PROCEEDINGS OF SCIENCE

Two New Developments on the Statistical Treatment of Flavour Tagging Uncertainties in ATLAS

Yan Ke,[a,*] **Ilaria Luise,**[a] **Giacinto Piacquadio**[a] **and Quentin Buat**[a,b]

[a] *Stony Brook University, Stony Brook, NY - USA*

[b] *University of Washington, Seattle, WA - USA*

E-mail: yan.ke@cern.ch

The document introduces two new methods on the implementation of flavour tagging uncertainties in ATLAS physics analyses. In order to reduce the number of flavor-tagging calibration uncertainties, the physics analyses use an eigenvector decomposition approach. However, the resulting flavour tagging eigenvectors are in general not the same across flavour tagging selections, so the uncertainties can not be directly correlated in combination analyses. A new method, called *eigenvector recomposition*, has been designed to overcome this problem. This proceeding describes the method and gives practical examples about its usage in physics analyses, focusing on the $VH, H \to b\bar{b}$ analysis.

The second development involves the flavour tagging uncertainties in analyses with high-transverse momentum jets. The in-situ calibration of the flavour tagging uncertainties is computed using events with jet-p_T spectra up to 140-250 GeV and used through all the jet-p_T spectrum. Therefore, at higher transverse momenta the calibration needs dedicated extrapolation uncertainties in order to account for possible deviations from the central value. The second part of the document describes the method used to extract these extrapolation uncertainties starting from Z' simulated events.

The 10th Annual Large Hadron Collider Physics Conference - LHCP 2022
16-21 May, 2022
Taipei, Taiwan

*Speaker

PoS(LHCP2022)322

1. Eigenvector recomposition

About ∼100 separate uncertainty sources are accounted for in the b-jet calibration analysis. As described in [4], an *eigenvector decomposition* method is used to reduce the number of flavor-tagging uncertainty components in the analyses to $O(10)$ while correctly preserving correlations across jet kinematic bins. The eigenvectors are closely related to the diagonalization of a specific covariance matrix, so they can not be directly correlated across analyses, as they correspond to different combinations of the same sources of uncertainties. This study introduces the concept of *eigenvector recomposition*.
The full mathematical derivation of the recomposition method can be found in Ref. [4]

1.1 Mathematical description

The covariance matrix V_{ij} of the scale factors between two different p_T bins ($i,j \in \{1,..., N\}$) is given by:

$$V_{ij} = \sum_{k=1}^{M} u_{ik}\, u_{jk}, \tag{1}$$

where i and j are p_T-bin indices, k denotes the uncertainty component and u the bin-dependent SF uncertainty amplitude. The diagonalisation of V_{ij} leads to N eigenvectors (v) and N eigenvalues (λ), one for each dimension of the covariance matrix. The i^{th} p_T bin of k^{th} eigenvector uncertainty is: $\tilde{u}_{ik} = \lambda_k v_{ik}$. The total SF variation in each p_T bin i, referred to as T_i, can be expressed as:

$$T_i = \sum_{k=1}^{L} \theta_k \lambda_k v_{ik} = \sum_{k=1}^{L} \theta_k \tilde{u}_{ik} \tag{2}$$

where $k \in \{1, ..., L\}$, is the eigenvector index and θ_k are the nuisance parameters used in the analysis. At the same time, T_i can also be expressed as a function of the *original* sources of systematic uncertainties:

$$T_i = \sum_{j=1}^{M} \eta_j u_{ij} \tag{3}$$

where η_j is one of the M nuisance parameters representing the original sources of systematic uncertainties. This leads to an equality that can be written shortly as: $\vec{\eta} \cdot U = \vec{\theta} \cdot \tilde{U}$, where $\tilde{U}$ is a matrix with rank L and its right inverse $\tilde{U}^{-1}_{\text{right}} = \tilde{U}^T(\tilde{U}\tilde{U}^T)^{-1}$ is a pseudo-inverse (Moore–Penrose inverse) of $\tilde{U}$ and therefore must be unique. Thus, multiplying the expression above by $\tilde{U}^{-1}_{\text{right}}$ leads to an expression for $\vec{\theta}$ in terms of $\vec{\eta}$: $\vec{\theta} = \vec{\eta} \cdot U \cdot \tilde{U}^{-1}_{\text{right}}$. One can hence define a dimension $M \times L$ matrix: $\mathfrak{R} \equiv U\,\tilde{U}^{-1}_{\text{right}}$ that translates the uncertainties between the original set and the eigenvector decomposed uncertainty set.

1.2 Implementation

Physics analyses using maximum-likelihood fit to extract their parameters of interest with the diagonalized b-tagging uncertainties encoded in the nuisance parameters θ can be expressed as:

$$\mathcal{L} = L(\vec{\mu}, \theta_1, \theta_2, ..., \theta_L)P(\theta_1)P(\theta_2)...P(\theta_L), \tag{4}$$

where $\vec{\mu}$ is the vector of the parameters of interest, and the $P(\theta_k)$ represent the Gaussian priors centered at zero and with variance set to unity for the flavour tagging eigenvectors. The aim of the eigenvector recomposition approach is to transform the likelihood function in Eq. 4 into likelihood function depending on original nuisance parameters η:

$$\mathcal{L} = L(\vec{\mu}, \eta_1, \eta_2, ..., \eta_M)P(\eta_1)P(\eta_2)...P(\eta_M), \tag{5}$$

The nuisance parameters θ (after diagonalization) and the nuisance parameters η (before diagonalization) represent the exact same uncertainties in the space of the b-tagging scale factors, i.e. the same central values and same covariance matrix. Concisely, in order to modify Eq. 4 to make it identical to Eq. 5 the eigenvector recomposition tool replace θ priors by η priors and replace θ in L by η using Eq. 3

2. High-p_T extrapolation uncertainties

The second development introduced in this note involves the treatment of the calibration uncertainties for high-p_T jets, in particular in the extrapolation regime above the highest p_T of the in-situ flavour tagging calibration. In these regions in fact the calibration uncertainties need to be complemented by additional high p_T extrapolation uncertainties calculated from simulation, in order to account for possible deviations from the Scale Factor (SF) values. extrapolated from the lower jet p_T.

2.1 Derivation of the extrapolation uncertainties

The p_T value that determines the division between high p_T region and the standard in-situ calibration region is called the p_T *reference point*. The SFs in the extrapolation region are related to the SFs at the highest p_T reference point of the in-situ calibration region by:

$$SF_{p_T} = SF_{p_T,ref} \cdot R(p_T) \tag{6}$$

where R is a factor depending on p_T. The value of R is assumed to be equal to one since the SFs do not have a strong dependence on p_T, as reported in Ref. [3, 5]. However, R has associated uncertainties that need to be evaluated since it is used to express $\Delta_{rel}(SF_{p_T})$, the relative uncertainty of the SF in the high-p_T extrapolation region:

$$\Delta_{rel}(SF_{p_T}) = \Delta_{rel}(SF_{p_T,ref}) + \Delta_{rel}(R) \tag{7}$$

The relative variation of R is evaluated from Monte-Carlo samples as described below[1]. It is assumed that the potential relative variations of the efficiency in data is equal to MC. Therefore,

[1]Please note that the statistical uncertainties in the high-p_T region are inherited from the statistical uncertainties of the highest p_T bin in the standard calibration region. Therefore in the current procedure, only the systematic variations are evaluated in the high p_T region.

given the variation of the MC b-tagging efficiency, corresponding to a specific source of uncertainty the scale factor is varied correspondingly by:

$$SF' = \frac{\epsilon_{data} + \Delta(\epsilon_{data})}{\epsilon_{MC}} = \frac{\epsilon_{data}(1 + \Delta_{rel}(\epsilon_{MC}))}{\epsilon_{MC}} = SF(1 + \Delta_{rel}(\epsilon_{MC})) \tag{8}$$

Based on the result above, the variation on R becomes:

$$R' \approx R \cdot (1 + \Delta_{rel}(\epsilon_{MC,p_T}) - \Delta_{rel}(\epsilon_{MC,p_T,ref})) \tag{9}$$

Therefore the relative uncertainty on R is:

$$\Delta_{rel}(R) \approx \Delta_{rel}(\epsilon_{MC,p_T}) - \Delta_{rel}(\epsilon_{MC,p_T,ref}) \tag{10}$$

where from Eq. 9 to 10 the assumption $R = 1$ is used. Combining Eq. 10 with Eq. 7, the SF extrapolation uncertainty is:

$$\Delta_{rel}(SF_{p_T}) = \Delta_{rel}(SF_{p_{T,ref}}) + \Delta_{rel}(\epsilon_{MC,p_T}) - \Delta_{rel}(\epsilon_{MC,p_{T,ref}}) \tag{11}$$

2.2 Results

The extrapolation uncertainties are computed using simulated $Z' \rightarrow q\overline{q}$ ($q = uds, c, b$) events with a Z' mass of 3 TeV to enhance the high-p_T region. The jets in the sample are reconstructed as Variable-Radius (VR) track-jets [7]. The extrapolation uncertainties are evaluated separately for truth b-, c- and $light$-jets.

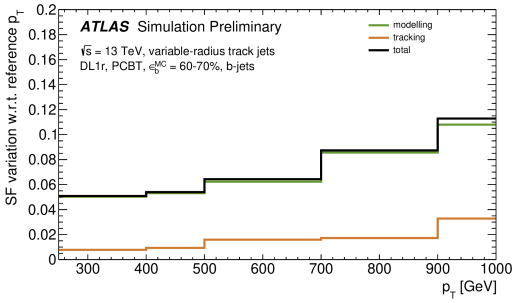

Figure 1: Monte Carlo based extrapolation uncertainties for the DL1r tagger applied to Variable Radius track jets as a function of the reconstructed b-jet pT (black curve). The contribution from experimental uncertainties affecting the reconstruction of tracks is shown in orange. Theoretical uncertainties on the modelling of the jet fragmentation are drawn in green. The total variation is obtained summing in quadrature the single contributions. The figure refers to the 60-70% Working Point interval of the pseudo-continuous b-tagging calibration, which is a simultaneous calibration of a binned version of the continuous b-tagging score distribution, expressed in intervals of the b-tagging efficiency [2]. The uncertainties are evaluated largely following the method described in Ref. [1]. The events are required to have b-hadron $p_T \leq 1000$ GeV. The monotonic requirement described in Ref. [1] is now dropped and the extrapolation uncertainties are applied to $pT \geq 250$ GeV jets.

Figure 1 shows an example of extrapolation uncertainties for b-jets. More plots are available in Ref. [6]. The total uncertainty is the quadrature sum of the individual uncertainties. Eigenvector decomposition is done separately on extrapolation uncertainties and in-situ calibration uncertainties. Two eigenvector set is ready to be used in physics analyses.

References

[1] Simulation-based extrapolation of b-tagging calibrations towards high transverse momenta in the ATLAS experiment. Technical report, CERN, Geneva, Mar 2021. URL http://cds.cern.ch/record/2753444. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-003.

[2] ATLAS Collaboration. Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s}$ = 13 TeV. *JHEP*, 08:089, 2018. doi: 10.1007/JHEP08(2018)089.

[3] ATLAS Collaboration. ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s}$ = 13 TeV. *Eur. Phys. J. C*, 79(11):970, 2019. doi: 10.1140/epjc/s10052-019-7450-8.

[4] ATLAS Collaboration. Eigenvector recomposition: a new method to correlate flavor-tagging systematic uncertainties across analyses. Technical report, CERN, Geneva, 2022. URL https://cds.cern.ch/record/2809651. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2022-024.

[5] ATLAS Collaboration. Measurement of the c-jet mistagging efficiency in $t\bar{t}$ events using pp collision data at $\sqrt{s}$ = 13 TeV collected with the ATLAS detector. *Eur. Phys. J. C*, 82(1):95, 2022. doi: 10.1140/epjc/s10052-021-09843-w.

[6] Yan Ke, Ilaria Luise, and Giacinto Piacquadio. Montecarlo-based extrapolation uncertainties for variable radius track jets with multiple b-tagging working points. Technical report, CERN, Geneva, 2022. URL https://cds.cern.ch/record/2808590.

[7] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jets with Variable R. *JHEP*, 06:059, 2009. doi: 10.1088/1126-6708/2009/06/059.