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

1 Fraction of double b-hadron jets in data **2**

1.1 Unbinned maximum likelihood fits 2

1.2 Fitting MC templates to data 6

1.3 Systematic uncertainties 11

1.4 Enriched samples in single and merged b-jets 12

Chapter 1

Fraction of double b-hadron jets in data

1.1 Unbinned maximum likelihood fits

The analysis of experimental data often involves the estimation of the composition of a sample, based on Monte Carlo simulations of the various sources.We measure a number of observables x_i and we want to determine one or more parameters p_i from the data, such as the number of signal and background events. The distribution of the observables is described by a probability density function (PDF), which is a function of both the observables and the parameters, $F(\vec{x}, \vec{p})$. We choose the PDF based on some guess about what function would match the data, and vary the parameters in order to make the PDF match the distribution of the observables as well as possible.

In the case of binned data into a histogram, one approch is to use a least-squares fitting technique to estimate the parameters. They are adjust to minimize

$$\chi^2 = \sum_i \frac{(d_i - f_i)^2}{d_i} \tag{1.1}$$

where d_i is the number of events in the real data that fall into bin i, and f_i the predicted number of events in bin i, defined by

$$f_i = \sum_{j=1}^{m} p_j \cdot a_{ji} \tag{1.2}$$

with p_j the proportions of the different m sources, normalized to the total number of events in the data and a_{ij} the number of Monte Carlo events from source j in bin i, with $i = 1, 2, ..., n$.

This χ^2 assumes that the distribution for d_i is Gaussian; it is of course Poisson, but the Gaussian is a good approximation to the Poisson for large numbers. Unfortunately it often happens that many of the d_i are small, making the χ^2 value given in Equation 1.1 inappropiate to describe the problem. Instead one can go back to the original Poisson distribution, and write down the probability for observing a particular d_i as

$$e^{f_i} \frac{f_i^{d_i}}{d_i!} \tag{1.3}$$

and the estimates of the proportions p_j are found by maximizing the total likelihood,

$$\mathcal{L} = \prod_{i=1}^{n} e^{f_i} \frac{f_i^{d_i}}{d_i!}. \tag{1.4}$$

This accounts correctly for the small numbers of data events in the bins. It is often referred to as a "binned maximum likelihood" fit[1].

[1]This formalism does not account for fluctuations in the a_{ji} due to finite Monte Carlo samples. A similar methodology that correctly describes this scenario exist, see Ref. [1]. The effects of finite MC data size can be considered small for MC samples ten times larger than the data sample.

The binned maximum likelihood fits is a technique in general use. Unfortunately this method does not behave well in problems where it is necessary to apply weights to the Monte Carlo, such as in our composite dijet sample. We will use instead a different technique for fitting, an "unbinned maximum likelihood fit", which does support weighted datasets.

The likelihood to be maximize in an unbinned dataset of events $\{x_k\}_{k=1}^{N}$ is

$$\tilde{\mathcal{L}}(\vec{x};\vec{p}) = \prod_{k=1}^{N} F(\vec{x};\vec{p}) \tag{1.5}$$

which, can be rewritten in terms of the probability of observing an event from source j in the sample,

$$\mathcal{L} = \prod_{k=1}^{N}\sum_{j=1}^{m} n_j \mathcal{P}_j \tag{1.6}$$

where $\mathcal{P}_j$ are the PDFs that represent the total probability for each of the m hypothesis, n_j is the number of events for the j^{th} hypothesis, and N is the total number of input data points.

Maximum likelihood information only parametrizes the shape of a distribution; that is, 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 a parameter that represents the number of events in the sample, N_{exp}. The extra term describes the probability of observing the actual number of events, N_{obs}, given this parameter. This probability is described by the Poisson distribution

$$P(N_{obs}, N_{exp}) \sim N_{exp}^{N_{obs}} \cdot e^{-N_{exp}}, \tag{1.7}$$

and we refer to the likelihood including this factor as the "extended likelihood"

$$\tilde{\mathcal{L}}(\vec{x}, N_{obs}; \vec{p}, N_{exp}) \equiv P(N_{obs}, N_{exp}) \cdot \mathcal{L}(\vec{x};\vec{p}) \tag{1.8}$$

The fit then finds the values of n_j, the number of events for each hypothesis j.

The fits were performed in this thesis by means of the RooFit Toolkit for data modelling [2]. Performing a fit consists of minimizing the negative log-likelihood of a PDF calculated over the data set (for simplicity we drop for a moment the extra term)

$$-\log \mathcal{L}(\vec{p}) = \sum_{k} F(\vec{x}_k; \vec{p}) \tag{1.9}$$

with respect to the model's parameters. The RooFitTools package uses the MINUIT[**?**] algorithms to find the minimum of this function and estimate the errors in each parameter. To increse the chances of proper convergence, it is important to provide reasonable initial estimates for the parameters to be fitted.

Most realistic data description models are sum of multiple components. Mathematically, the sum of two probability density functions is also a normalized probability density function as long as the coefficients add up to 1,

$$M(x) = f_{sig} \cdot S(x) + (1 - f_{sig}) \cdot B(x), \tag{1.10}$$

or generically for N components:

$$S(x) = c_0 \cdot F_0(x) + c_1 \cdot F_1(x) + ... + c_{n-1} F_{n-1}(x) + (1 - \sum_{i=1}^{n-1} c_i) F_n(x) \tag{1.11}$$

If the sum of these coefficients becomes larger than one, the remainder coefficient will be assigned a negative fraction. As long as the summed p.d.f is greater than zero everywhere, this is not ill-defined.

For the extended fit (for simplicity we take the example of signal and backgrond PDFs),

$$N_{sig} = f_{sig} \cdot N_{exp} \tag{1.12}$$

$$N_{bkg} = (1 - f_{sig}) \cdot N_{exp} \tag{1.13}$$

so that the extended ML procedure estimates the number of singal and background events rather than a signal fraction and a total number of events.

1.2 Fitting MC templates to data

The likelihood Monte Carlo templates were derived from the simulated dijet sample described in Section ??, from all jets passing the selection criteria defined in Section ??. Templates of likelihood were constructed for b, c, $b\bar{b}$, $c\bar{c}$ and light flavoured MV1 tagged jets separately, and these were fit to the likelihood distribution in data in order to obtain the fractions of single b, merged b, single c, merged c and light jets in the data sample. Merged c-jets (single c-jets) are defined as those jets matching exactly two (only one) "D" hadrons, the products of the fragmentation of c-quarks. A jet is classified as light when it has no B nor D hadron within a cone of 0.4 around its axis.

The likelihood template fits are performed using the unbinned maximum likelihood technique, in its extended version (see Section 1.1). A separate fit is carried out for each p_T bin. Different combinations of templates ("models" in the following) were used to fit the likelihood distribution in data. The first implemented model uses all five templates,

$$F(x) = n_s \cdot S(x) + n_m \cdot M(x) + n_l \cdot L(x) + n_{sc} \cdot S_c(x) + n_{mc} \cdot M_c(x) \tag{1.14}$$

with $S(x)$, $M(x)$, $L(x)$, $S_c(x)$ and $M_c(x)$ the likelihood PDFs for the different hypothesis; and n_s, n_m, n_l, n_{sc} and n_{mc} the free parameters representing the number of expected events for all components: single b, merged b, light, single c and merged c-jets, respectively. The initial estimates for the parameters were obtained from the PYTHIA Monte Carlo sample. The fractions derived in PYTHIA as a function of the jet p_T are shown in Fig. 1.1.

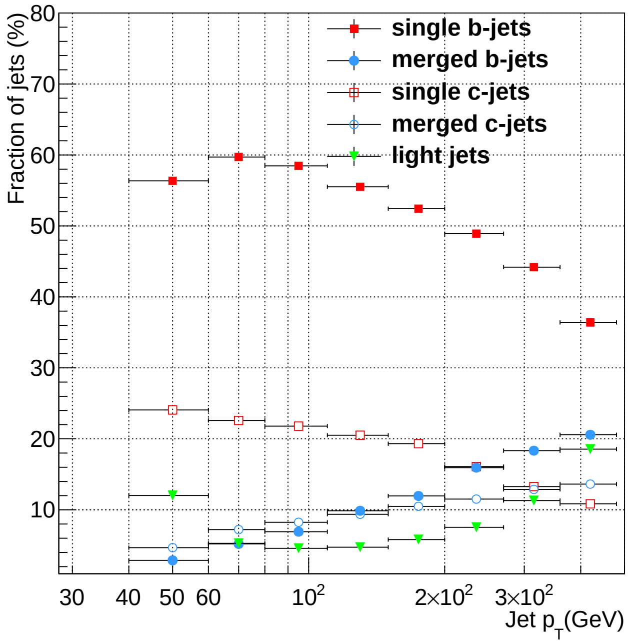

Figure 1.1: Pythia predictions of the fractions of MV1 tagged b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets in a Monte Carlo dijet sample.

The sensitivity of the fit result to fixing the ratio of the single c (merged c) fraction, for each p_T bin, and the single b (merged b) one to the value extracted from the simulation was investigated by carrying out separate fits with a model with three free parameters only. 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- and c-flavour fractions, caused by the high correlations between these components.

The results of the template fits to the likelihood distribution in data, using the three-parameter model, are shown in table 1.1. Examples of this set of fits are displayed in Figures 1.2 and 1.3.

Jet p_T	single b-jet		merged b-jet		light jet	
(GeV)	n_s	stat.err.	n_m	stat.err.	n_l	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 1.1: Measured fractions of single, merged and light b-tagged jets in experimental data from 2011 run.

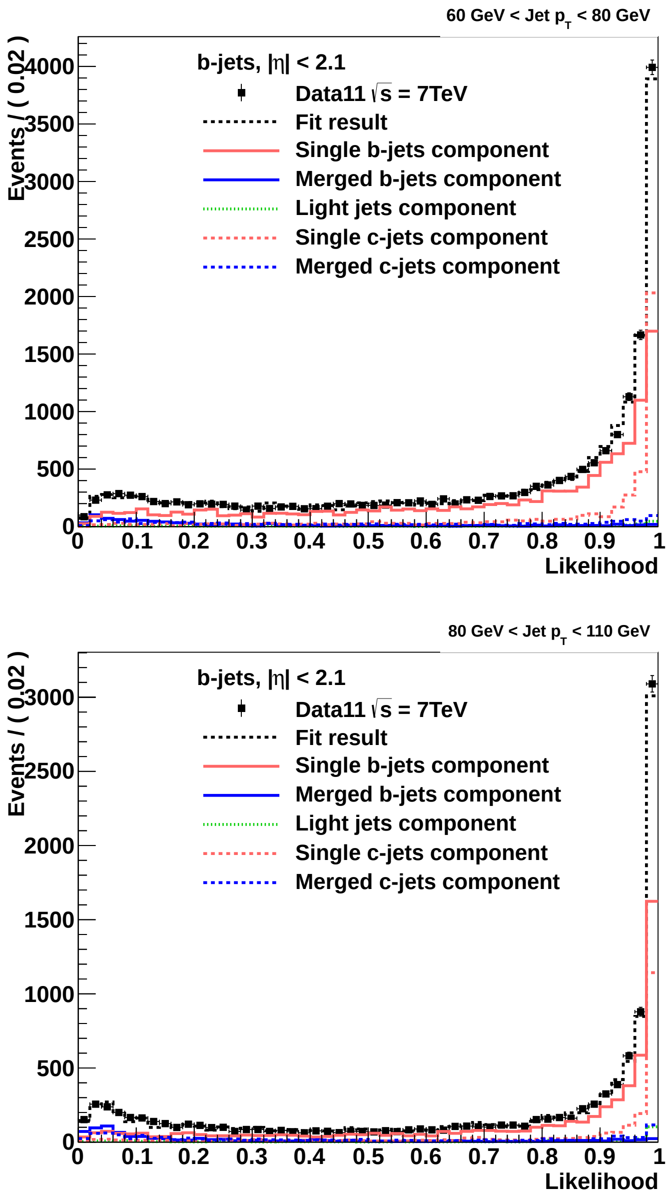

Figure 1.2: The results of template fits to the likelihood distribution in data. The fits shown here were performed on jets with p_T between 60 GeV and 80 GeV, and 80 GeV and 110 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

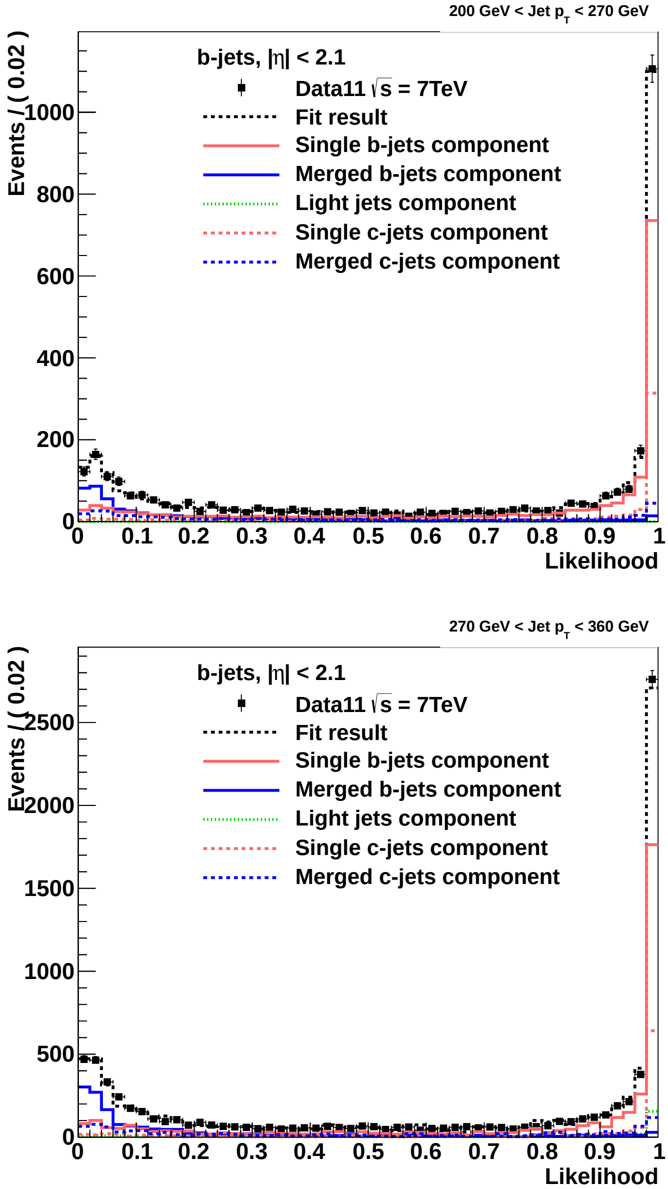

Figure 1.3: The results of template fits to the likelihood distribution in data. The fits shown here were performed on jets with p_T between 200 GeV and 270 GeV, and 270 GeV and 360 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

1.3 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
- uncertainty in the jet energy scale.
- uncertainty in the heavy flavor fraction

In order to calculate the contribution to the total systematic uncertainty from the the uncertainty in the track reconstruction efficiency the procedure described in Section **??** was followed. New likelihood tamplates were produced and new fits performed.

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.

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

The impact of the uncertainty in the knowledge of the flavour fractions in the simulation was evaluated by changing the ratio of merged c to merged b fraction 20%. This variation only produced a marginal effect on the fit results. The total number of merged c plus merged b did not change showing, that, although a separate value for the b- and c-flavoured components can be obtained, we are, in reality, 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 1.2. 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%
heavy flavour fraction	negligible

Table 1.2: Systematic uncertainties.

1.4 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 1.3 to 1.5. 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 1.4 and Figures 1.5.

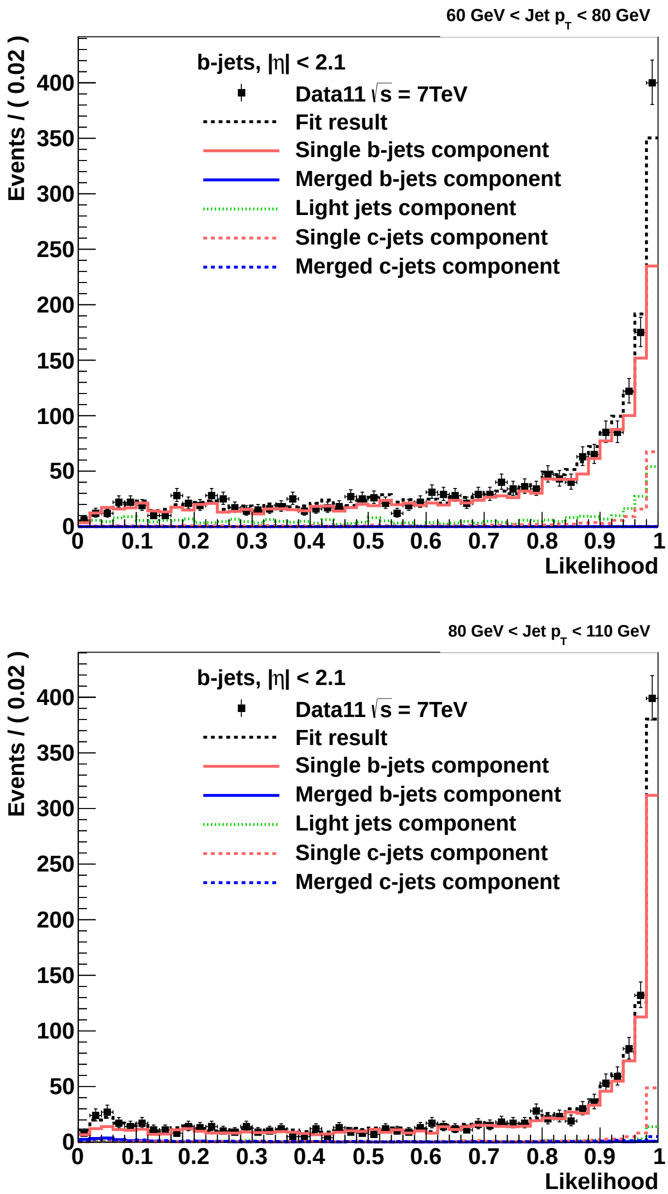

Figure 1.4: The results of template fits to the likelihood distribution in data enriched in single b-jets. The fits shown here were performed on jets with p_T between 60 GeV and 80 GeV, and 80 GeV and 110 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

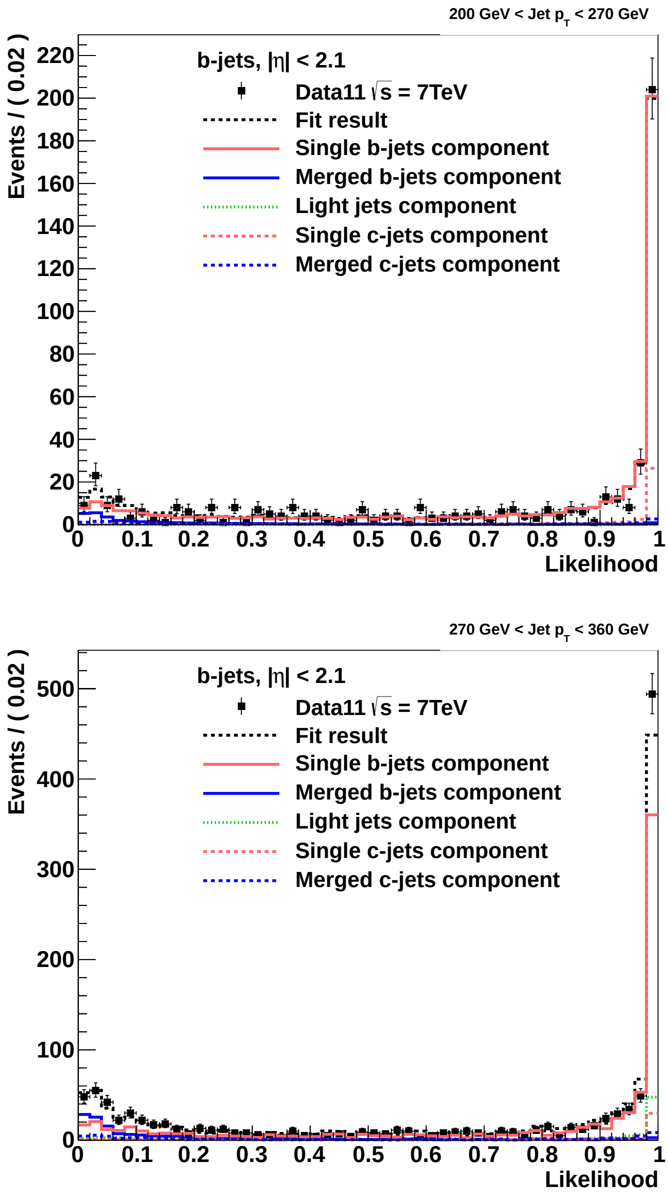

Figure 1.5: The results of template fits to the likelihood distribution in data enriched in single b-jets. The fits shown here were performed on jets with p_T between 200 GeV and 270 GeV, and 270 GeV and 360 GeV, using five templates of b-, $b\bar{b}$-, c-, $c\bar{c}$, and light jets. The ratio of the c- to b-flavour fractions was fixed to the values observed in the simulation. Uncertainties shown are for data statistics only.

Jet p_T (GeV)	single b-jet		
	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 1.3: Measured fractions of single b-jets in experimental data from 2011 run, enriched in single b-jets.

Jet p_T (GeV)	merged b-jet		
	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 1.4: Measured fractions of merged b-jets in experimental data from 2011 run, enriched in single b-jets.

Jet p_T (GeV)	light b-jet		
	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 1.5: Measured fractions of light b-jets in experimental data from 2011 run, enriched in single b-jets.

Bibliography

[1] Roger Barlow and Christine Beeston. Fitting using finite monte carlo samples. *Computer Physics Communications*, 77(2):219 – 228, 1993.

[2] W Verkerke and D. Kirby. Roofit users manual v2.07. 2006.