Identification of double b-hadron jets from gluon-splitting with the ATLAS Detector

María Laura González Silva

Doctoral Thesis in Physics

Physics Department

University of Buenos Aires

November 2012

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Física

Identificación de jets con hadrones *b* producidos por desdoblamiento de gluones con el detector ATLAS.

Trabajo de Tesis para optar por el título de
Doctor de la Universidad de Buenos Aires en el área Ciencias Físicas

por **María Laura González Silva**

Director de Tesis: Dr. Ricardo Piegaia
Lugar de Trabajo: Departamento de Física

Buenos Aires, Noviembre 2012

Agradecimientos

Quiero agradecer a mi director, Ricardo Piegaia, por darme la oportunidad de trabajar en el proyecto ATLAS, por su dedicación y su enseñanza constante; y a mis compañeros de grupo, Gastón Romeo, Gustavo Otero y Garzón, Hernán Reisin y Sabrina Sacerdoti por el trabajo compartido y por brindarme su amistad a lo largo de estos años. Quiero agradecer a Ariel Schwartzman por darnos este análisis, por su caudal inagotable de ideas y por su generosidad y la de todo su equipo. Agradezco al Laboratorio CERN, al Experimento ATLAS, a los programas HELEN y e-Planet, al CONICET y al Fundación Exactas por hacer posible la realización de esta tesis.

Quiero agradecer el apoyo de mis compañeros de la carrera, especialmente a mis amigos Cecilia, Tomás y Leandro. Quiero agradecer también a mis compañeros de grupo y oficina, Javier, Yann, Pablo, y Orel por estar siempre dispuestos a darme una mano. Quiero agradecer a mis colegas y amigos de la Universidad de La Plata, Fernando, Martín y Xabier por todos los momentos compartidos; y a los amigos que hice a lo largo de estos años en mis visitas al Laboratorio CERN, Dodo, Laura, Lucile, Bárbara, Teresa, Manouk, Alex, Olivier y Haris, por ser mi familia en la distancia.

Agradezco profundamente a mis amigos y a toda mi familia por su apoyo y aliento; y de manera especial a mamá y a Juan, por comprenderme y acompañarme en todo. A ellos les dedico esta tesis.

Identificación de jets con hadrones *b* producidos por desdoblamiento de gluones con el detector ATLAS.

Resumen

En esta tesis se presenta un estudio de la subestructura de jets que contienen hadrones b con el propósito de distinguir entre jets-b genuinos, donde el quark b se origina a nivel de elemento de matriz (por ejemplo, en decaimientos de top, W, o Higgs) y jets-b producidos en la lluvia partónica de QCD, por el desdoblamiento de un gluón en un quark y un antiquark b cercanos entre sí. La posibilidad de rechazar jets-b producidos por gluones es importante para reducir el fondo de QCD en análisis de física dentro del Modelo Estándar, y en la búsqueda de canales de nueva física que involucran quarks b en el estado final. A tal efecto, se diseñó una técnica de separación que explota las diferencias cinemáticas y topológicas entre ambos tipos de jets-b. Esta se basa en observables sensibles a la estructura interna de los jets, construídos a partir de trazas asociadas a éstos y combinados en un análisis de multivariable. En eventos simulados, el algoritmo rechaza 95% (50%) de jets con dos hadrones b mientras que retiene el 50% (90%) de los jets-b genuinos, aunque los valores exactos dependen de p_T, el momento transverso del jet. El método desarrollado se aplica para medir la fracción de jets con dos hadrones b en función del p_T del jet, con 4,7 fb^1 de datos de colisiones pp a $\sqrt{s} = 7$ TeV, recogidos por el experimento ATLAS en el Gran Colisionador de Hadrones en 2011.

Palabras clave: Experimento ATLAS, Jets, Subestructura de Jets, QCD, Producción de jets b, Etiquetado de Jets b.

Identification of double b-hadron jets from gluon-splitting with the ATLAS Detector.

Abstract

This thesis presents a study of the substructure of jets containing b-hadrons with the purpose of distinguishing between "single" b-jets, where the b-quark originates at the matrix-element level of a physical process (e.g. top, W or Higgs decay) and "merged" b-jets, produced in the parton shower QCD splitting of a gluon into a collimated b quark-antiquark pair. The ability to reject b-jets from gluon splitting is important to reduce the QCD background in Standard Model analyses and in new physics searches that rely on b-quarks in the final state. A separation technique has been designed that exploits the kinematic and topological differences between both kinds of b-jets using track-based jet shape and jet substructure variables combined in a multivariate likelihood analysis. In simulated events, the algorithm rejects 95% (50%) of merged b-jets while retaining 50% (90%) of the single b-jets, although the exact values depend on p_T, the jet transverse momentum. The method developed is applied to measure the fraction of double b-hadron jets as a function of jet p_T, using 4.7 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV collected by the ATLAS experiment at the Large Hadron Collider in 2011.

Keywords: ATLAS Experiment, Jets, Jet Substructure, b-jet Production, QCD, Gluon Splitting, b-tagging.

Contents

Chapter 1

The Multivariate Analysis

After the evaluation of the best discriminating variables, a tagging algorithm, capable of efficiently identifying single b-jets while rejecting merged b-jets, can be constructed using several different approaches. In this chapter we present the results of a multivariate likelihood method. Compared to single variable or 2-dimensional cuts analyses, with this technique the sensitivity is largely improved because several variables are combined to achieve the maximum separation power.

1.1 Multivariate methods

Multivariate data analysis refers to a statistical technique used to analyze data that is composed of more than one variable. Classification is done through learning algorithms that make use of training events, for which the desired output is known, to determine the mapping function that discribes a decision boundary. The following multivariate methods were explored:

- Likelihood ratio estimators (LLR)
- Neural Networks (NN)

- Boosted decision Trees (BDTs)

The method of LLR consists of building a model out of probability density functions (PDF) that reproduces the distributions of the input variables for signal and background. The likelihood ratio $y_L(i)$ for event i is defined by:

$$y_L(i) = \frac{L_S(i)}{L_S(i) + L_B(i)}, \tag{1.1}$$

In the case of poorly correlated variables, the likelihood for being of signal type is obtained by multiplying the signal probability densities of all input variables and normalising this by the sum of the signal and background likelihoods,

$$L_{S(B)}(i) = \prod_{k=1}^{n_{var}} p_{S(B),k}(x_k(i)), \tag{1.2}$$

Because correlations among the variables are ignored, this PDE approach is also called âĂIJnaive Bayes estimatorâĂİ. and where $p_{S(B),k}(x_k(i))$ is the signal (background) PDF for the kth input variable x_k. All the PDFs are normalized to one.

The parametric form of the PDFs is generally unknown, however it is possible to empirically approximate its shape by nonparametric functions. Nonparametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from the data sample used for training. A histogram is a simple example of a nonparametric estimate of a probability distribution. The nonparametric functions can be chosen individually for each variable and can be either polynomial splines of various degrees fitted to binned histograms[1] or unbinned kernel density estimators (KDE). The basic idea in the KDE approach is to model the distribution of

[1] A spline is a sufficiently smooth polynomial function that is piecewise-defined, and possesses a high degree of smoothness at the places where the polynomial pieces connect. It is often referred to as polynomial interpolation.

a sample of events as the sum of individual equal-area Gaussian kernels for each event For a PDF $p(x)$ of a variable x, one finds [1].

$$p(x) = \frac{1}{Nh}\sum_{i=i}^{N} K(\frac{x - x_i}{h}) = \frac{1}{N}\sum_{i=i}^{N} K_h(x - x_i), \tag{1.3}$$

where N is the number of training events, $K_h(t) = K(t/h)h$ is the kernel function, and h is the bandwidth of the kernel (also termed the *smoothing parameter*). For the present implementation a Gaussian form of K is used.

The smoothness of the kernel density estimate is evident compared to the discreteness of a histogram; kernel density estimates converge faster to the true underlying density for continuous random variables.

An artificial Neural Network (NN) is a nonlinear discriminant. It is, most generally speaking, a simulated collection of interconnected neurons, with each neuron producing a certain response at a given set of input signals. It can be viewed as a mapping from a space of input variables $x_1, ..., x_{n_{var}}$ onto, in the case of a signal-versus-background discimination problem, a one-dimensional output variable. The behaviour of an artificial neural network is determined by the layout of the neurons, the weights of the inter-neuron connections, and by the response of the neurons to the input, described by the neuron response function. The neuron response function maps the neuron input (in R^n) onto the neuron output (R); often it can be separated into a synapse function ($R^n \rightarrow R$) and a neuron activation function ($R \rightarrow R$). The neuron activation function can be a either a *linear*, *sigmoid*, *tanh*, or a *radial* function.

While in principle a neural network with n neurons can have n^2 directional connections, the complexity can be reduced by organising the neurons in layers and only allowing direct connections from a given layer to the following one. This kind of neural network is termed multi-layer perceptron. The first layer of a multilayer perceptron is the input layer, the last one the output layer, and all others are hidden layers. For a classification problem with n_{var} input variables the input layer consists of n_{var} neurons that hold the input values, $x_1, ..., x_{n_{var}}$, and one neuron in the output layer that holds the output variable, the neural net estimator y_{NN} .

A decision tree (BDT) is a binary tree structured classifier similar to the one sketched in Fig. 1.1. Repeated yes/no decisions are taken on one single variable at a time until a stop criterion is fulfilled. The phase space is split this way into many regions that are eventually identified as "signal-like" or "background-like", depending on the majority of training events that end up in the final *leaf* node. The boosting of a decision tree extends this concept from one tree to several trees which form a *forest.* Boosting increases the statistical stability of the classifier and typically also improves the separation performance compared to a single decision tree [2].

Training and testing the MVA methods

A sub-set of the dijet Monte Carlo sample was used for training the three methods in the context of the Toolkit for Multivariate Data Analysis, TMVA [3], written in C++ language. After the event and jet selections were performed, the b-tagged jets with $|\eta| < 2.1$ were classified as signal (single b-jets) or background (merged b).

Based on discrimination power, correlation and pile-up dependence three variables were selected for the training: the jet track multiplicity, the track-jet width and the ΔR between the axes of two k_t subjets in the jet. Event samples of single and merged b-jets, with the information of the value of the selected variables plus their correlations were given to the multivariate methods as input. Other variables such as τ_2 or $\max\{\Delta R(trk, trk)\}$ were

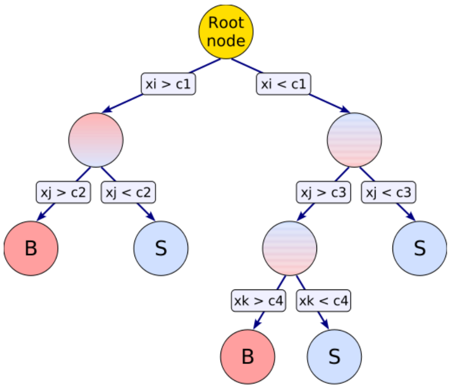

Figure 1.1: Schematic view of a decision tree. Starting from the *root node*, a sequence of binary splits using the discriminating variable x_i is applied to the data. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes. Image taken from Reference [3].

also tested leading to no gain in performance.

Different training options were evaluated for the likelihood and Neural Network classifiers. The final configuration for the likelihood estimator uses the KDE approach for PDF estimation of the input variables. The NN trained is a Multi-layer perceptron (MLP) with two hidden layer of n_{var} and $n_{var} - 1$ neurons respectively, using a sigmoidal neuron activation function. The Boosted Decision Tree approach was implemented with 400 trees in the BDT forest. Distributions of the final output for the three methods, evaluated in an orthogonal sample of simulated dijet events, are shown in Fig. 1.3 for a medium p_T bin. The testing sample satisfies the same selection used for the training sample.

The outputs of the MVA discriminants explored are different in terms of shape and range, although the latter could be rearranged with a suitable

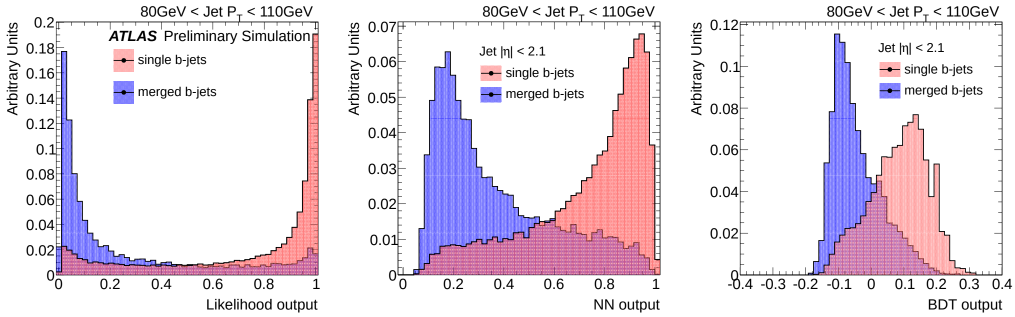

Figure 1.2: Distribution of the MVA discriminant outputs for the Likelihood (a), Neural Network (b) and Boosted Decision Trees (c) classifiers, for single and merged b-jets between 80 GeV and 110 GeV.

variable transformation. In spite of these distinct features the performances of the different methods, expressed in rejection of merged b-jets as a function of single b-jet efficiency (see next section), agrees within statistics, see Fig. 1.3.

As opposed to Neural Network discriminants with large number of training cycles, the training and the application of the likelihood are very fast operations that are suitable for very large data sets and tuning of the training parameters. Although also very fast, a shortcoming of decision trees is their instability with respect to statistical fluctuations in the training sample from which the tree structure is derived. If two input variables exhibit similar separation power, a fluctuation in the training sample may cause the tree growing algorithm to decide to split on one variable, while the other variable could have been selected without that fluctuation. In such a case the whole tree structure is altered below this node, possibly resulting also in a substantially different classifier response [3]. It is for these reasons that the likelihood classifier is the selected method for our tagger.

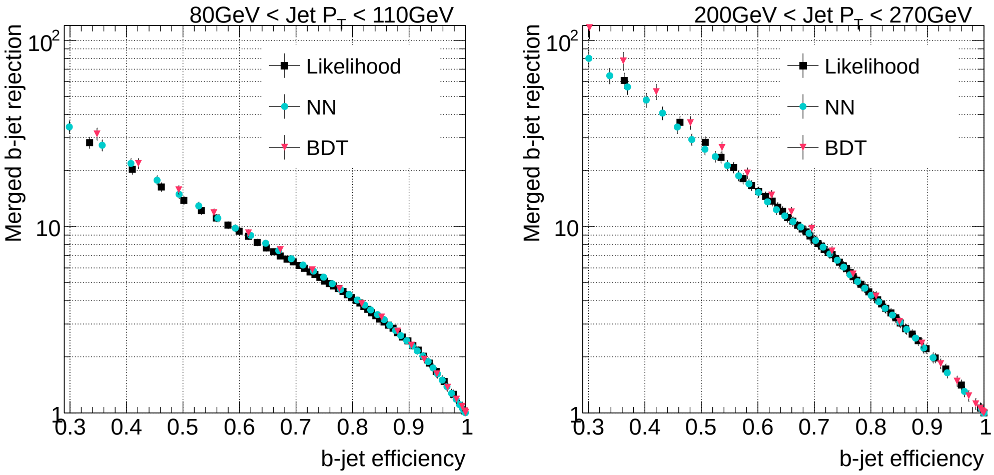

Figure 1.3: Rejection of merged b-jets as a function of single b-jet efficiency for the the different MVA methods evaluated for low and high jet p_T.

1.2 Likelihood training and performance

As indicated in section 1.1, a discriminant between single and merged b-jets was built by training a simple likelihood estimator, with the following three variables as input,

1. Jet track multiplicity
2. Track-jet width
3. ΔR between the axes of 2 k_t subjets within the jet

Given the correlation of the variables with the jet transverse momentum, the training sample was categorized in bins of calorimeter jet p_T, and independent likelihood classifiers were built for each category.

Due to the lack of statistics of merged jets in the low p_T bins, signal and background jets were not weighted by the dijet samples cross-sections to allow the contribution of subleading lower p_T jets from high p_T events. The gain in statistics in the first p_T bin was of more than 500%. It is important to

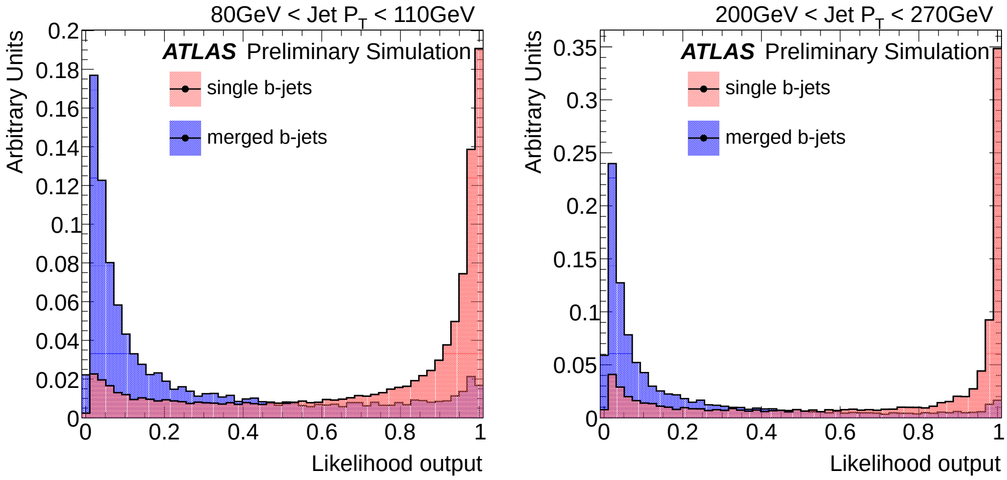

Figure 1.4: Distribution of the likelihood output for single and merged b-jets for medium and high p_T jets.

stress that, although essential for data to MC comparisons (see Section **??**), the weighting of the dijet Monte Carlo samples by their respective cross-sections is not necessary for studies performed at simulation level only. For the evaluation of the method the same procedure was followed.

The distribution of the likelihood output for single and merged b-jets is shown in Fig. 1.4 for low, medium and high transverse momentum jets.

The performance of the tagger in the simulation can be displayed in a plot of rejection of merged b-jets, $(1/\epsilon_{bkg})$, as a function of single b-jet efficiency, ϵ_{sig}; where ϵ_{bkg} (ϵ_{sig}) is the probability that a double (single) b-hadron jet passes the single b-jet tagger. This is shown in Fig. 1.5 for the eight bins of jet p_T mentioned in section **??**. The performance improves with p_T:

- $p_T > 40$ GeV: rejection above 8 at 50% eff.
- $p_T > 60$ GeV: rejection above 10 at 50% eff.
- $p_T > 200$ GeV: rejection above 30 at 50% eff.

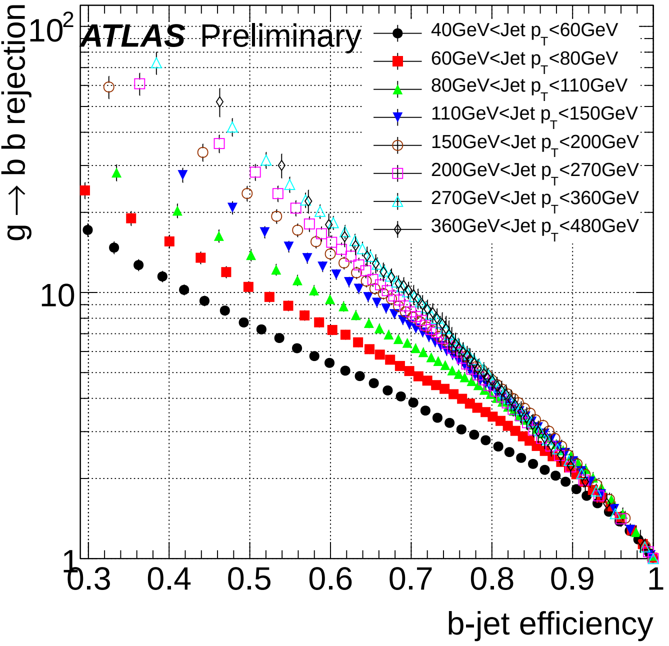

Figure 1.5: Rejection of merged b-jets as a function of single b-jet efficiency for dijet events in 8 jet p_T bins.

The rejection of merged jets attained as a function of p_T for the 50% and 60% single b-jet efficiency working points are summarized in Table 1.1, together with their relative statistical error. These are propagated from the Poisson fluctuations of the number of events in the merged and single b distributions. The error is slightly lower for the 60% efficiency working point because a higher efficiency allows for a greater number of Monte Carlo events to measure the performance.

Jet p_T	single b-jet efficiency 50%		single b-jet efficiency 60%	
(GeV)	Rejection	stat.err.	Rejection	stat.err.
40 - 60	8	4%	5	3%
60 - 80	10	4%	7	4%
80 - 110	14	5%	9	4%
110 - 150	19	5%	12	4%
150 - 200	23	5%	14	5%
200 - 270	30	7%	16	6%
270 - 360	36	7%	19	6%
360 - 480	41	8%	18	8%

Table 1.1: The merged b-jet rejection for the 50% and 60% efficiency working points in bins of p_T.

1.3 Systematic uncertainties

The development, training and performance determination of the tagger is based on simulated events. Although the agreement between simulation and data explored in section **??** is a necessary validation condition, it is also important to investigate how the tagger performance depends on the systematic precision with which the MC simulates the data. In particular we have considered:

- presence of additional interactions (pile-up);
- uncertainty in the b-jet tagging efficiency;
- uncertainty in the track reconstruction efficiency;
- uncertainty in the track transverse momentum resolution;
- uncertainty in the jet transverse momentum resolution;
- uncertainty in the jet energy scale.

I. Pile-up

The size of this effect was studied by comparing the performance of the likelihood discriminant with b-jets in events with small (1-9) and large (9-20) number of primary vertices. A comparison of the performance in these two sub-samples relative to the inclusive sample is shown in Fig. 1.6 for the two lowest p_T bins, where the effect of pile-up is more important. As expected from the use of tracking (as opposed to calorimeter) variables, no significant dependence with pile-up is observed. Performance differences between high and low number of primary vertices events are $\leq 2\%$. The impact of pile-up might be larger in 2012 data.

II. b-tagging efficiency

The performance of heavy-flavor tagging in Monte Carlo events is calibrated to experimental data by means of the scale factors (SFs). The SFs are defined as the ratio of the heavy-flavor tagging efficiency in data over that in Monte Carlo for the different jet flavors. They are measured by the ATLAS

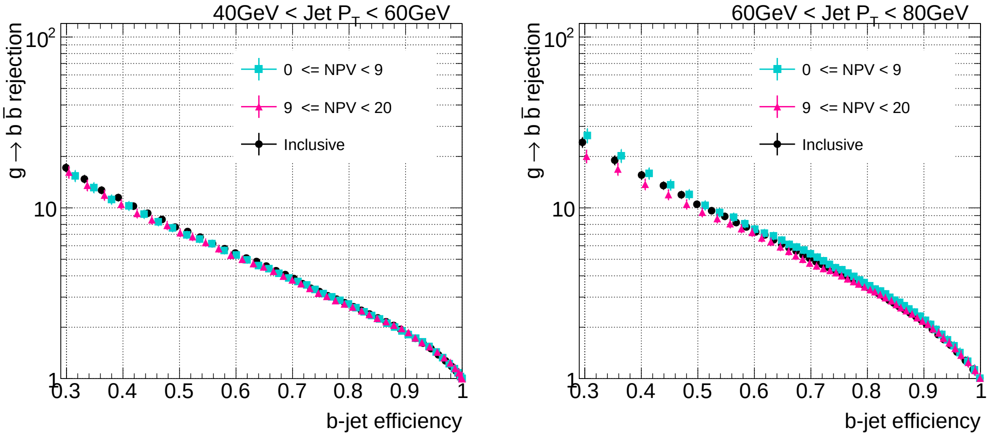

Figure 1.6: Rejection of merged b-jets as a function of single b-jet efficiency in bins of NPV for two low jet p_T bins.

Flavour Tagging Working group, and their measurement carries a systematic uncertainty.

To estimate the impact of this uncertainty a conservative approach is followed: the SFs are varied in all the p_T bins simultaneously by one standard deviation both in the up and down directions. The MC distributions weighted by the varied SFs show no major deviations from the nominal, see Fig. 1.7. In the same manner, the effect of the b-tagging calibration uncertainty on the likelihood peformance, shown in Fig. 1.8, is $< 1\%$, negligible with respect to the statistical uncertainty. This was indeed expected. The scale factors depend on the true flavor of the jet and on its p_T, but these are basically constant in the performance determination, which is based on single flavor (true b-) jets classified in p_T-bins.

III. Track reconstruction efficiency

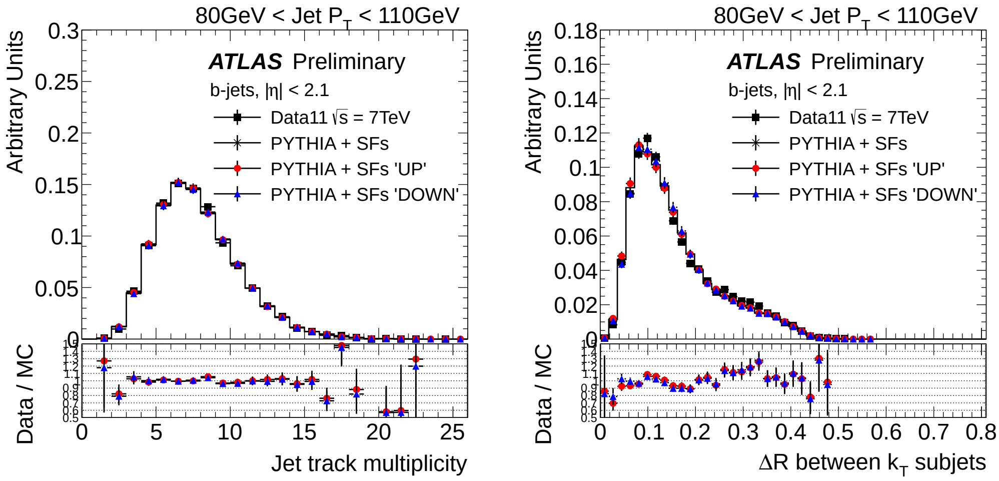

Figure 1.7: The effect of a variation in the b-tagging Scale Factors on the tracking variables distributions. Scale Factors were varied up (down) by 1-sigma to evaluate the systematic uncertainty from this source. The ratio data over MC is shown for MC PYTHIA with SFs varied up (circles) and down (triangles).

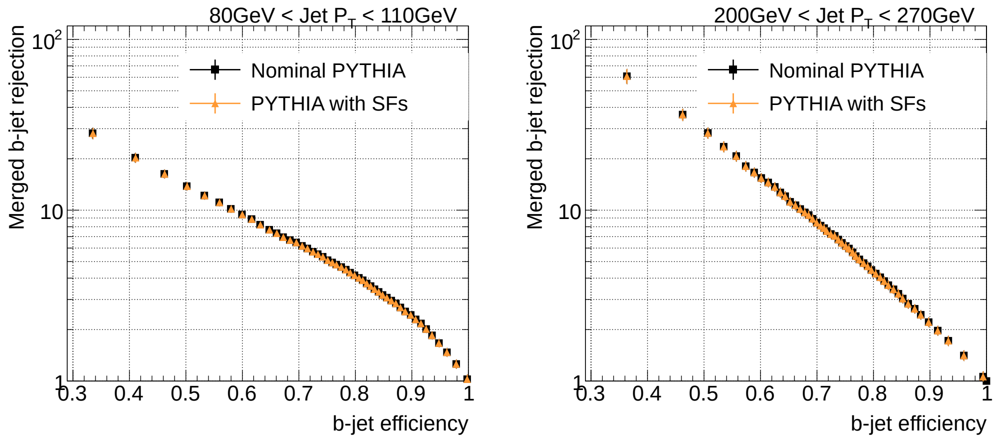

Figure 1.8: Rejection of merged b-jets as a function of single b-jet efficiency with and without scale factors as weights.

The track reconstruction efficiency, ϵ_{trk}, parametrised in bins of p_T and η, is defined as:

$$\epsilon_{trk} = \frac{N_{rec}^{matched}(p_T, \eta)}{N_{gen}(p_T, \eta)} \tag{1.4}$$

where $N_{rec}^{matched}$ is the number of reconstructed tracks matched to a generated charged particle, and $N_{gen}(p_T, \eta)$ is the number of generated charged particles in that bin[2]. As the track reconstruction efficiency is determined from MC, the main systematic uncertainty results from the level of agreement between data and MC. Since charged hadrons are known to suffer from hadronic interactions with the material in the detector, a good description of the material in MC is needed to get a good description of the track reconstruction efficiency. The uncertain knowledge of the material in the inner detector is the main source of systematic uncertainties in the tracking efficiency [4]. An increase (decrease) in material leads to an increase (decrease) in the number of hadronic interactions, hence to a decrease (increase) in the reconstruction efficiency.

The tracking efficiency systematics are given in bins of track η. For tracks with $p_{\mathrm{T}}^{\mathrm{track}} > 500$ MeV the uncertainties are independent of p_T: 2% for $|\eta^{\mathrm{track}}| < 1.3$, 3% for $1.3 < |\eta^{\mathrm{track}}| < 1.9$, 4% for $1.9 < |\eta^{\mathrm{track}}| < 2.1$, 4% for $2.1 < |\eta^{\mathrm{track}}| < 2.3$ and 7% for $2.3 < |\eta^{\mathrm{track}}| < 2.5$ [4]. All numbers are relative to the corresponding tracking efficiencies.

To test the impact of these uncertainties, a fraction of tracks determined from the track efficiency uncertainty was randomly removed. The tracking variables were re-calculated and the performance of the nominal likelihood was evaluated in the new sample with worse tracking efficiency. The

[2] The matching between a generated particle and a reconstructed track uses a cone-matching algorithm, associating the particle to the track with the smallest ΔR within a cone of radius 0.15.

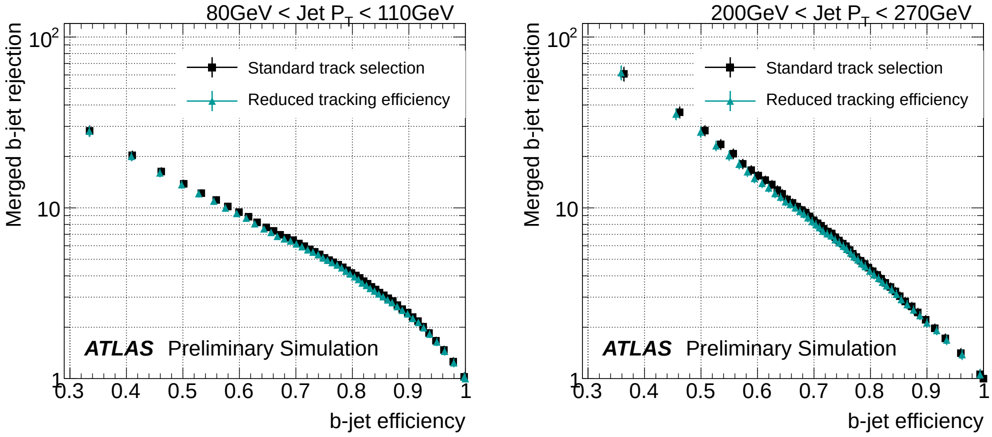

Figure 1.9: Rejection of merged b-jets as a function of the single b-jet efficiency showing shift in likelihood performance caused by a reduction in the tracking efficiency.

rejection-efficiency curves show a small degradation of the performance which is comparable to the statistical uncertainty. The effect is however systematically present over all 16 p_T bin/working points, without a clear p_T dependence. We have thus taken the average over p_T, and obtained a global systematic uncertainty of 4% both for the 50% and 60% efficiency working points. The performance comparison is shown in Fig. 1.9 for two p_T bins.

IV. Track momentum resolution

The knowledge of the track momentum resolution is limited by the precision both in the material description of the Inner Detector and in the mapping of the magnetic field. Its uncertainty propagates to the kinematic variables used in the double b-hadron jet tagger. In order to study this effect, track momenta are over-smeared according to the measured resolution uncertainties, before the track selection cuts are applied. The actual smearing is done in $1/p_T$, with an upper bound to the resolution uncertainty given by $\sigma(1/p_T) = 0.02/p_T$ [5].

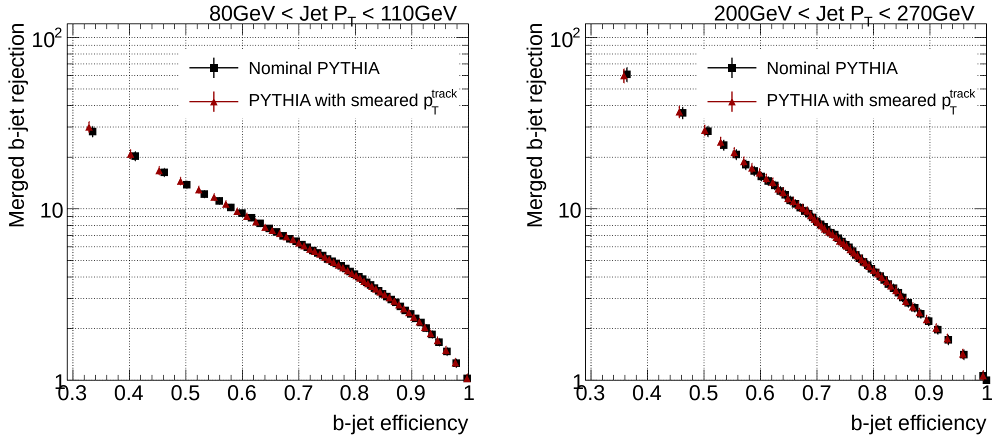

Figure 1.10: Rejection of merged b-jets as a function of the single b-jet efficiency showing the effect of the track momentum resolution uncertainty. It is found to be negligible with respect to the statistical uncertainty.

The effect is found to be negligible, see Fig 1.10.

V. Jet energy scale and momentum resolution

The jet energy scale (JES) uncertainty for light jets reconstructed with the anti-k_t algorithm with distance parameter $R = 0.4$ and calibrated to the EM+JES scale is between ∼4% at low p_T and ∼2.5% for jets with p_T >60 GeV in the central region [6]. In the case of b-jets, and additional uncertainty arising from the modelling of the b-quark production mechanism and the b-quark fragmentation was determined from systematic variations of the Monte Carlo simulation. The resulting fractional additional JES uncertainty for b-jets has an upper bound of 2% for jets with p_T ≤100 GeV and it is below 1% for higher p_T jets. To obtain the overall b-jet uncertainty this needs to be added in quadrature to the light JES uncertainty.

The systematic uncertainty originating from the jet energy scale is obtained by scaling the p_T of each jet in the simulation up and down by one

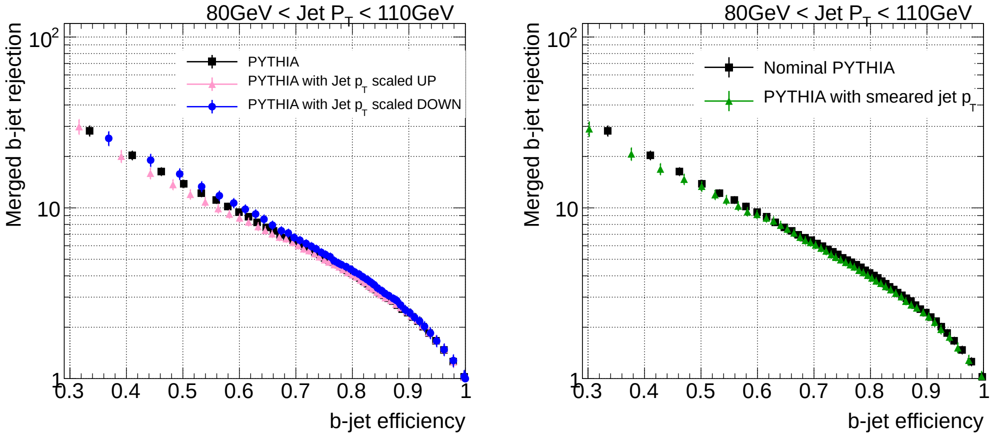

Figure 1.11: Rejection of merged b-jets as a function of single b-jet efficiency for (a) jets with smeared p_T and (b) for jets with varied energy scale compared to nominal.

standard deviation according to the uncertainty of the JES. The result is shown in Fig. 1.11a for a medium p_T bin. The effect on the likelihood performance is an average variation of 5% for the 50% and 60% efficiency working points.

The jet momentum resolution was measured for 2011 data and found to be in agreement with the predictions from the PYTHIA-based simulation [7]. The precision of this measurement, determined in p_T and η bins, is typically 10%. The systematic uncertainty due to the calorimeter jet p_T resolution was estimated by over-smearing the jet 4-momentum in the simulated data, without changing jet η or ϕ angles. The performance, shown in Fig. 1.11b, is found to globally decrease by 5%, without a particular p_T dependence.

The different contributions to the systematic uncertainty on the merged b-jet rejection are summarized in Table 1.2.

Although the likelihood training was peformed in EM+JES calibrated jets, the performance of the tagger was also evaluated in jets calibrated with

Systematic source	Uncertainty
pile-up	2%
b-tagging efficiency	neglible
track reconstruction efficiency	4%
track p_T resolution	neglible
jet p_T resolution	5%
jet energy scale	5%

Table 1.2: Systematic uncertainties in the merged b-jet rejection (common to both the 50% and the 60% efficiency working points).

the LC+JES scheme, described in Section **??**. A small degradation of the performance is observed, but comparable with the statistical uncertanties. A comparison of the performances is shown in Fig. 1.12 for two p_T bins, representative of the jet momentum range covered.

1.4 Other Monte Carlo generators

The development, training and performance determination of the tagger has been done using Monte Carlo events generated with the PYTHIA event simulator, interfaced to the GEANT4 based simulation of the ATLAS detector. An immediate question is what the performance would be if studied with a different simulation. In this section we investigate this question for the PYTHIA Perugia tune and the HERWIG++ event generators (Section **??**).

Fig. 1.13 shows a comparison of the likelihood rejection, at the 50% efficiency working point, between nominal PYTHIA and the alternative simulations as a function of the jet p_T . The larger errors are due to the reduced

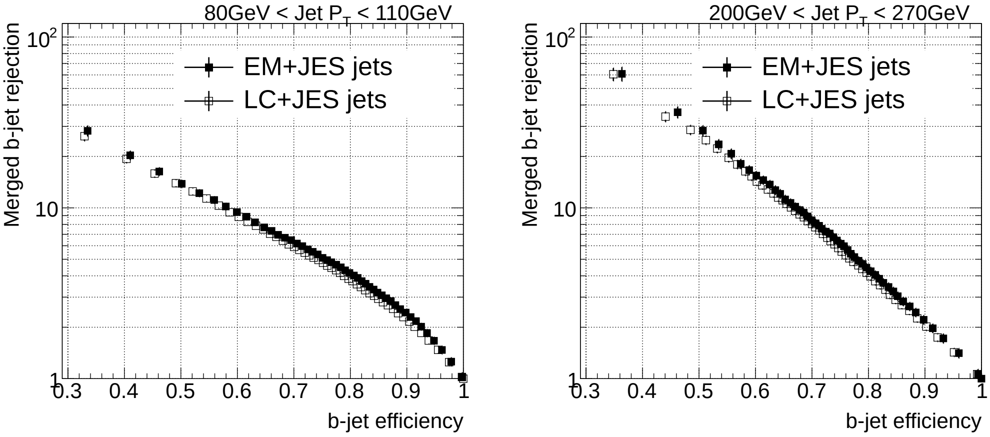

Figure 1.12: Rejection of merged b-jets as a function of single b-jet efficiency for jets calibrated to the EM+JES (LC+JES) scale, between 80 GeV and 110 GeV and 200 GeV and 270 GeV.

statistics available, which are even lower for the Perugia case than for HERWIG.

The performance in HERWIG shows a systematic trend, with agreement at low p_T and increasingly poorer performances compared to PYTHIA as p_T grows. For the Perugia tune, on the other hand, there is no definite behavior, with the performance fluctuating above or below the nominal simulation for different p_T bins consistently with the statistical uncertainties.

The reason for the systematic difference observed between the performances of PYTHIA and HERWIG can be traced to the extent with which jets are accurately modelled. Fig. 1.14 compares the measured jet track multiplicity distributions in b-tagged jets and the prediction from both simulations, for low and high p_T jets. It is observed that indeed HERWIG++ does not correctly reproduce the data, particularly at high p_T. The level of agreement is found to be better for track-jet width and the ΔR between the axes of the two k_t subjets in the jet, the two other variables used for discrimination.

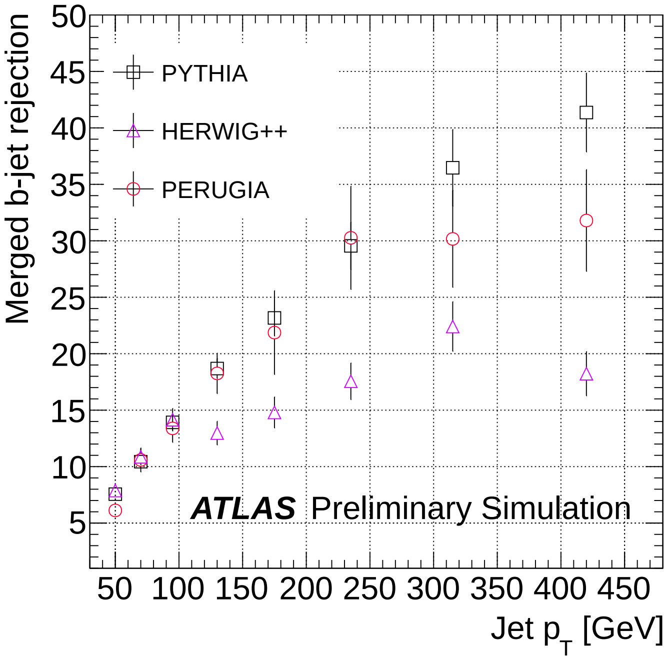

Figure 1.13: Rejection of merged b-jets as a function of jet p_T for different Monte Carlo generators, at the 50% efficiency working point.

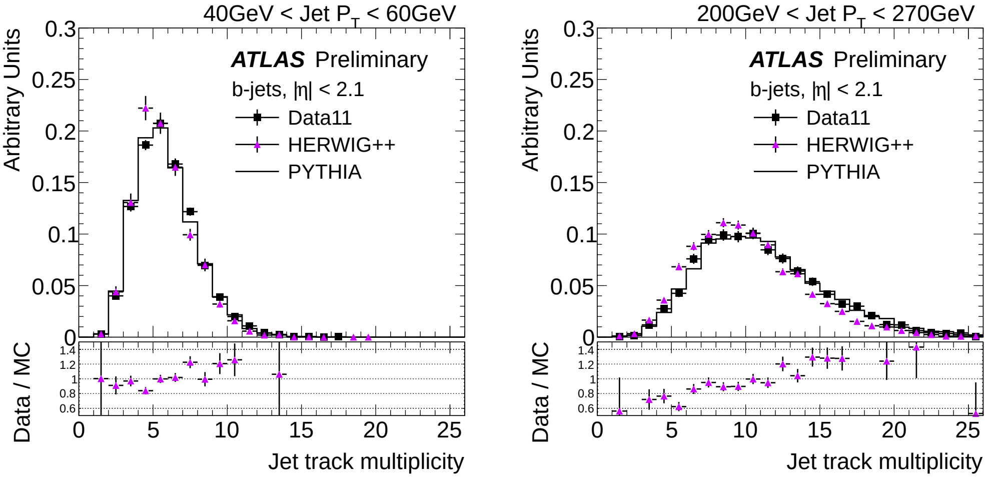

Figure 1.14: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected during 2011 (solid black points) and HERWIG++ events (solid violet triangules). The ratio data over HERWIG++ simulation is shown at the bottom of the plot. PYTHIA distribution is also shown for reference.

Chapter 2

Fraction of double *b*-hadron jets in data

2.1 Maximum likelihood fits

Maximum likelihood (ML) information only parametrizes the shape of a distribution (i.e. one can determine fraction of signal events from MC fits but no number of signal events).

The extended version of the maximum likelihood approach adds an extra term allowing the estimation of the absolute number of signal/background events.

For N p.d.f.s, there are N-1 fraction coefficients that should sum to less 1. The remainder is by construction 1 minus the sum of all other coefficients.

Binned or unbinned MF fit.

In most RooFit applications it doesn't matter. Internally binned data is represented the same way as unbinned data, a ROOT TTree with the bin coordinates.

Weights are supported in unbinned datasets. But use with care. Error analysis in ML fits to weighted unbinned data can be complicated...

Unbinned fits

Fitting and likelihood minimization

What happens when you do pdf->fitTo(*data)?

- Construct object representing -log of (extended) likelihood
- Minimize likelihood w.r.t floating parameters using MINUIT.

2.2 Results

The likelihood template fits are performed using an unbinned maximum likelihood technique.

The likelihood Monte Carlo templates were derived from the simulated dijet sample described in Section **??**.

Templates of likelihood were constructed for b, c, $b\bar{b}$, $c\bar{c}$ and light MV1 tagged jets separately, and these were fit to the likelihood distribution in data to obtain the fraction of single b, merged b, single c, merged c and light jets in the data sample. The fit is done by adjusting the relative constributions of the different templates such that the sum of them best describes the likelihood shape in data. A separate fit was performed for each p_T-bin. The templates were derived from all jets passing the selection criteria defined in Section **??**.

Different combinations of templates were used to fit the likelihood distribution in data. The sensitivity of the fit to fixing the ratio of the c- to b-flavour fractions to the value extracted from the simulation was investigated by carrying out separate fits with 5 and 3 free parameters. This was motivated by the fact that templates for single c- (merged c-) and single b-jets (merged b-jets) look very similar leading to inestabilities in the fitted b-

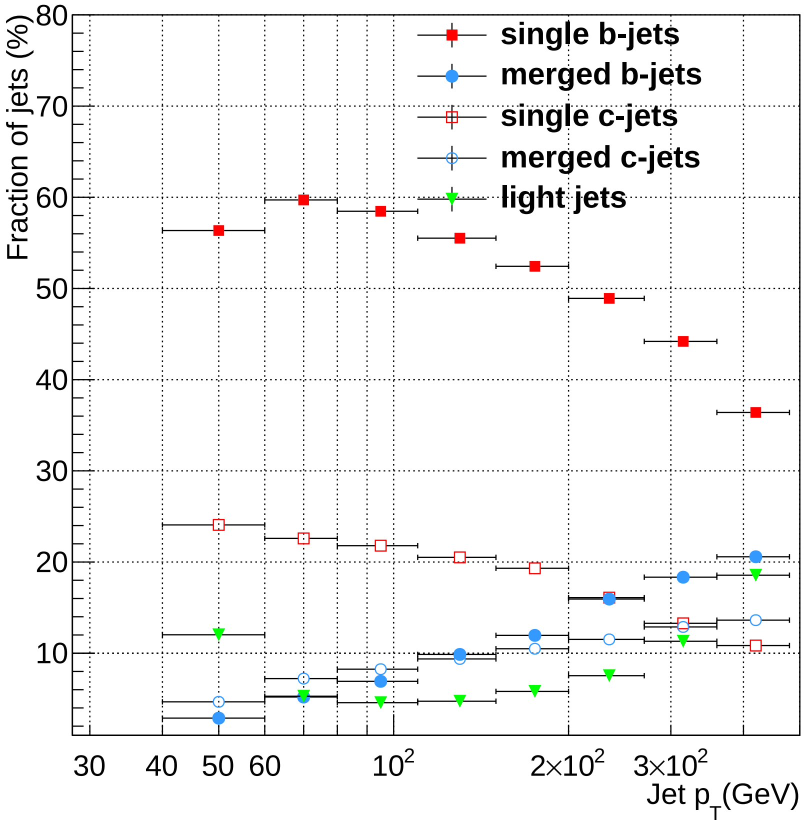

Figure 2.1: Pythia predictions of the fractions of MV1 tagged b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets in a Monte Carlo dijet sample.

and c-flavour fractions.

True fractions derived in Pythia simulation as a functin of the jet p_T is shown in Fig. 2.1.

The results of the 3-parameter fits for all bins of p_T are shown in table 2.1. The fit results are shown in Figures 2.2 and 2.3.

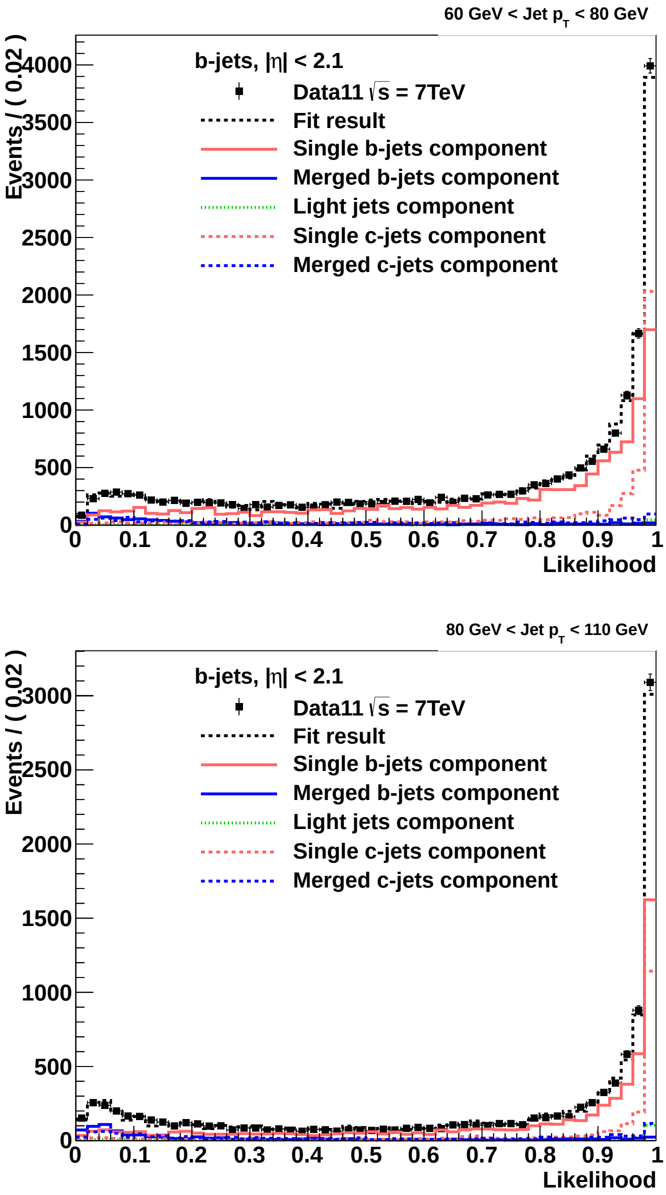

Figure 2.2: The results of template fits to the likelihood distribution in data. The fits shown here were performed on jets with p_T between 60 GeV and 80 GeV, and 80 GeV and 110 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

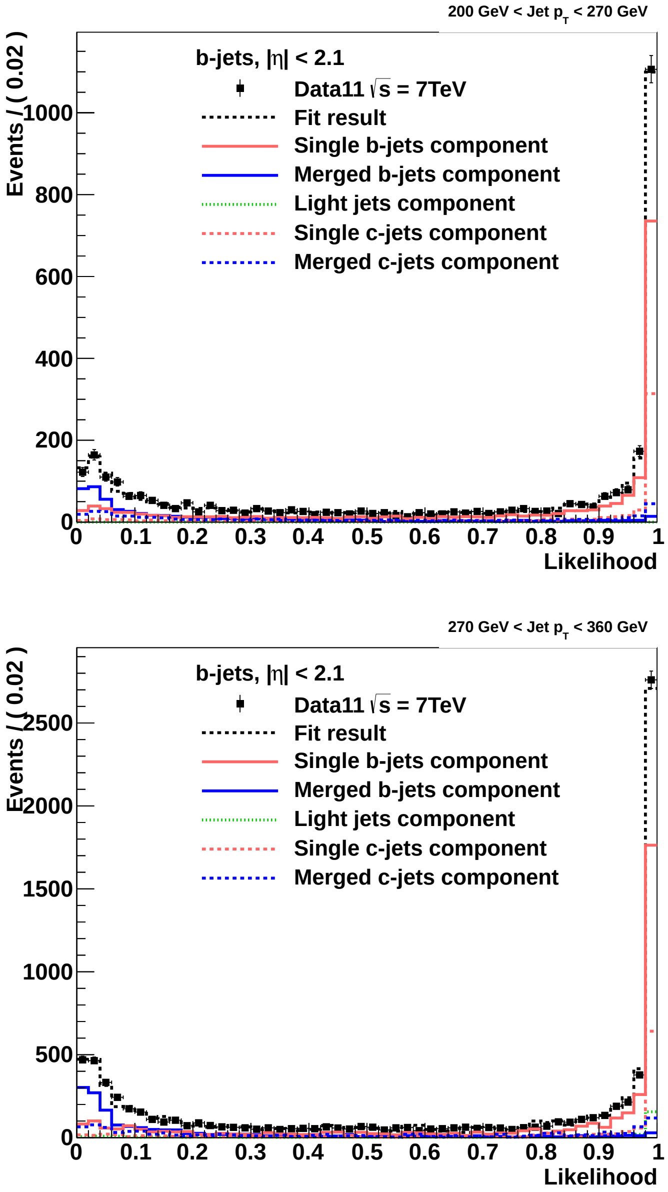

Figure 2.3: The results of template fits to the likelihood distribution in data. The fits shown here were performed on jets with p_T between 200 GeV and 270 GeV, and 270 GeV and 360 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

Jet p_T	single b-jet		merged b-jet		light jet	
(GeV)	fit result	stat.err.	fit result	stat.err.	fit result	stat.err.
40 - 60	62%	3%	3%	1%	4%	4%
60 - 80	62%	1%	5.2%	0.4%	2%	2%
80 - 110	57%	1%	8.5%	0.4%	3%	2%
110 - 150	55%	2%	13%	1%	1%	4%
150 - 200	53%	3%	15%	1%	0%	4%
200 - 270	53%	5%	17%	1%	-1%	7%
270 - 360	48%	3%	19%	1%	4%	4%
360 - 480	39%	5%	21%	1%	15%	6%

Table 2.1: Measured fractions of single, merged and light b-tagged jets in experimental data from 2011 run.

2.3 Systematic uncertainties

The systematic uncertainties affecting the method are mainly those that change the shape of the likelihood tamplates used to fit the sample composition. The following contributions were evaluated:

- uncertainty in the track reconstruction efficiency;
- uncertainty in the jet transverse momentum resolution **TO DO**;
- uncertainty in the jet energy scale.
- heavy flavor fraction.

The systematic uncertainty originating from the jet energy scale is obtained by scaling the p_T of each jet in the simulation up and down by one standard deviation, according to the uncertainty of the jet energy scale (see

Section 1.3), and redoing the likelihood fits on data with the modified b, c, $b\bar{b}$, $c\bar{c}$ and light templates.

The systematic uncertainty originating from the jet p_T resolution is obtained by smearing the calorimeter jet p_T in the simulation. The likelihood templates were rederived from this "smeared" sample, and the likelihood distribution in data was fit using these altered samples. The difference between the unsmeared and the smeared scenarios is taken as a systematic uncertainty.

Changing the ratio of merged c to merged b fraction in 20% only produced a marginal effect on the fit results. The total number of merged c plus merged b did not change showing that in reality we are measuring the fraction of merged $b + c$ together. The same result is expected if changing the single c/single b ratio.

The systematic uncertainties are summarized in Table 2.2. The largest ones arise from the jet energy scale and jet transverse momentum resolution.

Systematic source	Uncertainty
track reconstruction efficiency	negligible%
jet p_T resolution	2%
jet energy scale	2%
flavour fractions	negligible

Table 2.2: Systematic uncertainties.

2.4 Enriched samples in single and merged b-jets

Enriched sample in merged b-jets

Enriched sample in single b-jets

The results of performing the fits on an data sample enriched in single b-jets is shown in tables 2.3 to 2.5. The model fitted to the data agrees well within statistics and the result is in agreement with the predictions made by PYTHIA on a sample with the same level o of enrichment.

The fit results are shown in Figures 2.4 and Figures 2.5.

Jet p_T	single b-jet		
(GeV)	fit result	stat.err.	pythia prediction
40 - 60	99%	11%	84%
60 - 80	82%	5%	87%
80 - 110	84%	5%	88%
110 - 150	86%	8%	85%
150 - 200	89%	9%	83%
200 - 270	95%	15%	80%
270 - 360	67%	11%	81%
360 - 480	73%	16%	73%

Table 2.3: Measured fractions of single b-jets in experimental data from 2011 run, enriched in single b-jets.

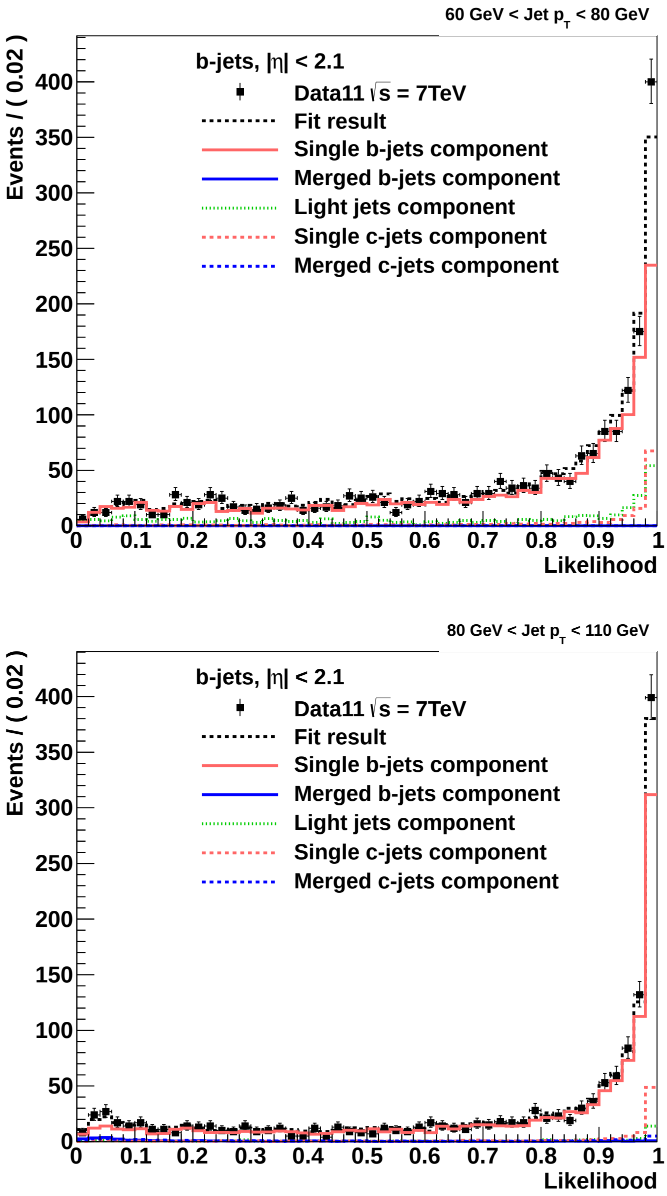

Figure 2.4: The results of template fits to the likelihood distribution in data enriched in single b-jets. The fits shown here were performed on jets with p_T between 60 GeV and 80 GeV, and 80 GeV and 110 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

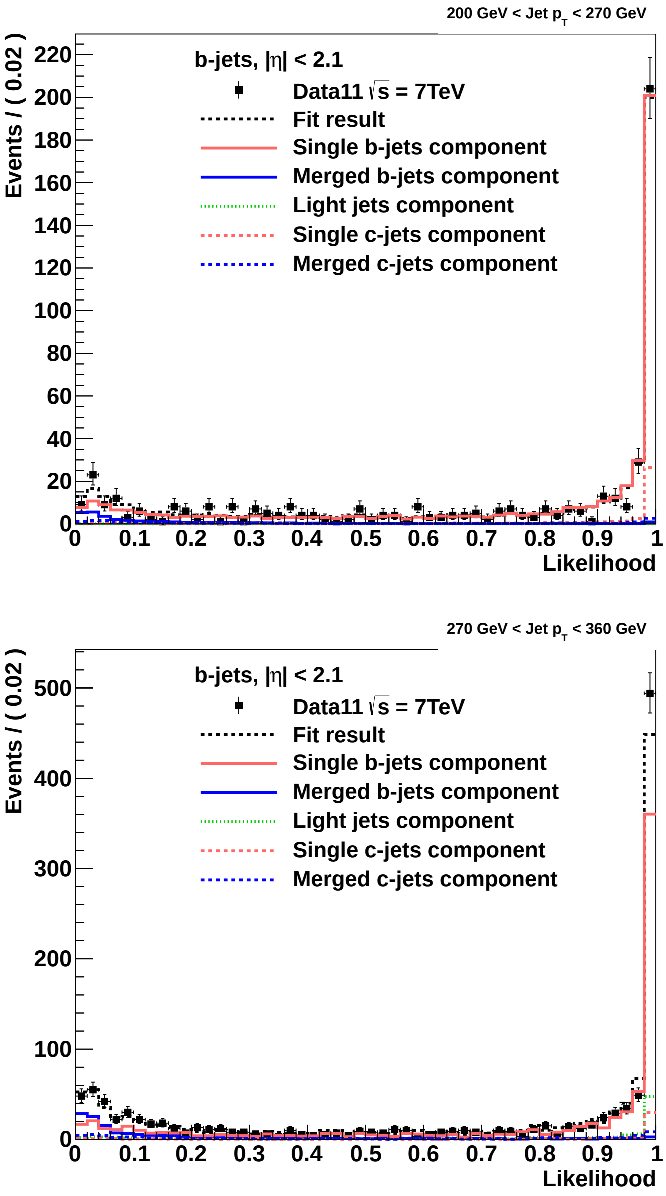

Figure 2.5: The results of template fits to the likelihood distribution in data enriched in single b-jets. The fits shown here were performed on jets with p_T between 200 GeV and 270 GeV, and 270 GeV and 360 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

Jet p_T	merged b-jet		
(GeV)	fit result	stat.err.	pythia prediction
40 - 60	-1%	1%	1%
60 - 80	-3%	1%	1%
80 - 110	2%	1%	1%
110 - 150	4%	2%	3%
150 - 200	4%	2%	3%
200 - 270	7%	2%	5%
270 - 360	12%	2%	6%
360 - 480	10%	1%	8%

Table 2.4: Measured fractions of merged b-jets in experimental data from 2011 run, enriched in single b-jets.

Jet p_T	light b-jet		
(GeV)	fit result	stat.err.	pythia prediction
40 - 60	-7%	11%	5%
60 - 80	17%	6%	2%
80 - 110	4%	6%	1%
110 - 150	-1%	9%	1%
150 - 200	-6%	10%	2%
200 - 270	-17%	17%	3%
270 - 360	9%	11%	4%
360 - 480	4%	16%	8%

Table 2.5: Measured fractions of light b-jets in experimental data from 2011 run, enriched in single b-jets.

Bibliography

[1] Scott, D.W. Multivariate Density Estimation: Theory, Practice, and Visualization. *John Wiley and Sons, Inc. United States*, 1992.

[2] Quinlan, J.R. Simplifying decision trees. *International Journal of Human-Computer Studies*, 51(2):497 – 510, 1999.

[3] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. Von Toerne, and H. Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[4] G. Aad et al. Charged-particle multiplicities in pp interactions measured with the ATLAS detector at the LHC. *New J.Phys.*, 13:053033, 2011.

[5] ATLAS Collaboration. Estimating Track Momentum Resolution in Minimum Bias Events using Simulation and K_s in $\sqrt{s} = 900$ GeV collision data. *ATLAS-CONF-2010-009*, 2010.

[6] Aad, G. and others. Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV. 2011.

[7] G. Romeo, A. Schwartzman, R. Piegaia, T. Carli, and R. Teuscher. Jet energy resolution from in-situ techniques with the atlas detector using proton-proton collisions at a center of mass energy $\sqrt{s} = 7$ tev. *ATL-COM-PHYS-2011-240*, 2011.