & (zcands.probe_pt >= lo) & (zcands.probe_pt < hi)],
        zcands.mass[~zcands.goodprobe & (zcands.probe_pt >= lo) & (zcands.probe_pt < hi)],
    )
    for lo, hi in zip(ptbins[:-1], ptbins[1:])
]
eff_per_bin_cut
```

```
[0.18972332015810275+/-0.004929998227297084,
 0.3931218307112209+/-0.003938292196515949,
 0.5013629177342855+/-0.002610467015270283,
 0.5600452609951464+/-0.0019153176942416127,
 0.6078392517438174+/-0.0015368114883751664,
 0.6914601529915649+/-0.0006316787983476892,
 0.7663328021852948+/-0.0045145241806082 6,
 0.8285714285714286+/-0.01839000869562255]
```

Cut and Count T&P

- Compared to binned approach, Cut and Count has an advantage over binned approach mainly for two reasons

1. Reduced complexity: Instead of dividing the data into numerous bins like the binned approach, only a few distinct regions are considered. This reduces the complexity of the analysis that arises from binning choices.

```
1  eff_per_bin
```

```
[0.29659522187813986+/-0.02300097550851576,
 0.38387381129747944+/-0.011911475390449584,
 0.539779086880948+/-0.013774650534399361,
 0.5806001341863474+/-0.01156264650924916,
 0.6543769696436891+/-0.010016099259790563,
 0.7102235361706415+/-0.00680810509767564,
 0.7900737810826105+/-0.007531843694984588,
 0.7990502812035124+/-0.015600457008654723]
```

```
[0.18972332015810275+/-0.004929998227297084,
 0.3931218307112209+/-0.003938292196515949,
 0.5013629177342855+/-0.002610467015270283,
 0.5600452609951464+/-0.0019153176942416127,
 0.6078392517438174+/-0.0015368114883751664,
 0.6914601529915649+/-0.0006316787983476892,
 0.7663328021852948+/-0.00451452418060826,
 0.8285714285714286+/-0.01839000869562255]
```

Cut and Count

2. Greater Sensitivity to Low mass Regions: The Cut and Count method can be more sensitive than the binned approach. By focusing on specific regions in the data, the Cut and Count method can enhance the power to detect rare signals, improving the overall sensitivity of the analysis.

```
# Calculate the efficiency for the current bin using the cut_and_count_efficiency function
eff_bin = cut(
        zcands.mass[zcands.goodprobe & (zcands.mass >= 80) & (zcands.mass <= 101)],
        zcands.mass[~zcands.goodprobe &   (zcands.mass >= 80) & (zcands.mass <= 101)],
    )
```

Unbinned T&P

- Unbinned T&P efficiency measurements do not rely on dividing the data into bins based on some observable (e.g., transverse momentum) like the binned approach. Instead, it evaluates the efficiency for each event individually.

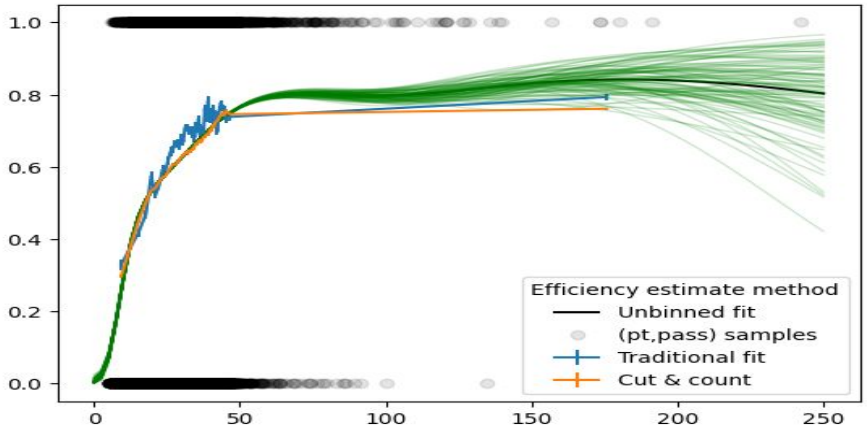
- This method uses Maximum Likelihood Estimation where the observable is Bernoulli-distributed according to a latent efficiency that is a function of certain variables, in this case probe pT.

- Since it does not involve binning, the dimensionality issue associated with high-dimensional binning in the binned approach is avoided.

Unbinned T&P

- The disadvantage to unbinned T&P is that it can be computationally more demanding than the binned approach, especially for large datasets.

- The unbinned efficiency estimation method demonstrates less susceptibility to fluctuations compared to other techniques.



Conclusion

Comparison of Tag & Probe Methods

1. Unbinned Method
   - **Advantages**: Increased precision via mass information; background-aware.
   - **Limit**: it is systematically off from the traditional fit, since we don't take into account the failing probes that are not true because the mass cut only approximates the likelihood of being true.
2. Binned Method
   - **Advantages**: Simplifies analysis, suitable for discrete efficiency regions, reduced data variation.
   - **Limit**: Electrons from failing probe do not get calculated in efficiency.
3. Cut & Count
   - **Advantages**: Quick estimation, straightforward approach for basic analyses.
   - **Limit**: misses electrons that are not in the cut

Future Work

- While our current analysis has successfully demonstrated the precision benefits of the unbinned efficiency estimation method, there remain areas for further refinement and development. In particular, addressing the observed bias in the unbinned approach and enhancing overall accuracy are crucial directions for future work.

- Improved Modeling: Exploring more sophisticated models for the efficiency distribution, such as incorporating machine learning techniques, could lead to more accurate characterizations. These models could better capture the complexities of the underlying physics processes.

Acknowledgment

Nick Smith : Your mentorship was like colliding particles in a particle accelerator – it created sparks of knowledge and a burst of appreciation. Thanks for the 'eureka' moments!

PURSUE Team : Just as matter and antimatter annihilate into pure energy, my gratitude for you is immeasurable and boundless. Thanks for making my internship electrifying!

Fellow Interns : You've accelerated my learning curve so fast that I feel like I've time-traveled through the internship! Thanks for making time fly!