Light and heavy jets combinatorics

Romain Madar

August 2018

Abstract

This note tries to understand how the correction works for the yields of events with a veto on b-tagged jets. In order to do so, we play with basic probabilities assuming a kinematically flat b-tagging efficiency ϵ and a given number of (b) jets at the truth level. First we start with only truth b-jets to look at exclusive and inclusive probabilities versus ϵ and the number of reconstruced b-jets. Then, we add light jets at the truth level with the corresponding mis-tag rate f.

Contents

1 Assuming only truth b-jets

Some definition to start with:

- $N_b^{truth} \equiv$ number of truth b-jets in the event.
- $N_b^{reco} \equiv$ number of b-tagged jets in the events
- $N_{N_b^{reco}} \equiv$ number of events with N_b^{reco} b-tagged jets

We can write the number of events with N_b^{reco} b-tagged jets knowing the probability $\mathscr{P}(N_b^{reco}|N_b^{truth},\epsilon)$ to tag N_b^{reco} b-jets among N_b^{truth} truth b-jets assuming a b-tagging effiency of ϵ:

$$N_{N_b^{reco}} = \sum_{Evt} \mathscr{P}(N_b^{reco}|N_b^{truth},\epsilon)$$

where

$$\mathscr{P}(N_b^{reco}|N_b^{truth},\epsilon) = \binom{N_b^{truth}}{N_b^{reco}} \times \epsilon^{N_b^{reco}} \times (1-\epsilon)^{N_b^{truth}-N_b^{reco}} \tag{1}$$

$$= \frac{N_b^{truth}!}{N_b^{reco}!(N_b^{truth}-N_b^{reco})!}\epsilon^{N_b^{reco}} \times (1-\epsilon)^{N_b^{truth}-N_b^{reco}} \tag{2}$$

(3)

1.1 Exclusive probabilities

```
def P(eff,n_reco,n_truth):
    return comb(n_truth,n_reco) * eff**n_reco * (1-eff)**(n_truth-n_reco)
```

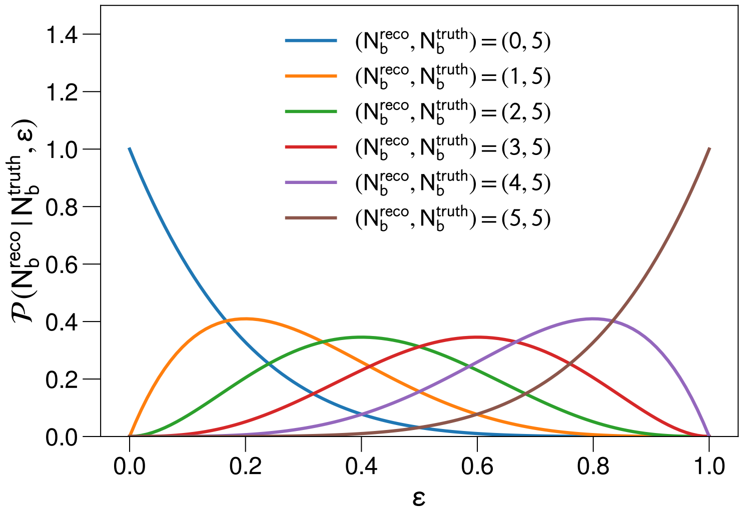

Figure 1: Probability to reconstruct N_b^{reco} for $N_b^{\mathrm{truth}} = 5$, as function of b-tagging efficiency.

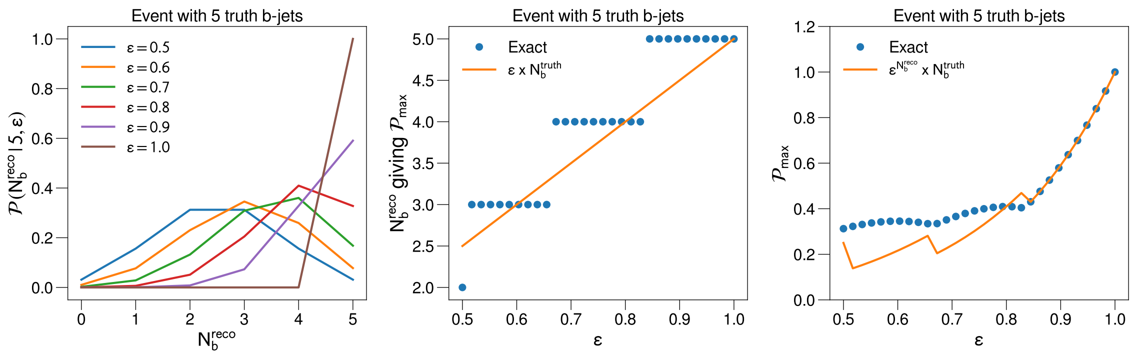

Figure 2: Probability as fuction of the reconstruced b-jets (left), number of b-tagged jets with the maximum probability as function of the efficiency (middle) maximum probability as function of the efficiency (right)

1.2 Inclusive probabilities

```
def InclusiveP(epsilon,nbmin,nj):
    return np.sum( [P(epsilon,nb,nj) for nb in range(nbmin,nj+1)] )
```

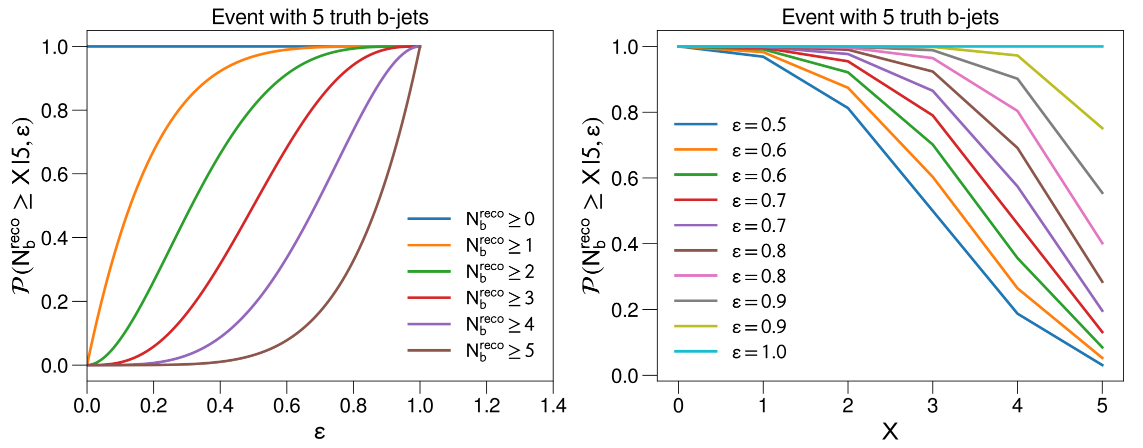

Figure 3: Inclusive probability as function of the efficiency (left) and as function of number of jets (right)

2 Assuming both truth b-jets and light jets

- $N_b^{truth} \equiv$ number of truth b-jets in the event.
- $N_b^{reco} \equiv$ number of b-tagged jets in the events
- $N_j^{truth} \equiv$ number of truth light jets in the event.
- $N_j^{reco} \equiv$ number of reconstructed light jets in the events

In principle, the interesting probability would be now $\mathscr{P}(N_b^{reco}, N_j^{reco} | N_b^{truth}, N_j^{truth}, \epsilon, f)$ where f stands for the mis-tag rate (ie the probability that a light jet is tagged as a b-jet). But since, we assume here a jet reconstruction efficiency of 100% (ie $N_b^{truth} + N_j^{truth} = N_b^{reco} + N_j^{reco}$), this reduces to $\mathscr{P}(N_b^{reco} | N_b^{truth}, N_j^{truth}, \epsilon, f)$.

There is an easy way to split the problem since we need to consider all configuration where the sum of tagging and mis-tagging gives the wanted number of b-tagged jets:

$$\mathscr{P}(N_b^{reco} | N_b^{truth}, N_j^{truth}) = \sum_{k+\ell=N_b^{reco}} \mathscr{P}(N_b^{truth} \to k, \epsilon) \times \mathscr{P}(N_j^{truth} \to \ell, f)$$

where $\mathscr{P}(N_b^{truth} \to k)$ is the probability to tag k reco b-jets (from truth b-jets) and $\mathscr{P}(N_j^{truth} \to \ell)$ is the probability to mis-tag ℓ reco b-jets (from truth light jet). These individual probabilities are easy to write:

$$\mathscr{P}(N_b^{truth} \to k, \epsilon) = \binom{N_b^{truth}}{k} \times \epsilon^k \times (1-\epsilon)^{N_b^{truth}-k} \mathscr{P}(N_j^{truth} \to \ell, f) = \binom{N_j^{truth}}{\ell} \times f^\ell \times (1-f)^{N_j^{truth}-\ell}$$

```
def proba_light_to_heavy(nbr,njtruth,f):
    '''
    Return the proba to have nbr mis-tagged jets
    and njtruth-nbr light jets from njtruh truth light jets
    '''
    if (nbr>njtruth): return 0.0
    return comb(njtruth,nbr) * f**nbr * (1-f)**(njtruth-nbr)

def proba_heavy_to_heavy(nbr,nbtruth,e):
    '''
    Return the proba to have nbr tagged jets and njtruth-nbr
    not tagged jets from nbtruh truth b-jets
    '''
    if(nbr>nbtruth): return 0.0
    return comb(nbtruth,nbr) * e**nbr * (1-e)**(nbtruth-nbr)

def ProbaFull(nbr,nbt,njt,e,f):
    '''
    Return the probability to have \'nbr\' b-tagged jets in a
    event with \'nbt\' truth b-jets and \'njt\' truth light jets.
    '''
    kl = [[nbr-i,i] for i in range(0,nbr+1)]
    each_proba = [
                    proba_heavy_to_heavy(k,nbt,e)*\
                    proba_light_to_heavy(l,njt,f)
                    for [k,l] in kl
                ]
    return np.sum(each_proba)
```

2.1 Exclusive probabilities

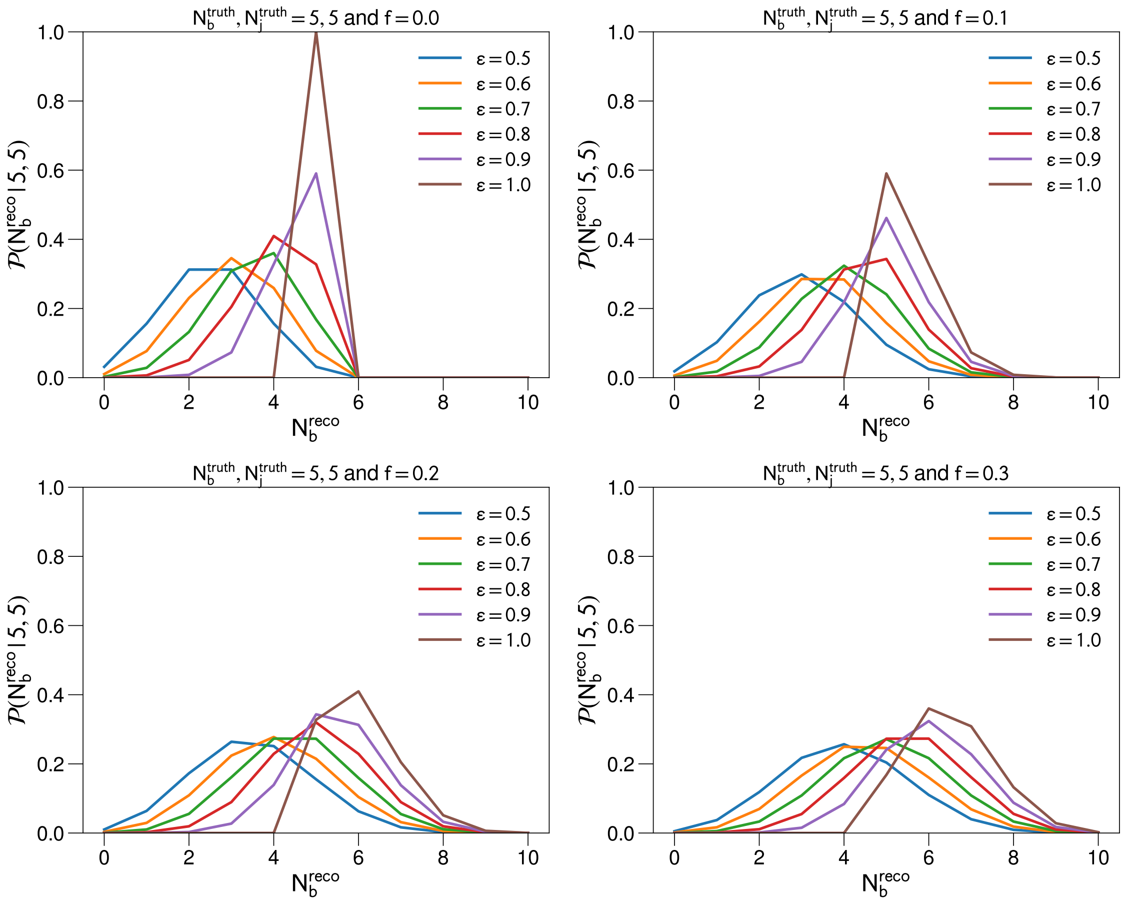

Figure 4: Probability to reconstruct N_b^{reco} as a function of N_b^{reco} for four different value of fake rate f for 5 truth b-jets and 5 truth lights jets

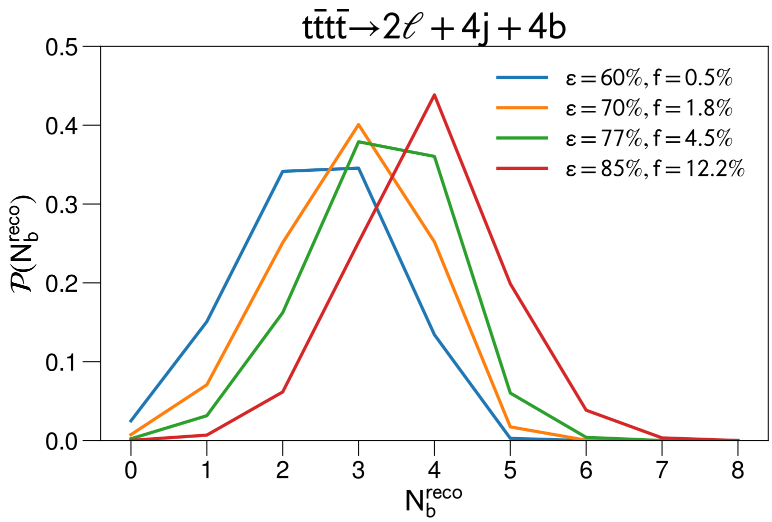

Figure 5: Probability to reconstruct N_b^{reco} for $t\bar{t}t\bar{t}$ events using typical working points of ATLAS b-tagging

One particular case of exclusive probabilities is the efficiency of a veto of the number of b-jets, particularly important for certain analysis. Figure 6 shows how this probability evolves with the b-tagging efficiency for different value of the fake rate f.

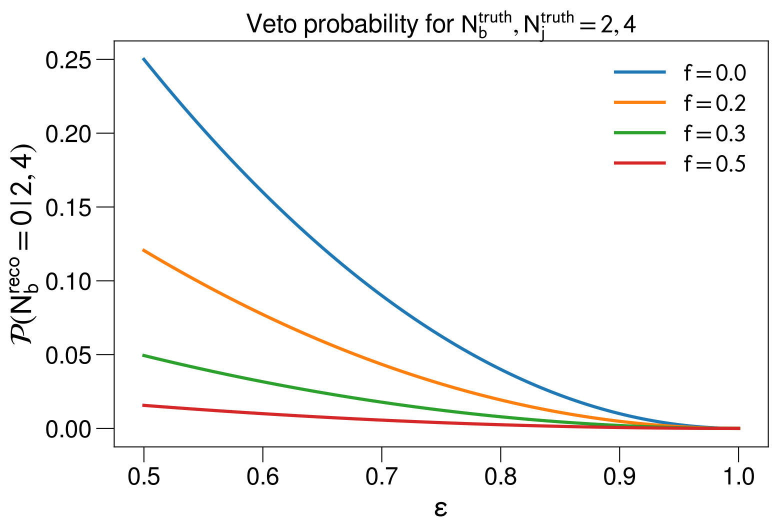

Figure 6: Probability to reconstruct exactly 0 b-jets as function of the efficiency ϵ for different fake rate values

2.2 Inclusive probabilities

```python
def InclusiveProbaFull(nbmin,nbt,njt,e,f):
    return np.sum( [ProbaFull(nbr,nbt,njt,e,f) for nbr in
    ↪  range(nbmin,nbt+njt+1)] )
```

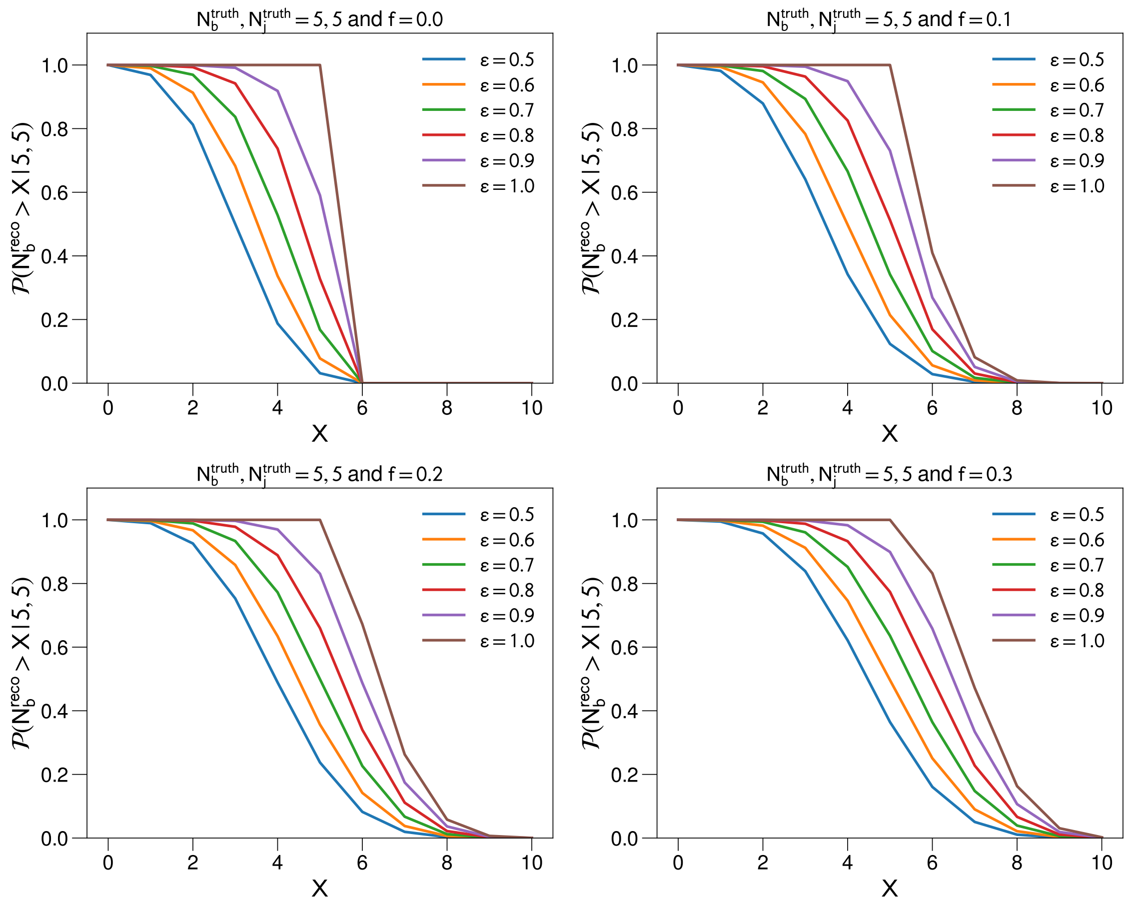

Figure 7: Probability to reconstruct at least X b-jets as function of X for different fake rate and efficiencies values.

3 Correcting of data to MC differences using a jet-based weight

3.1 Exact correction

Given the formulas described before, applying the **proper weight to each jet** allows to properly correct efficiency and fake rates. This weight changes depending on its true nature (b or light) and its reconstructed type (b or light). Four combinations are possible:

1. true positive (TP): true b-jet b-tagged → receive a factor $\epsilon_{data}/\epsilon_{mc}$
2. false positive (FP): true light jet b-tagged → receive a factor f_{data}/f_{mc}
3. true negative (TN): true light jet not b-tagged → receive a factor $(1-f_{data})/(1-f_{mc})$
4. false negative (FN): true b-jet not b-tagged → receive a factor $(1-\epsilon_{data})/(1-\epsilon_{mc})$

In that way, each probability in the MC is mathematically corrected to the one in data. The detail if given for $\mathscr{P}(N_b^{truth} \to k)$ but works in the same way for $\mathscr{P}(N_j^{truth} \to \ell)$.

$$\mathscr{P}^{corr}_{mc}(N_b^{truth} \to k,\, \epsilon_{mc}, \epsilon_{data}) = \binom{N_b^{truth}}{k} \times \epsilon_{mc}^k \times (1-\epsilon_{mc})^{N_b^{truth}-k} \times \left(\frac{\epsilon_{data}}{\epsilon_{mc}}\right)^k \times \left(\frac{1-\epsilon_{data}}{1-\epsilon_{mc}}\right)^{N_b^{truth}-k} \quad (4)$$

$$\mathscr{P}^{corr}_{mc}(N_b^{truth} \to k,\, \epsilon_{mc}, \epsilon_{data}) = \mathscr{P}_{data}(N_b^{truth} \to k,\, \epsilon_{data}) \quad (5)$$

```
def ProbaFull_corr(nbr,nbt,njt,eMC,fMC,eDATA,fDATA):
    kl = [[nbr-i,i] for i in range(0,nbr+1)]
    all_proba = [
                 proba_heavy_to_heavy(k,nbt,eMC) * (eDATA/eMC)**k *
 ↪  ((1-eDATA)/(1-eMC))**(nbt-k) * \
                 proba_light_to_heavy(l,njt,fMC) * (fDATA/fMC)**l *
 ↪  ((1-fDATA)/(1-fMC))**(njt-l)\
                 for [k,l] in kl\
                ]
    return np.sum(all_proba)

def InclusiveProbaFull_corr(nbmin,nbt,njt,eMC,fMC,eDATA,fDATA):
    return np.sum( [ProbaFull_corr(nbr,nbt,njt,eMC,fMC,eDATA,fDATA)
                    for nbr in range(nbmin,nbt+njt+1)] )
```

3.2 Efficiency ratio are not enough: absolute efficiencies matter too

Usually, we often think on correcting the efficiencies with a scale factor $SF = \epsilon_{data}/\epsilon_{mc}$ which might let think that corrections is independant from absolute efficiences values but only depend on the ratio. This is however not true because of the inefficiency correction. To illustrate this, let's assume that the efficiency is scaled by a factor k **in both data and simulation, letting** SF **constant**. How such a change will affect the various probabilities? This is looked at in the example of semi-leptonic decay of $t\bar{t} + 2j$.

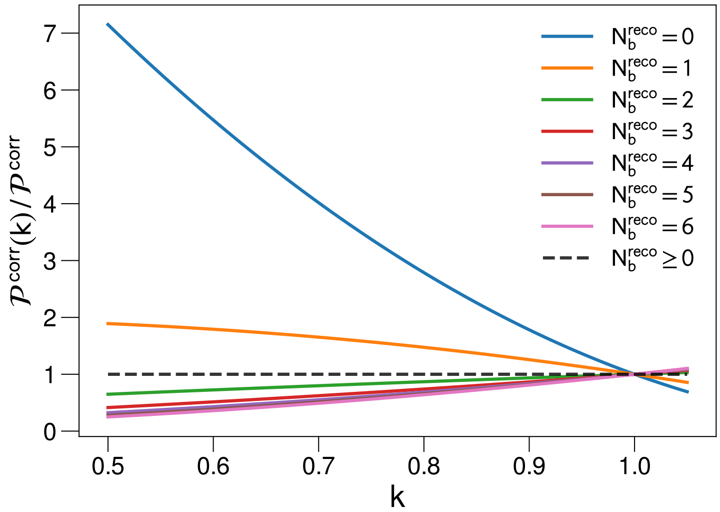

Figure 8: Evolution of the corrected probability when the efficiency is scaled by a factor k.

What we see in Figure 8 is actually easy to interpret: for low k values, the absolute effiency is low and it's more likely to observe no b-tagged jets. And in the current case, the correction is still applied properly since the correct efficiencies are used to correct simulation effiencies.

Now the question is: **what happens if the simulated efficiencies in the control region (CR) are different from the one in the signal region (SR)?** This would mean that the correction to apply is not consistent and a closure test can be performed to probe this inconsitency. In practice, we assume three types of efficiencies: $\epsilon_{MC,CR}$, ϵ_{DATA} and $\epsilon_{MC,SR}$. The corrections are computed using $\epsilon_{MC,CR}$ and ϵ_{DATA} while they are applied in the SR where simulated efficiency is $\epsilon_{MC,SR}$.

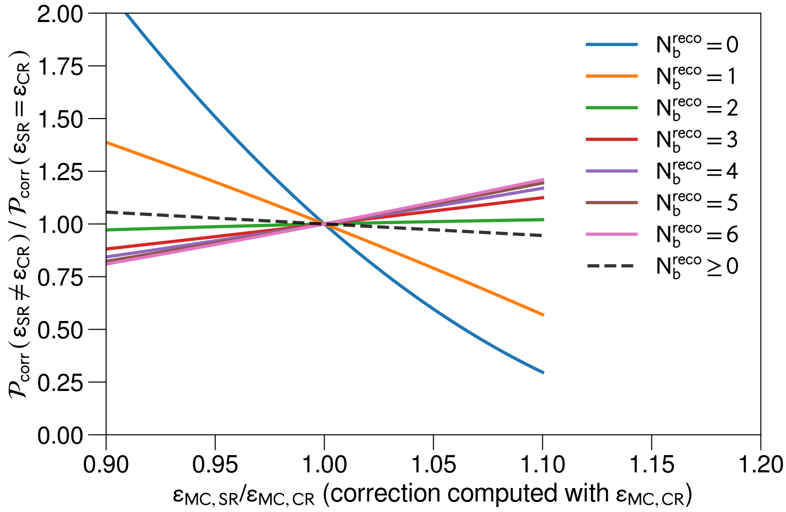

Figure 9: Ratio of corrected probability with the incorrect SF with the one corrected with the correct SF as a function of the SF ratio, for different numbers of reconstructed b-jets.

Figure 9 shows how the sum of the probability $\mathscr{P}(N_b^{reco} \geq 0)$ deviates from 1.0 when $\epsilon_{MC,SR}$ deviates from $\epsilon_{MC,CR}$ (assuming the fake rate SF are the same in both regions). Also, each exclusive probability deviation is shown and the 0-tag probability is has the largest effect (5% efficiency variation leads to $\sim$ 50% probability variation). Figure 10 shows the same information but as a function of a only accessible quantity, namely the CR-corrected probability.

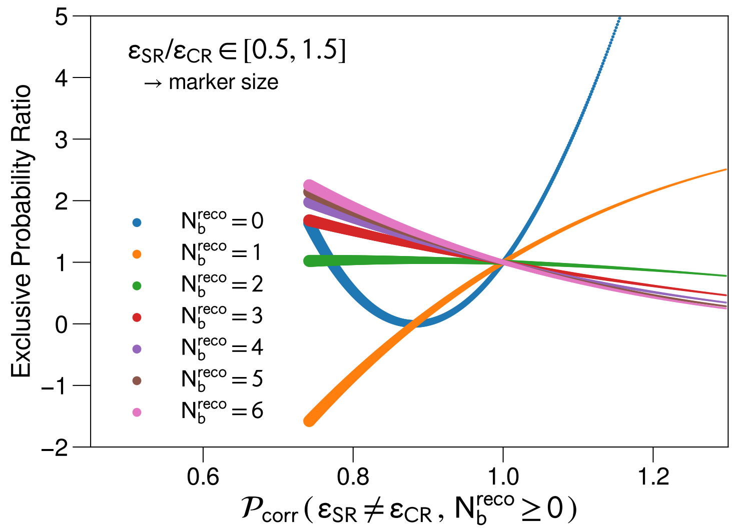

Figure 10: Ratio of corrected probability with the incorrect SF with the one corrected with the correct SF as a function wrongly-corrected probability (accessible experimentally),for different numbers of reconstructed b-jets.