Combining b-tagging calibrations in ATLAS

Gordon Watts (UW/Seattle)

For the ATLAS Collaboration

Abstract

The ATLAS collaboration calibrates its b-tagging algorithms using a number of data-driven methods. The b-tagging algorithms, used to separate b-quark jets from light-flavor jets, are widely used in top, Higgs, and Exotics analyses in ATLAS. The algorithm's performance is measured for light-flavor jets, charm jets, and bottom jets. In some cases multiple methods can be used to measure the performance: methods using di-jet data and a pure sample of top-antitop are being used to measure the b-jet performance. These methods have complementary strengths – the di-jet methods tend to have the smallest errors at low jet energy, and the $t\bar{t}$ methods tend to have smaller errors at high jet energy. We used a minimum likelihood method to fit the results, profiling common systematics, to combine the results. For results on common data we have also understood the statistical correlation of our methods. This poster will describe the details of the combination.

Where the Calibration Fits Into The Analysis

Efficiency or MC background predictions start from a MC sample containing b-quark jets (like $t\bar{t}$).

MC Sample

The b-tagging algorithm is applied directly to the MC

Apply b-tagging

This yields the number of events the analysis expects to see in Monte Carlo.

Predicted # Events in MC

A scale factor accounts for any differences between data and MC.

Data/MC Scale Factor

And that yields the expected number of events in data, along with additional errors.

Predicted # Events in Data

Find the true value of the b-tagging Calibration Scale Factor

The True Value - ϵ_b

p_T^{rel} Estimation

Statistical Error
JES Error
Template Error

$t\bar{t}$ Estimation

Statistical Error
JES Error
Luminosity Error
(this is will be included in future work)

System 8

Statistical Error
JES Error
Background Error

ATLAS makes a number of measurements of the b-tagging scale factor. Each measurement is a different way of getting at the true tagging efficiency. Accurately combining the various measurements should yield a better approximation of the true value. The current combination uses only the dijet methods, p_T^{rel} and System 8, as inputs. Future work will use a number of independent analyses from the $t\bar{t}$ channel in addition.

Tools

- RooFit – a fitting package, part of the ROOT distribution, is used for the fitting. Heavy use is made of the RooGaussian object. The Gaussians shown below in this poster are then combined using RooProduct and RooAddition. MINUIT is used to do the final fit and find the maximum and the best values of the scale factors. Some tuning was required to the fitting process to make sure there was convergence in as many cases as reasonable.
- Spirit – A parser-combinator library that is part of the boost library distribution. parser-combinator libraries allow you to construct a grammar using C++ method calls and for a small language are much more approachable and usable than a full blown parser like Bison. Used to parse the result files from each of the calibrations. And kept the inputs more readable than XML.
- C++ was heavily used to glue everything together. A better choice, in retrospect, would have been python because a great deal of string manipulation was required.

Combining Calibrations

p_T^{rel}, MV1, 70% operating point

Bin 1 | Bin 2 | Bin 3

Terms are added for each bin of each measurement for a common tagger and operating point and scale factor type (bottom, charm, or light flavor).

$G(SF_3|SF_{b3}f_{sys3}, \delta SF_3^{stat})$ - A Gaussian evaluated at the measured scale factor for this bin, SF_3, with an uncorrelated statistical error δSF_3^{stat}. This is a measurement of the true value of the scale factor in this kinematic bin, SF_{b3}. f_{sys3} is the systematic error factor, below.

$f_{sys3} = 1 + \delta m_3^{err_1} s_1 + \delta m_3^{err_2} s_2 + \cdots$ - The true value of the scale factor SF_{b3} is modified by all the systematic errors for this scale factor measurement. The contribution of each systematic is in the $\delta m_3^{err_i}$ term.

$G(0, s_1, 1)$ – A Gaussian term to keep the contributions of the systematic errors as they are intended. Implicit in this formalism are that all errors are Gaussian and symmetric. This method does not demand that, however. The s_i is common between measurements that share the same systematic error.

System 8, MV1, 70% operating point

Bin 1 | Bin 2 | Bin 3

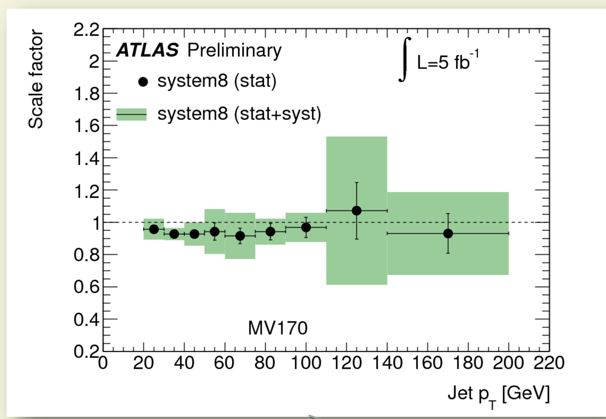
+ Additional Bins

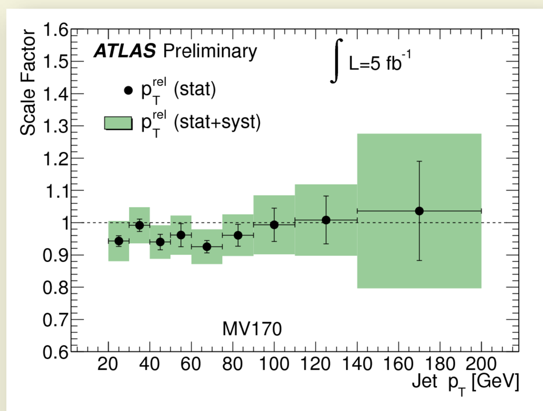
$$G\left(SF_2|SF_{b2}\left(1 + \delta m_2^{err_1} s_1 + \delta m_2^{err_2} s_2\right), \delta SF_2^{stat}\right)$$
$$\times\, G\left(SF_3|SF_{b3}\left(1 + \delta m_3^{err_1} s_1 + \delta m_3^{err_2} s_2\right), \delta SF_3^{stat}\right)$$
$$\times\, G(0, s_1, 1) G(0, s_2, 1)$$

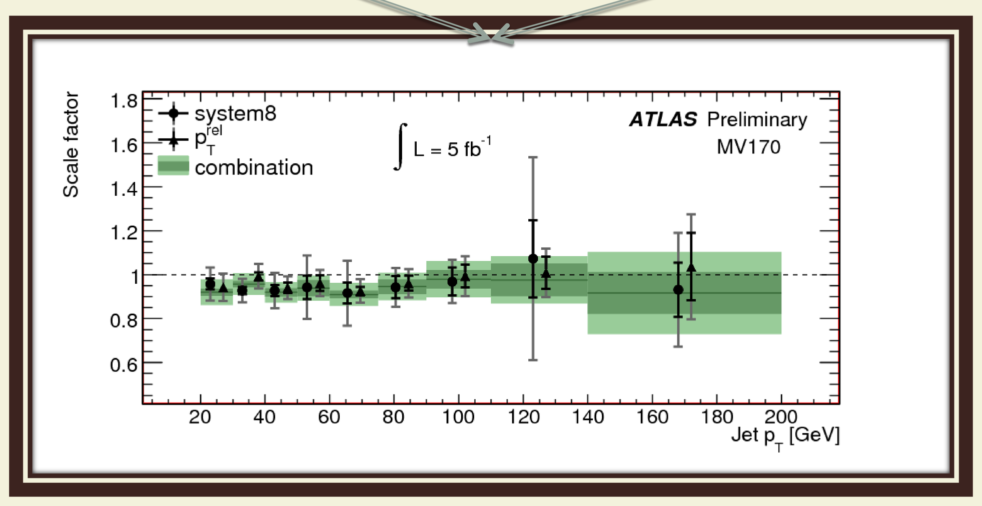
Maximize this function to find the best values for SF_{bi}.

Determining the errors:

- Analyses need individual error contributions to properly propagate the errors in their analyses and avoid double counting the error.
- Fix each s_i to zero in turn, and refit. The different, in quadrature, is the contribution of that systematic to the total error.
- Fix all systematic errors to zero and refit to determine the statistical error.







Calibration Methods

ATLAS is using two primary calibration methods for the efficiency and two for the fake rate. These are described in detail in ATLAS public notes and other posters here at PLHC. The fake rate methods are not described here (see the *Problems* section below). The efficiency methods are both based on dijet data:

- **The p_T^{rel} method** – About 15% of bottom quark decays result eventually in a muon. The relative p_T of the muon to the jet axis (p_T^{rel}) is a function of the mass of the object the muon decayed from. Bottom quarks (at 5 GeV) are much more massive than other objects that produce muons. Thus this distribution can be used as an independent way of measuring the efficiency. It is dependent on MC templates.
- **The System 8 method** – Two uncorrelated taggers, two data samples, and 8 equations with 8 unknowns that relate the data samples; one can solve for the tagging efficiency. The straight forward method needs very little in the way of input from MC, though a thorough understanding of the systematics does.

Each measurement comes with a long list of correlated and uncorrelated systematic errors.

Since both methods are based on the same dijet events and similar triggers, we must account for statistical correlation as well as systematic correlations.

Statistical Correlations

The two calibrations methods are statistical correlated because they are based on the same underlying jet sample. In order to combine them a study was done to determine the extend of the statistical correlation.

- In each kinematic bin the overlap in number of jets used in each calibration method was studied. Due to differences in the triggers, in some bins the overlap was close to 100%, and in other bins it was almost non-existent.
- For bins where the overlap was not negligible, a toy Monte Carlo was used to determine the correlation.
- The numbers of events in the two samples in System 8 along with the flavor composition and the b-tagging efficiency were taken for a particular tagger and working point.
- The expected p_T^{rel} distribution was taken for similar samples from MC.
- Using these numbers and distributions, one can now run a toy MC.
- Re-fit both p_T^{rel} and System 8 for the efficiency on each iteration. The scatter plot contains the correlation between the two taggers.

The statistical error is then split into a correlated part and an uncorrelated component and added to the fit above. The uncorrelated part is the new statistical error (δSF_2^{stat}) and the correlated part becomes a new systematic error.

Getting the statistical correlations correct with this method is a bit tricky. Our first attempt we mis-estimated how much data was shared in some of the bins resulting in the bins being more correlated than they should have been. This can lead to invalid results when results from the two methods are averaged.

Sample Input File

```
Analysis(pTrel,bottom,MV1,0.601713,AntiKt4Topo)
{
    bin(20<pt<30,0<abseta<2.5)
    {
        central_value(0.9429,0.0166)
        sys(jet vertex fraction,2.9360%)
        sys(modelling of the b-hadron direction,0.1395%)
        sys(modelling of b-production,0.4874%)
        sys(jet energy resolution,-0.1477%)
        sys(ptrel light template contamination,-0.1181%)
        sys(scale factor for inclusive b-jets,4.0000%)
        sys(charm-light ratio,3.0153%)
        sys(b-fragmentation fraction,0.1396%)
        sys(semileptonic correction,-0.5147%)
        sys(muon pT spectrum,0.0789%)
        usys(simulation tagging efficiency,0.9797%)
        sys(jet energy scale,0.0056%)
        sys(b-decay p* spectrum,-0.5902%)
        sys(modelling of c-production,0.1264%)
        usys(MC statistics,2.2405%)
        sys(b-decay branching fractions,0.0120%)
        sys(b-fragmentation function,0.0273%)
        sys(pileup mu reweighting,0.0938%)
        sys(fake muons in b-jets,0.0187%)
    }
...
```

Analysis, flavor, tag algorithm, and operating point

$\underbrace{\hphantom{Analysis, flavor, tag algorithm, and operating point}}_{\text{Everything with these in common is fit simultaneously}}$

Bin

Central Value and Systematic Error

Systematic error correlated across bins and analyses

Systematic error correlated across analyses but not bins

Comments & Future Work

RooFit does not always converge. In cases where the fit did not work we took the best measurement in the bin and used that as the final result.

Final results, with complete systematic errors, are written out to a ROOT files in the form of histograms which is used directly by ATLAS analyzers.

Future work is mostly driven by including new measurements in the combination. Specifically those from $t\bar{t}$. The scale factors from p_T^{rel} and System 8 are most powerful in the low jet p_T bins. We expect tt^- to be most powerful in the higher bins. We also expect to do further tuning to the algorithms to improve our ability to combine edge cases – like the fake rates.

We will also study the compatibility between the measurements – some regions will have 3 or more measurements